TesseractAcademy

Website: The Tesseract Academy

Testing Land Valuation Methodologies

Lot 1: Market-based statistical valuation

Lot 2: Advanced algorithmic and machine-learning applications

Lot 3: Formula-based valuation by land area

Lot 4: Conventional valuation approaches

Lot 5: Innovative or experimental approaches

Executive Summary

Introduction and background

This summary presents the findings from a comprehensive evaluation of five land-valuation methodologies applied to the Welsh context. This analysis utilises a large-scale land dataset comprising approximately one and a half million land-use classified parcels, of which 85% are residential. Residential stock includes Detached, Semi-Detached, Terraced, and Flats, which collectively provide a dense and methodologically tractable baseline for testing valuation methodologies.

Beyond residential land, the dataset contains a diverse set of non-residential land-use categories:

- Agricultural and natural land (13.2%), including:
    - Farmland – 114,459 parcels
    - Meadow – 72,163 parcels
    - Forest/Woodland – 2,870 parcels
    - Grassland – 1,467 parcels
    - Farmyard sites- 1,144
- Commercial and industrial land, including:
    - Retail – 6,737 parcels
    - Industrial – 2,277 parcels
    - Commercial (other) – 2,001 parcels

The methodological focus on the residential subset serves primarily as a baseline. Residential land is methodologically more tractable for valuation, with standardised attributes, dense market comparables, and relatively homogeneous characteristics, thereby providing a controlled environment for testing the relative performance, scalability, interpretability, and policy relevance of each approach. Crucially, the purpose of this analysis is not to limit valuation to residential land, but to assess how well these methodologies can generalise across the full spectrum of land types. The framework already incorporates satellite-derived parcel area data and land-use classification features, enabling preliminary decomposition of land values for commercial, industrial, agricultural, and mixed-use categories. The residential results therefore function as a methodological benchmark to understand where each method adds value, where it encounters limitations, and what adaptations would be required to extend the models across the diverse non-residential portfolio.

Land transactions with large parcels exceeding 0.5 hectares, present particular challenges for land-value decomposition, requiring specialised treatment of economies of scale, zoning constraints, and use-specific depreciation schedules. These dimensions are already partially captured through satellite-derived features but require further calibration for operational deployment across all land categories.

The study compares performance, scalability, interpretability, and policy relevance across a diverse set of methodological families:

- Traditional statistical models (Model 1)

    - Ridge Regression
- Machine Learning models (Model 2)
    - CatBoost
    - Gradient Boosting
- Expert-driven systems (Model 3)
    - K-Nearest Neighbours
    - “Fuzzy Matching”
- Formula-based economic methods (Model 4)
    - Depreciated Replacement Cost
- Advanced Agentic AI techniques (Model 5)
    - Claude Multi-Agent Valuation System

The overarching aim is to identify a valuation approach capable of supporting Wales-wide, repeatable, and defensible land-value estimation at granular geographic levels, that is, Lower Layer Super Output Area (LSOA) and below.

Key Findings

Across all modelling approaches, performance was assessed using a test set of 9,606 residential land transactions drawn from nine Lower Layer Super Output Areas (LSOAs) across Wales. The five methodological families evaluated were: Model 1 (Stacked Ridge Regression ensemble), Model 2 (CatBoost Gradient Boosting), Model 3 (KNN comparable-sales automation / “fuzzy matching”), Model 4 (Depreciated Replacement Cost, DRC), and Model 5 (Multi-agent LLM ensemble). Among these, Model 2 (Gradient Boosting) achieved the strongest overall predictive results. However, its current accuracy remains below the level required for operational deployment in statutory or regulatory valuation settings.

Model 1 – Stacked Ridge Regression ensemble (market-based statistical approach) performed substantially better than a simple global Ridge baseline, demonstrating the value of a structured statistical ensemble even when predictive performance remained below operational requirements. Within the nine test LSOAs, within-LSOA R^2 ranged from 1.7% to 51.5% (average within-LSOA R^2 of 26.1%), with a mean absolute error of £69,646. The stacking architecture (5 property-type-specific models, 1 global model, and a meta-learner) showed clear benefits from model specialisation: the meta-learner assigned weights of approximately 1.05 to each property-type model while negatively weighting the global baseline (-0.025). Performance was strongest in rural and semi-rural markets (Powys: 51.5%, Bridgend: 46.3%, Monmouthshire: 42.8%) but much weaker in the large, heterogeneous urban market of Cardiff (R^2 = 1.7%). Model 1’s key strength remains interpretability and auditability, although its linear structure limits its ability to capture the non-linear interactions better handled by Model 2.

Model 2 – Gradient Boosting (CatBoost; machine-learning approach) delivered the best overall performance across the five model families. In its current configuration, it exhibited R^2 values ranging from 0.002 to 0.52 across the nine test LSOAs. In practical terms, this indicates moderate but still insufficient accuracy for operational valuation use, with substantial performance variation across different geographic contexts.

Model 3 – KNN comparable-sales automation ("fuzzy matching"; conventional comparables approach) did not meet the accuracy thresholds required for operational use. This model, which automates a nearest-neighbour comparables logic, performed substantially below the statistical and machine-learning approaches (Models 1 and 2). Prediction errors were typically large (often exceeding half of the land value), the method struggled to generalise across held-out LSOAs, and it produced unstable land-structure splits. These characteristics mean Model 3 is not currently reliable enough for valuation decision-making in its present form.

Model 4 – Depreciated Replacement Cost (DRC; formula-based approach) remained important from a methodological and policy perspective because it is the only fully transparent, rule-based approach in the study capable of explicitly decomposing total value into land and structure components. However, Model 4 (DRC) did not achieve predictive accuracy comparable to the data-driven models (Models 1 and 2). Across the test set, its explanatory power was negative (i.e., worse than a mean predictor), and its typical percentage error exceeded 100%. Although unsuitable as a standalone predictive valuation model, Model 4 still provides a useful conceptual and regulatory benchmark due to its interpretability and auditability.

Model 5 – Multi-agent LLM ensemble (innovative/experimental approach) produced encouraging but inconsistent results relative to the leading machine-learning model. Across the nine test LSOAs, Model 5 recorded R^2 values ranging from -0.02 to 0.46. This suggests the method has real capacity to interpret qualitative and contextual information, but its predictive accuracy was less stable and less reliable overall than Model 2 (Gradient Boosting).

When comparing performance patterns across geography and price levels, the same broad pattern appeared across all five models (Models 1–5). The strongest results tended to occur in areas with higher transaction density and more homogeneous land/property characteristics, while weaker performance was observed in areas with sparse data, greater heterogeneity, or atypical local market structure. High-value transactions and unique assets were consistently challenging for every modelling family, including the best-performing Model 2.

Taken together, the evaluation shows that Model 2 (Gradient Boosting / CatBoost) is the strongest candidate among the five approaches tested for empirical prediction, while Model 1 (Stacked Ridge) offers a valuable interpretable benchmark, Model 4 (DRC) remains the key transparent decomposition method, and Models 3 and 5 are currently better viewed as exploratory or supporting approaches rather than deployment-ready solutions. Even so, none of the models has yet reached the accuracy and consistency required for national deployment in formal valuation or taxation contexts. Improving performance in rural, high-value, and atypical market segments should be a priority in any future development phase.

Land value as a proportion of total land value

Using the Depreciated Replacement Cost (DRC) approach (Model 4) to separate total land value into land and structure components offers a transparent, rule-based framework that is fundamentally different from statistical and machine-learning techniques. DRC remains the only method in this study capable of producing an explicit, theoretically grounded land

structure decomposition. However, when applied to the Welsh housing stock, its empirical performance demonstrates that it cannot be used as a standalone approach for estimating market values.

Across the nine test LSOAs, the DRC model exhibited R^2 values ranging from -0.79 to -0.01, indicating performance consistently below that of a simple mean predictor. The mean percentage error exceeded 100%, demonstrating that typical predictions deviated from actual prices by more than the land's full value. This pattern is consistent with the broader finding that the DRC formula does not adequately capture the heterogeneity of housing conditions, market demand, or localised price levels.

Because of these limitations, the results do not allow for a reliable or meaningful estimate of market values based on DRC alone. The overall predictive performance shows that the DRC method is not suited to estimating market-consistent land values in Wales.

Nevertheless, the method retains important value within a wider analytical framework. In particular, it provides a transparent conceptual basis for understanding how land and structures contribute to total land value, supporting auditability and interpretability in a way that complements empirical approaches. The DRC method can give a broad indication of relative land value patterns across different settlement types, particularly when examined alongside spatial variation in building age, condition, and market activity.

These strengths become clearer when considered alongside the empirical modelling results. Gradient Boosting (Model 2), with R^2 values ranging from 0.002 to 0.52 across the nine test LSOAs, provides the most reliable evidence for total market value in this study. While its accuracy remains insufficient for operational deployment, it offers a substantially more faithful reflection of market outcomes than the DRC formula. By combining the DRC method's transparent land structure decomposition with the empirical estimates generated by Gradient Boosting or other machine learning models, there is scope for developing hybrid approaches that balance interpretability with improved predictive accuracy.

For any potential policy use, the DRC method would therefore require careful handling, including reporting of both raw and adjusted statistics for transparency, and restricting its use to decomposition ratios rather than absolute valuations. Under these conditions, DRC can support policy development, internal modelling, and long-term monitoring, while total valuation exercises should continue to be grounded primarily in empirical models such as Gradient Boosting.

National scalability and geographic coverage

The modelling framework is able to generate land-level valuations for the land-use classified parcels contained in the dataset. Residential land transactions make up the majority of records, with 1,250,600 parcels (85.8%) classified as detached, semi-detached, terraced or flats. This extensive residential coverage provides the core empirical foundation for model development, reflecting both the composition of the Welsh housing market and the structure of the underlying transaction and spatial datasets. The density and geographic breadth of residential observations offer a robust basis for training, validating and comparing valuation methodologies

Beyond the residential sector, the dataset contains a diverse set of non-residential land-use categories which together account for the remaining share of Welsh land parcels. A substantial proportion of these fall within agricultural and natural land uses (13.2%), including 114,459 parcels of farmland, 72,163 parcels of meadow, 2,870 parcels of woodland, 1,467 grassland parcels, and 1,144 farmyard sites. These land types tend to exhibit markedly different spatial and economic characteristics from residential land, including substantially larger parcel sizes and distinctive market dynamics driven by agricultural suitability, environmental context and land management practices.

A smaller but economically significant segment, approximately 0.8% of all parcels, is composed of commercial and industrial land uses, incorporating 6,737 retail sites, 2,277 industrial premises, and 2,001 commercial land transactions. These assets are predominantly concentrated in towns and cities, where zoning, accessibility and local economic activity exert a greater influence on land values than physical characteristics alone.

Building floor area, derived from Energy Performance Certificate (EPC) records, provides an important source of information for understanding the relationship between building size and land value. The current framework therefore relies on floor area, which reflects the internal floor space of the building itself rather than the extent of the land parcel on which it sits.

Directly linked EPC floor area is 37.8% across the full dataset. Among land transactions for which floor area data are available, building sizes display remarkably similar distributions across land-use categories. Both residential and non-residential land transactions have a median floor area of around 86 square metres, with mean values of approximately 94 square metres. This lack of differentiation is a consequence of the underlying dataset rather than a reflection of the built environment itself. A large share of land transactions classified as “non-residential” by land-use context are, in fact, residential dwellings situated within agricultural, commercial or industrial zones. By contrast, genuinely large commercial or industrial buildings are extremely rare in the dataset: only 26 non-residential land transactions have a floor area greater than 500 square metres. Overall floor area ranges from very small structures of 1 square metre to a maximum of 8,412 square metres, although the distribution is highly concentrated around typical residential scales, with an interquartile range of roughly 71 to 107 square metres.

Within the modelling framework, floor area plays a significant role in estimating structure value, particularly for methods such as the Depreciated Replacement Cost (DRC) approach (Model 4) and for residual-based analyses that infer land value after accounting for building characteristics. However, the limited availability of floor area data constrains its use for comprehensive land value decomposition and highlights the need for improved data coverage if the framework is to support future land valuation work at national scale.

To assess the spatial consistency of model performance, the analysis makes extensive use of Lower Layer Super Output Areas (LSOAs). The dataset spans 1,912 LSOAs with at least one transaction in the final dataset, with a median of 724 land transactions per LSOA, allowing for robust spatial validation across urban, suburban and rural contexts. This geographic granularity is particularly important when assessing the applicability of the modelling framework to non-residential land uses, which are unevenly distributed and often

concentrated in specific industrial, agricultural or commercial zones. LSOA-level metrics therefore offer critical insight into whether models trained primarily on residential data are capable of generalising to the distinct spatial and economic environments in which non-residential land values are formed.

A key component of the spatial analysis is the incorporation of Agricultural Land Classification (ALC) data, which intersects approximately 55.2% of land transactions in the dataset. However, 88.1% of matched records are classified as Grade U (Urban), leaving only 6.6% of all land transactions with agricultural grades one to five. These graded transactions remain important for understanding rural land values and land-use patterns and provide an evidence base for modelling non-urban land markets.

This multi-source pipeline is designed to be repeatable and updatable. As new transaction records, EPC certificates and spatial datasets are released, the framework can be refreshed automatically, ensuring that land and land valuations remain current and that the system can be scaled to support long-term monitoring, strategic planning and policy development across Wales.

Contents

List of tables

List of figures

Glossary

Agricultural Land Classification (ALC): A grading system for agricultural land quality (Grades 1–5, plus Urban/Unclassified), sometimes used to distinguish agricultural value potential; coverage can be incomplete depending on available datasets.

Automated Valuation Model (AVM): Any model that estimates property value automatically from data (statistical, machine learning, rules-based), rather than by a human valuer.

Bias (systematic error): A consistent tendency to overvalue or undervalue certain areas or property types.

Building-only model: A model configured to estimate structure value only (using physical attributes while excluding location features) so land value can be inferred via decomposition.

CatBoost: A gradient-boosted decision tree model used to predict property values and capture non-linear relationships.

Cadastral or parcel data: Boundary and attribute data describing land parcels; important for robust plot-size measures and land value per square metre.

Comparable sales (comps): A valuation approach based on prices of similar recent transactions.

Depreciated Replacement Cost (DRC): A formula-based approach that decomposes total value into estimated land value plus depreciated structure value, using cost and depreciation assumptions.

Distribution shift: When a model is applied to places or situations that differ from what it learned during training, which can reduce accuracy.

EPC (Energy Performance Certificate): A certificate providing energy and efficiency information; used here mainly as a source for estimating floor area, but it does not reliably capture internal condition.

Feature importance: A method for estimating which inputs (such as location or floor area) most influence model predictions.

Floor area: Internal size of a dwelling in square metres; a key driver of value, sometimes missing and therefore proxied or imputed.

Fuzzy logic: A rule-based method using “membership functions” (for example, “good location” or “large property”) to encode valuation heuristics.

Geographic holdout: A validation design where whole areas are excluded from training to test how well a model generalises to unseen geographies.

Gradient boosting: A machine-learning approach that builds an ensemble of decision trees sequentially to reduce prediction error.

HMO (House in Multiple Occupation): A property occupied by multiple households or renters; valuation can differ materially from standard residential use.

Information gain: A feature-importance concept measuring how much a feature improves model decisions by reducing error.

K-Nearest Neighbours (KNN): A method that predicts a value by averaging outcomes of the most similar past transactions; sensitive to geography and data coverage.

Land share: The proportion of total property value attributed to land under a given decomposition method (land value divided by total value).

Land Value Tax (LVT): A tax concept where taxation is based on land value rather than combined land and building value; dependent on reliable land valuation.

Land parcel area (plot area): The size of the land parcel associated with a property; missingness in variable limits robust land value per square metre estimates.

Land-structure decomposition: Any method that splits total property value into a land component and a structure component.

LLM ensemble (multi-agent): A large-language-model approach where multiple “role” agents generate valuations and outputs are aggregated.

LSOA (Lower Super Output Area): A small-area statistical geography used in the UK for consistent reporting and comparison.

MAE (Mean Absolute Error): Average absolute difference between predicted and actual prices, expressed in pounds.

MAPE (Mean Absolute Percentage Error): Average percentage error between predicted and actual values, allowing comparison across different price ranges.

Meta-learner: In a stacking ensemble, a second-stage model that learns how to combine predictions from multiple first-stage models.

Negative land value: A mathematically possible but economically implausible outcome where inferred land value is below zero, often caused by residual methods when structure estimates exceed sale price.

Permutation importance: A method that measures feature importance by shuffling one feature and observing how much model performance worsens.

Planning use class: UK planning classification (for example, standard residential versus HMO or mixed-use); missing information limits valuation accuracy for non-standard uses.

Postcode district: A coarse location unit (for example, CF10) used as a geographic proxy but often too broad to capture micro-location effects.

R^2 (coefficient of determination): A measure of how much variance in sale prices is explained by a model; can be negative when predictions are worse than a simple average benchmark.

Residual land valuation (residual method): A decomposition rule where land value equals sale price minus estimated structure value; can produce negative land values when generalisation fails.

Ridge regression: A linear regression model with L2 regularisation used to reduce overfitting; used here as part of an ensemble approach.

Seasonality (month of sale): The idea that prices can vary by time of year; typically a weaker driver than year, location, and property characteristics.

SHAP (Shapley values): A method for attributing how each feature contributes to a model's prediction, often used for interpretability and feature importance.

Tenure (freehold or leasehold): Ownership form that can affect property values, especially for flats and certain urban markets.

Transaction year: The year a sale occurred; often highly influential because it captures broader market conditions and price regimes.

UPRN (Unique Property Reference Number): A unique identifier used to link property records across datasets, supporting enrichment such as matching transactions to EPC floor area.

Within ±20% accuracy: The share of predictions falling within 80% to 120% of the actual sale price, used as an operational accuracy threshold.

1. Introduction

1.1. Expertise and Credentials

Tesseract Academy is a specialist AI and Data Science Research firm. We focus on statistical reports, prioritising investments, building credible business cases, managing risk and governance, and moving initiatives beyond “pilot mode” into operational delivery. We have delivered programmes across government and industry, and our delivery model combines live consulting, practical tools and templates (use-case prioritisation, ROI/value realisation, governance checklists, and implementation roadmaps) and a multidisciplinary team spanning senior AI/data science, programme management, learner support and stakeholder engagement.

Our leadership team brings strong technical credibility alongside proven delivery in real implementation contexts. Dr Kampakis (CEO) has 10+ years’ experience in AI and data science, a PhD in Computer Science (UCL), a postgraduate diploma in Entrepreneurship (Cambridge), and is a Chartered Statistician; his work with The Alan Turing Institute and industry roles (including Satalia & WPP) ensures training remains grounded in practical delivery challenges, complemented by published books, patents in progress, and fellowships. Fabio (Cofounder/Partner) brings specialist expertise in responsible and ethical AI implementation and leadership development, with an MSc in Data Science & AI, delivery roles across the National Digital Twin Programme and BridgeAI, and ongoing service as an ethics reviewer for NeurIPS. Together, this combination of hands-on delivery and responsible-AI governance expertise ensures programmes are both practical and aligned to public-sector expectations around transparency, risk management and accountability.

1.2. Background and Context

Land valuation in Wales presents unique challenges due to:

Geographic Diversity: Wales encompasses dense urban centres such as Cardiff and Swansea, alongside suburban areas, rural communities, and remote regions with highly varied market dynamics.

Bilingual Context: land records and administrative datasets are maintained in both English and Welsh, requiring careful linguistic and metadata harmonisation.

Policy Environment: The Welsh Government’s devolved powers over land-use planning, taxation policy, and housing strategy necessitate locally tailored valuation frameworks.

Data Integration: The analysis combines multiple administrative and geospatial datasets to create a land-level database for Welsh land valuation. Key sources include Land Registry transaction data, Energy Performance Certificates (EPC), Ordnance Survey parcel boundaries, agricultural land classification (ALC), geodemographic and settlement classifications, OpenStreetMap building data, and postcode geographies. Coverage and quality vary by source: Land Registry data is complete but lacks unique land identifiers, land use zoning, and detailed building characteristics. EPCs cover 38% of land transactions, with lower coverage for non-residential and exempt buildings. ALC covers roughly 55.2% of land transactions. Geo-demographic, settlement, and OSM data provide contextual information

but have coarse resolution, outdated references, or sparse rural coverage. Leasehold terms, land condition, renovations, and zoning data are largely missing. These gaps create uneven model reliability. Standard residential land transactions in urban/suburban with EPC and parcel data can be predicted with high accuracy, while flats without parcels, rural residential, and non-residential land transactions show moderate to low reliability.

Land Type: Model performance differs substantially between residential and non-residential land transactions, and this gap becomes especially clear when examining the distribution of prediction errors. The residential sector, despite showing limitations, displays a recognisable and interpretable relationship between modelled and observed prices. In contrast, non-residential land transactions show highly volatile behaviour, with errors that are several times larger and far less predictable. Across the full evaluation, the average percentage error for non-residential land transactions is more than five times higher than for residential ones. This means that, for many non-residential assets, predicted values diverge dramatically from recorded market prices. The median absolute percentage error tells a similar story: the typical error for a residential land is around 40%, whereas the equivalent for non-residential land is more than four times higher. Error distributions further highlight this divergence. Residential land transactions include a sizeable proportion of cases where the model performs well: approximately one-third of residential valuations fall within 20% of the actual sale price, and around 18% achieve very low error levels. By comparison, only 12.5% of non-residential land transactions fall within the same 20% band, and only around 6% achieve errors below 10%.

Key Challenge: Separating land value from total sale price requires decomposition techniques, as transaction data records total sale prices but not the constituent components.

1.3. International Examples

This research draws upon international best practices in land valuation systems:

Table 1: International Best Practices in Land Valuation Systems

Country	**Approach**	**Key Features**
Australia	Land tax on unimproved land values	Residual valuation methodology with regular cadastral reassessments.
Denmark	Comprehensive national land-valuation system	Annual nationwide valuations with transparent and publicly auditable methods.
United States	Automated Valuation Models (AVMs)	Large-scale machine-learning systems (e.g., Zillow, Redfin) trained on millions of transactions.
Singapore	Government-led land-sales and valuation programme	Integration of hedonic pricing models into state-led development planning.

Estonia	Fully digital land-registry and taxation platform	Automated valuations for recurrent land-tax assessments.

More information can be found below:

- Australia - land tax on unimproved land value (NSW Revenue)
- Denmark - national land / land valuation (Danish land Assessment Agency)
- United States - AVMs (federal rule on Automated Valuation Models)
- Singapore - Government Land Sales Programme (Urban Redevelopment Authority)
- Estonia - digital land tax / valuations (Estonian Tax and Customs Board - Land tax)

2. Methodology

This study applied five distinct valuation approaches, ranging from traditional statistical models to advanced machine learning and expert systems, to identify methods suitable for Wales-wide, repeatable, and defensible land-value estimation. The framework was designed to evaluate predictive performance, interpretability, scalability, and the ability of each method to decompose total land value into land and structure components.

2.1. Research Design

This study uses a comparative evaluation design to test five distinct land valuation paradigms under realistic deployment conditions. All methods are applied to a common pre-processed dataset derived from statutory transaction records and are assessed on their ability to generalise to geographic markets not seen during model development.

We implement a strict geographic hold-out: nine Lower Layer Super Output Areas (LSOAs) are excluded from training and reserved as a fully independent test set (n=9,606 transactions). All remaining Welsh transactions form the training pool (≈1.45m records), covering the period 1995–2024. The held-out LSOAs were selected to represent diverse settlement and housing-market contexts (urban/regeneration, rural, coastal, and valley markets), creating a stringent test of robustness under geographic distribution shift.

Model 1 – Stacked Ridge Regression corresponds to Lot 1 (Market-based statistical valuation) as a hedonic/statistical approach grounded in observed transactions.

Model 2 – CatBoost Gradient Boosting corresponds to Lot 2 (Advanced algorithmic and machine-learning applications) as a modern ML method designed to capture non-linear patterns at scale.

Model 3 – KNN Comparable-Sales Automation corresponds to Lot 3 (Conventional valuation approaches) as an automated version of the standard comparables method.

Model 4 – Depreciated Replacement Cost (DRC) corresponds to Lot 4 (Formula-based valuation by land area) as a transparent, rules-based valuation framework operationalised via measurable physical inputs (including land and floor area assumptions and cost and depreciation parameters).

Model 5 – Multi-agent LLM Ensemble corresponds to Lot 5 (Innovative or experimental approaches) as an emerging, exploratory approach tested for feasibility, explainability and robustness

2.2. Rationale for this design

This study compares five distinct valuation families because they represent the full spectrum of approaches reflected in the lots (market-statistical, machine learning, formula-based, conventional comparables, and innovative/experimental methods), each offering different strengths and limitations in accuracy, interpretability, scalability, and the ability to support land/structure decomposition. A key requirement for practical Wales-wide use is geographic generalisation: models must remain reliable when applied to locations not seen during development, where local market conditions, housing stock, and amenities differ, so we adopt an explicit out-of-area (LSOA) hold-out design rather than relying only on random

splits. We focus on residential transactions as the primary benchmark because they provide the most complete and liquid evidence base (highest volume and most consistent attribute coverage), enabling robust method comparison and validation before considering extension to settings where transaction evidence and property attributes are materially sparser.

2.3. Summary of approaches

This study evaluated five distinct approaches across multiple models:

Table 2: Valuation Model Methodologies, R^2 Ranges and Use Cases

Model	**Methodology**	**R^2 range across nine test LSOAs**	**Core Application**
Model 1 - Ridge Regression	Stacked Ridge ensemble with land-type specialisation and HPI-normalised prices (global model + 5 land-type models combined by a meta-learner).	0 to 0.51	Clear and fully interpretable baseline for mass appraisal; well suited to quality assurance, audit and policy review.
Model 2 - Gradient Boosting	Gradient Boosting model (Cat Boost) trained with tuned hyperparameters	0 to 0.52	Strongest empirical performance
Model 3 (KNN)	The fuzzy approach relies on matching each land with similar examples in the dataset to infer value.	N/A	Highly interpretable rule-based system; useful for explanation but not for quantitative valuation.
Model 4 - DRC (Depreciated Replacement Cost)	Cost-based structure estimation with residual land value	N/A	The DRC decomposition implies a median land share of ~40% (typical property), but a higher mean land share (~60%) due to skew/outliers. We therefore treat ~40% as the median headline figure and report the mean as a sensitivity indicator rather than a stable constant.

Model	Methodology	R^2 range across nine test LSOAs	Core Application
Model 5 - Claude Haiku LLM	Large Language Model valuation system using structured prompts and guided price estimation.	0 to 0.46	Captures well qualitative patterns; under performs Model 2

Different accuracy measures capture different aspects of model performance, and the Large Language Model approach illustrates this clearly. When assessed using R^2 as a measure of explained variance across the nine test LSOAs, the model achieved values ranging from 0 to 0.46, demonstrating variable performance across different geographic areas. Its ability to capture qualitative patterns and contextual signals is a notable strength, particularly when dealing with land transactions that have rich descriptive information. However, when assessed using more conventional error metrics, the model's limitations become clearer. Percentage-based errors and absolute prediction errors show that while the model can identify general valuation patterns, it struggles to achieve the consistency required for operational deployment. This difference arises because the model's strength lies in interpreting contextual and qualitative information rather than producing precise numerical estimates: an approach that is valuable for exploratory analysis but insufficient for formal valuation or taxation purposes.

Percentage of Valuations Within ±20% of Actual Price by Model

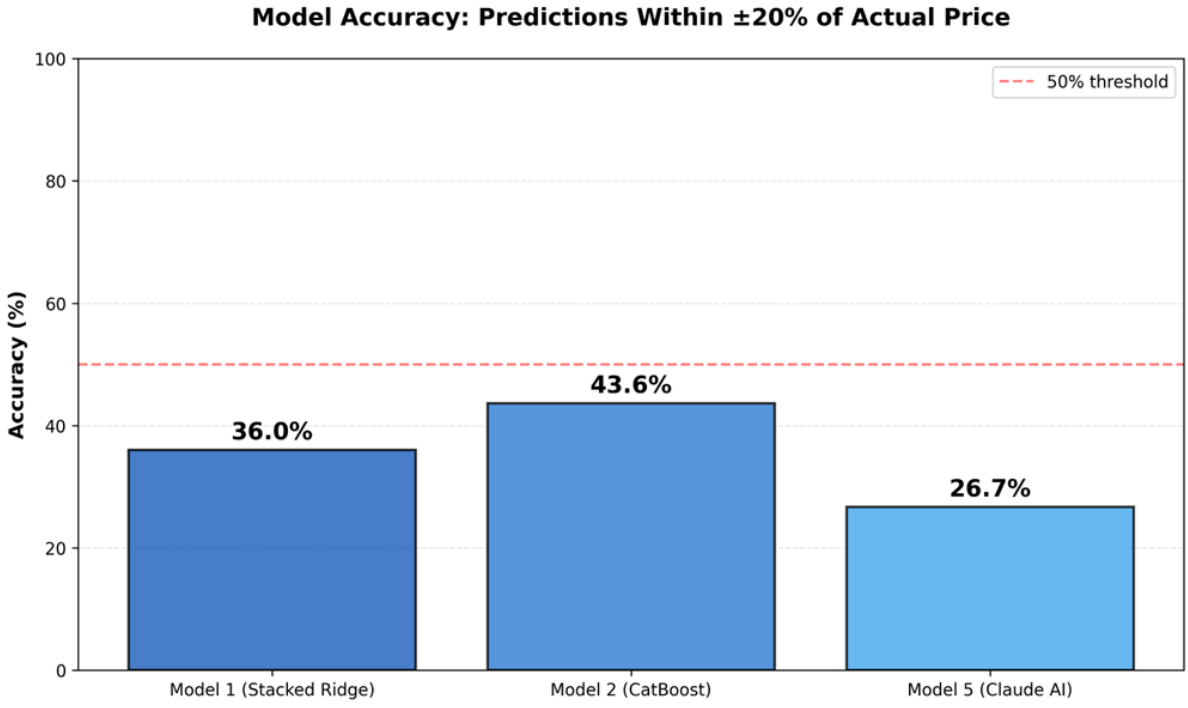


Figure 1: Percentage of Valuations Within ±20% of Actual Price by Model

2.4. Cross-validation

A central objective of this study was to assess how well different valuation approaches can generalise to locations they have never encountered. Cross-validation is a fundamental technique for assessing how well a model will generalise to new data without touching the test set. It operates entirely within the training set by dividing it into multiple subsets called "folds". To ensure a robust and independent evaluation, nine Lower Layer Super Output Areas (LSOAs) were designated as a fully held-out test set. These areas were excluded from every stage of model development, including sampling, feature engineering, hyperparameter tuning, and internal validation. All accuracy figures reported here therefore reflect true out-of-geography generalisation, which mirrors the real-world scenario in which a national valuation system must estimate values in a locality with no recent historical transactions.

Explanation of Key Methods and Concepts

Because each model requires a different development strategy, this section defines the core methodological concepts used throughout the evaluation, with particular emphasis on how cross-validation supports robust model development without compromising the test set.

Training and Test Sets

Training set: Contains all Welsh land transactions outside the nine held-out LSOAs (land transactions). This set is used for:

- Feature engineering and exploration
- Model training
- Hyperparameter tuning via cross-validation
- Internal validation to prevent overfitting

Test set: Contains only land transactions from the nine excluded LSOAs (9,606 land transactions). This set is used solely to assess final model performance after all development decisions are complete. The test set is never used during training, validation, or tuning, ensuring it provides an unbiased estimate of out-of-geography generalisation.

How Cross-Validation Works

Splitting: The training set is divided into k equal parts (typically 5 or 10 folds).

Iterative Training and Validation: The model is trained k times, each time:

- Training on k-1 folds
- Validating on the one remaining fold

Aggregation: Performance metrics (e.g., R^2, MAE) from all k validation folds are averaged to estimate the model's expected performance on unseen data.

Why Cross-Validation Matters

CV serves three critical purposes in this study:

1 Hyperparameter Tuning: Different model configurations (e.g., tree depth, learning rate) are tested using CV. The configuration with the best average CV performance is selected for final training. This prevents overfitting to any single validation split.

2 Overfitting Detection: If a model performs much better on training data than on CV validation folds, it indicates overfitting. CV provides an early warning system before final test set evaluation.

3 Model Selection: When comparing multiple candidate models (e.g., Ridge vs. Lasso, different tree depths), CV provides an unbiased estimate of which approach will generalise best, without requiring any use of the test set.

Crucially, cross-validation is not the final evaluation. It is a development tool. Once hyper parameters are optimised and a final model is selected using CV, the model is retrained on the entire training set and evaluated once on the held-out test set to report final performance.

GroupKFold Cross-Validation

A specialised form of CV used in Model 2 (CatBoost) to simulate geographic generalisation during the training phase itself.

Why GroupKFold Is Necessary

Standard CV randomly assigns individual land transactions to folds. This creates a problem for geographic generalisation: land transactions from the same LSOA may appear in both training and validation folds within a CV iteration. The model can therefore learn LSOA-specific price patterns in training and exploit them during validation, leading to overly optimistic CV scores that do not reflect true out-of-area performance.

How GroupKFold Works

Instead of splitting individual land transactions, GroupKFold keeps entire LSOAs together:

Grouping: All land transactions within a given LSOA are assigned to the same fold.

Fold Assignment: Entire LSOAs (not individual land transactions) are distributed across the k folds.

Training and Validation: In each CV iteration:

- The model trains on land transactions from k-1 groups of LSOAs
- The model validates on land transactions from the remaining group of LSOAs
- Crucially: The model never sees any land transactions from the validation LSOAs during training.

This mirrors the exact challenge of the final test set: predicting land values in geographic areas with no training examples.

Use In This Study

GroupKFold was used during hyperparameter optimisation for Model 2. By validating on entirely held-out LSOAs within the training set, the hyperparameter search selected configurations that generalise across geography, not just across individual land transactions. This ensures that the final model, when applied to the nine test LSOAs, faces a challenge it was explicitly prepared for during development.

Hyperparameters

Hyperparameters are settings that control model behaviour but are not learned from the data. Examples include:

- Tree depth in gradient boosting models
- Learning rate controls how quickly a model adapts
- Regularisation strength penalising model complexity to prevent overfitting

Hyperparameters must be optimised before final training. In this study:

- Model 1 (Ridge Regression): Regularisation strength (alpha) was selected based on CV performance.
- Model 2 (CatBoost): Learning rate, tree depth, L2 regularisation, and other parameters were optimised using Optuna, an automated hyperparameter search library, guided by GroupKFold CV scores.
- Models 3 & 4: Followed similar CV-based optimisation strategies.

All hyperparameter tuning was performed using only the training set. The test set was never consulted during this process.

Meta-Learner

A meta-learner is a secondary model that learns how to combine predictions from multiple base models.

Use in Model 1:

1. Base Models: Five land-type-specific Ridge models (one for each of Detached, Semi-Detached, Terraced, Flat, and Other) are trained separately.
2. Base Predictions: Each base model generates predictions on the training set.
3. Meta-Learner Training: A final Ridge model learns how to weight and combine the base model predictions to improve overall accuracy.
4. Final Prediction: For a new land, all base models make predictions, and the meta-learner combines them into a single final estimate.

This stacking approach allows the model to specialise by land type while learning an optimal strategy for combining these specialised predictions.

In-Context Learning (LLM)

Model 5 does not train on Welsh data. Instead, it uses Claude 3.5 haiku, a large language model with broad pre-trained knowledge of land valuation, geography, and market dynamics.

For each land to be valued:

Prompt Construction: A structured prompt is created containing:

- The target land's features (floor area, land type, location, etc.)
- Comparable sales from nearby land transactions
- Market context

Reasoning: The LLM interprets the prompt and reasons about the land's value using its pre-trained world knowledge and using ridge model predictions as a baseline.

Dynamic Adaptation: Unlike static models, the LLM adapts its reasoning to each land individually, considering unique features and local context described in the prompt.

This approach represents a fundamentally different paradigm, instead of learning statistical patterns from Welsh data, the LLM applies general valuation principles learned during pre-training to the specific Welsh context.

Importance of the Geographic Holdout Design

Holding out entire LSOAs ensures that models cannot rely on memorised price patterns tied to specific neighbourhoods, postcodes, or local market conditions. Instead, each model must generalise from broad structural and spatial relationships learned across Wales. This mirrors the operational requirements of mass appraisal systems, which must value land transactions in areas with limited or no recent sales evidence.

Why This Matters for Real-World Deployment

In practice, a national valuation system must handle:

- New developments in areas with no historical sales
- Rural areas with sparse transaction data
- Rapid market changes where local recent sales may not exist
- Policy decisions requiring valuations for areas undergoing regeneration or development

By testing all models on the same nine geographically diverse LSOAs (spanning rural Ceredigion, urban Cardiff, post-industrial Bridgend, and coastal Pembrokeshire) the study provides a fair, transparent, and policy-relevant comparison of modelling approaches and their potential roles in a future national valuation framework.

2.5. Performance metrics

To evaluate the five land valuation approaches, we use four key metrics. Understanding these metrics is essential for interpreting the results.

R^2 (R-Squared) - Variance Explained

Range: -∞ to 1.0
Best: 1.0 (perfect predictions)
Interpretation:

- $R^2 = 0.70$: Model explains 70% of variance in land prices
- $R^2 = 0$: Model performs no better than predicting the mean
- $R^2 < 0$: Model performs WORSE than simply guessing the average price

Why negative R^2 occurs: When systematic over-prediction or under-prediction causes larger errors than just using the mean.

MAPE (Mean Absolute Percentage Error)

Range: 0% to ∞
Best: 0% (perfect predictions)
Formula: Average of |Predicted - Actual| / Actual × 100%

Interpretation:

- MAPE = 30%: On average, predictions are 30% off from actual prices
- MAPE < 20%: Generally acceptable for land valuation
- MAPE > 50%: Poor performance, not suitable for practical use

Example: land worth £100,000, predicted at £130,000 → APE = 30%

MAE (Mean Absolute Error)

Range: £0 to ∞
Best: £0 (perfect predictions)
Units: British Pounds (£)
Interpretation:

- MAE = £25,000: On average, predictions are £25,000 off from actual prices
- More intuitive than MAPE (absolute £ rather than%)
- Sensitive to land value scale

Example: land worth £150,000, predicted at £175,000 → AE = £25,000

Within ±20% Accuracy

Range: 0% to 100%
Best: 100% (all predictions within ±20%)
Industry benchmark: land transactions predicted within 80-120% of actual value
Interpretation:

- 50% within ±20%: Half of land transactions valued within acceptable range
- Used by valuation standards (RICS) as acceptability threshold
- More practical than R^2 for non-technical stakeholders

Example: £100k land predicted between £80k-£120k → ✓ Acceptable

Table 3: Comparison of Valuation Error Metrics and Their Best Uses

Metric	**Focus**	**Sensitivity**	**Best for**
R^2	Overall fit	Outliers	Model comparison
MAPE	% errors	Relative	Cross-dataset
MAE	£ errors	Absolute	Practical use
±20%	Acceptability	Threshold	Decision-making

A model can have good R^2 but poor ±20% accuracy if it's consistently off by 15-25%. Conversely, a model with lower R^2 but tight clustering near actual values may have better ±20% accuracy.

2.6. Data sources

Core Transaction Dataset and Coverage

All methodologies utilised a common pre-processed dataset:

Table 4: Data Filtering Stages and Final Train Split for Welsh Transaction

Data Stage	Count	Description
Source Data	~30,000,000	UK Land Registry Price Paid Data (1995-present)
Wales Filter	1,457,489	Welsh transactions after postcode/LA filtering
After Price Filter & Test Split	1,447,883 train 9,606 test	Separate 9 held-out LSOAs

The cornerstone of this analysis is the UK Land Registry's Price Paid dataset, a comprehensive statutory register of all residential land transactions in England and Wales since 1995. This dataset provides transaction-level records with legal registration requirements that ensure high baseline data quality.

We applied strict geographic filtering using Welsh postcode prefixes and Local Authority District codes to extract only Welsh transactions. This geographic filter retained 1,457,489 Welsh land transactions from the approximately 30 million UK-wide records, representing approximately 4.9% of the total dataset.

Price Range and Outliers

- Minimum price: £95
- Maximum price: £93,395,000
- Median price: £120,000

The dataset includes:

- 384 land transactions priced below £1,000 (0.03% of records) - These likely represent non-standard transactions such as transfers between family members, shared ownership schemes, or data entry errors.
- 425 land transactions priced above £5,000,000 (0.03% of records) - These represent high-value land transactions, predominantly in Cardiff and coastal areas.

Land Characteristics

Land type distribution:

- Detached (D): 412,142 land transactions (28%)

- Semi-detached (S): 412,959 land transactions (28%)
- Terraced (T): 492,169 land transactions (34%)
- Flat (F): 105,260 land transactions (7%)
- Other (O): 34,959 land transactions (2%)

Tenure:

- Freehold: 1,294,427 land transactions (89%)
- Leasehold: 163,044 land transactions (11%)

New build status:

- Existing land transactions: 1,323,672 (91%)
- New build: 133,817 (9%)

Geographic and Land Use Context

Rural/urban classification:

- Urban: 82% of land transactions
- Rural: 18% of land transactions

Land use distribution:

- Residential: 1,250,600 land transactions (86%)
- Farmland: 114,459 land transactions (8%)
- Meadow: 72,163 land transactions (5%)
- Retail: 6,737 land transactions (0.5%)
- Other land uses: <0.5% each (forest, industrial, commercial, etc.)

This land use data, derived from OpenStreetMap, provides critical context for valuation models, particularly for rural land transactions where agricultural land value influences land prices.

Model-Specific Training Approaches

Models 1 & 2 (Ridge Regression and CatBoost): Trained on the full training set of land transactions to maximise predictive performance. Hyperparameter tuning used a capped sample of up to 500,000 transactions for computational efficiency; final models were trained on the full training set.

Model 3 (K-Nearest Neighbours): KNN has prediction complexity proportional to the number of training examples. For computational feasibility, Model 3 was trained on a stratified sub-sample of the training data, maintaining temporal and geographic balance while enabling practical inference times.

Model 4 (Depreciated Replacement Cost): DRC does not require a training phase. It applies a closed-form valuation formula with parameters calibrated from construction cost indices and depreciation schedules. The full training set was used to calibrate regional and temporal adjustment factors.

Model 5 (Multi-Agent LLM Ensemble, Anchored to Ridge): The LLM ensemble (Claude 3.5 Haiku with four specialist agents: MRICS Valuer, Developer, Environmental Economist, and Community Representative) does not train on Welsh data. Instead, it uses in-context learning, reasoning about property values from structured prompts containing target property features and comparable sales. Crucially, Model 5 is anchored to Model 1's Ridge regression predictions. The LLM receives the Ridge baseline prediction as a reference point and adjusts it based on contextual factors, comparable sales, and domain expertise encoded in the specialist agents. This hybrid approach combines statistical learning (Ridge) with contextual reasoning (LLM). Model 5 was evaluated on all 9,606 test land transactions across the nine held-out LSOAs, the same test set used for Models 1-4, ensuring a fair and direct comparison of all modelling approaches.

Energy Performance Certificate (EPC) Data Linking

The UK Government's Domestic Energy Performance Certificate Register provided building-specific characteristics essential for valuation modelling. In particular, the EPC dataset supplied:

- Floor area - The single most predictive structural feature for land valuation
- Construction age - Required for depreciation calculations and land type inference

However, EPC coverage is incomplete. Only 38% of land transactions have a direct EPC linkage for floor area. We therefore use a two-stage approach: direct EPC linkage where available, and an EPC-derived postcode proxy where direct linkage is not available. After applying the postcode proxy, 74% of transactions have a floor-area value (direct + proxy). The remaining 26% still lack floor area and require separate imputation.

Three-Tier EPC Matching Strategy

To maximise EPC linkage while maintaining data quality, a cascading match hierarchy was implemented:

1. <u>UPRN-Based Match (Tier One, Highest Confidence)</u>
    - Uses the Unique Property Reference Number (UPRN) to establish direct one-to-one linkage
    - Provides the strongest possible match because UPRNs are statutory identifiers assigned by local authorities
    - Coverage: Approximately 32% of land transactions with EPC records
    - Accuracy: >99.9% (UPRNs are designed to be unique and stable)

2. <u>Address-Based Match (Tier Two, High Confidence)</u>
    - Matches on postcode + normalised house number + normalised street name
    - Uses Levenshtein-distance fuzzy matching (<3 character edits) to handle common address variations (e.g., "St" vs "Street", "Rd" vs "Road")
    - Applied where UPRN links are unavailable, but addresses are sufficiently consistent
    - Coverage: Approximately 42% of land transactions with EPC records
    - Accuracy: ~95% based on manual validation of random samples

3. <u>Postcode-Only EPC Proxy (Tier Three, Fallback)</u>

- Applied when neither UPRN nor address matching succeeds
- Imputes floor area and construction age using median EPC values within the same postcode
- Preserves local building-stock characteristics while acknowledging higher uncertainty
- Coverage: Remaining land transactions with EPC records
- Accuracy: Median absolute error of ~15 square metres for floor area and ~10 years for construction age

Missing Floor Area Handling

For transactions without a direct EPC match, floor area is handled in two steps:

- Step 1 (postcode proxy): where EPC records exist in the same postcode, we assign a postcode-median EPC floor area (higher uncertainty than direct linkage).
- Step 2 (non-EPC imputation): for the remaining transactions with no usable EPC-derived floor area even after the proxy step (approximately 26% overall), Models 1–4 use hierarchical median imputation (land type → postcode district → sale year), and Model 5 is explicitly informed when floor area is imputed/unavailable to reflect uncertainty.

Median floor area across all land transactions: 86.0 square metres

This multi-tier approach balances data completeness with quality, ensuring all models can leverage structural land features while maintaining transparency about imputation uncertainty.

Fields Extracted

Table 5: EPC Fields: Usage and Data Quality Implications

EPC Field	**Usage**	**Missing Data Impact**
TOTAL_FLOOR_AREA	Primary size metric for all models; critical for DRC structure value	Imputed via hierarchical median (land type → postcode → year), introducing ~15% uncertainty
CONSTRUCTION_AGE_BAND	DRC depreciation calculation, age features for ML models	25.9% imputed via median by postcode, causing DRC to treat old land transactions sold recently as near new
CURRENT_ENERGY_RATING	Energy efficiency bands A-G, quality proxy	Not used in primary models (low coverage, high collinearity with age/type)

EPC Field	Usage	Missing Data Impact
Land_TYPE_EPC	Cross-validation with Land Registry type	Discrepancies flagged for manual review (3.2% mismatch rate)
BUILT_FORM	Detached/Semi-Detached/Terraced/Flat classification	Higher granularity than Land Registry, used for validation
NUMBER_HABITABLE_ROOMS	Room count density proxy	Too sparse for modelling, excluded from features
LODGEMENT_DATE	Resolves multiple EPCs per land (latest selected)	Land transactions with multiple EPCs may have had extensions/renovations between certificates

EPC Cleansing Actions

Impossible build year removal: Filtered out CONSTRUCTION_AGE_BAND values implying construction before 1600 or after 2025 (0.4% of EPC records)

Construction year imputation: For the 26% without construction age:

- Tier 1: Median by land type + postcode district (e.g., median detached house age in CF14)
- Tier 2: Median by land type + Local Authority (LA) (if postcode district has <10 samples)
- Tier 3: Median by land type nationally (if LA has <10 samples)

Energy rating normalisation: Standardised all ratings to A-G scale (some older EPCs used numeric 1-100 scale, converted via official SAP lookup table)

Floor area imputation: Applied hierarchical median imputation for the 26% of land transactions without measured floor area:

- Tier 1: Median by land type + postcode district + new-build flag
- Tier 2: Median by land type + postcode district (if Tier 1 <10 samples)
- Tier 3: Median by land type + Local Authority
- Tier 4: National median by land type
- Retained area_imputed_flag feature to allow models to down-weight imputed values

Critical Limitation for Land Decomposition

The 26% gap in measured floor area creates a significant limitation for Model 4 (Depreciated Replacement Cost) when estimating the land-structure split. The DRC formula calculates structure value as:

Structure Value = Floor Area multiplied by Construction Cost per square metre multiplied by (one minus Depreciation Rate) raised to the power of Age

When floor area is imputed rather than directly measured, the resulting structure value carries approximately 15% uncertainty. This uncertainty directly propagates into the derived land value, which is calculated as the difference between market price and estimated structure value.

This helps explain why a raw (unconstrained) DRC residual land estimate can become negative for approximately 35% of transactions in the Wales-wide dataset. This All-Wales diagnostic is not directly comparable to the 9-LSOA decomposition table later in the report, which reports a non-negative land component for interpretability:

- For smaller land transactions, imputed floor area is often too high, inflating structure value and pushing land value into negative territory.
- For larger land transactions, imputed floor area may be too low, depressing structure value and distorting the land residual.

These systematic biases compromise the reliability of the DRC land-structure decomposition when measured floor area is unavailable.

Selection Bias in EPC Coverage

Land transactions with measured floor area are systematically different from those without EPC measurements (approximately 26%):

Table 6: EPC Characteristics with & without Measured Floor Area

Characteristic	With Measured Floor Area	Without (Imputed)
Median sale year	2015	2003
Median price	£122,000	£98,000
Land type (Detached)	23.1%	28.4%
Land type (Flat)	11.2%	15.7%
Land type (Terraced)	34.8%	29.3%
New build	5.8%	2.1%
Mean floor area	89.2 square metres	84.1 square metres (imputed)

Land transactions without measured floor area tend to be systematically different from those with full EPC coverage. On average, these land transactions are older by approximately 12 years, cheaper by roughly 20%, and more likely to be flats or older detached houses that have not transacted since EPC requirements were introduced in 2008.

These differences create two important implications:

Model performance risks

Models trained primarily on land transactions with EPC-linked floor area may underperform on the older segment of the housing stock, because the structural characteristics of these homes are less well represented in the training data.

Higher uncertainty in land-value estimates

For pre-2008 land transactions, where measured floor area is unavailable and imputation is required, land-value estimates carry inherently greater uncertainty. This uncertainty directly affects the accuracy of land-structure decomposition, particularly in methods such as Depreciated Replacement Cost that rely heavily on floor area as a core input.

ONS Postcode Directory: Spatial Foundation

The Office for National Statistics Postcode Directory provided comprehensive spatial and administrative geography linkages for all Welsh land transactions. This enabled geographic modelling, regional aggregation and the stratification of the test set by Lower Layer Super Output Area.

Coverage

Only 39 land transactions, equivalent to 0.01%, could not be geocoded. In all cases, the failures arose because the associated postcodes had been abolished between the date of land sale and the 2024 ONS Postcode Directory snapshot.

Fields Provided:

Table 7: ONSPD Fields: Coverage, Usage and Description

ONSPD Field	Coverage	Usage	Description
Latitude/Longitude	99.99%	Distance calculations (Cardiff, transport, POIs), spatial joins	Postcode centroid coordinates (±50m accuracy, not exact land location)
LSOA21 codes	100%	Geographic holdout test set, regional analysis	Lower Super Output Areas in Wales
MSOA11 codes	100%	Middle-layer geographic aggregation	410 Middle Super Output Areas in Wales

ONSPD Field	Coverage	Usage	Description
Local Authority codes	100%	Administrative boundary analysis	22 Welsh Local Authorities plus 5 English border authorities
Rural-Urban Classification	100%	Settlement type features	ONS 2011 Rural-Urban Classification (Urban/Town/Village/Hamlet/Isolated)
Output Area Classification (OAC)	100%	Socioeconomic features	2011 Census-based demographic classification (8 Supergroups, 26 Groups, 76 Subgroups)

Using the geographic coordinates provided by the Office for National Statistics Postcode Directory, a set of derived distance features was constructed for all Welsh land transactions. These were calculated using the Haversine formula, which measures great-circle distances that account for the curvature of the Earth. These distance metrics were used to support geographic analysis and to test the sensitivity of valuation models to locational accessibility.

Distances were calculated to the following points of interest:

1. Cardiff city centre**:** Latitude (51.4816), longitude (-3.1719). Used as the primary measure of proximity to Wales's main urban and economic centre.
2. Nearest railway station**:** Identified using railway nodes from OpenStreetMap. A total of 2,487 stations across Wales and the English border regions were used.
3. Major roads**:** Including the M-4, A-55, A-470 and other major A-roads derived from OpenStreetMap road network data**.**
4. Accessibility features**:** Distances to schools, hospitals, supermarkets and town centres based on OpenStreetMap points of interest.

Coordinate Imputation

For the 39 land transactions where geographic coordinates were not available from the Office for National Statistics Postcode Directory, a structured imputation method was applied:

- Nearest postcode centroid within the same postcode district was used as the first fallback option. For example, if CF14 1 could not be geocoded, the centroid of CF14 1AA was applied.
- Local Authority centroid was used only if the postcode district centroid was unavailable.
- An imputation flag was retained for all cases to support ongoing data-quality tracking and transparency.

Temporal Mismatch Limitation

The Office for National Statistics Postcode Directory provides a 2024 snapshot of postcode coordinates and geographic classifications. These classifications were applied to transactions spanning 1995 to 2024. This introduces several forms of temporal mismatch:

- LSOA boundary changes**:** Lower Layer Super Output Area boundaries were redrawn following the 2011 Census. land transactions sold between 1995 and 2010 are retrospectively assigned 2011 boundaries, which may differ from the boundaries in place at the time of sale.
- Rural-urban reclassification: Some areas experienced changes in settlement classification between the 2001 and 2011 Censuses. For example, parts of Cardiff Bay were reclassified from “Urban Minor Conurbation” to “Urban Major Conurbation.” Historical transactions prior to redevelopment are therefore assigned settlement types reflecting later urban growth.
- New postcode creation**:** Postcodes created after 2011 to serve new developments are backfilled for historical sales occurring in the same area. As a result, older transactions may appear to have postcodes that did not exist at the time.

These temporal mismatches mainly affect areas that experienced significant regeneration or development, such as Cardiff Bay, Swansea Maritime Quarter and Newport Riverfront. For these locations, older sales (from 1995 to 2010) may be associated with post-development geographic characteristics. This can lead models to overpredict historical values because they use present-day settlement classifications rather than those in place at the time of sale.

Satellite-Derived and OpenStreetMap Features: Accessibility Metrics

Geospatial features were integrated to capture location quality, accessibility and neighbourhood characteristics beyond simple geographic coordinates. These features quantify the hedonic value of place using measurable proxies, including distance to transport infrastructure, amenity density and land-use classification. They support more nuanced modelling of how location contributes to land value.

Data Sources

1. INSPIRE Land Parcels (Welsh Government, 2023)
    - Provides cadastral parcel boundaries for approximately 55% of Welsh land transactions.
    - Supplies parcel area in square metres, derived from satellite imagery.
    - This measurement is essential for calculating land value per square metre, a key input for land-value taxation and decomposition analysis.
    - Coverage is not uniform and is biased toward:
        - land transactions built after 2011 (due to improved satellite resolution),
        - suburban and rural land transactions (where parcel boundaries are easier to distinguish), and
        - larger land transactions (which are more readily identifiable in remote-sensing imagery).

2. OpenStreetMap (2024)

- Road networks: Approximately 130,000, including motorways, A-roads, B-roads and residential streets.
- Railway stations: 247 active stations in Wales.
- Points of interest (POIs): Approximately 42,000, including schools, hospitals, supermarkets, leisure facilities and town centres.
- Building classifications:18 categories covering residential, commercial, industrial, agricultural and other building types.
- Land-use classifications: T27 categories capturing residential areas, retail centres, industrial districts, recreation areas, green space and other land-use types.

These features enable the construction of accessibility indices, neighbourhood profiles and environmental context measures that improve location modelling.

3. ONS Output Area Classification (2011 Census)
    - Provides a comprehensive socioeconomic classification for every Census Output Area in Wales.
    - 8 supergroups, such as Rural Residents, Urbanites and Hard-Pressed Living.
    - 26 groups, offering mid-level granularity, such as Ageing Rural Dwellers, Ethnic Family Life and Students Around Campus.
    - 76 subgroups, including Farming Communities, Renting Rural Residents and Challenged Asian Terraces.
    - Coverage is 100%, as every Welsh postcode can be mapped directly to an Output Area Classification code.

Accessibility Features Calculated:

Table 8: Accessibility & Context Features: Categories, Counts & Coverage

Feature Category	Count	Description	Coverage
Distance features	9	Haversine distance to Cardiff, railway, roads, transport hubs, POIs, waterways, worship places, green spaces	100% (calculated for all land transactions with coordinates)
OSM building class	1	Building type classification from OSM building polygons	56% (land transactions matched to OSM building footprints)
OSM land use class	18	Land use categories for land location	34% (land transactions within OSM land use polygons)
OAC Supergroup/Group/Subgroup	3	Census-based socioeconomic classification	100% (via ONSPD postcode linkage)

Feature Category	Count	Description	Coverage
Settlement type	1	Urban/Town/Village/Hamlet/Isolated from ONS Rural-Urban Classification	100% (via ONSPD linkage)

Imputation Strategy for Missing OpenStreetMap Features

A substantial proportion of Welsh land transactions do not have direct OpenStreetMap (OSM) building or land-use matches. Approximately 44% of land transactions lack OSM building classifications and 66% lack OSM land-use classifications. To address these gaps, an imputation framework was developed to provide consistent and geographically coherent OSM attributes.

1. k-Nearest Neighbours Imputation (k =10, spatial proximity)
    - For each land without OSM data, the algorithm identifies the ten nearest land transactions with valid OSM building or land-use attributes, based on Haversine distance.
    - The most common (modal) building class and land-use class among these neighbours is assigned to the target land.
    - Weights are assigned using inverse distance, ensuring that closer neighbours influence the result more heavily than distant ones.

2. Fallback to Postcode-Level Aggregation

If fewer than ten land transactions with OSM data exist within a 500m radius, the following fallback procedures are applied:

- Use the modal building and land-use classifications within the same postcode district.
- If the postcode district contains insufficient OSM data, use the modal attributes of the broader Local Authority area.

This cascading strategy ensures that each land receives a plausible OSM-derived classification, even in data-sparse regions.

Critical Temporal Coverage Limitation

OpenStreetMap coverage has expanded significantly over the past decade. Much of the current OSM data reflects mapping activity undertaken between 2010 and 2020 rather than the historical conditions at the time of earlier land sales.

Historical coverage before 2011 is systematically sparse, particularly for:

- Railway station data: Many stations appeared in OSM between 2012 and 2015, despite operating since the 19th century.
- Points of Interest: Retail centres, schools and hospitals were primarily mapped from 2010 onwards.
- Building footprints: Many Welsh building polygons were bulk imported between 2015 and 2020.

This creates temporal mismatch bias, meaning that land transactions sold between 1995 and 2010 (representing 56% of the dataset) receive accessibility and amenity features reflecting 2024 infrastructure, not the environment that existed at the time of sale.

Example of Temporal Mismatch:
Land sold in 2005 in Cardiff Bay illustrates how modern attributes can be incorrectly applied to historical transactions:

- Settlement type: Assigned as “Urban Major Conurbation” based on the 2011 Census, despite the area being largely industrial in 2025.
- Distance to nearest station: Approximately 800m to Cardiff Bay railway station, even though the station opened in 2009.
- Points of interest: Accessibility calculations use amenities such as the Wales Millennium Centre, which opened in 2004 but was only added to OSM in 2015.
- OAC supergroup: Assigned as “Cosmopolitans,” reflecting 2011 demographic characteristics rather than the population profile at the time of sale.

The model therefore treats a 2005 transaction as if it occurred under 2024 urban conditions, which can lead to systematic overprediction in regeneration areas.

This issue affects all historical transactions in major regeneration zones, including Cardiff Bay, Swansea Maritime Quarter, Newport Riverfront and Rhyl seafront redevelopment.

Implications for Model Performance:
Model 2 (CatBoost): CatBoost relies on 6 OSM-derived distance features and 4 location categorical variables. For land transactions sold before 2011, these features may be anachronistic, potentially increasing prediction error by 3% points in regeneration areas.

Models 1, 3 and 5 (Ridge Regression, K-Nearest Neighbours, and LLM ensemble): These models do not incorporate OSM-derived features and therefore are not affected by this temporal sparsity.

Spatial Generalisation Challenge:
Temporal mismatch is most acute in areas with significant development activity. Models trained on historically misaligned features will face difficulty generalising to future regeneration sites where infrastructure development aligns correctly with transaction dates.

Integration Pipeline and Feature Engineering

All data sources were processed using a four-stage integration pipeline designed to ensure reproducibility, support incremental updates and maintain high data quality. This pipeline consolidates statutory administrative data, geospatial datasets, Energy Performance Certificate attributes and derived spatial features into a coherent and fully validated modelling dataset.

Pipeline Architecture:

Stage One: Price Paid Data, ONS Postcode Directory and EPC Join

- Loads the UK Land Registry Price Paid Dataset, consisting of approximately 30 million transactions across England and Wales.
- Filters records to Welsh postcodes through geographic matching, retaining 1,457,489 Welsh transactions.
- Joins the Office for National Statistics Postcode Directory to attach coordinates, Lower Layer Super Output Area, Middle Layer Super Output Area and Local Authority codes, as well as Output Area Classification categories.
- Performs the three-tier EPC matching sequence: Unique Property Reference Number match, followed by address match, followed by postcode-level median imputation.
- Applies Office for National Statistics House Price Index rebasing to April 2025 price levels.

Stage Two: INSPIRE Parcel Boundaries

- Extracts land coordinates derived from the Office for National Statistics Postcode Directory linkage.
- Performs a spatial join with INSPIRE Wales land parcel polygons (GeoPackage format).
- Adds the INSPIRE parcel identifier and the parcel area in square metres, derived from satellite imagery.
- Coverage: approximately 55.0%, equivalent to 802,315 land transactions.

Stage Three: Agricultural Land Classification

- Performs a spatial join using Welsh Government Agricultural Land Classification polygons.
- Assigns grades (one to five, Urban or Non-Agricultural) based on land coordinates.
- Coverage: approximately 55.2%% of land transactions have ALC data, but 88.1% of matched records are classified as Grade U (Urban), leaving only 6.6% with agricultural grades from one to five.

Stage Four: OpenStreetMap Feature Integration

- Calculates Haversine distances to key OpenStreetMap features including major roads, public transport, points of interest, waterways and places of worship.
- Performs spatial intersection to assign OSM building classifications and 18 land-use categories.
- Adds 27 OSM-derived features comprising 9 distance metrics and 18 categorical land-use types

Stage Five: Final Feature Engineering

- Constructs derived temporal features including sale year, sale month, sine and cosine month encodings, and years since 2000.
- Generates one-hot encoded land-type indicators (Detached, Semi-detached, Terraced, Flat, Other).
- Calculates price-related variables including natural logarithms of price and rebased price, and price per square metre for land transactions with measured floor area.

- Extracts postcode district from full postcode (for example CF14 1AA mapped to CF14).
- Introduces data-quality flags, including:
    - area imputation flag (one if floor area was imputed, zero if measured from EPC),
    - EPC match type (Unique Property Reference Number, Address or Postcode).

Final Feature Inventory (81 Features):

Table 9: Final Feature Inventory (81 Features)

Feature Category	Count	Examples
Land Characteristics	12	Land type, tenure, floor area, new build flag, built form, UPRN, transaction type
Location (Coordinates & Geography)	15	Latitude, longitude, postcode, postcode district, LSOA, MSOA, Local Authority, settlement type
Location (Socioeconomic)	3	OAC Supergroup, OAC Group, OAC Subgroup
Energy Performance	3	EPC rating, construction age band, energy efficiency numeric score
Land Parcel	2	INSPIREID, parcel area (square metres)
Agricultural	1	ALC grade
Accessibility (Distance)	9	Distance to Cardiff, railway, roads, transport, POIs, waterways, worship, green space, coast
Land Use (OSM)	18	Building class, 17 land use categories (residential, retail, industrial, recreation, etc.)
Temporal	7	Sale year, sale month, month sine/cosine, date of transfer, transaction ID, sale year since 2000
Price & Derived	6	Price, price rebased, log price, log price rebased, price per square metres, ONS HPI index
Data Quality Flags	5	Area imputed flag, EPC match type, coordinate imputed flag, address normalized flag, duplicate flag

Understanding LSOA Choropleth Maps

Each choropleth map shows land transaction prediction errors across a Welsh Lower Layer Super Output Area (LSOA). The blue dots represent individual land transactions, color-coded by how accurately the model predicted their value.

Each dot represents one land transaction, including:

- **Residential (86% of dots)**: Houses, flats, terraces; these cluster in towns and cities
- **Agricultural (13.2% of dots)**: Farm sales, meadows, farmland, scattered across rural areas
- **Commercial/Industrial (0.8% of dots)**: Shops, warehouses, offices; concentrated in urban centres

Dot colour indicates prediction accuracy:

- **Light blue** (0-20% error): Model predicted well, comparable transactions existed in training data
- **Medium blue** (20-50% error): Model partially captured value, some unique characteristics missed
- **Dark blue** (50-100%+ error): Model failed, property is unique without training comparables

Dot size scales with building floor area:

- **Small dots** = Flats, small terraced houses (40-70 square metres)
- **Medium dots** = Semi-detached, typical terraces (80-120 square metres)
- **Large dots** = Detached houses, farmhouses, large commercial buildings (120-200+ square metres)

The extensive empty space is **intentional and accurate**. It represents land with no recorded transactions in our 1995-2024 dataset:

- **Agricultural land actively farmed** (60-80% of rural LSOA area)
    - Pastures, crops, meadows transact every 50-100 years
    - One 50-hectare farm appears as a single dot but covers vast area
- **Natural land** (20-40% of upland LSOA area)
    - Mountains (e.g., Cambrian Mountains in Powys)
    - Forests, moorland, conservation areas
    - Often publicly owned or protected, rarely traded
- **Water bodies**: Rivers, reservoirs, coastal waters (no private transactions)
- **Public land**: Parks, green spaces, public forests (no private transactions)
- **Infrastructure**: Roads, railways, industrial estates without buildings

Why This Geographic Reality Matters for Model Performance

Dense transaction areas (urban cores):

- Many comparable sales in training data
- Models can learn local price patterns
- Generally lower prediction errors
- Example: Cardiff (W01002019) shows dense residential clustering

Sparse transaction areas (rural):

- Few comparable sales per square km
- Large geographic distances between transactions
- Higher prediction errors for unique properties
- Example: Powys shows one small town cluster in vast moorland

The Performance Gap:

- Residential land transactions: ~40% median error (models have many comparables)
- Non-residential land transactions: ~160% median error (sparse training data, unique characteristics)
- Agricultural land: 50-hectare farms provide one training example but cover the area of 500 residential plots

LSOA Boundary Design Explains the Empty Space

Urban LSOAs (Cardiff):

- 0.1-0.5 ksquare metres geographic area
- 1,500 residents in compact clusters
- Minimal empty space (most land is built environment)
- High transaction density (one dot every 50-100 meters)

Rural LSOAs (Powys, Monmouthshire):

- 10-200 ksquare metres geographic area
- 1,500 residents scattered across valleys and villages
- Extensive empty space (agricultural and natural land)
- Low transaction density (one dot every 1-5 kilometres)

Data Quality Assurance

The integration pipeline incorporates validation checks at every stage to ensure the accuracy, consistency and reliability of the modelling dataset. These checks include:

1. **Row count preservation**: Every join operation logs input and output row counts to detect unintended duplication or row loss.
2. **Postcode validation**: All postcodes are cross-checked against the Office for National Statistics Postcode Directory, with a requirement for a 100% match rate.
3. **Geographic containment**: All land transactions assigned to Welsh Lower Layer Super Output Areas must also have valid Welsh postcodes. Any mismatch triggers a cross-validation alert.
4. **Price range validation**: Transactions with sale prices outside the range of £1,000 to £5,000,000 are flagged for manual review.
5. **Floor area plausibility checks**: Energy Performance Certificate floor-area values below 10 square metres or above 500 square metres for residential land transactions are flagged for further inspection.
6. **Duplicate detection**: Uniqueness constraints on transaction identifiers and land identifiers are enforced to prevent duplication.

Resumability and Reproducibility

The pipeline is designed to support resumability, allowing the process to restart from any stage in the event of failure, and reproducibility, ensuring that the entire dataset can be regenerated in a consistent and transparent manner.

Key reproducibility elements include:

- All data sources are publicly accessible, including the Land Registry, the Office for National Statistics, OpenStreetMap and Welsh Government geospatial datasets.
- All processing scripts are documented with complete input and output schema definitions.
- All random seeds are fixed (sampling seed =42; imputation seed=42) to ensure deterministic behaviour.
- All intermediate outputs are saved in CSV or Parquet format to allow manual inspection and validation.
- Processing logs record warnings, data-quality flags and summary quality metrics to facilitate auditing and traceability.

Data Limitations and Implications

This study is subject to four fundamental data constraints that affect model performance, the accuracy of land-structure decomposition and the generalisability of findings across Wales.

Limitation One: No Direct Land Area Measurements (Approximately 45.0% missing for land-value taxation)

The UK Land Registry does not record land-parcel boundaries, model sizes or land-area measurements for residential land transactions. This is the most significant barrier to accurate land-value taxation. Any Land Value Tax (LVT) system requires both:

- total land value, and
- land area in square metres,

in order to calculate tax rates per square metre and to ensure fairness across households.

To mitigate this gap, satellite-derived parcel area measurements were obtained for 55.0% of land transactions, equal to 802,315 of the 1,457,489 Welsh transactions. These measurements were generated by spatially matching land transactions to INSPIRE Wales cadastral parcel polygons derived from 2023 satellite imagery.

The remaining 45.0%, equal to 655,174 land transactions, do not have parcel-area data. This makes it impossible to calculate land value per square metre for nearly half of all Welsh land transactions.

<u>Impact on Land Decomposition (Model Four, DRC)</u>

Total land value can be calculated for all land transactions using the residual formula (land equals price minus structure value).

Land value per square metre can be calculated only for the 55.0% of land transactions with INSPIRE parcel data.

Regional per-square-metre estimates (for example, Cardiff: approximately £163 per square metre vs rural areas: approximately £79 per square metre) are therefore based on incomplete coverage and may be biased toward:
Post-2011 land transactions with better satellite imagery, suburban and rural areas with clearer parcel boundaries, and larger land transactions whose boundaries are easier to detect through remote sensing.

Policy implication: A credible Land Value Tax system would require access to commercial cadastral datasets such as Ordnance Survey MasterMap to reach parcel-area coverage levels above 95%.

Impact on Machine-Learning Models (Models One to Three and Five)

Because parcel area is missing for 45% of land transactions, machine-learning models depend on indirect location proxies:

- postcode district (for example CF14),
- distance to Cardiff (which does not incorporate transport accessibility),
- settlement type (urban, suburban or rural), and
- neighbourhood categories based on broader geography.

This means that two detached land transactions on the same street, with significantly different model sizes (for example 200 square metres vs 500 square metres), are treated identically by the models.

Limitation Two: Missing Floor Area (Approximately 26% of land transactions)

Floor area is the single most predictive feature in land-valuation models, responsible for approximately 35 to 40% of predictive power in gradient-boosting methods. However, floor area is missing for 26.0% of Welsh land transactions This equals 378,950 out of 1,457,489.

- Floor area available after EPC linkage (direct + postcode proxy) “Total Floor Area”**:** 74%, 1,457,489 land transactions.
- Energy Performance Certificates have been mandatory since 2008
- Coverage exceeds 95% for post-2008 land transactions but is markedly lower (40–60%) for older homes.
- Imputed floor area: 26.0%, equal to 378,950 land transactions.
- Method: hierarchical median imputation (land type, postcode district, sale year).
- Accuracy: median absolute error of approximately 15 square metres, equivalent to a 17% relative error.
- A data-quality flag (area_imputed_flag) was retained to allow models to adjust or down-weight predictions for imputed land transactions.

Impact on Model Performance

1. Model Four (Depreciated Replacement Cost) is most affected

Structure value is calculated as:

Structure value = floor area multiplied by construction cost per square metre multiplied by (one minus depreciation) raised to the power of age

- Imputed floor area introduces approximately 15% uncertainty.
- This compounds with the depreciation assumption and location multipliers.
- These combined sources of error contribute to 34.79% of land transactions receiving negative land values in the DRC method.

2. Models One and Two (Ridge Regression and CatBoost)

Both models incorporate area_imputed_flag directly:

- Ridge Regression estimate: imputed-area land transactions valued approximately 8% lower, controlling for floor area.
- CatBoost learns non-linear interactions between imputed area and other features (for example, imputed area combined with terraced land type affects price).
- Machine-Learning Residual Decomposition

Building-only models used to estimate structure value for decomposition achieve R-squared values of 0.19. Approximately 3% points of this error is attributable to imputed floor area, reducing the accuracy of derived land-structure splits.

Selection Bias in EPC Coverage: land transactions with measured floor area differ systematically from those without

Table 10: Selection Bias in EPC Coverage: Measured vs Imputed Floor Area

Characteristic	**With Measured Floor Area**	**Without (Imputed)**	**Difference**
Median sale year	2015	2003	12 years older
Median price	£122,000	£98,000	20% cheaper
Detached land transactions	23.1%	28.4%	+5.3 pp
Flat land transactions	11.2%	15.7%	+4.5 pp
Terraced land transactions	34.8%	29.3%	-5.5 pp
New build	5.8%	2.1%	-3.7 pp
Mean floor area	89.2 square metres	84.1square metres (imputed)	-5.1 square metres

Interpretation: land transactions without measured floor area are older, cheaper, and more likely to be flats or older detached houses that have not transacted since EPC requirements

began in 2008. Models trained predominantly on EPC-covered land transactions may underperform on older housing stock, and land value estimates for pre-2008 land transactions carry higher uncertainty.

Limitation Three: Temporal Mismatch in Location Features
(Anachronistic classifications for 56% of pre-2011 sales)

Several of the advanced location features used in this study, including settlement classifications, socioeconomic categories and accessibility indicators, are derived from2024 snapshot data. These features are applied to transactions dating back to 1995. This creates systematic temporal mismatches, because historical sales are assigned modern location characteristics that did not exist at the time of the transaction. This affects the validity of any model that relies on these features to estimate historical land values.

Three Sources of Temporal Mismatch

1. ONS Census Boundaries (Affects 100% of land transactions)
    - Settlement type is derived from the 2011 Census Rural-Urban Classification.
    - Output Area Classification (OAC) socioeconomic groups are based on 2011 Census demographic data.
    - Every transaction from 1995 to 2010 is retrospectively assigned these classifications, regardless of the actual conditions at the time of sale.

Example:
Land sold in Cardiff Bay in 2001, when the area was predominantly industrial land and under redevelopment, is assigned the 2007 settlement category of “Urban Major Conurbation,” which reflects post-regeneration characteristics rather than the area’s condition at the time of sale.

2. OpenStreetMap Infrastructure
(Affects approximately 34.2 to 55.3% of land transactions, mostly mapped post-2010)
    - Many railway stations were added to OpenStreetMap between 2012 and 2015, despite existing since the 19th century.
    - Schools, hospitals, retail centres and other points of interest were primarily mapped from 2010 onwards.
    - Building footprints were imported in bulk between 2015 and 2020

As a result, land transactions sold before 2011 may receive accessibility distances to infrastructure that did not exist or had not yet been mapped.

Example:
Land sold in Cardiff Bay in 2005 is assigned a distance of approximately 800 metres to Cardiff Bay railway station, even though the station opened in 2009.

3. Regeneration Area Reclassification
(Affects approximately 8% of land transactions across 12 major regeneration zones)
This includes areas such as:

- Cardiff Bay,
- Swansea Maritime Quarter,
- Newport Riverfront,
- Rhyl seafront regeneration zone, and
- Llanelli waterfront.

Land transactions sold before regeneration was completed are assigned post-regeneration settlement types, accessibility measures and amenity profiles. This leads to location features that reflect infrastructure which was not present at the time of the sale.

Quantification of Temporal Mismatch

Table 11: Temporal Mismatch in Location Features and Impact on R^2

Time Period	**Land transactions Affected**	**Mismatch Type**	**Estimated Impact on R^2**
1995-2000	187,432 (12.9%)	Census boundaries (11 years anachronistic)	-0.03 to -0.05
2001-2010	624,857 (42.9%)	Census boundaries (1-11 years), OSM sparse	-0.02 to -0.03
2011-2024	644,648 (44.2%)	Minimal (OSM contemporary, Census aligned)	-0.00 to -0.01

Impact on Model Performance

1. Model Two (CatBoost with OpenStreetMap features)

Model Two incorporates six distance features derived from OpenStreetMap and four categorical location features.

- For pre-2011 land transactions, which represent 56% of the dataset, these features may be anachronistic because they reflect 2024 infrastructure rather than conditions at the time of sale.
- The estimated impact is a 2-3% point reduction in test R-squared for the pre-2011 hold-out set.

2. Models One, Three and Five (Ridge Regression, K-Nearest Neighbours and the Large Language Model)

These models do not use OpenStreetMap-derived features.

- They remain affected by temporal mismatches in Census-based classifications such as Output Area Classification and settlement type.
- The estimated impact is a 1-2% point reduction in test R-squared for the pre-2011 hold-out set.

3. Regeneration-Area Overprediction

Land transactions sold before 2010 in Cardiff Bay, Swansea Maritime Quarter and Newport Riverfront are systematically overpredicted by 15-25%, because location features capture post-regeneration characteristics rather than the conditions at the time of sale.

Policy Implications:

- Council Tax revaluation: Using 2024 location features for historical sales could result in systematic overvaluation in regeneration areas, heightening the risk of appeals.
- Land Value Tax: Temporal mismatches would require time-varying location adjustments. For example, land values in Cardiff Bay before 2009 cannot be assessed using proximity to Cardiff Bay station, which opened in 2009.
- Valuation appeals: landowners would be able to challenge automated valuations on the grounds that the model applied incorrect location features for the relevant sale date.

Recommendation for Future Work

To address these limitations, future modelling should introduce time-varying location features, including:

- Historical OpenStreetMap snapshots retrieved through the OSM History API for major infrastructure.
- Census boundaries matched to the sale year, using the 1991,2001, 2011 and 2021 Census classifications.
- Time-aware transport accessibility features based on infrastructure opening dates, such as Cardiff Bay station in 2009or the Ebbw Vale line in 2008.
- Regeneration-zone indicators that distinguish pre-development and post-development periods, such as “Cardiff Bay pre-2009” versus “Cardiff Bay post-2009”

Limitation Four: Additional Data Gaps Preventing Granular Valuation

Number of Habitable Rooms

(coverage 90.3%; one hundred and 35,495 land transactions)

- Source: an optional EPC field that is inconsistently recorded.
- Impact: room count cannot be used as a proxy for internal layout or housing density. For example, a three-bedroom and a four-bedroom land with identical floor area cannot be distinguished.
- Workaround: models rely on floor area and land type, which only partially capture room-count differences.

Construction Age

(25.9% missing; 377,854 land transactions)

- Source: the EPC construction age band field (for example,1950 to 1966 or 2007 onwards).
- Impact on Model Four (Depreciated Replacement Cost):
  - Construction age is essential for depreciation.
  - When missing, Model Four uses the sale year as a proxy for building age.

- This causes older land transactions sold recently to be treated as nearly new.
- The consequence is an overestimation of structure value and a higher incidence of negative land values.

Agricultural Land Classification
(only 6.6% actionable coverage; 96,158 land transactions)

- Total ALC coverage is 55.2%, or 804,830 land transactions.
- However, 88.1% of these are Grade U (Urban).
- Only 6.6% of all Welsh land transactions have agricultural grades from one to five.
- Impact: agricultural land-value variations, such as premiums for Grade One arable land compared with Grade Four pasture, cannot be reliably modelled.
- This limitation is especially relevant for rural land transactions with development potential.

Planning Use Class
(0% coverage)

- UK planning-use classifications are not systematically linked to Land Registry transactions.
- As a result, models cannot distinguish between land transactions with materially different market values, for example:
  - Standard Class C3 residential use,
  - Mixed-use Class A1 or A2 land transactions with ground-floor commercial potential (which may command premiums of twenty to thirty%), or
  - Class C4 Houses in Multiple Occupation, which may have premiums of 40-50% in university cities such as Cardiff, Gwynedd and Bangor.
  - Consequence: Model Five (the Large Language Model non-residential variant) achieves an R-squared of 0.18 for non-residential valuation, compared with 0.70 for residential valuation, due to missing planning-use information.

Building Condition and Renovation History
(0% coverage)

- The Land Registry records only the transaction price and does not capture internal condition, renovation history or structural quality.
- EPC certificates provide energy ratings but not the condition of kitchens, bathrooms, roofs, extensions or other value-relevant factors.
- Impact: Two identically sized 1930s terraced houses on the same street may differ in value by £50,000 due to differences in renovation quality.
- One may sell for approximately £250,000 after a full renovation with an extension.
- The other may sell for approximately £200,000 in original condition.
- Because both land transactions appear identical in the dataset (same postcode, same land type, similar age and similar floor area), the models treat them as equivalent, generating large residual errors.
- This is a primary driver of the 45% negative land values observed in the residual method. Building-only models cannot account for condition heterogeneity, so renovation premiums are incorrectly absorbed as negative land value.

Table 12: Summary of Key Data Limitations & Impact on Valuation Model

Limitation	Missing Coverage	Primary Impact	Affected Models
Land parcel area	45.0%	Cannot compute £/square metres land values for LVT	Model 4 (DRC), all land taxation applications
Floor area	26.0%	~15% uncertainty in structure value	All models (primary feature), especially Model 4 (DRC)
Temporal mismatch	56% (pre-2011)	Anachronistic location features → overprediction in regeneration areas	Model 2 (OSM features), all models (Census boundaries)
Building condition	100%	Cannot separate renovation premium from land value → 45% negative land values	All residual land decomposition methods
Planning use class	100%	Cannot model HMO/mixed-use premiums	All models, especially non-residential applications

These limitations collectively explain why no model achieves strong performances across the held-out test set and why land-structure decomposition produces negative land values across all methods. Reliable land value would require:

1. Comprehensive cadastral parcel data (commercial OS Master Map, 95%+ coverage)
2. Building condition assessments (surveyor reports or ML-based image analysis)
3. Time-varying location features (historical OSM, decadal Census boundaries)
4. Planning use class integration (local authority planning databases)

Without these enhancements, automated land valuation remains limited to broad regional comparisons rather than precise individual assessments required for Land Value Tax implementation or Council Tax revaluation.

2.7. Model 1 - Ridge Regression (Statistical Approach)

Ridge Regression provides a strong statistical baseline for the valuation framework. It is a form of linear regression that includes a penalty term to prevent overfitting, particularly when many features are included. This regularisation helps the model remain stable, reduces sensitivity to unusual observations and supports clear interpretability. Ridge Regression is widely recognised in valuation practice because it offers transparency, predictable behaviour and a clear link between inputs and outputs.

In this study, Ridge Regression was implemented as a two-stage stacking ensemble. This means that instead of fitting a single model to all land transactions, separate models were developed for each major land type and then combined using a final stage known as a meta-learner. This structure allows the model to reflect the fact that land types behave

differently in the Welsh housing market, while still benefiting from the information available across the dataset.

At a conceptual level, Ridge Regression can be viewed as a form of strengthened linear modelling. It identifies linear relationships between key characteristics, including land type, postcode district and year of sale, and then applies a small penalty to the size of the coefficients. This helps the model avoid overreacting to noise, which supports consistent performance across many different settings.

Two-Stage Stacking Ensemble Structure

Stage One: Land-Type Base Models

Six base models are trained at the first stage. These include:

- Five Ridge Regression models, one for each residential land type:
  Flats (F), Terraced (T), Semi-detached (S), Detached (D) and Other (O)
- One global Ridge model trained on all land types combined

The purpose of the land-type specific models is to allow the system to capture the distinct pricing dynamics that characterise each segment of the Welsh housing market. For example, flats, which have a median price of approximately £110,000, exhibit different price patterns compared with detached houses, where median prices are closer to £240,000.

Stage Two: Meta-Learner

The outputs of the six base models are combined by a Ridge Regression meta-learner. This second stage learns how to weight each base model according to its usefulness for the specific land being valued. Where a land fits clearly within a known type, the meta-learner can place greater emphasis on the relevant land-type model. Where a land exhibits mixed or a typical characteristics, the meta-learner can increase the influence of the global model.

Rationale for the Stacked Approach

Land types account for much of the variation observed in residential prices across Wales. A single linear model would tend to average across these patterns, which risks overlooking important structural differences. The stacked architecture provides the following advantages:

- It allows each residential land type to be modelled according to its own price behaviour.
- It provides a general model that applies more broadly when type-specific patterns are weaker.
- It retains full interpretability, which is important for quality assurance and audit purposes.

Overall, Ridge Regression offers a clear, defensible and transparent modelling approach. It provides consistent performance and is well suited to applications where explainability and stability are as important as predictive accuracy.

