Used Car Valuation in the UK: A Predictive Approach

**Date:** 2025-08-09 06:51

This report summarizes data preparation, model selection, and evaluation results for predicting used car prices in the UK using a supervised regression pipeline. The workflow includes preprocessing (imputation + one-hot encoding), 5-fold cross-validated model selection with randomized hyperparameter search, and final hold-out testing.

# 1. Dataset

Source file: Used Cars Prices in UK export 2025-08-09 05-54-05.csv

Rows: 2,859 | Columns: 14

Target variable: price

Example features: title, mileage\_miles, registration\_year, previous\_owners, fuel\_type, body\_type, Gearbox, Doors, Seats, emission\_class ...

# 2. Methodology

Data were split into Train/Validation/Test with an overall ratio of 60/20/20. We trained multiple regressors with log-transformed targets for stability: LinearRegression, Ridge, Lasso, ElasticNet, RandomForest, HistGradientBoosting, SVR (RBF). Hyperparameters (where applicable) were tuned using 5-fold cross-validation with RandomizedSearchCV. The best model was chosen by validation RMSE, retrained on Train+Validation, and evaluated on the Test set.

# 3. Model Comparison (Validation)

Leaderboard sorted by Validation RMSE (lower is better).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| model | cv\_best\_rmse | val\_rmse | val\_mae | val\_r2 |
| HistGradientBoosting | 1470.73 | 1577.65 | 992.27 | 0.862 |
| FINAL\_TEST\_HistGradientBoosting |  | 1620.13 | 979.08 | 0.850 |
| RandomForest | 1606.32 | 1676.59 | 1063.38 | 0.844 |
| ElasticNet | 1513.34 | 1749.34 | 1117.59 | 0.830 |
| Lasso | 1512.19 | 1755.52 | 1125.25 | 0.829 |
| Ridge | 1524.79 | 1762.11 | 1119.21 | 0.828 |
| LinearRegression |  | 2009.12 | 1243.07 | 0.776 |
| SVR | 3896.42 | 4104.50 | 2728.98 | 0.064 |

# 4. Final Test Performance

Best model: HistGradientBoosting

Test RMSE: £1,620.13

Test MAE: £979.08

Test R²: 0.850

# 5. Diagnostic Figures

The following figures are generated from the final model on the test set.

