PropertyCompanion: Ensemble Machine Learning for Residential Property Valuation with Uncertainty Quantification

Application to North-West London and Surrounding Boroughs
(Properties priced between £150k–£3M)

Vishruth Dhamodharan
vishy.dhamo@gmail.com | github.com/vishruth-d/property-companion

January 2026

Abstract

Traditional automated valuation models provide single point estimates without quantifying prediction uncertainty, a significant limitation given the inherent heteroscedasticity in property prices. This paper presents PropertyCompanion, a production-grade machine learning system addressing these challenges through: (1) systematic comparison of five price-segmentation strategies across multiple gradient boosting frameworks, (2) distribution-free uncertainty quantification via conformal prediction, and (3) comparable transaction analysis. The system combines HM Land Registry Price Paid data with Energy Performance Certificates, requiring novel approaches to infer bedroom and bathroom counts. This report details the data pipeline, feature engineering, Bayesian hyperparameter optimisation, and statistical validation including Friedman-Nemenyi tests and two-way ANOVA.

Contents

1 Introduction

The UK residential property market transacts over £300 billion annually, yet valuation remains largely manual, expensive, and opaque. Traditional automated valuation models typically provide single point estimates with no indication of uncertainty, which is problematic given the inherent heteroscedasticity in property prices, particularly at distribution tails where variance is highest.

My family, on the side, invests into real estate properties. Through this experience I observed how property valuations are often based on intuition and informal market expectations rather than rigorous analysis. I wanted to automate their due diligence process through data-driven systems, replacing "in the air valuations" with objective, quantitative assessments that could identify genuine investment opportunities.

This project makes the following contributions: (1) **Comprehensive Model Selection** through systematic comparison of 5 segmentation strategies across 13 categorical configurations, 9 numerical configurations, and 8 model types; (2) **Calibrated Uncertainty** via distribution-free prediction intervals using conformal prediction with segment-specific thresholds; (3) **Statistical Validation** through rigorous hypothesis testing including Friedman-Nemenyi tests and two-way ANOVA; and (4) a **Production API** built with FastAPI supporting real-time and batch inference with comparable transaction analysis.

2 Data Pipeline

2.1 Data Sources

Two primary sources were combined: **HM Land Registry Price Paid** with 500,000+ transactions containing price, date, property type, tenure, and full address (2018-2025), and **Energy Performance Certificates** with 400,000+ records containing floor area, habitable rooms, energy rating, and built form (2018-2025). Records were merged on address fields, achieving a **100% match rate** for the final dataset.

2.2 Price Normalisation and Filtering

Historical transaction prices were normalised to present-day values using the **UK House Price Index** published by the Office for National Statistics. Transactions were filtered to the £150,000 to £3,000,000 range (approximately 2.3rd to 97.7th percentiles), retaining 97.7% of matched records.

2.3 Data Versioning

Three dataset versions were created: v1 (2-year, 56,181 transactions), v2 (extended, 89,432 transactions), and v3 (5-year comprehensive, 148,678 transactions) used as the production model.

3 Modelling Assumption: Estimating Bedrooms and Bathrooms

EPC data provides only *habitable rooms* (H), not bedrooms or bathrooms directly. A rule-based estimation system was developed based on EPC definitions, where habitable rooms include living rooms, sitting rooms, dining rooms, bedrooms, and studies, but *exclude* kitchens and bathrooms.

3.1 Key Assumption

Let H = habitable rooms, L = living rooms (estimated), B = bedrooms, A = floor area (m^2). We assume:

$$H \approx L + B \quad \Rightarrow \quad B = H - L \tag{1}$$

3.2 Living Room Estimation

Living rooms are estimated using property-type-specific rules with floor area thresholds for disambiguation.

Flats and Maisonettes:

$$L = \begin{cases} 0 & \text{if } H = 1 \\ 1 & \text{if } 2 \leq H \leq 5 \\ 2 & \text{if } H \geq 6 \end{cases} \tag{2}$$

Houses (Terraced, Semi-Detached):

$$L = \begin{cases} 0 & \text{if } H = 1 \\ 1 & \text{if } 2 \leq H \leq 4 \\ 2 & \text{if } H = 5 \wedge A < 100 \\ 1 & \text{if } H = 5 \wedge A \geq 100 \\ 2 & \text{if } 6 \leq H \leq 7 \\ 3 & \text{if } H \geq 8 \wedge A < 150 \\ 2 & \text{if } H \geq 8 \wedge A \geq 150 \end{cases} \tag{3}$$

Rationale for $H = 5$: Smaller houses (below $100m^2$) with 5 habitable rooms typically have 2 reception areas and 3 bedrooms. Larger houses ($100m^2$+) are more likely to be genuine 4-bedroom properties with 1 reception room.