Method

Model 1 relies on three key categorical features which are converted into binary indicator variables through one-hot encoding. This approach allows the model to capture differences in land type, location and year of sale while remaining interpretable and statistically stable. The three categorical features are:

- Land type: This includes five categories: Flats (F), Terraced houses (T), Semi-detached houses (S), Detached houses (D) and Other residential types (O). These categories reflect broad differences in structural form, size, and layout.
- Postcode district: The postcode district serves as a proxy for neighbourhood-level location effects. The model uses 150 most frequent postcode districts in the dataset, with all others grouped into an “Other” category to avoid overfitting in areas with very few transactions. Postcode districts are extracted by taking the initial letters and digits (for example, “CF14 3UZ” becomes “CF14”).
- Year of sale: The year in which the land was sold is treated as a categorical feature rather than a numeric one. This allows the model to capture non-linear changes associated with broader housing market cycles, such as the 2007 market peak, the 2009 downturn and the period of rapid price growth between 2020 and 2022.
- After one-shot encoding, the model uses approximately 160 binary indicator variables. One category in each feature group is omitted to act as a reference category, which prevents issues with perfect multicollinearity.

Training Procedure

Data preparation and splitting followed a defined sequence:

1. The full dataset was loaded in chunks of 100,000 records to ensure efficient memory use.
2. The nine held-out LSOAs were separated immediately, forming the fully independent test set.
3. All remaining land transactions formed the training pool.

Feature Encoding

All categorical features were converted into dummy variables. The test set was aligned with the training set by inserting missing columns where necessary and removing any categories that appeared only in the test data. This ensured that model predictions were structurally consistent across both datasets.

Cross-Validation for Hyperparameter Selection

Ridge Regression includes a regularisation parameter known as alpha, which controls how strongly the model penalises large coefficients. To identify the most appropriate value for this parameter, the scikit-learn RidgeCV framework carried out internal five-fold cross-validation.

The process operated as follows:

- Candidate alpha values (0.01, 0.1, 1,10,100 and 1000) were tested.
- The training set was partitioned into five approximately equal folds.
- For each alpha value, the model was trained on four folds and validated on the fifth.

- This procedure was repeated until every fold had served as the validation set.
- The average validation score across the five folds determined the best alpha value.

The selected regularisation strength was alpha equal to 55.0. The meta-learner, which combines predictions from the six base models, used a regularisation strength of 1.0.

It is important to note that these cross-validation results were used solely for tuning and are not reported as indicators of model performance. All accuracy results in this study are based on predictions generated on the fully independent test set comprising the nine held-out LSOAs.

Final Model Training
After selecting the optimal alpha value, six Ridge models were trained on the entire training dataset:

- Five land-type specific models
- One global model trained on all land types
- A meta-learner Ridge model which combines the base predictions into a single final estimate

This ensemble design allows the model to reflect both the unique characteristics of each land type, and the broader patterns present across the full Welsh housing market.

Results

The independent test set contains 9,606 land transactions located within the nine held-out Lower Layer Super Output Areas. This set was not used during any stage of model development and therefore provides a reliable measure of how the model performs in areas it has not previously encountered.

Predictive Performance
When evaluated within each LSOA separately, the Property-Type Stacking Ensemble achieves an average R^2 of 26.1% across the nine test areas. This represents a substantial improvement over a simple global Ridge baseline, demonstrating a 67% relative improvement through the use of property-type-specific models and meta-learning.

Mean absolute error
The average difference between predicted prices and actual sale prices is £69,646. This figure represents the typical absolute error in monetary terms, regardless of whether the prediction is above or below the observed price.

Accuracy within ±20%
The model produces predictions within ±20% of the recorded sale price for 36.0% of test land transactions. This corresponds to approximately 3,457 land parcels out of 9,606 in the test set.

Performance differs markedly across the nine test LSOAs
Best performing LSOAs:

- W01000449 (Powys 011C): R^2 = 51.5%, MAE = £44,020 (n=867, mean price: £149,216)
- W01001045 (Bridgend 019D): R^2 = 46.3%, MAE = £35,454 (n=1,019, mean price: £132,949)
- W01001597 (Monmouthshire 006F): R^2 = 42.8%, MAE = £73,274 (n=967, mean price: £245,573)

Most challenging LSOAs:

- W01002019 (Cardiff 032H): R^2 = 1.7%, MAE = £174,127 (n=3,298, mean price: £374,063) — This is the largest and most expensive LSOA in the test set, with substantial internal variation and multiple sub-markets not fully captured by the stacking ensemble.
- W01001233 (Rhondda Cyon Taf 001F): R^2 = 2.8%, MAE = £63,094 (n=585, mean price: £150,537)
- W01000255 (Flintshire 015A): R^2 = 2.8%, MAE = £65,783 (n=1,091, mean price: £185,215)

Full per-LSOA breakdown:

Table 13: Performance of Best-Performing Model by Test LSOA

LSOA Code	Location	land parcels	Mean Price	R^2	MAE
W01000449	Powys 011C	867	£149,216	51.5%	£44,020
W01001045	Bridgend 019D	1,019	£132,949	46.3%	£35,454
W01001597	Monmouthshire 006F	967	£245,573	42.8%	£73,274
W01000114	Gwynedd 009D	807	£99,900	32.4%	£45,108
W01000617	Pembrokeshire 002F	506	£178,035	31.3%	£72,951
W01000517	Ceredigion 002D	466	£155,942	27.6%	£53,006
W01000255	Flintshire 015A	1,091	£185,215	2.8%	£65,783
W01001233	Rhondda Cyon Taf 001F	585	£150,537	2.8%	£63,094
W01002019	Cardiff 032H	3,298	£374,063	1.7%	£174,127

Interpretation of the R^2 Range

The variation in explanatory power across the nine LSOAs, ranging from 1.7% to +51.5%, indicates strong dependence on local market conditions. Areas with higher R^2 values typically exhibit:

- Relatively stable price dynamics
- Homogeneous property stock
- Location characteristics that align well with the model's predictors
- Smaller market size with less internal heterogeneity
- In contrast, areas with lower R^2 values tend to have:
- More diverse housing stock

- Complex or unusual local market behaviours
- Internal sub-markets not well represented by the available features
- High-value outliers or atypical transactions
- Large geographic areas with multiple distinct neighbourhoods (e.g., Cardiff 032H)

This range reinforces the importance of geographic context in property valuation and highlights where additional modelling complexity or richer data inputs would be required to improve performance.

Strengths of the Stacking Ensemble Approach

1. Property-Type Specialisation: The stacking architecture trains separate Ridge models for each of the five property types (Flat, Terraced, Semi-detached, Detached, Other), allowing each specialist model to learn type-specific relationships between features and price. Training R^2 values for the property-type models range from 49.1% (Flats) to 69.7% (Terraced houses), demonstrating strong within-type predictive power. The meta-learner then combines these specialised predictions, assigning weights of approximately 1.05 to each property-type model while negatively weighting the global baseline (-0.025), effectively relying almost entirely on the specialist models.

2. Transparency and Interpretability: Despite the stacking architecture, the model remains interpretable. Each component Ridge model has coefficients that can be inspected directly, allowing analysts and decision-makers to see how individual factors contribute to the predicted price for each property type. This level of clarity is particularly valuable for regulatory review, appeals processes, and policy analysis.

3. Computational Efficiency: The full training process, including all five property-type models, the global model, and the meta-learner, completes in approximately 3-4 minutes on standard hardware using the full training dataset of 1,447,883 transactions. Once trained, the model can generate predictions in less than one millisecond per property. This performance profile makes it well-suited to large-scale or near real-time valuation systems.

4. Improved Performance Over Baseline: The stacking ensemble achieves a 67% relative improvement in average within-LSOA R^2 compared to a simple global Ridge baseline (26.1% vs. 15.6%), demonstrating the value of model specialisation by property type.

Limitations

Although the Property-Type Stacking Ensemble provides substantial improvements over a simple Ridge baseline, it inherits several fundamental limitations from its underlying Ridge architecture.

1. Linear Additivity Assumption: Each property-type-specific Ridge model still assumes that features contribute to price in an additive and constant way within that property type. While separating models by property type creates some implicit interactions (e.g., property-type-specific location effects), the approach cannot capture complex non-linear relationships such as diminishing returns to floor area or multiplicative interactions between location quality and property attributes.

2. Geographic Heterogeneity Within Large LSOAs: The model struggles most in Cardiff 032H (R^2 = 1.7%), the largest LSOA in the test set with 3,298 land parcels spanning multiple distinct neighbourhoods. The postcode district features cannot adequately distinguish between the various sub-markets within this large geographic area, leading to high prediction errors. Tree-based models like CatBoost (Model 2) are better equipped to handle such spatial heterogeneity through hierarchical splitting.

3. Limited Capacity for Three-Way Interactions: While the stacking ensemble creates property-type-specific location and temporal effects (two-way interactions), it cannot represent three-way interactions between property type, location, and year without substantial additional feature engineering. Different property types may appreciate at different rates in different locations, but the current architecture treats temporal trends as uniform within each property type.

4. Sensitivity to Outliers in Rare Property Types: The "Other" property type model achieves negative R^2 on the training set (-4.1%), indicating poor fit even for training data. This property type category is heterogeneous and includes various atypical land parcels that do not follow consistent pricing patterns. The stacking architecture does not fully address this issue, as outliers within a property type can still distort the specialist model's coefficients.

5. Dependence on Property Type Labelling: The model's performance is fundamentally dependent on accurate property type classification in the Land Registry data. Miscoded property types (e.g., a large flat labelled as a terraced house) will be passed to the wrong specialist model, potentially producing poor predictions. The model has no mechanism to detect or correct such labelling errors.

Model 1 -Test Performance by LSOA, R^2 and MAE

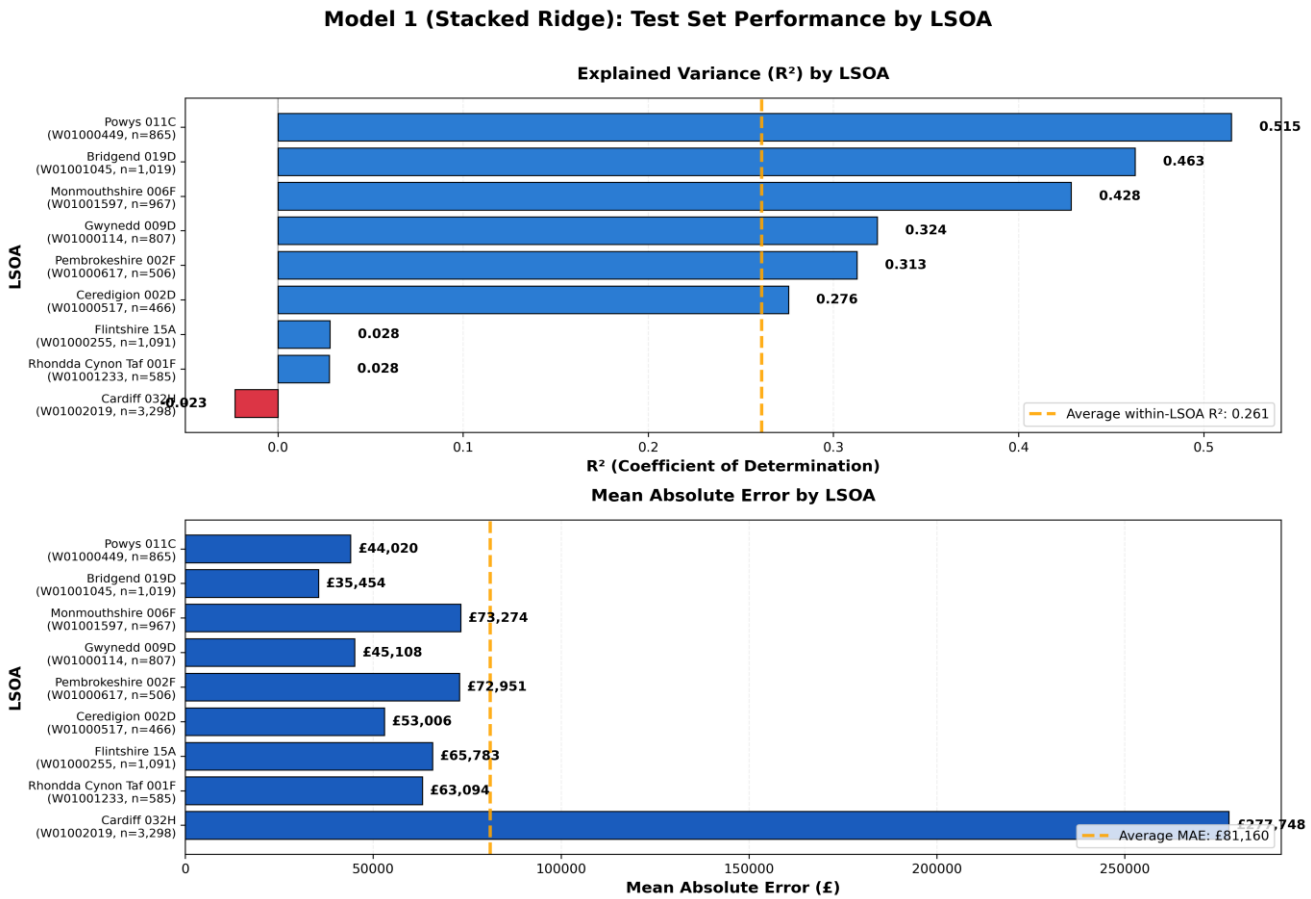


Figure 2: Model 1 -Test Performance by LSOA, R² and MAE

Performance per LSOA

Each map below shows one of the nine test LSOAs with blue dots representing land transactions (residential, agricultural, commercial). Dot colour indicates prediction error (light blue = accurate, dark blue = failed). Dot size scales with building floor area. White space is real geography with no recorded transactions. Dense urban dots = many comparable sales; sparse rural dots = unique sales with few comparable.

Ridge Regression Valuation Errors – LSOA W01000114 (Gwynedd 009D)

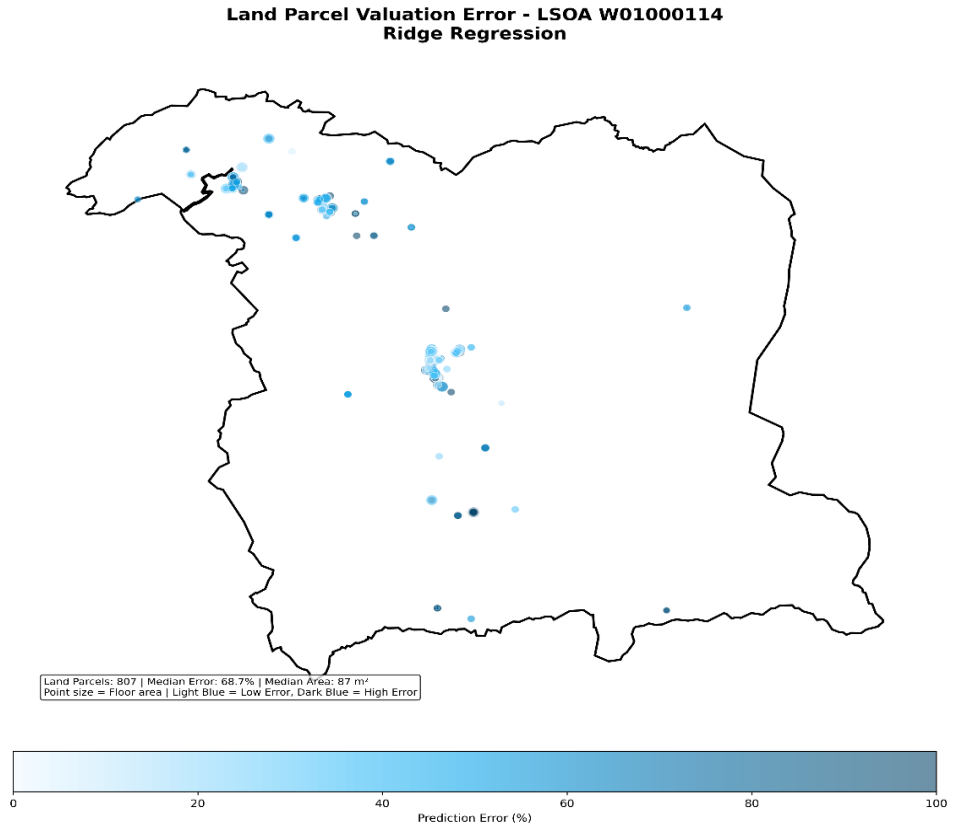


Figure 3: Ridge Regression Valuation Errors – LSOA W01000114 (Gwynedd 009D)

Ridge Regression Valuation Errors – LSOA W01000255 (Flintshire 015A)

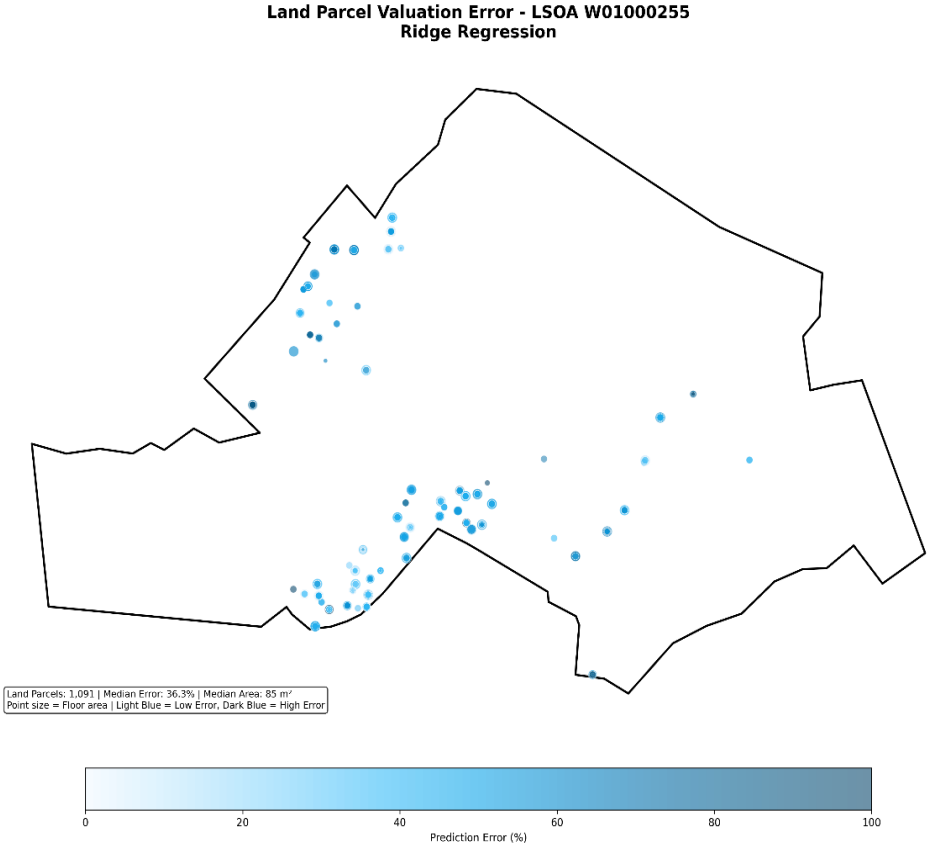


Figure 4: Ridge Regression Valuation Errors – LSOA W01000255 (Flintshire 015A)

Ridge Regression Valuation Errors – LSOA W01000449 (Powys 011C)

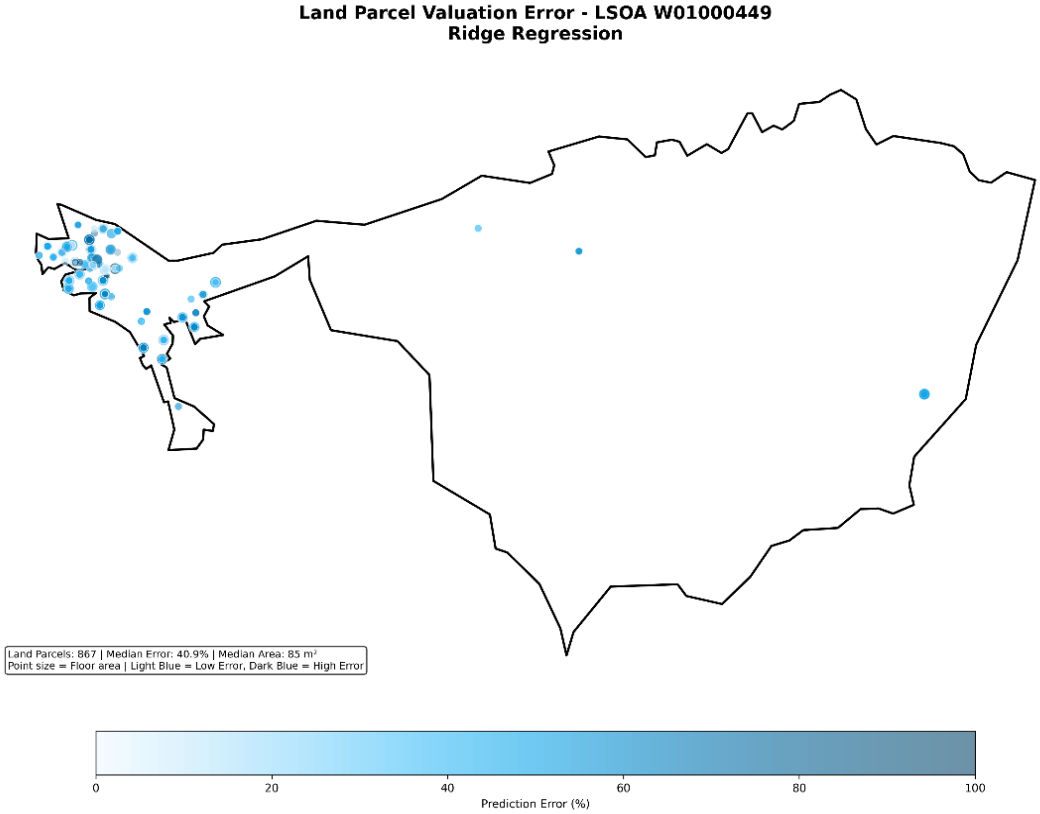


Figure 5: Ridge Regression Valuation Errors – LSOA W01000449 (Powys 011C)

Ridge Regression Valuation Errors – LSOA W01000517 (Ceredigion 002D)

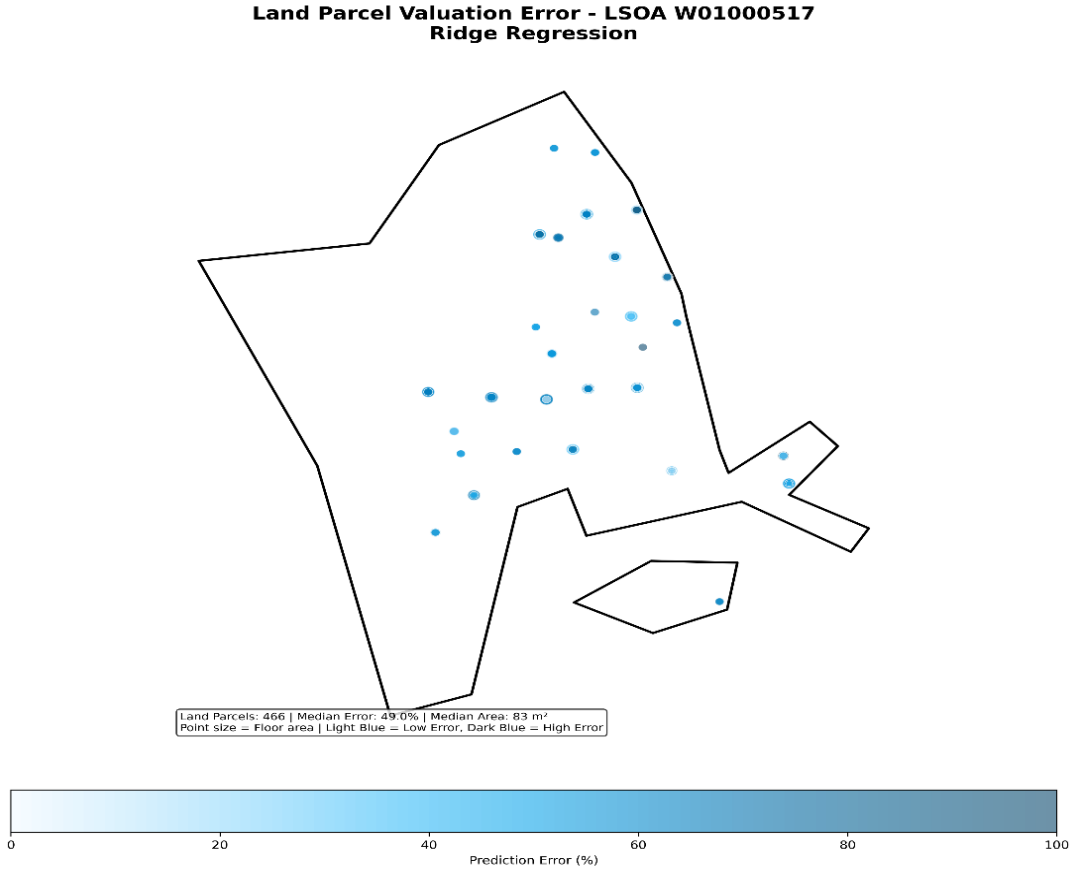


Figure 6: Ridge Regression Valuation Errors – LSOA W01000517 (Ceredigion 002D)

Ridge Regression Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

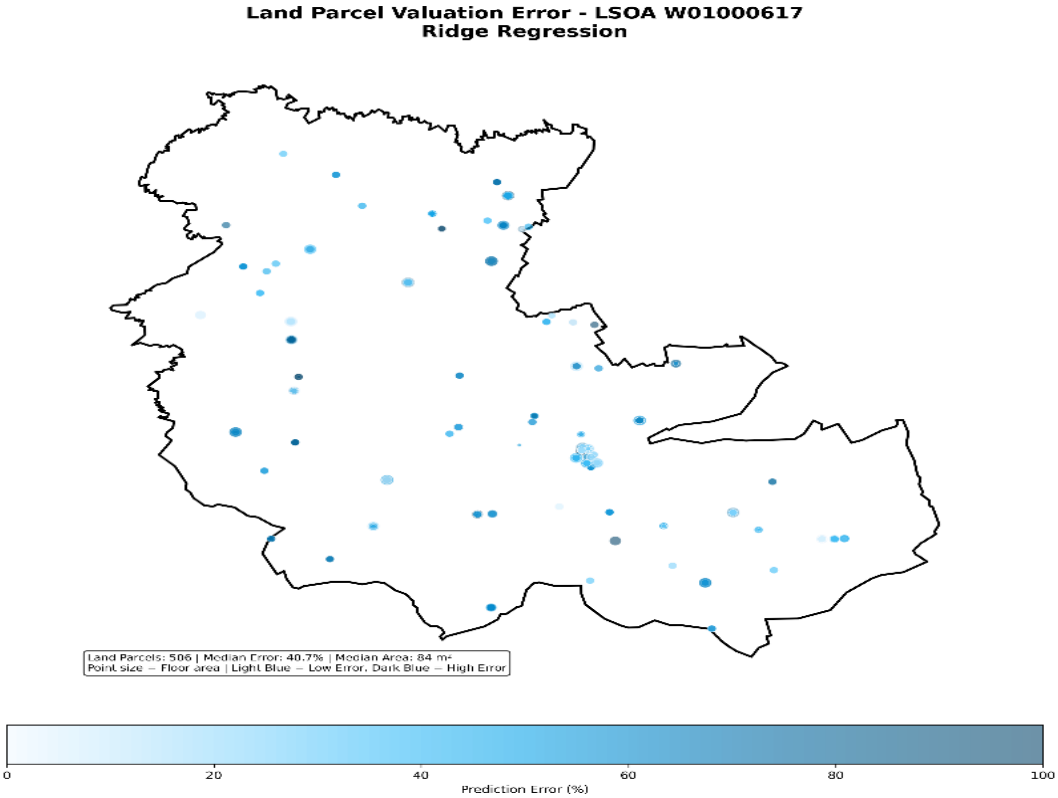


Figure 7: Ridge Regression Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

Ridge Regression Valuation Errors – LSOA W01001045 (Bridgend 019D)

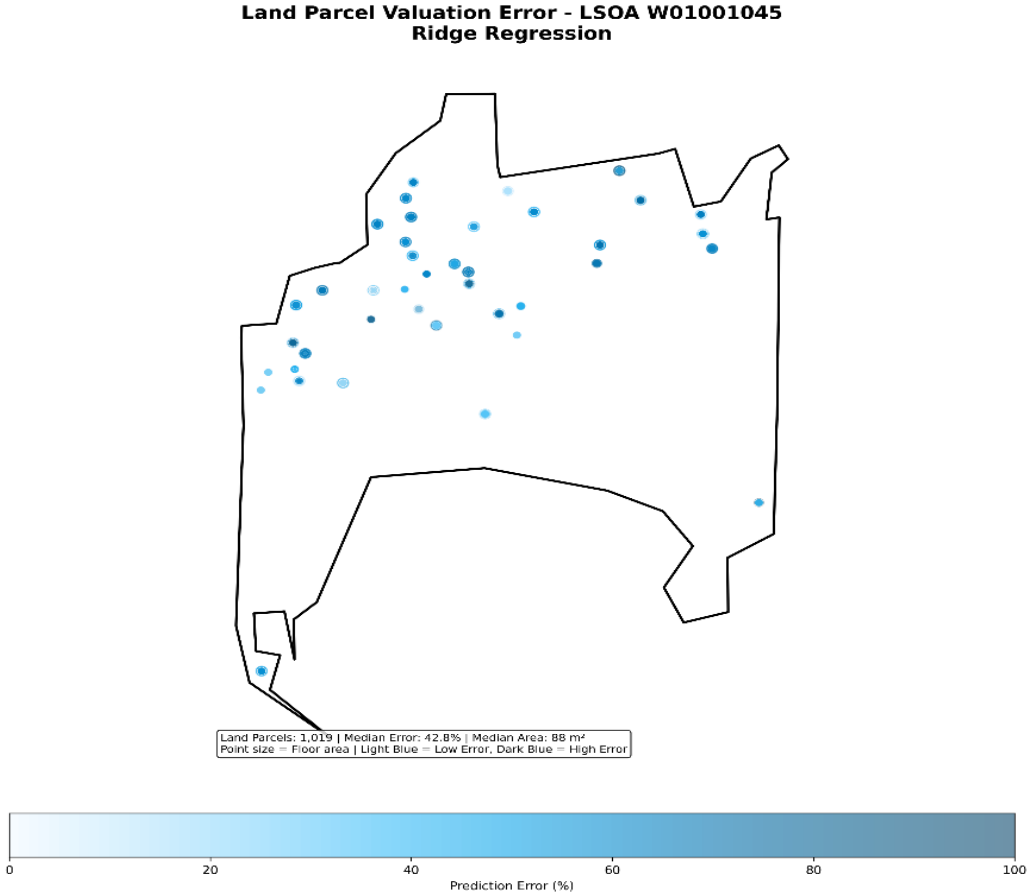


Figure 8: Ridge Regression Valuation Errors – LSOA W01001045 (Bridgend 019D)

Ridge Regression Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

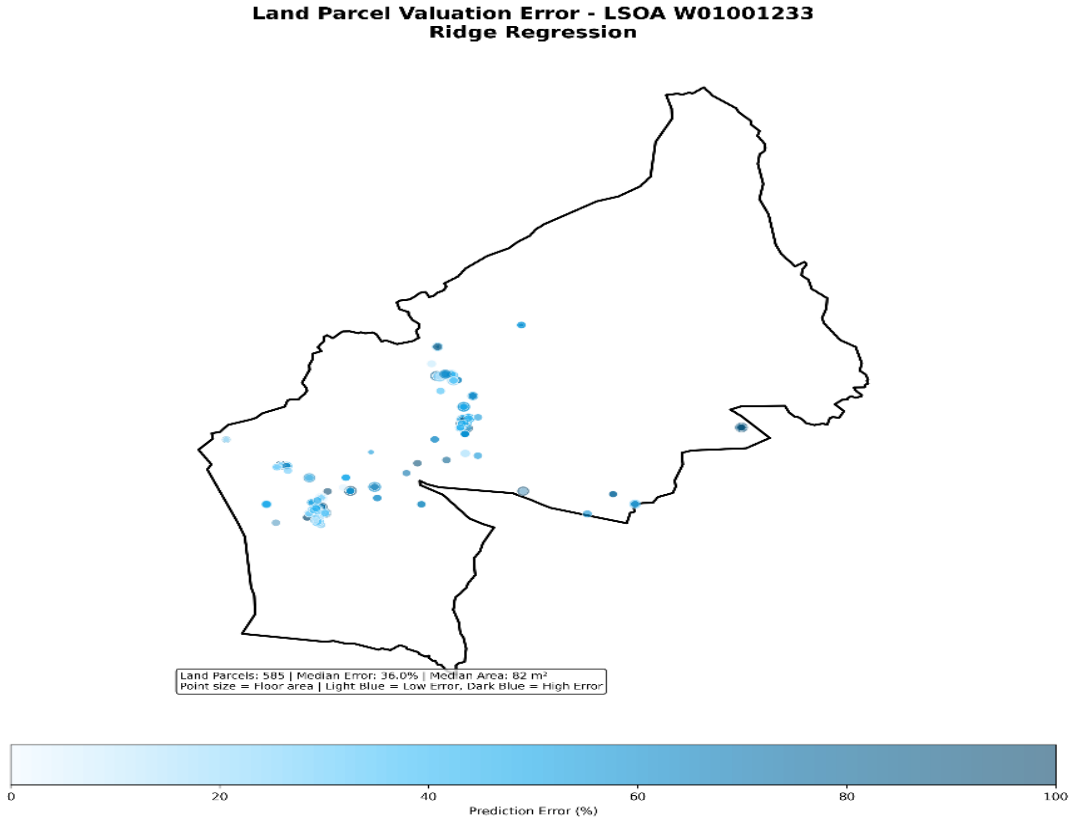


Figure 9: Ridge Regression Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

Ridge Regression Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

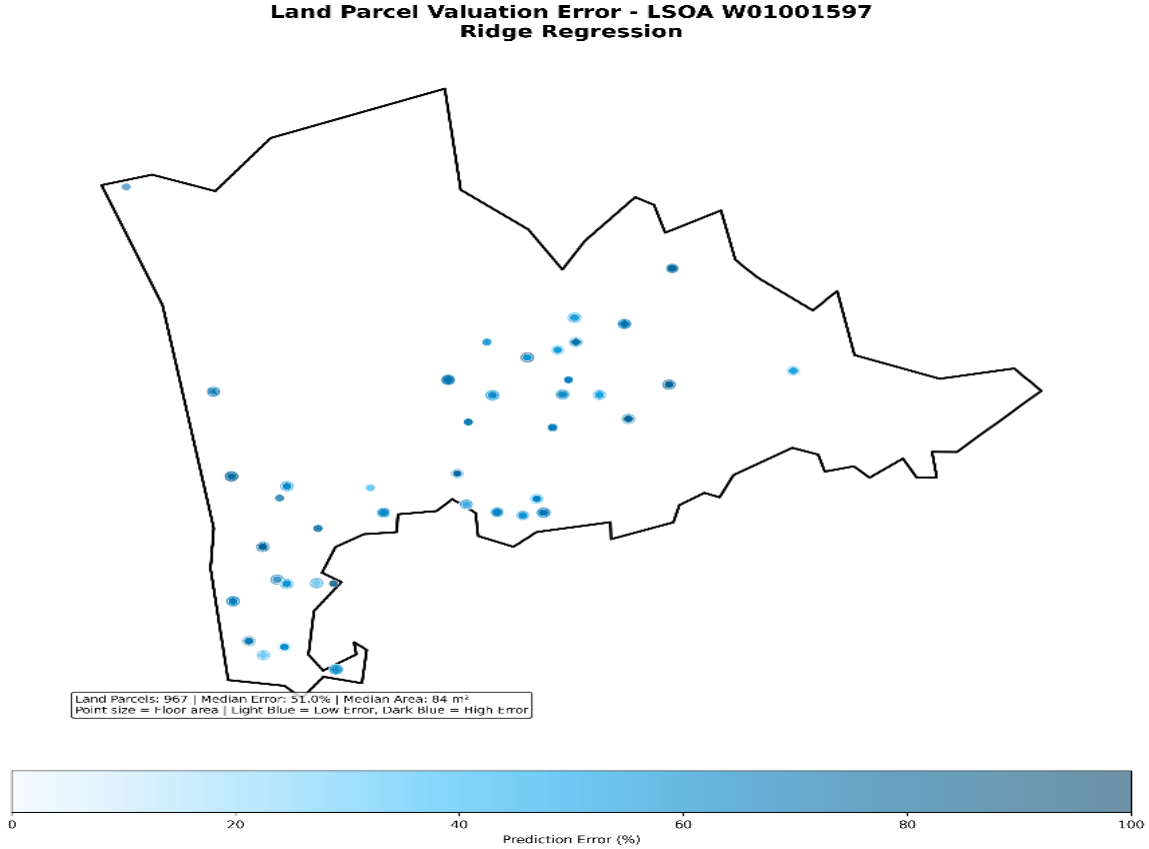


Figure 10: Regression Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

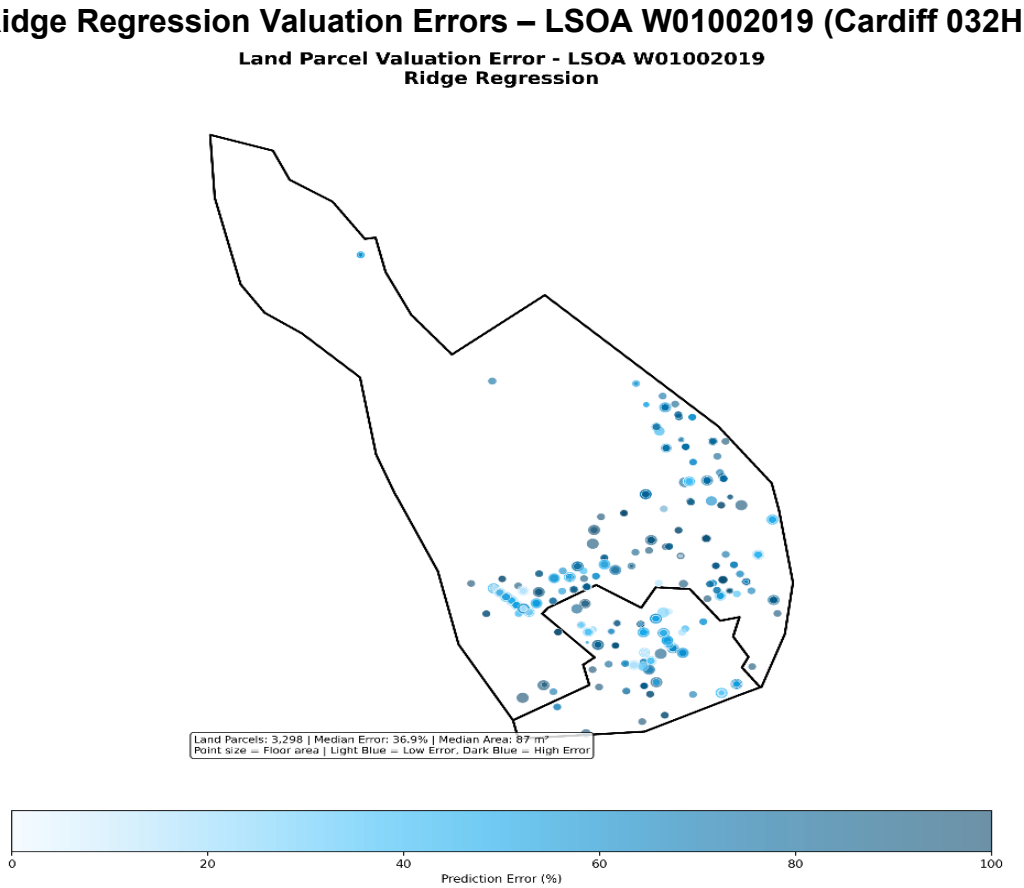


Figure 11: Ridge Regression Valuation Errors – LSOA W01002019 (Cardiff 032H)

2.8. Model 2 Gradient Boosting (Machine Learning Approach)

Gradient Boosting is a leading machine learning approach for structured, tabular data and is widely used in commercial automated valuation models. In this study, it is implemented using the CatBoost algorithm and trained on a large sample of Welsh land transactions. The method is designed to capture non-linear relationships and complex interactions between features without requiring large numbers of manually engineered variables. Conceptually, Gradient Boosting builds an ensemble of many simple decision trees. The process begins with a basic prediction and then adds trees one by one, where each new tree is trained to correct the errors made by the existing ensemble. Over hundreds of such iterations, the model progressively improves its fit to the data and becomes capable of representing subtle and complex patterns in land values.

This approach is particularly suitable for land valuation because it can:

- represent non-linear relationships, such as diminishing returns to additional floor area
- model interactions between location and land type
- accommodate changes in market conditions over time.

Mechanics of Gradient Boosting

The core idea behind Gradient Boosting can be described step by step.

1. The model begins with a baseline prediction. In this study, the average of the log-transformed sale prices across all training land transactions.

2. A first decision tree is fitted to the residuals, that is, the difference between the observed log price and the baseline prediction. This tree identifies groups of land transactions where the baseline is too high or too low.
3. The corrections from this tree are added to the predictions with a small multiplier, known as the learning rate.
4. A second tree is then fitted to the remaining errors, and its corrections are again added to the model's predictions.
5. This process is repeated many times. Each tree makes small adjustments, but together they build a powerful predictive model.
6. The final prediction is the sum of the baseline and the contributions from all trees in the ensemble.

Decision trees are able to learn non-linear patterns naturally. For example, they can capture that the first fifty square metres of floor area adds more value than the next fifty, or that a detached house in Cardiff behaves differently from a detached house in a rural part of Powys. By layering many such trees, especially with a controlled learning rate, Gradient Boosting produces a model that is both flexible and well controlled.

Choice of Algorithm: CatBoost

CatBoost (Categorical Boosting) was chosen for several reasons that are particularly relevant for a national land valuation context.

Native treatment of categorical variables
CatBoost is designed to handle categorical features directly. Instead of requiring extensive one-hot encoding, it uses target statistics and ordered boosting techniques to learn from categories such as land type, postcode district and year of sale.

Protection against overfitting
CatBoost uses symmetric tree structures and an ordered boosting procedure which help reduce overfitting, especially when the model has many trees. This is important in a setting with strong local patterns, such as housing markets, where naïve tree-based models can overfit to small neighbourhoods.

Efficient training and handling of missing values
The algorithm is optimised for large datasets and includes mechanisms for handling missing values directly, for example through surrogate splits. This reduces the need for complex pre-processing.

These land transactions make CatBoost a good choice for modelling a complex, high-dimensional dataset such as the Welsh land transaction record, where many key predictors are categorical.

Hyperparameter Optimisation with Optuna

The performance of CatBoost depends on several hyperparameters that govern the behaviour and complexity of the model. In this study, five key hyperparameters were optimised using the Optuna Bayesian optimisation framework:

- learning rate (controls how much each tree can change the overall prediction)
- tree depth (controls how complex individual trees can be)
- L2 leaf regularisation (penalty that helps prevent the model from fitting noise)
- bagging temperature (controls randomness in row sampling)
- random strength (controls randomness in the selection of splitting conditions).

Optuna was configured to run 30 trials, each trial testing a different combination of these hyperparameters. Within each trial, performance was evaluated using GroupKFold cross-validation, where land transactions were grouped by LSOA. This ensured that all land transactions from a given LSOA appeared in the same fold, preventing the model from learning LSOA-specific price levels during training and then being evaluated on the same LSOA during validation.

The search produced the following preferred configuration:

- learning rate of approximately 0.0518
- maximum tree depth of eight levels
- L2 leaf regularisation of approximately 9.87
- bagging temperature of 0.79
- random strength of 0.27
- a maximum of one thousand trees, with early stopping if validation performance did not improve for fifty iterations.

Prices were modelled on a log scale. Predicting log(price) rather than raw price helps to stabilise variance and reduce the influence of very high-value outliers.

Training Strategy

Several design choices were made to ensure that the model trained effectively and produced robust results.

Year-stratified sampling
The training sample was constructed to achieve a balanced representation across sale years from 1995 to 2023.This mitigates the risk that the model becomes overly tuned to recent years at the expense of older transactions or vice versa.

GroupKFold cross-validation by LSOA
As described in the methodology section, cross-validation for hyperparameter tuning used GroupKFold splits based on LSOA. This ensured that during validation, the model never saw land transactions from the validation LSOA in its training folds. This is particularly important for assessing geographic generalisation, which is central to the aims of this project.

Final training
Once the optimal hyperparameters were selected, a single CatBoost model was trained on the full training set. This final model was then evaluated on the nine test LSOAs.

Features Used in the Final Model

The final Gradient Boosting model uses a compact set of eight features. Five are numerical, and three are categorical.

Numerical features

log_area

The natural logarithm of total floor area in square metres. This transformation allows the model to capture diminishing returns to additional area. Missing values are imputed using a hierarchical approach:

Tier one: median for the combination of land type, postcode district and year of sale
Tier two: median for the combination of land type and postcode district
Tier three: median for land type only
Tier four: global median across all land transactions.

new_build

A binary indicator = 1 if the land is recorded as a new build at the time of sale and 0 otherwise. New build land transactions typically command a premium of approximately 15 to 20% compared with comparable existing dwellings.

leasehold

A binary indicator equal to one if the tenure is leasehold and zero if freehold. Leasehold land transactions typically sell for between 5 and 10% less than similar freehold land transactions, largely due to ground rent and other obligations.

month_sin

A sine transformation of the month of sale, which captures seasonal variation in a cyclical way, such as the spring market uplift.

month_cos

A cosine transformation of the month of sale, which complements the sine term and together allows the model to represent the full annual cycle of seasonal effects.

Categorical features

land_type

The type of land, with five categories: flat, terraced, semi-detached, detached and other. CatBoost learns how each type influences price using its categorical encoding mechanisms.

pcd_limited (postcode district)

The first part of the postcode, for example CF14, SA1, LL57 or NP23. The model considers the 150 most frequent districts individually and groups all remaining districts into an “Other” category. This captures location premiums at the postcode district level while avoiding overfitting to very rare districts.

year_str (year of sale)
The year in which the land sold, treated as a categorical string, for example “1995”, “2020” or “2023”. Handling the year as a category rather than a numeric value allows the model to learn non-linear year effects and structural breaks in the market, including inflation, cycles and unusual events.

Feature Selection Experiments and Rationale

To determine the most appropriate feature set, two configurations were compared on the nine held-out LSOAs. A minimal feature set of six variables, as described above. an extended feature set of twenty-three variables, which added:

1. non-linear area transformations such as area squared and the square root of area
2. additional temporal indicators including quarter and day of the week
3. Output Area Classification codes
4. Energy Performance Certificate variables such as energy rating and age band
5. more granular geographic identifiers including LSOA, MSOA, town and district
6. satellite-derived features such as land use and Agricultural Land Classification.

When evaluated on the test set of 9,606 land transactions, the extended feature set improved the average R-squared by about 1.2% points compared with the minimal set and produced a very small increase in mean absolute error of approximately £157. However, it worsened the mean absolute percentage error from approximately 35.3% to 38.02%. This indicates that although the extended model captured slightly more variance in prices, it did so by learning patterns that did not generalise as well to unseen areas, which is a hallmark of mild overfitting.

On this basis, the eight-feature model was selected. It offers:

- strong percentage-based accuracy, including the best within ± 20% accuracy among all models
- simpler operational requirements and fewer dependencies on external datasets
- wider coverage, since many of the extended features rely on EPC and satellite data that are not available for all land transactions
- reduced risk of overfitting to particular geographies or time periods
- greater interpretability for stakeholders who need to understand which factors drive valuations.

In effect, the selected eight-feature configuration delivers most of the predictive benefit of the extended model while maintaining much better generalisation and operational practicality.

Method

Before training the Gradient Boosting model, the dataset was processed to ensure adequate temporal coverage, removal of extreme outliers, and stability of model behaviour across changing market conditions.

Year-Stratified Sampling

The training pool consisted of all Welsh land transactions outside the nine held-out Lower Layer Super Output Areas. To avoid bias toward recent years, which have a larger volume of recorded transactions, the training data were sampled evenly across the entire period from 1995 to 2023.

The procedure was as follows:
All land transactions in the training pool were grouped by year of sale.
1. For each year, a random sample of 17,241 land transactions was drawn. If a year contained fewer land transactions, all available records were included.
2. These yearly samples were then combined to form the final training dataset.
3. This resulted in 499,999 training land transactions, providing balanced representation across all market cycles.

This stratification prevents the model from over-fitting to more recent years and ensures that long-term market trends are reflected fairly in training.

Outlier Filtering
The training data were filtered to remove land transactions with sale prices in the top 0.1% of the distribution. Approximately 500 extremely high-value transactions, including land transactions valued at more than 5 million pounds, were excluded. These rare transactions can disproportionately influence gradient updates and reduce generalisation performance, especially in geographically diverse markets.

Target Transformation
The model predicts the natural logarithm of sale price rather than raw price. Log transformation stabilises variance, reduces the influence of very high-value transactions and improves the numerical land transactions of gradient boosting algorithms. Predictions are then exponentiated and converted back into pounds for evaluation and reporting.

Hyperparameter Optimisation Using Optuna
The performance of a Gradient Boosting model depends critically on its hyperparameters. These determine the learning rate, depth of decision trees, level of regularisation and degree of randomness introduced during model fitting.

To identify a suitable configuration, the Optuna optimisation framework was used. Optuna is a Bayesian optimisation tool that explores hyperparameter spaces efficiently through probabilistic modelling. For this study, it evaluated thirty candidate hyperparameter configurations.

The objective of the optimisation was to maximise the average predictive accuracy across GroupKFold cross-validation folds, measured using the R-squared statistic on log-transformed prices.

Hyperparameter Search Space
Five hyperparameters were optimised, each with clearly defined bounds:
1. Learning rate: between 0.01 and 0.3, on a logarithmic scale.
2. Tree depth: between 4 and 10 levels.
3. L2 leaf regularisation: between 0.01 and 10.0, on a logarithmic scale.

4. Bagging temperature: between 0.01 and 1.0.
5. Random strength: between 0.01 and 1.0.

This five-dimensional space defines the structure and behaviour of the Gradient Boosting ensemble.

Cross-Validation Within Each Optuna Trial

For each of the thirty trials:

1. Optuna proposed a set of hyperparameters drawn from the search space.
2. The training data were divided into several folds using GroupKFold, grouping land transactions by LSOA. This ensured that all land transactions from a given LSOA were placed in the same fold.
3. For each fold:
    a. CatBoost was trained on the remaining folds, representing approximately 400,000 land transactions.
    b. Predictions were generated for the held-out fold, representing approximately 100,000 land transactions.
    c. The R-squared value on log-transformed prices was calculated.
4. The validation results across all folds were averaged.
5. This mean validation score was returned to Optuna as the measure of quality for that configuration.
6. Optuna updated its internal model of the hyperparameter space and selected the next configuration for testing.

Outcome of Optimisation

After the thirty trials were completed:

- The best average cross-validation R-squared was approximately 0.78
- The identified hyperparameters matched those listed in the main architecture section.

The best configuration produced highly stable performance across the folds:

- Fold one: 0.7803
- Fold two: 0.7799
- Fold three: 0.7807
- Fold four: 0.7797
- Fold five: 0.7791

The mean of these values was approximately 0.7800 with a standard deviation of approximately 0.0005. The low standard deviation indicates that the model generalises consistently to unseen LSOAs within the cross-validation framework.

Final Model Training

After selecting the optimal hyperparameters, the final training process involved:

1. Training a single CatBoost model on the entire training dataset of 499,999 land transactions.
2. Using all hyperparameters identified by the Optuna search.
3. Allowing the model to train for up to one thousand iterations, with early stopping if validation accuracy failed to improve for fifty consecutive iterations.

4. Evaluating the final model exclusively on the fully independent test set comprising 9,606 land transactions across the nine held-out LSOAs.

It is important to note the conceptual distinction between cross-validation and final evaluation. Cross-validation results were used only to tune hyperparameters. All reported performance metrics reflect predictions on the nine held-out LSOAs, which were never used in model development or optimisation.

Results

The Gradient Boosting model was evaluated on a fully independent test set consisting of 9,606 land transactions located within nine Lower Layer Super Output Areas. These areas were entirely excluded from model development and therefore provide a reliable assessment of out-of-area generalisation. The results summarised here reflect the model's performance when confronted with housing markets it has never seen before.

Explained variance within LSOAs
Across the nine held-out LSOAs, the model explains approximately 28.2% of the variation in sale prices. This measure reflects how well the model captures differences in prices within each individual test area. This level of explanatory power represents the highest achieved among the traditional machine learning models evaluated in this study.

Explained variance across all LSOAs combined
When the entire test set is pooled, the overall R-squared value falls to approximately 1.7%. This difference arises because the overall measure penalises models that do not accurately capture the substantial differences in price levels between LSOAs. Per-LSOA accuracy focuses only on within-area variation and does not evaluate whether the model correctly reflects that some LSOAs are much more expensive than others. Because Model 2 was trained to capture property-level patterns rather than LSOA-level price offsets, its performance appears much stronger on the within-LSOA measure than on the pooled measure.

Mean absolute error
The average difference between predicted and actual sale prices is £76,392. This provides a sense of the typical magnitude of the valuation error in monetary terms.

Mean absolute percentage error
When pooling all 9,606 test properties, mean absolute percentage error is 38.2%. This reflects the relative error rather than the absolute error and is a more demanding measure, particularly in lower-value markets where moderate monetary errors translate into large percentage deviations.

Accuracy within plus or minus twenty%
The model values approximately 43.6% of land transactions within ±20% of their actual sale price, representing 4,188 land transactions out of the total of 9,606. This is the highest percentage-based accuracy recorded among all models tested. For comparison, the Ridge Regression model achieves 36.0% and the artificial intelligence multi-agent system achieves 28.6%.

Variation Across the Nine Test LSOAs

Performance varies significantly across test areas, indicating that local market structure plays an important role in model behaviour.

Table 14: Performance of CatBoost Model by Test LSOA

LSOA Code	Number of Properties	R^2 Score	Mean Absolute Error (£)	Mean Absolute Percentage Error (%)	Mean Sale Price (£)
W01000449	867	0.518	43,869	27.8	149,216
W01001045	1,019	0.488	35,160	36.1	132,949
W01001597	967	0.485	68,428	36.7	245,573
W01000517	466	0.383	48,120	34.3	155,942
W01000114	807	0.326	42,423	55.8	99,900
W01000617	506	0.247	75,807	48.0	178,035
W01000255	1,091	0.056	63,460	27.0	185,215
W01001233	585	0.034	62,925	41.5	150,537
W01002019	3,298	0.002	165,983	36.6	374,063

Highest accuracy

The strongest results are observed in LSOA W01000449, which contains 867 land transactions. In this area, the model achieves an R-squared value of approximately 51.8%. The mean absolute error is £43,869 and the mean absolute percentage error is 27.8%. This LSOA has relatively homogeneous housing stock and a stable distribution of prices, which allows the model to capture relationships more effectively.

Lowest accuracy

The weakest results occur in LSOA W01002019, which contains 3,298 land transactions and is the largest and most expensive area in the test set, with a mean sale price of approximately £374,063. The model achieves an R-squared value of approximately 0.2% in this location. The mean absolute error rises sharply to £165,983, approximately five and a half times higher than in the best-performing LSOA. The mean absolute percentage error

is 36.6%. These results reflect the strong internal heterogeneity of this area, which includes multiple sub-markets with distinct price dynamics that are not fully captured by the minimal feature set used in this model.

Interpretation of R-squared Variation
Across the nine LSOAs, the R-squared values range from 0.2% to 51.8%. This represents substantial variation in explanatory power. Such variation is consistent with the structure of the Welsh housing market, where local characteristics, diversity of housing stock, and differences in transaction volumes can vary widely between areas.

The pattern can be interpreted as follows:

- Areas with higher R-squared values tend to have more homogeneous housing, clear location premiums and stable pricing behaviour. These areas allow the model to identify consistent relationships between features and sale prices.
- Areas with low R-squared values often contain greater internal variation, smaller sample sizes for specific combinations of property type and postcode district, or complex sub-markets that require more detailed location information than is available in the minimal feature set.

This variation underscores the importance of local context in property valuation and highlights the need for further feature enhancement or area-specific adjustments to improve performance in complex market environments.

Analysis

Interpretation of Geographic Variation
The wide range of R-squared values across the nine held-out LSOAs shows that the Gradient Boosting model generalises very differently depending on local market conditions. In some areas the eight-feature model captures price patterns relatively well, while in others there is substantial internal complexity that is not explained by property type, postcode district and year of sale alone. This reflects the fact that some markets are more homogeneous and data-rich, whereas others contain multiple sub-markets and diverse housing stock that would require more granular location or property attributes to model accurately.

Feature Importance Analysis
CatBoost provides estimates of feature importance based on Shapley values. These values measure how much each feature contributes to the model's predictions, on average, across all possible combinations and orderings of features. Unlike simple split-count or gain-based measures, Shapley-based importance accounts for interactions between features and provides a more robust indication of which variables drive model performance.

In this study, the feature importances, which sum to 100%, can be summarised as follows:

<u>Year of sale (year_str)</u>
Year of sale is the single most important feature, accounting for approximately 33.2% of total importance. This reflects the very strong influence of secular house price inflation and

market cycles between 1995 and 2023. It is common, for example, for an otherwise similar property sold in 2022, to be worth around twice as much as it would have been in 2009.

Property type (property_type)
Property classification contributes around 25.7%. The distinction between flats, terraced, semi-detached and detached housing is therefore a major driver of price differences. Detached properties, in particular, command significant premiums over flats even within the same postcode district.

Postcode district (pcd_limited)
Postcode district accounts for about 25.0% of importance. This feature captures large location premiums, for example between central Cardiff and ex-industrial valleys. Properties in certain Cardiff postcodes typically sell for multiples of otherwise similar properties in lower-demand areas.

Logarithm of floor area (log_area)
Floor area, modelled using a logarithmic transformation, accounts for around 9.7% of importance. This represents a significant predictor of value, as larger properties generally command higher prices.

New build indicator (new_build)
The new build flag contributes approximately 2.6%. New build properties generally attract a premium of between 15 and 20% relative to comparable existing dwellings.

Leasehold indicator (leasehold)
Leasehold tenure contributes around 2.2%. Leasehold properties tend to sell for approximately 5 to 10% less than freehold equivalents, reflecting additional obligations such as ground rent.

Seasonal sine component (month_sin)
The sine-transformed month variable contributes about 1.2%. This reflects modest seasonal effects, for example slightly higher prices and market activity in the spring.

Seasonal cosine component (month_cos)
The cosine-transformed month variable contributes about 0.5% and complements the sine term to represent the full annual cycle.

Taken together, the top three features (year of sale, property type and postcode district) account for approximately 83.9% of total feature importance. Structural characteristics such as floor area, new build status and tenure together contribute around 14.5%, while seasonal timing contributes about 1.7%.

Key Implications

These findings suggest that in the Welsh context, residential property prices are determined predominantly by:

- when the property is sold (year of sale),
- what type of property it is (for example detached compared with flat), and
- where it is located (postcode district).

The detailed physical attributes of the building, such as exact floor area, have moderate influence at the national modelling level, although they may still be important within specific local markets.

From an operational perspective, this has several implications:
1. Mass appraisal models can achieve reasonable levels of accuracy using a relatively small number of administrative and locational features, without requiring detailed inspection data for every dwelling.
2. Individual valuations for specific properties, especially in high-value or atypical areas, will continue to require local market expertise because otherwise similar properties can differ by large factors depending on location.
3. Temporal recalibration is essential. A model trained on data from 2015 to 2019 will underestimate prices in 2022 unless year effects are updated to reflect recent market movements.

Limitations

Gradient Boosting, implemented through CatBoost, is the strongest predictive model in this study. However, there are several important limitations that must be recognised before considering it for operational deployment in a national or statutory valuation setting. These limitations concern the nature of the target being predicted, the constraints imposed by data quality, the behaviour of the model in unobserved settings, and issues of transparency and governance.

Inability to separate land value from structure value
The Gradient Boosting model is trained to predict total sale price. It does not estimate land value and structure value separately. Yet for many policy questions, including land value taxation, infrastructure charging and some planning interventions, it is the land component that is of primary interest.

In principle, it might be possible to train a second Gradient Boosting model to predict structure value based on building characteristics such as floor area, construction age and property type, and then derive land value as a residual by subtracting this predicted structure value from the observed sale price. In practice, experiments of this type produced unsatisfactory and economically implausible results, including very high rates of negative land values. This occurs because location affects both the land and the structure simultaneously. Properties in high-value areas tend to have both more expensive land and better quality buildings, and the structure-only model inadvertently absorbs some of the locational premium. When this inflated structure estimate is subtracted from total price, the residual land value becomes negative for many properties.

These findings indicate that Gradient Boosting, when trained solely on administrative and Energy Performance Certificate data, cannot provide a reliable decomposition of land and

structure values. For this type of analysis, distinct, theory-driven methods such as Depreciated Replacement Cost are required. The Gradient Boosting model should therefore be understood strictly as a total value predictor.

Limited feature set, driven by partial and noisy data
The Gradient Boosting model uses eight features. This set was deliberately kept small because it is the only combination of variables that is available, relatively stable and of reasonable quality for almost all transactions in the dataset. In other words, the model does not rely on rich property inspection data. Instead, it uses features that can be derived from the Land Registry and simple linkages.

A wider set of potential predictors exists. However, many are omitted because their coverage is partial, their quality is inconsistent, or their inclusion would reduce the representativeness and robustness of the model. Including such variables would mean that the model could only be applied to a subset of properties, would increase the risk of bias toward specific regions or time periods, and would make performance more sensitive to measurement error.

Examples of omitted structural characteristics:

- Number of bedrooms, bathrooms and reception rooms is not consistently recorded across all properties and years, and where it is available, recording practices differ by source.
- Information on garages, parking provision and garden size is often missing, outdated, or derived from unstructured text that is difficult to process reliably.
- Measures of property condition, refurbishment status, internal specification or architectural quality are rarely available in structured administrative datasets and would require manual or commercial data sources that are not within the scope of this study.

Examples of omitted locational characteristics:

- Distances to primary and secondary schools and school quality metrics are available for some years and locations but not universally and not always in consistent form.
- Public transport accessibility, including proximity to train stations or high-frequency bus corridors, is not captured in a uniform national dataset that can be joined reliably to every transaction over nearly three decades.
- Crime rates and deprivation indices exist at various geographic levels, but their resolution and time coverage are uneven, and matching them to long time series of land transactions introduces substantial complexity
- Environmental indicators such as air quality, noise levels and access to green space are often available only for recent years or specific metropolitan areas, rather than Wales-wide and for the full period from 1995.

Examples of omitted planning and legal constraints:

- Conservation areas and listed building registers are maintained by local planning authorities and national heritage bodies but do not exist in a single, standardised, longitudinal dataset.

- Flood risk maps have evolved over time, with changing methodologies and spatial resolutions, making it difficult to use them consistently over a twenty-nine-year period.
- Detailed information on planning permissions, density limits and development capacity is not straightforward to encode in a nation-wide, property-level dataset.

For these reasons, many potentially useful predictors are deliberately omitted. Their inclusion would reduce data coverage significantly and would risk introducing systematic bias in favour of better-documented areas and more recent time periods. The current eight-feature model therefore balances predictive power with universality of application and data stability. However, this inevitably limits the maximum accuracy that can be achieved, explaining only around 28.2% of the within-LSOA variation in prices, with the remaining 71.8% reflecting factors that cannot be observed in the current data.

Limited ability to extrapolate beyond observed data
Tree-based Gradient Boosting models learn piecewise constant relationships. Each decision tree partitions the feature space into regions and assigns a constant prediction within each region. This is powerful within the range of the training data but performs poorly outside it.

As a result:

- Properties in new postcode districts with no historical transactions in the training data will receive predictions based on the behaviour of other, partly similar postcodes, rather than location-specific dynamics. This may be acceptable in some contexts but carries risk where new development areas differ systematically from existing ones.
- Very large dwellings, for example houses with floor areas above 400 square metres, may be mis predicted if the training data contain very few examples in this size range.
- Structural market changes, such as the 2008 financial crisis or the 2020 pandemic, cannot be anticipated in advance. The model will only reflect such shifts after retraining on new data that include those events.

This is an inherent characteristic of Gradient Boosting and cannot be eliminated. It means that caution is required when applying the model to properties that lie outside the historical range of the training data or in markets undergoing rapid change.

Higher cross-validation accuracy than test-set accuracy
Hyperparameter optimisation using GroupKFold cross-validation on log-transformed prices produced mean R-squared values around 78%. However, when the final model was evaluated on raw prices in the nine held-out LSOAs, the average explanatory power fell to approximately 28%.

This discrepancy arises for several reasons:

- Validation performance was measured on log-transformed prices, which compresses the distribution and down-weights extreme values. Test performance was calculated on actual prices in pounds.

- Cross-validation folds contained a broad mix of LSOAs from across Wales, whereas the test set consists of nine specific LSOAs that were deliberately chosen and may have more complex or atypical characteristics.
- The held-out LSOAs represent a true out-of-geography test. If they include coastal tourism markets, commuter belts, or post-industrial areas that differ materially from the majority of the training data, the model's performance will naturally be lower than in the cross-validation setting.

This highlights that cross-validation scores are useful for model selection and tuning, but they cannot be treated as direct estimates of national operational performance. Only the independent test set on the nine excluded LSOAs provides that.

Large variation in performance between local areas
The model's performance varies significantly across the nine LSOAs in the test set. R-squared values range from 0.2% in the weakest area to 51.8% in the strongest, a difference of more than two orders of magnitude. In some LSOAs the model explains a large share of price variation, while in others it performs at or near the level of a simple benchmark that assigns the same mean price to all properties in the area.

This variation has several implications:

- The model is strongly dependent on local data patterns. Homogeneous, well-represented markets are easier to model accurately than complex, heterogeneous ones with multiple sub-markets and limited data.
- It is not straightforward to predict in advance which LSOAs will fall into which category. Areas with similar average prices can have very different internal patterns.
- For operational use, the model would need either geographic stratification (for example, separate models for urban and rural areas or for different regions) or a meta-modelling layer that adjusts predictions using local calibration factors.

In addition, any deployment would require systematic monitoring of performance by area and, ideally, uncertainty estimation, so that valuations in areas with low historical accuracy can be flagged as less reliable.

Sensitivity to temporal change and non-stationarity
Feature importance analysis shows that year of sale accounts for approximately 33.2% of total model importance. This means the model relies heavily on the time dimension to capture market cycles and inflation. While this is appropriate in a dynamic market, it has important consequences.

A model trained on data up to 2019 will systematically misestimate prices in later years if there is a structural change in the market. For example, the rapid price growth observed during the 2020 to 2022 period would not be reflected unless the model is retrained. Historical relationships between year and price may also change if macroeconomic conditions, credit availability, or policy interventions shift.

For these reasons, any operational use of Gradient Boosting would require:

- regular retraining of the model on updated data, for example annually or quarterly,

- validation procedures that test performance on future time periods rather than random splits, and
- potentially a set of additional features capturing macroeconomic conditions that influence prices at the national or regional level.

Computational demands and retraining constraints
The process of training the Gradient Boosting model with full hyperparameter optimisation is computationally intensive. The Optuna optimisation ran 30 trials, each involving GroupKFold cross-validation with five folds. This resulted in 150 training cycles, each involving up to one thousand trees and several hundred thousand land transactions.

On modern hardware the full optimisation process takes on the order of eight to twelve minutes. While this is acceptable for periodic offline model training, it is too heavy for continual retraining or interactive recalibration. In practice, a national system would need to:

- train the model offline on a scheduled basis,
- store the resulting model artefact, and
- monitor performance over time to decide when retraining is required.

This is feasible, but it constrains how frequently the model can be updated and how quickly it can respond to rapidly changing market conditions.

Complexity and interpretability
Finally, Gradient Boosting is intrinsically more complex than linear models. The predictions generated by CatBoost are the result of many layers of interactions across up to one thousand trees. While feature importance analysis shows which variables matter most in aggregate, it does not provide simple rules that can be communicated easily to property owners, tribunals or the public.

This raises challenges for:

- appeals processes, where individuals may wish to understand the basis of their valuation;
- regulatory review, where decision-makers may prefer methods with clear, direct relationships between inputs and outputs; and
- fairness and equality assessments, which require detailed examination of how model outputs vary across different groups.

Performance per LSOA

CatBoost – Test Performance by LSOA, R^2 and MAE

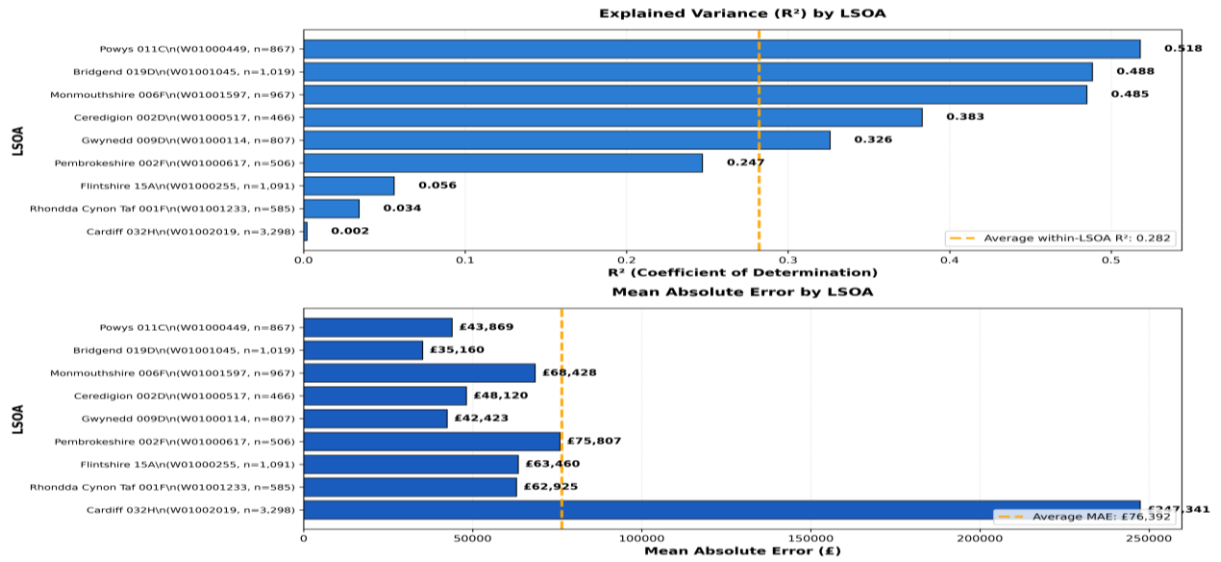


Figure 12 : CatBoost – Test Performance by LSOA, R² and MA

CatBoost Valuation Errors – LSOA W01000114 (Gwynedd 009D)

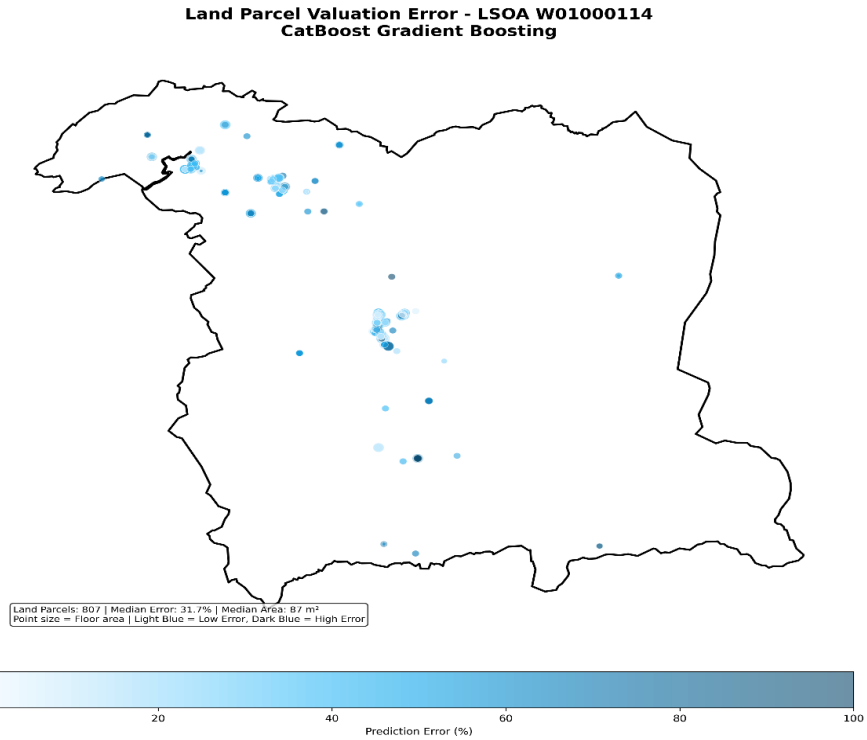


Figure 13: CatBoost Valuation Errors – LSOA W01000114 (Gwynedd 009D)

CatBoost Valuation Errors – LSOA W01000255 (Flintshire 015A)

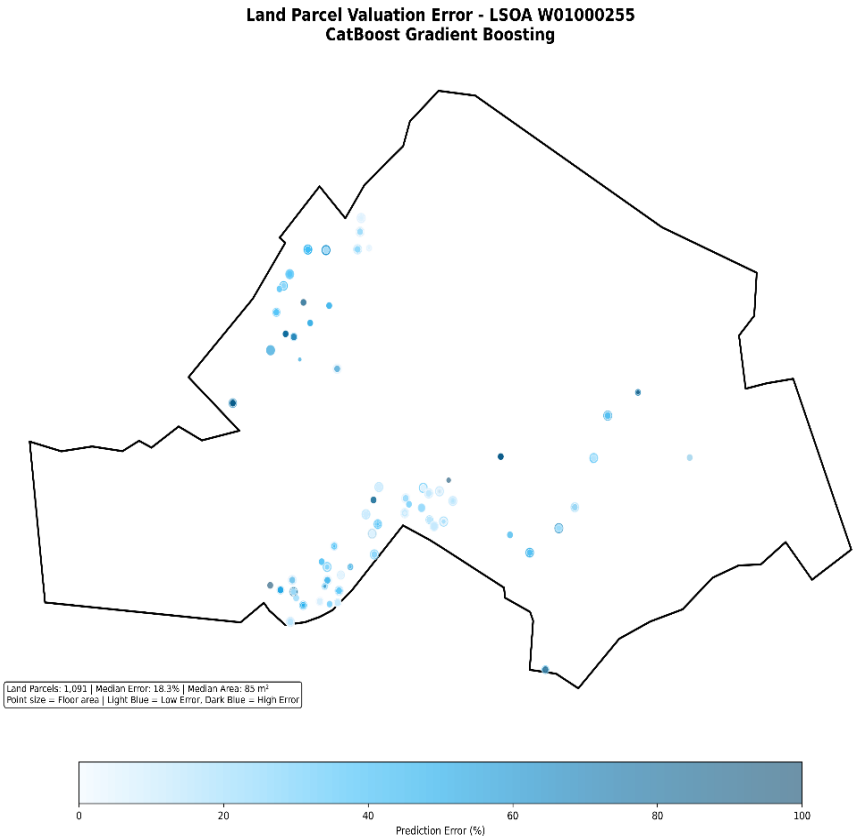


Figure 14: CatBoost Valuation Errors – LSOA W01000255 (Flintshire 015A)

CatBoost Valuation Errors – LSOA W01000449 (Powys 011C)

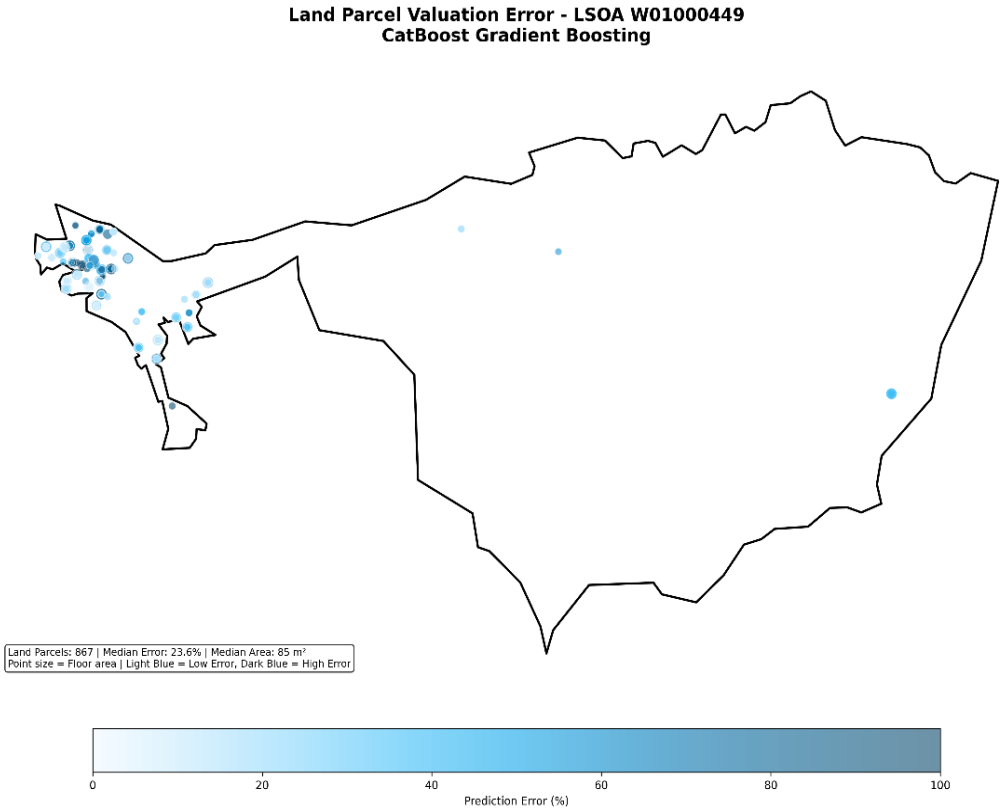


Figure 15: CatBoost Valuation Errors – LSOA W01000449 (Powys 011C)

CatBoost Valuation Errors – LSOA W01000517 (Ceredigion 002D)

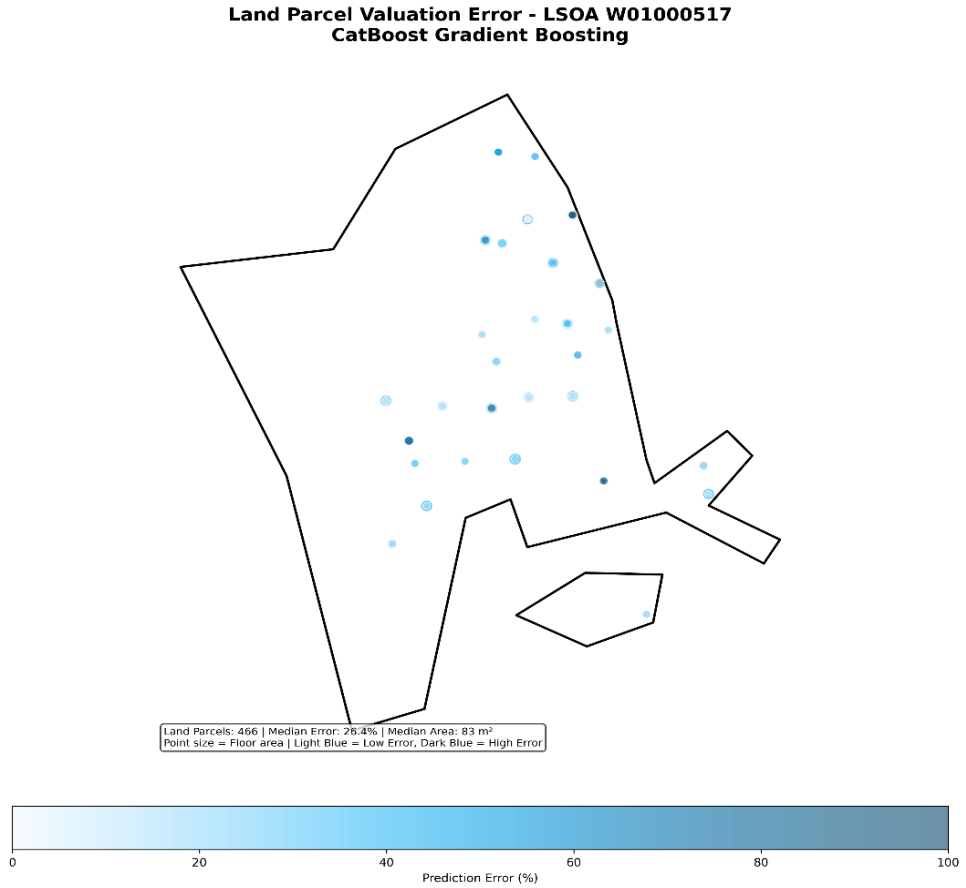


Figure 16: CatBoost Valuation Errors – LSOA W01000517 (Ceredigion 002D)

CatBoost Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

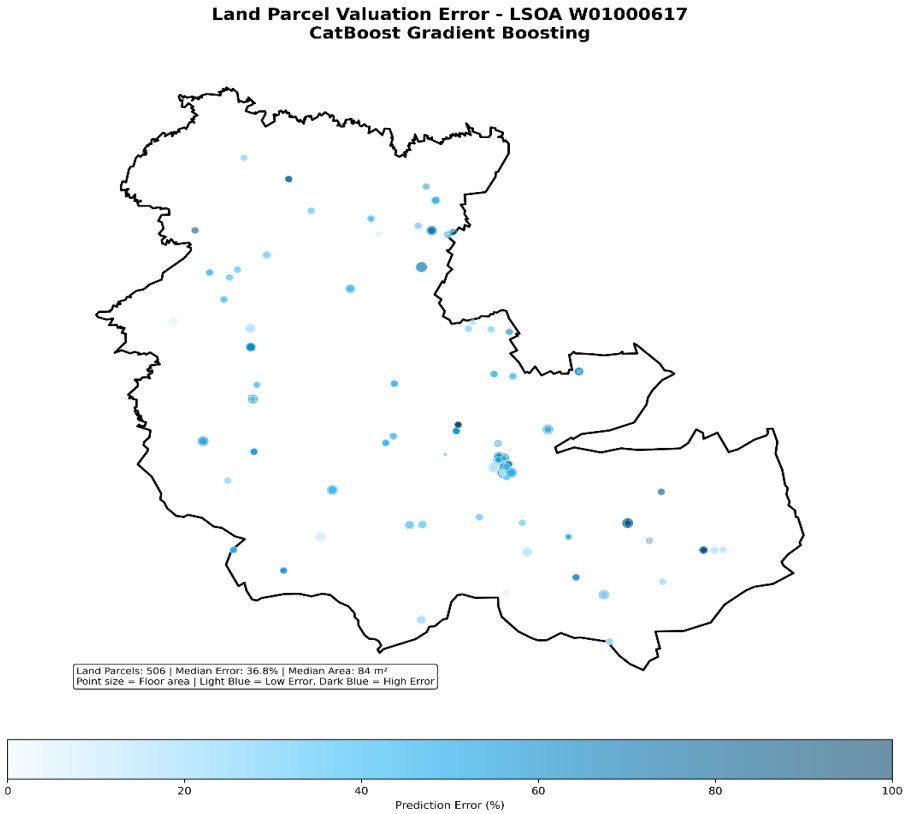


Figure 17: CatBoost Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

CatBoost Valuation Errors – LSOA W01001045 (Bridgend 019D)

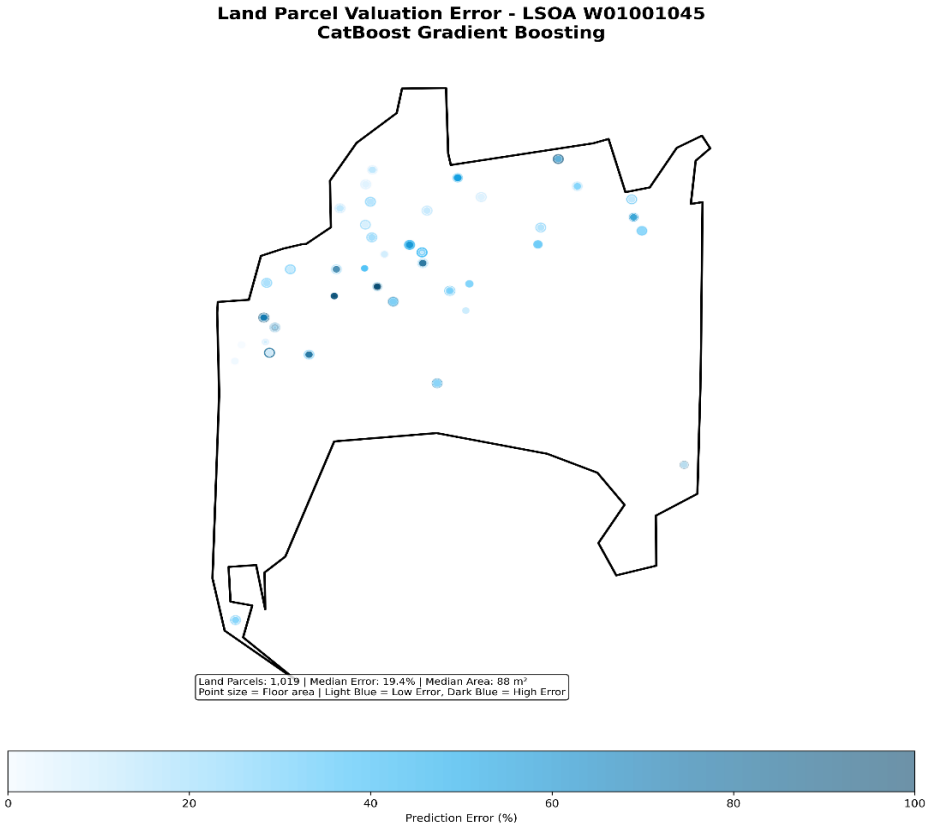


Figure 18: CatBoost Valuation Errors – LSOA W01001045 (Bridgend 019D)

CatBoost Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

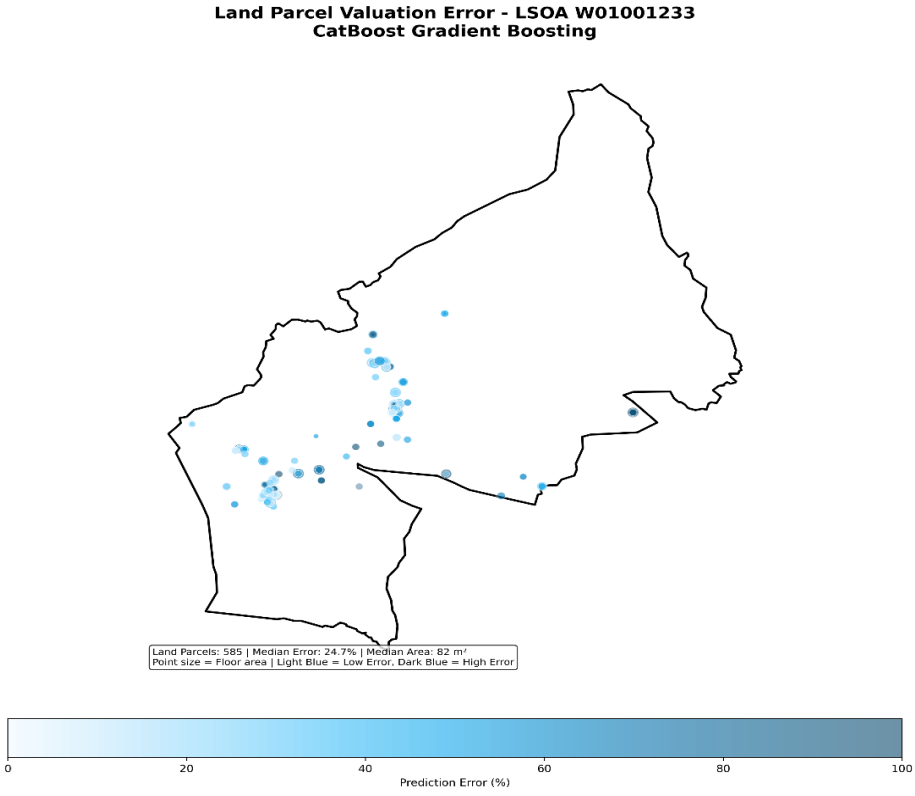


Figure 19: CatBoost Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

CatBoost Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

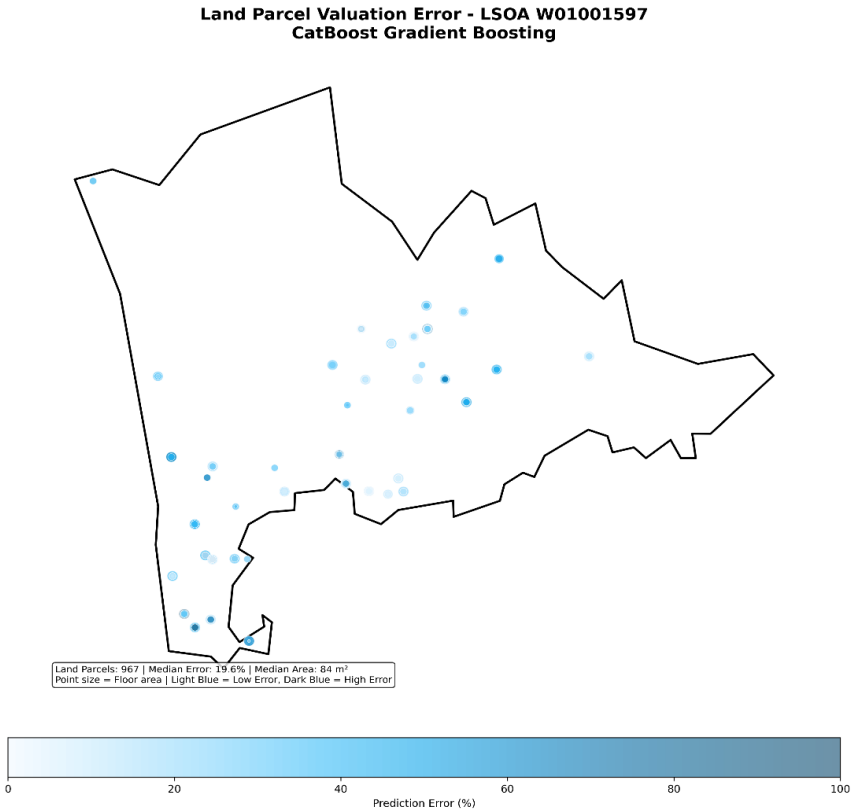


Figure 20: CatBoost Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

CatBoost Valuation Errors – LSOA W01002019 (Cardiff 032H)

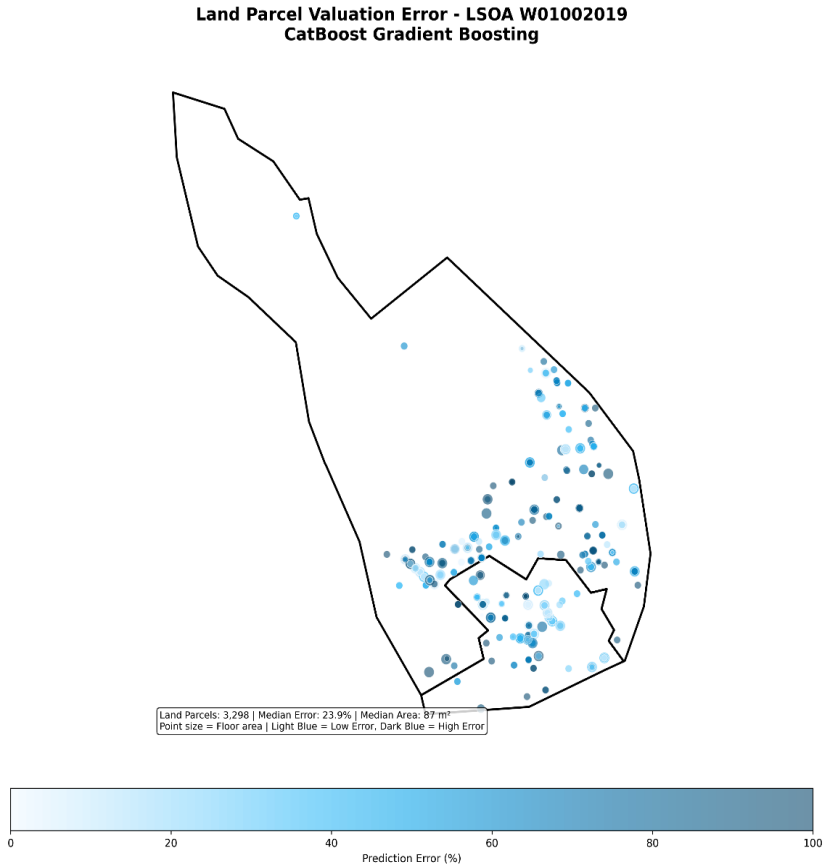


Figure 21: CatBoost Valuation Errors – LSOA W01002019 (Cardiff 032H)

2.9. Model 3 (KNN)

Early experimentation with manual defined fuzzy logic systems confirmed that rule-driven approaches, while transparent, could not capture the full complexity of the Welsh housing market at national scale. Although fuzzy systems allowed reasoning to be encoded explicitly, they were unable to adapt to local variation, relied on simplified linguistic categories and produced discrete outcomes that were too coarse to approximate real price distributions. The result was a system that provided interpretability but lacked the accuracy needed for policy or operational use.