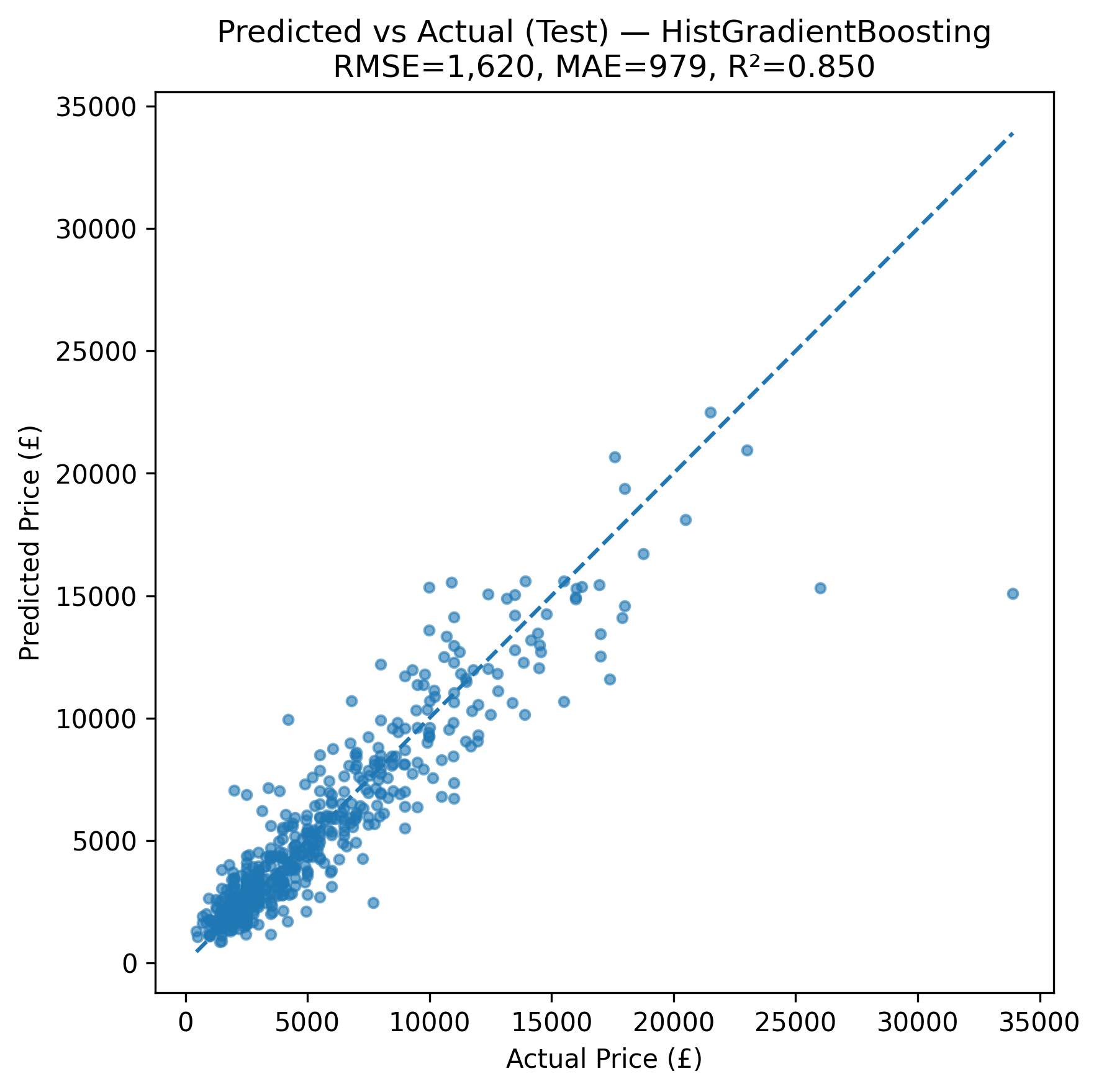


Figure 1. Predicted vs. Actual prices with y=x reference.