3.3 Bathroom Estimation

$$\text{Bathrooms} = \begin{cases} 1 & \text{if } B \leq 2 \\ \lceil B/2 \rceil & \text{if } B \geq 3 \end{cases} \tag{4}$$

3.4 Validation

Sanity checks against floor area bounds were applied. On fully cleaned data, **93.4%** passed sanity checks. Manual verification ($n = 82$): 25.6% exact match, 65.9% off by 1 bedroom, 8.5% off by 2+. Overall, **94% of estimates were realistic**.

4 Feature and Model Engineering

4.1 Feature Summary

Type	Features
Categorical	Postcode sector, local authority, property type, built form, tenure, old/new, EPC rating (7 features)
Numerical	Total floor area, bedrooms, bathrooms, plus engineered ratios, logs, and interactions
Target	log(Price) to normalise right-skewed distribution

4.2 Configuration Space

Categorical configurations (13 variants): Full set of all 7 features, ablation tests systematically dropping one feature, and strategic subsets including location-only and property-characteristics-only.

Numerical configurations (9 variants): Base features (floor area, bedrooms, bathrooms), log transforms, ratios (area per bedroom, area per bathroom), polynomial terms, and interaction features (area $\times$ bedrooms).

Model types (8 algorithms): CatBoost, LightGBM, XGBoost (gradient boosting with native categorical support), Random Forest, Gradient Boosting Regressor, and linear baselines (Ridge, Lasso, ElasticNet).

This configuration yielded **272 feature-model combinations per segment**. Across all 5 segmentation options and 3 data versions, over 8,160 configurations were evaluated.

5 Machine Learning Methodology

5.1 Segmentation Rationale

Property prices exhibit **heteroscedasticity**, meaning variance increases substantially at distribution tails. Residual diagnostics confirmed this: the Breusch-Pagan test yielded $p = 4.19 \times 10^{-11}$, strongly rejecting homoscedasticity. A single model struggles to simultaneously capture dynamics across £150k flats and £2m+ estates. Price-based segmentation has been shown to improve prediction accuracy in real estate applications (4), with segment-specific models capturing local patterns more effectively than global approaches.

5.2 Five Modelling Options

Option	Strategy	Segments	Rationale
1	Vanilla	1	Baseline benchmark
2	Stratified sampling	1	Balanced price representation
3	3-Segment (10/80/10)	3	Tail-focused separation
4	3-Segment (20/60/20)	3	Expanded tail coverage
5	5-Segment	5	Fine-grained specialisation

Option 5 segment boundaries: 0-15th, 15-50th, 50-80th, 80-95th, 95-100th percentiles.

5.3 Bayesian Hyperparameter Optimisation

After initial model selection, best-performing configurations underwent Bayesian hyperparameter optimisation using **Optuna** (1).

5.3.1 Limitations of Traditional Approaches

Grid Search exhaustively evaluates all parameter combinations. For k hyperparameters with n values each, the computational complexity is:

$$\text{Grid Search Complexity} = O(n^k) \tag{5}$$

This becomes intractable as the search space grows. For example, with 8 hyperparameters and 5 values each, grid search requires $5^8 = 390,625$ evaluations.

Random Search, as demonstrated by **(author?)** (3), samples uniformly from the parameter space. While more efficient than grid search, it does not learn from previous evaluations, treating all regions equally regardless of observed performance.

5.3.2 Tree-structured Parzen Estimator

Optuna implements the Tree-structured Parzen Estimator, introduced by **(author?)** (2). Given observations $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where x_i are hyperparameter configurations and y_i are corresponding validation MAPEs, TPE models two distributions based on a quantile threshold y^* (typically the best 25% of observations):

$$p(x|y) = \begin{cases} \ell(x) & \text{if } y < y^* \quad (\text{“good” configurations}) \\ g(x) & \text{if } y \geq y^* \quad (\text{“poor” configurations}) \end{cases} \tag{6}$$

where $\ell(x)$ models the density of hyperparameters producing good results and $g(x)$ models those producing poor results.

5.3.3 Acquisition Function

The next configuration to evaluate is selected by maximising the Expected Improvement. In TPE, this is proportional to:

$$\text{EI}(x) \propto \frac{\ell(x)}{g(x)} \tag{7}$$

This ratio is high when x is likely under the “good” distribution and unlikely under the “poor” distribution, efficiently directing search toward promising regions (7).