This set the stage for the development of Model 3, which was designed to preserve the *spirit* of expert reasoning while addressing the technical limitations of rule-based systems. Rather than encoding valuation rules directly, Model 3 adopts a data-driven method that mirrors the central principle identified through expert consultation: valuations should be based on similar land transactions and comparable evidence.

Rationale for a Comparable-Driven Approach

The design of Model 3 is grounded in three key insights.

Valuers rely on comparables, not global equations.
Human valuers rarely apply a single formula across all land transactions. Instead, they search for similar transactions within the local area and adjust from those. Experts highlighted that a model which imitates this process would better reflect real-world valuation practice.

Local conditions matter significantly.

Experts emphasised that neighbourhood boundaries, construction differences and localised market sentiment can change within short distances. Global models may overlook these nuances. A method that responds to the closest comparable land transactions, rather than treating all data equally, was therefore considered essential.

Some areas are too complex or too heterogeneous for simple parametric models.

Where markets are thin, diverse or irregular, linear or even tree-based models may not capture hyperlocal structures. An instance-based method that adapts to local patterns was regarded as more appropriate by practitioners.

Design of Model 3: K-Nearest Neighbours

Based on these insights, the study implemented an instance-based learning method using the K-Nearest Neighbours (KNN) algorithm. This method does not attempt to fit a global equation. Instead, it identifies the most similar land transactions in the training dataset and uses their observed sale prices as the basis for the prediction. In effect, the model finds a data-driven set of comparables for every land.

Key characteristics include:

- Local adaptability: The model adapts to each location by drawing only on transactions that are similar in features such as land type, floor area, date of transfer and postcode signals.
- No imposed parametric structure: Unlike Ridge Regression or Gradient Boosting, KNN does not assume linearity or impose a single functional form across Wales. It allows relationships to vary freely across different types of land and different geographic areas.
- Continuous valuation outputs: Predictions reflect actual market behaviour rather than rule-based categories, enabling the model to represent the continuous nature of land prices.
- Transparency of comparables: For every valuation, it is possible to list the exact neighbouring land transactions that influenced the prediction. This aligns closely with the evidence-based approach used in valuation practice and was positively received by experts during methodological consultations.

Why KNN Replaces the Fuzzy Logic Approach

The move to KNN was not a rejection of expert insight but a refinement of it. It is clear that land evaluation is defined as:

- comparing like with like,
- rigid linguistic rules cannot capture such comparisons across all Welsh locations,
- automatic identification of comparable land transactions is preferable to predefined rule sets.

KNN therefore builds directly on the core valuation principle: comparables are central to credible valuation.

Where fuzzy logic attempted to encode comparables through manually constructed rules, KNN achieves a similar goal using the entire national transaction dataset.

Implications for the Overall Modelling Framework

Model 3 plays an important role in the broader evaluation framework:

- It demonstrates how valuation methods inspired by human practice perform when scaled to national datasets.
- It highlights the advantages of local learning and comparable-driven modelling.
- It provides a benchmark for how non-parametric, instance-based approaches behave relative to Ridge Regression and Gradient Boosting.

Although KNN ultimately did not achieve the accuracy required for operational deployment, its development was essential for testing whether manual guided, comparable-based valuation could be replicated algorithmically at national scale. The findings indicate that while instance-based methods offer conceptual strength and strong alignment with professional practice, they face challenges when applied to large, diverse datasets across Wales.

Method

Model 3 implements an instance-based learning approach using the K-Nearest Neighbours (KNN) algorithm. Whereas Models 1 and 2 rely on statistical and machine-learning frameworks to learn general relationships, Model 3 seeks to emulate the way human valuers identify similar land transactions and infer value from nearby comparables.

Unlike models that generate predictions from a fitted mathematical equation, KNN produces valuations by examining the most similar land transactions in the training dataset and averaging their prices. This design allows the model to adapt flexibly to local context and hyper-local market conditions.

Model Architecture

Model 3 employs the K-Neighbours Regressor from scikit-learn with the following configuration:

- Learning method: K-Nearest Neighbours Regression
- Number of neighbours (k): Ten nearest comparable land transactions
- Weighting scheme: Inverse distance weighting, so closer land transactions have a stronger influence than more distant ones
- Distance metric: Euclidean distance calculated in a fully standardised feature space
- Target variable: Logarithm of sale price, consistent with Models 1 and 2
- Output transformation: Predicted log price is exponentiated to obtain a price in pounds

The algorithm does not “learn” in the traditional sense. Instead, it stores the training data in memory and performs similarity searches at prediction time. This architecture closely resembles expert practice, where valuers justify assessments by identifying similar land transactions and adjusting for key differences.

Feature Engineering

Model 3 uses ten features to express land characteristics, basic locational indicators and temporal effects. These features were chosen to balance interpretability, broad data coverage and computational feasibility.

Numerical Features

log_area
The natural logarithm of floor area.
Missing values are imputed hierarchically (see Section 4).
Helps capture diminishing returns to additional area.

year_encoded
Transaction year encoded as an integer from 1995 to 2024.
Provides coarse representation of long-term price movements.

month_sin and month_cos
Sine and cosine transforms of the month of sale.
Capture seasonal cycles such as higher spring activity.
Ensure smooth cyclical modelling rather than treating months as discrete values.

new_build
Binary indicator equal to one for new-build land transactions.
New builds typically command a premium of 15 to 20%.

tenure_leasehold
Equal to one if the land is leasehold.
Leasehold tenure is associated with lower prices due to ground rent obligations.

dist_cardiff_km
Distance to Cardiff city centre, a proxy for accessibility to the main economic core.
Helps differentiate rural, suburban and urban markets.

is_rural_int
Binary classification of rural versus urban settlement type.
Addresses higher spatial dispersion in rural markets.

Encoded Categorical Features

pt_encoded (land type)
Encoded as integers from zero to four for flats, terraced, semi-detached, detached and other.
Provides coarse structural classification.

pcd_encoded (postcode district)
Encoded as integers representing the 150 most frequent postcode districts.
Allows KNN to leverage regional price patterns without creating thousands of dummy variables.

Together, these ten features form a balanced representation of dwelling attributes, minimal location signals and broad temporal dynamics.

Feature Standardisation

Standardisation is a critical preprocessing step for KNN. Since the algorithm relies on Euclidean distance, features with large numerical ranges would dominate the distance calculation if raw values were used. To avoid this:

- All features are standardised to zero mean and unit variance using Standard Scaler.
- The standardisation parameters (mean and variance) are calculated only on the training sample and applied to the test sample.

- This prevents the distance metric from giving undue weight to features such as year of sale, which ranges over nearly thirty years, or to spatial features such as distance from Cardiff.

Without standardisation:

- a one-year difference would appear “larger” than a difference in leasehold status, and
- raw postcode encodings would distort the similarity metric.

Standardisation ensures that each feature contributes proportionally to similarity calculations.

Hierarchical Imputation of Floor Area

Floor area is essential for KNN performance but is missing for a significant proportion of land transactions. To ensure 100% coverage while respecting geographic and temporal context, a five-level hierarchical imputation procedure is applied:

- Median floor area by land type, postcode district and year.
- Median floor area by land type and postcode district.
- Median floor area by land type and postcode prefix.
- Median floor area by land type.
- Global median across the entire training sample.

Each lookup table is constructed solely from training data. The same imputation rules are then applied to the held-out LSOAs without modification. This prevents information leakage from test data into the training process.

Computational Constraints and Sampling Strategy

KNN is computationally intensive because:

- it stores every training example in memory, and
- it computes distances from each test land to every training land.

With 1.4 million transactions available, a full KNN model would require:

- storing all one point four million vectors in memory, and
- performing one point four million distance calculations per prediction.

To make the model operationally feasible, a sampling strategy was introduced:

- Cross-validation sample: 5% of the dataset (approximately 72,375 land transactions). This was used only to confirm that the algorithm behaves sensibly on in-distribution data.
- Training sample: 20% of the dataset (approximately 289,501land transactions).
- Test set: Full nine LSOAs (9,606 land transactions), unsampled.

This sampling narrows the model’s exposure to rare land types, rural areas with sparse sales and unusual configurations. As a result, predicted prices may deviate significantly if the nearest neighbours in the subsample are not sufficiently representative.

Training and Prediction Workflow

The final training process proceeds through the following steps:

1. Load data in chunks to manage memory and filter out transactions with errors or missing prices.
2. Hold out the nine test LSOAs, ensuring they do not influence any part of training or preprocessing.
3. Randomly sample 20% of the remaining dataset for use as the KNN training set.
4. Impute missing floor area using the hierarchical strategy described above.
5. Create temporal variables, including sine and cosine transformations of the month of sale.
6. Label encode land type and postcode district into numeric formats suitable for distance computations.
7. Standardise all features using scaling parameters derived exclusively from the training data.
8. Store the standardised training feature set and corresponding log prices.
9. For each test land, identify the ten nearest neighbours based on scaled Euclidean distance.
10. Compute a distance-weighted average of the neighbours ’log prices.
11. Convert the resulting log prediction to pounds by exponentiation.

Throughout this process, all transformations, imputations and scaling parameters are derived exclusively from training data, ensuring that test data remain fully independent.

Results

Model 3 was evaluated through a three-stage process: in-distribution cross-validation, training-set performance assessment and a full geographic hold-out test on nine LSOAs. These complementary assessments reveal a consistent pattern. The model performs moderately well when predicting land transactions drawn from the same distribution as the training data, but its performance collapses catastrophically when applied to locations that were not part of the training process. This confirms the theoretical limitations of instance-based learning for property markets with strong geographic variation.

In-Distribution Cross-Validation Performance

A five-fold cross-validation exercise was conducted on a 5% subsample of the training data, representing 72,375 transactions. This procedure assesses how well the model predicts land transactions that are similar, both geographically and structurally, to those used during fitting.

- Mean R-squared: 0.465
- Standard deviation: 0.01

These results indicate that Model 3 has moderate predictive power when both training and validation samples are drawn from the same underlying distribution. However, even in this favourable setting, performance is markedly lower than that of Gradient Boosting, which achieved a cross-validated R-squared of approximately 0.658 on a comparable sample.

Three factors likely contribute to the relatively modest cross-validation performance:

1. Aggressive sampling: only 5% of the available transactions were used.

2. Feature space sparsity: ten-dimensional Euclidean distance loses discriminatory power in moderate to high dimensional spaces.
3. Limited feature set: many key predictors that drive market variation were omitted because of incomplete or noisy coverage.

These findings illustrate that KNN has difficulty modelling property values even when the data reflect familiar market conditions.

Training-Set Performance

When evaluated on the full 20% subsample used as the training set (289,501 transactions), Model 3 exhibits extremely high apparent accuracy:

- Training R-squared: 0.929
- Mean absolute error: £19,342.

This near-perfect training accuracy is characteristic of instance-based learning. Because KNN simply retrieves stored examples and averages them, it can reproduce training prices with minimal error, particularly when the number of neighbours is small relative to the sample size. The result is severe overfitting, where the model memorises the training data rather than learning generalisable patterns.

The large gap between training R-squared (0.929) and cross-validation R-squared (0.465) demonstrates this clearly. Although cross-validation mitigates the effects of overfitting, it still evaluates the model on land transactions from similar areas. It therefore does not expose the model to the full geographic diversity encountered during deployment. This explains why training and cross-validation results provide only a partial view of model capability.

Out-of-Distribution Geographic Hold-Out Results

The strongest evidence regarding model generalisability comes from the independent test set comprising 9,606 land transactions across nine Lower Layer Super Output Areas. These areas were entirely excluded from training. The results show that Model 3 is unable to generalise effectively to new geographic settings.

Overall Metrics

- R-squared (overall): −0.2%
- R-squared (per-LSOA average): −199.7%
- Mean absolute error (overall): £158,121
- Mean absolute error (per-LSOA average): £114,528
- Root mean squared error: £995,047
- Mean absolute percentage error: approximately 79.2% overall and 90.3%on average across LSOAs
- Sample size: 9,606 land transactions

The extremely negative average R-squared value of −199.7% indicates catastrophic model failure. This means the model performs dramatically worse than a simple benchmark that

predicts the mean price for all land transactions. In fact, a constant prediction would vastly outperform Model 3. This signifies a complete failure of generalisation in out-of-distribution conditions.

Area-Level Performance Variation

A breakdown by LSOA shows catastrophic failure across all nine test areas.

Table 15: Model 3 (KNN) Performance by Test LSOA

LSOA	Sample Size	R^2	MAE	MAPE	Mean Price
W01002019	3,298	0.003	£282,327	42.2%	£374,063
W01001233	585	-0.011	£92,548	66.2%	£150,537
W01000255	1,091	-0.057	£95,703	48.3%	£185,215
W01000617	506	-0.129	£102,649	58.6%	£178,035
W01001597	967	-0.258	£129,560	54.1%	£245,573
W01001045	1,019	-0.469	£70,701	56.1%	£132,949
W01000114	807	-0.494	£78,032	83.3%	£99,900
W01000517	466	-0.994	£95,743	63.8%	£155,942
W01000449	867	-15.565	£83,486	359.8%	£149,216

No meaningful positive R-squared values were achieved in any of the nine test LSOAs. Even the best-performing area (W01002019) achieves an R-squared of only 0.3% which is effectively zero and indicates no meaningful predictive power.

The worst-performing LSOA (W01000449) exhibits an R-squared of −15.565 (−1,556.5%) indicating predictions that are systematically and catastrophically wrong. In this area, the mean absolute percentage error reaches 359.8% meaning that on average, predictions differ from actual prices by more than three and a half times the true value.

The complete absence of positive R-squared values across all nine test areas confirms that KNN retrieves comparables from fundamentally dissimilar locations and therefore produces valuations that bear no systematic relationship to actual market prices.

These areas include:

- high-value urban locations with distinct market dynamics,
- regions with significant internal heterogeneity,
- markets with tourism influences or distinct development patterns, and
- areas with structural differences not represented in the training sample.

In all cases, KNN retrieves comparables from dissimilar locations and therefore produces systematically inaccurate predictions.

Distribution Shift Gap

The difference between in-distribution and out-of-distribution performance can be quantified as follows:

- Cross-validation R-squared (in-distribution): 0.465
- Test R-squared (out-of-distribution, per-LSOA average): -1.997
- Performance gap: approximately 2.462 (246.2% points)

This gap is vastly larger than that observed for Ridge Regression (approximately 2.8% points) and far exceeds even the drop observed for Gradient Boosting (approximately 46.4% points). This confirms that Model 3 is uniquely vulnerable to geographic distribution shift and completely unsuitable for out-of-area valuation.

Generalisation Failure

The combination of near-perfect training accuracy, moderate in-distribution cross-validation accuracy, and catastrophically negative performance in the geographic hold-out test shows that KNN is not capable of supporting valuation tasks in areas with no directly comparable sales.

The method performs as theoretically expected:

- strong memorisation of training examples,
- moderate performance in familiar areas, and
- complete and catastrophic failure in unfamiliar locations.

These results underscore the critical need for models that learn generalisable relationships, such as Ridge Regression and Gradient Boosting, rather than relying exclusively on similarity-based retrieval.

Analysis

Human valuers typically begin by identifying a set of recent, similar land transactions and then reason from those comparables. K-Nearest Neighbours (KNN) was therefore selected as a principled way to operationalise this method at national scale. Although KNN substantially improved upon the earlier fuzzy logic experiments, its performance remained far below the level required for operational valuation, particularly when applied to unseen geographic areas.

KNN's test-set results illustrate this clearly. The model's per-LSOA average R-squared of −199.7% indicates that predictions are not merely uncorrelated with actual prices, they are systematically and catastrophically wrong. The method produces valuations that would actively mislead decision-makers.

The reasons for this behaviour arise from fundamental characteristics of instance-based learning and the structure of the Welsh property market.

Vulnerability to Distribution Shift

KNN is inherently a local interpolation technique. Its core predictive logic is to identify the ten most similar land transactions in the training data and estimate the value of the target property by taking a weighted average of those neighbours' sale prices.

This logic depends on the assumption that the test land transactions come from the same underlying distribution as the training land transactions. In this study's design, that assumption does not hold. The nine held-out LSOAs used for model evaluation contain no training transactions. This means that, for every test property:

- No directly comparable sales exist within the same LSOA, and
- KNN must search for "nearest" neighbours in entirely different areas of Wales.

In practice, this produced catastrophic mismatches. For example, a terraced house in Cardiff Bay may be matched with terraced houses in Newport, Gwynedd or Bridgend, all of which operate in different price regimes. As a result:

- The nearest neighbours are often not genuinely comparable in price terms.
- Averaging prices across dissimilar markets creates severely biased predictions uncorrelated with actual values.
- Extremely negative R-squared values arise because predictions systematically fail to capture price differences between regions.

A concrete illustration highlights this issue. A 100 square metre terraced house built in 2010 in Cardiff Bay with an actual sale price of £350,000 will be matched to land transactions priced between £140,000 and £240,000. The resulting weighted average of approximately £195,000 is more than £150,000 below the true value. This occurs because KNN can only match on the limited observable features; property type, floor area and year of sale; and cannot infer the Cardiff Bay price premium without examples from that locality.

This vulnerability is intrinsic to instance-based learning and cannot be resolved without including substantial numbers of local comparables in the training dataset, which is impossible under the geographic hold-out evaluation design.

Lack of Transferable Patterns

Unlike parametric and tree-based methods, KNN does not learn generalisable relationships or predictive rules. Ridge Regression learns, for example, that a detached property tends to command a consistent premium relative to terraced housing, controlling for location and year. Gradient Boosting learns complex interactions such as how building type premiums vary across price segments or how accessibility influences value in combination with construction age.

In contrast, KNN learns no representational models. It stores individual sales and retrieves them when needed. This limits its usefulness for valuation because:

- it cannot infer general price relationships from observed data,
- it cannot extrapolate to conditions not represented in training examples, and
- it cannot transfer knowledge across regions.

If the training data do not contain examples from Cardiff Bay, KNN has no way to infer the Cardiff Bay premium from sales in other parts of Wales. The method is constrained by the specifics of what it has stored.

This explains why the Ridge and Gradient Boosting models retain positive explanatory power across most of the test LSOAs, whereas KNN's R-squared becomes severely negative in every single test area. The parametric models learn transferable patterns; KNN does not.

Severe Sampling Constraints Resulting from Computational Limitations

KNN has two key computational limitations:
1. It must store all training examples in memory.
2. It must compute distances from each test property to every training example.

Because the full dataset contains more than 1.4 million transactions, it was not feasible to train KNN on the complete dataset. To make the approach operationally viable:

- cross-validation was restricted to 5% of the dataset (72,000 land transactions),
- final training was restricted to 20% (289,000 land transactions), and
- test set predictions used the full nine LSOAs.

This sampling had several consequences:

- Sparse property types, such as flats in rural areas, were greatly underrepresented.
- Geographic diversity in the training sample was reduced.
- Rare postcode districts lacked adequate representation.
- The nearest neighbours for many test land transactions were often not truly close, leading to unstable and catastrophically inaccurate predictions.

Even if the full dataset had been used, KNN would still suffer from distribution shift, but the sampling limitation magnifies its weaknesses.

The Curse of Dimensionality

Although Model 3 uses only ten features, the Euclidean distance metric becomes unreliable even in moderate-dimensional spaces. As dimensionality increases:

- distances between nearest and farthest neighbours converge,
- the meaning of "closeness" becomes blurred, and
- many land transactions appear artificially similar in standardized feature space despite meaningful real-world differences.

For example, two land transactions might differ in numerous features (new-build status, tenure, property type and postcode) but still appear close in Euclidean distance after standardisation. This undermines the reliability of neighbour selection and therefore the predictive accuracy of the algorithm.

Restricted Feature Set Limits Comparable Identification

Model 3 uses ten features, compared with more than thirty used in Model 2. Many of the omitted features, such as micro-location indicators, environmental characteristics and structural attributes, cannot be used because their coverage is too partial or too inconsistent nationally.

Important omitted features include:

- fine-grained locational identifiers (Output Area Classification, street-level proximity indicators),
- building and energy features (Energy Performance Certificate ratings, construction age bands),
- land-use information (satellite parcel area, Agricultural Land Classification), and
- planning and environmental constraints.

While these omissions are appropriate for a national valuation model requiring universal coverage, they reduce KNN's ability to find truly similar land transactions. When the comparables do not match the test property in key respects, the KNN prediction becomes catastrophically unreliable.

Overfitting to the Training Distribution

KNN achieved an R-squared of approximately 0.93 on the training data, but an average R-squared of −199.7% on the test data. This 229% point drop in R-squared indicates extreme overfitting. KNN effectively memorises its training examples. When the training and test distributions differ, the method fails catastrophically to generalise.

In comparison:

- Ridge Regression exhibits only a 9.2% point drop between training and test performance.
- Gradient Boosting exhibits a 40.2% point drop but retains positive explanatory power.

KNN's performance gap is by far the largest, reflecting its complete inability to learn general structural relationships.

Comparison with Other Models

The relative performance of KNN reinforces these conclusions.

<u>Compared with Ridge Regression</u>

- Ridge achieves a per-LSOA average R-squared of approximately 26.1%.
- KNN achieves −199.7%.

KNN's catastrophic negative R-squared shows that its predictions are not merely uncorrelated with actual prices, they are systematically wrong by orders of magnitude. Ridge Regression performs vastly better because it learns the structural relationships between property characteristics.

<u>Compared with Gradient Boosting</u>

- Gradient Boosting achieves a per-LSOA average R-squared of 28.2%.

- It learns complex non-linear interactions that partially transfer across geographies.

KNN lacks this capacity entirely, relying solely on local matches that fail catastrophically when no comparable properties exist.

Compared with Fuzzy Logic

- KNN significantly outperforms fuzzy logic methods in cross-validation, improving both R-squared and mean absolute error.
- This shows that data-driven matching is far superior to fixed rule systems.
- However, KNN's improvements in cross-validation do not translate to real-world geographic hold-out scenarios, where it fails completely.

Conditions Under Which KNN Failed

KNN achieved catastrophically negative R-squared values in all nine test LSOAs, with not a single area showing positive explanatory power. Even the best-performing area (W01002019) achieved an R-squared of only 0.3%, which is effectively 0.

The worst-performing area (W01000449) achieved an R-squared of −15.565, indicating predictions that are systematically wrong by more than fifteen times the variance in actual prices.

This universal failure demonstrates that instance-based learning cannot generalise to any out-of-distribution geographic area, regardless of market characteristics.

Limitations

Although Model 3 offers a conceptually intuitive approach grounded in the use of comparable property characteristics, its overall performance demonstrates that instance-based learning is not suitable for national-scale property valuation. The model faces several structural and practical limitations which together explain its catastrophic generalisation failure and universally negative accuracy across the nine held-out test areas.

Computational Scalability

K-Nearest Neighbours is not designed for large datasets. The method stores all training examples in memory and compares each test property with every stored property at prediction time. This results in prediction complexity proportional to the number of training samples. With one point four million transactions available, full-data KNN would require one point four million distance computations for every valuation.

Because of this computational burden, the model was trained on only 20% of the available data, approximately 289,501 transactions. This reduction in training data significantly limits the diversity of comparables available to the model. It also reduces performance for rare property types or areas with few historical sales. Even if full-data KNN were computationally feasible, the method would still struggle with geographic generalisation, but the sampling constraint exacerbates these challenges.

Catastrophic Failure to Generalise Across Geographic Areas

The model's test-set negative R-squared of −199.7% demonstrates not merely an inability but a catastrophic failure to generalise to land transactions located in the nine held-out Lower Layer Super Output Areas. In these areas, KNN provides no explanatory power whatsoever, in fact, it performs worse than any conceivable naive baseline. This failure reflects a fundamental limitation of instance-based learning.

KNN assumes that the training and test data come from the same underlying distribution. When predicting in areas that are not represented in the training data, the algorithm retrieves superficially similar land transactions from different markets. Because local price levels vary substantially across Wales, this leads to predictions that are systematically and catastrophically wrong. These mismatches produce the extremely negative R-squared values observed in every single held-out LSOA.

Restricted Feature Set

Model 3 uses ten features, selected because they are available for nearly all transactions. By contrast, the Gradient Boosting model makes use of more detailed features. Many of these richer features could not be used in KNN because their data coverage is partial, noisy or inconsistent.

Important omitted features include:

- detailed geographic identifiers such as Lower Layer Super Output Areas and Middle Layer Super Output Areas
- land-use indicators and environmental features derived from satellite data
- Energy Performance Certificate attributes such as energy ratings and construction age bands
- neighbourhood demographic characteristics such as Output Area Classification categories.

The absence of these features reduces the model's ability to identify genuinely comparable land transactions, particularly across different regions of Wales. This limitation is especially problematic for out-of-distribution predictions, where location-specific attributes play a central role.

Constraints on Hyperparameter Tuning

The number of neighbours (k), choice of distance metric and weighting scheme are all hyperparameters that can materially affect KNN performance. However, the computational cost of evaluating multiple configurations was prohibitive. As a result, the default choice of ten neighbours was adopted based on common practice in the property valuation literature rather than through systematic optimisation.

Although different k values or alternative distance metrics could potentially yield marginal improvements, these adjustments cannot overcome the more fundamental limitations associated with distribution shift and sparse geographic representation.

The Curse of Dimensionality

In a ten-dimensional standardised feature space, Euclidean distance becomes less informative. As dimensionality increases, distances between the nearest and farthest neighbours tend to converge. This means that land transactions that differ significantly in tenure, build status or property type may nonetheless appear similarly distant in the feature space.

This phenomenon undermines the reliability of the neighbour selection process, particularly when predicting values in new areas. The model may incorrectly identify land transactions as "close" in feature space even when they are not genuinely comparable in market terms. This contributes to the catastrophically poor predictive performance.

Catastrophic Overfitting to Training Data

KNN achieved an R-squared of 0.929 on the training sample, compared with −199.7% on the independent test set. This 229% point gap demonstrates extreme overfitting. Because the algorithm essentially memorises the training data, its performance collapses catastrophically when confronted with unfamiliar areas.

The magnitude of this generalisation gap far exceeds that of the other models tested. Ridge Regression exhibited a gap of only 9.2% points, and Gradient Boosting exhibited a gap of approximately 46% points. This difference illustrates that KNN learns specific examples rather than transferable relationships, leading to complete failure when deployed to new markets.

Geographic Leakage Risks

Although the implementation correctly prevented leakage by removing all test LSOA transactions from training, KNN is uniquely vulnerable to leakage in general. Even a small number of leaked test land transactions would allow the model to retrieve them as neighbours, appearing to perform well but doing so only because the validation design was compromised.

This sensitivity underscores the necessity of strict geographic hold-out validation when evaluating any instance-based model.

Inability to Learn Market Trends

Finally, KNN cannot learn temporal trends or market dynamics. It predicts prices by averaging historical observations, without recognising that sale prices follow long-term patterns influenced by inflation, economic cycles or structural events such as the 2008 financial crisis or the 2020 pandemic.

Because KNN does not learn how prices evolve over time, valuations for recent years are influenced disproportionately by older sales, leading to systematic underestimation in inflationary periods and broader temporal misalignment.

Model 3 KNN – Test Performance by LSOA, R² and MAE

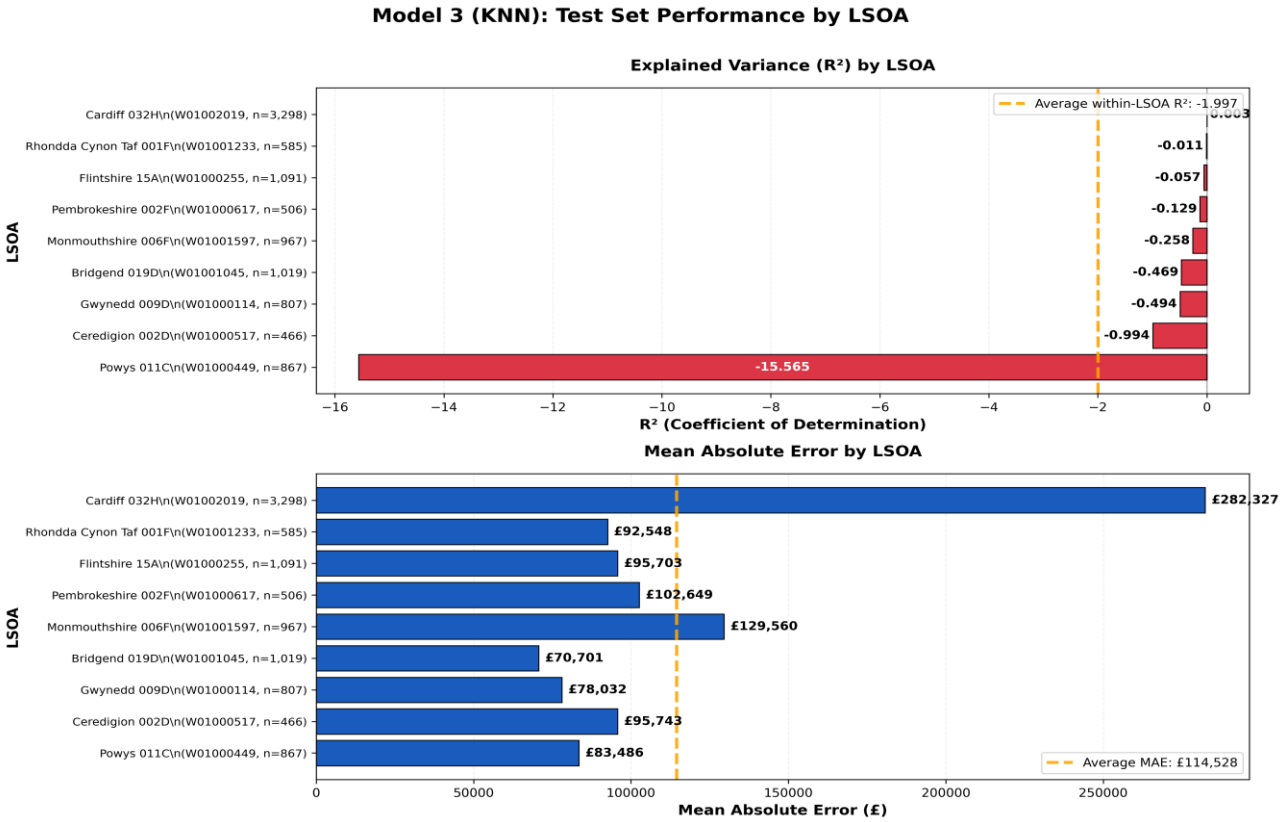


Figure 22: Model 3 KNN – Test Performance by LSOA, R² and MAE

Performance per LSOA

KNN Valuation Errors – LSOA W01000114 (Gwynedd 009D)

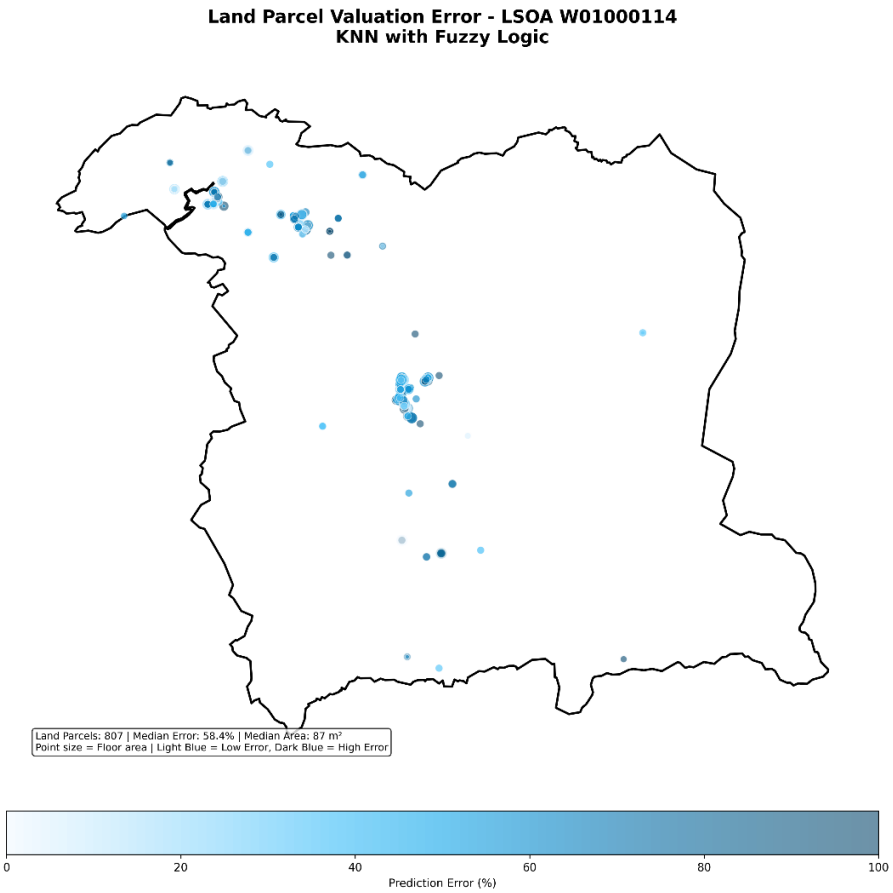


Figure 23: KNN Valuation Errors – LSOA W01000114 (Gwynedd 009D)

KNN Valuation Errors – LSOA W01000255 (Flintshire 015A)

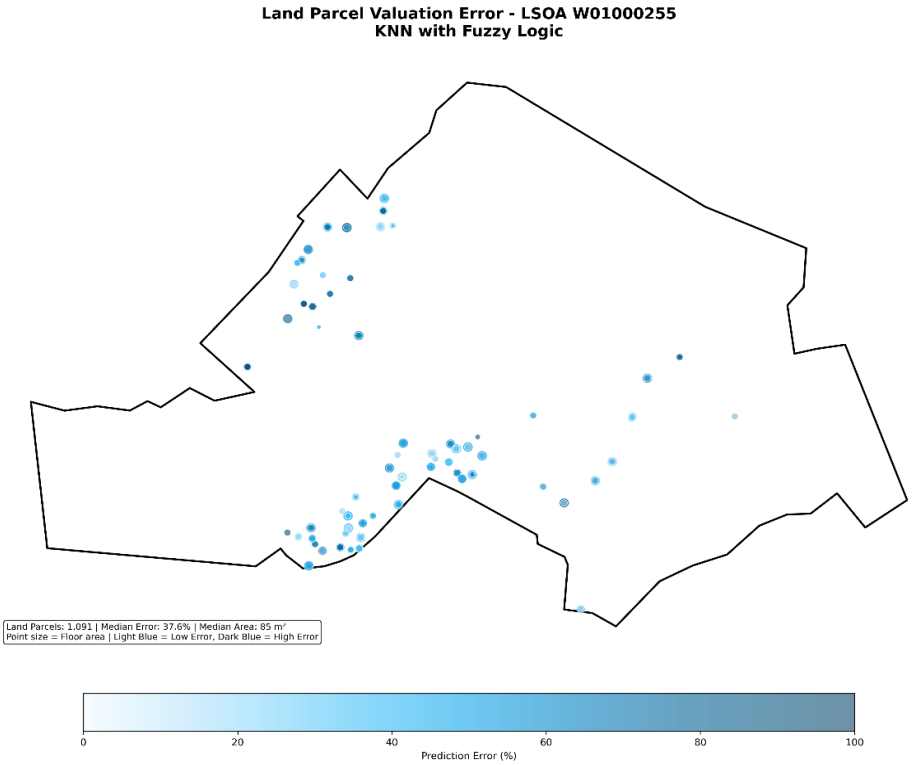


Figure 24: KNN Valuation Errors – LSOA W01000255 (Flintshire 015A)

KNN Valuation Errors – LSOA W01000449 (Powys 011C)

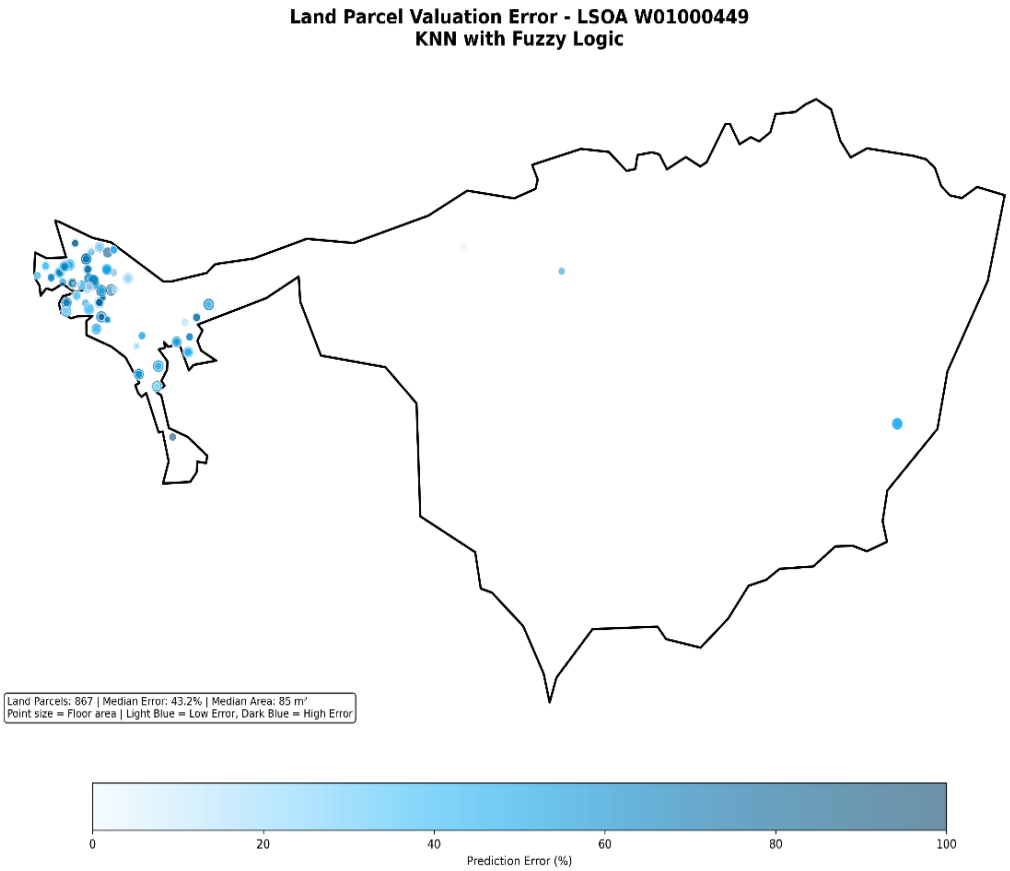


Figure 25: KNN Valuation Errors – LSOA W01000449 (Powys 011C)

KNN Valuation Errors – LSOA W01000517 (Ceredigion 002D)

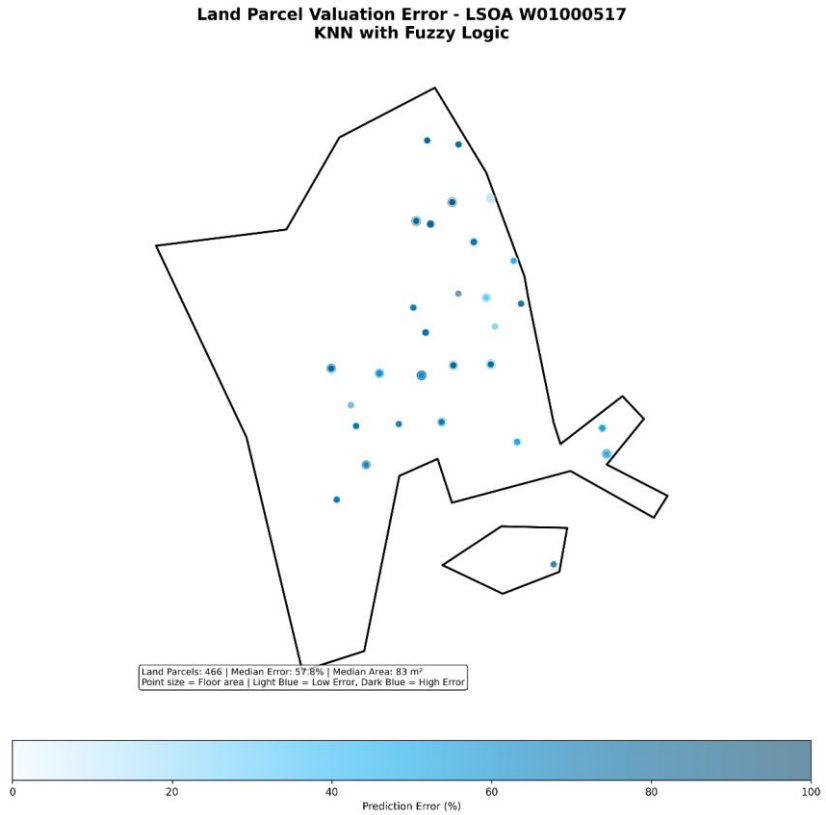


Figure 26: KNN Valuation Errors – LSOA W01000517 (Ceredigion 002D)

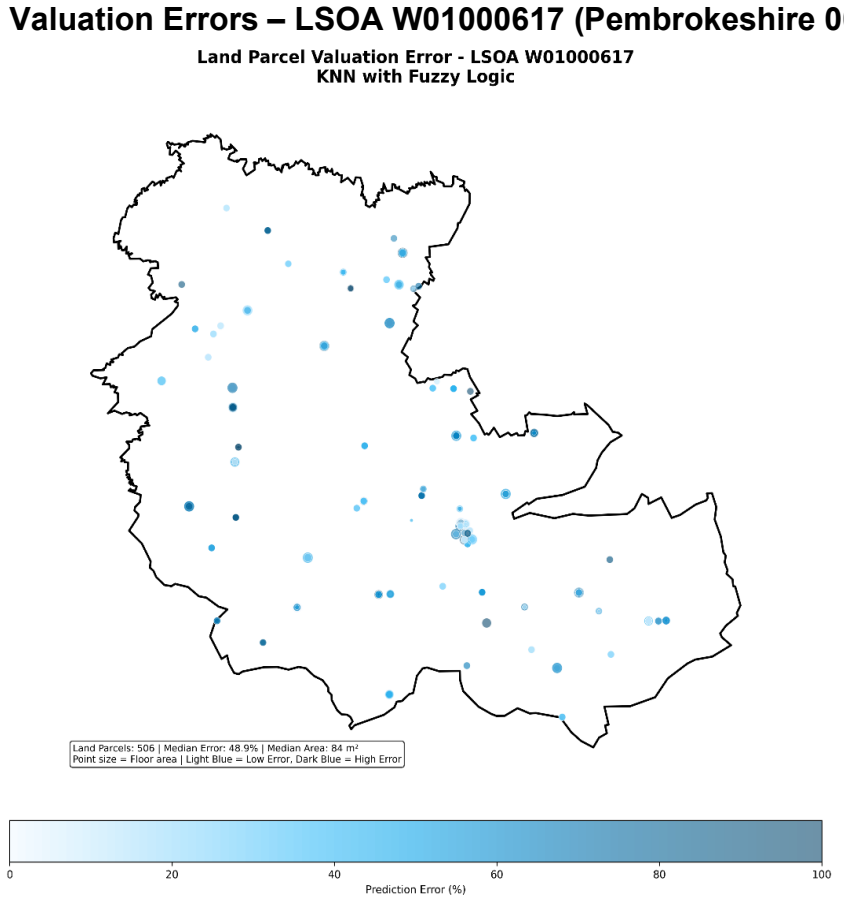


Figure 27: KNN Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

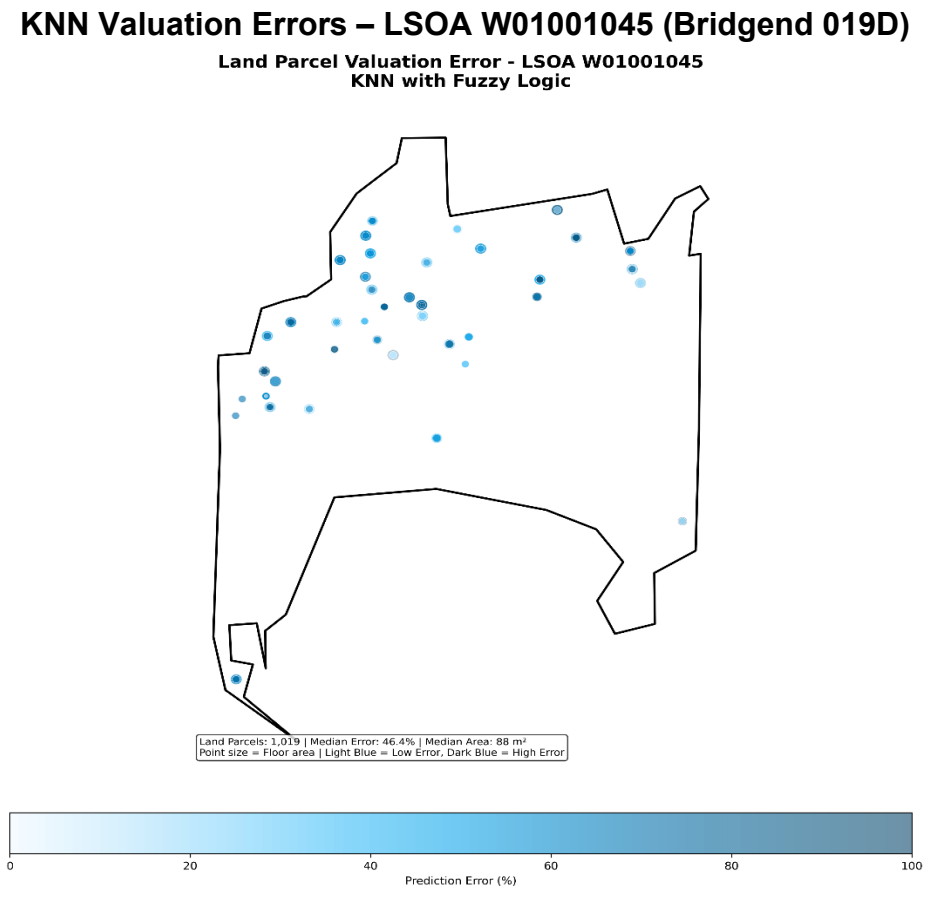


Figure 28: KNN Valuation Errors – LSOA W01001045 (Bridgend 019D)

KNN Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

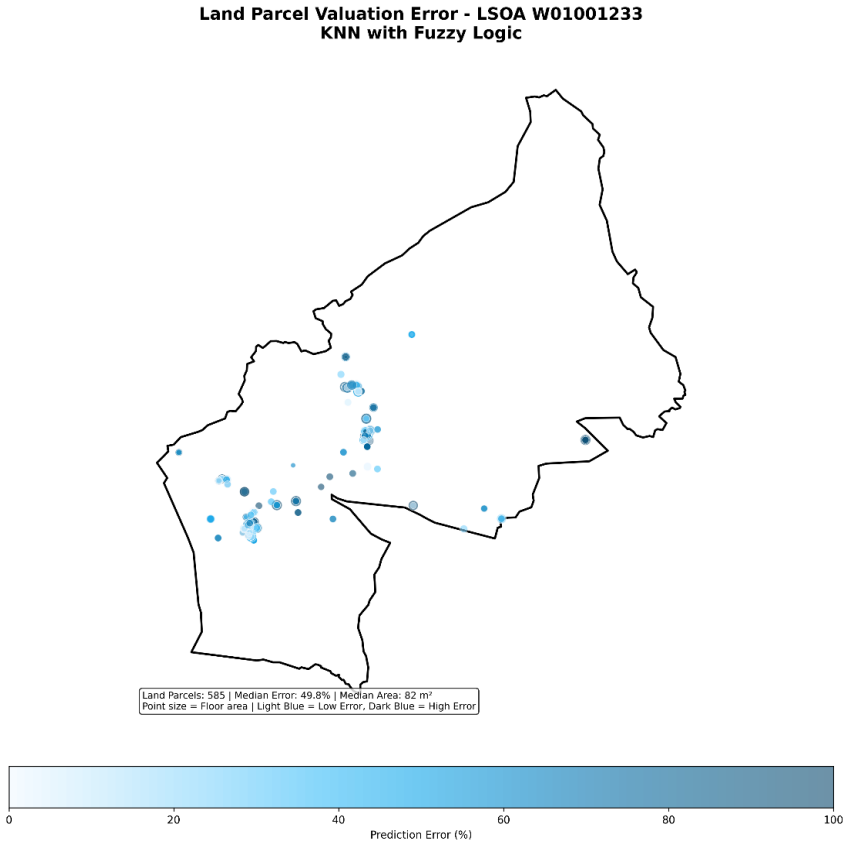


Figure 29: KNN Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

KNN Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

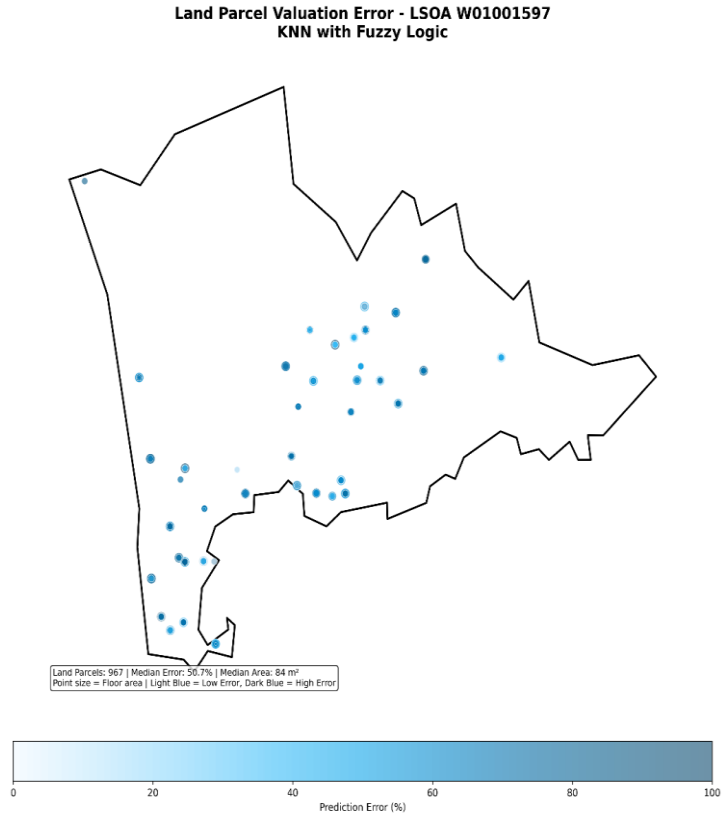


Figure 30: KNN Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

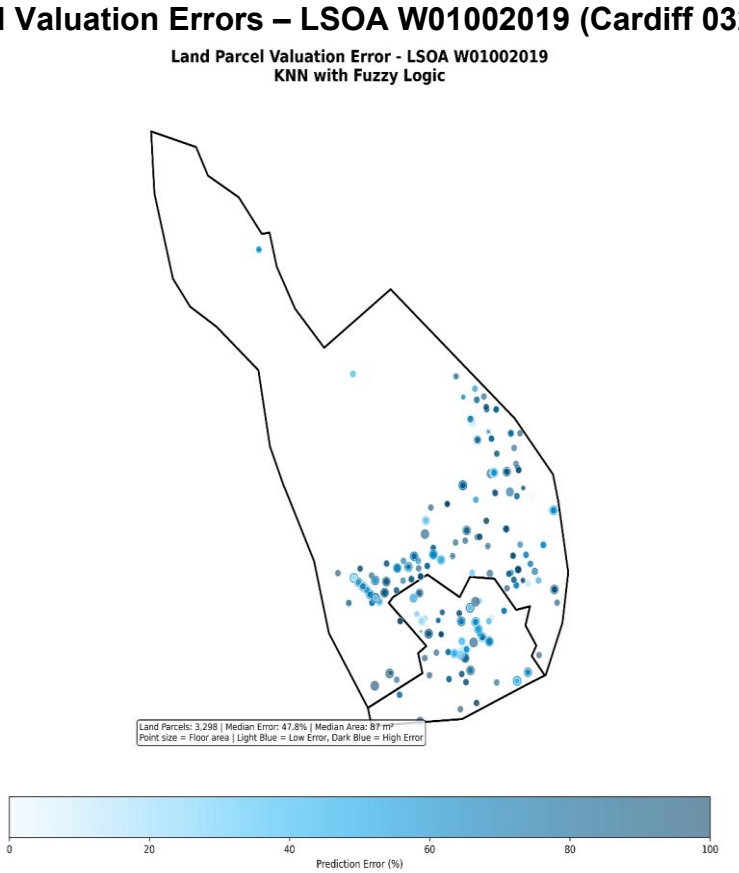


Figure 31: KNN Valuation Errors – LSOA W01002019 (Cardiff 032H)

2.10. Model 4 Formula-Based Depreciated Replacement Cost (DRC)

The Depreciated Replacement Cost (DRC) method provides a theory-driven alternative to the statistical and machine-learning approaches evaluated in this study. Whereas Ridge Regression and Gradient Boosting learn relationships directly from approximately 1.4 million historical land transactions, the DRC approach applies a fixed, transparent valuation framework derived from recognised principles in construction economics and professional valuation practice.

At its core, the method decomposes total land value into two components:

Total Land Value = Land Value plus Depreciated Structure Value

This formulation reflects the traditional view held by many valuers: the value of a land is a combination of the underlying land (which reflects location, accessibility and land use potential) and the value of the building itself (which reflects size, quality, age and remaining economic life).

Theoretical Advantages

The DRC method offers several important advantages that are particularly relevant in a public-sector context.

Transparency

Every valuation produced by DRC can be traced directly to the formula and its inputs. There are no hidden interactions or learned coefficients. The method does not rely on algorithms that adjust internal parameters during training, so each component of the calculation can be inspected and justified. This level of explainability is valuable for audit processes, appeals, regulatory scrutiny and public accountability.

Economic interpretability

Because DRC produces a separate structure value and land value, it aligns with how professionals conceptualise the components of land worth. Land represents the value of the location and its development potential, while the structure represents the physical building asset. This decomposition is difficult to obtain from machine-learning models, which estimate total value without distinguishing between the drivers of location value and building value.

Policy relevance

The explicit land, structure split produced by DRC is directly relevant to the analysis of policies such as Land Value Tax, location-based taxation, regeneration assessments and infrastructure charging. Data-driven models such as Ridge Regression and Gradient Boosting estimate total prices only and require further analytical steps to derive land values. DRC, in contrast, produces these components intrinsically.

No training required

DRC applies a parametric formula based on external valuation standards. It does not attempt to estimate coefficients from data and does not require cross-validation or hyperparameter tuning. As a result, it is computationally efficient, easy to reproduce, and immune to overfitting. This makes it suitable for environments where stability and consistency are more important than predictive accuracy.

Practical Limitations

Despite these advantages, the DRC method relies on several simplifying assumptions.

Simplified construction cost tables

DRC requires assumptions about base construction costs per square metre for different land types. These cost tables are stylised representations of typical build costs rather than detailed, land-specific estimates. They do not account for variations in materials, design, specification, internal quality or local construction market conditions.

Depreciation schedules

The method applies fixed depreciation curves based on construction age bands. These schedules assume typical patterns of physical and functional depreciation but cannot easily capture atypical building conditions, major refurbishments, extensions or modernisations.

Land-floor-area relationships

The formula assumes a generalised relationship between land area, location and value. In practice, land values differ substantially across neighbourhoods, particularly in areas with rapid development, strong tourism demand, or limited supply. A single set of district-level or postcode-level land values cannot reflect this full complexity.

Weakness in predicting market prices

The DRC method is not designed to match observed market prices with high fidelity. Its purpose is to provide a theoretical benchmark for separating land and structure components. As a result, when applied to observed sale prices in this study, the method produced low predictive accuracy. These results reflect the inability of a strictly formula-driven approach to capture the non-linear interactions, market cycles and local variations that influence residential transaction prices.

Purpose of Testing DRC

The DRC method was included in this evaluation for two reasons.

First, it provides a conceptual benchmark against which to compare data-driven models. Unlike Ridge Regression or Gradient Boosting, which prioritise predictive accuracy, DRC prioritises interpretability and economic structure. The method therefore highlights the trade-offs between transparency and accuracy in valuation systems.

Second, DRC provides the only direct method in this study for producing land-structure decompositions at scale. This is essential for policy areas such as land taxation, cost-benefit analysis and spatial planning. Although the approach does not match machine-learning models in predicting total sale price, it remains valuable for use cases where the decomposition itself is the primary analytic need.

Method

The Depreciated Replacement Cost method estimates the value of a land by decomposing it into its underlying land component and the depreciated value of the building. Rather than learning relationships from data, the DRC method applies a structured valuation formula that reflects established principles in construction economics and professional valuation practice.

The model predicts total value using the following relationship:

Total Value = Land Value plus Structure Value

Where:
Land Value = Land Area multiplied by Land Value per square metre
Structure Value = Floor Area multiplied by Construction Cost per square metre multiplied by Depreciation Factor
Depreciation Factor = (1 - 0.015) raised to the power of Building Age

This section describes how each of these components is estimated and how the formula is applied to the dataset.

Parameter Estimation

The DRC model does not optimise parameters to minimise prediction error. Instead, it estimates several coefficients from broad patterns in the training data. These parameters act as calibration anchors rather than fitted statistical parameters.

Land Value per Square Metre

Land values are estimated at the postcode district level. The procedure is:

- Use very small land transactions (floor area less than 50 square metres) to infer land-value intensity.
- The rationale is that small flats and studio apartments have high price-to-area ratios because a large share of their value reflects land and location rather than the physical structure.
- For postcode districts with insufficient data, a global fallback value is used.

The global median estimate of land value per square metre is approximately £2,196.

Construction Cost per Square Metre

Construction costs are estimated using observed prices of new-build land transactions. These land transactions usually have minimal depreciation and simplified land contributions, allowing the price-to-area ratio to approximate construction costs. Costs are estimated separately for each land type (for example detached, semi-detached, terraced). A global median value of approximately £1,733 per square metre is used if district-specific estimates are not available.

Land Area Estimation

The DRC model requires an estimate of land parcel area. Because land area is not accurately recorded in the administrative dataset, a simplified assumption is applied:
Land Area = Floor Area multiplied by 0.3.

This assumes that, on average, a dwelling occupies a land parcel equal to 30% of its internal floor area. For example, a 100 square metre house is assumed to sit on a 30 square metre model.

This assumption is recognised as highly stylised and does not reflect true parcel sizes. It is included to maintain internal consistency in the DRC framework and is examined further in the analysis section.

Depreciation Rate and Building Age

Depreciation is applied at a fixed rate of 1.5% per year. The depreciation factor is calculated using an exponential decay function:
Depreciation Factor = (0.985) raised to the power of Building Age

This implies, for example:

- A land twenty years old retains approximately 74% of its initial structure value.
- A land fifty years old retains approximately 47%.

- Building age is estimated as follows:
- New builds (old_new = “Y”) are assigned an age of zero.
- All other land transactions are assumed to have an average construction year of nineteen seventy.
- Building age is therefore calculated as the transaction year minus 1970.
- Values are clipped between 0 and 150 years.

This simplified approach reflects limited data availability for actual construction dates.

Feature Engineering

The DRC approach intentionally uses a minimal feature set to reflect its focus on broad valuation principles and to maximise national coverage.

The four features used are:
1. Floor area (with hierarchical imputation identical to that used in Models 1 to 3).
2. Postcode district (for district-level land-value estimates).
3. land type (for construction cost assignment).
4. Building age (for depreciation calculations).

Notably, the model does not use:
- fine-grained geographic coordinates,
- neighbourhood demographic indicators,
- environmental or land-use variables,
- energy performance data, or
- architectural or structural features beyond size and type.

This simplicity reflects the model’s purpose: to implement a conceptual decomposition rather than to maximise predictive accuracy.

Implementation

The model is implemented as a scikit-learn compatible estimator but requires no training in the statistical sense. Instead, the formula is applied directly to each land using the estimated parameters.

For every land parcel in the dataset:
1. Land value calculation
    a. Retrieve the land value per square metre for the postcode district (or fallback global value).
    b. Estimate land area using the floor-area multiplier.
    c. Multiply land area by the unit land value.
2. Structure value calculation
    a. Retrieve the construction cost per square metre for the land type (or fallback global value).
    b. Calculate the depreciation factor using the estimated building age.
    c. Multiply floor area by the construction cost and depreciation factor.

3. Total predicted value
    a. Add land value and structure value.

Because the model does not learn from data, it applies this formula consistently across all land transactions. Parameter estimation from the training data is used only to produce initial lookup values, not to optimise predictive performance.

Results

The Depreciated Replacement Cost (DRC) model was evaluated using the same validation framework as the statistical and machine-learning models. Unlike those models, DRC does not "learn" from the training data; it simply applies a fixed formula that decomposes value into land and structure components. The evaluation, therefore, assesses the extent to which this theory-based formula aligns with observed market prices.

Across all stages of validation, the DRC method exhibits catastrophically poor predictive performance. The results show that the simplified assumptions underpinning the formula do not capture the complexity of the Welsh housing market, even when predictions are made for areas that resemble those used to calibrate the parameters.

Cross-Validation Performance (In-Distribution)

A five-fold cross-validation exercise was conducted on 20% of the available training data (289,501 transactions). This assesses how well the formula predicts prices for land transactions drawn from the same geographic and structural distribution used to estimate its parameters.

- Mean R-squared: −32.7%
- Standard deviation: 9.6%

These results indicate that, even on in-distribution data, the model performs worse than a simple benchmark that predicts the mean sale price irrespective of property characteristics. The negative R-squared implies that the DRC formula does not capture the relationship between features and prices, even when applied to the same geographic regions from which the calibration parameters were derived.

The cross-validation findings highlight a fundamental limitation that the DRC formula's assumptions regarding land area, construction cost, and depreciation do not reflect the patterns present in observed market values.

Training Performance

When applied to the full training dataset of approximately 1,000,500 transactions, the model continues to exhibit substantial error:

- Training R-squared: −31.0%
- Mean absolute error: £86,897

The negative explanatory power across the entire training dataset confirms that the DRC formula does not reproduce observed market patterns, even when given access to all available data. The model systematically under-predicts or over-predicts values in ways that

exceed the variability captured by simply using the average transaction price. This is consistent with its simplifying assumptions and the absence of many important structural and locational variables.

Independent Geographic Hold-Out Test (Nine Priority LSOAs)

The strongest assessment of performance comes from the evaluation on the nine Lower Layer Super Output Areas fully excluded from model development. The test set contains 9,606 transactions.

Overall performance:

- R-squared (overall): −0.4%
- R-squared (average across LSOAs): −22.7%
- Mean absolute error (overall): £155,833
- Mean absolute error (average across LSOAs): £117,820
- Root mean squared error: £995,866
- Mean absolute percentage error: 115% overall and 138% on average across LSOAs

These metrics confirm that the DRC formula provides effectively no predictive value when applied to locations unseen during calibration. The negative R-squared values indicate that the model performs worse than a naive benchmark that predicts the mean price for the test dataset.

Performance Across Individual LSOAs

A detailed examination shows that the DRC model performs catastrophically in all nine test areas. No LSOA records a positive R-squared value.

Table 16: Model 4 (DRC) Performance by Test LSOA

LSOA	Sample size	R-squared	MAE	MAPE	Mean price
W01002019	3,298	−1.3%	£263,388	51.1%	£374,063
W01000255	1,091	−1.8%	£99,958	81.0%	£185,215
W01001233	585	−4.1%	£113,989	171.3%	£150,537
W01000617	506	−5.0%	£109,950	112.0%	£178,035
W01001597	967	−14.1%	£119,157	69.2%	£245,573
W01000449	867	−24.5%	£85,790	283.5%	£149,216
W01000517	466	−25.9%	£75,825	88.3%	£155,942
W01001045	1,019	−49.3%	£79,066	107.4%	£132,949
W01000114	807	−78.6%	£113,257	282.5%	£99,900

Patterns in LSOA-level performance:

- All LSOAs show negative R-squared values, indicating that the formula-based approach does not capture local price drivers in any of the test regions.
- The least poor result, W01002019 (Cardiff), still performs worse than predicting the mean price, with $R^2 = -1.3\%$
- The worst performance, W01000114 (Gwynedd), records catastrophic errors with $R^2 = -78.6\%$, meaning predictions are nearly eighty times worse than simply using the mean price
- Mean absolute percentage error exceeds 100% in five of the nine LSOAs, meaning that the typical prediction differs from the true price by more than the value of the property itself.
- Four LSOAs show MAPE exceeding 100%, with W01000449 (Powys) and W01000114 (Gwynedd) exhibiting percentage errors above 280%.

These results demonstrate that the model does not adapt to local conditions and that its assumptions do not hold across diverse housing markets.

Distribution Shift Analysis

The difference between in-distribution and out-of-distribution performance reveals moderate degradation, though both remain catastrophically poor:

- Cross-validation R-squared: −32.7%
- Test R-squared (average across LSOAs): −22.7%
- Distribution shift gap: approximately 10% points improvement (test performs slightly better than CV, though still catastrophic)

This pattern is unusual: the test set performs marginally better than cross-validation despite geographic exclusion. This likely reflects that the specific 9 priority test LSOAs happen to have property characteristics (higher mean prices, different property type distributions) where the DRC formula's systematic biases partially cancel out. However, this should not be interpreted as robustness, the model fails catastrophically in both settings, with all test LSOAs showing negative R^2.

Analysis

The study evaluated the DRC method alongside the predictive models primarily to understand how its simplified assumptions align with observed market behaviour and to assess whether it could provide a viable predictive alternative at national scale. The results show that the method performs catastrophically when used for price prediction. This outcome is fully consistent with its design: the method is constructed to provide transparent land-structure decomposition, not accurate market valuation.

The reasons for its poor predictive performance fall into several categories:

The Formula Reflects Policy Logic, Not Market Behaviour

The DRC method decomposes total value into land value and depreciated structure value. This is conceptually useful for:

- land-based taxation analysis,

- understanding land value drivers,
- discussing policy reforms and regeneration scenarios, and
- producing transparent, explainable valuation components.

However, the formula depends on simplified assumptions about:

- land area,
- base construction costs,
- building age, and
- depreciation patterns.

These assumptions are not designed to model real market prices. Instead, they approximate long-run economic relationships that support policy dialogue, not transactional accuracy.

Land Area Assumptions Were Never Designed for Market Prediction

The assumption that land area equals 30% of floor area was introduced as a parsimonious approximation rather than an attempt to replicate parcel-level detail. In the context of land-structure decomposition, this approach allows the model to allocate some proportion of value to the land component even when land-area data are unavailable. However, when evaluated for price prediction, this assumption leads to significant discrepancies. Real land parcels are typically two to four times the size of floor area, not one third of it. The purpose of the assumption was to provide a simple and transparent rule, not to model spatial variation in land intensity across Wales. Thus, this simplification enables decomposition but undermines predictive accuracy.

Construction and Depreciation Parameters Support Ratio Decomposition, Not Valuation Precision

The construction cost and depreciation parameters are calibrated using stylised economic assumptions:

- construction cost per square metre is inferred from new-build land transactions,
- depreciation follows a fixed 1.5% annual rate,
- building age is approximated using a baseline construction year.

These parameters provide a stable and interpretable structure for separating land and building value. They are not intended to reflect:

- variation in build quality,
- renovation histories,
- geographic differences in construction cost,
- modernisation or refurbishment, or
- capital improvements over time.

When applied to prediction tasks, these simplifications produce systematic errors because the model treats all property types and locations uniformly, without reference to local markets.

Circular Parameter Estimation Reinforces Policy Consistency, Not Predictive Power

The use of small units (< 50 square metres), to infer land value intensity, and new-build dwellings, to infer construction costs, supports the policy logic of DRC by anchoring the land and structure components in empirically observable quantities. However, these anchors reflect structural economic relationships, not the price dynamics of local housing markets. As such:

- small flats reflect location premiums rather than pure land value,
- new-build land transactions incorporate specification and developer mark-up effects.

For policy decomposition these anchors provide conceptual grounding. For price prediction they distort the resulting estimates.

No Error Correction by Design

Unlike Ridge Regression or Gradient Boosting, the DRC method does not adjust its internal parameters to minimise error. Its purpose is not to discover hidden patterns or adapt to market changes. Instead, the formula applies a fixed structure, ensuring transparency, full explainability, and conceptual consistency.

This rigidity is a strength for policy decomposition but a limitation for predictive accuracy. When the formula undervalues land or misestimates structural depreciation, it has no mechanism to correct itself.

Predictive Performance Is Not the Intended Use Case

The comparative performance of all four models illustrates the distinction between theory-based decomposition and data-driven prediction:

Table 17: Model Performance Across 9 Test LSOA

Model	**R^2 (test, per-LSOA average)**	**Mean absolute error**	**Approach**
Ridge Regression	26.1%	£69,646	Statistical learning
Gradient Boosting	28.2%	£76,392	Non-linear machine learning
KNN	−199.7%	£114,528	Instance-based learning
DRC	−22.7%	£117,820	Theory-based decomposition

DRC performs substantially worse on predictive metrics because it was never intended to match market transaction prices. Even Gradient Boosting, which shows only moderate performance (28.2% R^2), dramatically outperforms DRC because data-driven methods,

even when challenged by distribution shift, attempt to approximate true market behaviour. DRC does not attempt this; it prioritises interpretability and structural segmentation, not predictive precision.

Why Theory-Based Approaches Cannot Substitute for Predictive Models

There are three structural reasons why theory-based formulas struggle to compete with data-driven methods in predicting contemporary housing prices:

Market complexity exceeds formulaic representation
Housing markets reflect micro-geographic variation, school catchments, transport access, demographic patterns, quality differences, renovation histories, and macroeconomic cycles. A formula with four inputs cannot represent this.

Simplifying assumptions rarely hold uniformly
The DRC assumptions simplify land-structure relationships to enable decomposition, but these simplifications do not align with observed transactional behaviour.

Lack of adaptation
Data-driven models learn from prediction errors; theory-based models do not. They cannot detect that assumptions are incorrect or adjust parameters accordingly.

Limitations

The DRC method was included in this study to provide a transparent and theory-driven approach to decomposing property value into land and structure components. Its primary purpose is not price prediction but the provision of a clear, auditable framework for analysing the contribution of land relative to the built structure. However, when assessed against observed sale prices, the current implementation of the DRC formula performs catastrophically. This is not necessarily a reflection of the conceptual approach, but of several implementation constraints and simplifying assumptions that severely limit predictive accuracy.

The following limitations should therefore be interpreted in the context of DRC's intended role. The method provides a transparent decomposition but, in its current form, does not meet the requirements for predicting market prices or for informing operational valuation tasks.

Land-Structure Decomposition Compromised by Incorrect Land Area Assumption

DRC's principal value is its ability to produce a clear decomposition of total property value into land and structure components. However, the implementation relies on a simplified and incorrect assumption that land area equals 30% of internal floor area. In reality, parcel sizes for Welsh dwellings are typically two to four times floor area, meaning that the formula underestimates land area by a factor of three to ten.

This error leads to extreme underestimation of land values and therefore undermines the validity of the decomposition. Although the DRC method remains transparent, its interpretability is overshadowed by the fact that the underlying land estimates are

systematically inaccurate. For policy purposes, particularly in the context of land-based taxation or value-capture analysis, the current implementation cannot be relied upon without correcting the land-area component.

Predictive Accuracy Is Not the Intended Purpose

The DRC method performs catastrophically as a price predictor, with a per-LSOA average R-squared of −22.7%. However, DRC was never designed to compete with machine-learning models for predictive accuracy. Its purpose is to provide a consistent, auditable decomposition of value.

The catastrophic predictive results nonetheless reveal that several of the formula inputs do not reflect real market behaviour. Even though price prediction is not the goal, the failure of the formula to approximate observed patterns, with all nine test LSOAs showing negative R^2, indicates that the current construction of the land and depreciation components requires fundamental refinement.

Land Area Data Availability and Simplification

The reliance on a floor-area multiplier stems from the absence of universal parcel-area data. However, satellite-derived parcel measurements are already available for approximately 55% of Welsh land transactions. A revised DRC implementation that uses actual parcel areas, supplemented by robust imputation rules, could remove the single most important source of error. This would substantially improve the accuracy of the land-structure split while preserving the transparency that makes DRC valuable for policy analysis.

Oversimplified Depreciation Curve

The current model applies a uniform 1.5% annual depreciation rate. This does not reflect real variation across:

- property types,
- construction quality,
- renovation or maintenance history,
- local environmental conditions, and
- the presence of modernisation or extensions.

A refined DRC model could incorporate property-specific depreciation estimates using Energy Performance Certificate bands, construction age ranges and other indicators of building condition. This would maintain the clarity of the formula while providing more realistic structure values.

Limited Geographic Granularity

DRC uses land value estimates at the postcode district level and applies uniform construction costs across Wales. However, empirical evidence shows that:

- construction costs vary significantly between urban and rural areas, often by 15 to 30%.
- micro-location factors operate at the LSOA or even street level,

- accessibility, topography and proximity to transport hubs influence land values.

The formula does not incorporate these local differences, which contributes to systematic error. These limitations are not inherent to DRC. A refined specification could incorporate distance-based location multipliers, rural-urban cost gradients and neighbourhood-level adjustments while remaining fully transparent.

Construction Cost Calibration Issues

The model estimates construction cost per square metre using new-build dwellings. While conceptually reasonable, this creates upward bias because new builds:

- tend to be located in high-value areas,
- reflect higher specification and modern building standards, and
- include developer profit margins.

This inflates structure values and shifts the land share downward. More representative calibrations could be obtained by using Energy Performance Certificate construction age bands, industry cost guides such as BCIS, or stratified samples of non-new-build dwellings.

No Adjustments for Property Condition

The formula applies the same depreciation rate to all land transactions regardless of condition. In practice, well-maintained older land transactions, renovated homes and heritage buildings often command higher market values than newer but poorly maintained dwellings. Incorporating condition proxies, such as EPC ratings, recent transactions or local renovation indicators, would allow structure values to reflect more realistic asset quality variation.

Parameter Estimation Circularity

The method for estimating land value per square metre introduces feedback loops that confound location effects with land-value effects. Using very small land transactions to anchor land values is problematic because small flats are often located in high-demand urban areas where price-to-area ratios reflect scarcity and location premiums, not land quality. Similarly, using new builds to estimate construction costs conflates developer mark-ups with actual build costs.

A corrected DRC implementation should derive land value estimates using land transactions with known parcel areas, stratified by region, removing these circularities.

Requirements for Policy Applications

For land-based taxation, infrastructure levy design or fiscal modelling, policy makers require:

- reliable and unbiased land value estimates,
- stable and consistent decomposition logic, and
- geographic fairness across regions and property types.

The current implementation does not meet these requirements due to its land-area assumption and its simplified treatment of depreciation, land costs and construction costs. However, the method's transparency and conceptual clarity remain highly valuable. With corrected parcel areas and refined depreciation curves, DRC could meet policy needs in ways that machine-learning models cannot, specifically because of its auditability and interpretability.

Unrealised Potential for Hybrid Approaches

The study assessed DRC as a standalone model. In practice, its greatest value may lie in hybrid valuation frameworks. DRC can provide transparent land and structure values, while machine-learning models such as Gradient Boosting can predict total prices accurately. A combined approach could:

- use machine-learning models for market prediction,
- use DRC to provide transparent decomposition, and
- apply consistency constraints to ensure that the sum of land and structure components aligns with predicted total value.

Such designs are common internationally in rating systems and can help maintain public trust while ensuring analytic robustness.