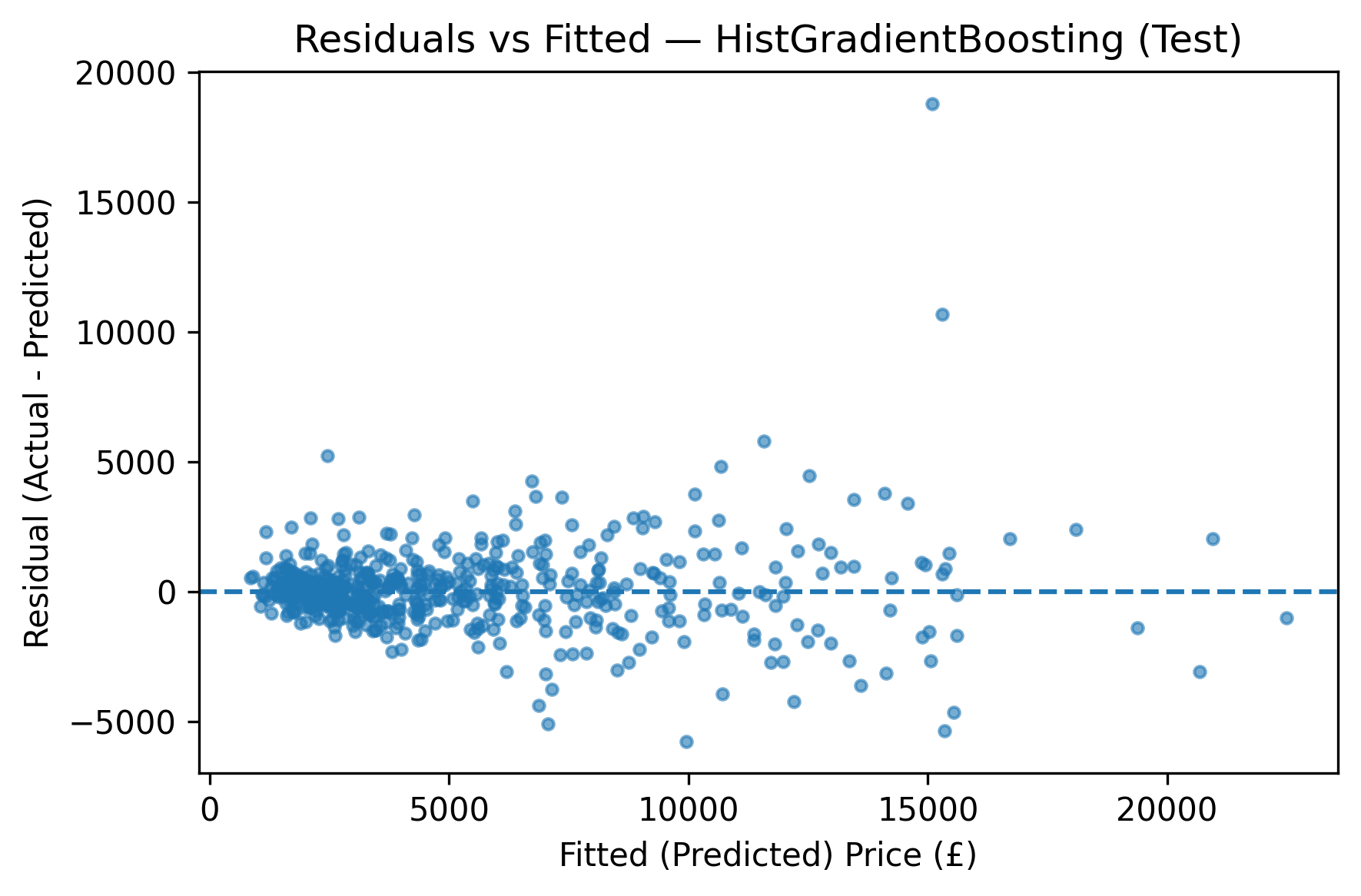


Figure 2. Residuals vs. Fitted values.