For each segment, 100 optimisation trials were conducted with search ranges including learning rate [0.01, 0.3] (log-uniform), estimators [300, 1500], max depth [4, 12], and regularisation parameters.

5.4 The R-squared versus MAPE Tradeoff

A critical methodological decision was optimising for **minimal MAPE rather than maximum $\mathbf{R^2}$**. While R^2 measures variance explained globally, MAPE directly reflects the user-facing metric: percentage error on individual predictions.

In segmented models, R^2 can be misleading. Each segment has reduced variance by construction, leading to lower within-segment R^2 even when predictions improve. For example, the vanilla model

achieves $R^2 = 0.832$ but MAPE = 14.87%, while Option 5 achieves lower $R^2 = 0.390$ but substantially better MAPE = 9.95%.

This apparent paradox arises because segmentation removes between-segment variance from each model's purview. The production system prioritises MAPE as it directly translates to valuation accuracy.

6 Uncertainty Quantification

6.1 Motivation

Point predictions alone are insufficient for practical property valuation. Users need plausible ranges for negotiation, and model confidence should inform decision-making. Residual diagnostics confirmed persistent heteroscedasticity (Breusch-Pagan $p = 4.19 \times 10^{-11}$, Shapiro-Wilk $p = 3.05 \times 10^{-23}$), motivating distribution-free methods.

6.2 Quantile Regression

After Optuna optimisation of point prediction models, separate **quantile regression models** were trained for the 10th and 90th percentiles. Unlike standard regression minimising squared error, quantile regression minimises the pinball loss:

$$L_\tau(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}) & \text{if } y \geq \hat{y} \\ (1 - \tau)(\hat{y} - y) & \text{if } y < \hat{y} \end{cases} \tag{8}$$

where $\tau \in \{0.1, 0.9\}$ for the lower and upper bounds respectively.

6.3 Conformal Calibration

Raw quantile predictions were calibrated using **split conformal prediction** (5; 6). Building on foundational work by **(author?)** (8), this approach provides finite-sample coverage guarantees without distributional assumptions.

For each segment s, conformal thresholds were computed on a held-out calibration set:

$$\hat{q}^{(s)}_{1-\alpha} = \text{Quantile}_{1-\alpha}\left(\{R_i : x_i \in \text{Segment}_s\}\right) \tag{9}$$

where $R_i = |y_i - \hat{y}_i|/\hat{y}_i$ are relative residuals.

6.4 Calibration Results

Segment	Nominal	Empirical	Miss Rate	Mean Width (£)
Seg1 (0-15%)	80%	81.0%	19.0%	164,752
Seg2 (15-50%)	80%	84.0%	16.0%	155,989
Seg3 (50-80%)	80%	82.0%	18.0%	227,940
Seg4 (80-95%)	80%	74.0%	26.0%	429,559
Seg5 (95-100%)	80%	61.0%	39.0%	779,431
Overall	80%	78.8%	21.2%	240,793

7 Statistical Validation

7.1 K-Fold Cross-Validation Stability

Five-fold cross-validation verified reproducibility:

Segment	Mean MAPE	Std	CI Lower	CI Upper	CV%
Seg1 (0-15%)	13.10%	0.15%	12.97%	13.23%	1.13%
Seg2 (15-50%)	7.86%	0.04%	7.82%	7.89%	0.57%
Seg3 (50-80%)	7.74%	0.02%	7.72%	7.76%	0.26%
Seg4 (80-95%)	9.43%	0.09%	9.35%	9.51%	0.97%
Seg5 (95-100%)	11.84%	0.25%	11.62%	12.06%	2.13%

Average coefficient of variation of **1.01%** indicates highly stable error estimates.

7.2 Friedman-Nemenyi Test

To determine whether Option 5 significantly outperforms alternatives, a **Friedman test** was conducted on $n = 500$ samples: Friedman statistic = 22.67, $p = 1.19 \times 10^{-5}$, Kendall's $W = 0.0227$ (small but significant effect).

Post-hoc Nemenyi comparisons (critical difference = 0.148) revealed Option 5 significantly outperforms Options 3 and 4 ($p < 0.05$), while Options 3 and 4 do not differ significantly from each other.

7.3 Two-Way ANOVA

To verify that segmentation benefits generalise across price bands:

Effect	F	p-value	η^2	Interpretation
Option	9.80	5.91×10^{-5}	0.012	Small, significant
Price Band	43.24	5.56×10^{-19}	0.054	Small, significant
Option × Price Band	0.93	0.447	0.002	Negligible, not significant

The non-significant interaction ($p = 0.447$) confirms that segmentation benefits generalise across all price bands.