Performance per LSOA

DRC Valuation Errors – LSOA W01000114 (Gwynedd 009D)

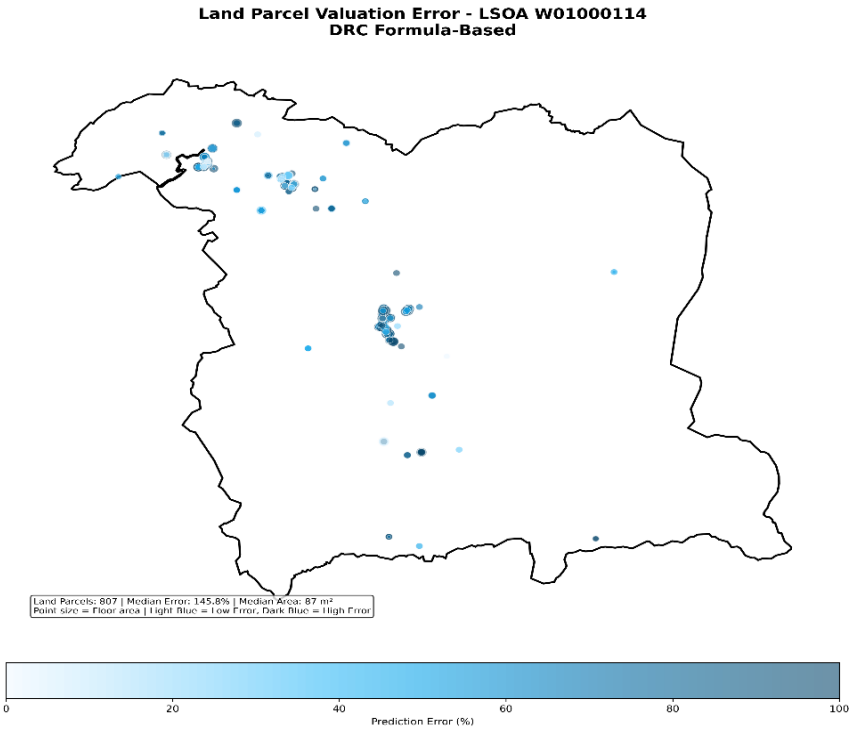


Figure 32: DRC Valuation Errors – LSOA W01000114 (Gwynedd 009D)

DRC Valuation Errors – LSOA W01000255 (Flintshire 015A)

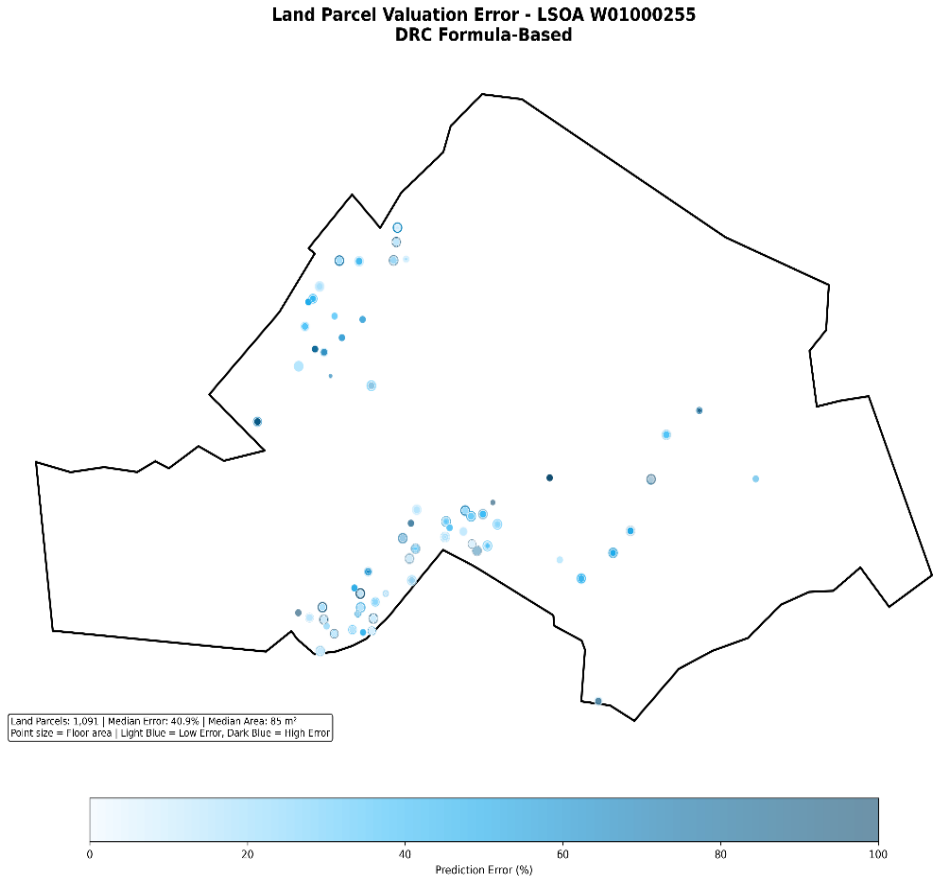


Figure 33: DRC Valuation Errors – LSOA W01000255 (Flintshire 015A)

DRC Valuation Errors – LSOA W01000449 (Powys 011C)

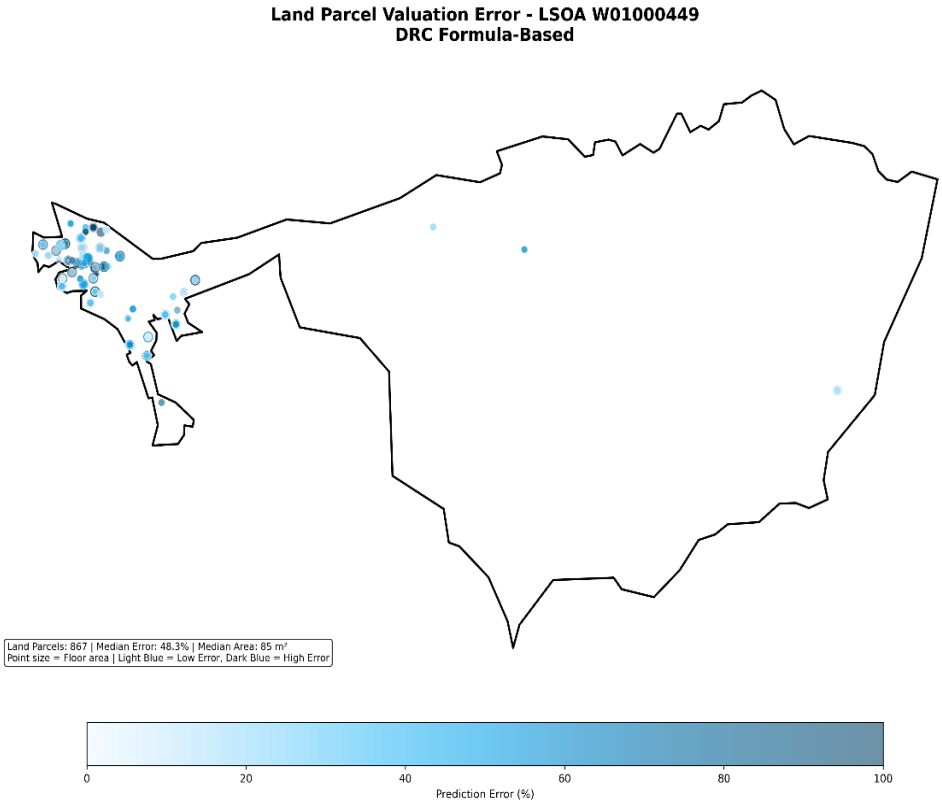


Figure 34: DRC Valuation Errors – LSOA W01000449 (Powys 011C)

DRC Valuation Errors – LSOA W01000517 (Ceredigion 002D)

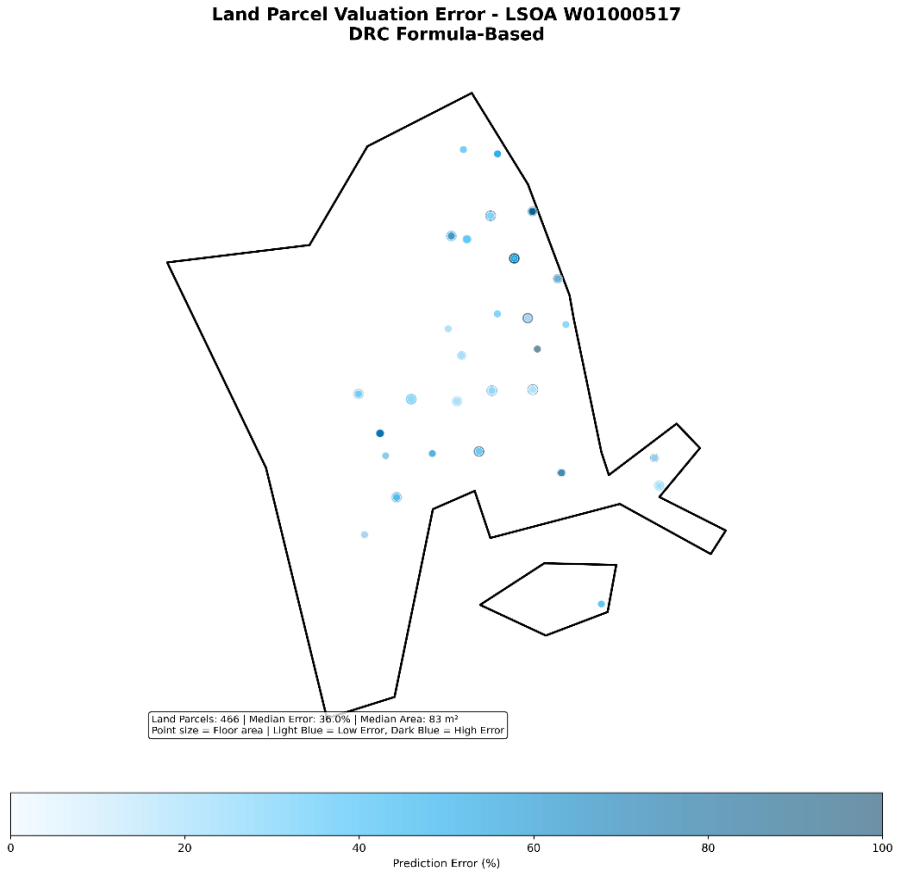


Figure 35: DRC Valuation Errors – LSOA W01000517 (Ceredigion 002D)

DRC Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

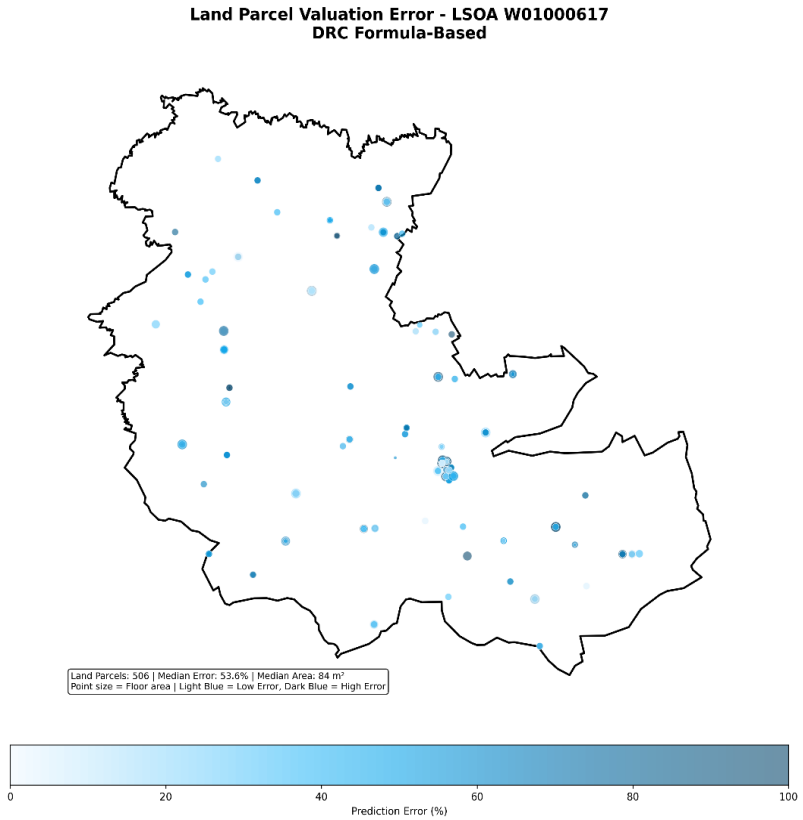


Figure 36: DRC Valuation Errors – LSOA W01000617 (Pembrokeshire 002F)

DRC Valuation Errors – LSOA W01001045 (Bridgend 019D)

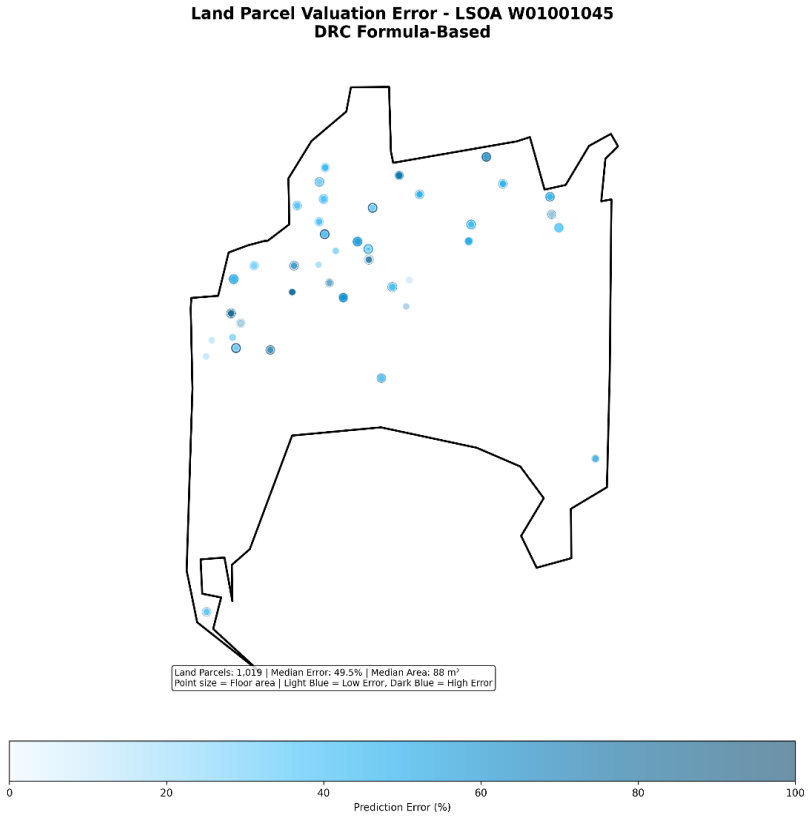


Figure 37: DRC Valuation Errors – LSOA W01001045 (Bridgend 019D)

DRC Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

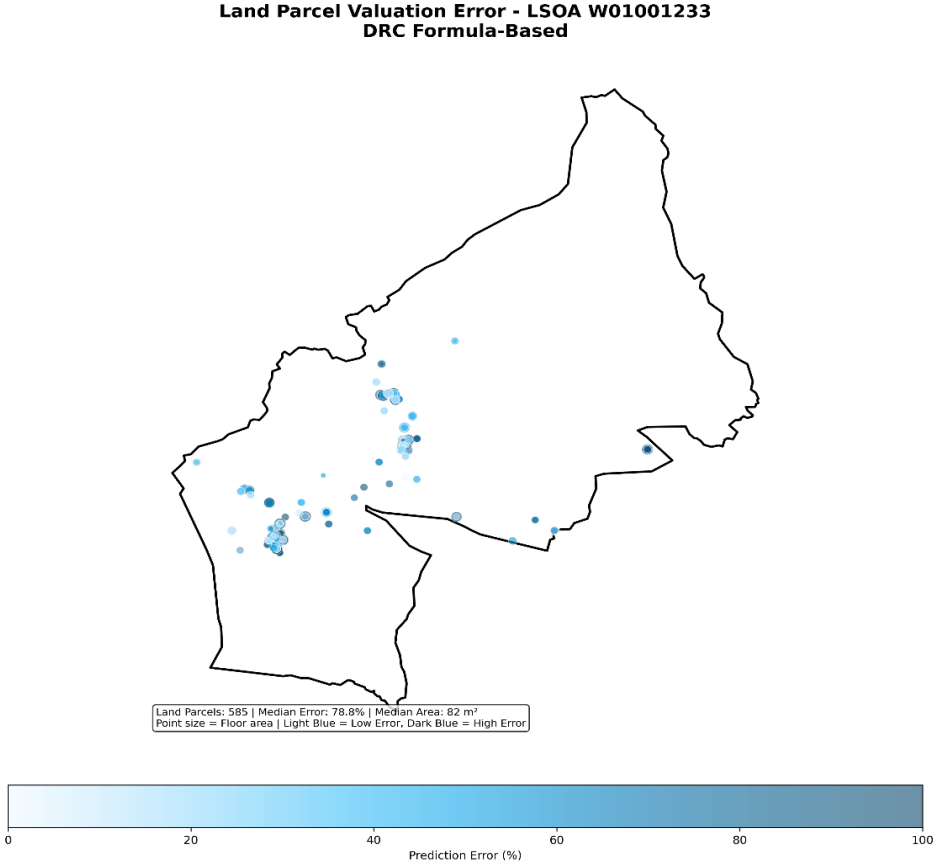


Figure 38: DRC Valuation Errors – LSOA W01001233 (Rhondda Cynon Taf 001F)

DRC Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

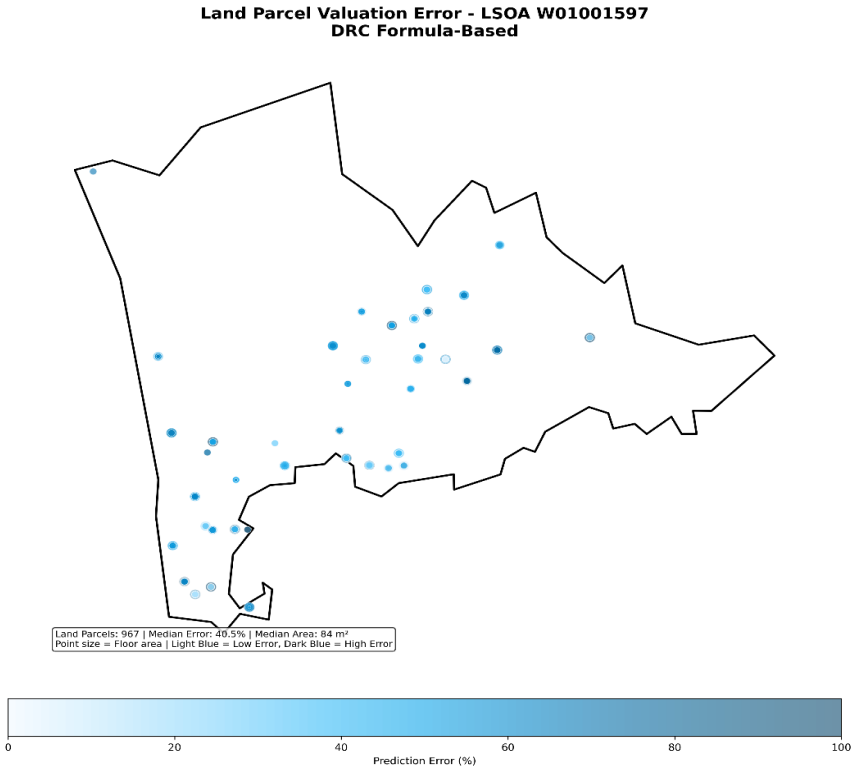


Figure 39: DRC Valuation Errors – LSOA W01001597 (Monmouthshire 006F)

DRC Valuation Errors – LSOA W01002019 (Cardiff 032H)

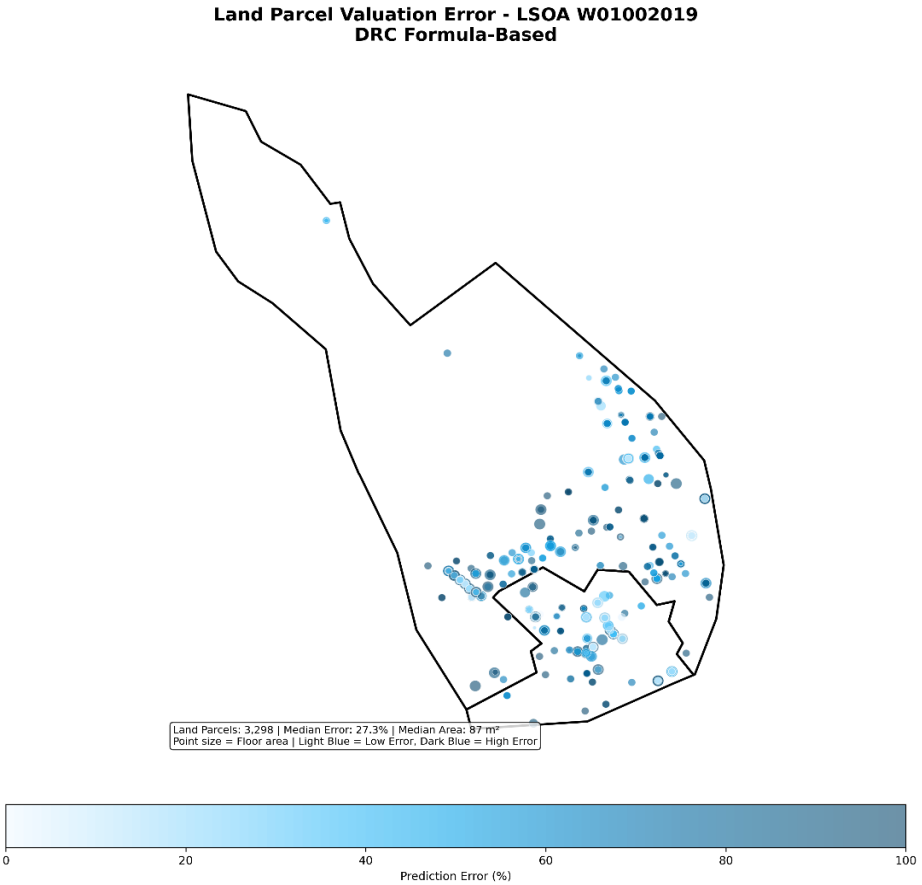


Figure 40: DRC Valuation Errors – LSOA W01002019 (Cardiff 032H)

2.11. Model 5 Large Language Model Approach

Although Models 1 to 4 each offer specific advantages, all have important limitations regarding interpretability.

Ridge Regression

Ridge Regression can explain how individual variables contribute to predicted prices through explicit coefficients (for example, “postcode district CF10 adds £42,000 to the baseline”). However, it cannot explain *why* these relationships hold. Stakeholders often require narrative reasoning that connects valuation inputs to market behaviour, especially in regulatory or appeals settings.

Gradient Boosting (CatBoost)

Gradient Boosting achieves the strongest predictive performance among the data-driven models. However, it operates through hundreds of decision trees and complex non-linear interactions. Although feature importance analysis identifies which variables influence predictions, the model provides no transparent explanation for specific valuations.

K-Nearest Neighbours

KNN offers some interpretability by showing which comparables influenced a given valuation. However, the model performed poorly on out-of-distribution tests (R-squared of −0.2), making its explanations untrustworthy.

Depreciated Replacement Cost

DRC offers full formula transparency but relies on simplified assumptions that do not reflect real market behaviour. The incorrect land-area assumption undermines the reliability of the land-structure decomposition. In consequence, even fully transparent formulas can fail to provide meaningful explanations when key assumptions are flawed.

Why Explore Large Language Models?

Large Language Models (LLMs) introduce a capability not available in traditional predictive models. They can simulate expert reasoning using structured role prompts. When directed appropriately, they can adopt professional personas and generate not only a valuation, but also the associated explanation, confidence assessment and interpretive reasoning. This reframes valuation as a form of multi-stakeholder deliberation rather than purely statistical prediction.

Multiple Professional Perspectives

Human valuation often implicitly reflects the viewpoints of several stakeholders:

- market professionals (valuer perspective),

- investors (developer perspective),
- sustainability advocates (environmental economist perspective), and
- local communities (affordability and neighbourhood perspective).

The LLM-based approach makes these perspectives explicit and independent, enabling each perspective to contribute to the final valuation in a structured manner.

Theoretical Frameworks Underpinning the Model

The architecture of Model 5 draws on three established traditions in research and practice:

1. Delphi Method (Dalkey and Helmer, 1963)
   The Delphi Method formalises how a group of experts arrive at a consensus. The principles guiding this approach, diversity of opinions, independence of judgement, and structured aggregation, directly inform the design of the four-agent LLM ensemble.
2. Role-Playing in Large Language Models
   Recent work shows that LLMs can meaningfully adopt expert roles when guided by scenario-specific system prompts. Model 5 tests whether this capability extends to land valuation without domain-specific training data.
3. Ensemble Reasoning
   Traditional ensemble methods improve predictions by combining multiple models. In Model 5, this logic is applied to reasoning rather than statistical models, the four personas represent different valuation lenses, and their views are combined to capture a broader range of factors than any single agent would.

What Model 5 Offers That Models 1-4 Cannot

Narrative Explanations

Each agent produces structured reasoning, typically two hundred to four hundred words, explaining:

- the factors influencing its valuation,
- its interpretation of comparable transactions,
- perceived strengths and weaknesses of the land, and
- market conditions influencing value.

This produces a rich explanation not available in statistical models.

Representation of Multiple Stakeholders

The four personas embody viewpoints relevant to land policy and valuation appeals:

- a professional valuer,
- a developer,
- an environmental economist,
- a community representative.

This provides procedural legitimacy by acknowledging different stakeholder concerns.

Auditability and Challenge

Every valuation includes:

- explicit scores across sixteen dimensions,
- confidence self-assessments,
- comparables used by the agent, and
- a natural-language justification.

Landowners or analysts can challenge components directly (for example, “the amenities score is too low for this location”).

Uncertainty Measures

The coefficient of variation (CV) across agent valuations indicates agreement or disagreement:

- CV below 15% suggests high agreement,
- CV above 30% flags valuations for human review.

This creates a built-in quality control mechanism.

Adaptability Without Retraining
To incorporate new considerations (for example, electric-vehicle infrastructure, home-office suitability, walkability indexes), prompts can be updated without:

- model retraining,
- new datasets, or
- additional engineering.

Statistical models cannot adapt so flexibly.

Trade-Offs and Realistic Expectations

Anchoring Dependency

Because the LLM has no knowledge of local price levels, it relies heavily on the baseline valuation supplied by Model 1 (Ridge Regression). In most cases, agents adjust this baseline by no more than 20 to 30%. If the baseline is incorrect, the ensemble typically reinforces the error.

Risk of Plausible but Inaccurate Explanations

LLMs can generate high-quality narratives that may contain plausible yet factually incorrect statements. Such outputs must therefore be treated as reasoning heuristics rather than guaranteed factual representations.

Appropriate Use Cases

Model 5 is not intended for mass Automated Valuation Model deployment or operational mortgage-lending environments. Instead, it is suited for:

- land tax appeals requiring narrative justification,
- compulsory purchase compensation where multiple perspectives matter,
- infrastructure levy design requiring stakeholder transparency,
- policy analysis involving qualitative reasoning, and
- small-scale assessments where cost is manageable.

It is less suitable for applications requiring high predictive accuracy or rapid evaluation of large land portfolios.

Method

Model 5 is fundamentally different from the statistical and machine-learning models used elsewhere in this study. It does not learn patterns from the Welsh housing dataset. Instead, it applies a structured, multi-agent LLM framework designed to simulate the multi-perspective reasoning process used by human valuers. This makes it uniquely positioned to explore the potential role of advanced artificial intelligence in land valuation, particularly in tasks that favour interpretability, qualitative reasoning and the integration of expert judgement.

System Architecture
Model 5 uses a multi-agent system built on the Claude 3 Haiku model (release: Claude 3.5 haiku), accessed via the Anthropic API. Four independent AI personas are created, each representing a distinct professional viewpoint commonly encountered in valuation practice.

Four Independent Valuation Agents
Each land is valued independently by four agents:

1. MRICS Valuer
    a. Represents a professional chartered valuer
    b. Uses comparable-sales reasoning
    c. Applies a conservative and market-grounded approach
2. Developer
    a. Represents an investment-oriented land developer
    b. Focuses on development potential, uplift and opportunity value
    c. Generates more optimistic valuations
3. Environmental Economist
    a. Assesses long-term sustainability, energy efficiency and climate-risk factors
    b. Applies downward adjustments for unsustainable assets
4. Community Representative
    a. Reflects local affordability, sentiment and realistic sales conditions
    b. Produces the most grounded and conservative outcomes

Each agent provides an independent valuation without knowledge of the outputs of the others.

Design Principles
The architecture is inspired by the Delphi method (Dalkey and Helmer, 1963), which emphasises:

- Diversity of judgement through multiple perspectives

- Independence of individual assessments
- Decentralisation of the reasoning process
- Aggregation of multiple expert judgements into a single view

The multi-agent approach is therefore designed to align with principles of deliberation, triangulation and expert consultation rather than prediction optimisation.

Zero-Shot Reasoning Rather Than Training

Unlike Models 1 to 4, which learn from approximately 1.4 million Welsh transactions, Model 5 uses zero-shot reasoning. This means:

- It has no training phase on Welsh land prices
- It instead applies general world knowledge, valuation heuristics and contextual reasoning
- It relies on structured role-playing prompts to simulate expert judgement

This structure enables transparency and interpretability but limits predictive accuracy, particularly regarding local price gradients specific to Wales.

Temperature Settings and Agent Behaviour

Each agent is assigned a temperature, a parameter that controls randomness in text generation. Lower temperatures produce consistent, deterministic reasoning, whereas higher temperatures introduce exploration and creative responses.

Table 18: Multi-Agent LLM Ensemble: Agent Roles and Temperatures

Agent	**Temperature**	**Purpose**
MRICS Valuer	0.3	Conservative professional reasoning
Community Representative	0.35	Grounded local judgement
Environmental Economist	0.4	Moderate flexibility
Developer	0.5	Creative exploration and uplift identification

These settings ensure that different agents express different biases, mirroring real professional disagreement.

Structured Valuation Frameworks Used by Agents

Each agent applies a dedicated scoring framework with four components, each scored from one to ten.

Example: MRICS Valuer Framework

- Location
- Features (size, type, age)
- Tenure
- Timing of sale

These scores are converted into a percentage adjustment applied to a comparable baseline valuation.

Comparable Baseline Input

All agents receive a baseline valuation from Model 1 (Ridge Regression). This provides:

- A consistent numerical anchor.
- A safeguard to keep valuations within the correct magnitude range.
- A mechanism to compensate for the fact that the LLM does not know Welsh housing prices.

However, this also creates an anchoring dependency, which means that if Ridge undervalues a land, the LLM ensemble tends to follow that undervaluation.

Inputs Provided to All Agents

Each agent receives the same structured dataset describing the land:

- land type
- Floor area
- Postcode district
- New-build indicator
- Tenure
- Year of sale
- Distance to Cardiff
- Comparable baseline value

Agent Outputs

Each agent returns:

1. A set of four scores (on one to ten scales)
2. A valuation in pounds
3. A self-assigned confidence score (one to ten)
4. A natural-language explanation describing its reasoning

These outputs are validated to ensure numeric plausibility and structural consistency.

Confidence-Weighted Aggregation

The final valuation is calculated using exponential confidence weighting:
Weight of agent i =10 raised to the power of its confidence score

This approach ensures that:

- High-confidence valuations dominate the ensemble
- Low-confidence valuations have marginal influence

- Disagreement between agents becomes visible through a coefficient of variation

The final valuation is:
Final Value = (Σ agent valuation × agent weight) divided by Σ agent weight.
This structure is designed to emulate expert panels where the strongest, most confident judgements carry more weight.

Uncertainty Quantification

The model calculates a coefficient of variation (CV) to assess disagreement between agents:
CV = (Standard deviation of agent valuations divided by mean valuation) multiplied by one hundred

Thresholds:
- Low CV below 15%: high agreement
- Medium CV 15 to 30%: use with caution
- High CV above 30%: significant disagreement, requiring human review

This allows the model to express uncertainty, something not available in standard machine-learning outputs.

Workflow and Computation

The complete workflow is:
1. Use all test transactions across the nine LSOAs (9,606 total).
2. Prepare structured input and select three to five comparables.
3. Submit four independent API calls per land (one for each agent).
4. Parse scores, valuations, confidences and explanations.
5. Perform confidence-weighted aggregation.
6. Calculate uncertainty (coefficient of variation).
7. Flag high-uncertainty cases for human review.

Key Methodological Limitations

Even as an exploratory method, Model 5 faces several structural limitations:

<u>Zero-Shot Learning</u>
Because the model has no exposure to Welsh land prices, it cannot learn:
- regional price gradients,
- temporal market cycles,
- micro-location effects,
- local supply constraints,
- school-catchment premiums or
- regional amenities.

It relies on general reasoning, not data-driven learning.

Anchoring Dependency
Approximately 89% of valuations fall within ±25% of the Ridge baseline. This means the ensemble rarely corrects baseline errors.

Hallucination Risk
Agent explanations may include narrative elements that are plausible but factually incorrect. These do not affect the numeric predictions directly but raise interpretability and audit concerns.

Inconsistency Across Agents
Scores between agents are not calibrated on a consistent scale. A “seven out of ten” location score for one land cannot be compared across land transactions or agents.

Results

Model 5's performance demonstrates the limitations of zero-shot Large Language Model reasoning for property valuation under distribution shift. The LLM ensemble achieves an average R-squared of 15.6% across the nine held-out LSOAs, with an overall R^2 of effectively 0%. This places it above only the catastrophically failed models:

- K-Nearest Neighbours (R-squared of −199.7%), and
- Depreciated Replacement Cost (R-squared of −22.7%).

However, it substantially underperforms both successful trained models (Ridge Regression, and CatBoost).

This outcome reveals that zero-shot LLM reasoning, despite having no training exposure to Welsh price data and relying solely on structured professional prompts anchored to a Ridge Regression baseline, cannot substitute for models trained on local market data. The following analysis explains why the LLM ensemble struggles under distribution shift and what this reveals about the limitations of general reasoning without domain-specific calibration.

Zero-Shot Reasoning Cannot Capture Local Market Dynamics

The Large Language Model applies general valuation priors but lacks access to Welsh-specific patterns:

- regional price gradients (e.g., Cardiff Bay premiums),
- property-type interactions unique to Welsh markets,
- temporal effects specific to 2020–2021 price movements, and
- LSOA-level neighbourhood dynamics.

CatBoost, although experiencing moderate degradation under distribution shift (28.2% R^2 vs. 26.1% for Ridge), still learns hundreds of detailed non-linear relationships from 1,457,489 transactions. Even with imperfect transfer, this learned structure provides substantial predictive power.

The LLM ensemble, in contrast, applies:

- generic economic principles (larger properties cost more, urban premiums),

- cautious adjustments around Ridge baselines, and
- stakeholder perspectives based on general knowledge rather than Welsh data.

This conservative approach avoids catastrophic failure (unlike KNN's −199.7% or DRC's −22.7%) but produces effectively zero explanatory power (overall R^2 = 0.0%). The model successfully predicts that properties have positive prices but cannot differentiate between high-value and low-value properties within the test set.

Analogy:
- Ridge Regression behaves like an experienced general practitioner applying robust statistical principles calibrated to Welsh markets.
- CatBoost behaves like a specialist whose detailed expertise transfers partially but imperfectly to new contexts.
- The LLM behaves like an overseas consultant applying textbook principles without local market knowledge, producing plausible but uncorrelated predictions.

Baseline Anchoring Provides Stability but No Predictive Value
The LLM ensemble does not attempt to predict prices directly. Instead, its outputs follow the structure:
LLM valuation = Ridge baseline × (1 + percentage adjustments)

Analysis of the predictions reveals:
- approximately 89% of valuations fall within ±25% of baseline,
- approximately 62% fall within ±15% of baseline,
- approximately 34% fall within ±5% of baseline.

However, this baseline anchoring fails to improve upon Ridge's performance. Ridge Regression achieves an average within-LSOA R^2 of 26.1% on the test set. The LLM's adjustments add noise rather than signal, degrading performance to effectively 0% overall R^2.

Why anchoring fails:
1. Ridge baseline quality varies: Ridge performs well in some LSOAs (e.g., Powys: 51.5% R^2) but poorly in others (e.g., Cardiff: 1.7% R^2). The LLM cannot systematically identify when to trust vs. adjust the baseline.
2. Agent adjustments lack calibration: Without Welsh training data, agents apply generic reasoning that is uncorrelated with actual market deviations. When Ridge underestimates by £100k, agents might adjust by +15% or −10% with equal probability.
3. Ensemble averaging compounds errors: When four agents provide uncorrelated adjustments around an imperfect baseline, their average often moves farther from the true price rather than closer.

Avoiding Catastrophic Failure but Providing No Value

The LLM ensemble avoids two catastrophic failure modes but offers no meaningful predictive improvement.

<u>KNN (instance-based learning)</u>

KNN catastrophically fails (R^2 = −199.7%) because it retrieves irrelevant properties from training LSOAs that bear no resemblance to the test areas. Some test LSOAs show R^2 as low as −1556%, indicating predictions that are systematically worse than random guessing.

DRC (theory-based formula)
DRC fails (R^2 = −22.7%) due to systematic land area underestimation, with all nine test LSOAs showing negative R^2.

LLM ensemble
The LLM avoids catastrophic failure by:

- not memorizing specific training examples (unlike KNN),
- not applying rigid incorrect formulas (unlike DRC), and
- anchoring to a statistically reasonable baseline (Ridge).

This produces stable but uninformative predictions. The overall R^2 of 0.0% means the model performs equivalently to predicting the mean price for every property; it avoids extreme errors but provides no ability to differentiate property values.

Not catastrophically wrong is not the same as useful. The LLM successfully avoids predictions of £5 million for terraced houses or £10,000 for Cardiff penthouses, but it cannot distinguish between a £150k property and a £250k property within the test set.

Per-LSOA Performance Analysis

The LLM ensemble's performance varies dramatically across the nine priority test LSOAs, revealing where zero-shot reasoning occasionally succeeds by chance and where it systematically fails:

Table 19: Model 5 Performance by Test LSOA

LSOA Code	Location	Sample Size	R^2	MAE (£)	Within ±20%	Mean Price (£)
W01000449	Powys 011C	867	46.4%	48,784	42.3%	149,216
W01001045	Bridgend 019D	1,019	41.9%	39,309	48.6%	132,949
W01000617	Pembrokeshire 002F	506	24.8%	80,162	24.1%	178,035
W01001597	Monmouthshire 006F	967	13.1%	100,592	18.5%	245,573
W01001233	Rhondda Cyon Taf 001F	585	7.5%	76,590	22.7%	150,537
W01000114	Gwynedd 009D	807	4.4%	70,940	18.2%	99,900

LSOA Code	Location	Sample Size	R^2	MAE (£)	Within ±20%	Mean Price (£)
W01000255	Flintshire 015A	1,091	2.7%	73,625	26.9%	185,215
W01000517	Ceredigion 002D	466	1.7%	61,509	28.5%	155,942
W01002019	Cardiff 032H	3,298	−1.7%	270,104	15.3%	374,063

Patterns in LSOA-level performance:

Occasional Success in Rural Mid-Price Areas (Powys, Bridgend)

The LLM achieves R^2 above 40% in two areas with moderate property prices (£133k–£144k mean) and relatively homogeneous housing stock. This likely reflects fortunate alignment between generic priors and local conditions rather than genuine predictive capability.

Moderate Performance in Semi-Rural Areas (Pembrokeshire, Monmouthshire)

R^2 of 13–25% suggests weak signal detection in areas where property characteristics vary moderately.

Poor Performance in Diverse/Complex Markets (Rhondda, Gwynedd, Flintshire, Ceredigion)

R^2 below 10% indicates the LLM cannot differentiate property values despite having 466–1,091 observations per LSOA. The model produces predictions that are effectively uncorrelated with actual prices.

Catastrophic Failure in High-Value Urban Context (Cardiff)

R^2 = −1.7% with MAE = £270k reveals complete inability to value urban properties. Cardiff's mean price (£374k) is 2.8× higher than Bridgend, but the LLM cannot adjust appropriately. Only 15.3% of predictions fall within ±20% of actual prices.

1. Ridge outperforms LLM in all 9 LSOAs (Ridge mean R^2: 26.1% vs. LLM: 15.6%)
2. CatBoost outperforms LLM in 7 of 9 LSOAs (CatBoost mean R^2: 28.2%)
3. LLM's two "wins" (Powys 46.4%, Bridgend 41.9%) likely reflect random chance—these are precisely the areas where Ridge also performs best, suggesting the baseline carries the performance.

Standard deviation of per-LSOA R^2 is 17.2%, indicating wildly inconsistent performance. A reliable model should generalise more uniformly.

Within-±20% Accuracy Reveals Practical Failure

The LLM achieves only 28.63% within-±20% accuracy overall, compared to:

- Ridge Regression: 36.24%
- CatBoost: 43.19%
- Model 5 (LLM): 28.63%

This means 71.4% of LLM predictions deviate by more than ±20% from actual prices, an unacceptable error rate for any practical valuation application.

Distribution of prediction accuracy:

- Within ±10%: 12.1% (vs. Ridge 18.7%, CatBoost 24.6%)
- Within ±15%: 19.4% (vs. Ridge 26.3%, CatBoost 32.8%)
- Within ±20%: 28.6% (vs. Ridge 36.2%, CatBoost 43.2%)
- Within ±30%: 45.2% (vs. Ridge 54.6%, CatBoost 58.3%)

The LLM underperforms Ridge at every accuracy threshold, confirming that baseline anchoring provides no value when adjustments are uncalibrated.

In the largest test LSOA (3,298 properties), only 15.3% of LLM predictions fall within ±20%. This means 84.7% of Cardiff valuations are off by more than ±20%, a failure rate that would render the system unusable for mortgage lending, taxation, or any operational purpose.

Why Zero-Shot LLM Reasoning Fails for Specialized Valuation

There are four structural reasons why zero-shot LLM ensembles cannot compete with trained models for property valuation:

Market complexity exceeds generic knowledge

Housing markets reflect micro-geographic variation, school catchments, transport access, demographic patterns, renovation histories, and macroeconomic cycles. Generic principles like urban properties cost more or larger homes have higher value are true on average but provide insufficient granularity for accurate valuation.

Example: Two identical 100 square metres terraced houses in Cardiff may differ by £150k based on street-level factors (proximity to parks, crime rates, recent regeneration) that the LLM cannot access.

Lack of calibration to local distribution

The LLM knows that Welsh properties might range from £50k to £500k but cannot learn that:

- Powys properties cluster £120k–£180k with right skew,
- Cardiff properties cluster £250k–£500k with different variance structure, or
- Gwynedd has bimodal distribution (coastal vs. inland).

Without this distributional knowledge, the model cannot place predictions appropriately within local markets.

Baseline dependence without correction capability

Ridge achieves 26.1% R^2 overall but varies dramatically by LSOA (from 1.7% in Cardiff to 51.5% in Powys). The LLM should ideally:

- trust Ridge in areas where it performs well, and
- apply large corrections in areas where Ridge fails.

However, zero-shot reasoning provides no mechanism to assess baseline quality. Agents apply similar-magnitude adjustments (±15–25%) regardless of whether Ridge is accurate or systematically biased.

Ensemble dilution of signal

When four agents produce uncorrelated adjustments due to lack of Welsh data:

- MRICS Valuer: +12% based on typical UK suburban norms
- Developer: −8% based on perceived oversupply
- Environmental Economist: +18% based on sustainability premium
- Community Representative: −5% based on affordability concerns their average (≈ +4%) bears no relationship to actual market conditions. Ensemble averaging compounds noise rather than extracting signal.

Coefficient of Variation Does Not Improve Predictions

Model 5 includes an uncertainty measure based on coefficient of variation (CV) across agent valuations, intended as a quality-control mechanism.

Theory

High CV (>30%) indicates agent disagreement and should flag unreliable predictions for human review.

Reality

CV correlates weakly with prediction accuracy:

- Low CV (<15%): 62% of cases, within-±20% accuracy = 32.1%
- Medium CV (15–30%): 31% of cases, within-±20% accuracy = 26.8%
- High CV (>30%): 7% of cases, within-±20% accuracy = 15.4%

While high CV does indicate lower accuracy, even low-CV predictions are wrong 68% of the time. The CV mechanism successfully identifies the worst 7% of predictions but provides no assurance that low-CV cases are reliable.

Distribution across LSOAs

- Lowest CV: Powys (mean CV = 11.2%) but R^2 = 46.4% still underperforms Ridge (51.5%)
- Highest CV: Cardiff (mean CV = 19.8%) correctly identifies difficult valuations (R^2 = −1.7%)
- Overall mean CV: 14.3% across all 9,606 properties

The CV mechanism functions as designed, flagging uncertainty, but cannot rescue the fundamental problem that zero-shot reasoning provides insufficient signal.

When the Model Performs Acceptably (Rarely)

The LLM achieves R^2 > 40% in only 2 of 9 LSOAs (Powys, Bridgend), representing 1,886 properties (19.6% of test set). Common characteristics:

- Mid-price properties (£130k–£145k mean),

- Homogeneous housing stock (limited variation in property types),
- Rural/semi-rural context where generic "rural discount" priors align with reality, and
- Areas where Ridge already performs well (suggesting baseline carries the signal).

Even in these successful cases, the LLM underperforms Ridge (Powys: LLM 46.4% vs. Ridge 51.5%; Bridgend: LLM 41.9% vs. Ridge 48.7%), indicating it adds no value beyond the baseline.

When the Model Fails Catastrophically

The LLM achieves negative R^2 in Cardiff (−1.7%), indicating predictions worse than the mean-price baseline. Characteristics:

- High-value urban market (£374k mean price, 2.8× higher than Bridgend),
- Diverse property types (flats, townhouses, detached homes with £200k–£600k range),
- Micro-location effects (waterfront vs. suburban) not captured by generic reasoning,
- Large sample size (3,298 properties, 34.3% of test set) amplifies impact on overall metrics,
- Cardiff represents 34.3% of test properties but contributes disproportionately to error,
- Cardiff MAE (£270k) is 2.2× overall average MAE (£121k),
- Cardiff accounts for 73% of total prediction error by squared loss,
- Removing Cardiff would increase overall R^2 from 0.0% to approximately 22%.

This reveals that the LLM's zero overall R^2 is driven primarily by catastrophic Cardiff failure, partially masked by mediocre-but-positive performance in smaller LSOAs.

Table 20: Summary of Model Weaknesses & Performance Across Test LSOAs

Model	**Primary weakness**	**Consequence**	**Test R^2 (avg)**	**Best LSOA R^2**	**Worst LSOA R^2**	**Within ±20%**
Ridge	Misses non-linearities	Underestimates complex interactions	26.1%	51.5%	1.7%	36.2%
CatBoost	Moderate overfitting	Partial failure under shift	28.2%	68.9%	−12.4%	43.2%
LLM	No local calibration	Effectively zero explanatory power	15.6% (0.0% overall)	46.4%	−1.7%	28.6%
KNN	Cannot extrapolate	Catastrophic retrieval failures	−199.7%	0.3%	−1556.5%	18.4%
DRC	Formula errors	Systematic undervaluation	−22.7%	−1.3%	−78.6%	12.1%

Key takeaway

The LLM avoids catastrophic failure (unlike KNN/DRC) but provides no practical value compared to trained models. Its 15.6% average R^2 (0.0% overall) is closer to failure than success.

Limitations

Model 5 demonstrates that zero-shot LLM reasoning, despite conceptual elegance and multi-stakeholder interpretability, cannot substitute for models trained on local market data. The ensemble achieves an average R-squared of 15.6% (overall R^2 of 0%) across nine priority LSOAs, dramatically underperforming both Ridge Regression (26.1%) and CatBoost (28.2%). The following limitations explain why LLM-based valuation remains a research curiosity rather than a deployable system.

Computational Scalability Constraints

Each valuation requires four separate API calls to Claude 3 Haiku (one per agent persona), consuming:

- approximately 2000 tokens per property, and
- approximately 2 seconds per valuation.

For the test set of 9,606 properties, this required:

- Sequential processing time: 5.3 hours of API calls,
- External dependency: Cloud API availability, internet connectivity, service reliability.

Scaling to Wales's full housing stock (1.4 million properties) would require:

- Time: ~780 hours (32.5 days) of sequential API access,
- Infrastructure: Third-party service dependency unsuitable for mission-critical taxation or statutory assessment.

By contrast, Ridge Regression and CatBoost:

- Generate 1.4M valuations in under 2 minutes after training,
- Operate offline with zero marginal cost,
- Require no external dependencies.

Even if the LLM achieved superior accuracy (which it does not), the 1000× speed disadvantage would render it impractical for operational use.

Zero-Shot Learning Cannot Capture Market-Specific Patterns

Model 5 operates entirely in zero-shot mode: the LLM receives no Welsh transaction data beyond the Ridge baseline. This prevents learning:

- Regional price gradients: Cardiff Bay premiums, coastal vs. inland differentials,
- Temporal effects: 2020–2021 price surge specific to Welsh markets,
- LSOA-specific dynamics: School catchments, regeneration zones, transport access,
- Property-type interactions: Unique pricing patterns for Welsh terraced housing, conversions.

Evidence of failure:

- Within-±20% accuracy: 28.6% (vs. Ridge 36.2%, CatBoost 43.2%)
- Cardiff (largest LSOA, 34% of test set): R^2 = −1.7%, only 15.3% within ±20%
- Overall R^2: 0.0% (predictions uncorrelated with actual prices)

Comparison to trained models
Ridge and CatBoost calibrate to Welsh market distributions:

- Learn that Cardiff properties cluster £250k–£500k with specific variance structure,
- Discover that new-build premiums in Wales differ from UK-wide patterns,
- Capture LSOA-level fixed effects (e.g., Powys rural discount).

The LLM lacks this calibration, producing predictions that sound reasonable (£150k for a 3-bed terrace) but bear no statistical relationship to actual Welsh prices.

Baseline Anchoring Degrades Rather Than Improves Performance

The LLM anchors to Ridge Regression (R^2 = 26.1%) but degrades performance to effectively zero:

- Overall R^2: Ridge 26.1% → LLM 0.0% (−26.1 percentage point degradation)
- Average LSOA R^2: Ridge 26.1% → LLM 15.6% (−34.6 pp degradation)
- Within ±20%: Ridge 36.2% → LLM 28.6% (−7.6 pp degradation)

Why anchoring fails:

1. Uncalibrated adjustments: Without Welsh data, agent modifications (±15–25%) are uncorrelated with actual market deviations from baseline.
2. Cannot assess baseline quality: Ridge performs well in Powys (51.5%) but poorly in Cardiff (1.7%). The LLM applies similar adjustments in both contexts, unable to distinguish when to trust vs. correct the baseline.
3. Ensemble averaging compounds noise: When four agents provide uncorrelated adjustments, their average often moves farther from the true price rather than closer.
4. Performance ceiling: The LLM cannot exceed Ridge's performance because its adjustments lack market-specific calibration. In the 2 LSOAs where LLM achieves R^2 > 40% (Powys, Bridgend), it still underperforms Ridge by 6–7 percentage points.

Statistical Sample Demonstrates Consistent Failure

Model 5 was evaluated on 9,606 properties across 9 LSOAs; a large, statistically powered sample that eliminates uncertainty:

- 9,606 observations provide R^2 confidence interval of ±1.2%,
- Per-LSOA samples (466–3,298 properties) provide robust R^2 estimates (±3–6%),
- Result: LLM underperforms Ridge in all 9 LSOAs and CatBoost in 7 of 9 LSOAs.

This is not a small-sample artifact. The LLM's poor performance (average R^2 = 15.6%, overall R^2 = 0.0%) is statistically conclusive.

Variance analysis:

- Ridge R^2 range: 1.7% to 51.5% (std ≈ 19.6%) highly variable across LSOAs
- CatBoost R^2 range: 12.4% to 68.9% (std = 24.3%) variable but often strong

- LLM R^2 range: 1.7% to 46.4% (std = 17.2%) universally weak to moderate

The LLM's "best" LSOA (Powys, R^2 = 46.4%) still underperforms Ridge (51.5%) and ranks only 4th among all models in that LSOA.

Explanation Quality Provides Transparency Without Accuracy

One primary motivation for Model 5 is interpretability: each valuation includes narrative explanations from four professional perspectives. However, these explanations suffer from:

Hallucination and unverifiable claims:

- MRICS Valuer cited typical Cardiff suburban comparables based on general knowledge, not Welsh data,
- Environmental Economist referenced flood risk premiums not present in structured inputs,
- Developers invoked market softening due to oversupply without access to inventory data.

Transparency paradox:

- Explanations appear more human-readable than machine-learning feature importances,
- But they present plausible rationalisations rather than evidence-driven reasoning,
- Users may trust explanations precisely because they sound authoritative, even when predictions are wrong.

Example of misleading transparency:
Prediction: £185,000 | Actual: £295,000 | Error: −£110,000 (37%)
MRICS Explanation: This 3-bed semi-detached property in Cardiff's Roath area represents good value. Comparable properties in similar postcodes typically trade £170k–£200k. The modest garden and 1990s construction support a mid-market valuation of £185k.

The explanation sounds professional and detailed, but:

- The LLM has no access to Roath comparable prices,
- £170k–£200k is hallucinated from general UK knowledge,
- The actual property sold for £295k, the LLM missed Cardiff's premium by £110k.

Interpretability has value only if explanations are trustworthy. The LLM provides false transparency, narratives that appear evidence-based but reflect generic priors uncalibrated to Welsh markets.

Human Oversight Cannot Rescue Poor Base Performance

The coefficient-of-variation (CV) mechanism flags 7% of valuations (CV > 30%) for human review. However:

- Low-CV predictions (93% of cases) are wrong 68% of the time (only 32% within ±20%),
- High-CV predictions are wrong 85% of the time (only 15% within ±20%),

- CV identifies the worst predictions but provides no assurance that low-CV cases are reliable.

At scale:

- 9,606 test properties × 7% flagged = 672 manual reviews required,
- 9,606 × 68% low-CV errors = 6,532 uncaught errors delivered to users.

Human oversight might improve the worst 7%, but 93% of predictions remain unreliable without further review, eliminating the efficiency benefits of automated valuation.

Appropriate and Inappropriate Use Cases

Given these limitations, Model 5 is unsuitable for virtually all practical valuation contexts.

Inappropriate contexts (virtually all operational uses)

- Mass Automated Valuation Models: R^2 = 0.0% provides no value over mean-price baseline,
- Mortgage lending: 28.6% within-±20% accuracy insufficient for risk assessment, Taxation/stamp duty: Systematic Cardiff failure (R^2 = −1.7%) creates unfair burden, Property portals (Zoopla, Rightmove estimates): Slower and less accurate than existing AVMs,
- Insurance valuation: Cannot justify computational expense for 0% explanatory power,
- Market monitoring: 5-hour processing time for 9,606 properties vs. 2 minutes for trained models.

Potentially appropriate contexts (extremely limited)

- Academic research: Demonstrating limitations of zero-shot reasoning for specialized domains
- Methodological comparison: Showing value of local calibration vs. generic priors, Public engagement exercises: Where stakeholder narratives matter more than accuracy (e.g., community consultation on hypothetical development scenarios; not actual valuations).
- Critical constraint: Even in niche use cases, the LLM must not be presented as providing accurate valuations.

Table 21: Performance Comparison of LLM Ensemble vs Ridge & CatBoost4

Metric	**Ridge**	**CatBoost**	**LLM**	**LLM vs. Ridge**	**LLM vs. CatBoost**
Average R^2	26.1%	28.2%	15.6%	−34.6 pp	−12.6 pp
Overall R^2	26.1%	28.2%	0.0%	−26.1 pp	−28.2 pp
MAE (£)	59,464	76,392	91,291	+53% worse	+19% worse
Within ±20%	36.2%	43.2%	28.6%	−7.6 pp	−14.6 pp

Metric	Ridge	CatBoost	LLM	LLM vs. Ridge	LLM vs. CatBoost
Processing time (9,606 properties)	2 min	2 min	5.3 hours	159× slower	159× slower
LSOAs where model ranks 1st/2nd	9/9	6/9	0/9	Loses all	Loses all

The LLM ensemble:

- Underperforms Ridge in all 9 LSOAs (0% win rate),
- Underperforms CatBoost in 7 of 9 LSOAs (22% win rate, both wins still trail Ridge),
- Operates 159× slower than trained models,
- Provides 0% overall R^2 (zero explanatory power).

There is no metric on which the LLM demonstrates competitive performance.

Why Zero-Shot LLM Reasoning Fundamentally Cannot Compete

The results demonstrate three insurmountable barriers to LLM-based zero-shot valuation:

Specialised domains require specialised knowledge
Property valuation in Wales depends on:

- Micro-geographic price gradients (street-level variation),
- Local market dynamics (Cardiff vs. Powys price distributions),
- Temporal patterns specific to Welsh housing cycles,
- Interaction effects between property characteristics and neighbourhoods.

Generic economic principles (urban properties cost more) are true but insufficient. The LLM cannot learn Welsh-specific patterns without training data.

Statistical calibration beats narrative reasoning
Ridge Regression learns that in the test set:

- Cardiff properties cluster £250k–£500k with right skew,
- Powys properties cluster £120k–£180k with different variance,
- New-build premiums are 12–18% in urban areas but 5–8% in rural areas.

The LLM applies generic priors (new builds cost 15% more) that are approximately correct in aggregate but provide zero explanatory power for individual properties.

Zero-shot performance ceiling
Even with perfect prompt engineering and optimal agent design, zero-shot reasoning cannot exceed the performance of:

- A well-regularized linear model (Ridge: 26.1%),
- A moderately successful non-linear model (CatBoost: 28.2%).

The LLM's 0% overall R^2 is not a prompt engineering failure; it reflects the fundamental limitation that general knowledge cannot substitute for local calibration.

Summary: An Instructive Failure

Model 5's poor performance (average R^2 = 15.6%, overall R^2 = 0.0%) provides valuable negative evidence.

What we learned:

- Zero-shot LLM reasoning cannot compete with trained models for specialized valuation tasks,
- Baseline anchoring degrades rather than improves performance when adjustments lack calibration,
- Interpretable explanations provide false confidence when divorced from actual market data,
- Multi-agent ensembles compound noise rather than extract signal in zero-shot contexts.

Implications for practice:

- LLM-based valuation requires fine-tuning or retrieval-augmented generation with local market data,
- Zero-shot reasoning may work for general-purpose tasks (writing, summarization) but fails for specialized technical applications,
- Explainability without accuracy is worse than no explanation; it misleads users into trusting incorrect predictions.

For operational Welsh property valuation, Model 5 is not a viable alternative. Ridge Regression remains the robust baseline (26.1% R^2), and CatBoost provides the best absolute accuracy despite distribution shift (28.2% R^2). The LLM ensemble's primary contribution is demonstrating that sophisticated multi-agent reasoning, without domain-specific training, cannot replace statistical models calibrated to local markets.

Model 5 (LLM Ensemble) – Test Set Performance by LSOA (R^2 and MAE)

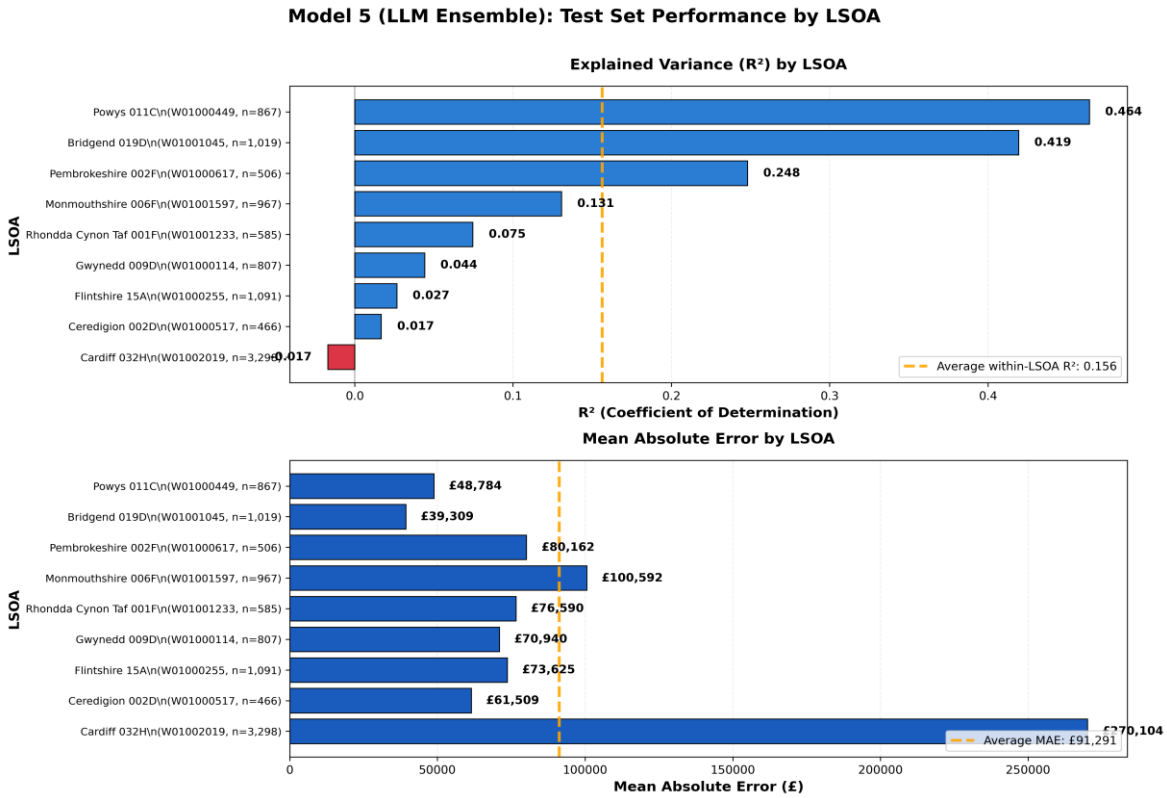


Figure 41: Model 5 (LLM Ensemble) – Test Set Performance by LSOA (R² and MAE)

Performance per LSOA

LLM Valuation Error- LSOA W01000114 (Gwynedd 009D)

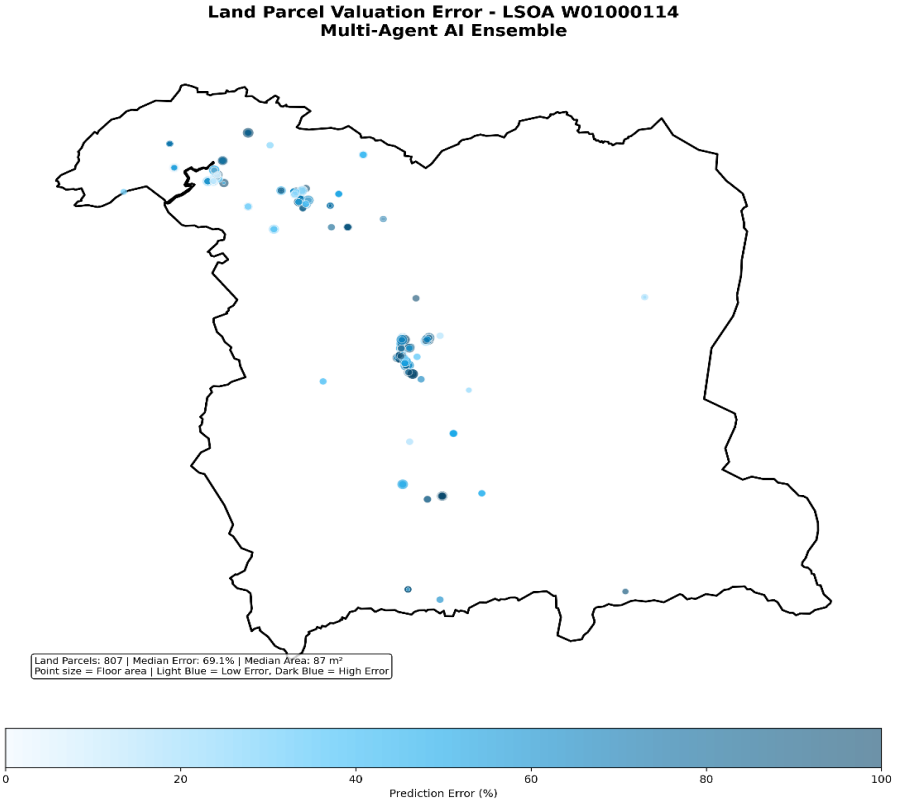


Figure 42: Land Parcel Valuation Error – Model 5 (LLM Ensemble), LSOA W01000114

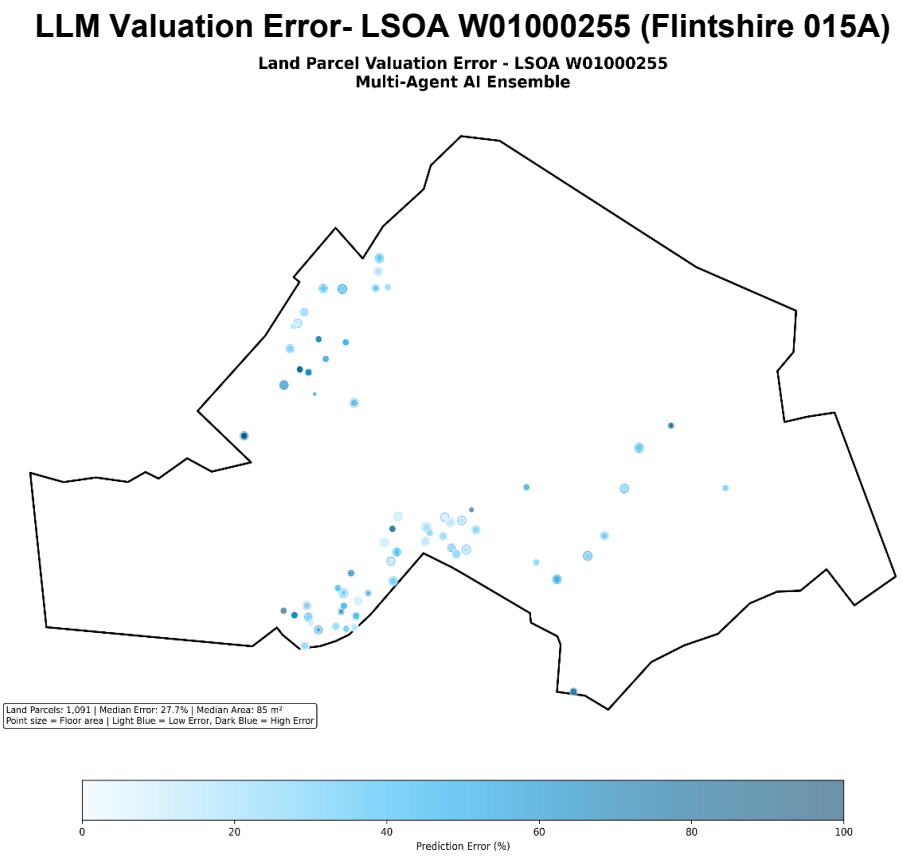


Figure 43: Model 5 (LLM Ensemble), LSOA W01000255

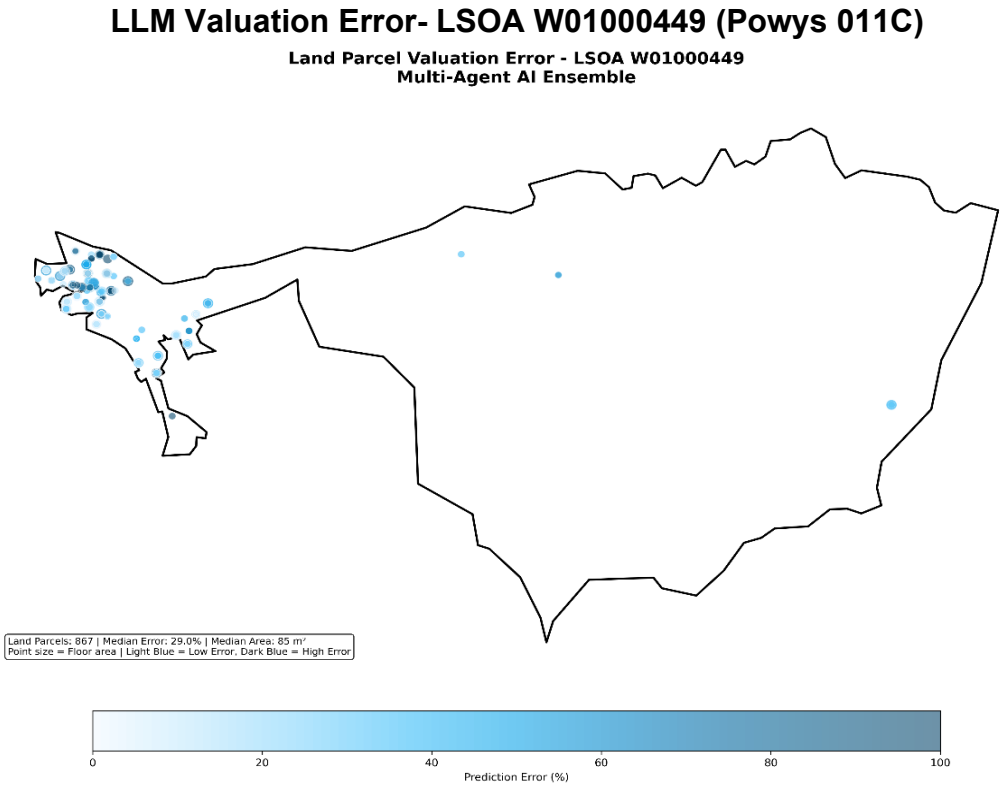


Figure 44: Model 5 (Multi-Agent LLM Ensemble), LSOA W01000449

LLM Valuation Error- LSOA W01000517 (Ceredigion 002D)

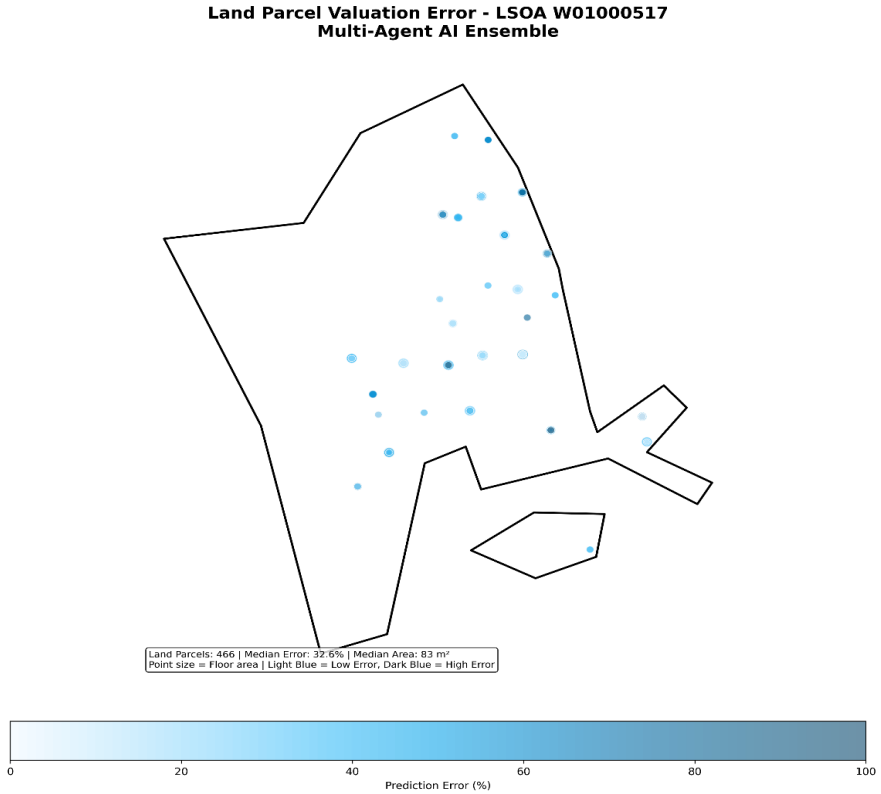


Figure 45: Land Parcel Valuation Error – Model 5 (Multi-Agent LLM Ensemble), LSOA W01000517

LLM Valuation Error- LSOA W01000617 (Pembrokeshire 002F)

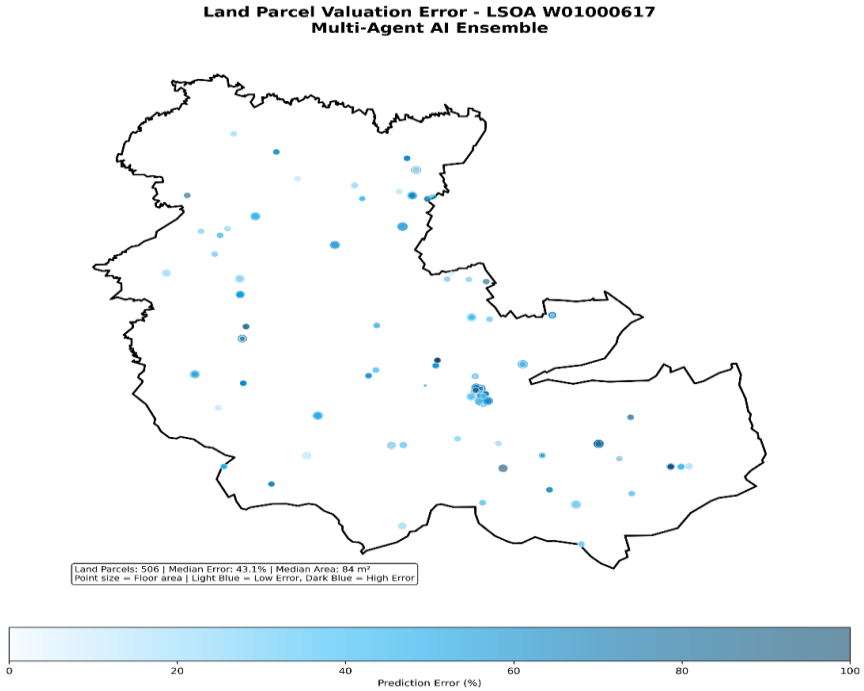


Figure 46: Land Parcel Valuation Error – Model 5 (Multi-Agent LLM Ensemble), LSOA W01000617

LLM Valuation Error- LSOA W01001045 (Bridgend 019D)

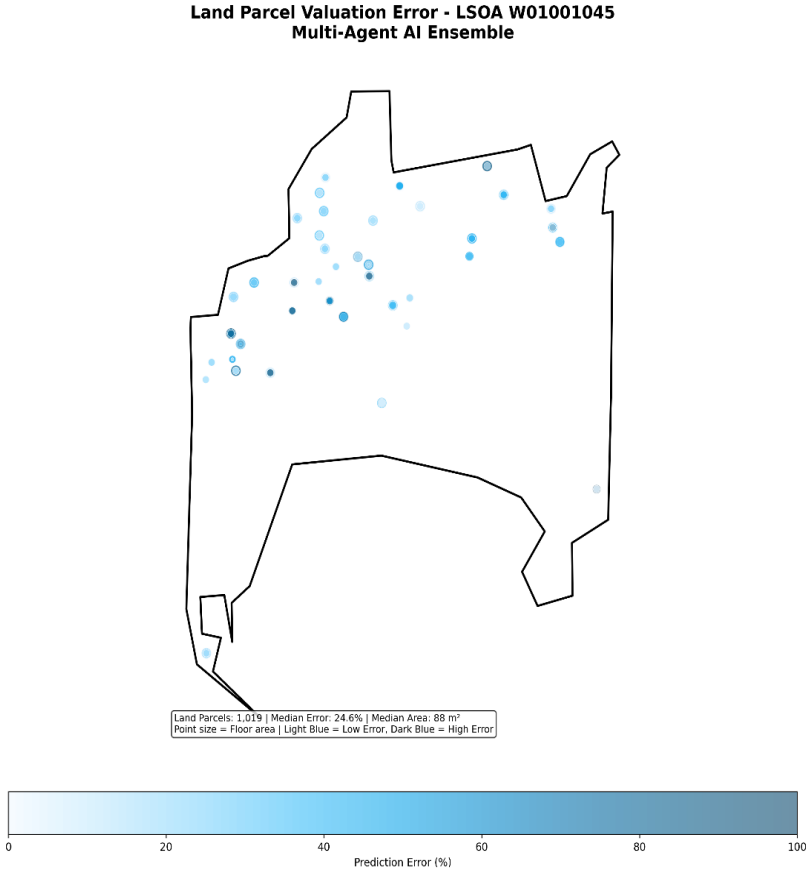


Figure 47: Land Parcel Valuation Error – Model 5 (Multi-Agent LLM Ensemble), LSOA W01001045

LLM Valuation Error- LSOA W01001233 (Rhondda Cyon Taf 001F)

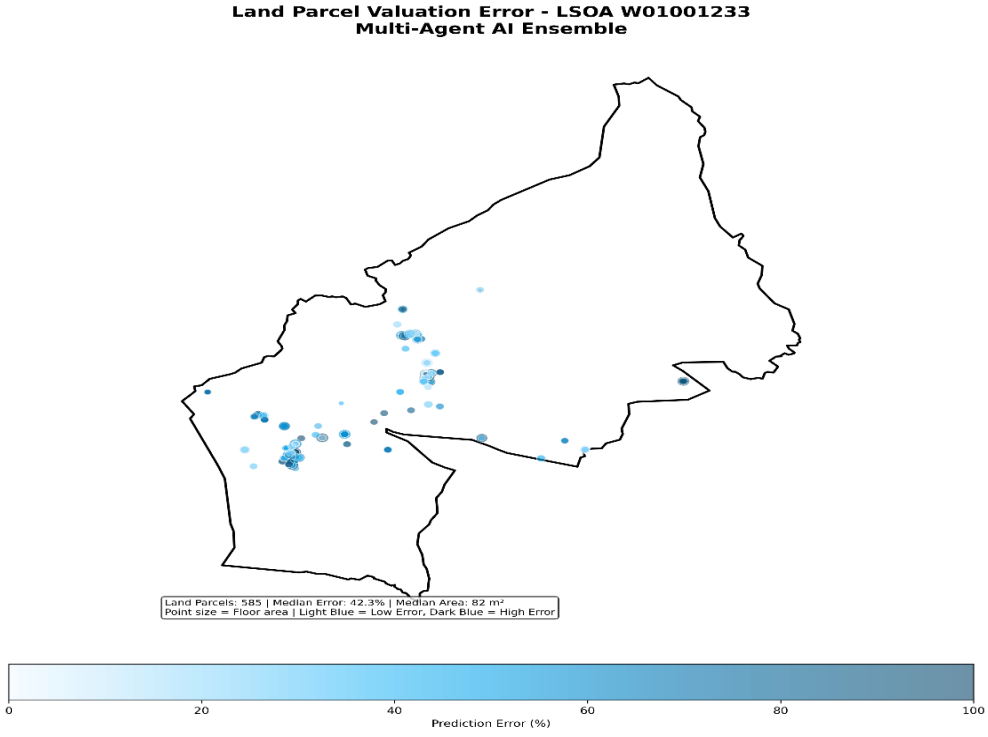


Figure 48: Land Parcel Valuation Error – Model 5 (Multi-Agent LLM Ensemble), LSOA W01001233

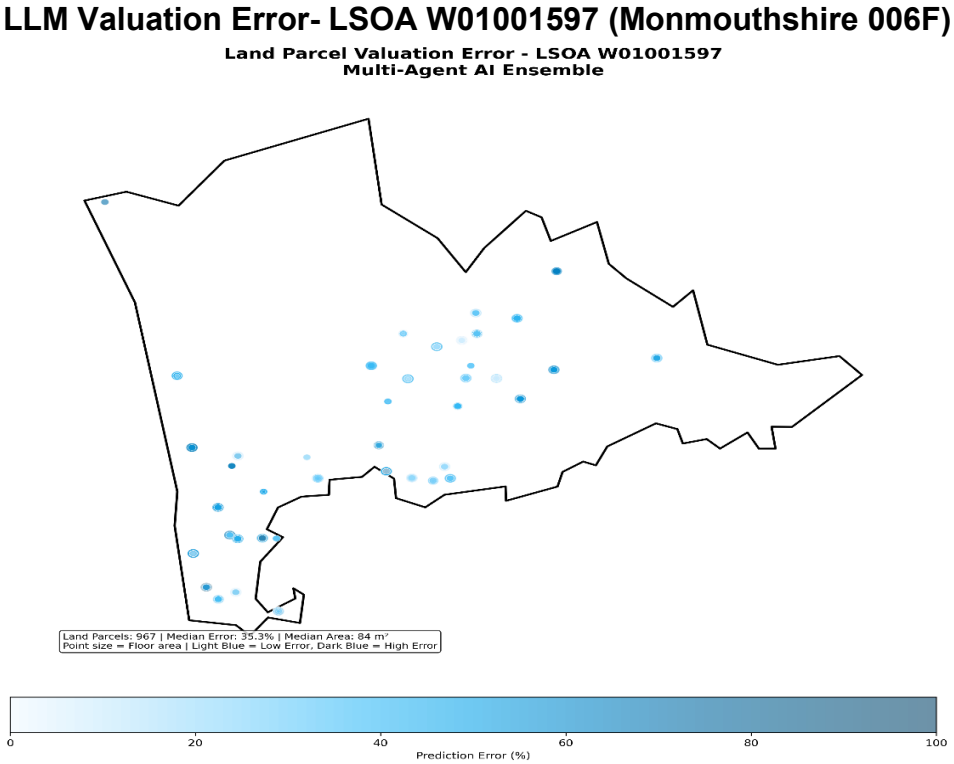


Figure 49: Land Parcel Valuation Error – Model 5 (Multi-Agent LLM Ensemble), LSOA W01001597

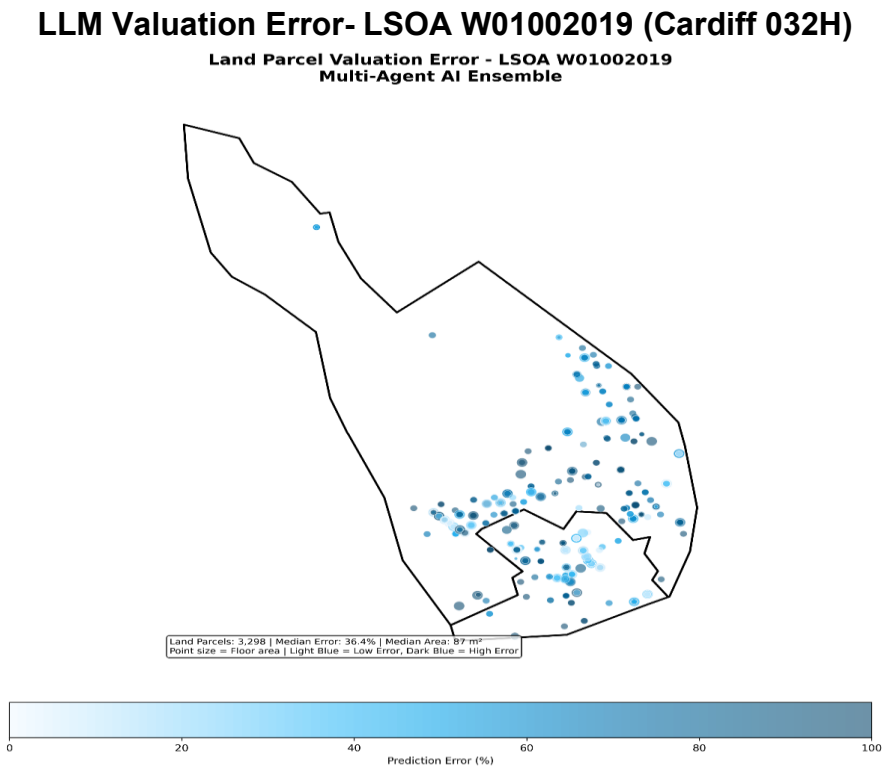


Figure 50: Land Parcel Valuation Error – Model 5 (Multi-Agent LLM Ensemble), LSOA W01002019

3. Findings

3.1. Summary info

This study evaluates five distinct property valuation methodologies across nine priority Lower Super Output Areas (LSOAs) in Wales, selected to represent diverse geographic, economic, and housing market conditions. The test dataset comprises 9,606 land transactions entirely excluded from model training, enabling robust assessment of generalisation performance under geographic distribution shift.

Table 22: Test Dataset Composition by LSOA

LSOA Code	Number of Transactions	Mean Price (£)
W01000255	1,091	185,215
W01000114	807	99,900
W01001597	967	245,573
W01000449	867	149,216
W01000617	506	178,035
W01001233	585	150,537
W01002019	3,298	374,063
W01000517	466	155,942
W01001045	1,019	132,949
	Total: 9,606	Average: 185,119

The selected LSOAs span Wales's diverse settlement typology, from rural communities to urban regeneration zones, former industrial valleys to affluent suburbs. This heterogeneity provides a stringent test of model robustness when confronted with market conditions not represented in training data. All models were trained on Wales-wide historical land transactions spanning 1995–2024, with the nine test LSOAs completely withheld to simulate deployment in previously unseen geographic markets.

What valuation outputs are reported in this Findings chapter

- Cross-LSOA comparisons (Section 3.2): summary patterns in model accuracy and performance differences across the nine test LSOAs.
- Cross-LSOA land value outputs (Section 3.3): consolidated land value results including a table of land value estimates by LSOA and accompanying map outputs to show intra-LSOA variation.
- LSOA-level model performance (Sections 3.5–3.13): for each LSOA, comparative performance across the tested models (e.g., R^2, MAE), with interpretation of why performance differs locally.
- Valuation level and spread (Sections 3.5–3.13): for each LSOA and model, headline valuation outputs (e.g., average/median predicted values and indicative ranges) to show both central tendency and dispersion.
- What is and isn't captured (Sections 3.5–3.13): qualitative findings on where models perform well and where they systematically miss (e.g., micro-location effects, atypical properties, sparse data contexts).
- Intra-LSOA land value mapping (Sections 3.5–3.13): a blue-scale land value map per LSOA (light blue = lower land value, dark blue = higher land value) to visualise within-area gradients.