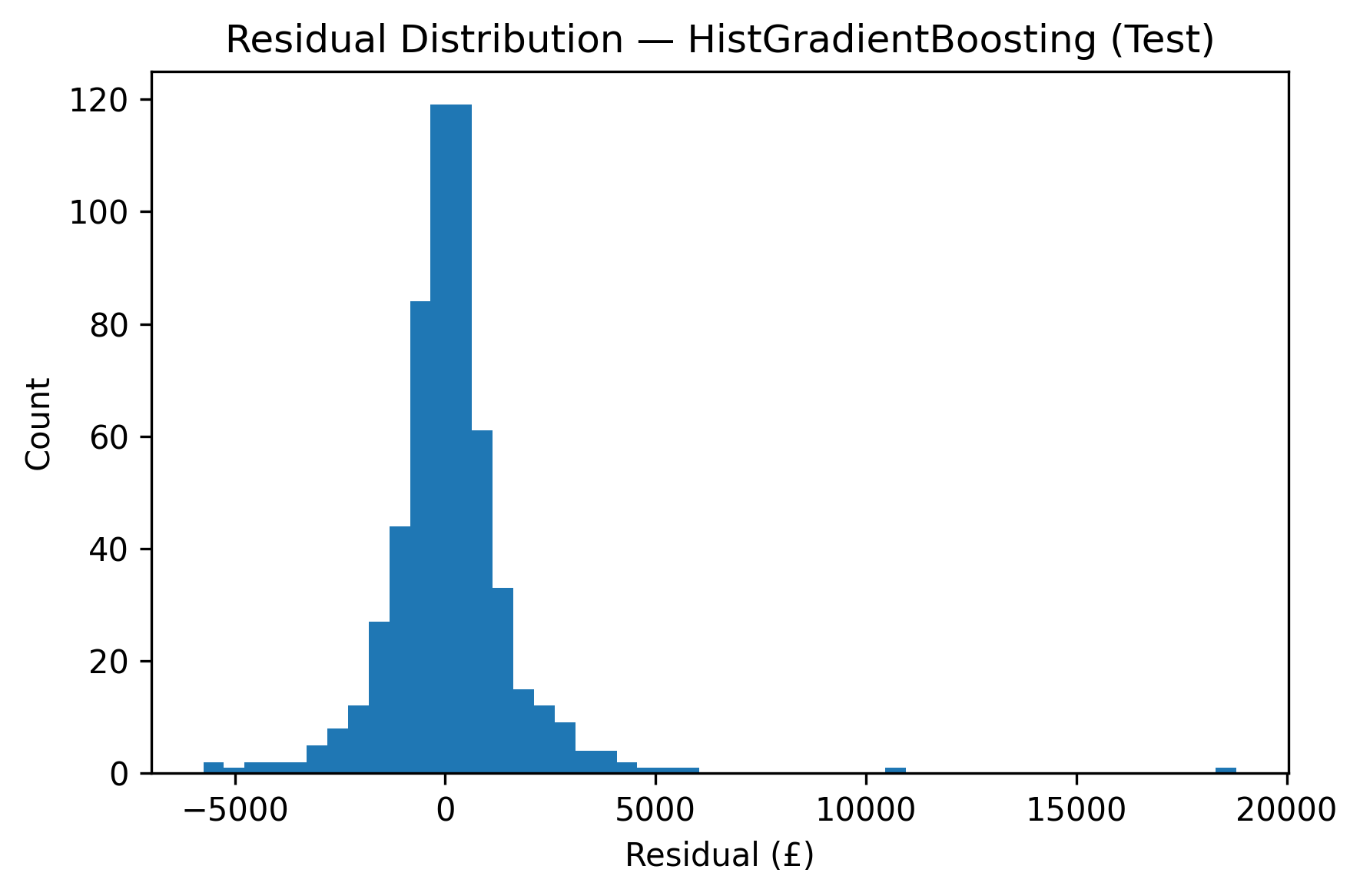


Figure 3. Residual distribution (histogram).

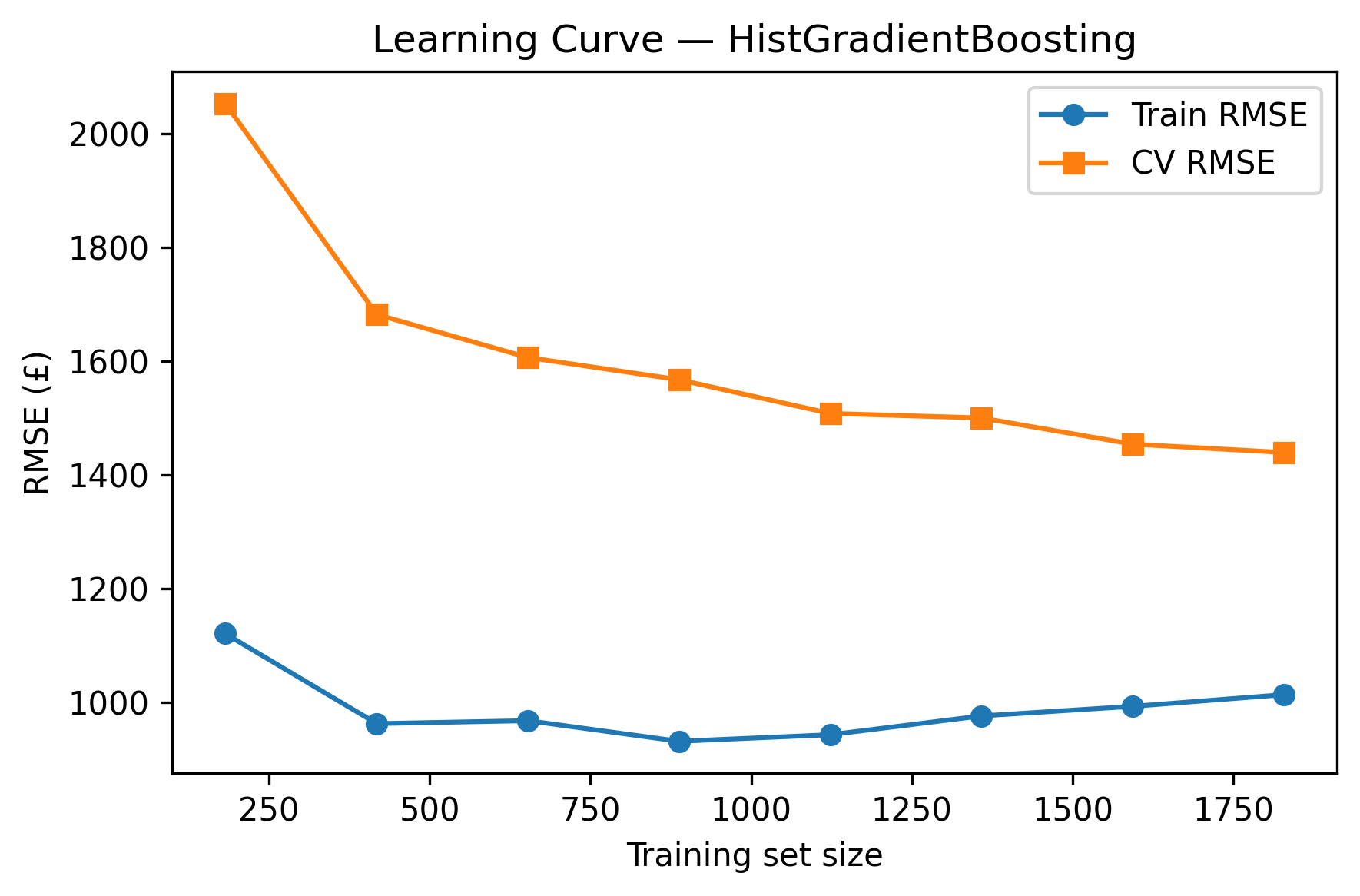


Figure 4. Learning curve (Train and CV RMSE).