8 Results

8.1 Model Performance Comparison

Option	V1 MAPE	V2 MAPE	V3 MAPE	Best
Option 1 (Vanilla)	14.98%	15.79%	14.87%	V3
Option 2 (Stratified)	15.09%	16.16%	14.75%	V3
Option 3 (3-Seg 10/80/10)	12.05%	11.51%	12.26%	V2
Option 4 (3-Seg 20/60/20)	12.12%	12.28%	11.87%	V3
Option 5 (5-Segment)	10.01%	10.31%	**9.95%**	**V3**

8.2 Segment-Level Performance

Segment	V1	V2	V3	Best Model
Seg1 (0-15%)	13.10%	12.48%	13.32%	LightGBM
Seg2 (15-50%)	8.05%	9.26%	**7.86%**	LightGBM
Seg3 (50-80%)	7.74%	8.93%	**7.66%**	LightGBM
Seg4 (80-95%)	9.61%	9.48%	**9.24%**	LightGBM
Seg5 (95-100%)	11.53%	11.41%	11.66%	CatBoost
Average	10.01%	10.31%	**9.95%**	–

8.3 Feature Importance

SHAP analysis with 50 bootstrap iterations reveals consistent feature rankings:

Feature	Avg Importance	Avg Rank	Segments in Top-5
POSTCODE_SECTOR	22.4%	1.9	5/5
TOTAL_FLOOR_AREA	29.5%	2.0	3/5
LOCAL_AUTHORITY_LABEL	16.7%	2.6	5/5
PROPERTY_TYPE	11.3%	5.1	4/5
BUILT_FORM	8.8%	5.2	4/5

8.4 Residual Diagnostics

Metric	V3 Value
Mean Residual	-0.0012 (near zero, unbiased)
Std Residual	0.117
Skewness	-0.404
Kurtosis	1.52 (leptokurtic)
Shapiro-Wilk p	3.05×10^{-23} (non-normal)
Breusch-Pagan p	4.19×10^{-11} (heteroscedastic)
Best-fit Distribution	Non-central t (AIC = -22,760)

9 Limitations

1. **Geographic coverage**: The model is trained exclusively on North-West London data. Performance in other UK regions may differ due to distinct market dynamics.
2. **Backtesting constraints**: True validation would require listings data showing asking prices that subsequently became transactions. Without access to such data, model evaluation relied on transaction-to-transaction accuracy rather than the more relevant listing-to-transaction prediction. To improve, one should collect paired listing-transaction data to measure whether the model successfully identifies undervalued properties before sale.
3. **Tail segment accuracy**: Properties at distribution extremes (Seg1, Seg5) show higher MAPE (11-13%) compared to middle segments (7-8%), reflecting inherent valuation difficulty.
4. **Missing features**: Property condition, garden size, local amenities, and transport proximity are not captured in available data sources.
5. **Heteroscedasticity**: Despite segment-specific models, Breusch-Pagan tests indicate residual heteroscedasticity ($p < 0.001$). Quantile regression partially addresses this.

10 Conclusions

This project developed PropertyCompanion, a production-grade machine learning system for UK residential property valuation. Key findings:

1. **Segmentation significantly improves accuracy**: Option 5 achieves 9.95% MAPE compared to 14.87% for vanilla, a 33% relative improvement (Friedman $p = 1.19 \times 10^{-5}$)
2. **Benefits generalise across price bands**: Two-way ANOVA confirms no overfitting to specific segments (interaction $p = 0.447$)
3. **Conformal prediction provides reliable uncertainty**: 78.8% empirical coverage with 0.8% calibration error
4. **LightGBM dominates**: Selected as best model for 4 of 5 segments
5. **Location is paramount**: Postcode sector and local authority consistently rank as top features

11 Future Work

1. **Listings-to-transaction validation**: Collect paired data of property listings and subsequent sale prices
2. **LLM integration**: Analyse listing descriptions and photographs for qualitative red flags
3. **Sub-8% MAPE target**: Explore custom loss functions and conformalized quantile regression
4. **Geographic expansion**: Extend coverage to Greater London and other UK regions

References

[1] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of KDD*.

[2] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems*, 24.

[3] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281-305.

[4] Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence. *Journal of Real Estate Research*, 32(2), 139-160.

[5] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.

[6] Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized Quantile Regression. *NeurIPS*, 32.

[7] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *NeurIPS*, 25.

[8] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.