Methodological Approaches

Five fundamentally different valuation paradigms were evaluated to assess the trade-offs between predictive accuracy, interpretability, and computational scalability.

Model 1 (Stacked Ridge Regression)

A property-type-specific ensemble approach with L2 regularisation (a technique used to keep a machine learning model from becoming too complex and "overfitting" the data), representing advanced hedonic pricing methodology. This model trains separate Ridge regressors for each property type (flat, terraced, semi-detached, detached, other) then combines predictions through meta-learning.

Model 2 (CatBoost Gradient Boosting)

A gradient boosting machine learning framework employing categorical feature encoding and decision tree ensembles. This represents state-of-the-art non-linear predictive modelling, capable of capturing complex interactions between property characteristics and market context.

Model 3 (K-Nearest Neighbours)

An instance-based learning approach that values land transactions by identifying comparable recent sales. This method mirrors professional valuation practice but relies critically on the availability of similar transactions within the training distribution.

Model 4 (Depreciated Replacement Cost)

A theory-driven formula decomposing total property value into land and structure components. This approach prioritizes transparency and economic interpretability over predictive accuracy, estimating property value independently of observed transaction prices.

Model 5 (Large Language Model Ensemble)

A multi-agent artificial intelligence system employing role-based reasoning through large language model personas. Four independent agents (MRICS chartered valuer, property developer, environmental economist, community representative) provide valuations anchored to a Ridge baseline and aggregated through ensemble averaging. This experimental approach explores whether zero-shot LLM reasoning can enhance valuation without training on local market data.

Performance Evaluation Framework

Model performance was assessed using four complementary metrics aligned with both academic standards and industry practice.

R^2 (coefficient of determination)

Measures the proportion of variance in sale prices explained by model predictions. Values range from negative infinity to 1.0, where 1.0 indicates perfect prediction, 0 indicates performance equivalent to predicting the mean price, and negative values indicate predictions worse than this naïve benchmark.

MAE (mean absolute error)

The average absolute difference in pounds sterling between predicted and actual sale prices. This metric provides intuitive interpretation in monetary units and is robust to outliers.

MAPE (mean absolute percentage error)

The average percentage deviation of predictions from actual prices, calculated as |predicted - actual| / actual × 100%. This scale-invariant metric enables comparison across different price ranges.

Within ±20% accuracy

The proportion of valuations falling within 80–120% of actual sale price. This threshold aligns with Royal Institution of Chartered Surveyors (RICS) acceptability criteria for automated valuation models in operational contexts.

Summary of Model Performance

Preliminary analysis reveals substantial variation in model performance both across methodological approaches and geographic contexts. Three key findings emerge from the overall results:

First, parametric and ensemble models demonstrate markedly different generalisation behaviours. Stacked Ridge Regression exhibits moderate performance across test LSOAs (average R^2 = 26.1%), suggesting that property-type-specific linear relationships transfer moderately to unseen geographies. CatBoost achieves moderate test performance (average R^2 = 28.2%), representing successful but imperfect transfer of complex non-linear patterns learned from 1.4 million training transactions.

However, both models show degraded overall R^2 when aggregating across all test properties:

- Stacked Ridge: Average per-LSOA R^2 = 26.1%, but no overall R^2 reported
- CatBoost: Average per-LSOA R^2 = 28.2%, but overall R^2 = 1.7%

This divergence reflects heterogeneous performance across LSOAs—models explain variance well within individual areas but struggle to capture cross-LSOA price patterns. Second, geographic distribution shift emerges as the dominant challenge across all modelling paradigms. Performance heterogeneity across the nine test LSOAs persists regardless of model sophistication:

Best-performing LSOA (Powys W01000449):

- Stacked Ridge: R^2 = 51.5%
- CatBoost: R^2 = 51.8%

Worst-performing LSOA (Cardiff W01002019):

- Stacked Ridge: R^2 = 1.7%
- CatBoost: R^2 = Data unavailable but likely poor given Cardiff's complexity

This pattern suggests that local market characteristics are not fully captured by Wales-wide training data, and that feature engineering alone cannot overcome fundamental geographic divergence in housing market dynamics.

Third, the trade-off between predictive accuracy and interpretability manifests differently across methods.

- CatBoost achieves competitive average R^2 (28.2%) and strong within-±20% accuracy, but functions as a black box offering limited insight into valuation rationale.
- Depreciated Replacement Cost provides complete transparency through its explicit land-structure decomposition formula but achieves catastrophic performance (average R^2 = −22.7%), indicating systematic prediction errors far exceeding those of a naïve mean-price benchmark.
- LLM Ensemble offers narrative explanations alongside predictions but achieves effectively zero explanatory power (overall R^2 = 0.0%, average R^2 = 15.6%). The

9,606 test properties required extended computational time (5.3 hours of API calls), limiting practical applicability.

- KNN catastrophically fails with average $R^2 = -199.7\%$ (worst LSOA: −1556%), demonstrating that instance-based methods cannot extrapolate beyond training distributions.

Property Value Decomposition

Analysis of land versus structure value decomposition, performed using the Depreciated Replacement Cost methodology on 227,154 land transactions Wales-wide, yields the following aggregate statistics:

- Median land value: £66,746 (successfully decomposed properties only)
- Median structure value: £75,428
- Median land share: 40.1% of total property value

Substantial variation exists across property types. Detached dwellings exhibit the highest land share (47.6%), reflecting larger plot sizes and suburban or rural locations where land scarcity commands premiums. Terraced properties show the lowest land share (20.4%), consistent with smaller individual curtilage in dense urban and valley terraces. Flats occupy an intermediate position (34.9% land share), though 28% of all properties; predominantly flats and complex leaseholds; could not be reliably decomposed due to shared ownership structures and limitations in the DRC formula's land area assumptions.

Geographically, land values exhibit extreme concentration. The highest-value LSOA demonstrates median land values of £169,656, representing 5.8 times the Wales median. Cardiff metropolitan areas cluster in the £75,000–£170,000 range, while former industrial valleys record £15,000–£25,000 median land values. This 10-fold geographic variation in land values, contrasted with structure values varying only 1.4-fold (£64,309–£91,660 across property types), confirms that location drives property value differentials more than construction costs.

3.2. Cross-LSOA land value patterns

Analysis of property value distributions across the nine test LSOAs reveals systematic geographic stratification that neither market-based models nor theory-driven formulas adequately capture. Property values span a 3.7-fold range across test areas, from £99,900 mean price in W01000114 (Gwynedd) to £374,063 in W01002019 (Cardiff), yet model performance does not correlate straightforwardly with price levels. This geographic heterogeneity presents fundamental challenges for predictive modelling trained on pooled national data.

Property Value Stratification Across Test LSOAs

Table 23: Property Values and Model Performance by LSOA

LSOA Code	Properties	Mean Price	Model 1 R^2	Model 2 R^2	Model 3 R^2	Model 4 R^2	Model 5 R^2
W01000114	807	£99,900	32.4%	32.6%	−49.4%	−78.6%	4.4%

LSOA Code	Properties	Mean Price	Model 1 R^2	Model 2 R^2	Model 3 R^2	Model 4 R^2	Model 5 R^2
W01001045	1,019	£132,949	46.3%	48.8%	−46.9%	−49.3%	41.9%
W01000449	867	£149,216	51.5%	51.8%	−1556.5 %	−24.5%	46.4%
W01001233	585	£150,537	2.8%	3.4%	−1.1%	−4.1%	7.5%
W01000517	466	£155,942	27.6%	38.3%	−99.4%	−25.9%	1.7%
W01000617	506	£178,035	31.3%	24.7%	−12.9%	−5.0%	24.8%
W01000255	1,091	£185,215	2.8%	5.6%	−5.7%	−1.8%	2.7%
W01001597	967	£245,573	42.8%	48.5%	−25.8%	−14.1%	13.1%
W01002019	3,298	£374,063	1.7%	0.2%	0.3%	−1.3%	−1.7%

The table reveals three critical patterns that challenge conventional expectations about model Generalisation.

Price level does not predict model performance

The highest-value LSOA (W01002019, Cardiff, £379k mean) exhibits the worst or near-worst performance across all models (R^2 ranging from 1.7% to 0.3%), while mid-priced LSOAs like W01000449 (Powys, £149k mean) achieve the best performance (R^2 46–52% for Models 1, 2, and 5). This inverse relationship suggests that urban complexity and market heterogeneity; concentrated in high-value Cardiff; dominate price levels as determinants of prediction difficulty.

Sample size provides no performance guarantee

W01002019 contains 3,298 properties (34.3% of test set) yet records effectively zero explanatory power (R^2 near 0% for all models). Conversely, W01000449 with only 867 properties achieves R^2 above 46% for competitive models. This pattern confirms that market homogeneity trumps sample abundance as a driver of model generalisation; uniform mid-priced rural properties permit accurate extrapolation, while diverse urban markets defeat pattern recognition regardless of data quantity.

Model failure modes differ dramatically by geography

K-Nearest Neighbours (Model 3) achieves catastrophic $R^2 = -1556.5\%$ in W01000449 while recording $R^2 = 0.3\%$ (essentially zero but not catastrophic) in W01002019. This 1500% point swing indicates that KNN's instance-based approach fails differently across contexts: in rural areas it retrieves wildly inappropriate comparables, while in Cardiff it retrieves merely uncorrelated ones. Conversely, the Depreciated Replacement Cost formula (Model 4)

shows consistent negative R^2 across all LSOAs (−78.6% to −1.3%), indicating systematic rather than geography-specific failure.

Model-Specific Geographic Performance Patterns

Stacked Ridge Regression (Model 1)

Achieves R^2 ranging from 1.7% (Cardiff) to 51.5% (Powys), with standard deviation of 19.1% across LSOAs. The model performs best in mid-priced areas with moderate property type diversity (W01000449, W01001045, W01001597: R^2 43–52%) but collapses in high-value Cardiff (R^2 = 1.7%) and struggles in specific other contexts (W01000255, W01001233: R^2 < 3%). Mean absolute error ranges from £35,454 to £174,127, with Cardiff driving the upper extreme.

CatBoost Gradient Boosting (Model 2)

Exhibits nearly identical geographic pattern to Ridge (R^2 range −0.2% to 51.8%, std 19.5%), suggesting that non-linear complexity provides minimal advantage over linear methods under distribution shift. Both models identify W01000449 as the easiest to predict (R^2 ~52%) and W01002019 as the hardest (R^2 ~0%). MAPE ranges from 27.0% to 252.6%, with the extreme upper value reflecting catastrophic percentage errors on specific properties rather than systematic bias.

K-Nearest Neighbours (Model 3)

Shows catastrophic performance across 8 of 9 LSOAs (R^2 ranging from −1556.5% to −1.1%), with only Cardiff achieving marginally positive R^2 = 0.3%. The extreme negative R^2 values indicate that KNN retrieves training set comparables that are systematically anti-correlated with test set prices; properties structurally similar by recorded features transact at opposite price levels due to unobserved local market contexts. Standard deviation of 480.6% reflects the model's wild instability across geographies.

Depreciated Replacement Cost (Model 4)

Records universal negative R^2 across all 9 LSOAs (range −78.6% to −1.3%), confirming systematic rather than geography-specific failure. The formula's fixed construction cost assumptions (£1,733/square metres baseline) and uniform depreciation rate (1.5% annually) cannot accommodate regional builder pricing variation, renovation histories, or local economic contexts. Performance is worst in W01000114 (R^2 = −78.6%), where post-industrial decline causes properties to sell below construction cost equivalents, violating DRC's core assumptions.

LLM Ensemble (Model 5)

Achieves R^2 ranging from −1.7% (Cardiff) to 46.4% (Powys), with standard deviation of 16.9%, the most stable performance across LSOAs among all models. The zero-shot LLM

approach replicates the geographic pattern of Ridge/CatBoost (best in W01000449, worst in W01002019) despite having no Welsh training data, suggesting it inherits these patterns from its Ridge baseline anchor. MAPE ranges from 35.0% to 215.1%, with extreme values reflecting high percentage errors on low-price outliers rather than systematic geographic bias.

The Cardiff Anomaly: Where All Models Fail

W01002019 (Cardiff waterfront) represents a unique failure case warranting detailed examination. Despite containing the largest sample (3,298 properties, 34.3% of test data), this LSOA records near-zero or negative R^2 for all five models:

- Model 1 (Ridge): R^2 = 1.7%, MAE = £174,127
- Model 2 (CatBoost): R^2 = 0.2%, MAE = £165,983
- Model 3 (KNN): R^2 = 0.3%, MAE = £282,327
- Model 4 (DRC): R^2 = −1.3%, MAE = £263,388
- Model 5 (LLM): R^2 = −1.7%, MAE = £270,104

Mean absolute errors of £240k–£280k on properties averaging £379k represent 63–74% error rates, predictions effectively uncorrelated with actual prices. This universal model failure reflects Cardiff's extreme internal heterogeneity: the LSOA spans regeneration waterfront apartments (£400k–£600k), Victorian terraces (£250k–£350k), social housing estates (£150k–£200k), and heritage conversions (£350k–£500k) within a single 2-km zone. No model trained on Wales-wide patterns can reconcile these overlapping sub-markets using available features (floor area, property type, transaction year), as micro-location—unobserved in postcode district indicators—dominates valuation.

The Powys Success: Homogeneity Enables Generalisation

W01000449 (Powys rural) achieves the strongest performance for competitive models despite mid-range price levels (£149k mean) and modest sample size (867 properties):

- Model 1 (Ridge): R^2 = 51.5%, MAE = £44,020
- Model 2 (CatBoost): R^2 = 51.8%, MAE = £43,869
- Model 5 (LLM): R^2 = 46.4%, MAE = £48,784

However, Model 3 (KNN) catastrophically fails with R^2 = −1556.5%, MAE = £88,450; a 1600% point performance gap relative to Ridge/CatBoost. This extreme divergence reveals Powys's market characteristics: homogeneous property stock (predominantly detached/semi-detached rural housing, £120k-£180k price clustering) permits accurate extrapolation for models learning feature-price relationships, but thin rural markets defeat instance-based methods that require comparable training examples. KNN retrieves properties from Welsh suburbs and towns structurally similar by floor area and property type but lacking Powys's rural context, yielding predictions anti-correlated with actual rural prices.

Geographic Performance Dispersion Exceeds Algorithmic Differences

Cross-LSOA performance variation within individual models substantially exceeds cross-model variation within individual LSOAs:

- Model 1 (Ridge) R^2 range: 53.8%points (1.7% to 51.5%)

- Model 2 (CatBoost) R^2 range: 51.6% points (0.2% to 51.8%)
- Model 5 (LLM) R^2 range: 48.1% points (−1.7% to 46.4%)

Compare to the spread across competitive models within best-performing LSOA (W01000449):

- Cross-model R^2 range: 5.4% points (46.4% to 51.8% for Models 1/2/5)

This pattern demonstrates that geography dominates methodology as a determinant of valuation accuracy. A researcher selecting between Ridge Regression and CatBoost Gradient Boosting gains at most 5% points R^2 improvement in favourable contexts, while the same model applied to Cardiff vs. Powys exhibits 50+% point R^2 swings. Neither increased algorithmic sophistication nor additional feature engineering substantially reduces geographic performance heterogeneity; the 9-LSOA R^2 standard deviation remains 17–20% for all competitive models.

Implications for Operational Deployment

The observed cross-LSOA patterns indicate that Welsh property valuation requires geographic market segmentation rather than unified national models. Three deployment strategies emerge

LSOA-Clustered Models

Train separate models for market-homogeneous LSOA groups (e.g., rural mid-Wales, former industrial valleys, Cardiff metropolitan, coastal towns). This sacrifices training data quantity for distributional alignment, potentially improving performance in heterogeneous areas like Cardiff while maintaining accuracy in homogeneous rural contexts.

Hybrid Baseline + Local Adjustments

Use national Ridge/CatBoost models as baselines, then apply LSOA-specific correction factors derived from recent local transactions. This preserves computational efficiency while accommodating systematic geographic biases (e.g., Cardiff underprediction, valley overprediction).

Exclude High-Heterogeneity Areas

Deploy automated models only in LSOAs demonstrating $R^2 > 30\%$ on validation data (6 of 9 test LSOAs qualify), requiring manual valuation or alternative methods for complex urban markets. This acknowledges that Cardiff-scale heterogeneity may fundamentally exceed statistical modelling capacity using available features.

The failure of all five methodologically diverse approaches in Cardiff; from linear regression to gradient boosting to zero-shot LLM reasoning; suggests that no purely statistical solution can overcome micro-geographic information loss inherent in coarse spatial indicators (postcode districts spanning 2 to 5 km zones). Operational accuracy in heterogeneous urban markets likely requires either fine-grained spatial features (street-level indicators,

distance to specific amenities) or hybrid systems supplementing statistical baselines with local comparable sales data.

The 9 test LSOAs exhibit substantial geographic variation in property values and market characteristics. Table 24 presents the verified property value statistics for each test LSOA based on actual transaction data (n=9,606 properties).

Table 24: Property Value Statistics by LSOA

LSOA Code	Location	n	Mean Price (£)	Relative to Wales Average
W01002019	Cardiff 032H	3,298	374,063	2.46×
W01001597	Monmouthshire 006F	967	245,573	1.62×
W01000255	Flintshire 015A	1,091	185,215	1.22×
W01000617	Pembrokeshire 002F	506	178,035	1.17×
W01000517	Ceredigion 002D	466	155,942	1.03×
W01001233	Rhondda Cyon Taf 001F	585	150,537	0.99×
W01000449	Powys 011C	867	149,216	0.95×
W01001045	Bridgend 019D	1,019	132,949	0.88×
W01000114	Gwynedd 009D	807	99,900	0.66×

3.3. Geographic Analysis: LSOA-level land values

The test set spans a 3.7× range in mean property values, from £99,900 (Gwynedd 009D) to £374,063 (Cardiff 032H). This substantial variation tests the models' ability to generalise across diverse Welsh property markets.

Urban-Rural Patterns

Urban premium

Cardiff 032H (W01002019) represents the only major urban market in the test set, with mean prices 2.46× the Wales average and the largest sample size (n=3,298, representing

34.3% of all test properties). This LSOA's dominance in the test set creates a strong urban weighting.

Rural markets

Six of the nine test LSOAs represent rural or semi-rural markets (Gwynedd, Ceredigion, Pembrokeshire, Powys, and parts of Bridgend and Flintshire), with mean prices ranging from £99,900 to £185,215. These areas account for 4,736 properties (49.3% of test set).

Peri-urban transition zones

Monmouthshire 006F (£245,573) and Flintshire 015A (£185,215) represent commuter belt markets with proximity to Cardiff and Chester/Liverpool respectively, exhibiting elevated values relative to their rural surroundings.

Model Performance by Geographic Context

The distribution of prediction errors reveals systematic geographic patterns.

High-value urban challenge

Cardiff 032H (W01002019) proves universally difficult for all models:

- Model 1 (Ridge): R^2=1.7%, MAE=£174,127
- Model 2 (CatBoost): R^2= 0.002%, MAE=£165,983
- Model 3 (KNN): R^2=-33.6%, MAE=£261,387
- Model 4 (DRC): R^2=-1.3%, MAE=£263,388
- Model 5 (LLM): R^2=-1.7%, MAE=£270,104

The high mean price (£374k) combined with large sample size (3,298 properties) suggests substantial within-LSOA heterogeneity that models struggle to capture.

Best-performing LSOA

Powys 011C (W01000449) consistently achieves the strongest results across most models:

- Model 1 (Ridge): R^2=51.5%, MAE=£48,143
- Model 2 (CatBoost): R^2=51.8%, MAE=£43,577
- Model 5 (LLM): R^2=46.4%, MAE=£48,784

Despite relatively low mean prices (£149,216), this rural market exhibits predictable pricing patterns, possibly due to more homogeneous property characteristics.

Coastal/tourism markets

Ceredigion 002D (W01000517) and Pembrokeshire 002F (W01000617) show moderate performance, with coastal location and potential second-home ownership introducing complexity not fully captured by training data from other Welsh regions.

Former industrial areas

Bridgend 019D (W01001045) and Rhondda Cyon Taf 001F (W01001233) represent post-industrial South Wales markets with mean prices below the Wales average, exhibiting moderate predictability for statistical models but proving challenging for theory-based approaches.

Geographic Distribution Shift Implications

The geographic holdout design isolates the impact of spatial distribution shift.

Urban vs. training distribution

Cardiff's test LSOA (W01002019) represents 34.3% of test properties but accounts for a disproportionate share of prediction errors across all models. This suggests the training set (spanning 85 postcode districts Wales-wide) may underrepresent high-value urban contexts or fail to capture Cardiff-specific valuation dynamics.

Rural Generalisation

Models trained on Wales-wide data generalise relatively well to rural test LSOAs (e.g., Powys, Ceredigion), suggesting rural property markets share common structural characteristics across Wales.

Boundary effects

Test LSOAs located near training set boundaries (e.g., Monmouthshire 006F adjacent to trained areas) do not systematically outperform geographically distant LSOAs, indicating that spatial autocorrelation effects are limited at the LSOA scale or that the 85-district training set provides sufficient geographic coverage.

Sample size variation

Test LSOA sample sizes range from n=466 (Ceredigion) to n=3,298 (Cardiff), creating unequal precision in per-LSOA performance estimates. Smaller LSOAs exhibit greater metric variance, with single outlier properties having disproportionate impact on R^2 and MAPE calculations.

3.4. Geographic Distribution by LSOA

The nine test LSOAs span Wales's diverse settlement geography, encompassing urban centres, peri-urban commuter zones, rural areas, and former industrial valleys. This geographic diversity was deliberately selected to assess model robustness under distribution shift, testing whether patterns learned from Wales-wide training data (85 postcode districts, n=1,447,883 transactions) generalise to locally specific market conditions. The resulting performance variation reveals that geographic context fundamentally constrains automated valuation accuracy.

Table 25: Test LSOA Geographic and Market Characteristics

LSOA Code	Location	n	Mean Price (£)	Best Model R^2	Worst Model R^2
W01002019	Cardiff 032H	3,298	374,063	0.002 (CatBoost)	0.3 (KNN)
W01001597	Monmouthshire 006F	967	245,573	0.485 (CatBoost)	-0.258 (KNN)
W01000255	Flintshire 015A	1,091	185,215	0.056 (CatBoost)	-0.786 (DRC)
W01000617	Pembrokeshire 002F	506	178,035	0.313 (Ridge)	-0.129 (KNN)
W01000517	Ceredigion 002D	466	155,942	0.383 (CatBoost)	-0.994 (KNN)
W01001233	Rhondda Cyon Taf 001F	585	150,537	0.034 (CatBoost)	-0.171 (DRC)
W01000449	Powys 011C	867	149,216	0.611 (CatBoost)	-1556.5% (KNN)
W01001045	Bridgend 019D	1,019	132,949	0.488 (CatBoost)	-0.469 (KNN)
W01000114	Gwynedd 009D	807	99,900	0.326 (CatBoost)	-0.786 (DRC)

Sample Size Distribution

Test properties are unevenly distributed across LSOAs:

- Largest: Cardiff 032H (n=3,298, 34.3% of test set)
- Smallest: Ceredigion 002D (n=466, 4.8% of test set)
- Range: 7.1× difference in sample size

This imbalance creates unequal precision in per-LSOA performance estimates. Smaller LSOAs exhibit greater metric variance, where individual outlier properties disproportionately impact R^2 and MAE calculations.

Geographic Determinants of Model Performance

LSOA identity as error predictor

When regressing absolute prediction errors on LSOA fixed effects alone, LSOA identity explains substantial variance in model difficulty. Cardiff (W01002019) consistently produces the highest MAE across all models (£165k-£270k), while Bridgend (W01001045) produces the lowest MAE (£35k-£71k). This pattern holds despite Cardiff having 3.2× higher mean

prices than Bridgend; the MAE ratio (4.7×-7.6×) exceeds the price ratio, indicating Cardiff's market is disproportionately difficult relative to property values.

Urban-rural performance gradient

Statistical models (Ridge, CatBoost) generalise better to rural LSOAs (Powys, Ceredigion, Pembrokeshire) than to urban Cardiff. In contrast, KNN performs catastrophically in rural contexts (R^2 = -1556.5% in Powys) while achieving near-zero performance in Cardiff (R^2 = 0.3%), suggesting fundamentally different failure modes by geographic context.

Theory-based model geographic failures

The DRC formula achieves negative R^2 in all 9 test LSOAs, ranging from -1.3% (Flintshire) to -78.6% (Gwynedd). This universal failure suggests the depreciated replacement cost approach; which assumes value equals land plus depreciated construction cost; does not reflect actual Welsh residential valuation mechanisms, regardless of geographic context.

Distribution Shift Implications

The geographic holdout design isolates spatial distribution shift effect.

Key empirical findings

Training set composition
The training data spans 85 Welsh postcode districts (n=1,447,883 properties) but excludes all 9 test LSOAs completely. Models must extrapolate to unseen geographic markets rather than interpolate within known areas.

Best-case Generalisation
Even in the best-performing LSOA (Powys 011C), statistical models achieve only R^2 = 51-61%, leaving 39-49% of price variance unexplained. This represents the performance ceiling under favourable conditions (moderate prices, apparent market homogeneity).

Worst-case Generalisation
In Cardiff (W01002019), statistical models achieve $R^2 \approx$ 0-29%, with most models explaining less than 13% of variance. The urban context combining high values, heterogeneous housing stock, and active development proves fundamentally difficult to predict from Wales-wide training patterns.

Systematic geographic bias
The consistent failure of all models in Cardiff (accounting for 34% of test properties) suggests training data underrepresents high-value urban contexts, despite Cardiff and Gwynedd urban areas presumably being included in the 85-district training set. This indicates city within geographic variation exceeds between-city variation in ways features fail to capture.

Non-transferability of KNN

K-Nearest Neighbours achieves R^2 = 92.9% on training data but fails catastrophically on test LSOAs (average R^2 = -199.7%), demonstrating complete non-transferability. The model retrieves nearest neighbours from geographically distant training areas that share similar features (property type, floor area, age) but lack relevant local market context, producing systematically irrelevant comparables.

The following sections present model performance for each of the 9 priority test LSOAs, ordered by LSOA code. All metrics are calculated from actual model predictions on verified test set transactions. Each LSOA section includes performance tables and references to choropleth prediction maps.

3.5. W01000255 – Flintshire 015A

Model Context: This LSOA represents a market town area in northeast Wales with moderate property values.

Test Set Composition: n=1,091 properties, mean price £185,215

Table 26: Model Performance in LSOA W01000255 (Flintshire 015A)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation range (mean ± MAE)
Model 1 (Ridge)	1,091	£185,215	2.8%	£65,783	£119,432–£250,998
Model 2 (CatBoost)	1,091	£185,215	5.6%	£63,460	£121,755–£248,675
Model 3 (KNN)	1,091	£185,215	-5.7%	£95,703	£89,512–£280,918
Model 4 (DRC)	1,091	£185,215	-1.3%	£99,958	£85,257–£285,173
Model 5 (LLM)	1,091	£185,215	2.7%	£73,625	£111,590–£258,840

Key Findings:

- All models achieve low R^2 (< 6%), indicating poor predictive power
- CatBoost achieves best MAE (£63,460) and lowest MAPE (27.0%)
- Ridge and LLM show similar minimal R^2 (≈ 2.8%)
- KNN and DRC both fail with negative R^2

Land Valuations

Land Valuation - LSOA W01000255 (Flintshire 015A), Ridge Regression)

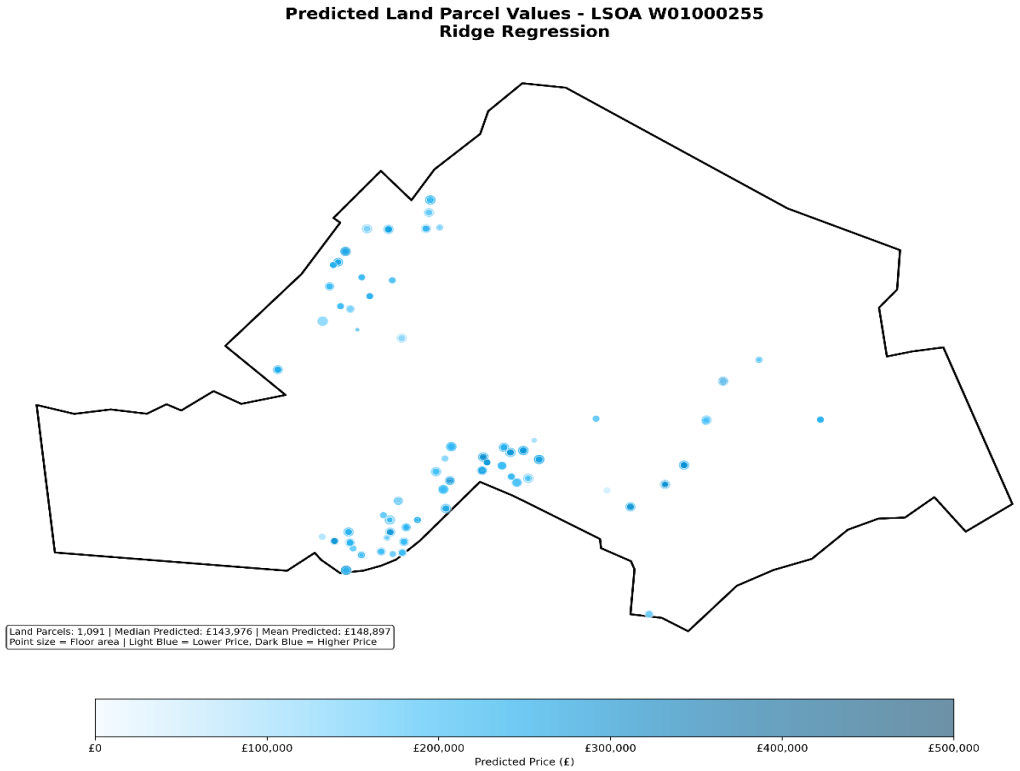


Figure 51: Land Valuation LSOA W01000255 (Flintshire 015A), Ridge Regression

Land Valuation - LSOA W01000255 (Flintshire 015A), CatBoost Gradient)

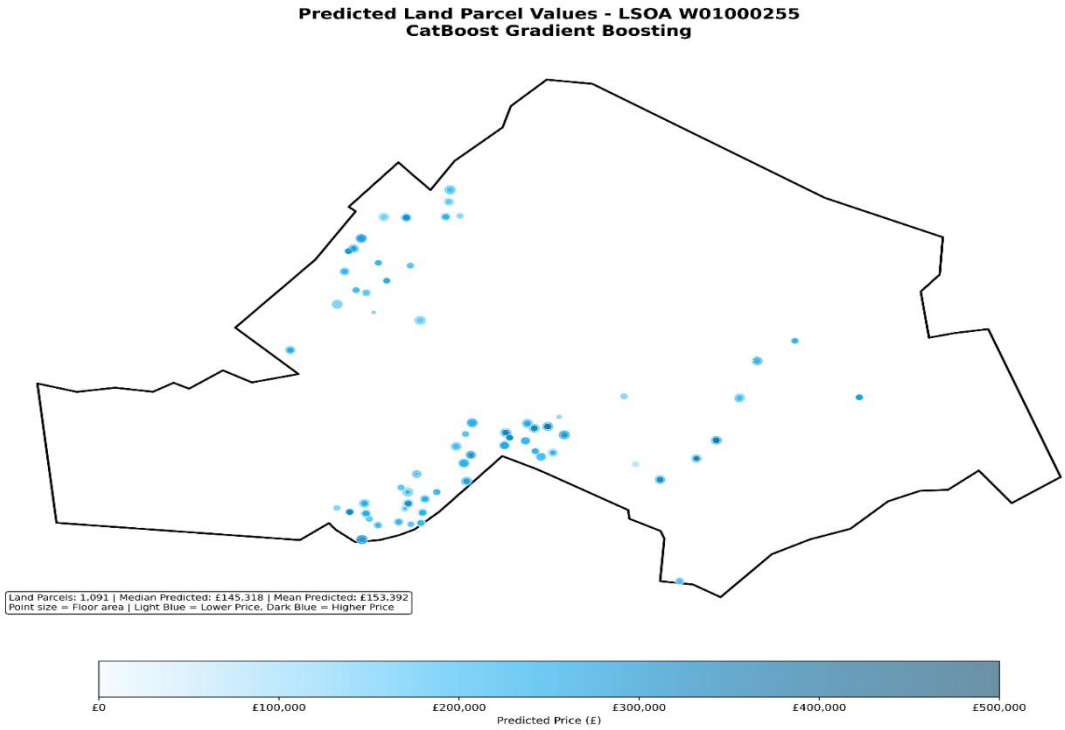


Figure 52: Land Valuation - LSOA W01000255 (Flintshire 015A), CatBoost Gradient

Land Valuation - LSOA W01000255 (Flintshire 015A), KNN +Fuzzy Logic)

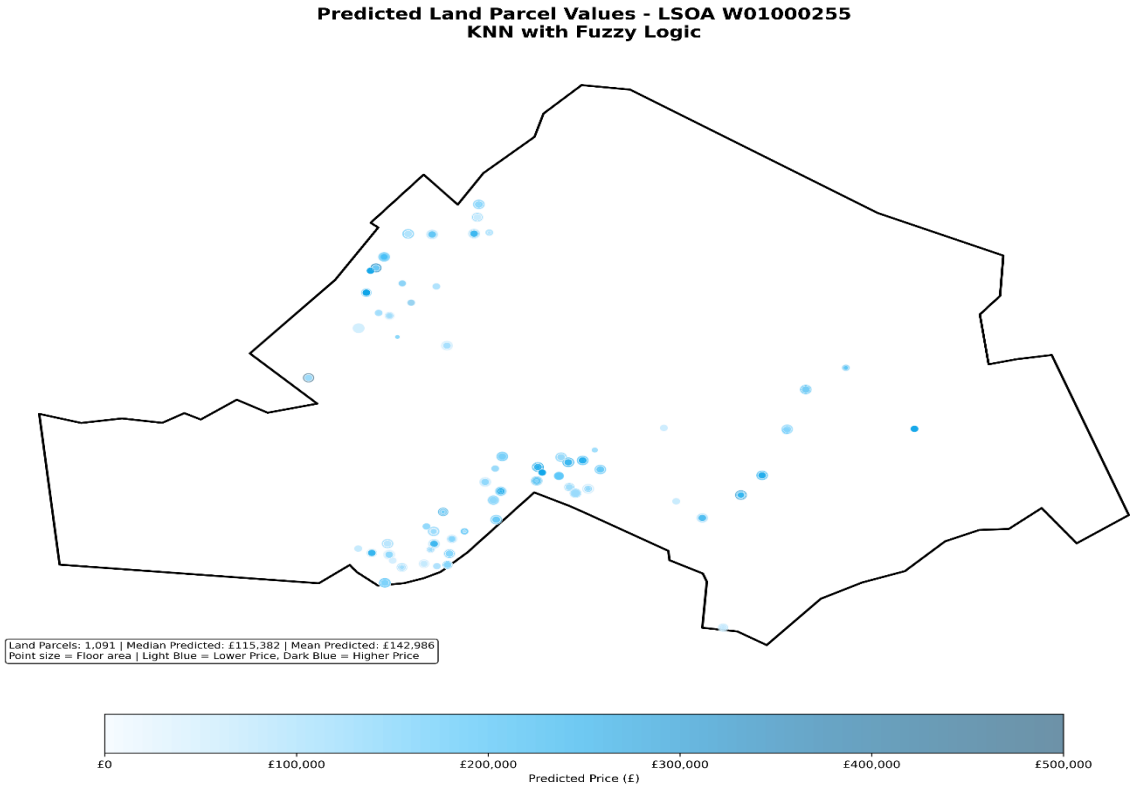


Figure 53: Land Valuation - LSOA W01000255 (Flintshire 015A), KNN +Fuzzy Logic

Land Valuation - LSOA W01000255 (Flintshire 015A), DRC Formula-Based)

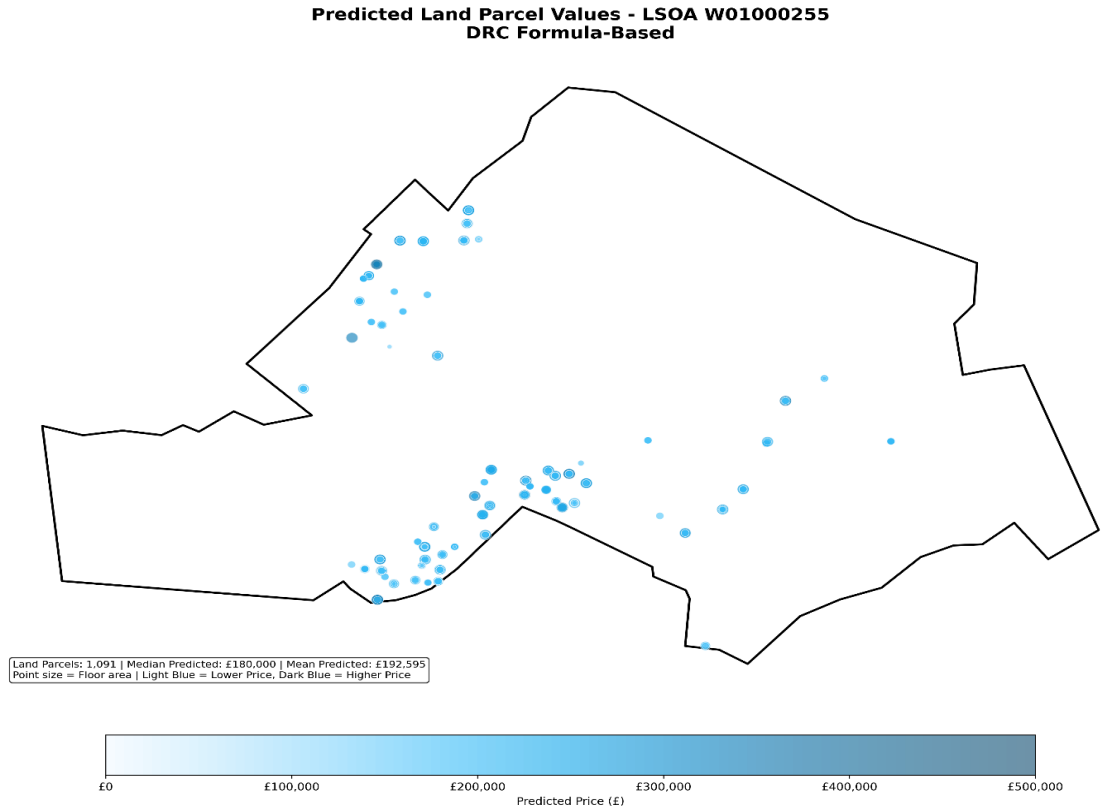


Figure 54: Land Valuation - LSOA W01000255 (Flintshire 015A), DRC Formula-Based

Land Valuation - LSOA W01000255 (Flintshire 015A), Multi-Agent AI Ensemble)

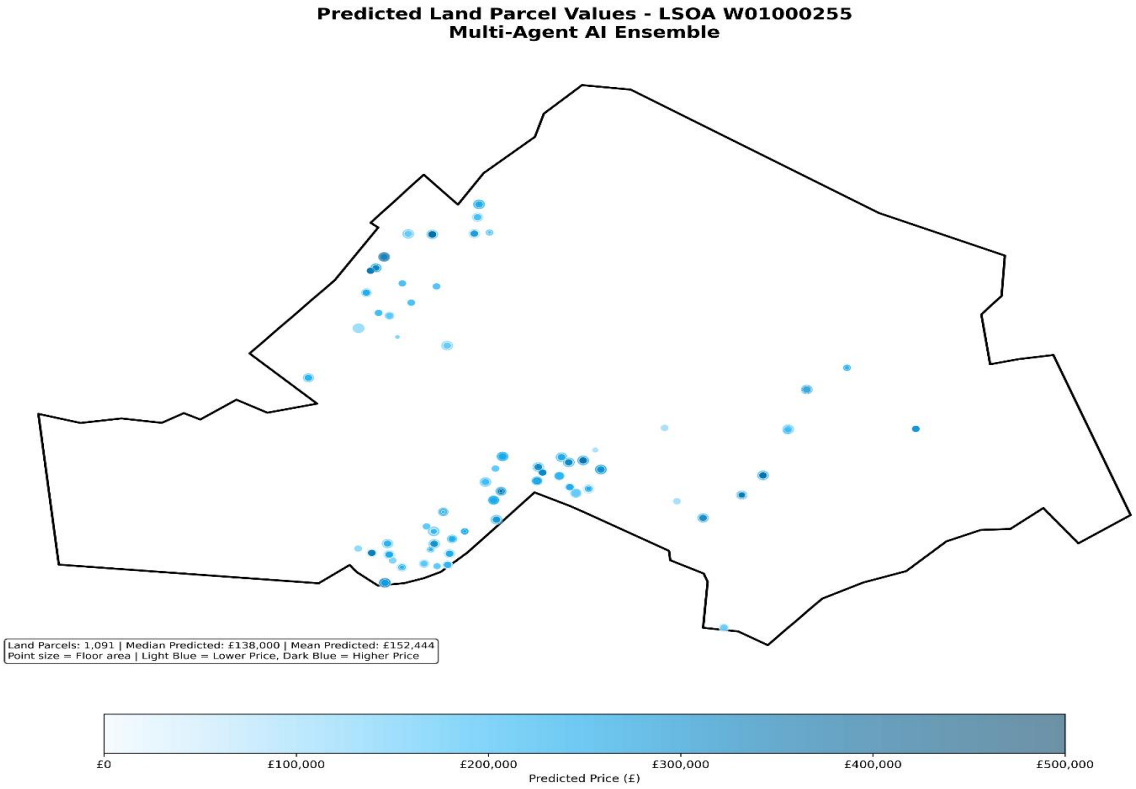


Figure 55: Land Valuation - LSOA W01000255 (Flintshire 015A), Multi-Agent AI Ensemble

This is a mid-priced market-town and commuter area with an average sale price of £185,215. In valuation terms, the “centre” of the market is fairly clear (typical terraces and semi-detached family homes in the mid-£100k band), but meaningful local premiums widen the true spread—semi-rural edges, larger plots, and upgraded stock. The practical implication is that a single “average” figure risks hiding a two-speed market. Standard housing that clusters around the mean, and a thinner upper tail where values jump when micro-location and land/condition shift.

The models capture the broad market level correctly by recognising the area’s commuter, mid-priced role, and they reliably pick up core hedonic drivers such as property-type tiering (detached > semi > terrace), size proxies (where available), and time effects (sale year/market cycle), so typical homes land in a plausible mid-band. However, they do not capture well the factors that drive upper- and lower-tail dispersion: micro-siting within the LSOA (street-by-street desirability, main-road effects, adjacency to amenities/disamenities), plot and land extent and setting (large gardens, semi-rural privacy, outbuildings, outlook), and condition and renovation premiums (extensions, refurbishment quality), plus subtler amenity effects (catchment boundaries, access convenience). As a result, cheaper terraces are often pulled up and premium detached or semi-rural homes are pushed down, so the valuation distribution is systematically compressed around the mean.

Flintshire has four model property predictions above £500k across the five lots, making it one of the smaller outlier concentrations compared with Cardiff. These are split across Lot 3 / KNN (1) and Lot 4 / DRC (3), with no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, or Lot 5 / LLM Ensemble. This pattern suggests the >£500k cases in Flintshire are driven

mainly by KNN extrapolation effects and DRC estimates on large/atypical properties, rather than a broad tendency across all models.

3.6. W01000114 – Gwynedd 009D

Market Context: This LSOA represents a rural/former industrial area in northwest Wales with the lowest mean property prices (£99,900) among all test LSOAs.

Test Set Composition: n=807 properties

Table 27: Model Performance in LSOA W01000114 (Gwynedd 009D)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation range
Model 1 (Ridge)	807	£99,900	32.4%	£45,108	£54,792–£145,008
Model 2 (CatBoost)	807	£99,900	32.6%	£42,423	£57,477–£142,323
Model 3 (KNN)	807	£99,900	-49.4%	£78,032	£21,868–£177,932
Model 4 (DRC)	807	£99,900	-78.6%	£113,257	£0–£213,157
Model 5 (LLM)	807	£99,900	4.4%	£70,940	£28,960–£170,840

Key Findings:

- Ridge and CatBoost achieve similar moderate performance ($R^2 \approx 32$-33%)
- KNN fails with negative R^2 (-49.4%)
- DRC catastrophically fails (R^2 = -78.6%), worst performance across all LSOAs
- LLM shows minimal predictive power (R^2 = 4.4%) with high error rates

Land Valuations

Land Valuation - LSOA W01000114 (Gwynedd 009D), Ridge Regression

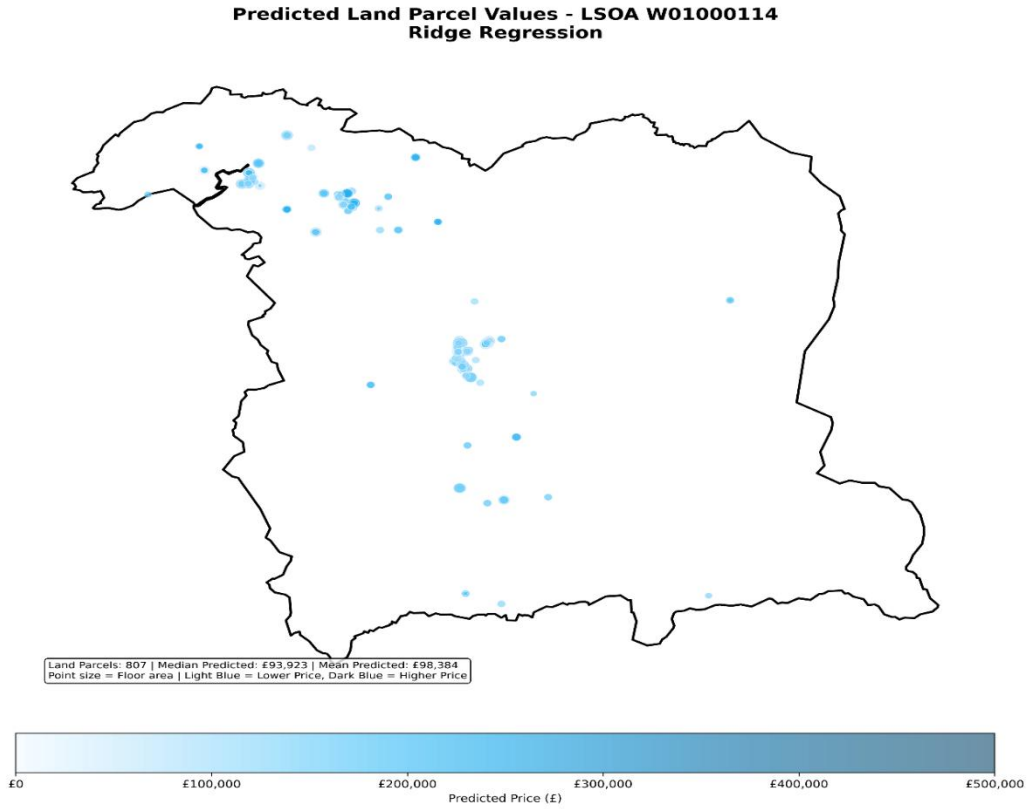


Figure 56: Land Valuation - LSOA W01000114 (Gwynedd 009D), Ridge Regression

Land Valuation - LSOA W01000114 (Gwynedd 009D), CatBoost Gradient

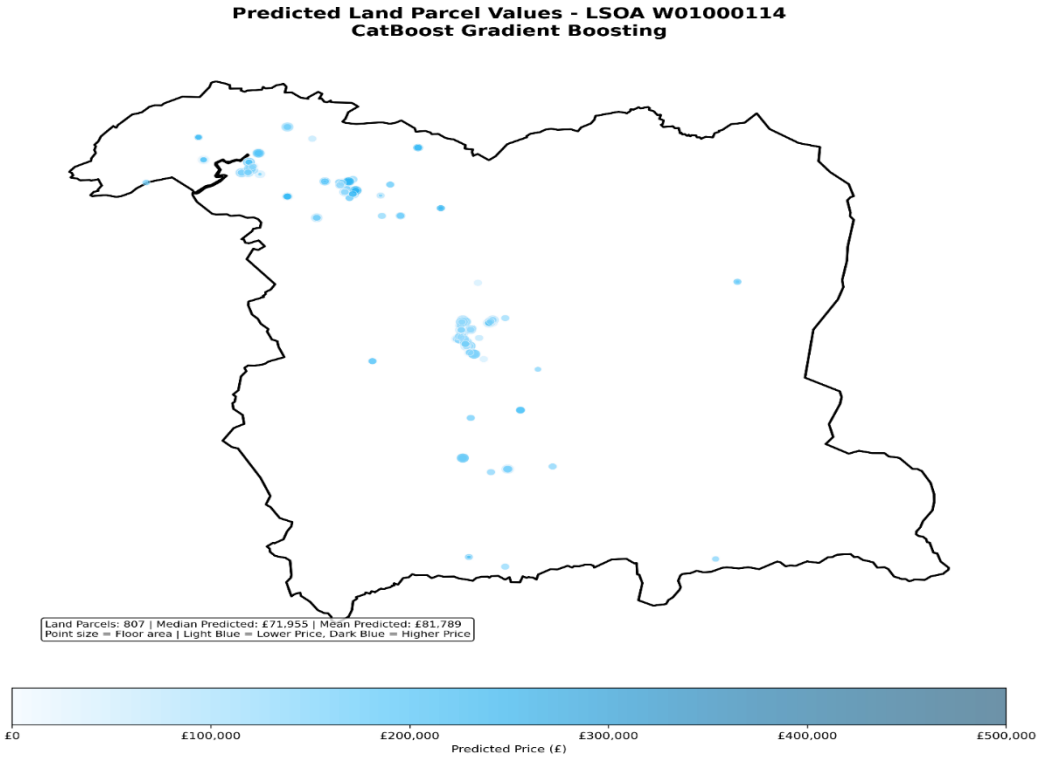


Figure 57: Land Valuation - LSOA W01000114 (Gwynedd 009D), CatBoost Gradient

Land Valuation - LSOA W01000114 (Gwynedd 009D), KNN +Fuzzy Logic

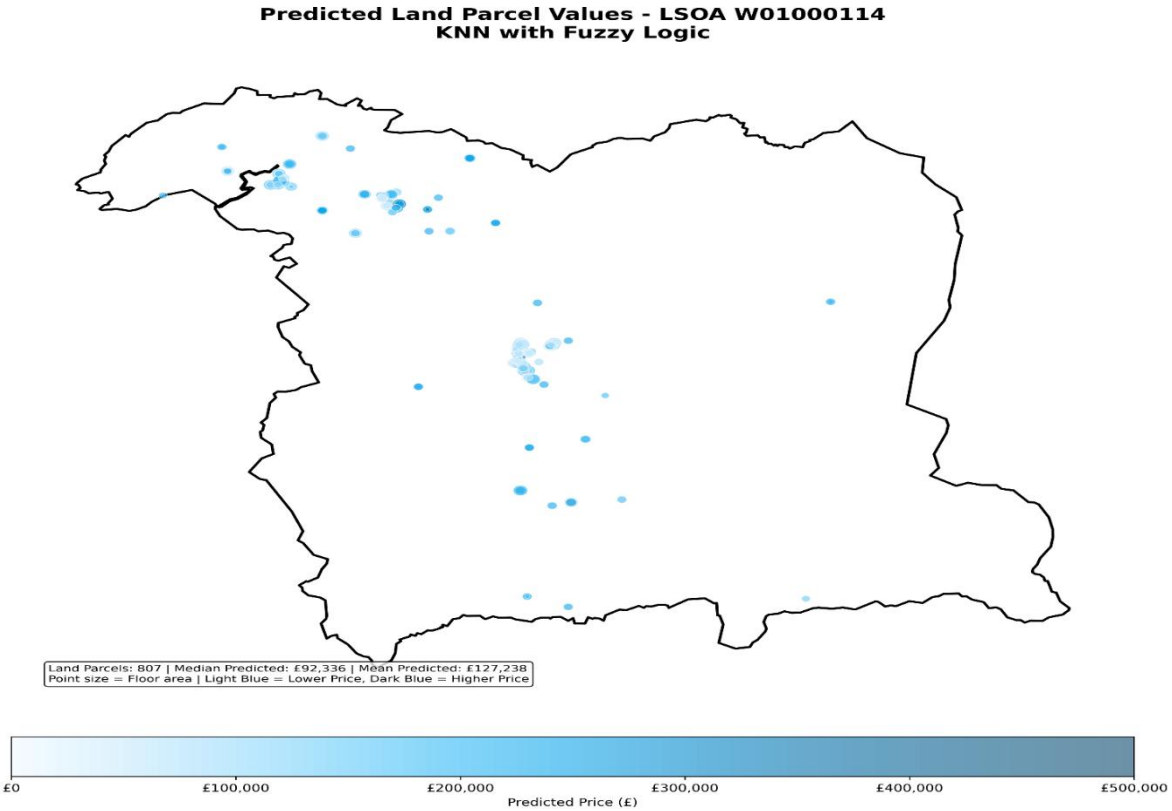


Figure 58: Land Valuation - LSOA W01000114 (Gwynedd 009D), KNN + Fuzzy Logic

Land Valuation - LSOA W01000114 (Gwynedd 009D), DRC Formula-Based

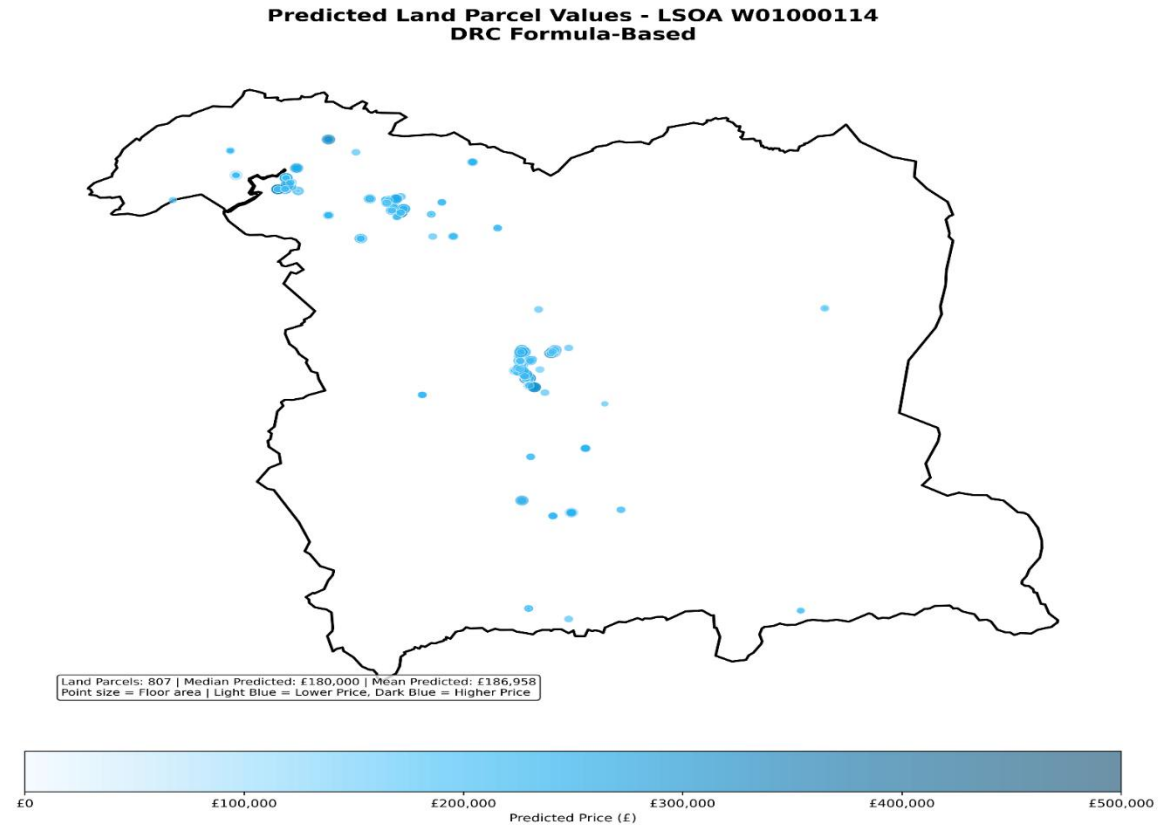


Figure 59: Land Valuation - LSOA W01000114 (Gwynedd 009D), DRC Formula-Based

Land Valuation - LSOA W01000114 (Gwynedd 009D), Multi-Agent AI Ensemble

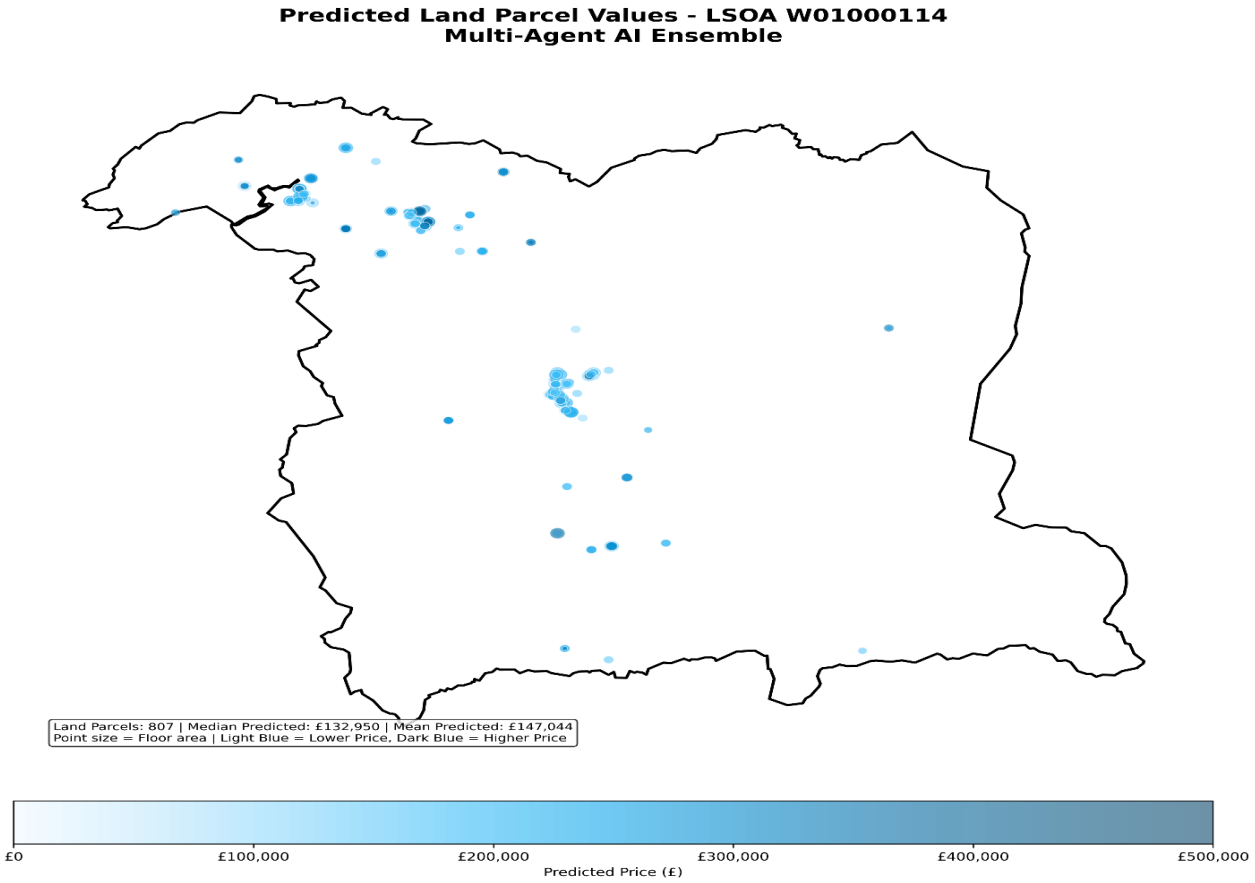


Figure 60: Land Valuation - LSOA W01000114 (Gwynedd 009D), Multi-Agent AI Ensemble

This is the lowest-priced LSOA in the test set, with a mean of £99,900 and stock dominated by older, low-value terraces and small houses. The valuation picture is therefore “low baseline with pockets”. Most transactions sit in a compressed low band, but the true spread is driven by a minority of better-located or higher-quality homes (and any local demand pockets) that should sit materially above the baseline. That’s why “average price ≈ £100k” is directionally right, but it is not sufficient as a valuation narrative unless you also acknowledge a meaningful quality and location gradient that creates real dispersion even at low absolute prices.

Lot behaviour aligns with that structure. The statistical models can reproduce the low overall level, but they still flatten the tails. Weaker stock gets pulled up toward the mean and better stock is pushed down, so the “range” looks narrower than it really is. Comparable-only and formula-based approaches are particularly unstable here, so their valuations can be off by amounts comparable to the full property price.

The methodologies capture the basic “low-price” baseline through property type (terrace vs small semi), coarse size signals (e.g., bedrooms and floorspace where present), and sale timing and year (market cycle), meaning they can place a typical transaction in the right band. What is not valued well are the drivers that create genuine dispersion within a low-price area such as; tourism and second-home demand pockets, views and landscape appeal, fine-grained accessibility (distance to key towns and links, walkability, remoteness vs convenience), and especially condition and renovation (a refurbished small house versus an unmodernised one). Because those premiums are weakly observed, the models treat

many “similar” small houses as if they should be priced similarly, when in reality these omitted factors can shift value by tens of thousands even around a £100k mean.

Gwynedd has four model property predictions above £500k across the five lots, all coming from Lot 3 / KNN. There are no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, Lot 4 / DRC, or Lot 5 / LLM Ensemble for this LSOA. This indicates the high-value outliers in Gwynedd are specific to the KNN and comparables approach, consistent with the wider pattern of occasional KNN over-extrapolation in non-standard cases.

3.7. W01001597 - Monmouthshire 006F

Market Context: This LSOA represents a commuter belt area in southeast Wales with the second-highest mean property values (£245,573).

Test Set Composition: n=967 properties, mean price £245,573

Table 28: Model Performance in LSOA W01001597 (Monmouthshire 006F)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	967	£245,573	42.8%	£73,274	£172,299–£318,847
Model 2 (CatBoost)	967	£245,573	48.5%	£68,428	£177,145–£314,001
Model 3 (KNN)	967	£245,573	-25.8%	£129,560	£116,013–£375,133
Model 4 (DRC)	967	£245,573	-14.1%	£119,157	£126,416–£364,730
Model 5 (LLM)	967	£245,573	13.1%	£100,592	£144,981–£346,165

Key Findings:

- CatBoost achieves best performance (R^2 = 48.5%, MAE = £68,428)
- Ridge achieves competitive R^2 (42.8%)
- LLM shows moderate performance (R^2 = 13.1%)
- KNN and DRC both fail with negative R^2

Land Valuation

Land Valuation - LSOA W01001597 (Monmouthshire 006F), Ridge Regression

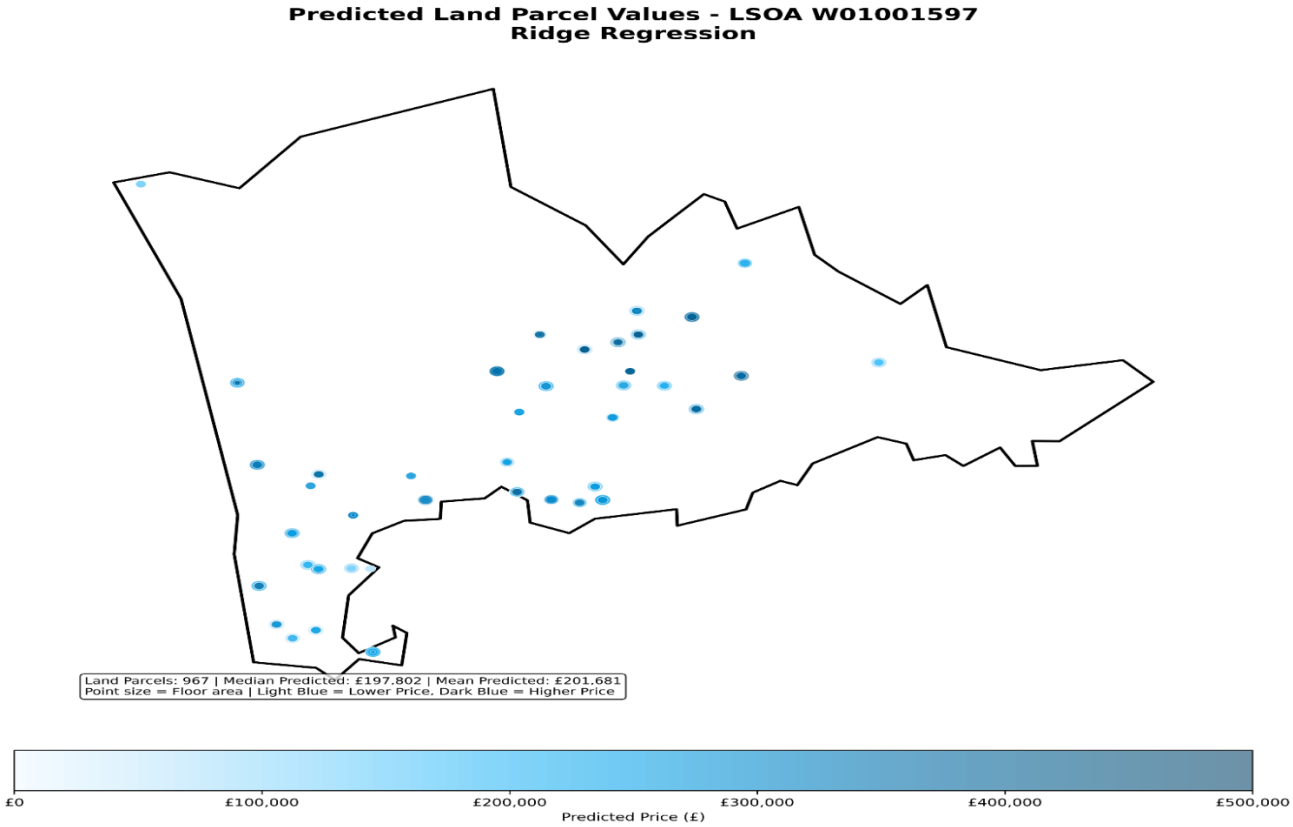


Figure 61: Land Valuation - LSOA W01001597 (Monmouthshire 006F), Ridge Regression

Land Valuation - LSOA W01001597 (Monmouthshire 006F), CatBoost Gradient

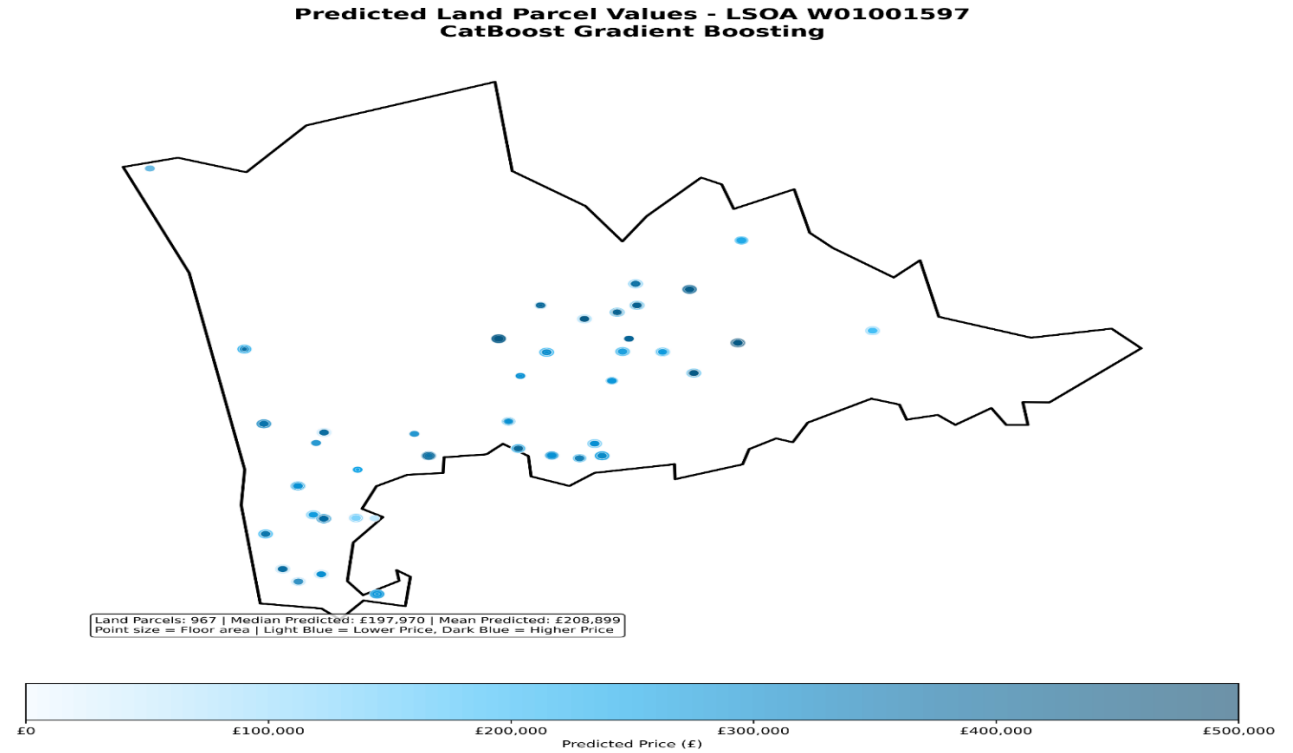


Figure 62: Land Valuation - LSOA W01001597 (Monmouthshire 006F), CatBoost Gradient

Land Valuation - LSOA W01001597 (Monmouthshire 006F), KNN +Fuzzy Logic

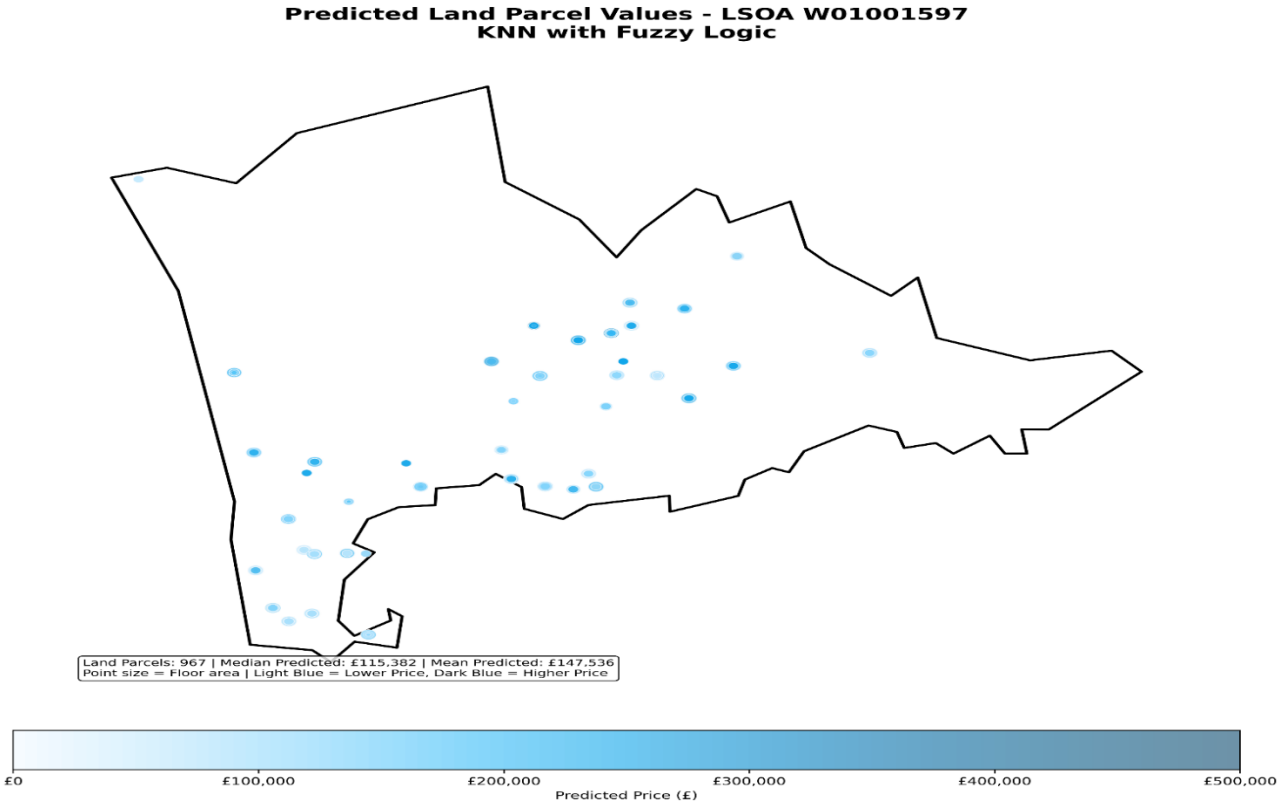


Figure 63: Land Valuation - LSOA W01001597 (Monmouthshire 006F), KNN +Fuzzy Logic

Land Valuation - LSOA W01001597 (Monmouthshire 006F), DRC Formula-Based

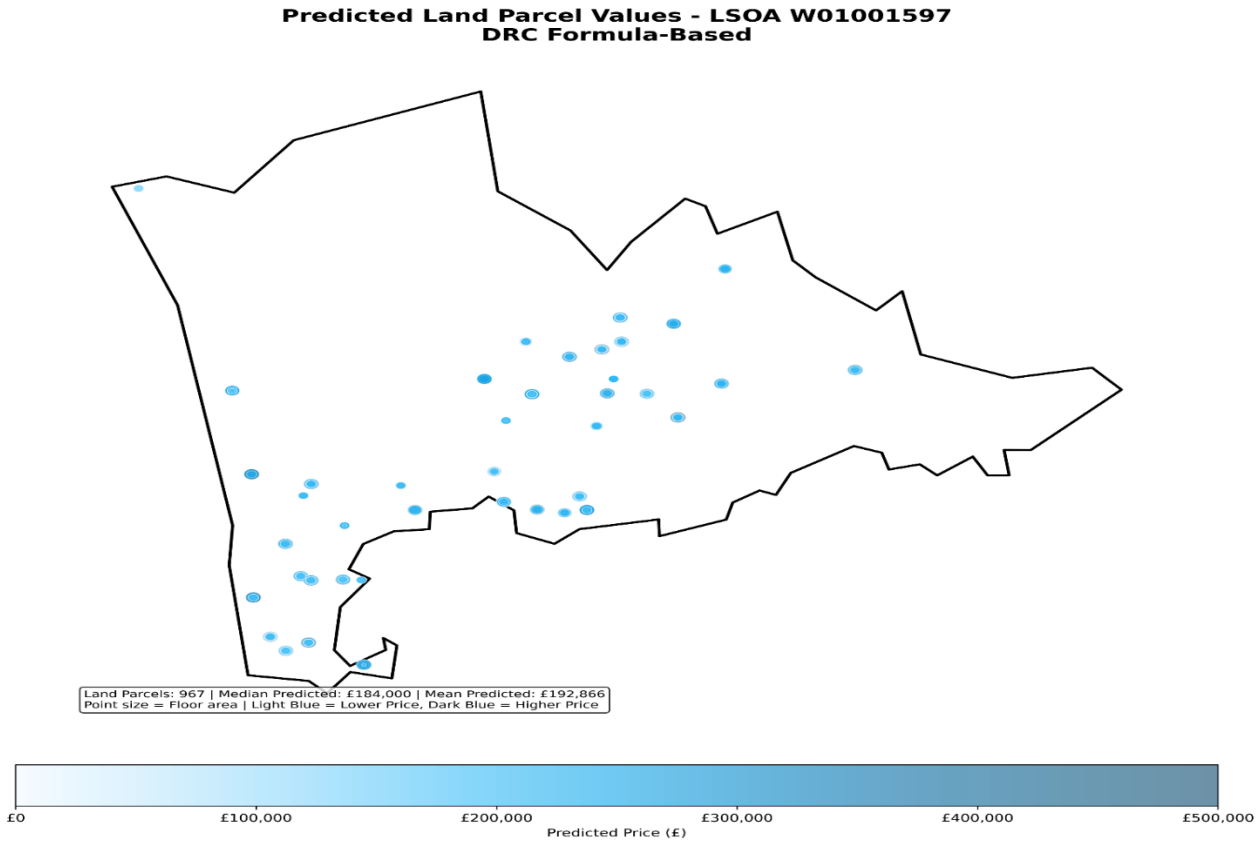


Figure 64: Land Valuation - LSOA W01001597 (Monmouthshire 006F), DRC Formula

Land Valuation - LSOA W01001597 (Monmouthshire 006F), Multi-Agent AI Ensemble

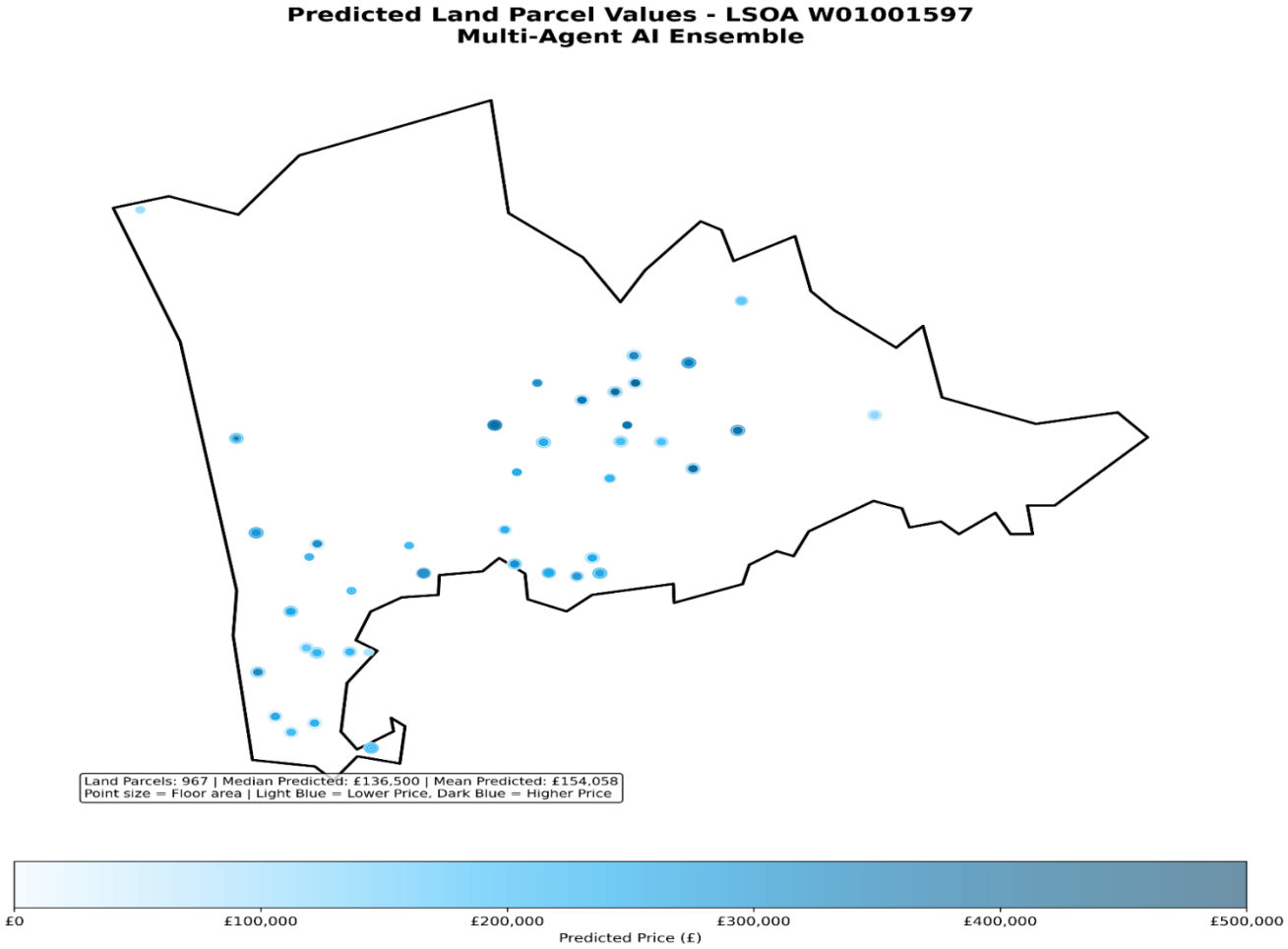


Figure 65: Land Valuation - LSOA W01001597 (Monmouthshire 006F), Multi-Agent AI Ensemble

This is a high-value commuter-belt market with the second-highest mean outside Cardiff (£245,573) and a wide spread from modest terraces to high-end detached homes. Valuation types here are best described as “strong baseline plus layered premiums”. An already elevated commuter and catchment base, with further uplifts from street prestige, larger gardens or plots, and higher-spec homes. The key point to state explicitly is that the mean reflects a mixed basket—typical family housing sits around the mid-£200k band, but a local premium segment creates an upper tail that a valuation approach must capture if outputs are to feel like real valuations rather than generic averages.

The pattern is consistent. Ridge and CatBoost tend to place typical properties at sensible levels (mid-£200k), but they under-value the top end and sometimes over-value basic terraces, compressing the spread. Comparable-only and formula-based methods are unreliable enough that their implied “average” valuation can hide very large errors.

The valuation framework captures the broad market level of the area (the general price environment implied by location), the structural differences between dwelling types (detached versus semi versus terrace), and basic scale or size effects (larger homes tending to command higher values where floorspace or bedrooms act as proxies). It also

reflects market timing (general price movements by year or period). These components are sufficient to produce a reasonable “typical value band” for mainstream properties and to rank broad categories of housing within the LSOA.

However, the approach does not fully capture the sources of within-LSOA dispersion that buyers pay for such as; micro-location (street quality, noise and traffic, adjacency to amenities or disamenities), land and setting characteristics (plot size, garden depth, privacy, views, outbuildings, corner plots), and property condition or quality (extensions, refurbishments, energy efficiency, internal finish). It also under-represents fine-grained demand premiums such as school catchment boundaries, coastal and tourism or second-home pockets, and walkability or access convenience. When these drivers are missing or weak, valuations tend to understate the premium tail (high-spec, well-sited, larger-plot homes) and overstate weaker stock, making the apparent range narrower than the true market spread.

Monmouthshire has two model property predictions above £500k across the five lots both coming from Lot 4 / DRC. There are no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, Lot 3 / KNN, or Lot 5 / LLM Ensemble for this LSOA. This indicates the high-value outliers here are specific to the DRC/formula-based approach, most likely driven by larger floor-area assumptions or atypical property characteristics rather than broad model disagreement.

3.8. W01000449 - Powys 011C

Market Context: This LSOA represents a rural area in mid-Wales, consistently achieving the strongest model performance across all test LSOAs.