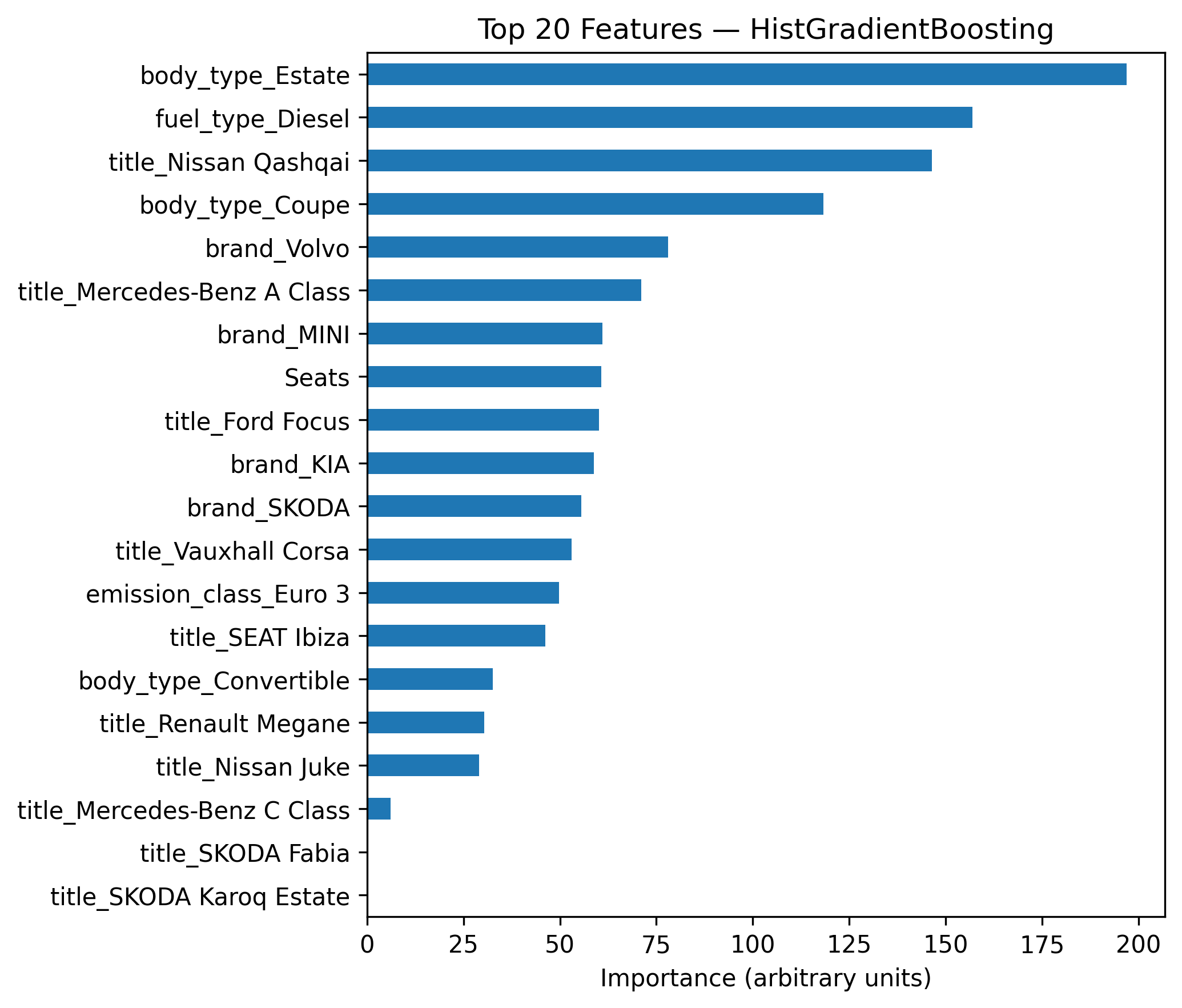


Figure 5. Top-20 feature importances (tree/permutation).

# 6. Conclusions

The selected model demonstrates competitive generalization on the hold-out test set. Diagnostic plots indicate the error structure and areas for improvement (e.g., mild heteroscedasticity if residual spread grows with price). Future work may include richer feature engineering (e.g., age from registration year, interaction terms), geographic aggregation (postcode-level effects), nested cross-validation for unbiased model selection, and fairness/robustness checks.