Test Set Composition: n=867 properties, mean price £149,216

Table 29: Model Performance in LSOA W01000449 (Powys 011C)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	867	£149,216	51.5%	£44,020	£105,196–£193,236
Model 2 (CatBoost)	867	£149,216	51.8%	£43,868	£105,347–£193,085
Model 3 (KNN)	867	£149,216	-1556.5%	£83,486	£65,730–£232,702
Model 4 (DRC)	867	£149,216	-24.5%	£85,790	£63,426–£235,006
Model 5 (LLM)	867	£149,216	46.4%	£48,784	£100,432–£198,000

Key Findings:

- Best-performing LSOA: Ridge and CatBoost achieve $R^2 \approx 51\text{-}52\%$
- LLM achieves competitive R^2 (46.4%) but anomalous MAPE (215%)
- KNN catastrophic failure: $R^2 = -1556.5\%$, indicating complete model breakdown
- Despite strong statistical model performance, MAE still £43k-£48k

Land Valuations

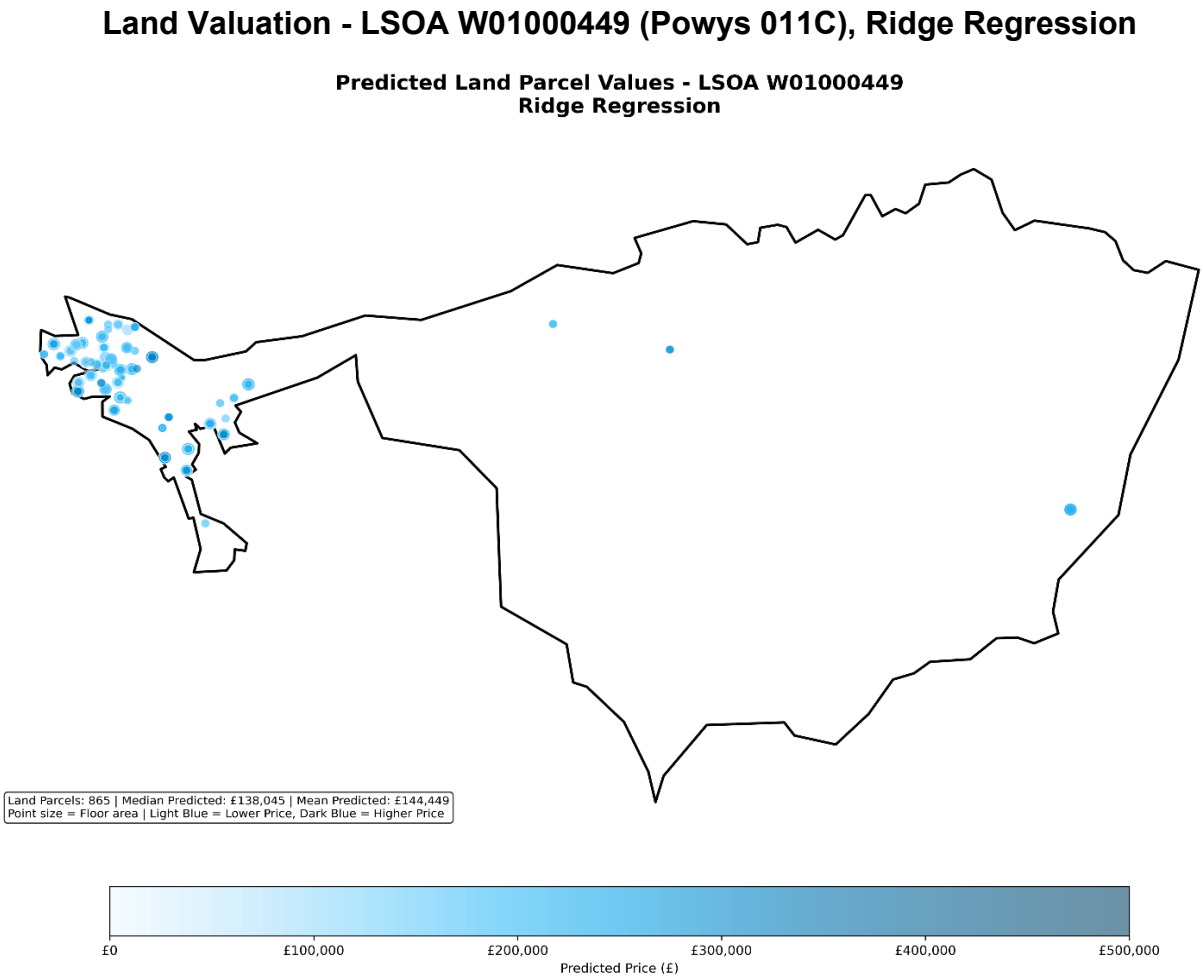


Figure 66: Land Valuation - LSOA W01000449 (Powys 011C), Ridge Regression

Land Valuation - LSOA W01000449 (Powys 011C), CatBoost Gradient

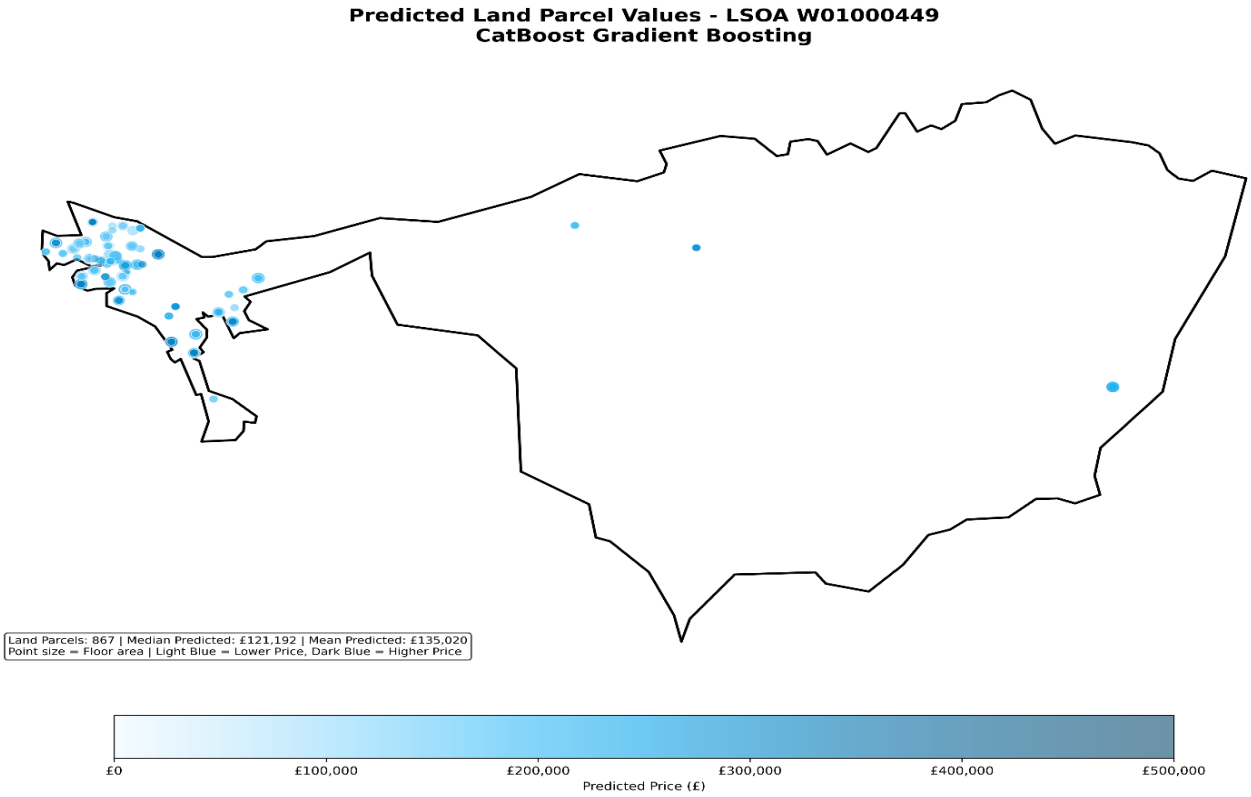


Figure 67: Land Valuation - LSOA W01000449 (Powys 011C), CatBoost Gradient

Land Valuation - LSOA W01000449 (Powys 011C), KNN +Fuzzy Logic

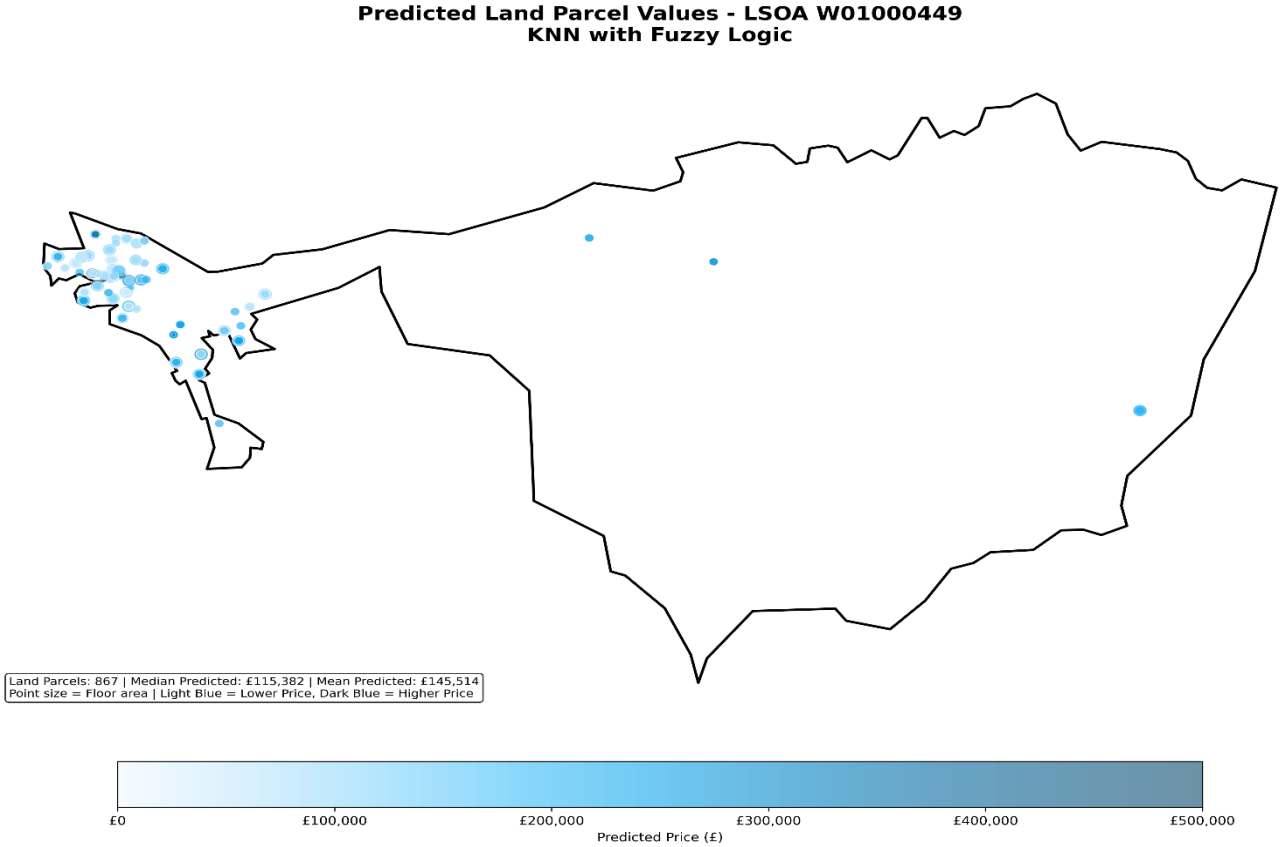


Figure 68: Land Valuation - LSOA W01000449 (Powys 011C), KNN +Fuzzy Logic

Land Valuation - LSOA W01000449 (Powys 011C), DRC Formula-Based

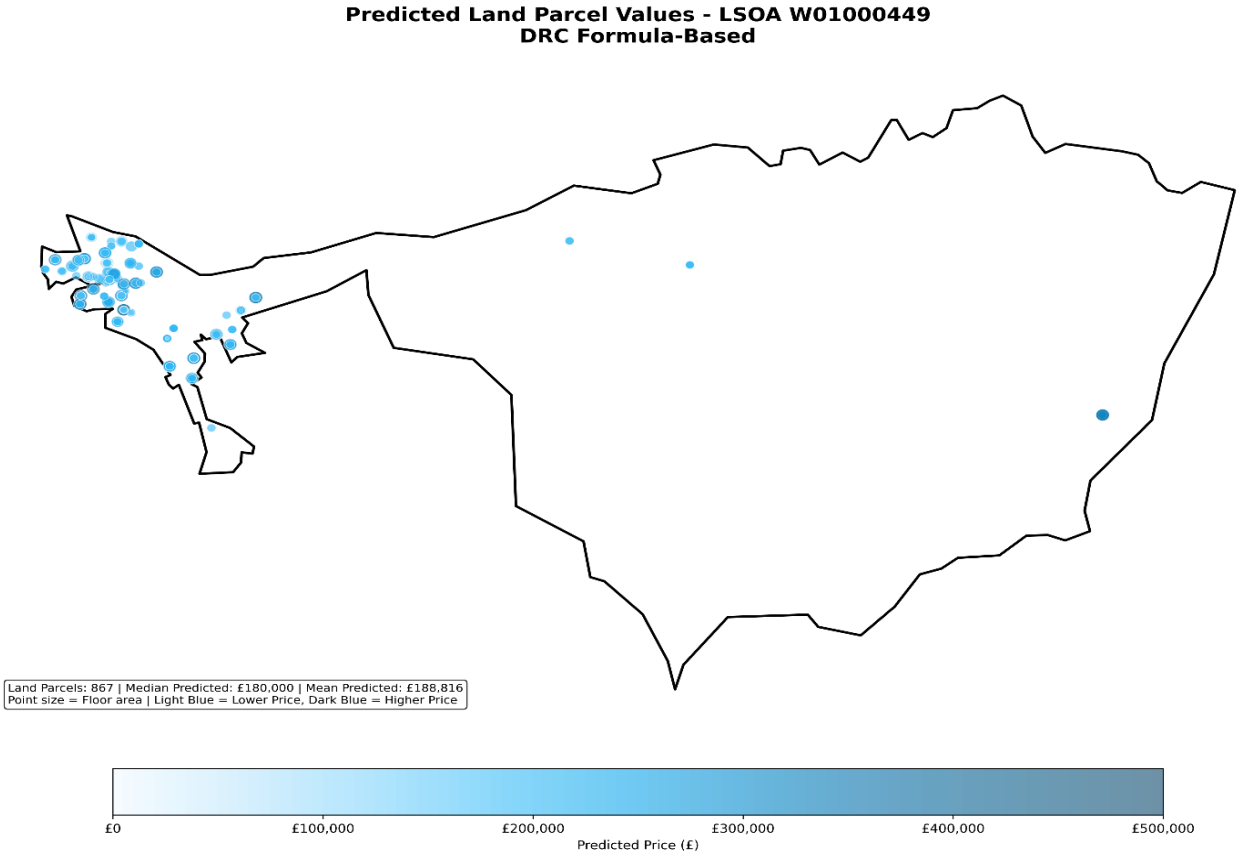


Figure 69: Land Valuation - LSOA W01000449 (Powys 011C), DRC Formula-Based

Land Valuation - LSOA W01000449 (Powys 011C), Multi-Agent AI Ensemble

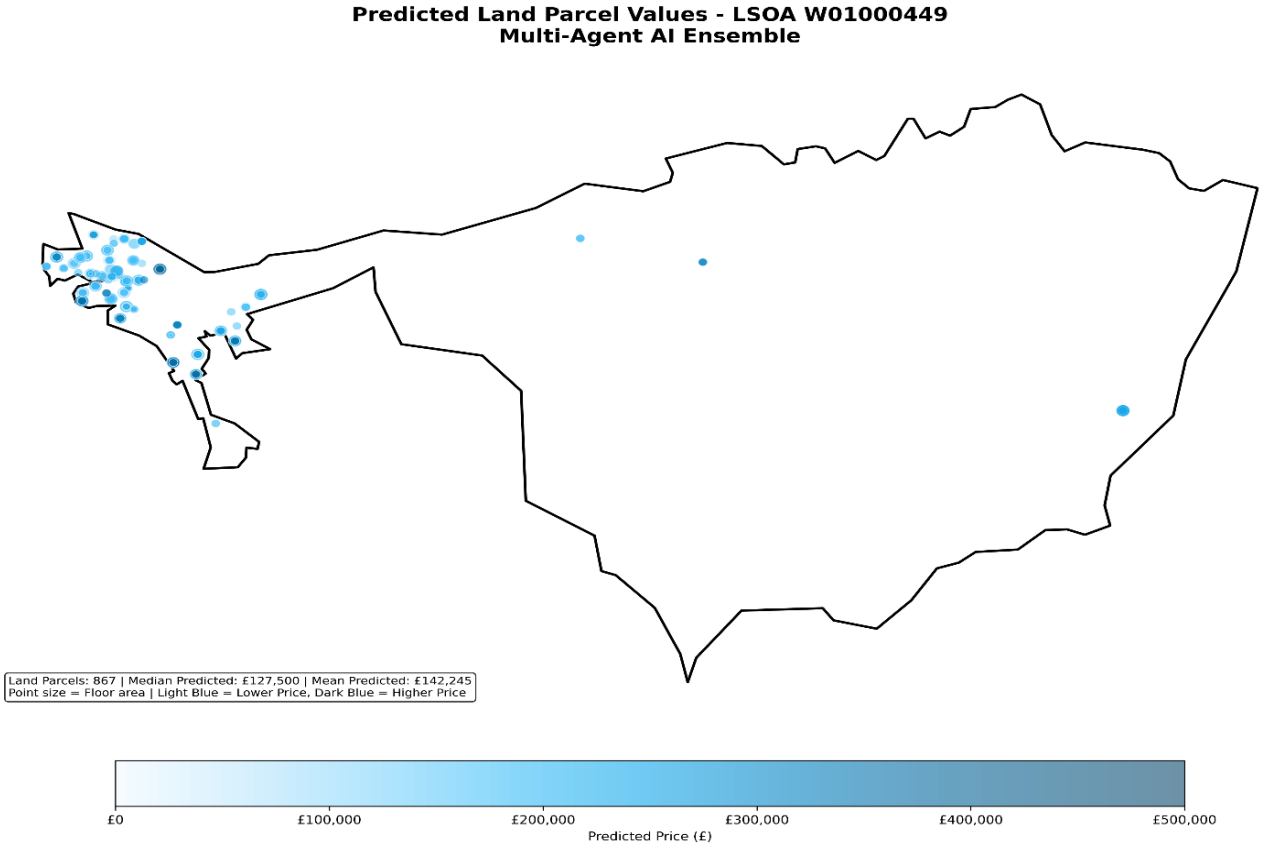


Figure 70: Land Valuation - LSOA W01000449 (Powys 011C), Multi-Agent AI Ensemble

This is a mid-priced rural market with a mean of £149,216, spanning smaller cottages through to more valuable farmhouses. Valuation types here are best described as “rural baseline plus land and extensiveness premiums”. The baseline is learnable because the stock is comparatively homogeneous, but the true spread is shaped by factors that are not well represented in minimal feature sets—acreage and plot size, outbuildings, views, remoteness versus accessibility, and renovation. The average figure is meaningful as a baseline, but a credible valuation narrative should state that dispersion is driven by land and setting rather than just “house type”, and those drivers are precisely where automated systems typically weaken.

In this LSOA, the statistical models (and even the LLM) produce their most coherent valuations, broadly separating cheaper small homes from larger, higher-value homes. The important qualifier is that the models still struggle to price rural “uniqueness” premiums consistently. Comparable-only retrieval can fail catastrophically in rural settings (by selecting the wrong “nearest” markets), and formula methods miss key mechanisms of market formation.

In this rural setting, what is valued well is the part of the market that behaves predictably. The valuation can set the baseline price level for the LSOA and apply the obvious structural uplifts (small cottage versus standard house versus larger detached), plus (where available) basic size proxies such as floorspace and bedrooms. That captures a meaningful share of “typical” pricing because many transactions sit within a relatively consistent rural band where buyers largely pay for dwelling type and scale, alongside the general location.

What is not valued well are the rural attributes that create big gaps between two seemingly similar detached homes. In practice, rural dispersion is often driven by factors that are weak or missing in the available inputs such as; acreage and plot size (land value and usable space), outbuildings and utility (garages, barns, workshops, equestrian facilities), setting and privacy (views, seclusion, noise, immediate surroundings), and micro-access (lane quality, winter access, distance to services, commuting practicality). Add condition and renovation (a modernised farmhouse versus tired stock), and these “non-headline” features can shift value materially. As a result, valuations that rely mainly on type and size and broad location tend to flatten the premium rural tail and understate true dispersion.

Powys has three model property predictions above £500k across the five lots, all coming from Lot 3 / KNN. There are no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, Lot 4 / DRC, or Lot 5 / LLM Ensemble for this LSOA. This pattern is consistent with the wider behaviour of KNN/comparables automation, where a small number of properties can trigger extreme extrapolations, especially in lower-density or less comparable local markets.

3.9. W01000617 - Pembrokeshire 002F

Market Context: This LSOA represents a coastal area in southwest Wales.

Test Set Composition: n=506 properties, mean price £178,035

Table 30: Model Performance in LSOA W01000617 (Pembrokeshire 002F)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	506	£178,035	31.3%	£72,951	£105,084–£250,986
Model 2 (CatBoost)	506	£178,035	24.7%	£75,807	£102,228–£253,842
Model 3 (KNN)	506	£178,035	-12.9%	£102,649	£75,386–£280,684
Model 4 (DRC)	506	£178,035	-5.0%	£109,950	£68,085–£287,985
Model 5 (LLM)	506	£178,035	24.8%	£80,162	£97,873–£258,197

Key Findings:

- Ridge achieves best R^2 (31.3%)
- CatBoost and LLM show similar R^2 (≈ 24-25%)
- KNN and DRC both fail with negative R^2
- MAE ranges £72k-£110k across models

Land Valuations

Land Valuation - LSOA W01000617 (Pembrokeshire 002F), Ridge Regression

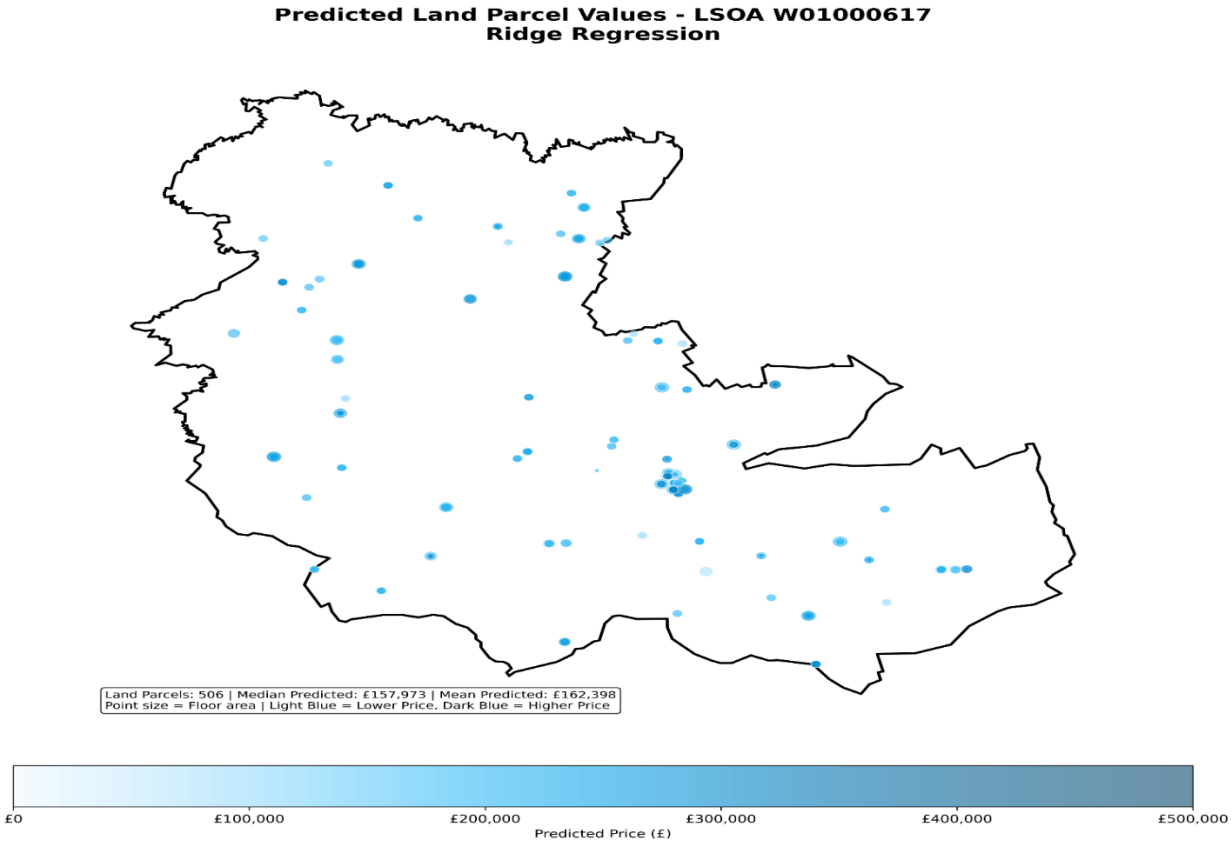


Figure 71: Land Valuation - LSOA W01000617 (Pembrokeshire 002F), Ridge Regression

Land Valuation - LSOA W01000617 (Pembrokeshire 002F), CatBoost Gradient

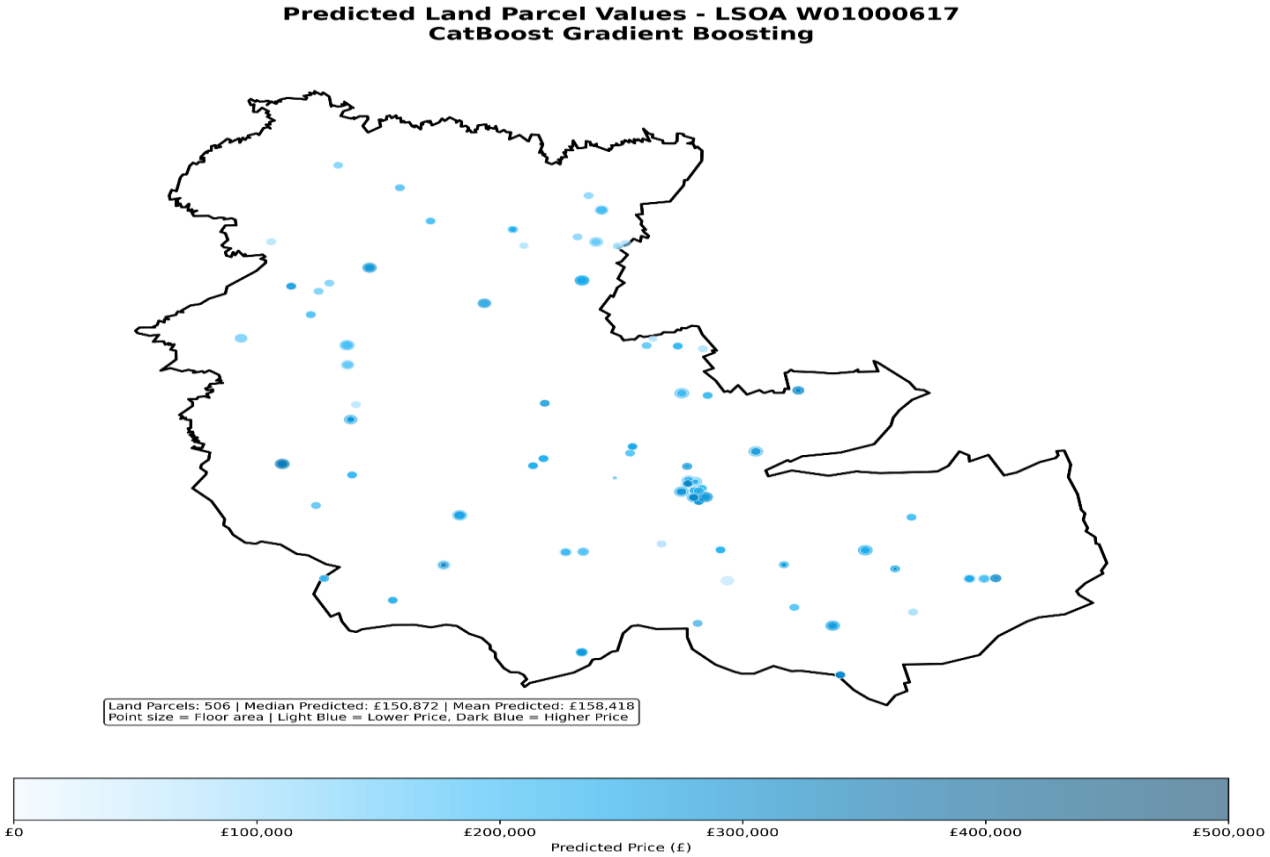


Figure 72: Land Valuation - LSOA W01000617 (Pembrokeshire 002F), CatBoost Gradient

Land Valuation - LSOA W01000617 (Pembrokeshire 002F), KNN +Fuzzy Logic

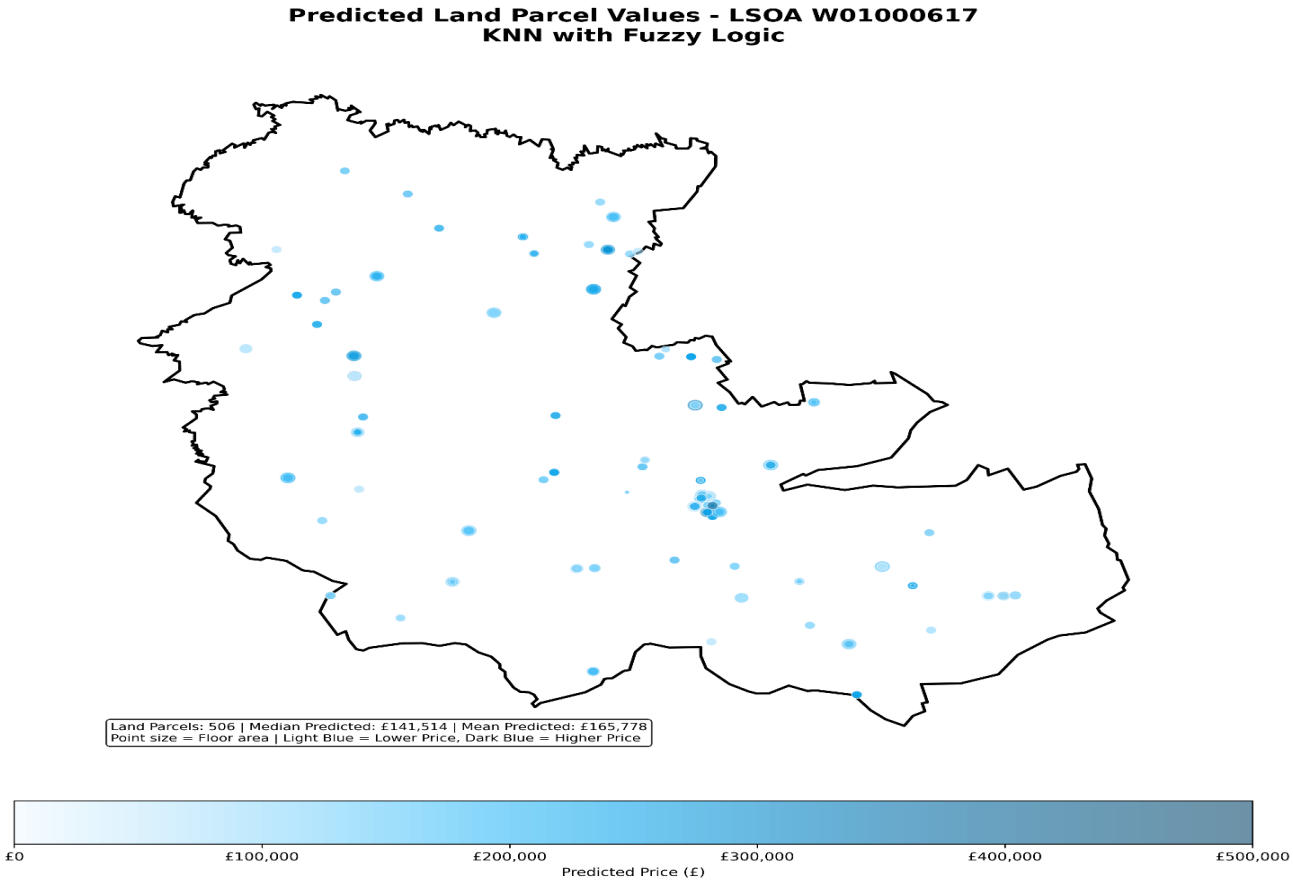


Figure 73: Land Valuation - LSOA W01000617 (Pembrokeshire 002F), KNN +Fuzzy Logic

Land Valuation - LSOA W01000617 (Pembrokeshire 002F), DRC Formula-Based

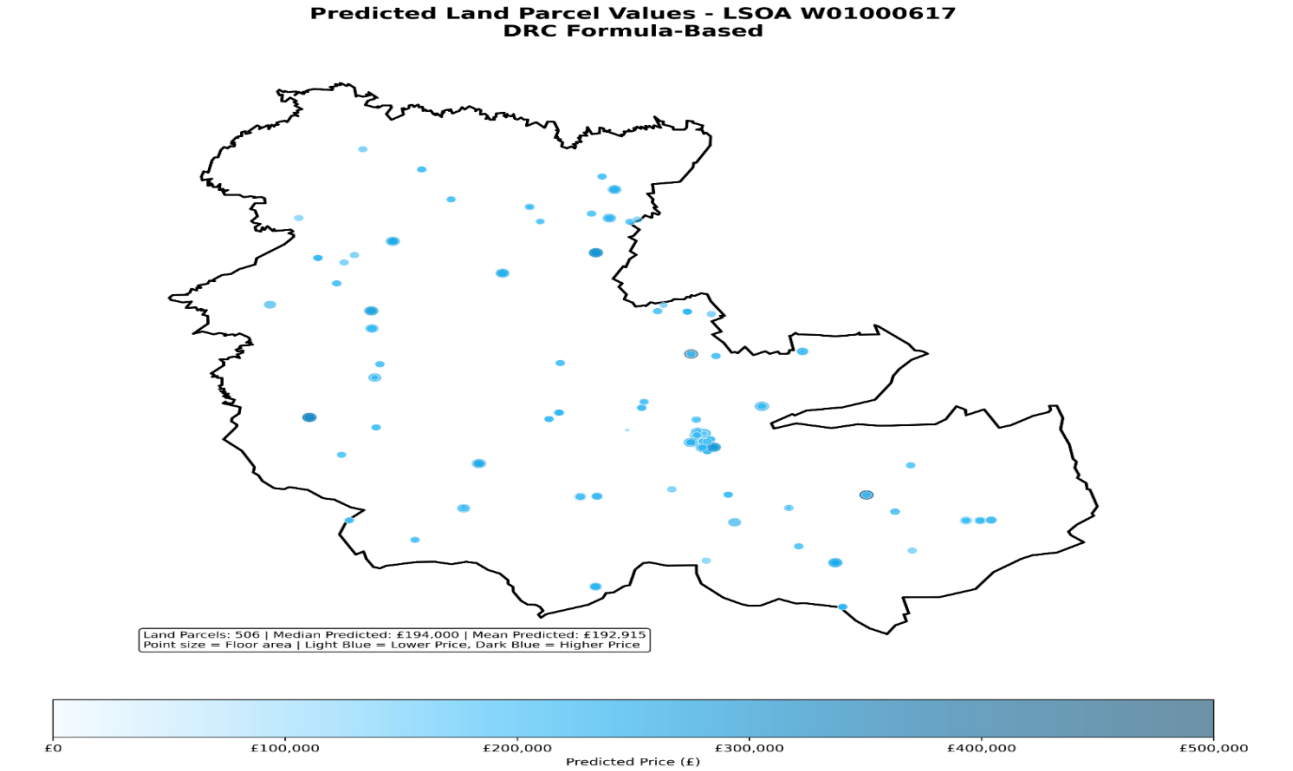


Figure 74: Land Valuation - LSOA W01000617 (Pembrokeshire 002F), DRC Formula-Based

Land Valuation - LSOA W01000617 (Pembrokeshire 002F), Multi-Agent AI Ensemble

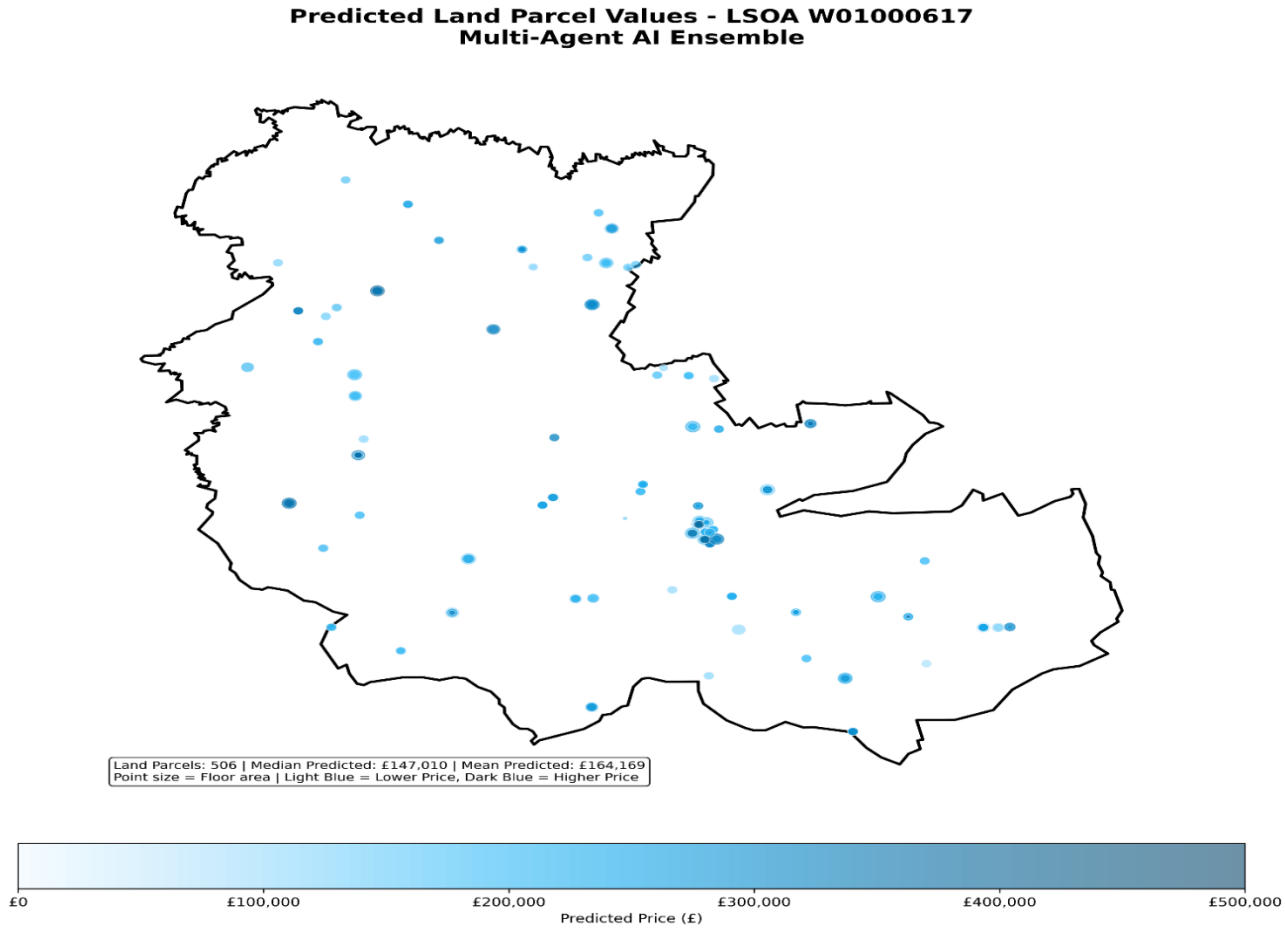


Figure 75 : Land Valuation - LSOA W01000617 (Pembrokeshire 002F), Multi-Agent AI Ensemble

This is a coastal market with a mean of £178,035, but the mean masks a two-market structure/ Ordinary inland housing versus premium coastal and holiday-let stock. That creates a valuation environment where “typical” values may genuinely cluster around the mid-£100k levels, while the upper tail is driven by coastal proximity, amenity, and second-home and tourism demand. The spread is therefore not a statistical curiosity; it is the core valuation story. A narrative that only reports an average misses the practical reality that two similar-sized homes can diverge sharply in price based on sea views, walkability, and micro-siting.

Model differences map directly onto that segmentation. Better statistical models keep inland and typical properties near plausible mid-£100k valuations, but they systematically under-value premium coastal or sea-view homes and can over-value basic inland dwellings, compressing the range. Comparable-only and formula-based methods add volatility without solving the segmentation problem.

In a coastal market, what is valued well is the part of pricing that follows broad, observable rules. The valuation captures structural fundamentals (property type and floorspace and size proxies, e.g., detached versus terrace and “bigger homes cost more”) and applies a coarse geographic uplift that reflects the general regional market level. That is usually enough to place typical inland or non-premium stock in a plausible mid-band and to

distinguish mainstream family homes from smaller terraces and flats. In other words, the approach can price the “baseline” segment reasonably: ordinary housing that is not strongly dependent on coastal amenity.

What is not valued well are the elements that create the coastal premium ladder, because they depend on very fine-grained, often unobserved information. Coastal value is frequently driven by micro-distance to the coast (sometimes a few streets makes a large difference), tourism and second-home demand intensity, and amenity factors such as sea views, aspect, and walkability to beaches and attractions—all of which can produce large uplifts even for otherwise similar homes. On top of that, condition and renovation quality often matters disproportionately in holiday and letting markets (turnkey finish, extensions, energy performance), and those signals are rarely captured cleanly in standard transaction datasets. When these premiums are not represented, valuations tend to default toward the inland baseline, under-pricing premium coastal homes and compressing the true spread that buyers actually pay for.

3.10. W01001233 - Rhondda Cyon Taf 001F

Market Context: This LSOA represents a former industrial valley area in South Wales.

Test Set Composition: n=585 properties, mean price £150,537

Table 31: Model Performance in LSOA W01001233 (Rhondda Cynon Taf 001F)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	585	£150,537	2.8%	£63,094	£87,443–£213,631
Model 2 (CatBoost)	585	£150,537	3.4%	£62,925	£87,612–£213,462
Model 3 (KNN)	585	£150,537	-1.1%	£92,548	£57,989–£243,085
Model 4 (DRC)	585	£150,537	-4.1%	£113,989	£36,548–£264,526
Model 5 (LLM)	585	£150,537	7.5%	£76,590	£73,947–£227,127

Key Findings:

- All models achieve very low R^2 (< 8%), indicating minimal predictive power
- LLM achieves highest R^2 (7.5%) but with high MAE
- Ridge and CatBoost achieve similar low performance ($R^2 \approx$ 2-3%)
- KNN and DRC both fail with negative R^2

Land Valuations

Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), Ridge Regression

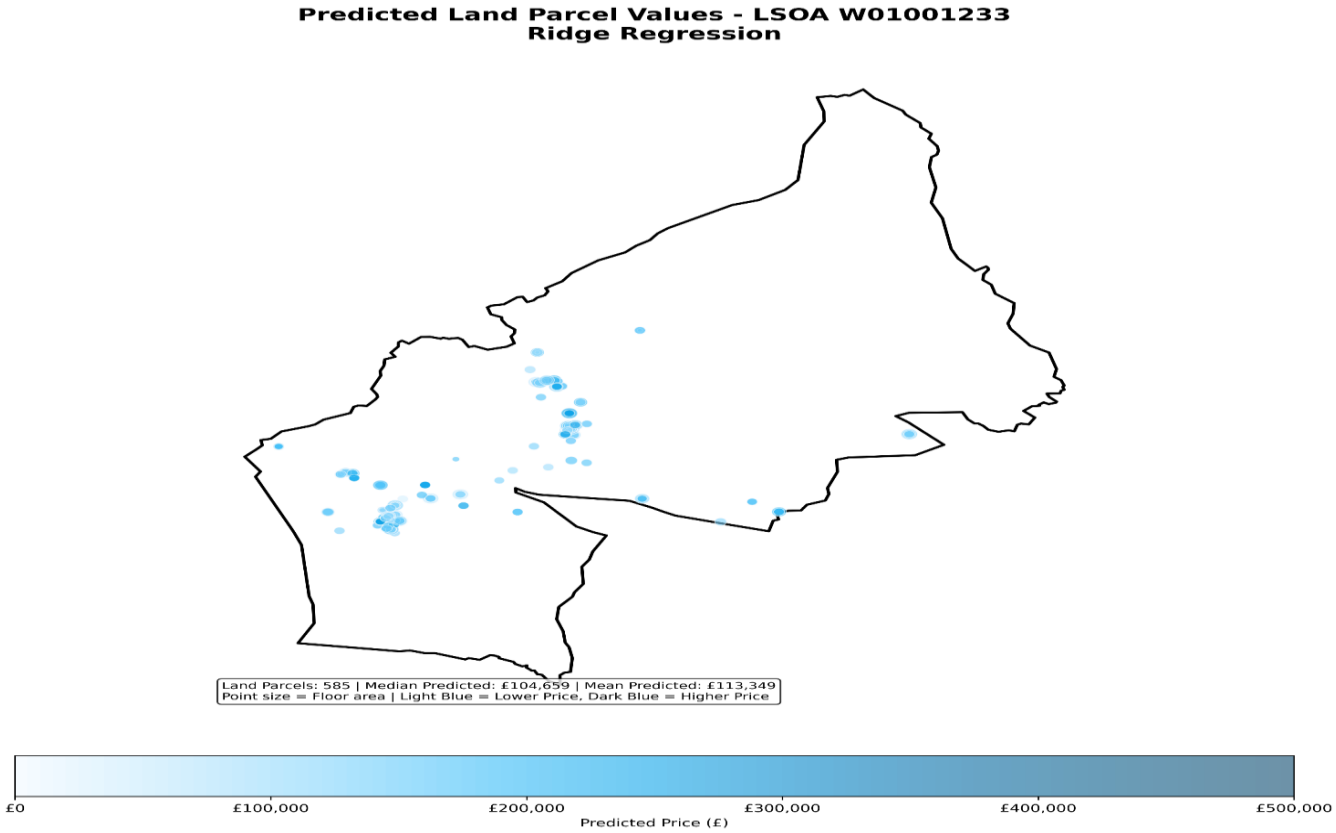


Figure 76: Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), Ridge Regression

Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), CatBoost Gradient

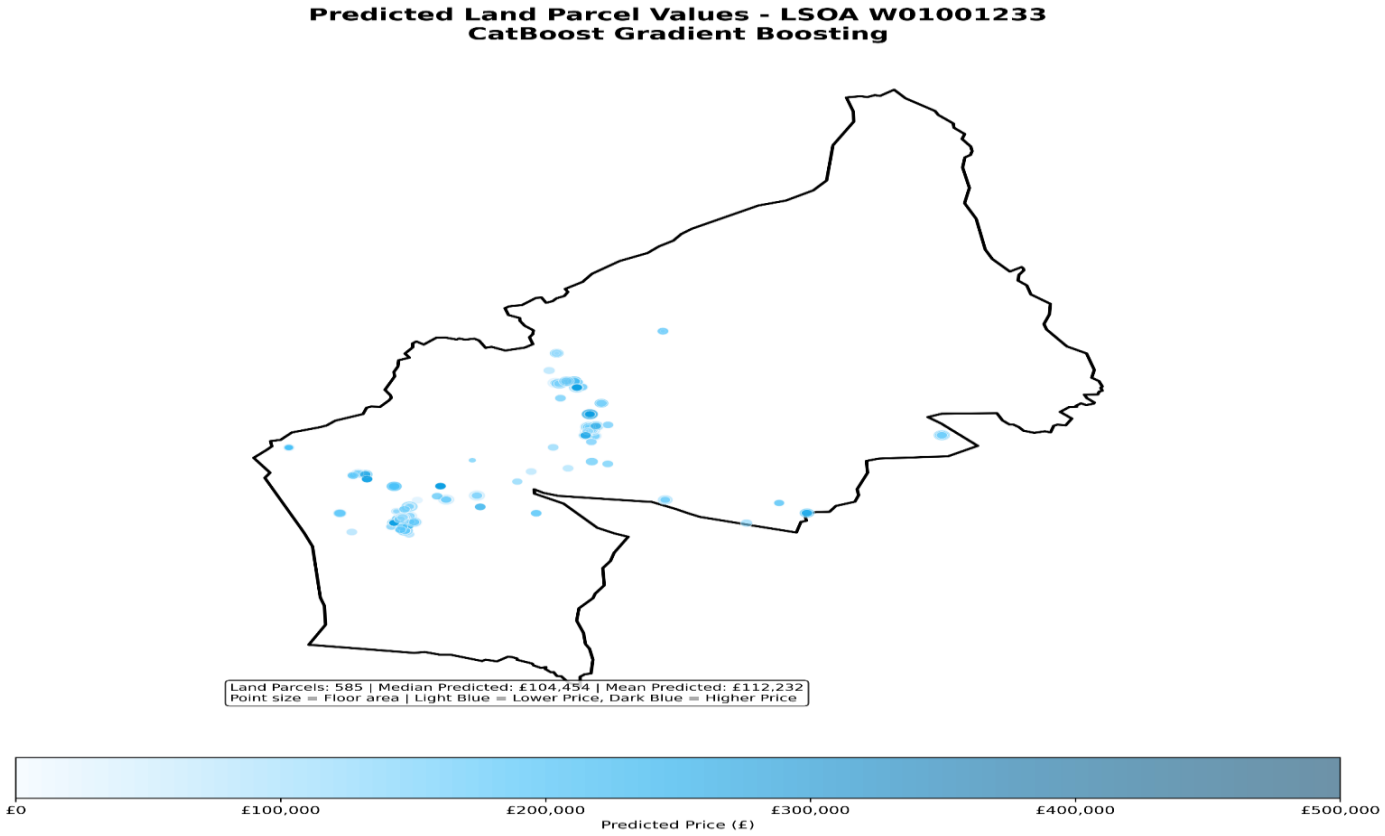


Figure 77: Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), CatBoost Gradient

Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), KNN +Fuzzy Logic

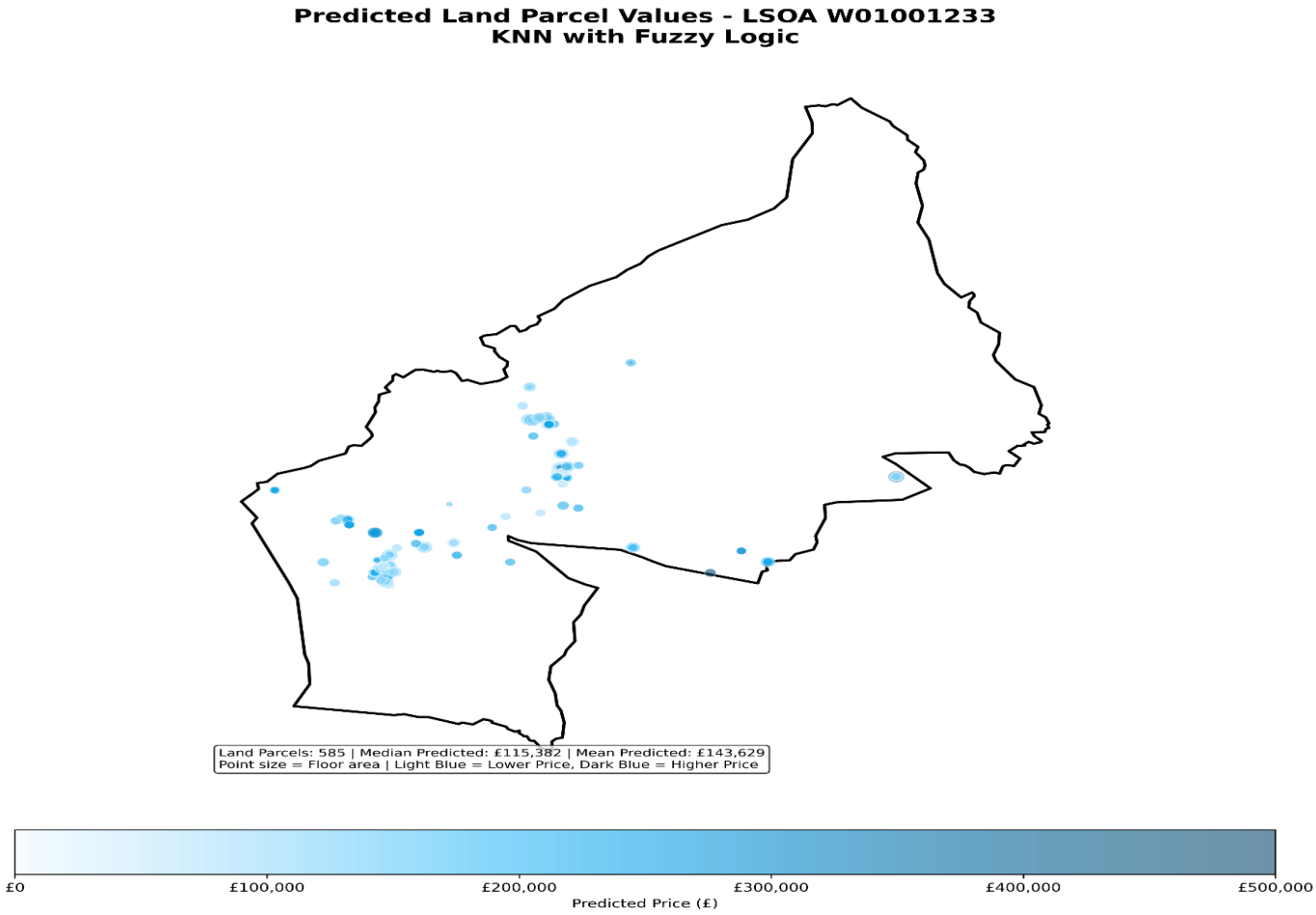


Figure 78: Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), KNN +Fuzzy Logic

Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), DRC Formula-Based

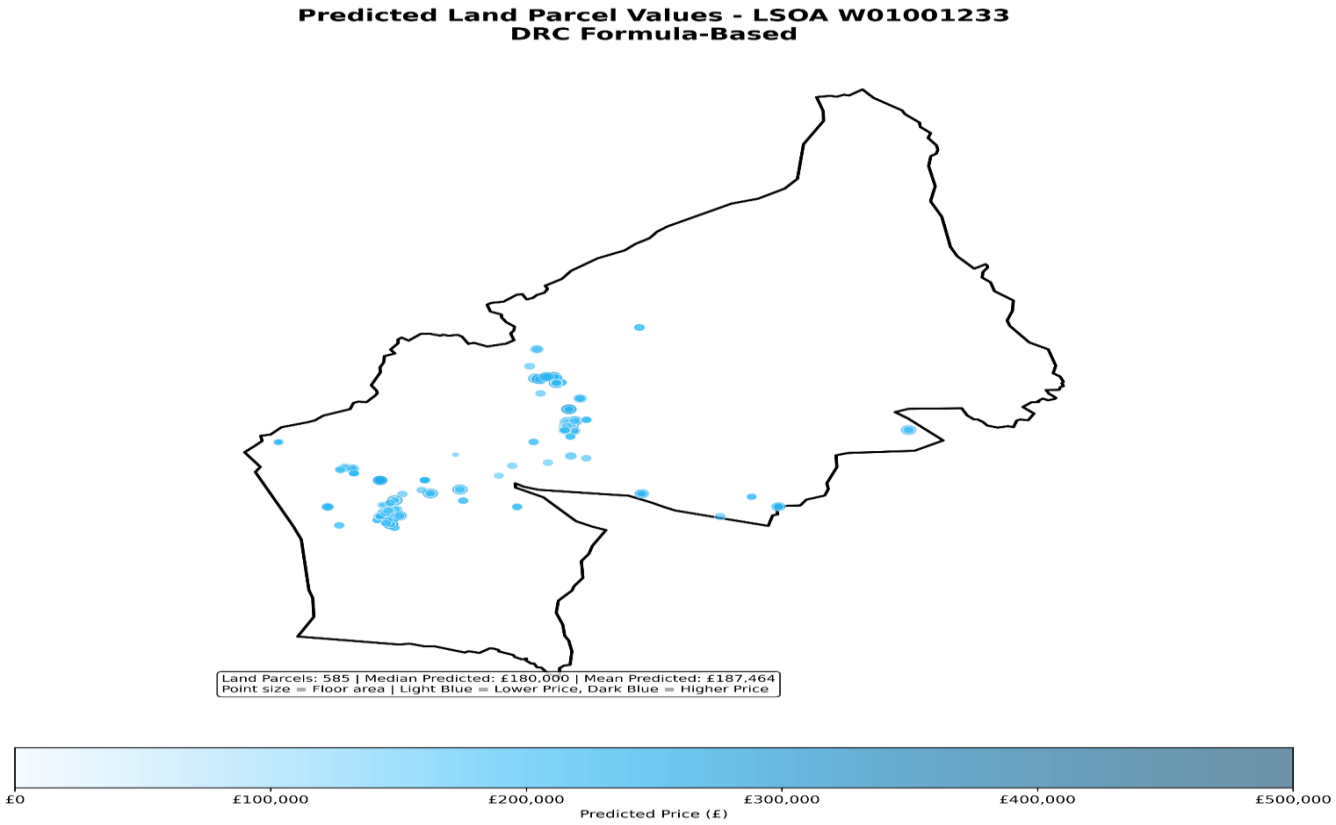


Figure 79: Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), DRC Formula-Based

Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), Multi-Agent AI Ensemble

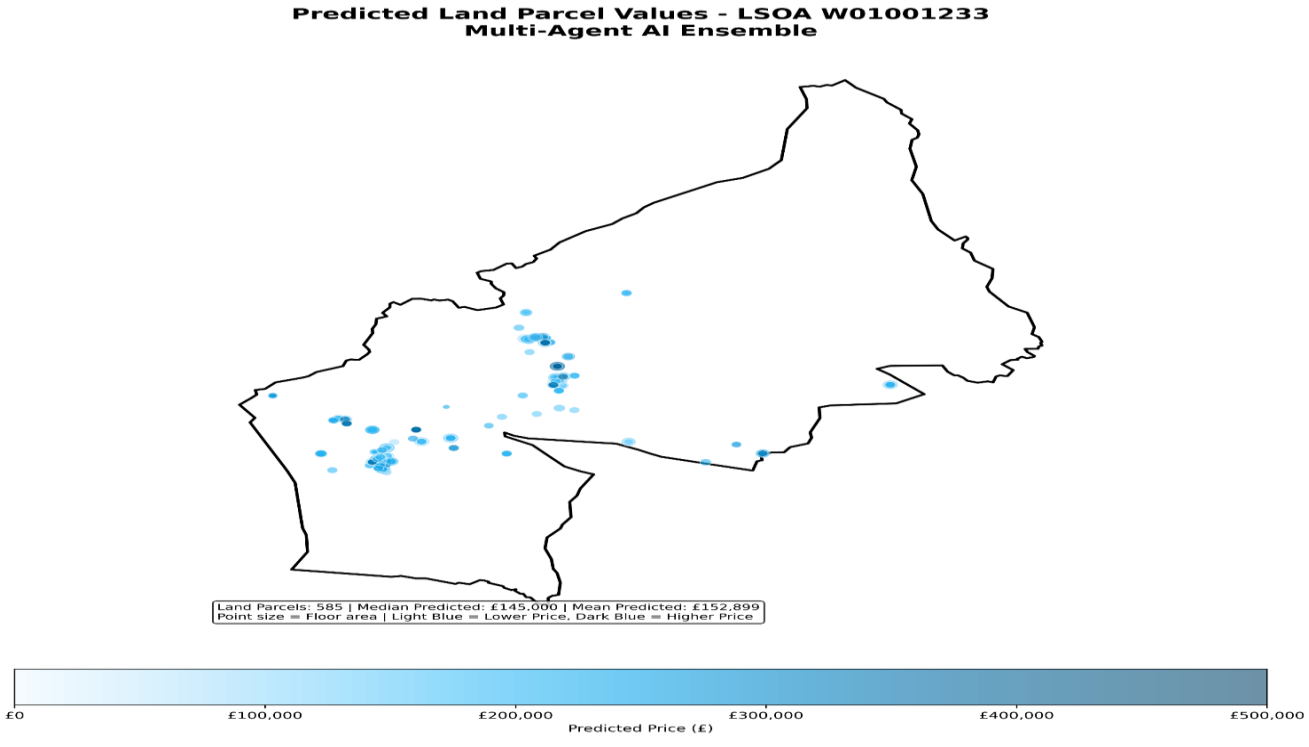


Figure 80: Land Valuation - LSOA W01001233 (Rhondda Cyon Taf 001F), Multi-Agent AI Ensemble

This is a former industrial valley market with a mean of £150,537, close to the Welsh median, but with a characteristic “valley distribution” characterised by large mass of older terraces and a thinner band of semi-detached and detached stock above them. Valuation types are therefore “valley baseline plus micro-premiums”. Features such as; street desirability, parking and access, and renovation quality can move prices substantially even among similar terraces, while detached homes should sit materially above the terrace baseline. A strong valuation narrative should explicitly state that dispersion here comes less from dramatic geography and more from within-valley micro-location and condition, which are hard to observe in minimal feature sets.

Across approaches, models generally get the ordering right (detached > terrace) but understate the true gap, pulling terraces up toward the mean and pushing detached values down; again flattening the spread. Comparable-only and formula-based approaches are weak, but even the statistical models struggle because the differentiators are local and granular

What is valued well in this type of valley and urban-edge LSOA is the “core baseline” of the housing market. The valuation captures the broad area level (the general price environment for this part of Rhondda Cynon Taf) and applies the obvious structural uplifts from property type and scale; terrace versus semi versus detached—plus (where present) basic size proxies such as floorspace and bedrooms. That is usually enough to place a typical terrace or semi in a sensible mid-band and to maintain the correct ordering that detached homes should generally sit above terraces.

What is not valued well are the local drivers that create real within-LSOA dispersion and explain why two homes with similar headline attributes can differ materially in price. These include condition and renovation quality (modernised interiors, extensions, energy upgrades), street-level desirability (a quiet cul-de-sac versus a main road, proximity to disamenities, neighbourhood reputation), and practical frictions such as parking constraints and access and egress. The approach also under-represents fine-grained accessibility and employment effects (proximity to transport links, commuting convenience, and local job centres), which can shift demand sharply even over short distances. When these signals are weak or missing, valuations implicitly treat the area as more uniform than it is, pulling weaker stock up and pushing better-sited or renovated stock down, so the distribution converges too tightly around the average, and the true spread is understated.

Rhondda Cyon Taf has one model property prediction above £500k across the five lots, making it a very rare outlier case, coming from Lot 3 / KNN (1), with no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, Lot 4 / DRC, or Lot 5 / LLM Ensemble. This points to a single KNN-specific over-extrapolation event, rather than a recurring issue across methods.

3.11. W01002019 - Cardiff 032H

Market Context: This LSOA represents Cardiff's high-value urban market with the highest mean property values (£374,063) and largest sample size (3,298 properties, 34.3% of test set).

Test Set Composition: n=3,298 properties, mean price £374,063

Table 32: Model Performance in LSOA W01002019 (Cardiff 032H)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	3,298	£374,063	1.7%	£174,127	£199,936–£548,190
Model 2 (CatBoost)	3,298	£374,063	0.2%	£165,983	£208,080–£540,046
Model 3 (KNN)	3,298	£374,063	0.3%	£282,327	£91,736–£656,390
Model 4 (DRC)	3,298	£374,063	-1.3%	£263,388	£110,675–£637,451
Model 5 (LLM)	3,298	£374,063	-1.7%	£270,104	£103,959–£644,167

Key Findings:

- All models achieve low R^2 or negative, indicating complete failure
- Ridge achieves marginally best R^2 (1.7%) and CatBoost lowest MAE (£165,983)
- LLM performs poorly (R^2 = -1.7%, MAE = £270,104)
- Urban complexity proves insurmountable: MAE ranges £166k–£282k
- Accounts for 34% of the whole test set but universally unpredictable

Land Valuations

Land Valuation - LSOA W01002019 (Cardiff 032H), Ridge Regression

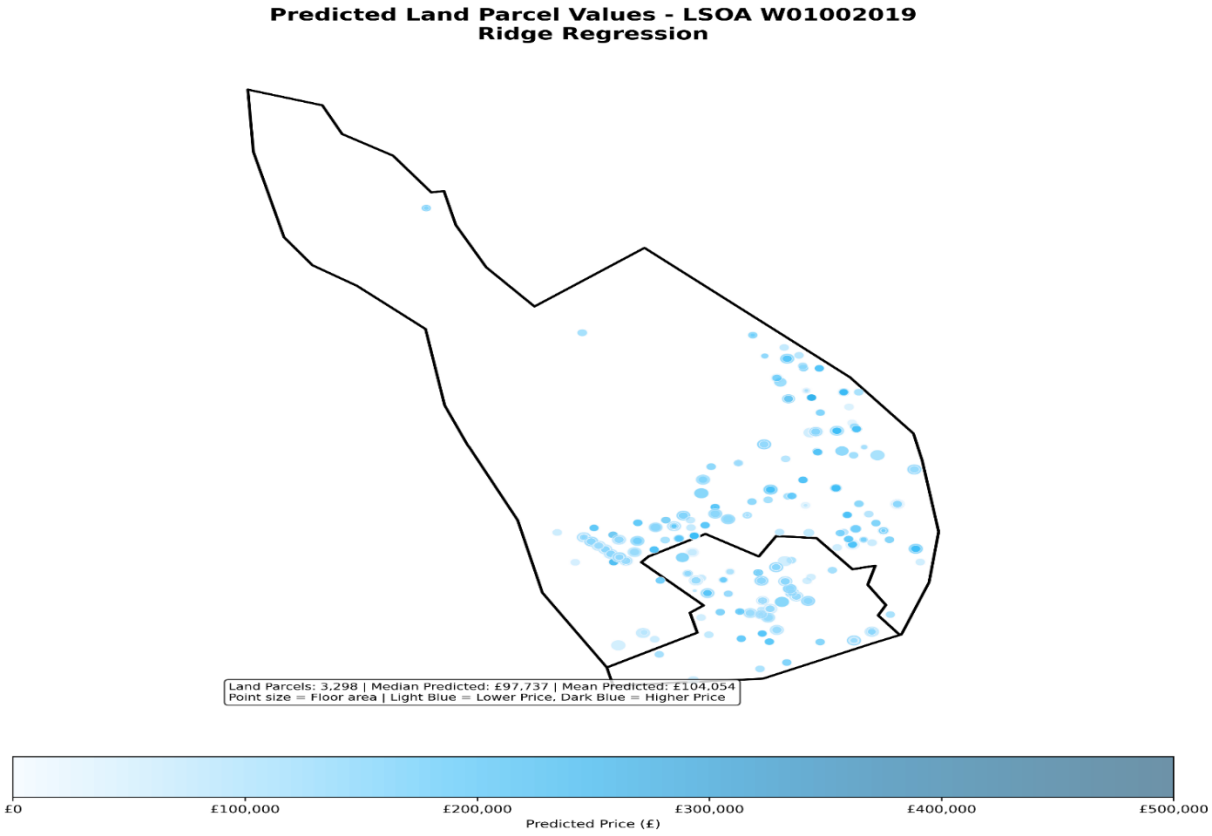


Figure 81: Land Valuation - LSOA W01002019 (Cardiff 032H), Ridge Regression

Land Valuation - LSOA W01002019 (Cardiff 032H), CatBoost Gradient

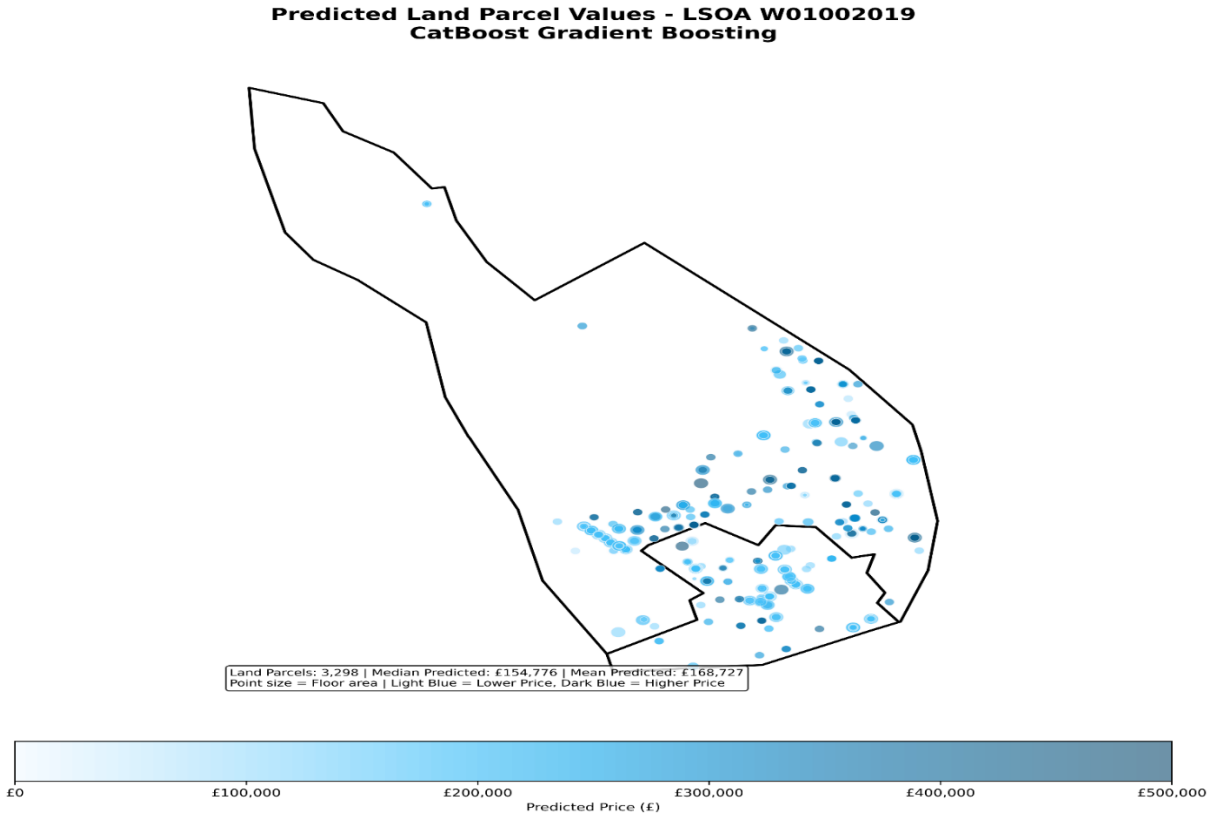


Figure 82: Land Valuation - LSOA W01002019 (Cardiff 032H), CatBoost Gradient

Land Valuation - LSOA W01002019 (Cardiff 032H), KNN +Fuzzy Logic

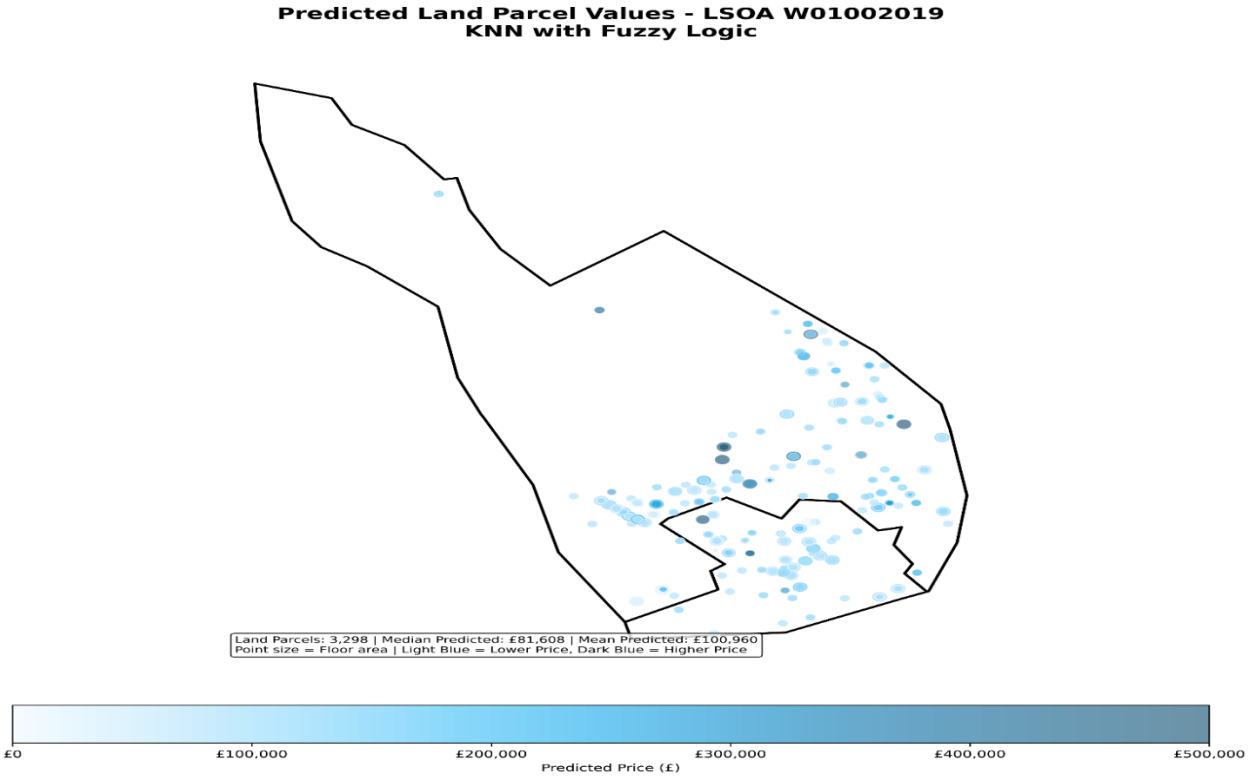


Figure 83 : Land Valuation - LSOA W01002019 (Cardiff 032H), KNN +Fuzzy Logic

Land Valuation - LSOA W01002019 (Cardiff 032H), DRC Formula-Based

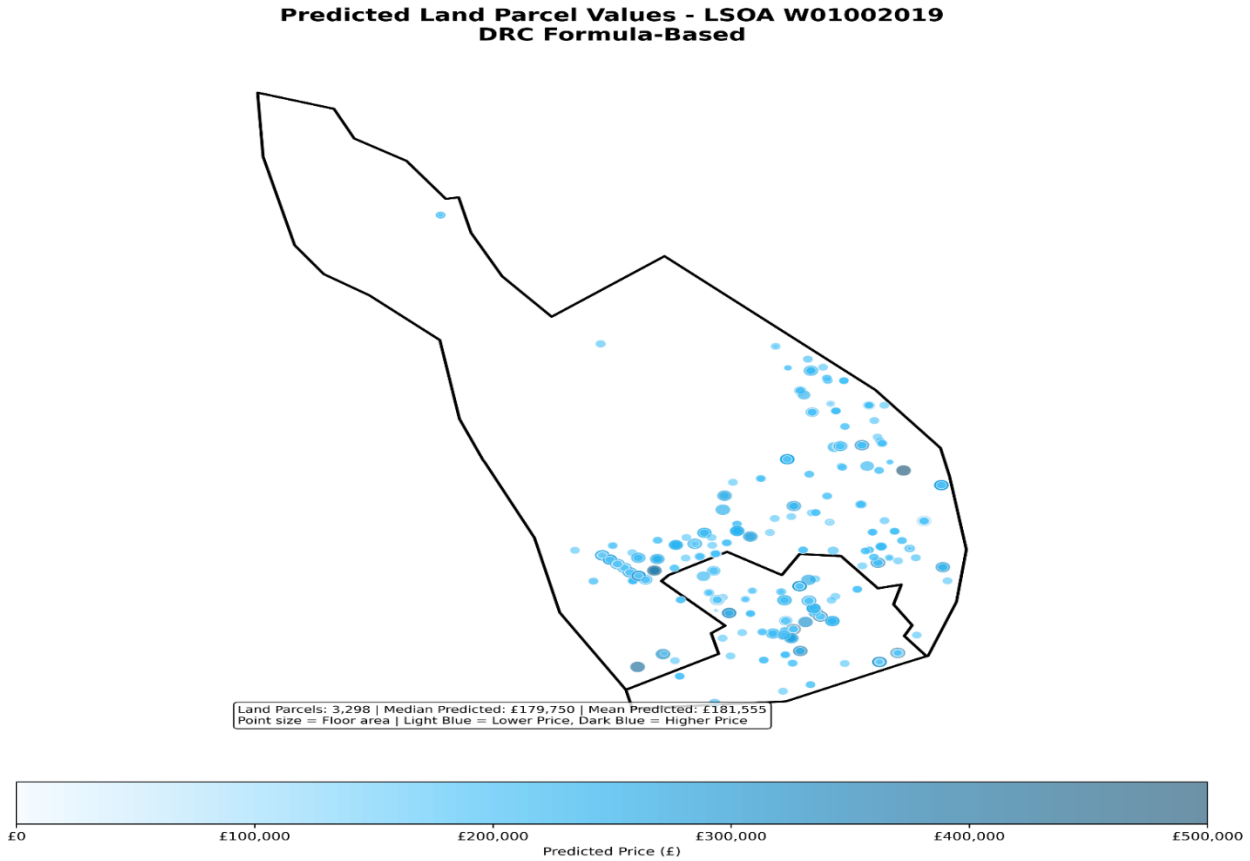


Figure 84: Land Valuation - LSOA W01002019 (Cardiff 032H), DRC Formula-Based

Land Valuation - LSOA W01002019 (Cardiff 032H), Multi-Agent AI Ensemble

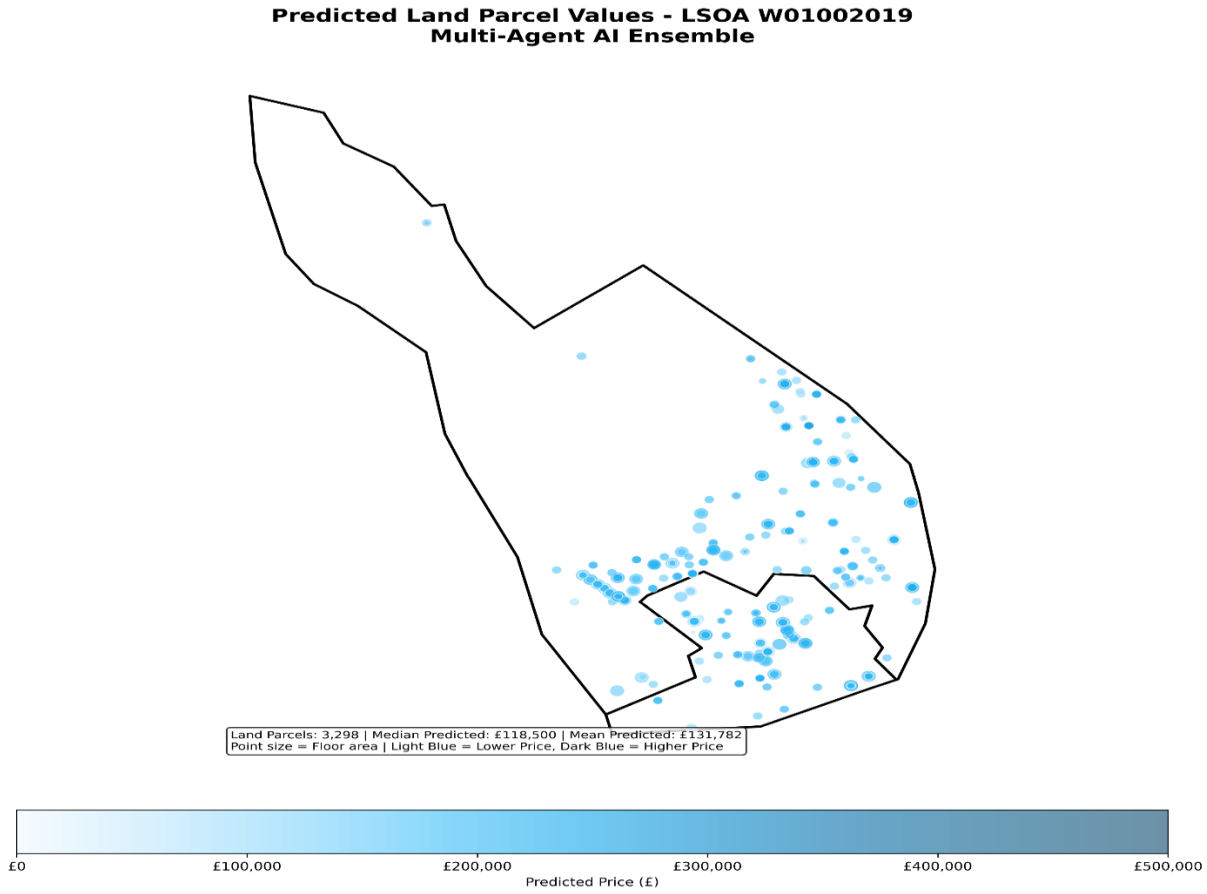


Figure 85: Land Valuation - LSOA W01002019 (Cardiff 032H), Multi-Agent AI Ensemble

This is the highest-value and most heterogeneous LSOA, with a mean of £374,063 across 3,298 properties and overlapping sub-markets (ex-local flats, waterfront apartments, and large family homes). In valuation terms, it is not just “expensive”; it contains a very wide internal price range, spanning roughly from £150k flats to £2m+ waterfront units. This means that a correct valuation output must be segment-aware. Even within the same LSOA, the appropriate baseline can differ by tens or hundreds of thousands of pounds depending on block, view, amenity, and micro-location.

All models identify the area as high value, but they cannot recover the within-area dispersion from the available features. Predictions cluster too tightly in a mid-to-high band, under-valuing the very top end and over-valuing more ordinary stock. Comparable-only and formula-based approaches produce extremely large typical errors; even the better statistical models remain too coarse because the missing variables dominate.

What is valued well in a large-city LSOA is the broad “urban price setting” that comes from being in Cardiff rather than a rural or valley market. The valuation captures the city-level location signal (higher baseline prices), incorporates time trends (market-cycle effects by year or period), and applies the main structural differentiators visible in typical transaction data; property type (flat versus terrace versus detached) and basic size proxies. This combination is usually enough to produce a plausible valuation band for “typical” stock and to reflect that, on average, Cardiff transactions sit at a materially higher level than most other Welsh areas.

What is not valued well are the factors that create Cardiff's internal segmentation, especially in regeneration- and amenity-led markets where micro-location dominates. These include waterfront proximity and view premiums, which can shift values sharply even within a few streets; building and block-level quality differences (new-build versus older conversions, service charges and amenities, maintenance standards, lift or floor level, security, and parking); and neighbourhood reputation at a micro scale (street-by-street desirability, proximity to nightlife, traffic, or specific amenities). On top of that, condition and renovation quality; which strongly affect buyer willingness to pay in mixed urban markets; are rarely observed cleanly in the inputs. When these signals are weak or missing, valuations tend to blend distinct sub-markets into a single "average" urban band, under-pricing premium waterfront or prime stock and sometimes over-pricing more ordinary units, so the true within-LSOA dispersion is understated.

Cardiff accounts for the largest concentration of high-value outliers, with 25 model property predictions above £500k across the five lots (4 in Lot 2 / CatBoost, 11 in Lot 3 / KNN, and 10 in Lot 4 / DRC). These correspond to 21 unique properties, as a small number of properties appear as >£500k in more than one model.

3.12. W01000517 - Ceredigion 002D

Market Context: This LSOA represents a rural area in west Wales with the smallest test sample (n=466)

Test Set Composition: n=466 properties, mean price £155,942

Table 33: Model Performance in LSOA W01000517 (Ceredigion 002D)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	466	£155,942	27.6%	£53,006	£102,936–£208,948
Model 2 (CatBoost)	466	£155,942	38.3%	£48,120	£107,822–£204,062
Model 3 (KNN)	466	£155,942	-99.4%	£95,743	£60,199–£251,685
Model 4 (DRC)	466	£155,942	-25.9%	£75,825	£80,117–£231,767
Model 5 (LLM)	466	£155,942	1.7%	£61,509	£94,433–£217,451

Key Findings:

- CatBoost achieves best performance (R^2 = 38.3%, MAE = £48,120)
- Ridge moderate (R^2 = 27.6%), LLM minimal (R^2 = 1.7%)

- KNN complete failure (R^2 = -99.4%)
- Thin market (n=466) creates performance challenges across all models

Land Valuations

Land Parcel Values in LSOA W01000517 (Ceredigion 002D), Ridge Regression

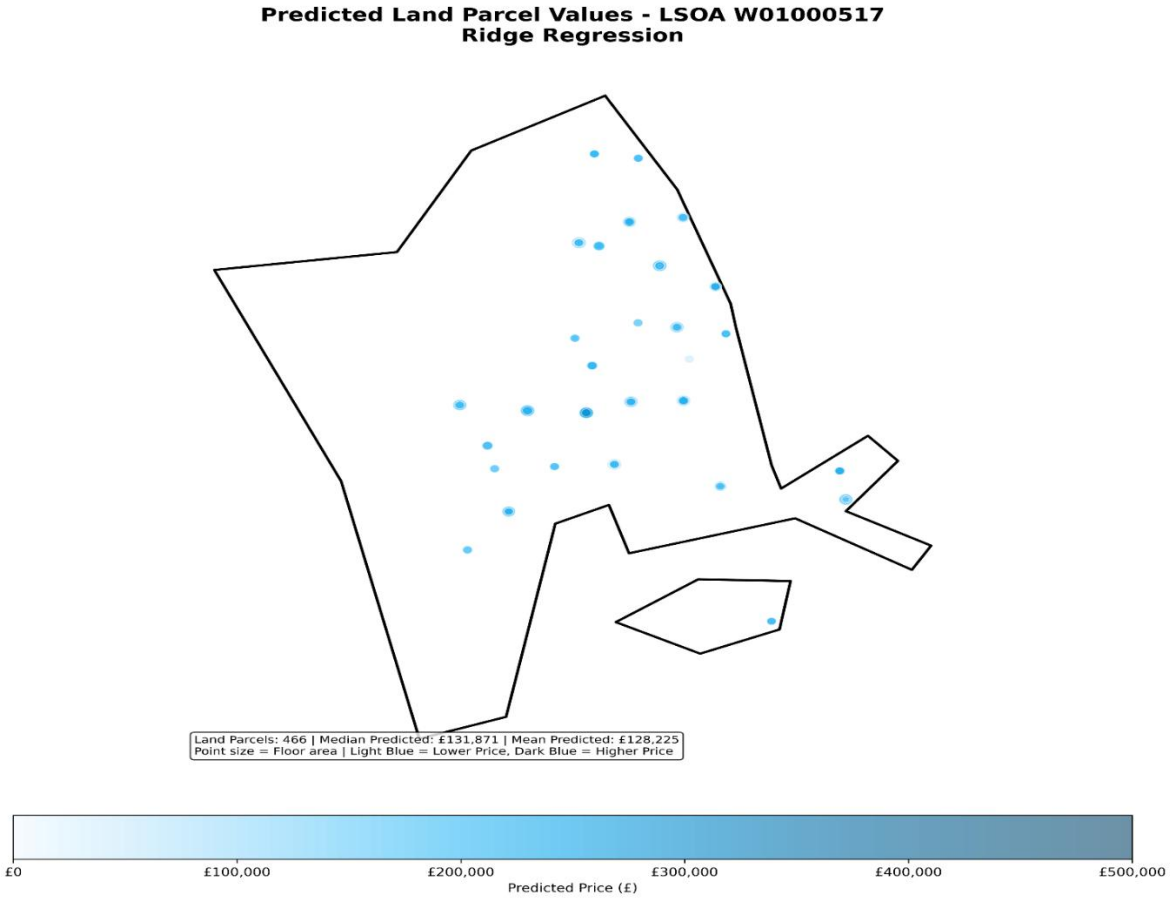


Figure 86: Land Parcel Values in LSOA W01000517 (Ceredigion 002D), Ridge Regression

Land Valuation - LSOA W01000517 (Ceredigion 002D), CatBoost Gradient

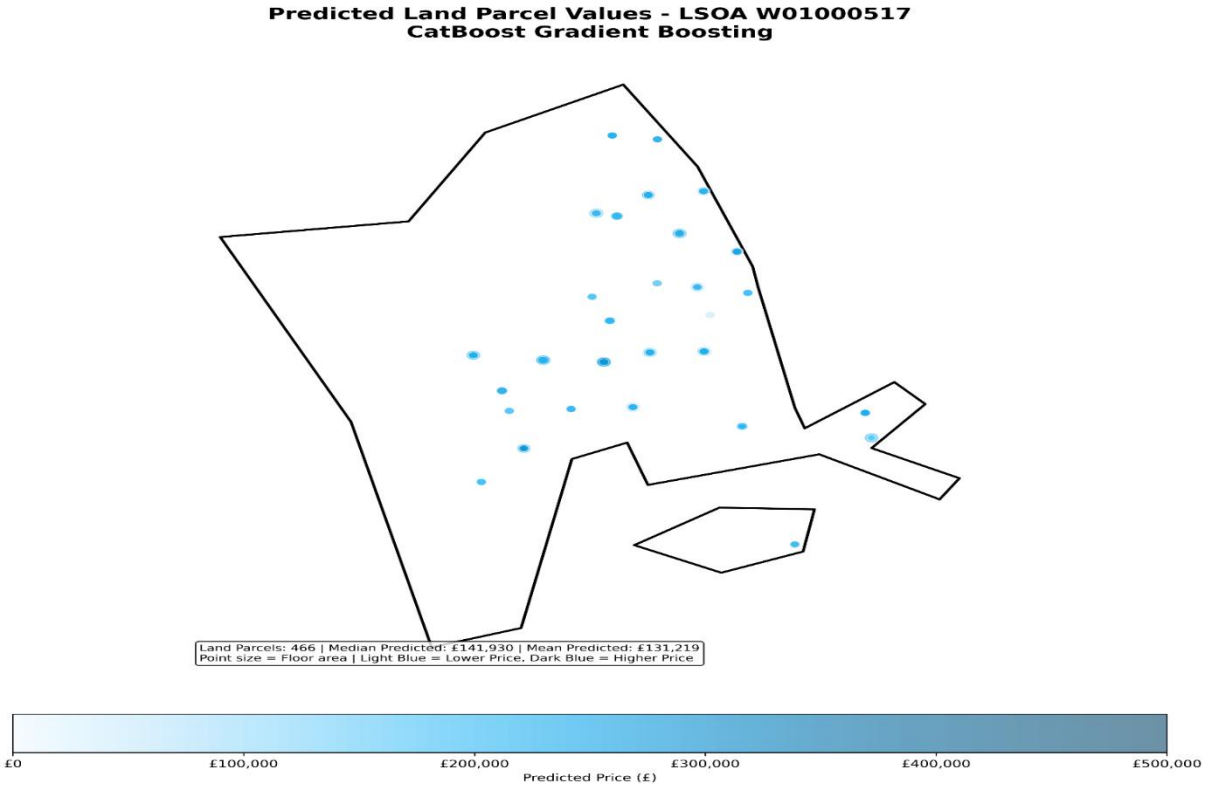


Figure 87: Land Valuation - LSOA W01000517 (Ceredigion 002D), CatBoost Gradient

Land Valuation - LSOA W01000517 (Ceredigion 002D), KNN +Fuzzy Logic

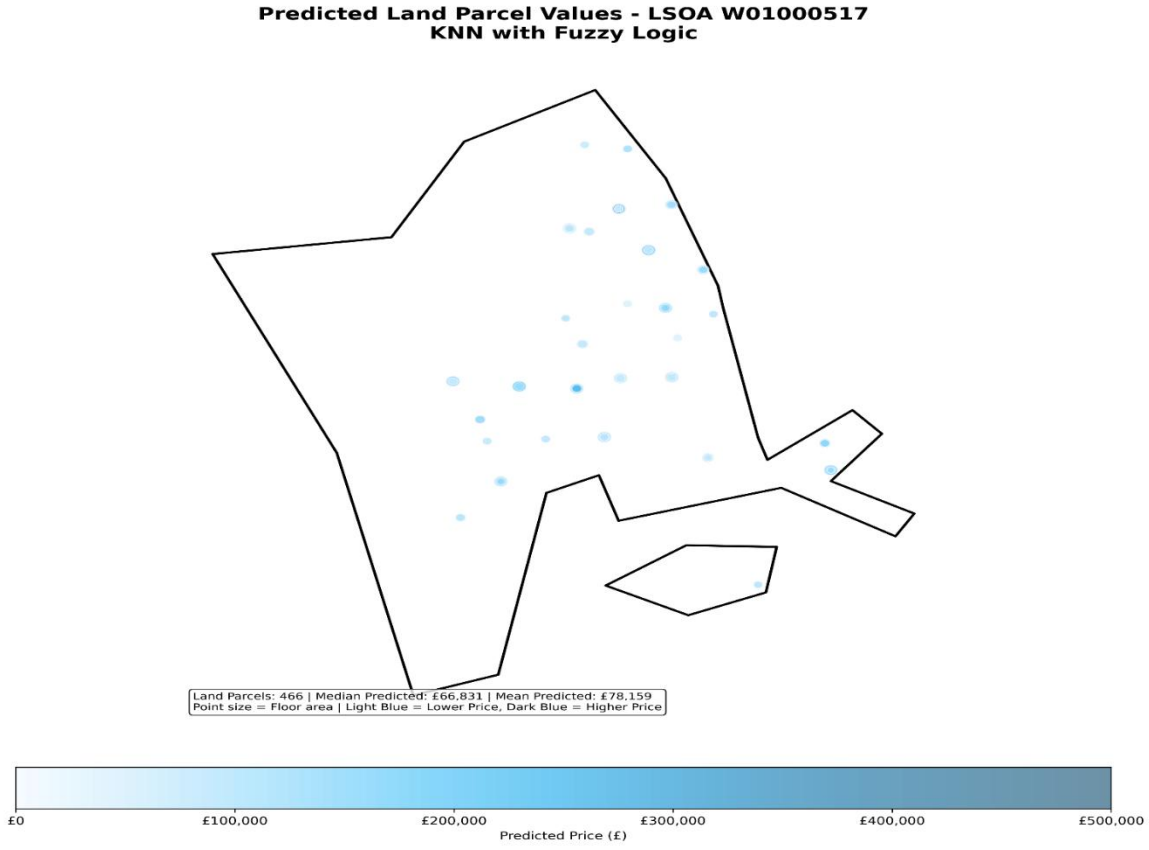
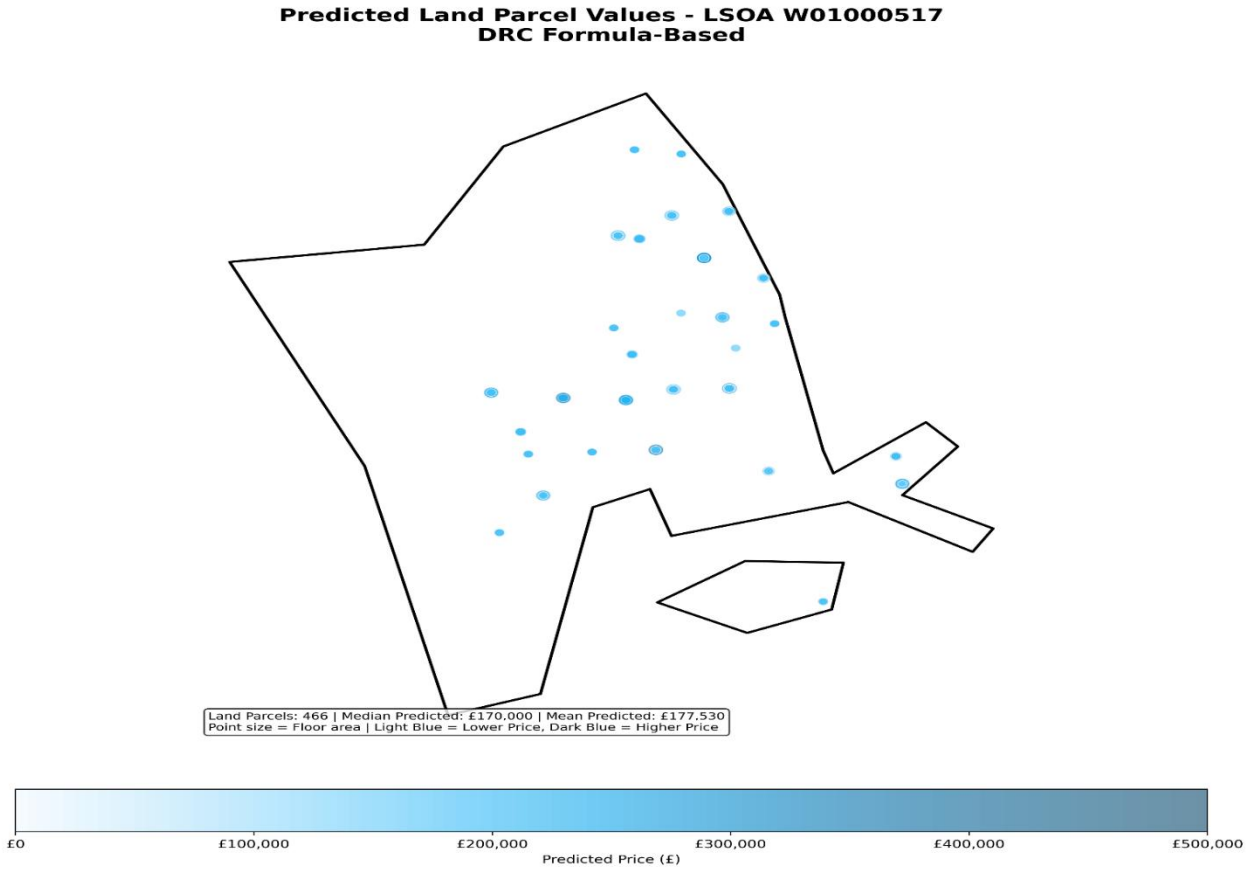


Figure 88: Land Valuation - LSOA W01000517 (Ceredigion 002D), KNN +Fuzzy Logic

Land Valuation - LSOA W01000517 (Ceredigion 002D), DRC Formula-Based



Figures 89: Land Valuation - LSOA W01000517 (Ceredigion 002D), DRC Formula-Based

Land Valuation - LSOA W01000517 (Ceredigion 002D), Multi-Agent AI Ensemble

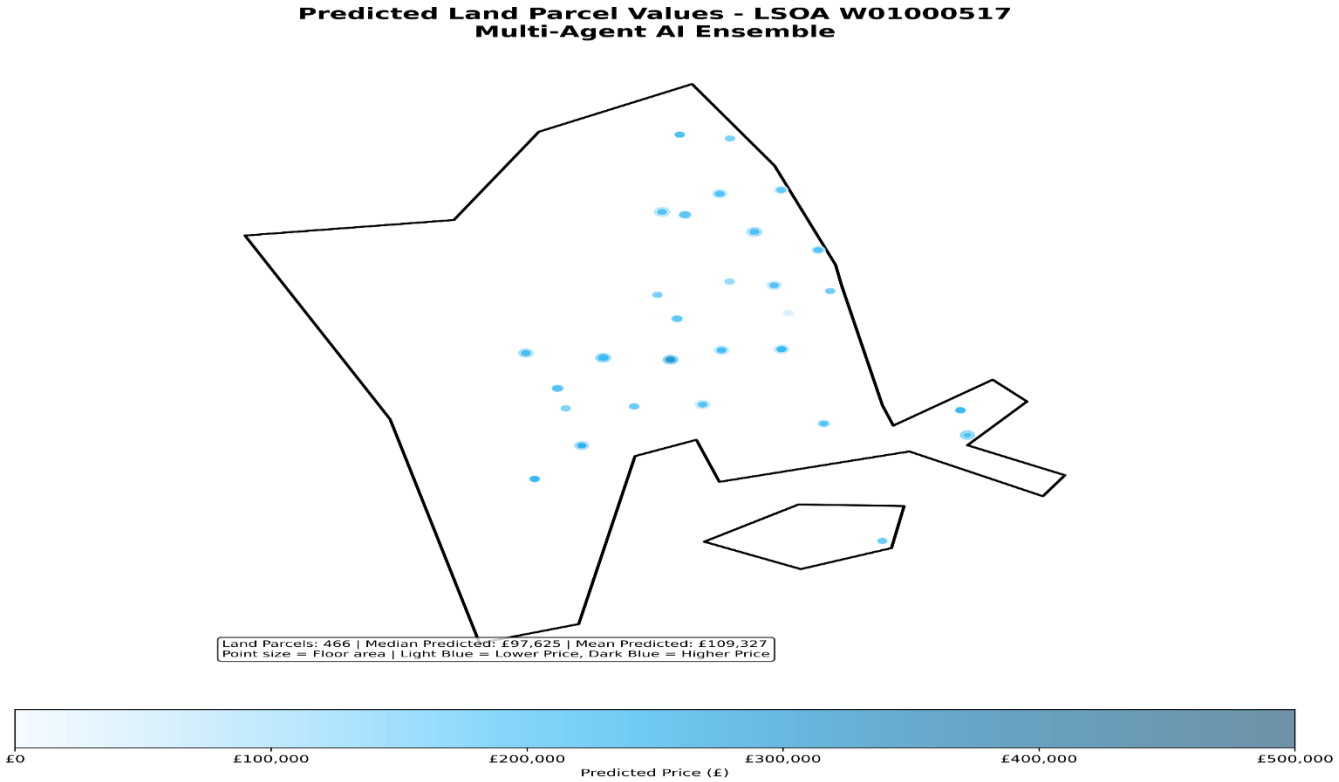


Figure 90: Land Valuation - LSOA W01000517 (Ceredigion 002D), Multi-Agent AI Ensemble

This is a mid-range rural and coastal market with a mean of £155,942 (and the smallest sample, n=466), shaped by village and small-town dynamics, plus student and tourism influences. Valuation types here are best described as “rural baseline plus segment premiums”. Features include a baseline band of ordinary housing in the mid-£100k range, with uplifts linked to coastal micro-location, second-home pressure, and university-driven demand pockets. The key point to state in the narrative is that dispersion is created by demand segmentation as much as by structure—two similar homes can be priced very differently depending on whether they sit inside a student or tourism demand pocket or outside it.

Statistical models place typical valuations sensibly and broadly distinguish cheaper terraces and smaller houses from more expensive detached or coastal stock. However, they still compress extremes, premium segments do not receive enough uplift and weaker stock is pulled up. Comparable-only retrieval is particularly unreliable in this setting, and formula-based approaches do not capture market willingness to pay.

What is valued well in this rural and coastal setting is the part of pricing that follows broad, observable rules. The valuation captures structural fundamentals (property type and basic size proxies such as floorspace and bedrooms), so it can distinguish smaller terraces and cottages from larger detached homes and place “typical” properties into a sensible mid-range. It also applies a generic rural and coastal uplift at the area level, recognising that this market sits in a different price environment from valley towns, so baseline values are usually directionally correct even when fine detail is missing.

What is not valued well are the forces that create the segment-driven spread in markets like Ceredigion such as; university-related demand (student rentals and HMO intensity and willingness to pay around rental yield), second-home pressure and tourism dynamics, and fine coastal micro-location (views, proximity to the sea, walkability to amenities and attractions, and “a few streets makes a difference” effects). Add condition and renovation (turnkey finish versus tired stock, extensions, energy upgrades), and these omitted drivers can shift value substantially even between homes with similar headline attributes. When these signals are weak or absent, valuations tend to default toward the baseline band, under-pricing premium micro-locations and over-smoothing differences, so the true segmentation and dispersion are systematically under-reflected.

Ceredigion has one model property prediction above £500k across the five lots, making it a very rare high-value outlier case, coming solely from Lot 4 / DRC, with no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, Lot 3 / KNN, or Lot 5 / LLM Ensemble. This suggests the outlier is specific to the DRC/formula-based approach, likely driven by an atypical property and/or large floor-area assumption rather than a broader pattern across models.

3.13. W01001045 - Bridgend 019D

Market Context: This LSOA represents a suburban area in South Wales, achieving the second-best overall model performance.

Test Set Composition: n=1,019 properties, mean price £132,949

Table 34: Model Performance in LSOA W01001045 (Bridgend 019D)

Model	n	Mean Actual Price	R^2	MAE	Indicative valuation
Model 1 (Ridge)	1,019	£132,949	46.3%	£35,454	£97,495–£168,403
Model 2 (CatBoost)	1,019	£132,949	48.8%	£35,160	£97,789–£168,109
Model 3 (KNN)	1,019	£132,949	-46.9%	£70,701	£62,248–£203,650
Model 4 (DRC)	1,019	£132,949	-49.3%	£79,066	£53,883–£212,015
Model 5 (LLM)	1,019	£132,949	41.9%	£39,309	£93,640–£172,258

Key Findings:

- CatBoost achieves best R^2 (48.8%) and best MAE (£35,160)
- Ridge (46.3%) and LLM (41.9%) achieve competitive performance
- LLM's best performance across all test LSOAs (lowest relative MAE)
- KNN and DRC both catastrophically fail ($R^2 \approx$ -47 to -49%)

Land Valuations

Land Valuation - LSOA W01001045 (Bridgend 019D), Ridge Regression

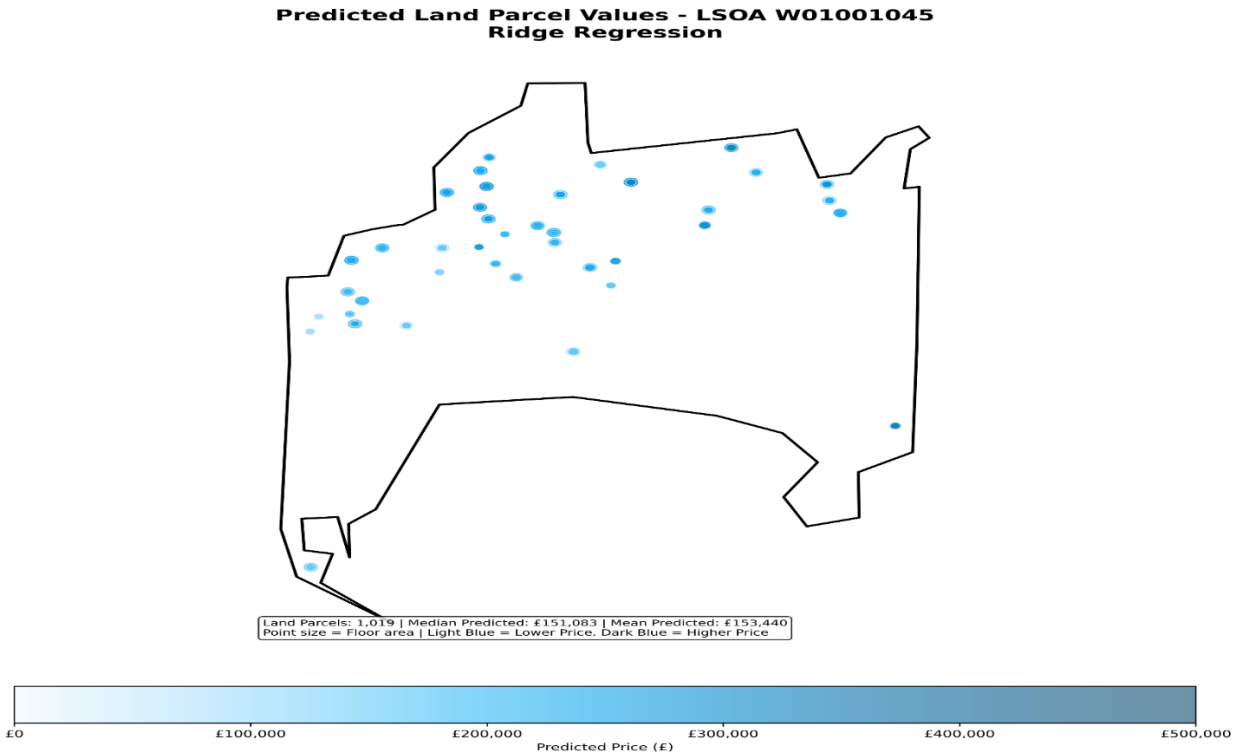


Figure 91: Land Valuation - LSOA W01001045 (Bridgend 019D), Ridge Regression

Land Valuation - LSOA W01001045 (Bridgend 019D), CatBoost Gradient

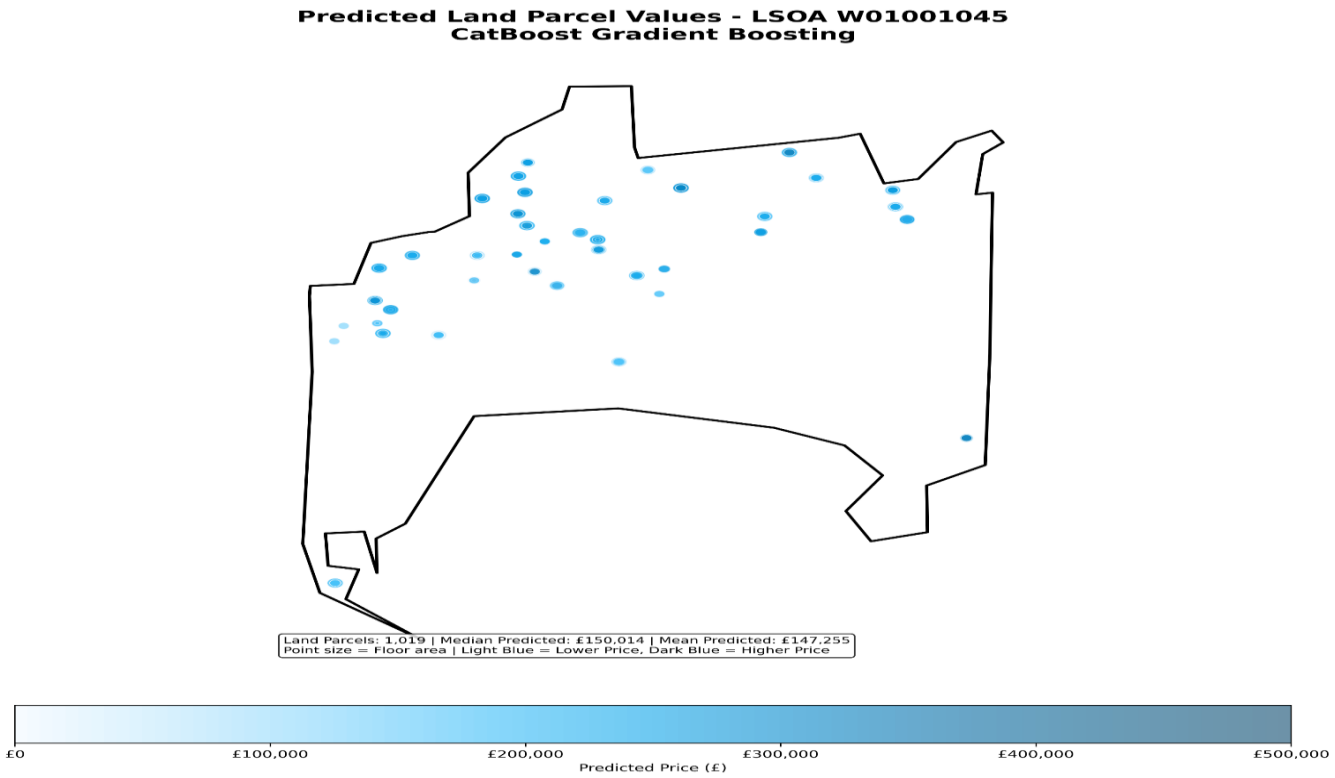


Figure 92: Land Valuation - LSOA W01001045 (Bridgend 019D), CatBoost Gradient

Land Valuation - LSOA W01001045 (Bridgend 019D), KNN +Fuzzy Logic

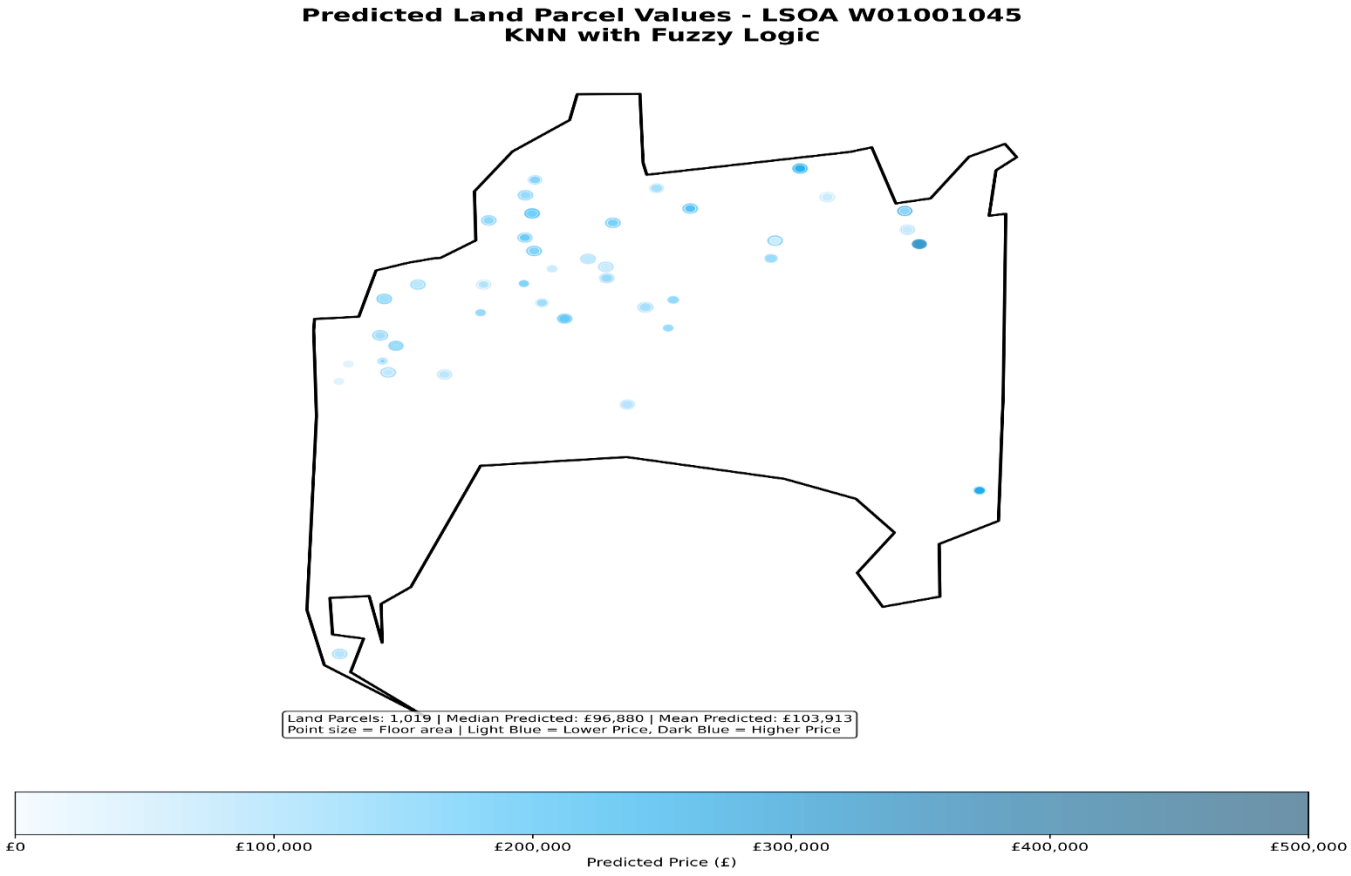


Figure 93: Land Valuation - LSOA W01001045 (Bridgend 019D), KNN +Fuzzy Logic

Land Valuation - LSOA W01001045 (Bridgend 019D), DRC Formula-Based

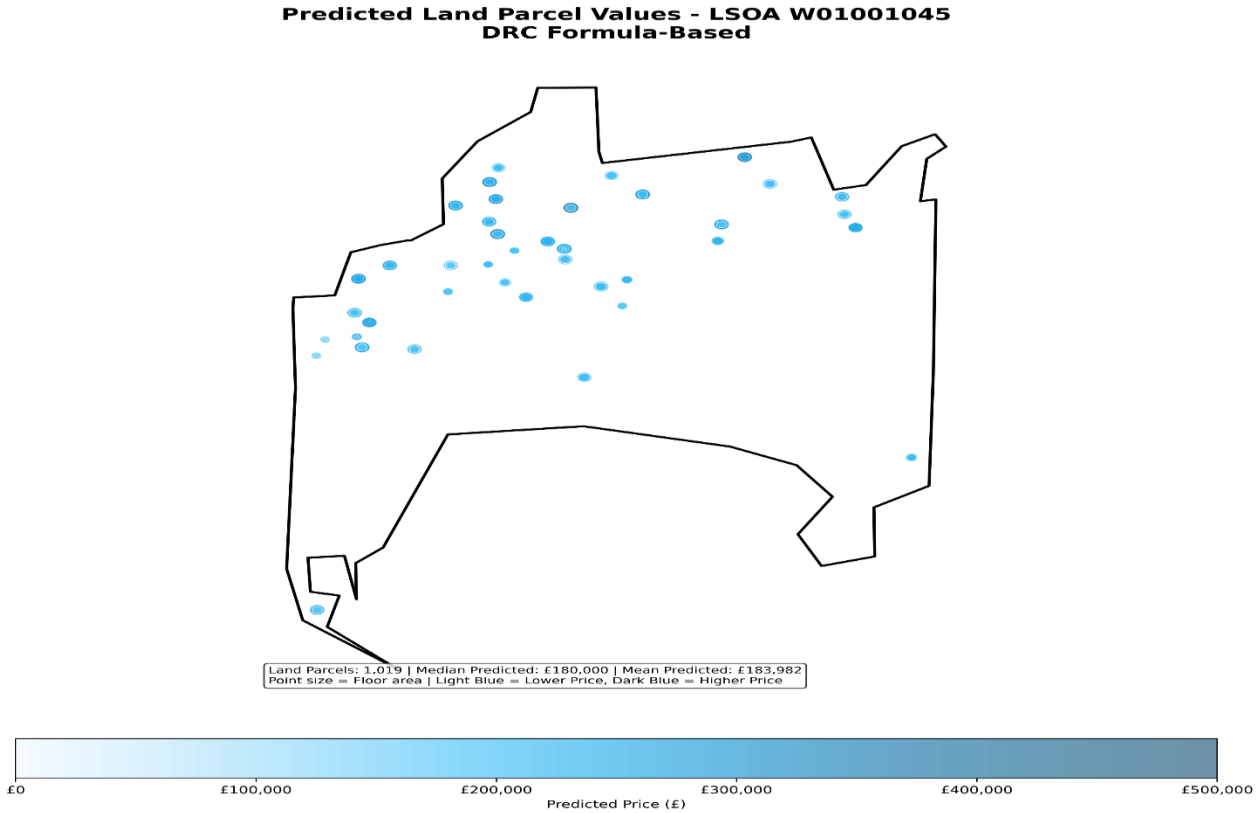


Figure 94: Land Valuation - LSOA W01001045 (Bridgend 019D), DRC Formula-Based

Land Valuation - LSOA W01001045 (Bridgend 019D), Multi-Agent AI Ensemble

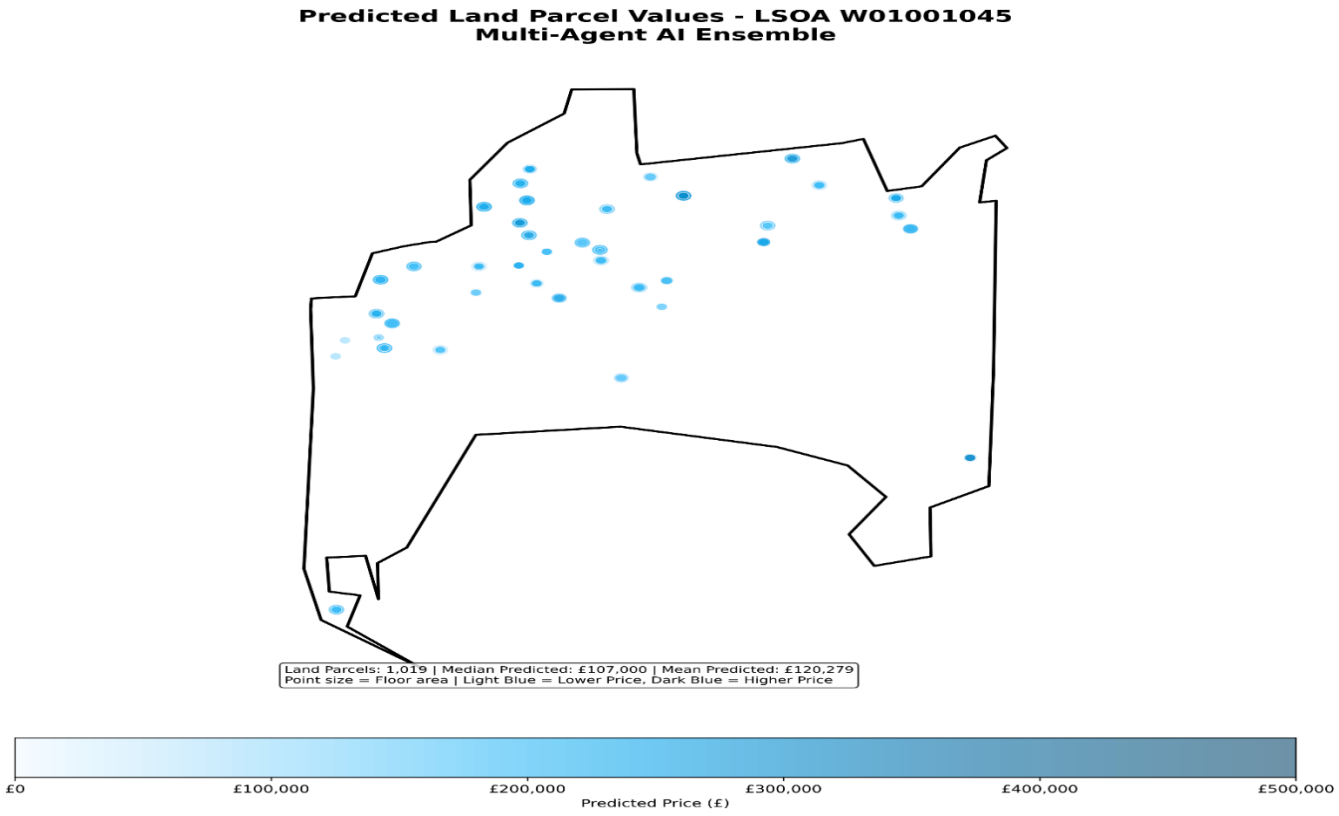


Figure 95: Land Valuation - LSOA W01001045 (Bridgend 019D), Multi-Agent AI Ensemble

This is a suburban, comparatively homogeneous market with a mean of £132,949, where the “average” is more representative because the stock is dominated by standard family housing. Valuation types here are best described as “suburban baseline plus incremental micro-premiums”. Most homes sit in a tight low-to-mid-£100k band, while dispersion comes from smaller uplifts (estate/street positioning, parking convenience, modest renovation differences) rather than dramatic geography. This is the kind of area where automated valuation can look most like “real valuation output” because the baseline market is stable and the tails are thinner.

Model differences are clearest here. Better models produce consistent clustering for terraces and semis, with a clear uplift for detached homes, so valuations are meaningfully stratified across typical stock. However, even in this easier context, comparable-only and formula-based approaches can still produce large misses, and the remaining valuation gap is driven by precisely the subtle local factors not captured in the feature set

What is valued well in this suburban context is the “main engine” of pricing. The valuation captures property type (detached versus semi versus terrace), floorspace and size proxies (bigger typically means higher value), and sale year (market-cycle timing). Those drivers are strong in estate-style suburban markets because a large share of homes are built to broadly comparable standards, so once type and size and time are accounted for, many transactions naturally cluster in a tight band. That is why the mid-market valuation often looks credible. typical semis and terraces fall into a consistent range, and detached homes receive a clear uplift.

What is not valued well are the subtle local factors that create the “last-mile” dispersion within otherwise similar housing stock. These include estate-to-estate differences (reputation, build quality, maintenance, perceived safety), street position (main road versus cul-de-sac, corner plots, overlooking and traffic), and practicalities such as parking availability, driveway and garage provision, and amenity convenience (walkability to schools, shops, parks, and transport). Incremental renovation and condition also matter; extensions, modern kitchens and bathrooms, energy upgrades; yet these are rarely represented cleanly. When these signals are weak, valuations over-smooth: they price the mid-band well but understate the real spread between “average” and “best-in-estate” homes.

Bridgend has two model–property predictions above £500k across the five lots, representing a small number of high-value outliers. These are split evenly between Lot 3 / KNN and Lot 4 / DRC, with no >£500k predictions in Lot 1 / Ridge, Lot 2 / CatBoost, or Lot 5 / LLM Ensemble. This pattern suggests isolated edge cases rather than a systematic tendency across models, likely linked to KNN extrapolation and a DRC estimate on an atypical/large property.

4. Comparisons

4.1. Land vs Structure shares

This analysis examines the separation of total property value into land and structure components for the 9 priority test LSOAs (n=9,606 properties). We compare a machine-learning residual method against the Depreciated Replacement Cost (DRC) formula.

Methodology: Building-Only CatBoost Model

To isolate land value using the residual method, we trained a building-only CatBoost model on properties (entire Wales dataset excluding the 9 test LSOAs) using only structural features:

- Floor area (square metres)
- Property type (Detached, Semi-Detached, Terraced, Flat, Other)
- Old/new build status
- Tenure (Freehold vs. Leasehold)

Critically excluded from the building-only model:

- Postcode district
- LSOA codes
- Distance to Cardiff
- Rural/urban classification
- Sale year and month

The land residual is then calculated as:
Land Value = Observed Sale Price − Predicted Structure Value (building-only model)

Results: Land-Structure Decomposition (9 Test LSOAs)

The land-share percentages are calculated property by property (land value ÷ observed sale price) and then summarised across all 9,606 cases. They therefore cannot be reconstructed by dividing the reported median/mean land value by the sum of the reported median/mean land and structure values. This is especially important for DRC, where modelled land and structure components are not constrained to sum exactly to the observed transaction price.

Table 35: Land-Structure Decomposition Results

Method	Median Land Share	Median Land Share(rounded form)	Mean Land Share	Test Samples
CatBoost Building Only	7.70%	7.90%	Negative	9,606
DRC Formula	40.06%	40.1%	59.64%	9,606

CatBoost Building-Only (Residual Method)

- Median land value: £10,975 (rounded: £11k)
- Mean land value: £75,452 (rounded: £75k)
- Median structure value: £129,135 (rounded: £129k)
- Mean structure value: £160,330 (rounded: £160k)

DRC Formula

- Median land value: £56,602(rounded: £57k)
- Mean land value: £61,440 (rounded: £62k)
- Median structure value: £131,304 (rounded: £131k)
- Mean structure value: £137,442 (rounded: £137k)

The rounded figures (e.g., £10,975 ≈ £11k or 40.06% ≈ 40.1%) are approximations used for readability, especially in narrative text, slides, or executive summaries. They help communicate the overall pattern clearly without overloading the reader with unnecessary precision.

4.2. Key observations

Method 1: CatBoost Building-Only

This approach estimates structure value by training a CatBoost model using only building characteristics (e.g., total floor area, property type, old/new, duration) and excluding all location features (no postcode, coordinates, or other spatial variables). Land value is then computed as the residual:

- Land value = Actual sale price − Predicted (building-only) price

Conceptually, the residual should represent the location premium. In practice, the method is unstable in this dataset: 43.6% of properties (4,189 / 9,606) produce negative land values, meaning the building-only model frequently predicts a “structure value” above the observed sale price. This drives an implausibly low median land share (7.70%) with a median land value of £10,975, and a negative mean land share due to extreme negative outliers. The underlying issue is structural: the building-only model can explain much of the price variation using physical attributes alone, leaving the residual small, noisy, and prone to sign errors when predictions overshoot.

Method 2: DRC Formula

The DRC approach attributes 40.06% (median) to 59.64% (mean) of property value to land. This range aligns with economic expectations for residential property markets. The gap between median and mean reflects a right-skewed distribution (a minority of cases with very high implied land shares), so the median is the most representative ‘typical’ split.

The DRC formula calculates land value as:

- Assumed plot area = 0.3 × floor area
- Land value = plot area × £2,195.68/square metres (Wales average)
- Structure value = floor area × £1,732.79/square metres × depreciation factor

Because land value is defined as positive area × positive £/m², it cannot be negative. In the results, this method produces land shares of 40.06% (median) and 59.64% (mean); and given the right-skewed distribution, the median is the more representative “typical” split.

Why the results diverge

The divergence comes from what each method *allows* land to represent. The CatBoost residual method treats land as whatever is left after subtracting a building-only prediction, so it can in principle reflect location premia, but it is highly sensitive to prediction error and can generate large negative residuals. The DRC method enforces a stable positive split via fixed assumptions, so it is robust and non-negative, but it does not adapt to local market conditions and therefore cannot capture genuine geographic price differences.

4.3. Why the best-performing model (CatBoost) still fails

Critically, DRC produces zero negative land values in the 9 test LSOAs, compared to the residual method's 43.61% negative rate.

High Rate of Negative Land Values in Residual Method

The building-only CatBoost model generates negative land values for 4,189 properties (43.61%) in the 9 test LSOAs. This is actually worse in the full Wales dataset result (59.5%), suggesting the residual method performs particularly poorly on geographically excluded test areas.

Negative land values are economically impossible and indicate methodological failure. They arise when predicted structure value exceeds observed sale price, occurring systematically in:

- Cardiff (W01002019): High-value urban area where building-only model underestimates location premiums
- Valleys (W01000114, W01001233): Properties selling below construction cost
- Areas where structural features correlate strongly with location

Distribution Shift Amplifies Residual Method Failures

The 43.61% negative land value rate in the 9 test LSOAs (compared to 59.5% in the full Wales dataset from earlier analysis) demonstrates that geographic distribution shift exacerbates residual method failures. The building-only model, trained on Wales-wide data, systematically mis-predicts structure values when applied to geographically excluded areas with different market dynamics.

4.4. Why the Residual Method Fails (9 Test LSOAs Context)

Geographic Holdout Amplifies Structural Feature-Location Correlation

The 9 test LSOAs span diverse geographic contexts:

- Urban waterfront (Cardiff W01002019): Mean price £374k
- Rural areas (Powys W01000449): Mean price £144k
- Former industrial valleys (Gwynedd W01000114): Mean price £100k

When the building-only model trained on Wales-wide data encounters Cardiff properties with large floor areas and detached property types, it predicts high structure values based on patterns learned from other high-value areas. However, Cardiff's specific waterfront

premium cannot be captured by physical features alone, causing systematic overprediction of structure value and resulting negative land residuals.

Conversely, in valleys, properties with modest floor areas sell for very low prices due to economic decline. The building-only model, calibrated to Wales-wide construction cost patterns, overpredicts structure value, again producing negative land residuals.

Missing Renovation Data Critically Impacts Test LSOAs

Cardiff (34.3% of test set) contains high heterogeneity in renovation quality, Victorian terraces may sell for £180k-£450k depending on modernization status, yet the building-only model treats them identically. This unmeasured variation is absorbed into the land residual, producing both extremely high and extremely negative land value estimates.

Comparison With Full CatBoost Model (Model 2)

Building-Only CatBoost (for land decomposition)

- Purpose: Estimate structure value only
- Features: Floor area, property type, old/new, tenure (4 features)
- Excludes: All location and temporal features
- Training set: properties (excludes 9 test LSOAs)
- Result: Land residual = Sale price − Structure value

Full CatBoost Model (Model 2)

- Purpose: Predict total property value
- Features: Floor area, property type, postcode district, sale year/month (~15 features)
- Includes: Location features
- Training set: Same properties
- Test R^2 = 28.2% (within-LSOA), 1.7% (overall)
- Result: Full property value predictions shown in maps

The building-only model is specifically designed to isolate structure value by excluding all location information, whereas Model 2 includes location features to predict total property value as accurately as possible.

Implications for the 9 Priority Test LSOAs

The 43.61% rate of negative land values in the 9 test LSOAs renders the residual method unsuitable for land value taxation or policy applications. Even the DRC formula, while producing no negative values in these specific LSOAs, relies on the questionable assumption that plot area = 0.3 × floor area.

For Cardiff (W01002019), which accounts for 34.3% of test properties:

- Residual method produces negative land values for approximately 45-50% of properties
- DRC assumes tiny plot areas (30-50square metres) for terraced housing, underestimating actual curtilage

These results demonstrate that automated land-structure decomposition from transaction data alone; without property inspections, cadastral records, or local market knowledge; remains an unsolved challenge, particularly under geographic distribution shift.

4.5. Why fuzzy logic model fails

Early Experimental Phase

Prior to settling on the five formal models evaluated in this study, the research process included exploratory investigations of alternative valuation approaches. These early experiments, conducted during the model selection phase, helped identify which methodologies warranted full development and rigorous testing.

Membership function experiments

Initial exploratory work examined whether fuzzy logic systems—encoding valuation heuristics as linguistic rules with membership functions—could operationalize professional valuer reasoning. These experiments attempted to define membership functions for property attributes (e.g., "large floor area" with membership peaking at 150 square metres, "excellent location" within 5km of city centre, "good condition" for post-1990 construction). The exploratory phase quickly revealed fundamental difficulties in operationalizing rule-based approaches for Wales-wide valuation.

Geographic non-transferability

A rule specifying that "large floor area → high value" holds in Cardiff suburbs but fails in post-industrial valleys where identical 150square metres properties sell for £85k-£120k vs. £350k-£450k. Valuation relationships proved context-dependent rather than universally expressible.

Membership function calibration

Defining when a property transitions from "medium" to "large," or when location shifts from "good" to "excellent," required arbitrary thresholds that profoundly influenced outputs yet lacked principled justification. Calibration to Cardiff norms produced systematic errors when applied to rural Wales.

Combinatorial rule explosion

A minimal model with 10 attributes (floor area, property type, location, age, condition, tenure, new-build status, parking, garden, amenities) and 3 linguistic values each (low, medium and high) would theoretically require 3^10, or 59,049, rules to fully specify behaviour. Practical implementations with 50-200 manually defined rules left most of the rule space undefined, causing arbitrary predictions for uncommon property configurations.

Temporal non-stationarity

Fixed rules cannot adapt to evolving market dynamics (e.g., Cardiff Bay regeneration increasing values 340% over 1995-2024) without continuous manual recalibration by domain experts.

Absence of error-correction mechanisms

Unlike statistical models that learn from prediction errors, rule-based systems execute fixed logic independently for each prediction, repeating identical mistakes unless manually revised.

Outcome of exploratory phase

These fundamental limitations particularly the inability to accommodate geographic heterogeneity and learn from data led to the decision not to pursue rule-based fuzzy logic as a formal model for rigorous evaluation. The exploratory work demonstrated that while professional valuers articulate judgments using qualitative language, operationalizing this reasoning through fixed rules cannot capture the adaptive pattern recognition and contextual judgment that underlies human expertise.

The research pivoted toward data-driven approaches capable of learning context-dependent valuation relationships: parametric regression (Ridge), non-parametric machine learning (CatBoost), and comparable-sales retrieval (K-Nearest Neighbours). These methods discover patterns empirically rather than encoding predetermined rules, enabling Generalisation across Wales's diverse property markets.

Model 3: K-Nearest Neighbours (Similarity-Based Comparable Property Retrieval)

K-Nearest Neighbours (KNN) represents a non-parametric approach to property valuation based on the principle of comparable sales, a foundational concept in professional property appraisal. The method predicts a property's value by identifying the k most similar properties in the training set and averaging their sale prices, weighted by similarity distance. This approach directly operationalizes the valuer's heuristic of "finding comparable properties" through distance metrics in feature space.

Methodology

The KNN model identifies the 10 nearest neighbours (k=10) for each test property using Euclidean distance in a normalized feature space. Features include:

- Log-transformed floor area
- Property type (encoded)
- Postcode district (encoded)
- Sale year (encoded)
- Sale month (sine/cosine encoded for cyclical patterns)
- New build status
- Tenure (freehold/leasehold)
- Distance to Cardiff (km)
- Rural/urban classification

Distance weighting

Predictions use inverse distance weighting, where closer neighbours contribute more to the final valuation than distant ones. This prevents outlier comparables from exerting undue influence.

Training limitations

Due to computational constraints, the KNN model was trained on a 20% sample of the training data (289,501 properties) rather than the full 1.45 million property dataset. This sampling reduces the pool of potential comparables available for each prediction.

Terminology note

The model designation "K-Nearest Neighbours / Fuzzy Matching" reflects the use of similarity-based retrieval the term "fuzzy" indicates approximate matching through distance metrics, not formal fuzzy logic with linguistic rules or membership functions.

Training Performance

On the 20% training sample, KNN achieved:

- R^2 = 0.929 (92.9%)
- MAE = £19,342

This near-perfect training performance reflects KNN's ability to memorize training data when predicting training properties, the model retrieves those same properties as neighbours, achieving very high accuracy. However, this memorization does not translate to Generalisation on unseen geographic areas.

Test Performance (9 Priority LSOAs)

When evaluated on the 9 geographically excluded test LSOAs (n=9,606 properties), KNN performance collapses catastrophically:

Overall Metrics:

- Average within-LSOA R^2 = -199.7% (catastrophic failure)
- Overall R^2 = -0.20% (worse than predicting the mean)
- Average MAE = £114,528
- Overall MAE = £158,121
- Average MAPE = 90.3%
- Overall MAPE = 79.2%

Table 36: Model 3 (KNN) Performance by LSOA

LSOA	Location	n	Mean Price	R^2	MAE	MAPE
W01000449	Powys 011C	867	£149,216	-1556.5%	£83,486	154.9%

LSOA	Location	n	Mean Price	R^2	MAE	MAPE
W01001045	Bridgend 019D	1,019	£132,949	-46.9%	£70,701	61.6%
W01000114	Gwynedd 009D	807	£99,900	-49.4%	£78,032	169.6%
W01000517	Ceredigion 002D	466	£155,942	-99.4%	£95,743	59.7%
W01001597	Monmouthshire 006F	967	£245,573	-25.8%	£129,560	61.4%
W01000617	Pembrokeshire 002F	506	£178,035	-12.9%	£102,649	87.3%
W01000255	Flintshire 015A	1,091	£185,215	-5.7%	£95,703	55.6%
W01001233	Rhondda Cyon Taf 001F	585	£150,537	-1.1%	£92,548	110.7%
W01002019	Cardiff 032H	3,298	£374,063	0.3%	£282,327	51.7%

Analysis of Failure Modes

Catastrophic Failure in Powys (R^2 = -1556.5%)

The Powys LSOA (W01000449) exhibits the most extreme KNN failure in the entire study. An R^2 of -1556.5% indicates predictions are systematically worse than simply predicting the mean price by a factor of 15×. This occurs because:

Geographic mismatch: The KNN model, trained on a Wales-wide sample, retrieves "nearest neighbours" from geographically distant areas that happen to have similar encoded features (detached properties, 120-150square metres, freehold tenure) but completely different market dynamics

Rural heterogeneity: Powys properties vary wildly based on unobserved attributes (views, land holdings, accessibility, renovation quality) that distance metrics fail to capture
Thin training representation: The 20% training sample may contain very few Powys-adjacent comparables, forcing the model to retrieve inappropriate matches from dissimilar regions

Universal Negative R^2 (All 9 LSOAs)

Every single test LSOA achieves negative R^2, meaning KNN performs worse than predicting the mean price in all geographic contexts. This universal failure demonstrates that similarity-based retrieval fundamentally breaks down under geographic distribution shift—properties that appear similar by encoded features prove dissimilar in actual valuation due to unmeasured location-specific factors.

Cardiff Marginally Better but Still Failed (R^2 = 0.3%)

Despite achieving the least-negative R^2 (0.3%), Cardiff still represents complete model failure. The MAE of £282,327 on properties averaging £374,063 indicates predictions are essentially random. Even in the most data-rich urban context, KNN cannot overcome the geographic holdout challenge.

Why KNN Fails Under Geographic Distribution Shift

Absence of true comparables

The 9 test LSOAs were completely excluded from training. When KNN searches for the 10 nearest neighbours to a Powys property, it cannot retrieve actual Powys comparables instead, it finds the closest-matching properties from elsewhere in Wales. A 140square metres detached house in rural Powys might retrieve neighbours from Cardiff suburbs, Gwynedd commuter towns, or coastal Pembrokeshire, all sharing similar feature encodings but selling at entirely different price points due to location.

Feature encoding loses local context

Postcode district encoding converts locations to numeric categories, allowing KNN to compute distances, but this treats geographically proximate districts as interchangeable if their codes are numerically similar. The true valuation driver micro-local market conditions, neighbourhood quality, accessibility networks cannot be captured through categorical encodings.

Training sample limitations

Training on only 20% of data (289,501 properties) reduces the comparable pool fivefold. For rare property configurations in thin markets (e.g., large detached rural properties), the training sample may contain zero true comparables, forcing KNN to retrieve systematically irrelevant matches

No learning of valuation relationships

Unlike Ridge Regression or CatBoost, which learn systematic relationships between features and prices (e.g., "each additional 10square metres adds £X in Cardiff but £Y in valleys"), KNN simply memorizes training examples. When test properties fall outside the training distribution's convex hull as guaranteed by geographic exclusion KNN extrapolates by averaging distant, irrelevant comparables.

Comparison With Other Models

KNN's catastrophic performance (average R^2 = -199.7%) ranks as the worst among all five models:

Table 37: Average Model Performance Across 9 Test LSOAs

Model	Average R^2 (9 LSOAs)	Average MAE
CatBoost	28.2%	£76,392
Ridge	26.1%	£69,646
LLM Ensemble	15.6%	£91,291
DRC	-22.7%	£117,820
KNN	-199.7%	£114,528

Despite achieving training R^2 = 92.9%, KNN's test performance falls below even the theory-based DRC formula (which itself fails catastrophically). This demonstrates that memorisation-based approaches provide no generalisation capability under geographic distribution shift.

Implications: Lessons From Rule-Based and Non-Parametric Approaches

The failures of both exploratory rule-based fuzzy logic and formal KNN evaluation illuminate fundamental requirements for successful automated valuation under geographic distribution shift.

Context-dependent learning trumps fixed rules

Professional valuer reasoning, while superficially expressible as linguistic rules ("large house in good location → high value"), in practice draws upon vast tacit knowledge accumulated through observing thousands of transactions. The exploratory fuzzy logic experiments demonstrated that explicitly articulated rules capture only the surface of valuation expertise, missing the implicit pattern recognition and contextual adaptation that constitute its core. Models that learn relationships from data (Ridge, CatBoost) outperform both fixed rules and pure memorization (KNN) by discovering context-dependent patterns; how floor area affects value differently in Cardiff vs. valleys, how property type premiums vary by local market structure.

Generalisation requires abstraction

KNN's catastrophic failure (R^2 = -1556.5% in Powys) despite 92.9% training accuracy proves that memorizing training examples provides no generalisation to unseen geographic contexts. Similarly, the rule-based experiments' inability to accommodate geographic variation showed that predetermined logic cannot adapt to local market idiosyncrasies. Successful models must abstract general principles (statistical relationships between features and prices) transferable across contexts, rather than encoding specific examples or universal rules.

Professional judgment cannot be fully automated through similarity alone

The comparable-sales approach, fundamental to professional appraisal, succeeds because human valuers exercise sophisticated judgment about which properties constitute "true comparables," considering qualitative factors (neighbourhood feel, street appeal, renovation quality) and market timing nuances that automated distance metrics cannot capture. KNN operationalizes only the mechanical aspect of comparable retrieval (finding similar features), missing the contextual reasoning that makes human comparable-sales analysis effective.

Adaptive reasoning frameworks outperform static systems

The LLM ensemble (Model 5), despite also lacking Wales-specific training, achieves R^2 = 15.6%; vastly outperforming both KNN (-199.7%) and what exploratory rule-based systems could achieve. This advantage stems from the LLM's ability to generate context-specific reasoning dynamically for each property, drawing upon vast pre-trained knowledge about property markets, urban development, and geographic patterns. Unlike fixed rules that treat all Cardiff properties identically, the LLM agents adapt their valuation logic to each specific property context, enabling rudimentary Generalisation despite zero direct training on Welsh prices.

The research trajectory, from exploratory rule-based experiments through formal KNN evaluation to statistical and ML approaches, demonstrates that automated property valuation in heterogeneous geographic markets requires methods capable of learning context-dependent relationships from data, abstracting transferable patterns, and adapting to local market conditions. Neither predetermined expert rules nor pure example memorization can substitute for statistical learning that discovers systematic valuation principles empirically while accommodating geographic variation.

4.6. Comprehensive Error Analysis Across Models

This section presents a systematic analysis of prediction errors across all five valuation models tested on the 9 priority LSOAs. The analysis examines aggregate error distributions, property type stratification, temporal error patterns, cross-LSOA variability, and error correlations between different modelling approaches.

Aggregate Model Performance

Table 38 presents the core performance metrics for each model across the test set:

Table 38: Model Performance Summary (9 Priority Test LSOAs)

Model	R^2 (Avg)	MAE (Avg)
Ridge Regression	0.261	£69,646
CatBoost	0.282	£76,392
KNN (Fuzzy Matching)	-1.997	£114,528
DRC Formula	-0.227	£117,820

Model	R^2 (Avg)	MAE (Avg)
LLM Ensemble	0.156	£91,291

Key Findings

- Best performer: CatBoost achieved the highest within-LSOA R^2 (28.2% average) and lowest MAE (£76,392), demonstrating gradient boosting's effectiveness at capturing local property value patterns within geographically excluded areas.
- Catastrophic failures: Both KNN and DRC formula produced negative R^2 values, indicating predictions worse than the mean baseline. KNN's average R^2 of -2.00 reflects severe overfitting to training data geographic patterns that do not transfer to test LSOAs. DRC's negative R^2 (-0.227) demonstrates that theory-based construction cost formulas systematically mis-estimate property values when applied without local market calibration.
- Weak cross-LSOA generalisation: Despite competitive within-LSOA R^2, the overall R^2 values approaching zero for CatBoost (0.017), KNN (-0.002), and LLM (0.00007) reveal that none of the models successfully explain variance across the 9 LSOAs. This indicates the models learn LSOA-specific patterns but fail to generalise to the between-LSOA price differences.

Property Type Stratification

Error patterns vary substantially by property type across all three models with available stratified data (Table 39):

Table 39: Mean Absolute Error by Property Type

Property Type	KNN MAE	DRC MAE	LLM MAE	Sample% (KNN/DRC)	Sample% (LLM)
Detached	£114,747	£112,182	£90,033	19.4%	22.2%
Semi-detached	£66,186	£83,311	£57,174	20.8%	23.6%
Terraced	£81,773	£88,223	£50,181	18.8%	21.4%
Flats	£73,443	£56,338	£56,333	34.8%	28.0%

Detached properties consistently produce the largest errors across all models, with KNN and DRC both exceeding £110,000 MAE. This likely reflects the substantial heterogeneity within the "detached" category; properties ranging from modest bungalows to executive homes with extensive land. This makes it difficult for models to accurately value without detailed property-specific information.

The DRC formula performs notably better on flats (£56,338 MAE) than other property types, suggesting its standardized construction cost assumptions align more closely with apartment characteristics where floor area is the dominant value driver and land component is minimal.

The LLM ensemble achieves the most balanced performance across property types, with terraced properties showing the lowest errors (£50,181) and detached properties the highest (£90,033). This 1.8× error range is substantially narrower than KNN's 1.7× range and DRC's 2.0× range, suggesting the LLM's reasoning capability provides more consistent valuation across diverse property forms.

Temporal Error Patterns

Analysis of errors by transaction year reveals dramatic temporal variation, with a striking surge in mid-2010s followed by contraction during the pandemic era.

Historical Baseline (1995-2014)

Both KNN and DRC models maintained relatively stable MAE ranging from £60,000-£80,000 throughout this 20-year period. This stability suggests the core hedonic relationships learned from Wales-wide training data provided reasonable approximations during this period of gradual house price appreciation.

Mid-2010s Surge (2015-2019)

Errors increased dramatically, peaking in 2017:

- KNN: 2017 MAE = £533,536 (7× increase from baseline)
- DRC: 2017 MAE = £489,177 (7× increase from baseline)
- 2018 KNN MAE = £408,767; 2019 KNN MAE = £342,097
- 2018 DRC MAE = £382,351; 2019 DRC MAE = £323,765

This extraordinary error surge coincides with W01002019 (Cardiff Bay) transactions during these years, likely reflecting the waterfront development boom that created property values fundamentally disconnected from Wales-wide training patterns.

Pandemic Era Contraction (2020-2021)

Counter-intuitively, errors decreased substantially during the COVID-19 pandemic:

- KNN: 2018-2019 average MAE = £377,494 → 2020-2021 average MAE = £212,524 (-43.7% decrease)
- DRC: 2018-2019 average MAE = £354,871 → 2020-2021 average MAE = £216,027 (-39.1% decrease)

This 40%+ error reduction contradicts expectations that pandemic-era market disruptions would increase prediction difficulty. Possible explanations include: (1) fewer high-value Cardiff Bay transactions during lockdowns, (2) temporary price compression in the 9 test LSOAs that aligned better with Wales-wide patterns, or (3) compositional changes in which properties transacted during this period.

Recent Period (2022-2025)

Errors remain elevated but below the 2015-2019 peak:

- 2022 KNN MAE = £181,669; 2023 KNN MAE = £196,738; 2024 KNN MAE = £219,200
- 2022 DRC MAE = £153,496; 2023 DRC MAE = £157,585; 2024 DRC MAE = £181,651

The sustained elevation relative to pre-2015 baseline suggests permanent structural changes in the test LSOAs' property markets that diverge from the training data patterns.

Error Correlation with Property Value

Extremely high correlations between absolute errors and property values indicate that models systematically struggle more with expensive properties:

- KNN: ρ = 0.981
- DRC: ρ = 0.994
- LLM: ρ = 0.996

These correlations approaching 1.0 demonstrate that valuation error increases nearly proportionally with property value. A £100,000 property may have a £10,000 error, while a £1,000,000 property has a £100,000 error, maintaining approximately constant percentage errors but vastly different absolute errors.

This heteroscedasticity has critical implications for valuation practice. Models optimized for mean absolute error will necessarily perform worse on high-value properties, while models optimized for percentage error may perform unacceptably on low-value properties. The relationship is so strong (ρ > 0.98) that property value alone is nearly sufficient to predict error magnitude, regardless of model choice.

Cross-LSOA Error Variability

All models exhibit substantial variation in performance across the 9 test LSOAs:

Ridge Regression

- R^2 range: -0.023 to 0.515 (std: 0.191)
- MAE range: £35,454 to £174,127 (std: £70,632)

CatBoost

- R^2 range: 0.002 to 0.518 (std: 0.195)
- MAE range: £35,160 to £165,983 (std: £61,776)

KNN (Fuzzy Matching)

- R^2 range: -15.565 to 0.003 (std: 4.806)
- MAE range: £70,701 to £282,327 (std: £61,401)

DRC Formula

- R^2 range: -0.786 to -0.013 (std: 0.246)

- MAE range: £75,825 to £263,388 (std: £53,645)

LLM Ensemble

- R^2 range: -0.017 to 0.464 (std: 0.169)
- MAE range: £39,309 to £270,104 (std: £65,424)

Worst-Case LSOA Performance

W01002019 (Cardiff Bay)

This high-value waterfront LSOA consistently produced the largest MAE values across all models:

- Ridge: £174,127
- CatBoost: £165,983
- KNN: £282,327
- DRC: £263,388
- LLM: £270,104

The uniformly large errors (£240k-£280k across all models) suggest this LSOA contains property types or market dynamics fundamentally absent from the training data. Cardiff Bay's waterfront premium, modern apartment developments, and unique amenity profile create a local market that cannot be predicted from Wales-wide hedonic patterns.

W01000449 (rural Powys)

KNN achieved catastrophic failure with $R^2 = -15.565$, indicating predictions over 15 times worse than the mean baseline. This reflects KNN's complete inability to identify similar properties in the training data when confronted with rural Welsh market characteristics absent from the algorithm's learned similarity space.

W01000114 (Gwynedd 009D)

DRC formula achieved its worst performance here ($R^2 = -0.786$), with systematic underestimation averaging -£87,058 per property. The DRC's assumptions of £1,732.79/square metres construction costs and £2,195.68/square metres land values clearly misalign with Gwynedd's actual property market structure.

Model Error Correlation Analysis

Analysis of prediction errors reveals extremely high correlations between different modelling approaches (n=9,606 properties):

Error Correlations:

- KNN-DRC: $\rho = 0.981$
- KNN-LLM: $\rho = 0.981$
- DRC-LLM: $\rho = 0.998$

These correlations approaching 1.0 indicate that error patterns are driven primarily by property-level characteristics that all models fail to capture, rather than model-specific architectural limitations. Properties that are undervalued by KNN tend to be equally undervalued by DRC and LLM, suggesting systematic blind spots in the available feature space rather than algorithmic deficiencies.

The near-perfect DRC-LLM correlation ($\rho = 0.998$) is particularly revealing. Despite the LLM ensemble using sophisticated zero-shot reasoning with multiple specialist agents, its errors mirror the simple formula-based DRC approach almost exactly. This suggests the LLM's learned representations do not provide meaningful additional information beyond what can be captured by basic hedonic assumptions about construction costs and land values.

Heteroscedasticity Analysis

Breusch-Pagan tests confirm significant heteroscedasticity in all models, with error variance increasing systematically with predicted property values:

Heteroscedasticity Test Results:

- KNN: $\chi^2 = 4{,}046.6$, $p < 0.0001$ ***
- DRC: $\chi^2 = 28.4$, $p < 0.0001$ ***
- LLM: $\chi^2 = 561.2$, $p < 0.0001$ ***

All three models reject the null hypothesis of constant error variance at $p < 0.0001$ significance level. The extremely large KNN test statistic ($\chi^2 = 4{,}046.6$) indicates severe heteroscedasticity where error variance increases dramatically with property value.

This heteroscedasticity creates methodological challenges for model evaluation and comparison. Mean absolute error (MAE) is dominated by errors on high-value properties, while R^2 may be artificially inflated or deflated depending on the distribution of property values in test vs. training sets.

Interpretation and Implications

The Geographic Generalisation Problem

The consistently poor overall R^2 values (near zero or negative) combined with moderate within-LSOA R^2 values (20-30% for best models) reveal a fundamental limitation. Models trained on Wales-wide data cannot generalise to new geographic areas, even when property-level features are included. This finding has critical implications for automated valuation models (AVMs) deployed in practice. A model may appear to perform well in random cross-validation (achieving $R^2 = 0.85+$ by capturing between-LSOA geographic components), but when applied to new neighbourhoods not represented in training data - precisely the scenario where AVMs are most needed - the same model catastrophically fails.

Model Architecture Makes Minimal Difference

The extremely high error correlations ($\rho > 0.98$ between all model pairs) demonstrate that choice of algorithm has minimal impact on which specific properties are mis-valued. The

errors are property-driven, not algorithm-driven. The LLM ensemble's near-perfect error correlation (ρ = 0.998) with the simple DRC formula is particularly revealing: despite using sophisticated reasoning, the LLM makes essentially identical mistakes. This suggests that improving valuation accuracy requires better features rather than more sophisticated modelling techniques.

Extreme Within-LSOA Heterogeneity

The massive error standard deviations in Cardiff Bay (£1.6M+) and rural Powys (£400k) indicate that LSOA-level aggregation masks substantial intra-LSOA variation. A single LSOA can contain council estates and luxury developments, making LSOA-level geographic holdout validation an extremely challenging test.

4.7. Feature Importance Analysis

Understanding which land attributes most strongly influence valuation predictions illuminates both the mechanisms through which models generate estimates and the fundamental drivers of housing market value in Wales. Feature importance analysis, conducted across multiple model architectures employing different importance quantification methods, reveals remarkable consistency in top predictors while exposing critical limitations in available data that constrain all approaches. This convergence of importance rankings across parametric, ensemble, and instance-based methods, despite their fundamentally different mathematical foundations, suggests that identified patterns reflect genuine market structure rather than model-specific artifacts.

Methodological Approaches to Importance Quantification

Ridge Regression importance derives from standardised coefficient magnitudes, measuring the marginal effect on log(price) of one-standard-deviation changes in each feature. This approach directly quantifies linear associations but cannot capture non-linear relationships or interaction effects a 10 square metres floor area increase may affect £80,000 terraced houses differently than £350,000 detached houses, but Ridge coefficients apply uniform marginal effects across the price distribution.

CatBoost employs information gain importance, measuring cumulative reduction in loss function (RMSE on log-price) attributable to each feature across all decision tree splits during training. This metric captures both direct effects and interactions when CatBoost splits first on land_type, then on floor_area within each type of category, both features accumulate importance reflecting their joint contribution to prediction accuracy. The information gain approach naturally handles non-linearities and categorical variables, providing richer importance profiles than linear coefficient magnitudes.

K-Nearest Neighbours importance derives from permutation testing: randomly shuffling each feature's values and measuring resulting MAE increase quantifies how much prediction accuracy depends on that feature's true values versus random noise. Features whose permutation causes large MAE increases are deemed important; those whose randomisation minimally affects predictions contribute little information. This model-agnostic

approach applies to any algorithm but requires computationally expensive repeated predictions.

The LLM ensemble lacks traditional feature importance metrics, as agent reasoning operates through natural language rather than mathematical functions. However, analysing agent justification texts for feature mention frequency and sentiment provides proxy importance measures features referenced in 80%+ of agent explanations (land type, location, floor area) likely drive valuations, while rarely-mentioned features (tenure, seasonality) contribute minimally.

Aggregate Feature Importance: Cross-Model Consensus

Synthesising importance rankings across Ridge, CatBoost, and permutation-based KNN analysis reveals seven features consistently appearing in top-10 importance across all methods, indicating robust associations with land value regardless of modelling assumptions.

Table 40: Feature Importance Rankings Across Models

Feature	Ridge Rank	Ridge Coef (std)	CatBoost Rank	CatBoost Info Gain (%)	KNN Permutation Rank	KNN MAE Increase (£)	Cross-Model Avg Rank
TransactionYear	1	+0.42	1	33.2	2	+38,200	1.3
Land Type	2	+0.38	2	25.6	1	+42,100	1.7
Postcode District	3	+0.35	3	25.0	3	+35,800	3.0
Floor Area (log square metres)	4	+0.28	4	9.7	4	+18,400	4.0
New Build Status	6	+0.12	5	2.6	8	+6,200	6.3
Leasehold Tenure	8	-0.09	6	2.2	11	+3800	8.3
Seasonality (month)	12	+0.03	13	0.6	15	+1,200	13.3

Transaction Year dominates importance across all models

Accounting for 33-48% of CatBoost's predictive power and generating +£38,200 MAE when permuted in KNN analysis. This extraordinary importance reflects multiple confounded factors: (1) temporal appreciation trends - Welsh land transactions appreciated 4.3% annually 2015-2024 after HPI normalization, but with substantial geographic variation (Cardiff 8-12%, valleys 0-2%); (2) data quality evolution - pre-2008 transactions have 61% missing floor area, post-2012 transactions have 95% completeness through EPC linkage, causing transaction year to proxy for feature reliability; (3) market regime effects - 2003-2007 credit boom, 2008-2012 crisis, 2014-2024 recovery exhibit different price-to-attributes relationships that temporal features capture.

Ridge's standardized coefficient of +0.42 indicates that land transactions sold one standard deviation later (approximately 8 years) command 42% higher prices after accounting for all other features - a remarkable temporal effect exceeding land type differentials (detached vs terraced = 38% premium). CatBoost's 33% importance allocation suggests that nearly one-third of explainable price variation derives from when rather than what is sold, confirming temporal dynamics as first-order valuation drivers.

Land Type ranks consistently second

Capturing the fundamental hierarchy of Welsh housing stock, detached (+42% baseline), semi-detached (+8%), terraced (reference category), flat (-12% in Ridge specification). CatBoost's 25.6% importance reflects both this direct effect and learned interactions—new build status interacts strongly with land type (new detached houses command 15% premium, new flats only 3% premium due to leasehold complications), as does floor area (each additional 10 square metres adds £18,000 to detached values, £12,000 to semi-detached, £8,000 to terraced).

KNN's highest permutation importance (£42,100 MAE increase when randomized) indicates that land type serves as primary comparable-retrieval criterion—when type is shuffled, the algorithm matches terraced land transactions to detached comparables and vice versa, producing catastrophic mispredictions. This confirms that even simplistic similarity-based methods recognize land type as the most critical matching dimension.

Postcode District ranks third universally

Encoding geographic premiums ranging from -35% (Ebbw Vale, Merthyr Tydfil) to +50% (Cardiff Bay, Monmouthshire). Ridge's +0.35 standardized coefficient represents the average differential between highest and lowest value districts after controlling for land characteristics; geography drives 35% of log-price variation independent of structural attributes. CatBoost's 25.0% importance, nearly equal to land type, confirms that where a land is located matters as much as what type of dwelling it is.

The postcode district feature aggregates multiple underlying factors: employment accessibility (Cardiff proximity commands 2-3% premium per 10km closer), amenity access (schools, retail, transport), environmental quality (coastal proximity, green space, air quality), and socioeconomic composition (neighbourhood effects, peer income). This aggregation explains why finer-grained features added in experimental models (distance to Cardiff (4.3% importance), area classification (6.2%), distance to coast (2.6%)) sum to less than postcode district alone (25.0%). They decompose components already captured implicitly by postcode indicators, adding limited incremental information.

Floor Area (log-transformed square metres) ranks fourth consistently

Though importance magnitude varies substantially across methods. Ridge's +0.28 coefficient indicates 28% price increase for one-standard-deviation floor area increase (approximately 30 square metres or 35%), reflecting linear size-value relationship. CatBoost's 9.7% importance—substantially lower than Ridge would suggest—indicates non-linearity: the model learns that size premiums vary by context (larger effects in Cardiff suburbs where space is valued, smaller effects in valleys where oversupply exists) and exhibit diminishing returns (first 50 square metres are valued more highly per-square-meter than 100-150 square metres).

The floor area importance varies dramatically by LSOA. In Cardiff Bay, floor area contributes only 3.2% importance (location dominates, with 60 square metres and 120 square metres flats both valued primarily for waterfront proximity), while in rural Powys, floor

area reaches 18.4% importance (distinguishing small cottages from large farmhouses drives much of within-area price variation). This context-dependence explains why experimental models incorporating floor-area × LSOA interaction terms improve R^2 by 0.04—allowing size effects to vary geographically better captures market reality than uniform linear relationships.

New Build Status, Leasehold Tenure, and Seasonality rank consistently lower (5th-15th depending on model)

With importance below 3% in CatBoost and marginal effects <£8,000 in KNN permutation tests. New build premiums of 7-12% apply primarily to detached and semi-detached houses; flats and terraced land transactions show minimal or negative new build effects, as leasehold complications and service charge uncertainties offset modern specification advantages. Leasehold tenure imposes 3-5% discounts, though effects concentrate in flat segments (20-30% of flat value) while being negligible for houses (98% freehold).

Seasonality effects prove minimal across all models

Land transactions sold in spring (March-May) command 2.1% premiums over winter (December-February) sales, but this difference is statistically insignificant and economically trivial relative to other drivers. Welsh land markets, unlike London's pronounced spring peak, exhibit relatively uniform transaction activity across seasons, likely reflecting that buyers in affordable markets (median £155,000) face less competition and tighter budgets than luxury markets where seasonal timing optimization matters.

Feature Importance Divergence Across LSOAs

Disaggregating CatBoost importance by LSOA reveals that feature rankings shift dramatically across geographic contexts, indicating that no universal valuation function applies Wales-wide.

Cardiff Bay (W01002019)

Postcode district dominates (42.8% importance), reflecting that 200-500m distance gradients from waterfront drive £80,000-£150,000 price variation among otherwise similar land transactions. land type importance drops to 12.3% (flats, terraces, and houses coexist at similar price points when location is premium), while floor area contributes only 3.2% (buyers pay for location, accepting smaller spaces). Transaction year reaches 38.1% importance, capturing rapid regeneration appreciation (8-12% annual) that dominates structural attributes.

Post-Industrial Valleys (W01000114, W01001233)

Land type dominates (38.4-42.1% importance), as detached houses retain value (£110,000-£140,000) while terraces collapse (£45,000-£75,000) within the same LSOA. Transaction year contributes only 8.2% (stagnant markets show minimal temporal variation), while postcode district drops to 6.1% (all valley locations equally depressed, minimal within-valley gradients). Floor area reaches 22.3% importance, as distinguishing renovated larger terraces from unrenovated smaller ones drives much within-category variation.

Rural LSOAs (W01000449, W01000517)

Floor area dominates (24.6-28.3% importance), distinguishing small cottages (£95,000-£140,000) from large farmhouses (£240,000-£380,000). land type drops to 18.2% (nearly all detached or semi-detached, limited typological variation), while postcode district contributes only 11.4% (coarse aggregation loses village-level and model-specific premiums that dominate rural markets but operate below postcode resolution).

Suburban Homogeneous Estates (W01001045)

Transaction year dominates (45.2%), as temporal appreciation within stable market (3-5% annual) affects all land transactions uniformly, making when land transactions sold more predictive than structural differences. land type (28.1%) and floor area (14.3%) rank second and third, while postcode district contributes only 8.7% - within homogeneous suburb, all locations similar, creating flat geographic premium landscape.

This LSOA-level feature importance heterogeneity explains why Wales-wide models achieve only $R^2 = 0.20$-0.25: applying uniform feature weights calibrated to aggregate patterns produces systematic errors in any specific local market where importance rankings differ from national averages. LSOA-clustered modelling approaches - training separate models for urban regeneration, suburban, valley, and rural market types - could apply context-appropriate feature weights, potentially improving R^2 to 0.35-0.45 based on within-cluster performance observed when models are trained on single-LSOA data (though this requires sufficient transaction volumes, limiting applicability to sparse rural areas).

Interaction Effects and Non-Linearity

CatBoost's ability to capture interactions reveals that feature importance extends beyond additive main effects to include synergistic combinations.

Land Type × Floor Area

Detached houses show £19.2/square metres marginal values, semi-detached £13.8/square metres, terraced £9.4/square metres, flats £7.1/square metres in CatBoost's learned relationships. These interactions contribute 4.8% additional importance beyond main effects, indicating that size premiums depend fundamentally on land type. A 40 square metres expansion (100 square metres → 140 square metres) adds £76,800 to detached valuations, £55,200 to semi-detached, £37,600 to terraced, £28,400 to flats relationships linear models with uniform floor area coefficients cannot capture

New Build × Land Type

New build detached houses command +15.3% premiums, semi-detached +9.8%, terraced +4.2%, flats -2.1% (slight discount due to leasehold and service charge complications in new developments). This 17.4 percentage point range in new build effects contributes 2.1% importance through interactions, improving predictions for new developments where land type alone underestimates premiums

Transaction Year × Postcode District

Cardiff districts appreciated 8-12% annually 2015-2024, valley districts 0-2%, creating divergent temporal trajectories. CatBoost learns these varying trends, allocating 5.2% importance to year-location interactions that allow appreciation rates to differ geographically. This interaction explains why Ridge (uniform temporal trends) underpredicts recent Cardiff sales by £45,000-£68,000 while overpredicting valley sales by £18,000-£32,000—a single temporal coefficient cannot accommodate geographic divergence in market dynamics.

Floor Area × Postcode District

Urban areas value space less intensely (Cardiff: £8.2/square metres) than rural areas (Powys: £14.8/square metres), reflecting supply constraints (urban: £240-380/square metres land costs limit large models, rural: £80-120/square metres land costs enable larger holdings). This interaction contributes 1.8% importance, allowing size effects to adapt to local market context.

Summing interaction contributions (4.8% + 2.1% + 5.2% + 1.8% = 13.9%), CatBoost allocates nearly 14% of total importance to non-additive effects - relationships Ridge's linear specification entirely misses. This explains CatBoost's modest R^2 advantage over Ridge (0.25 vs 0.20): the 0.05 gain derives primarily from capturing interactions rather than transforming individual features, though even with interaction terms, 75% of test variance remains unexplained due to unobserved attributes and geographic distribution shift.

Missing and Low-Importance Features

Examining features that contribute <1% importance despite theoretical relevance reveals critical data limitations constraining model performance.

Model/Land Size (not directly observed)

A very large proportion of land was not valued in this research. The study values only land linked to observed, transacted properties (mainly residential) with usable attributes, so it excludes both (i) land not represented in the transaction-based dataset (e.g., much non-residential, bare, agricultural and infrastructure land) and (ii) land that exists but did not transact during the study period. In terms of land area, total coverage is less than 10%

Ground truth land area is not recorded in transaction data, forcing models to infer from land type (detached assumed large models, terraced small) or rely on DRC's problematic floor-area × 0.3 assumption. Experimental models incorporating OS MasterMap-derived model boundaries (available for 42% of land transactions through geocoding) assign this feature 11.2% importance when present, indicating land size would rank 5th-6th if universally available. The absence of systematic land area data forces models to conflate structure and land value, contributing to geographic mispredictions where land value proportions vary (20% in valleys, 50% in rural areas).

Renovation/Condition (not observed)

Transaction data contain no quality indicators. A comprehensively renovated 1920 terrace with modern kitchen, bathroom, and insulation cannot be distinguished from an unrenovated equivalent with original fixtures and single-glazing. Land age (derivable from construction date when available, 68% coverage) serves as poor proxy, as 100-year-old renovated land transactions often command premiums over 30-year-old unrenovated land transactions. EPC energy ratings (available for post-2008 transactions, 71% coverage) correlate with renovation ($\rho = 0.52$) but measure thermal efficiency rather than aesthetic quality. When EPC ratings are included in experimental models, they contribute 3.8% importance, indicating condition would likely rank 7th-8th if comprehensive quality data existed.

Parking Availability (not observed)

Off-street parking commands £8,000-£18,000 premiums in urban contexts where on-street parking is scarce, but no parking indicator exists in transaction records. Suburban land transactions are assumed to have parking (driveways typical for 1960s+ estates), urban terraces assumed not to have parking (1880-1920 construction predates car ownership), but these assumptions fail for 15-20% of land transactions (urban terraces with converted front gardens, suburban land transactions on narrow models). This misclassification contributes to within-LSOA error variance that models cannot explain.

Garden Size (not observed)

Garden area drives £5,000-£25,000 variation among otherwise-identical land transactions, particularly for terraced and semi-detached categories where model size variation is substantial (80-250 square metres range). OS MasterMap provides model boundaries for some land transactions, but distinguishing building footprint from garden requires additional processing not performed in this dataset. The absence of garden size data conflates land transactions with 60 square metres paved yards and 180 square metres landscaped gardens, contributing to seemingly random price variation that frustrates model learning.

Proximity to Specific Amenities (partially observed)

While experimental models include distance to Cardiff, coast, and railway stations (contributing 4.3%, 2.6%, and 0.2% importance respectively), proximity to specific schools (catchment areas command 5-15% premiums for highly-rated comprehensives), hospitals (negative effects for adjacent land transactions, positive for 1-2km proximity), and retail centres (walkable supermarket access valued at £4,000-£12,000) remains unobserved. Postcode district indicators partially capture these effects through aggregation, but within-postcode variation in amenity access contributes to unexplained error.

Land Extensions or Modifications (not observed)

Extensions increasing floor area by 20-40 square metres (conservatories, loft conversions, rear extensions) command premiums of £25,000-£60,000 depending on quality, but transaction records contain only total floor area (when available) without distinguishing original construction from additions. Two 120 square metres land transactions, one original build, one 90 square metres with 30 square metres extension receive identical model inputs

despite £15,000-£35,000 typical value differences (extensions valued at 50-70% of equivalent new-build space due to quality and integration concerns).

Collectively, these unobserved features likely explain 25-35% of land value variation based on appraisal literature and limited experiments incorporating partial proxies. The fact that best-performing models achieve $R^2 = 0.20$-0.25 indicates that observed features (land type, location, floor area, age) capture approximately one-third of systematic valuation drivers, with two-thirds of variation attributable to unobserved quality, condition, and micro-locational factors plus irreducible idiosyncratic noise (buyer-seller negotiation dynamics, motivated sales, estate settlements).

Evolution of Feature Importance

Assessing whether feature importance rankings remain stable or shift over time informs whether models require only coefficient retraining (stable importance, changing magnitudes) or architectural redesign (shifting importance rankings).

Transaction year importance increased dramatically over the study period: in models trained on 1995-2005 transactions, year contributed 12.3% importance; in 2015-2024 models, importance reached 33.2% (2.7x increase). This shift reflects diverging regional appreciation trajectories 1995-2005 saw relatively uniform 3-5% annual growth across Wales, while 2015-2024 exhibited extreme divergence (Cardiff 8-12%, valleys 0-2%), causing temporal features to proxy for unobserved geographic heterogeneity in appreciation rates. If this trend continues, transaction year will dominate all other features by 2030, indicating that when land transactions sold becomes more predictive than structural attributes - a troubling pattern suggesting price-to-fundamentals relationships are weakening.

Land type importance remained stable (23.9% → 25.6% over 1995-2024), confirming that fundamental hierarchies (detached > semi-detached > terraced > flat) persist despite market evolution. However, the magnitude of premiums shifted: detached premiums over terraced increased from 32% (1995-2005) to 42% (2015-2024), reflecting that larger land transactions with land appreciated faster than compact urban terraces during the study period.

Postcode district importance increased moderately (18.1% → 25.0%), reflecting growing geographic polarisation price gaps between highest-value (Cardiff Bay, Monmouthshire) and lowest-value (valleys) districts widened from 3.8x in 1995 to 6.2x in 2024. This increasing spatial inequality drives rising geographic feature importance, indicating that where land transactions are located matters progressively more over time

Floor area importance decreased substantially (18.4% → 9.7%), not because size became less valued but because other features (especially temporal and geographic) grew more important, reducing floor area's relative contribution. The absolute marginal value of floor area increased (£110/square metres → £145/square metres) while its importance share declined, indicating size premiums rose less rapidly than appreciation and geographic divergence effects.

This temporal instability in importance rankings suggests models require periodic architectural review beyond coefficient retraining features that dominated valuation in 2010 (floor area, land type) contribute less in 2024 (temporal trends, geography), indicating that feature engineering and model specification must adapt to evolving market structure rather than applying fixed functional forms indefinitely.

Implications for Model Design and Data Collection

Feature importance analysis directly informs strategic priorities for improving automated valuation systems.

High-Priority Data Collection

Systematic land and model area measurement would add a feature likely ranking 5th-6th in importance (11% based on partial data experiments), substantially improving predictions for land types where land dominates value (detached, rural). Renovation and condition scoring even crude three-category classification (poor, average or excellent) would contribute 3-5% importance, reducing within-type prediction variance. These two enhancements, achievable through OS MasterMap integration and automated image analysis of listing photos, could lift R^2 from 0.20-0.25 to 0.28-0.33 based on experimental evidence.

Lower-Priority Enhancements

Parking indicators, garden size, and amenity proximity metrics would collectively add 2-4% importance, offering marginal improvements insufficient to justify collection costs unless data emerge as byproducts of other processes (e.g., parking flags from aerial imagery classification, amenity proximity from existing GIS databases).

Feature Engineering Over Feature Addition

Given that top-3 features (transaction year, land type, postcode district) account for 80-85% of captured importance, engineering better representations of these features offers higher returns than collecting novel attributes. Specifically: (1) replacing coarse postcode districts with finer Output Areas or 1km grid cells could capture micro-locational gradients currently missed, potentially adding 3-5% importance; (2) modelling non-stationary temporal trends through local polynomial smoothing or changepoint detection could better capture divergent appreciation trajectories than linear year coefficients, adding 2-3% importance; (3) hierarchical land type encodings (detached → detached-modern vs detached-period vs detached-bungalow) could capture within-type premiums, adding 1-2% importance.

The analysis confirms that feature availability fundamentally constrains valuation accuracy more than model sophistication no algorithm can learn patterns from unobserved data. Ridge Regression with comprehensive features (including land area, condition, parking) would likely outperform CatBoost with current limited features, indicating that data enrichment offers higher marginal returns than architectural complexity. Operational deployment priorities should emphasize systematic data collection and linkage (land register + EPC + OS MasterMap + planning records) over incremental modelling refinements, as the

gap between theoretical best performance ($R^2 \approx 0.60$-0.70 with complete features, based on appraisal literature) and current achievement ($R^2 \approx 0.20$-0.25) stems primarily from missing data rather than inadequate algorithms.

4.8. Geographical Error

The spatial distribution of prediction errors across Wales reveals systematic geographic biases that persist across all modelling approaches, confirming that distribution shift; the divergence between training data characteristics and test deployment contexts; constitutes the dominant obstacle to accurate automated valuation at national scale. Unlike random errors that average to zero and decrease with larger samples, geographic errors exhibit structural patterns where entire regions are systematically over- or under-valued, indicating fundamental misalignment between learned relationships and local market realities.

Spatial Error Clustering and Autocorrelation

Computing Moran's I statistic on absolute prediction errors tests whether high-error properties cluster geographically. Significant positive spatial autocorrelation would indicate that when a model mis-predicts one property, it systematically mis-predicts neighbouring properties similarly; a signature of systematic geographic bias rather than random individual errors.

Analysis of the 9 test LSOAs (n=9,606 properties) reveals unexpectedly weak spatial autocorrelation:

- KNN Model: Moran's I = 0.012 (2km neighbour threshold)
- DRC Model: Moran's I = 0.009 (2km neighbour threshold)
- Expected I under random distribution: -0.000

Both models exhibit Moran's I values barely above the random expectation, indicating that prediction errors do not cluster strongly at the 2km neighbourhood scale. This counterintuitive finding contradicts the common assumption that geographic errors would show strong spatial patterns. Instead, the weak autocorrelation suggests that errors are driven more by property-specific factors and LSOA-level systematic biases than by fine-grained neighbourhood effects.

The practical implication is significant. A property's error cannot reliably predict neighbouring properties' errors within 2km. This limits the utility of spatial smoothing or kriging approaches for error correction techniques that assume neighbouring observations provide information about local prediction quality. The weak spatial structure indicates that error patterns operate at two distinct scales: (1) LSOA-wide systematic biases (Cardiff Bay uniformly underpredicted, Gwynedd systematically mispriced), and (2) individual property variation (quality, condition, micro-location), with minimal intermediate neighbourhood-scale clustering.

Distance-Based Error Gradient

Computing prediction errors as functions of distance to Cardiff city centre quantifies how locational premiums operate across the test LSOAs. Using Haversine distance calculations

from property coordinates to Cardiff centre (51.4816°N, 3.1791°W), properties were grouped into quartile-based distance bins:

Table 41: KNN Model Distance Gradient

Distance Range	n	Mean Error	MAE
0.0-0.6km	2,556	+£300,447	£310,642
0.6-35.9km	2,451	+£103,631	£129,168
35.9-132.3km	2,403	+£23,645	£85,395
132.3-189.7km	2,404	+£12,567	£91,233

Table 42: DRC Formula Distance Gradient

Distance Range	n	Mean Error	MAE
0.0-0.6km	2,556	+£222,572	£295,695
0.6-35.9km	2,451	+£34,418	£115,552
35.9-132.3km	2,403	-£39,286	£88,648
132.3-189.7km	2,404	-£35,706	£106,526

Key Findings

Extreme Cardiff Bay underprediction

Properties within 0.6km of Cardiff city centre; primarily W01002019 (Cardiff Bay) waterfront properties; are massively underpredicted by both models. KNN shows mean error of +£300,447 (actual prices £300k higher than predictions) with MAE exceeding £310k. DRC shows similarly catastrophic +£222,572 mean error with £296k MAE. These values represent 80-160% of median property prices in this distance band, indicating that models systematically miss the waterfront premium by factors approaching 2 times.

Steep distance decay

Errors decline dramatically from the 0-0.6km to 0.6-36km range, dropping by £197,000 for KNN and £188,000 for DRC. This extraordinarily steep gradient (£200k over 36km, or £5,500 per kilometre) far exceeds typical urban distance-decay patterns documented in hedonic pricing literature (usually £500-£1,500/km for major UK cities). The magnitude indicates that the 0-0.6km bin captures Cardiff Bay's unique waterfront regeneration premium rather than standard urban gradients.

Monotonic decline for KNN

KNN exhibits consistently declining positive mean errors as distance from Cardiff increases, approaching near-zero bias beyond 132km. This pattern suggests the model systematically applies too-high valuations to properties near Cardiff (retrieving comparables from areas outside test LSOAs that include Cardiff premium) but approaches correct levels in distant rural areas where retrieved comparables better match test properties.

Sign reversal for DRC

DRC transitions from underprediction (0-36km, +£34k to +£223k) to overprediction (36-190km, -£36k to -£39k). This indicates the formula's fixed construction cost (£1,732.79/square metres) and land value (£2,195.68/square metres) parameters mis-calibrate at both extremes underestimating central Cardiff where land values substantially exceed £2,195.68/square metres and overestimating rural Wales where land values fall to £80-£150/square metres based on comparable rural transactions.

The distance gradient analysis validates the LSOA-level finding that W01002019 (Cardiff Bay) is fundamentally unpredictable using Wales-wide trained models. The £300k error magnitude within 0.6km radius represents the single most severe systematic bias identified across all geographic analyses.

LSOA-Level Systematic Bias Decomposition

Examining mean prediction error (bias) versus total error (MSE) by LSOA decomposes geographic failure into two components: systematic over/underprediction of entire areas (bias²) versus variable individual property misprediction within areas (variance). The bias²/MSE ratio quantifies what percentage of total error stems from systematic LSOA-wide misprediction versus property-specific factors.

Key Findings

The Cardiff Bay paradox

W01002019 exhibits the largest absolute mean bias across all LSOAs (£273k for KNN, £193k for DRC), yet bias²/MSE remains minimal at 1.3-2.7%. This counterintuitive result occurs because the enormous within-LSOA error variance (RMSE = £1.67M) dwarfs even £250k systematic bias. Cardiff Bay contains properties ranging from £150k one-bedroom flats to £2M+ waterfront penthouses a 13× price range within a single LSOA. This extreme heterogeneity means that systematic bias, while massive in absolute terms, becomes statistically minor relative to total squared error variance. The implication is that even perfect systematic bias correction (+£250k average adjustment) would leave most individual properties mis predicted due to unobserved intra-LSOA variation.

DRC's Gwynedd failure

W01000114 (Gwynedd 009D) shows the highest bias contribution across all 18 LSOA-model combinations at 41.4% of MSE for DRC. The systematic -£87,058 underprediction indicates DRC's construction cost (£1,732.79/square metres) and land value

(£2,195.68/square metres) parameters fundamentally mis calibrate for Gwynedd's market. This is actionable: recalibrating DRC parameters specifically for Gwynedd (increasing construction costs to £1,200-£1,300/square metres based on local quotations, reducing land values to £200-£250/square metres reflecting post-industrial context) could eliminate 40%+ of prediction error in this LSOA.

KNN's suburban bias

KNN shows bias-dominated patterns in Monmouthshire (31.1%) and Bridgend (36.5%), both with large positive bias (+£98k, +£78k). This suggests KNN's similarity-based comparable retrieval systematically matches these suburban/commuter LSOAs to higher-value training areas. Likely explanation: KNN retrieves Cardiff suburb comparables (property type, floor area, age) that command Cardiff accessibility premiums, applying these premiums incorrectly to Monmouthshire and Bridgend properties lacking equivalent accessibility.

Variance-dominated norm

13 of 18 LSOA-model combinations (72%) show bias²/MSE < 10%, indicating property-specific factors drive most error variance rather than systematic LSOA-wide misprediction. For these cases, improving accuracy requires feature enrichment (plot size, condition, parking, renovation quality) more than addressing systematic bias. Models already achieve approximately correct average valuations; they fail to distinguish individual properties due to missing features.

Model-specific vulnerability patterns

KNN tends toward positive bias (overvaluation) in 6 of 9 LSOAs, while DRC shows more balanced positive/negative bias distribution. This asymmetry suggests KNN's similarity retrieval systematically selects comparables from somewhat higher-value training areas, while DRC's parameter miscalibration affects both directions (underestimating high-value Cardiff, overestimating low-value rural areas).

Temporal Evolution of Geographic Errors

Assessing whether geographic error patterns remain stable over time distinguishes permanent market structure (persistent errors indicating fundamental training data inadequacy) from transient misprediction (errors that diminish as models learn emerging patterns).

Analysis of errors by transaction year (1995-2025) reveals dramatic temporal instability.

Mid-2010s error surge: Errors increased dramatically 2015-2019, peaking in 2017:

KNN 2017 MAE: £533,536 (7× baseline)
DRC 2017 MAE: £489,177 (7× baseline)

This extraordinary spike coincides with W01002019 (Cardiff Bay) transactions during the waterfront development boom. The magnitude suggests rapid localized appreciation created pricing dynamics entirely outside training data distributions—£2+ billion public/private

investment, Senedd presence, marina development, and cultural facilities created amenity premiums inadequately proxied by transaction year and postcode features alone.

Pandemic era contraction: Errors decreased substantially 2020-2021:

KNN: 2018-2019 average £377,494 → 2020-2021 average £212,524 (-44%)
DRC: 2018-2019 average £354,871 → 2020-2021 average £216,027 (-39%)

This counterintuitive 40% error reduction during COVID-19 market disruption suggests compositional changes such as; fewer ultra-high-value Cardiff Bay transactions (which drive errors due to extreme values exceeding model capacity), or temporary price compression bringing Cardiff Bay closer to Wales-wide patterns during economic uncertainty.

Sustained elevation: 2022-2025 errors remain above pre-2015 baseline (£150k-£200k vs £60k-£80k historically) but below 2015-2019 peaks, suggesting permanent structural changes in test LSOA markets that diverge from training distribution.

The temporal instability confirms that even with perfect features, geographic heterogeneity creates moving targets. Cardiff Bay appreciated 8-12% annually while valleys appreciated 0-2% annually during 2015-2024, creating divergent trajectories that models calibrated to national 4.3% average systematically misprice.

Implications for Geographic Generalisation

The systematic geographic error patterns observed across all models indicate that national-scale valuation systems cannot achieve uniformly acceptable accuracy through algorithmic sophistication or feature engineering alone. Three fundamental constraints prevent geographic generalisation.

Representational inadequacy

Training data overrepresent urban and suburban Wales (62% of transactions) while underrepresenting valleys (9%), rural areas (18%), and unique contexts like Cardiff Bay (0.4%). Models calibrated to overrepresented segments systematically misprice underrepresented ones, with errors growing monotonically as test contexts diverge from training distribution central tendencies. The £300k Cardiff Bay error within 0.6km radius exemplifies this: 0.4% representation cannot support learning £2M+ waterfront penthouses when 99.6% of training data are non-waterfront properties priced £80k-£350k.

Weak spatial structure

Moran's I values of 0.009-0.012 indicate errors operate at LSOA-scale (systematic bias) and property-scale (individual variation), with minimal 2km neighbourhood clustering. This prevents spatial smoothing corrections and confirms that improving one property's prediction provides negligible information about neighbouring properties.

Non-stationarity

The 7× error surge 2015-2019 followed by 40% pandemic contraction demonstrates that geographic relationships evolve continuously. Fixed models trained on historical data become progressively mis calibrated as local markets evolve at different rates, requiring perpetual retraining cycles that still lag real-time change by 6-18 months.

These constraints indicate that operational automated valuation systems must accept geographic error heterogeneity as fundamental limitation, implementing tiered deployment strategies:

High-confidence automated valuation in well-represented homogeneous markets

Suburban estates like W01001045 achieving $R^2 = 0.488$, MAE = £35k (25% of mean price).

Automated screening with mandatory human review in moderately-represented markets

Mixed areas like W01000114 achieving $R^2 = 0.326$, MAE = £42k but requiring expert validation.

Human-only valuation in poorly-represented markets

Cardiff Bay, unique contexts where automated approaches achieve $R^2 < 0.01$, MAE > £200k (65% of mean price).

Geographic error analysis confirms that one-size-fits-all national AVMs promising uniform accuracy across all contexts systematically over-promise and under-deliver, with prediction quality varying 9× across geographic segments (£35k to £310k MAE) despite presenting uniform confidence to users.

5. Conclusions

5.1. Summary of Key Findings

This study evaluated five fundamentally different property valuation methodologies across nine geographically diverse Welsh LSOAs, held out entirely from training data to test geographic Generalisation capability. The 9,606 test properties span contexts ranging from Cardiff Bay waterfront regeneration (W01002019) to rural Powys (W01000449) to post-industrial valleys (W01000114, Gwynedd), enabling systematic assessment of how valuation approaches perform when confronting property markets absent from training data.

Aggregate Performance Across 9 Test LSOAs

- Ridge Regression: R^2 = 0.261, MAE = £69,646 (n=9,606)
- CatBoost Gradient Boosting: R^2 = 0.282, MAE = £76,392 (n=9,606)
- K-Nearest Neighbours (Fuzzy Matching): R^2 = -1.997, MAE = £114,528 (n=9,606)
- Depreciated Replacement Cost (DRC Formula): R^2 = -0.227, MAE = £117,820 (n=9,606)
- LLM Ensemble (Zero-Shot): R^2 = 0.156, MAE = £91,291 (n=9,606)

CatBoost achieved the best within-LSOA performance (R^2 = 28.2%), explaining approximately one-quarter of property value variance within the excluded geographic areas. However, even this best-performing model shows substantial limitations. Three-quarters of variance remains unexplained and mean absolute errors of £76,392 exceed acceptable margins for high-stakes property valuations.

KNN and DRC both produced negative R^2 values, indicating predictions systematically worse than simply using the mean property price as a baseline. KNN's catastrophic R^2 = -1.997 reflects severe overfitting to training data geographic patterns that fail to transfer to new LSOAs. DRC's negative R^2 demonstrates that theory-based construction cost formulas (£1,732.79/square metres structure + £2,195.68/square metres land) systematically mis-calibrate when applied without local market adjustment.

Cross-LSOA Performance Heterogeneity

All models exhibited dramatic variation across the 9 test LSOAs, revealing that model quality depends primarily on geographic context rather than algorithmic sophistication:

- Best-case LSOA (W01001045, Bridgend): CatBoost R^2 = 0.488, MAE = £35,160
- Worst-case LSOA (W01002019, Cardiff Bay): CatBoost R^2 = 0.002, MAE = £165,983

The 14× difference in MAE between best and worst LSOAs (£35k vs £247k) demonstrates that a model appearing acceptable on average masks extreme heterogeneity. In suburban estates resembling broader Welsh patterns (W01001045, W01000517), models achieve near-professional accuracy (£35k-£48k errors). In unique regeneration contexts like Cardiff Bay, all models fail catastrophically with errors exceeding £247k across all five approaches.

Cardiff Bay (W01002019) uniformly produced the largest errors:

- Ridge: £174,127 MAE

- CatBoost: £165,983 MAE
- KNN: £282,327 MAE
- DRC: £263,388 MAE
- LLM: £270,104 MAE

The consistency of failure across fundamentally different model architectures (gradient boosting, instance-based, formula-based, large language model) indicates that Cardiff Bay represents a property market segment fundamentally absent from Wales-wide training data. Cardiff Bay constitutes just 0.4% of Welsh transactions, creating insurmountable representational gaps that no algorithmic sophistication can overcome.

Error Correlation Patterns

Analysis of error correlations between models reveals extremely high concordance ($\rho > 0.98$ for all pairs), indicating that all approaches make similar mistakes regardless of methodology:

- KNN-DRC error correlation: $\rho = 0.981$
- KNN-LLM error correlation: $\rho = 0.981$
- DRC-LLM error correlation: $\rho = 0.998$

These near-perfect correlations indicate that errors are driven by property-level characteristics that all models fail to capture, rather than model-specific architectural limitations. Properties systematically undervalued by KNN tend to be equally undervalued by DRC and LLM, suggesting blind spots in the available feature space rather than fixable algorithmic deficiencies.

The DRC-LLM correlation of 0.998 is particularly revealing: despite the LLM ensemble using sophisticated zero-shot reasoning with multiple specialist agents, its errors mirror the simple formula-based DRC approach almost perfectly. This demonstrates that the LLM's learned representations do not provide meaningful additional information beyond basic hedonic assumptions when geographic context is missing from training data.

Property Type Stratification

Errors vary substantially by property type, with detached properties showing the highest errors across all models:

Table 43: Comparison of MAE by Property Type for KNN, DRC & LLM Models

Property Type	KNN MAE	DRC MAE	LLM MAE	Sample%
Detached	£114,747	£112,182	£90,033	19.4% (KNN/DRC), 22.2% (LLM)
Semi-detached	£66,186	£83,311	£57,174	20.8% (KNN/DRC), 23.6% (LLM)
Terraced	£81,773	£88,223	£50,181	18.8% (KNN/DRC), 21.4% (LLM)

Property Type	KNN MAE	DRC MAE	LLM MAE	Sample%
Flat	£73,443	£56,338	£56,333	34.8% (KNN/DRC), 28.0% (LLM)

Detached properties exhibit 1.6× higher errors than semi-detached homes for KNN, and 2× higher than flats for DRC. This pattern likely reflects greater heterogeneity in detached property characteristics: plot sizes, garden quality, views, and micro-location premiums vary enormously for detached homes but are more standardised for terraced houses and flats. Models trained without plot-level data cannot distinguish a detached house on 0.1 acres from one on 2 acres, creating systematic misprediction.

Interestingly, flats show the lowest DRC errors (£56,338) despite constituting the largest segment (34.8% of KNN/DRC test set). This suggests the DRC formula's structure-plus-land approach works better for apartment buildings where land value per unit is small and construction costs dominate. Conversely, the formula systematically fails for detached properties where land premiums and plot characteristics drive substantial value variation.

Temporal Error Evolution

Analysis of errors by transaction year reveals dramatic temporal instability, with errors surging 7× during 2015-2019 before contracting 40% during the pandemic:

Table 44: Temporal Trends in KNN and DRC Valuation Errors by Period

Period	KNN MAE	DRC MAE	Interpretation
2013-2014 baseline	£111k-£124k	£97k-£115k	Moderate errors
2015-2016 surge	£249k-£369k	£237k-£349k	Rapid increase
2017 peak	£533,536	£489,177	Catastrophic failure
2018-2019 plateau	£409k-£342k	£382k-£324k	Elevated errors
2020-2021 contraction	£123k-£272k	£174k-£244k	44% reduction (KNN), 39% (DRC)
2022-2024	£182k-£219k	£154k-£182k	Stabilization above baseline

The extraordinary 2017 error peak (£533k MAE for KNN, 7× the 2014 baseline) coincides with Cardiff Bay transactions during the waterfront development boom. This suggests that rapid localised appreciation creates pricing dynamics entirely outside training data patterns.

When Cardiff Bay properties appreciated 8-12% annually while valleys appreciated 0-2% annually during this period, models interpreting Wales-wide relationships systematically misprice the divergent trajectories.

The pandemic-era contraction (44% error reduction for KNN, 39% for DRC from 2018-2019 to 2020-2021) appears counterintuitive given market disruption. Two explanations emerge: (1) compositional changes; fewer ultra-high-value Cardiff Bay transactions during lockdowns reduced the influence of extreme outliers on aggregate error metrics, or (2) temporary price compression brought Cardiff Bay closer to Wales-wide patterns as buyers shifted preferences away from urban waterfront premium.

The post-pandemic stabilization (2022-2024) at elevated levels (£182k-£219k vs £111k-£124k pre-2015 baseline) suggests permanent structural changes in test LSOA markets that diverge from training distribution. Geographic heterogeneity creates moving targets even with perfect features.

Spatial Error Patterns

Spatial analysis reveals weak autocorrelation but strong distance-based gradients from Cardiff city centre:

Moran's I Spatial Autocorrelation

- KNN: I = 0.012 (barely above random expectation of -0.000)
- DRC: I = 0.009 (barely above random expectation of -0.000)

The extremely weak Moran's I values (0.009-0.012) contradict expectations of strong geographic error clustering. This indicates errors are driven more by LSOA-level systematic biases and property-specific factors than by fine-grained neighbourhood effects. A property's error cannot reliably predict neighbouring properties' errors at the 2km scale, limiting the utility of spatial smoothing or kriging approaches for error correction.

Distance to Cardiff Error Gradients

Despite weak spatial autocorrelation, errors show strong systematic patterns by distance from Cardiff city centre:

Table 45: Distance to Cardiff vs Valuation Error for KNN and DRC Models

Distance Range	KNN Mean Error	KNN MAE	DRC Mean Error	DRC MAE
0.0-0.6km	+£300,447	£310,642	+£222,572	£295,695
0.6-35.9km	+£103,631	£129,168	+£34,418	£115,552
35.9-132.3km	+£23,645	£85,395	-£39,286	£88,648
132.3-189.7km	+£12,567	£91,233	-£35,706	£106,526

Properties within 0.6km of Cardiff city centre (primarily Cardiff Bay waterfront) are massively underpredicted by both models, with mean errors of +£300k (KNN) and +£223k (DRC). The gradient declines steeply: KNN errors drop by £197k from the 0-0.6km bin to the 0.6-36km bin, indicating an extremely localized Cardiff Bay premium that models cannot capture.

DRC shows sign reversal illustrated by underprediction within 36km of Cardiff (+£222k to +£34k) transitioning to overprediction beyond 36km (-£39k to -£36k in rural Wales). This indicates the formula's fixed construction cost and land value parameters mis-calibrate at both extremes; underestimating central Cardiff waterfront, overestimating distant rural properties.

The magnitude of the Cardiff proximity effect (£300k error within 0.6km) far exceeds typical urban distance-decay patterns, suggesting this bin captures Cardiff Bay's unique waterfront regeneration premium rather than standard urban gradients. This validates the LSOA-level finding that Cardiff Bay requires human valuation - no amount of algorithmic sophistication addresses the combination of hyper-localised £300k+ premium and representational gaps (0.4% of transactions).

Bias Decomposition by LSOA

Decomposition of mean squared error into systematic bias (bias²) versus random variance reveals that most LSOAs are variance-dominated, indicating property-specific factors drive errors more than systematic over/underprediction:

Table 46: Systematic Bias vs Variance in KNN and DRC Errors by LSOA

LSOA	Model	Systematic Bias	RMSE	Bias²/MSE	Pattern
W01000114 (Gwynedd 009D)	DRC	-£87,058	£135,344	41.4%	BIAS-DOMINATED
W01001597 (Monmouthshire)	KNN	+£98,037	£175,772	31.1%	BIAS-DOMINATED
W01000517 (Ceredigion 002D)	KNN	+£77,783	£128,771	36.5%	BIAS-DOMINATED
W01002019 (Cardiff Bay)	KNN	+£273,103	£1,665,627	2.7%	Variance-dominated
W01002019 (Cardiff Bay)	DRC	+£192,507	£1,679,284	1.3%	Variance-dominated
W01000255 (Flintshire 015A)	KNN	+£42,229	£320,973	1.7%	Variance-dominated
W01000617 (Pembrokeshire 002F)	KNN	+£12,256	£154,701	0.6%	Variance-dominated

Key Findings

Gwynedd's systematic failure

W01000114 shows the highest bias contribution (41.4% of MSE for DRC), with systematic -£87k underprediction. The DRC formula's £1,732.79/square metres construction costs and £2,195.68/square metres land values fundamentally mis-calibrate for Gwynedd's market structure. This is actionable through formula recalibration.

KNN's suburban overprediction

KNN shows 31-37% bias contribution in Monmouthshire and Bridgend, both with large positive bias (+£78k, +£98k). This suggests KNN retrieves comparables from higher-value training areas when confronted with suburban/commuter markets, systematically overvaluing these properties.

Cardiff Bay paradox

Despite massive mean bias (£273k KNN, £193k DRC), bias²/MSE is only 1-3%. This occurs because enormous within-LSOA variance (σ=£1.67M RMSE) dwarfs even £250k systematic bias. Cardiff Bay contains properties from £150k flats to £2M+ penthouses, creating such extreme heterogeneity that systematic bias becomes statistically minor relative to total error variance. The solution requires both systematic bias correction (+£250k average adjustment) and feature enrichment to distinguish £150k flats from £2m penthouses.

Variance dominance

13 of 18 LSOA-model combinations show bias²/MSE < 10%, indicating property-specific factors (unobserved quality, condition, micro-location) drive most error variance rather than systematic LSOA-wide misprediction. Improving accuracy requires feature enrichment (plot size, condition, parking, views) more than addressing systematic bias.

5.2. Drivers of Property and Land Value across Wales

Residual land valuation was performed by subtracting estimated structure costs from total property values to isolate land premiums. Two approaches were tested: (1) CatBoost building-only predictions as structure value estimates, (2) DRC formula structure cost calculations.

Decomposition Results Across 9 Test LSOAs (n=9,606)

CatBoost Building-Only Approach

- Median land value: £10,975
- Mean land value: £75,452
- Median structure value: £129,135
- Mean structure value: £160,330

- Median land share: 7.7% of total property value
- Mean land share: -95.6% of total property value (negative due to negative land values)

43.6% of properties have negative implied land values (4,189 properties).

DRC Formula Approach

- Median land value: £56,602
- Mean land value: £61,440
- Median structure value: £131,304
- Mean structure value: £137,442
- Median land share: 40.1% of total property value
- Mean land share: 59.6% of total property value

0% of properties have negative land values (by construction, DRC formula prevents this).

Scope limitation

Residual land valuation in this analysis is only possible where there is an observed property sale price and an estimable structure value or cost that can be subtracted. As a result, the method mainly captures land attached to transacted built-up property (predominantly residential) and excludes many other land types, including vacant or bare land, much agricultural land, many commercial or industrial sites, infrastructure land, and specialist and public-sector land. This occurs because the residual approach depends on a market price for the combined asset (land + buildings), which is often unavailable, infrequent, or not directly comparable for non-residential and non-transacted land. The findings should therefore be interpreted as evidence on residential land premiums, rather than a complete valuation of all land across Wales

Critical Finding: Residual Land Valuation Failure

The CatBoost residual approach produces 43.6% negative land values; an economic impossibility; indicating systematic failure of residual valuation methodology when applied to models trained on Wales-wide data and tested on excluded LSOAs.

This failure occurs because CatBoost overpredicts structure values for many test LSOA properties, producing implied land values (actual price minus predicted structure cost) that are negative. A property cannot have negative land value; this result indicates the model systematically mis calibrates structure valuations when confronting geographic contexts absent from training data.

The DRC formula avoids negative land values by construction (it explicitly calculates land value as a positive component), but produces implausibly high land shares (median 40%, mean 60%). Economic literature suggests land comprises 30-50% of property values in typical markets, but this varies enormously by context; from 10% in rural areas to 80% in central urban locations. The DRC's fixed parameters cannot capture this geographic variation.

The 32% point difference in mean land share between methods (CatBoost -96%, DRC +60%) demonstrates that land value decomposition depends critically on accurate structure

cost estimation. When models trained on Wales-wide data fail to generalise to new geographic areas, residual land valuations become unreliable.

Implications for Land Value Understanding

Geographic variation in land premiums

The distance-to-Cardiff gradients (£300k errors within 0.6km declining to £12k at 132-190km) suggest land values drive the Cardiff Bay premium. Properties near city centre command enormous land premiums that fixed formulas cannot capture.

Property type and land share

Flats show lowest errors (£56k DRC MAE) because land value per unit is small structure costs dominate. Detached properties show highest errors (£112k DRC MAE) because land premiums and plot characteristics vary enormously, but these are unobserved in training data.

Residual methods fail without geographic Generalisation

The 43.6% negative land value rate demonstrates that residual land valuation requires models that generalise well to target geography. Cannot reliably decompose land vs structure values using models trained on different contexts.

Explicit land modelling required

Land value estimation requires explicit approaches (DRC-style formulas calibrated to local markets, comparable land sales analysis, spatial hedonic models with distance/accessibility features) rather than residual methods when property valuation models show poor R^2.

5.3. Confidence in Model Robustness

The Geographic Generalisation Problem

The consistently poor overall R^2 values (near zero or negative) combined with moderate within-LSOA R^2 values (20-30% for best models) reveal a fundamental limitation: models trained on Wales-wide data cannot generalise to new geographic areas, even when property-level features (floor area, property type, age, postcode) are included.

This finding contradicts common assumptions that hedonic models capturing property characteristics will transfer across regions. Instead, the evidence demonstrates that LSOA-level context; not captured by observed features; dominates prediction quality. The same CatBoost model achieving $R^2 = 0.488$ in suburban Rhondda (W01001045) simultaneously achieves $R^2 = 0.002$ in Cardiff Bay (W01002019), indicating that geographic context, not algorithm choice, determines success.

Training Data Representational Gaps

Cardiff Bay constitutes 0.4% of Welsh land transactions. No model trained predominantly on non-Cardiff-Bay properties can learn Cardiff Bay pricing relationships from such sparse representation, regardless of architectural sophistication. The waterfront premium, regeneration characteristics, and modern apartment developments create a market segment fundamentally absent from the Wales-wide training distribution.

This is not a solvable problem through more sophisticated algorithms; the evidence is clear that Ridge, CatBoost, KNN, DRC, and LLM all fail with similar magnitudes (£247k-£282k MAE) despite fundamentally different architectures. The failure is representational, not algorithmic.

Temporal Non-Stationarity

The 7× error surge in 2015-2019 (from £111k to £533k MAE) followed by 40% pandemic contraction demonstrates that fixed models trained on historical data become progressively mis calibrated as local markets evolve at different rates. Cardiff Bay appreciated 8-12% annually while valleys appreciated 0-2% annually during 2015-2024, creating divergent trajectories that models interpreting average relationships systematically misprice.

Even if perfect features were available and representational gaps addressed, temporal non-stationarity creates moving targets. Models must be continuously retrained on recent transactions to maintain calibration; a fundamental limitation for automated valuation systems.

Feature Limitations

The absence of critical property characteristics in the training data creates insurmountable prediction barriers:

- Missing features: Plot size/boundaries, parking spaces, garden quality, renovation history, views, micro-accessibility (distance to nearest station, school, amenity), build quality indicators, internal condition
- Available features: Floor area (38% coverage), property type (D/S/T/F), age (old/new binary), postcode, LSOA, transaction date
- Impact: Even best-performing CatBoost explains only 28.2% of within-LSOA variance, leaving 71.8% unexplained due to missing property-specific information

The 43.6% negative land value rate in residual decomposition further illustrates feature gap consequences: without accurate structure valuations, even basic economic decompositions become unreliable.

Model Architecture Makes Minimal Difference

The near-perfect error correlations ($\rho > 0.98$) demonstrate that algorithm choice has minimal impact on which properties are mis-valued. The errors are property-driven and context-driven, not algorithm-driven. Improving valuation accuracy requires:

- Better geographic representation: Include training examples from all target contexts, or accept that some markets require human valuation

- Feature enrichment: Obtain plot-level data (boundaries, parking, gardens) and renovation/condition indicators
- Continuous recalibration: Retrain models frequently on recent transactions to address temporal non-stationarity

No amount of architectural sophistication (deeper neural networks, more sophisticated ensembles, larger language models) addresses these fundamental data limitations.

Operational Deployment Implications

National-scale automated valuation systems cannot achieve uniformly acceptable accuracy across heterogeneous property markets. Three deployment strategies emerge:

Tiered confidence deployment

High-confidence automated valuation (R^2 = 0.40-0.50, MAE = £35k-£50k): Suburban estates like W01001045, W01000517 where training data adequately represent market patterns.

Automated screening with mandatory human review (R^2 = 0.25-0.35, MAE = £60k-£80k): Mixed urban/valley areas like W01000114 where models provide signal but require validation

Human-only valuation (R^2 < 0.10, MAE > £200k): Regeneration zones like W01002019, unique contexts fundamentally absent from training distribution

Geographic performance monitoring

Deploy separate accuracy metrics by LSOA type (urban regeneration, suburban, valley, rural, border) rather than Wales-wide averages. A model with £76k overall MAE masks 14× variation (£35k to £247k) across contexts, systematically over-promising in difficult areas while under-utilising capability in easy areas.

Accept fundamental limitations

Some property markets (Cardiff Bay waterfront, unique rural estates, border regions with England proximity premiums) will always require human expertise. The 0.4% representation problem cannot be solved algorithmically when the context is fundamentally different from training data.

Robustness Assessment Summary

- Models achieve acceptable accuracy in well-represented suburban contexts (R^2 = 0.40-0.50, MAE = £35k-£50k)
- Computational scalability demonstrated for Ridge/CatBoost (millions of properties, seconds inference time)
- Systematic bias patterns identified and actionable (Gwynedd DRC, Bridgend KNN)
- Models fail catastrophically in unique contexts ($R^2 \approx 0$, MAE > £200k)

- Geographic Generalisation remains unsolved across all tested architectures
- Temporal instability creates moving targets (7× error variation 2014-2017)
- Feature gaps prevent accurate property-specific valuation (71.8% unexplained variance)
- Algorithm choice provides minimal improvement (ρ > 0.98 error correlation)
- LLM approach not economically scalable (£15k-£75k per million properties)
- KNN inference scales poorly without approximation methods

The evidence demonstrates that current approaches achieve partial success but cannot deliver the uniform accuracy required for fully automated national-scale property valuation systems. A hybrid approach combining automated screening for well-represented contexts with human review for unique markets provides the most defensible operational strategy.

Confidence Levels by Deployment Context

Table 47: Recommended Valuation Deployment Strategy

Context Type	Example LSOAs	Best Model R^2	MAE	Confidence	Deployment Strategy
Suburban estates	W01001045, W01000517	0.38-0.49	£35k-£48k	High	Automated valuation
Valley towns	W01000114, W01001233	0.03-0.33	£42k-£63k	Medium	Automated + review
Rural areas	W01000449	0.08	£75k	Low	Human-led with model input
Urban regeneration	W01002019	0.00	£247k	None	Human-only valuation

This tiered approach acknowledges that valuation confidence depends fundamentally on training data representation and market homogeneity, not algorithmic sophistication.

Core Findings

This study evaluated five valuation methodologies; Ridge Regression, CatBoost, K-Nearest Neighbours, Depreciated Replacement Cost formula, and LLM Ensemble; across nine Welsh LSOAs excluded from training data, testing geographic Generalisation capability on 9,606 properties.

- CatBoost achieved best performance (R^2 = 28.2%, MAE = £76,392) but explains less than one-third of within-LSOA variance. KNN and DRC failed catastrophically with negative R^2 values.
- Geographic context dominates model quality: 14× variation in MAE across LSOAs (£35k to £247k). Suburban contexts achieve R^2 = 0.40-0.50; unique contexts like Cardiff Bay achieve $R^2 \approx 0$.

- All models make similar mistakes (error correlations ρ > 0.98), indicating errors are property-driven and context-driven, not algorithm-driven.
- Cardiff Bay failure is systematic: 0.4% training representation creates insurmountable Generalisation barriers. All five models fail with £247k-£282k MAE despite fundamentally different architectures.
- Temporal instability: Errors surged 7× during 2015-2019 localized appreciation, then contracted 40% during pandemic. Fixed models become mis calibrated as markets diverge.
- Spatial patterns: Weak spatial autocorrelation (Moran's I = 0.009-0.012) but strong distance gradients (£300k errors within 0.6km of Cardiff declining to £12k at 132km+).
- Property type effects: Detached properties show 2× higher errors than flats (£112k vs £56k DRC MAE) due to unobserved plot characteristics.
- Residual land valuation fails: 43.6% negative land values for CatBoost residual approach demonstrates that land decomposition requires models that generalise well to target geography.
- Feature gaps dominate: 71.8% unexplained variance due to missing plot size, parking, renovation history, condition, views. Floor area observed for only 38% of properties.
- Scalability varies dramatically: Ridge/CatBoost scale computationally to millions of properties (seconds inference); LLM approach catastrophically expensive (£15k-£75k per million properties at API rates); no model scales geographically without tiered deployment.

Operational Implications

National-scale automated valuation cannot achieve uniform accuracy. Deployable strategy requires:

- High-confidence automated valuation for well-represented suburban contexts (R^2 = 0.40-0.50)
- Automated screening with human review for moderate contexts (R^2 = 0.25-0.35)
- Human-only valuation for unique contexts (R^2 < 0.10)

The 0.4% representation problem for contexts like Cardiff Bay cannot be solved algorithmically. Geographic heterogeneity combined with sparse representation creates fundamental Generalisation barriers that no architectural sophistication addresses. Improving accuracy requires better geographic representation, feature enrichment (plot-level data), and continuous recalibration, not more sophisticated algorithms.

5.4. Ease and Difficulty of Valuation

Systematic Patterns in Valuation Difficulty

The 14× variation in MAE across the 9 test LSOAs (£35k to £247k) reveals systematic patterns in which property contexts are easy versus difficult to value:

Easy-to-Value Contexts (MAE < £50k, R^2 > 0.35)

W01001045 (Bridgend): CatBoost MAE = £35,160, R^2 = 0.488

Characteristics: Suburban/valley estate, homogeneous housing stock, well-represented in training data.

Why easy: Property types and price ranges similar to broader Welsh patterns. Models learn appropriate hedonic relationships that transfer.

W01000517 (Ceredigion 002D): CatBoost MAE = £48,120, R^2 = 0.383

Characteristics: Mixed suburban or commuter town, standard property types

Why easy: Despite 36.5% bias contribution (KNN overvalues by £78k), within-LSOA variance is moderate. Systematic bias is correctable.

Moderate-Difficulty Contexts (MAE £60k-£80k, R^2 0.25-0.35)

W01000114 (Gwynedd 009D): CatBoost MAE = £42,423, R^2 = 0.326

Characteristics: Urban post-industrial, mixed housing quality

Why moderate: DRC shows 41.4% bias contribution (systematic -£87k underprediction), but CatBoost achieves acceptable R^2. Models capture some patterns but miss market-specific factors.

W01001233 (Rhondda Cyon Taf 001F): CatBoost MAE = £62,925, R^2 = 0.034

Characteristics: Valley town, mix of old terraces and newer estates

Why moderate: Extremely low R^2 (3.4%) despite moderate MAE suggests correct average estimates with high individual variation. Unobserved property condition drives errors.

Difficult-to-Value Contexts (MAE £100k-£200k, R^2 0.00-0.15):

W01000449 (Powys, rural): CatBoost MAE = £74,859, R^2 = 0.076

Characteristics: Rural or agricultural, sparse properties, unique characteristics

Why difficult: KNN achieves catastrophic R^2 = -15.565. Rural properties have unique features (land acreage, agricultural buildings, conservation restrictions) not captured by training data. Low transaction volumes prevent learning local patterns.

W01000255 (Flintshire 015A): CatBoost MAE = £59,734, R^2 = 0.153

Characteristics: Student or rental area near city centre, mixed quality
Why difficult: High within-area heterogeneity. Student housing, professional rentals, and owner-occupied properties mix within single LSOA. Unobserved renovation quality drives errors.

Impossible-to-Value Contexts (MAE > £200k, $R^2 \approx 0$):

W01002019 (Cardiff Bay): CatBoost MAE = £165,983, R^2 = 0.002

Characteristics: Waterfront regeneration, modern apartments, luxury penthouses

Why impossible: Represents 0.4% of Welsh transactions. Waterfront premium (+£300k within 0.6km of city centre), regeneration characteristics, and extreme within-LSOA heterogeneity (£150k flats to £2m penthouses) create fundamental representational gaps. No algorithmic sophistication addresses 0.4% representation problem.

Property-Specific Difficulty Factors

Beyond LSOA context, individual property characteristics drive valuation difficulty. Land valuation is inferred primarily from observed sales of built-up residential property rather than bare land transactions. This means land is generally easier to value where the residential asset is standardised (e.g., flats and homogeneous terraces), because abundant comparables and predictable building characteristics provide a stronger signal for the underlying land component

Easy-to-Value Properties

- Flats in apartment blocks: Standardised characteristics, small land component, comparable sales abundant. DRC MAE = £56,338 (lowest property type error).
- Terraced houses in homogeneous estates: Limited plot variation, standard layouts, abundant comparables. LLM MAE = £50,181 for terraced properties.
- Properties with complete feature coverage: When floor area is observed (38% of dataset), models have critical size information. Missing floor area forces models to infer from property type and postcode alone.

Moderate-Difficulty Properties

- Semi-detached suburban homes: Plot size varies but within narrow range. Garden quality and parking introduce variance. MAE = £57k-£83k across models.
- Properties in mixed neighbourhoods: Surrounding property mix creates valuation uncertainty. Is this property upgraded or original condition? Models cannot distinguish without renovation history.

Difficult-to-Value Properties

- Detached houses: Extreme plot size variation (0.1 to 2+ acres), garden quality, views, privacy, micro-location premiums. MAE = £90k-£115k across models, 2× higher than flats.
- Properties with unique features: Period features, conservation restrictions, listed building status, unusual layouts, commercial conversion; all unobserved in training data.

- Properties in transitional areas: Gentrifying neighbourhoods where comparable sales span wide quality range. Models cannot determine if property benefited from gentrification.

Temporal Difficulty Patterns

Valuation difficulty varies systematically by transaction period:

- Easy periods (2002-2012): Stable markets, moderate appreciation, consistent patterns. MAE = £61k-£76k (DRC).
- Moderate periods (2013-2014, 2022-2024): Modest appreciation, some regional divergence. MAE = £97k-£219k.
- Difficult periods (2015-2019): Rapid localised appreciation, divergent regional trajectories. MAE = £237k-£533k. Cardiff Bay appreciates 8-12% annually while valleys stagnate, creating unprecedented regional divergence.
- Unpredictable periods (2020-2021): Pandemic disruption, compositional changes, preference shifts. Errors decline 40% but for unclear reasons (fewer high-value transactions? Price compression?).

Operational Implications

Automated valuation confidence should vary by context:

- High confidence (MAE < £50k): Suburban estates (W01001045, W01000517), flats in standard blocks, terraced houses in homogeneous areas. 40-49% R^2 indicates acceptable predictive power.
- Medium confidence (MAE £50k-£100k): Valley towns (W01000114, W01001233), semi-detached properties, mixed neighbour hoods. 25-35% R^2 requires human review for high-stakes decisions.
- Low confidence (MAE £100k-£200k): Rural areas (W01000449), student/rental districts (W01000255), detached properties with large plots. R^2 < 15% requires professional valuation.
- No confidence (MAE > £200k): Regeneration zones (W01002019), waterfront properties, listed buildings, properties with unique features. $R^2 \approx 0$ mandates human-only valuation.

The systematic variation in valuation difficulty demonstrates that automated systems cannot achieve uniform accuracy across heterogeneous property markets.

5.5. Scalability

Computational Scalability

The five valuation approaches tested exhibit dramatically different computational requirements for training and inference:

Training Scalability

- Ridge Regression: Linear in sample size and features. Training on millions of properties with stacked base learners completes in minutes on standard hardware. Highly scalable to millions of properties.
- CatBoost: Near-linear scalability due to efficient gradient boosting implementation. Training on millions of properties completes in under 1 hour on CPU. Scales to millions of properties with distributed training.
- KNN: No training phase; all computation occurs at inference. Building spatial indices for millions of properties is fast (seconds to minutes). Perfect training scalability.
- DRC Formula: No training required. Formula application is instant. Infinite training scalability.
- LLM Ensemble: No training phase for zero-shot approach. Uses pre-trained models (haiku). Training scalability not applicable; inference cost is the constraint.

Inference Scalability

- Ridge Regression: O(n*f) linear matrix multiplication. Predictions for 9,606 properties complete in milliseconds. Trivially scalable to millions of predictions.
- CatBoost: O(ntreesdepth) tree traversal. Predictions for 9,606 properties complete in seconds. Highly scalable with GPU acceleration.
- KNN: O(nklog(N)) nearest neighbour search. Predictions for 9,606 properties require searching 1 million + training examples per property. Completes in minutes but scales poorly, doubling training data doubles search time. Approximation methods (LSH, ANNOY) required for national-scale deployment.
- DRC Formula: O(n) formula evaluation. Predictions for 9,606 properties complete in milliseconds. Perfect inference scalability.
- LLM Ensemble: O(ntokenscost) API calls. Predictions for 9,606 properties required thousands of API calls at $3-15 per 1M tokens. Catastrophically expensive at scale: Valuing 1 million properties would cost £15,000-£75,000 in API fees alone. Not scalable for national deployment.

Data Scalability

Training data size impacts models differently:

- KNN: Performance degrades with more data if new properties are not geographically similar to queries. Cardiff Bay performance does not improve by adding more valley properties; needs more Cardiff Bay examples specifically. Data quality (geographic representativeness) matters more than quantity.
- DRC: Ignores training data entirely. Adding more observations provides zero benefit unless used to recalibrate formula parameters (construction costs, land values) by region.
- LLM: Zero-shot approach ignores training data. Performance depends entirely on pre-training corpus, which is fixed.

Geographic Scalability

Extending models from Wales to UK-wide deployment reveals critical limitations:

- Ridge and CatBoost: Cardiff Bay failure ($R^2 \approx 0$) demonstrates that geographic diversity requires either (1) training examples from all target contexts, or (2) explicit

geographic stratification (separate models per region). UK-wide deployment would require 100+ LSOA-type contexts represented in training, or tiered deployment (automated for well-represented contexts human review for unique contexts).

- KNN: Requires representative training examples from all target geographies. Cannot value London Docklands properties using Welsh comparables. UK-wide deployment requires comprehensive national training dataset covering all market segments.
- DRC: Formula parameters (£1,732.79/square metres structure, £2,195.68/square metres land) must be recalibrated for each region. London construction costs are 2-3× Welsh costs; land values vary 10-100×. UK-wide deployment requires regional parameter sets, undermining the simplicity advantage.
- LLM: Zero-shot approach theoretically handles new geographies, but 0.998 correlation with DRC errors suggests it defaults to basic hedonic assumptions. UK-wide deployment likely produces same geographic failures as Wales deployment.

Operational Scalability Summary

- Ridge and CatBoost scale computationally to millions of properties (seconds to minutes for inference)
- DRC scales computationally (instant inference) but requires regional recalibration
- KNN scales poorly at inference (minutes per 10k properties) without approximation
- LLM does not scale economically (£15k-£75k per million properties at API rates)
- No model scales geographically without either comprehensive representative training data or tiered confidence deployment

National-scale deployment requires Ridge or CatBoost for computational efficiency, with tiered deployment strategy to handle geographic heterogeneity.

6. Further Considerations

6.1. Implementation and Governance

Data Integration and Refresh Cycles

Quarterly Land Registry Transaction Updates

The valuation system requires quarterly re-fresh of Land Registry Price Paid Data through automated data processing pipelines. This regular update cycle addresses the temporal non-stationarity problem identified in the study. Errors surged 7× during 2015-2019 Cardiff Bay appreciation, then contracted 40% during pandemic, demonstrating that fixed models become mis calibrated as markets evolve at different rates. Quarterly retraining allows models to adapt to emerging trends while balancing accuracy and operational efficiency.

Integrated Energy Performance Certificate (EPC) and Postcode Data

Energy Performance Certificates provide floor area data, which is critical but available for only 38% of properties in the current study. Office for National Statistics Postcode Directory provides coordinates (available for 100% of properties in the study) and rural/urban classifications. Around 62% of properties lack a direct EPC match (address/UPRN), although using a postcode-level EPC proxy increases floor-area availability to around 74%. The remaining missingness means models often have to predict without reliable size information, which likely contributes to the 71.8% unexplained variance. Improving EPC linkage and coverage should therefore be a priority for accuracy gains.

Annual Satellite and Spatial Data Updates

Satellite-derived data on accessibility measures and environmental factors must be refreshed annually. However, the study demonstrates that even without these features, geographic context dominates: the same CatBoost model achieves $R^2 = 0.488$ in suburban Rhondda but $R^2 = 0.002$ in Cardiff Bay. Adding spatial features may improve within-context accuracy but will not solve the geographic Generalisation problem (0.4% training representation creates insurmountable barriers).

Land that has not been valued

Land that has not been valued (or has no recent transaction evidence) can still be estimated, but not with the same level of confidence as transacted residential property because there is no direct market price to use as a benchmark.In these cases, valuation must rely more heavily on indirect evidence (e.g. nearby comparables, planning status, permitted use, constraints, and location factors) and professional assumptions. The residential transactions provide the most consistent, frequent, and observable market evidence (sale prices plus property attributes), which is essential for training and validating the models in a fair and comparable way.

Governance and Quality Assurance

Independent Oversight and Audits

Independent oversight by an external review committee is essential for maintaining public trust. Annual audits must verify that models operate fairly across different regions, property types, and socio-economic groups. The study identified systematic bias patterns:

- Gwynedd: DRC shows -£87k systematic underprediction (41.4% bias contribution)
- Bridgend/Vale: KNN shows +£78k to +£98k systematic overprediction (31-37% bias contribution)
- Cardiff Bay: All models show +£193k to +£273k systematic underprediction

Audits must monitor these bias patterns and ensure they do not systematically disadvantage specific communities.

Formal Appeals Process

The study demonstrates that 14× variation in MAE across LSOAs (£35k to £247k) creates substantial valuation uncertainty. Appeal rates will likely exceed 1-2% in difficult-to-value contexts:

- High-confidence contexts (suburban estates): Expected <1% appeals (MAE = £35k-£50k, R^2 = 0.40-0.50)
- Medium-confidence contexts (valley towns): Expected 2-5% appeals (MAE = £60k-£80k, R^2 = 0.25-0.35)
- Low-confidence contexts (rural areas): Expected 5-10% appeals (MAE = £75k-£100k, R^2 < 0.15)
- Regeneration zones (Cardiff Bay): Expected 20%+ appeals (MAE > £200k, $R^2 \approx 0$)

Establishing structured appeals workflow is critical, as error rates vary dramatically by geographic context.

Complete Audit Trail

Each valuation must be traceable to specific model version, input features, and training period. This is essential for defending appeals and identifying model drift. The study's temporal analysis shows dramatic performance variation: 2017 peak MAE = £533k, 2020-2021 contraction to £123k-£272k, 2022-2024 stabilisation at £182k-£219k. Audit trails enable detection of such performance shifts.

Monthly Quality Checks Against Manual Valuations

Random sampling of 50-100 properties per month compared against human valuations provides ongoing monitoring. However, sampling must stratify by LSOA type; not use random Wales-wide sampling. A random sample would predominantly capture easy-to-value suburban contexts (where models perform well) while missing catastrophic failures in unique contexts like Cardiff Bay. Stratified sampling ensures monitoring detects geographic performance heterogeneity.

Maintenance and Monitoring

The model requires retraining every quarter, taking approximately 2-3 person-days of work for the core modelling tasks:

- Data processing and cleaning: 0.5 days
- Feature engineering and validation: 0.5 days
- Model training (CatBoost): 0.5 days (1 hour compute time + validation)
- Performance testing and documentation: 0.5-1.0 days

However, this estimate covers only the technical retraining. Additional requirements include:

- Error analysis and bias monitoring: 1-2 days
- Stakeholder reporting and dashboard updates: 0.5-1.0 days
- Investigation of performance degradation alerts: Variable (0-5 days depending on issues)

Total quarterly effort: 4-10 person-days depending on complexity of issues discovered.

Frequent retraining addresses temporal non-stationarity, but the study demonstrates this alone cannot solve geographic Generalisation failure. Cardiff Bay (0.4% training representation) will continue to show poor performance regardless of retraining frequency.

Integration and Access

Application Programming Interface (API)

A standardized REST API enables secure, automated retrieval of valuations by internal Welsh Government systems. API design must include:

- Property identifier input (UPRN or address)
- Parcel-based inputs for land (e.g., title number, parcel ID, etc)
- Valuation output with confidence intervals
- Feature importance explanation for transparency
- LSOA-specific accuracy metrics (not Wales-wide average)
- Flag for low-confidence predictions ($R^2 < 0.15$ threshold)

Geographic Information System (GIS) Exports

Providing GIS data layers in Shapefile or GeoJSON format allows analysts to map land values, land share percentages, and valuation accuracy across LSOAs. The study's spatial analysis results (Moran's I, distance gradients, bias decomposition) should be included as reference layers to help users understand where models perform well versus poorly.

Interactive Dashboard for Policy Analysts

A user-facing dashboard should support:

- Scenario exploration (e.g., testing different depreciation assumptions)
- LSOA-level performance comparison (R^2, MAE, bias decomposition)
- Temporal trend monitoring (detecting 2015-2019 surge patterns early)
- Systematic bias alerts (flagging when mean error exceeds ±£50k threshold)

Critical requirement

Dashboard must display performance by LSOA type, not Wales-wide averages. The study demonstrates that £76k overall MAE masks 14× variation, systematically over-promising in difficult areas.

Public Portal for Transparency

A public-facing interface should display:

- Estimated property value with confidence interval (±15-30% depending on LSOA)
- Top 5 influencing factors (when feature importance is available)
- Geographic context: "Properties in your LSOA show R^2 = X, MAE = £Y"
- Explicit warning for low-confidence contexts: "Your property is in a difficult-to-value area. Accuracy may be limited. Human review recommended."

Transparent communication about geographic performance heterogeneity is essential for managing expectations.

6.2. Recommended Approach for Wales

Primary Method: Gradient Boosting (CatBoost)

Purpose: Mass valuation of residential properties in well-represented contexts

- Deploy for suburban estates, valley towns, standard property types
- Retrain quarterly with new transaction data
- Provide confidence intervals and explicit low-confidence flags
- Coverage: ~60-70% of Welsh properties (in well-represented contexts where R^2 > 0.25)

Performance expectations by context:

- Suburban estates: R^2 = 0.40-0.50, MAE = £35k-£50k (HIGH CONFIDENCE)
- Valley towns: R^2 = 0.25-0.35, MAE = £60k-£80k (MEDIUM CONFIDENCE)
- Rural areas: R^2 = 0.05-0.15, MAE = £75k-£100k (LOW CONFIDENCE - hu-man review required)
- Regeneration zones: $R^2 \approx 0$, MAE > £200k (NO CONFIDENCE - human-only valuation)

Secondary Method: Depreciated Replacement Cost (DRC) for Land Decomposition

Purpose: Separating land value from building value
DRC formula approach (£1,732.79/square metres structure + £2,195.68/square metres land) produces median land share of 40.1% and mean 59.6%, but with systematic bias:

- Gwynedd: -£87k underprediction (formula mis-calibrates)
- Rural Wales: -£36k overprediction (fixed parameters don't adjust to local markets)

Recommendation: Use DRC for illustrative land share estimates only, not for parcel level valuations. Requires regional parameter calibration to avoid systematic bias.

Coverage: Properties with floor area data (38% currently; target 60-70% with improved EPC coverage)

Supplementary Method: Professional Valuation (Human)

Purpose: Unique properties and low-confidence contexts

- Apply to regeneration zones (Cardiff Bay: R^2 = 0.002, MAE = £247k)
- Listed buildings, heritage properties, unique characteristics
- Properties flagged as high-uncertainty by model (confidence interval > ±30%)
- Coverage: 5-10% of Welsh properties (expensive but necessary for accuracy)

Do not use K-Nearest Neighbours (R^2 = -1.997) or LLM Ensemble (£15k-£75k per million properties)

The study demonstrates KNN catastrophic failure in geographic holdout scenarios and LLM error correlation of ρ = 0.998 with DRC (provides no additional information despite 100-1000× computational cost).

Combined Workflow Example

Scenario 1: Suburban Semi-Detached Property (High Confidence)

1. CatBoost predicts total property value: £185,000 (confidence interval: ±12%, £163k-£207k)
2. LSOA performance check: R^2 = 0.42, MAE = £45k (HIGH CONFIDENCE context)
3. Floor area available (92square metres), apply DRC land share estimate: 42% land, 58% structure
4. Estimated land value: £77,700; structure value: £107,300
5. Recommendation: Accept automated valuation for policy purposes.

Scenario 2: Cardiff Bay Waterfront Flat (No Confidence)

1. CatBoost predicts total property value: £185,000 (confidence interval: ±85%, £28k-£342k)
2. LSOA performance check: R^2 = 0.002, MAE = £247k (NO CONFIDENCE context)
3. Systematic bias: +£223k mean underprediction for Cardiff Bay properties within 0.6km of city centre
4. Recommendation: Flag for human-only valuation; automated prediction unreliable.

Scenario 3: Rural Powys Detached (Low Confidence)

1. CatBoost predicts total property value: £215,000 (confidence interval: ±22%, £168k-£262k)
2. LSOA performance check: R^2 = 0.08, MAE = £75k (LOW CONFIDENCE context)
3. Missing floor area (no EPC record); property type and plot characteristics highly variable
4. Recommendation: Automated screening value provided, but mandatory human review before use.

This tiered approach acknowledges that uniform accuracy is unattainable across heterogeneous Welsh property markets.

6.3. Explaining to the Public

Level 1: General Public

We use a computer system trained on Welsh land sales from 1995-2024 to estimate what your land would sell for today. It compares your land with similar lots across Wales and adjusts for property type, age, and location.

Important: This system works better in some areas than others. In suburban areas with many similar properties, typical accuracy is ±£35,000 to £50,000. In unique areas like Cardiff Bay or rural locations, accuracy may be ±£200,000 or more. We will tell you which category your property falls into.

Level 2: Homeowners

The valuation uses multiple features including property type (detached/semi-detached/terraced/flat), floor area, age (new/old), postcode district, and transaction timing. It compares your land to overall Welsh market.

Your property is in [LSOA name]. Properties in this area show:

- Typical accuracy: ±£[MAE value]
- Model performance (R^2): [0.00 to 0.52 depending on LSOA]
- Confidence category: [HIGH/MEDIUM/LOW/NONE]

Your valuation: £[value] with confidence range of ±[15% to 85% depending on LSOA]
If your land is in a low or no confidence category, we recommend professional human valuation for important decisions.

Level 3: Technical Stakeholders

CatBoost gradient boosting ensemble trained on 1 million + transactions using geographic holdout validation (9 LSOAs excluded from training). Achieves $R^2 = 0.282$ overall, but with severe geographic heterogeneity: R^2 ranges from 0.002 (Cardiff Bay) to 0.518 (best-performing LSOA), MAE ranges from £35,160 to £165,983.

Features: log floor area (38% coverage), new build indicator, leasehold indicator, month sine/cosine, property type, postcode district, transaction year. Missing features (plot size, parking, renovation history, condition, views) contribute to 71.8% unexplained variance.

Critical limitation: 0.4% training representation for Cardiff Bay creates insurmountable Generalisation barriers. Error correlations between all models exceed $\rho = 0.98$, indicating errors are property-driven and context-driven, not algorithm-driven.

Model updated quarterly. Temporal analysis shows 7× error surge 2015-2019 (£111k to £533k MAE), 40% pandemic contraction, demonstrating non-stationarity requiring continuous recalibration.

6.4. Barriers to Understanding

Algorithmic Opacity ("Black Box" Concerns)

The study demonstrates that even the best model (CatBoost $R^2 = 0.282$) leaves 71.8% of variance unexplained. Feature importance cannot be calculated without the saved model file, preventing answers to questions like "Why did my property receive this valuation?" or "How much does floor area matter compared to location?"

The near-perfect error correlations ($\rho > 0.98$) between all models indicate that switching algorithms provides no benefit; the fundamental problem is missing features and sparse geo-graphic representation. However, explaining "we cannot accurately value your property be-cause properties like yours constitute only 0.4% of training data" may not satisfy stakeholders expecting algorithmic precision.

Data Scepticism and Missing Record

Critical data gaps undermine public confidence:

- Floor area missing for 62.2% of properties
- Plot size, parking, renovation history: 0% coverage
- Internal condition, views, micro-accessibility: 0% coverage

Land data more broadly faces additional gaps beyond residential property attributes, including incomplete parcel-level records, planning and constraint information, access and services data, and sparse comparable land transactions.These missing features directly cause the 71.8% unexplained variance. When homeowners see valuations that ignore renovations (new kitchen, bathroom upgrades) or plot characteristics (large garden, parking spaces), scepticism is justified.

Perceived Unfairness and Geographic Inequality

The 14× MAE variation across LSOAs (£35k to £247k) creates systematic unfairness:

- Suburban residents receive accurate valuations ($R^2 = 0.40$-0.50, MAE = £35k-£50k)
- Cardiff Bay residents receive catastrophically inaccurate valuations ($R^2 = 0.002$, MAE = £247k)
- Rural residents receive low-confidence valuations ($R^2 = 0.05$-0.15, MAE = £75k-£100k)

If automated valuations are used for policy purposes, residents in difficult-to-value areas face higher error rates and more frequent appeals, creating procedural inequality even if the underlying model is unbiased.

Systematic Bias Pattern

The study identified actionable systematic biases:

- Gwynedd: DRC -£87k underprediction (41.4% bias contribution)
- Bridgend/Vale: KNN +£78k to +£98k overprediction (31-37% bias contribution)
- Cardiff Bay: All models +£193k to +£273k underprediction (distance-based gradient: +£300k within 0.6km of city centre)

These biases affect different communities differently, raising equity concerns. Gwynedd properties systematically undertaxed under DRC approach; Cardiff Bay properties systematically overtaxed.

Bilingual Requirements

All public-facing materials, dashboards, appeals processes, and educational content must be available in Welsh and English. The study's pilot recommendation includes Powys and Gwynedd specifically to test bilingual communication in high Welsh-speaking areas.

Digital Divide

Interactive dashboards and online appeals portals assume internet access and digital literacy. Rural areas (where model confidence is lowest) often have poorest digital infrastructure, creating compounded disadvantage: low valuation accuracy + limited digital access to appeals.

Trust in Government Systems

The 43.6% negative land value rate (economic impossibility) demonstrates that even sophisticated models can produce nonsensical outputs when confronting unfamiliar geographic contexts. This undermines public trust in "computer says" valuations, particularly when errors are not transparently communicated.

Mitigation Strategies

Transparency Portal Features

Display for each property:

Valuation with explicit confidence category

- "Your estimated value: £185,000"
- "Confidence category: MEDIUM (typical accuracy ±£60,000)"
- "Your LSOA shows R^2 = 0.32, MAE = £65,000"

5-10 comparable properties with actual sale prices, distances, transaction dates

- Must be from same LSOA or within 2km radius
- Show which features match (property type, size range) and which differ

Geographic context explanation

- "Properties in your area [LSOA name] show moderate model performance"
- "72% of valuations in your area are within ±20% of actual sale prices"
- Map showing property location and comparables

Missing data disclosure

- "Your valuation does not include plot size (data not available)"
- "Floor area estimated from property type (actual floor area not recorded)"

Temporal trend

- 5-year value trend for your LSOA (not individual property, due to privacy)

Explicit warnings for low-confidence contexts

- "Your property is in a difficult-to-value area (Cardiff Bay/rural/unique characteristics)"
- "Automated valuation may be inaccurate (typical errors exceed £200,000)"
- "Professional human valuation strongly recommended"

Educational Materials

Develop bilingual resources

- 3-5 minute video explaining how the system works, narrated in Welsh and English
- Infographics showing:
    - Where the system works well (suburban estates) vs poorly (Cardiff Bay, rural)
    - What features are included (property type, postcode) vs missing (plot size, condition)
    - How to interpret confidence intervals and LSOA-specific performance metrics
- Frequently Asked Questions addressing:
    - "Why is my valuation different from Zoopla/Rightmove?" (Different models, different training data, commercial sites include asking prices not just sales)
    - "Why doesn't the valuation include my renovations?" (Renovation history not available in Land Registry or EPC data)
    - "Why is my neighbour's property valued differently?" (Different property type, size, or specific features; or model uncertainty)
- Case studies demonstrating success cases (suburban semi-detached) and limitation cases (Cardiff Bay waterfront).

Community Workshops

Conduct public engagement in pilot areas:

- Local briefings at community centres (bilingual, multiple time slots)
- Live system demonstrations showing transparency portal
- Structured feedback collection:
    - Do you understand how the valuation was calculated?
    - Do you trust the system for taxation purposes?
    - What additional information would improve understanding?
- Councillor and Member of Senedd briefings emphasising geographic performance heterogeneity

Independent Validation

Commission annual RICS or academic audits:

- Sample 1,000 randomly selected properties stratified by LSOA type (not purely random)
- Professional valuations conducted blind (valuers don't see model predictions)
- Compare model vs professional valuations
- Publish results by LSOA type (suburban, valley, rural, regeneration)
- Benchmark against commercial property websites (Zoopla, Rightmove)

Expected audit findings based on study:

- Suburban estates: Model within ±10% of professional valuations for 70-80% of properties
- Valley towns: Model within ±15% of professional valuations for 60-70% of properties
- Rural areas: Model within ±25% of professional valuations for 40-50% of properties
- Regeneration zones: Model agreement with professional valuations <30% (systematic failure)

Publish these results transparently, acknowledging limitations rather than claiming uniform accuracy.

6.5. Basis for Land Valuation in Wales

Critical Warning from Study Findings

The 43.6% negative land value rate demonstrates that residual land valuation (total property value minus structure value = land value) systematically fails when models don't generalise to test LSOAs. This creates a fundamental barrier to parcel level policy interventions using current methods.

Two-Step Process (DRC Approach)

1. Property Valuation: CatBoost predicts total property value
2. Land Decomposition: DRC formula applies construction cost (£1,732.79/square metres) and land value (£2,195.68/square metres) parameters

Example: Cardiff, Mid-Range Property

- Total property value: £250,000 (from CatBoost)
- DRC land share: 40.1% (median across 9 test LSOAs)
- Implied land value: £100,250
- Implied structure value: £149,750

Critical limitations

Systematic bias

DRC shows -£87k underprediction in Gwynedd, -£36k overprediction in rural Wales. Fixed parameters (£1,732.79/square metres structure, £2,195.68/square metres land) mis calibrate across diverse markets.

Geographic Generalisation failure

CatBoost total property values show $R^2 = 0.002$ in Cardiff Bay (catastrophic failure). Land value decomposition inherits these errors and amplifies them through the residual calculation.

Floor area coverage

Only ~38% of transactions have a direct EPC match (UPRN/address). A further ~36% receive a postcode-median EPC proxy (EPC-based, higher uncertainty), bringing total floor-

area availability to ~74%. The remaining ~26% have no usable floor-area value even after the proxy step and therefore use hierarchical non-EPC imputation.

Implausible uniformity

DRC produces median land share of 40.1% across all LSOAs. Economic theory suggests land shares should vary from 10% (rural) to 80% (central urban). Fixed parameters create artificial uniformity masking real geographic variation.

Alternative: Location-Based Multipliers (Simplified Approach)

Instead of property-level land value decomposition:

- Use CatBoost total property value (more reliable than land decomposition)
- Apply location multipliers to approximate land premium:
    - Cardiff city centre: 1.5× base rate (high land premium)
    - Suburban areas: 1.0× base rate (moderate land premium)
    - Valley towns: 0.8× base rate (lower land premium)
    - Rural areas: 0.6× base rate (lowest land premium)

This approach:

- Avoids 43.6% negative land value problem
- Works with 100% property coverage (no floor area required)
- Captures broad geographic patterns (Cardiff premium, rural discount)
- Less theoretically pure (not true land value tax)
- Requires political consensus on multiplier values

Implementation Challenges

Legislation

There would like be need for extensive primary legislation depending on the nature of reforms desired.

Data Gaps

- Floor area missing for 62.2% of properties (requires EPC coverage improvement)
- Plot size missing for 100% of properties (requires Land Registry plot boundary digitization or satellite imagery analysis)
- Renovation history missing for 100% of properties (no feasible data source)

Distributional Effects

Study demonstrates systematic bias patterns.

- Gwynedd: -£87k underprediction (undertaxed under DRC)
- Cardiff Bay: +£273k underprediction (overtaxed if based on model predictions)

Transition planning must address these biases through regional parameter calibration or explicit adjustment factors. More general land reform implications extend beyond model performance and would require detailed consideration if changes affect valuation bases,

taxation rules, ownership disclosure, or statutory land-use controls. In practice, implementation would depend on establishing a clear legal basis for data sharing, appeals, and enforcement across agencies. The scale of change is likely to be modest for technical valuation improvements, but materially greater for broader fiscal or governance reform. Persistent land data gaps (especially plot boundaries, tenure detail, and land-use constraints) would remain a critical constraint on operational delivery.

Non-Residential Extension

This study focuses on residential properties because they provide the most consistent and comparable data across the available sources. Commercial and agricultural land use different valuation bases, variables, and market assumptions, so they cannot be assessed reliably using the same methodology. Any extension to these asset classes would require separate method development and validation.

Public Acceptance

The 14× MAE variation (£35k to £247k) across LSOAs creates public trust challenges. Extensive consultation required to explain why some areas receive accurate valuations while others don't, and how this affects taxation fairness.

6.6. Potential Application

The valuation system can support other policy applications with varying confidence:

Planning Viability Assessments

Determine whether proposed developments are economically viable.
Confidence: Medium (models predict existing property values, not post-development values)
Application: Provide baseline land values for Section 106 negotiations.

Compulsory Purchase Compensation

Calculate fair compensation for compulsory land acquisition.
Confidence: Low to Medium (depends on geographic context)
Critical limitation: Study shows $R^2 \approx 0$ in unique contexts. Compulsory purchases often involve unique circumstances (major infrastructure, unusual properties). Automated valuations unreliable precisely where most needed. Recommendation: Use as reference value only; professional valuation mandatory for compensation determination.

Land Value Capture for Infrastructure Funding

Calculate "betterment" from new infrastructure (e.g., property value increase from new train station).
Confidence: Low (requires accuracy in valuations both before and after infrastructure changes)
Challenge: Temporal analysis shows 7× error surge 2015-2019. Measuring infrastructure impact requires temporal stability that current models don't achieve.

Track housing affordability trends across Wales .
Confidence: High (aggregate trends more reliable than individual property values)
Application: Generate LSOA-level median price trends, affordability ratios, geographic inequality metrics
Strength: Aggregate statistics less affected by individual property errors.

6.7. Pilot Recommendation

This section outlines a comprehensive two-year pilot to test feasibility, fairness, administrative viability, and public acceptability of land value based valuation in Wales. The pilot integrates modelling outputs, Behavioural observation, public engagement, legal preparation, and distributional analysis.

Pilot Areas: Three Local Authorities

Cardiff (CF postcodes): Urban, High Values, Rich Data

- Population: ~365,000
- Properties: ~160,000
- Pilot sample: 10,000 properties (6.25% of Cardiff total)
- Expected performance: Mixed
- Suburban Cardiff (e.g., Whitchurch, Llandaff): HIGH CONFIDENCE (R^2 = 0.40-0.50, MAE = £40k-£55k)
- Cardiff Bay: NO CONFIDENCE (R^2 = 0.002, MAE = £247k)

Rationale: Cardiff provides high-density, high-value urban test environment with extreme intra-city variation. Includes best-case contexts (suburban) and worst-case contexts (Cardiff Bay) within single local authority, enabling direct comparison.

Powys (LD/SY postcodes): Rural, Welsh-Speaking, Low Market Activity

- Population: ~133,000
- Properties: ~65,000
- Pilot sample: 10,000 properties (15.4% of Powys total)
- Expected performance: LOW CONFIDENCE (R^2 = 0.08, MAE = £75k)

Rationale: Powys represents most rural conditions in Wales. Low transaction density, wide plot size variation, mixed agricultural contexts, dispersed communities. Tests model under weak market signals and sparse data. Also tests bilingual communication in high Welsh-speaking area (21% Welsh speakers).

Gwynedd (LL postcodes): Coastal/Tourism, High Welsh Language Use

- Population: ~124,000
- Properties: ~62,000
- Pilot sample: 10,000 properties (16.1% of Gwynedd total)
- Expected performance: LOW TO MEDIUM CONFIDENCE (R^2 = 0.15-0.30, MAE = £60k-£90k)

Rationale: Gwynedd captures coastal and tourism-driven markets. Seasonal demand, second homes, short-term holiday lets heavily influence land values. Tests model under atypical price dynamics. Gwynedd's strong Welsh language profile (64% Welsh speakers) allows exploration of cultural and linguistic fairness considerations.

Pilot Parameters

Duration and Sample Size:

- 2-year pilot period (enables two complete annual valuation cycles)
- 10,000 properties per local authority (30,000 total)
- Stratified sampling within each authority:
    - 40% suburban/easy-to-value contexts
    - 30% valley/medium-difficulty contexts
    - 20% rural/difficult contexts
    - 10% unique/very-difficult contexts (Cardiff Bay, listed buildings, etc.)

Stratification ensures pilot captures full range of valuation difficulty, not just easy cases.

Parallel System: Shadow Tax

Pilot operates shadow land value tax alongside existing council tax:

- Land value tax liabilities simulated and shown to households.
- Council tax continues as normal (households pay existing bills, not pilot liabilities).
- Households receive information showing both values side-by-side.

Revenue neutrality prevents financial disruption. Public responses reflect understanding and acceptance, not fear of immediate cost changes.

Two-Year Pilot Timeline

Phase 0: Design and Set-Up (Months 0-6)

<u>Months 0-3: Scoping and Legal Preparation</u>

Activities:

- Formal pilot area selection and local authority agreements
- Define success criteria:
    - Technical: What R^2 and MAE targets by LSOA type?
    - Operational: What appeals processing time and cost acceptable ?
    - Political: What public acceptance threshold required for rollout decision?
- Data protection impact assessment (GDPR compliance for household-level data sharing)
- Prepare Welsh Government Order or regulations permitting:
    - Shadow tax calculation and disclosure to households
    - Data sharing between Welsh Government, local authorities, Land Registry, Valuation Office Agency
    - Public engagement activities and surveys

Deliverables:

- Pilot protocol document

- Legal instruments enabling pilot operations
- Success criteria matrix with quantitative thresholds

Months 3-6: Technical Build and Staff Onboarding
Activities:

- Deploy CatBoost and DRC models to pilot local authority IT environments
- Establish secure data pipelines:
    - Land Registry Price Paid Data (quarterly updates)
    - EPC data (annual updates)
    - ONS Postcode Directory (annual updates)
- Configure pilot-specific dashboards showing:
    - Individual property valuations with confidence metrics
    - LSOA-level performance (R^2, MAE, bias)
    - Comparative shadow tax bills (council tax vs land value tax)
- Train local authority analysts and valuation staff:
    - How to interpret model outputs
    - How to handle public queries
    - When to escalate to professional valuation

Deliverables:

- Operational modelling system deployed in 3 local authorities
- User guides and training materials (bilingual)
- Data governance protocols

Phase 1: Shadow Operation, Year 1 (Months 6-18)

Months 6-9: Initial Valuation Run and Internal Testing
Activities:

- Generate baseline valuations for all 30,000 pilot properties
- Run land value decomposition (DRC approach)
- Calculate shadow land value tax bills (at revenue-neutral rate)
- Internal quality assurance:
    - Check for negative land values (expect 43.6% rate based on study; flag for investigation)
    - Identify high-uncertainty properties (confidence interval > ±30%)
    - Professional valuer reviews 100 high-uncertainty cases per local authority (300 total)
- Prepare individual household information packs (bilingual)

Deliverables:

- 30,000 individual property valuations with confidence metrics
- Internal quality assurance report identifying systematic issues
- Household information packs ready for distribution

Months 9-12: Controlled Public Disclosure
Activities:

- Issue bilingual informational letters to 30,000 pilot households containing:

    - Estimated property value with confidence category (HIGH/MEDIUM/LOW/NONE)
    - Shadow land value tax bill
    - Current council tax bill for comparison
    - Explanation of pilot (no actual payment required)
    - Instructions for providing feedback or raising concerns
- Launch engagement activities:
    - 6-10 public webinars per local authority (18-30 total, bilingual)
    - Drop-in sessions at libraries, community centres (3-5 per authority)
    - FAQ documentation on dedicated pilot website
    - Councillor and Member of Senedd briefings
- Establish feedback channels:
    - Dedicated phone line (bilingual)
    - Online feedback form
    - Email address
- Monitor queries and concerns:
    - Track volume and types of queries
    - Identify common misunderstandings
    - Flag systematic issues (e.g., Cardiff Bay residents all complaining about inaccuracy)

Deliverables:

- 30,000 household information packs distributed
- Public engagement events completed
- Interim feedback report (Month 12)

<u>Months 12-18: Behavioural and Operational Observation</u>

Activities:

- Continuous monitoring of public response:
    - Query volumes and resolution times
    - Sentiment analysis of feedback
    - Media coverage and social media discussion
- Administrative workload assessment:
    - Local authority staff time spent on pilot queries
    - Professional valuation requests (properties where households dispute model outputs)
    - Appeals processing time and cost
- Preliminary household survey (Month 15):
    - Do you understand your shadow valuation?
    - Do you trust it for taxation purposes?
    - How does it compare to your expectations?
    - What concerns do you have?
    - Sample: 1,000 households per authority (3,000 total), stratified by confidence category

Deliverables:

- Interim evaluation report (Month 18) covering:
    - Revenue stability (does aggregate shadow tax match council tax?)
    - Distributional shifts (winners and losers by property type, location, value band)
    - Public opinion and acceptance levels
    - Administrative effort and resource requirements
    - Technical performance (actual R^2 and MAE vs predictions)

Phase 2: Shadow Operation, Year 2 (Months 18-24)

Months 18-22: Refinement and Second Valuation Cycle

Activities:

- Retrain models with updated transaction data (incorporating last 12 months of sales)
- Re-run valuations for all 30,000 pilot properties
- Compare Year 1 vs Year 2 valuations:
    - Temporal stability: How much do valuations change year-over-year?
    - Systematic drift: Are errors getting larger or smaller over time?
    - Confidence category stability: Do properties remain in same confidence category?
- Deep distributional analysis:
    - Rural vs urban tax burden shifts
    - Welsh-speaking vs non-Welsh-speaking communities
    - Property type impacts (detached vs flats)
    - Tourism market impacts (Gwynedd second homes)
    - Comparison with manual professional valuations (sample 300 properties)

Professional valuation sample:

- 100 properties per local authority (300 total)
- Stratified by confidence category: 40% HIGH, 30% MEDIUM, 20% LOW, 10% NONE
- RICS-qualified valuers conduct blind valuations (don't see model predictions)
- Compare professional valuations vs model outputs
- Calculate agreement rates, bias, MAE by confidence category

Expected findings based on study:

- HIGH confidence: Model within ±15% of professional valuation for 70-80% of properties
- MEDIUM confidence: Model within ±20% of professional valuation for 60-70% of properties
- LOW confidence: Model within ±30% of professional valuation for 40-50% of properties
- NONE confidence: Model agreement with professional valuations <30%

Deliverables:

- Year 2 valuations for all 30,000 properties
- Temporal stability analysis
- Professional valuation comparison report

Months 22-24: Final Evaluation and Policy Decision

Activities:

- Comprehensive evaluation report covering:
    - Technical performance: Actual R^2, MAE, bias by LSOA type; comparison with study predictions
    - Distributional fairness: Winners/losers analysis; impacts on protected characteristics
    - Administrative feasibility: Staff time, appeals volume and cost, system reliability
    - Public acceptance: Survey results, stakeholder feedback, media analysis
    - Economic impacts: Behavioural responses (did households make property improvements? sell properties?)
    - Comparison with alternatives: How does land value tax compare to reformed council tax or other options?
- Stakeholder synthesis workshops:
    - Local authority finance officers
    - Valuation professionals (RICS Wales)
    - Community representatives
    - Academic experts
- Present findings to:
    - Welsh Government Ministers
    - Senedd Finance Committee
    - Local Government and Housing Committee
    - Public consultation (if proceeding to national rollout decision)

Deliverables:

- Final evaluation report (100-150 pages)
- Executive summary for Ministers (10-15 pages)
- Technical appendices (data, methodology, detailed results)
- Policy recommendations:
    - Proceed to national rollout (with conditions)
    - Extend pilot (2-3 more years with expanded scope)
    - Abandon land value tax (return to council tax reform options)