London House Price Regression Project

Abstract

This project focused on predicting London house prices using regression models. As housing prices in London are both a key economic indicator and a subject of public interest, developing accurate predictive models can provide valuable insights for buyers, sellers, and policymakers. The dataset contained over 418,000 property records with 28 features, including geographic information, property characteristics, and historical valuation data. A structured workflow was followed, beginning with data cleaning and exploratory analysis, followed by feature engineering and preprocessing to address issues such as multicollinearity, skewed distributions, and outliers. Predictive modeling was then carried out using both Linear Regression and XGBoost, with the latter achieving far superior performance. The study highlights the importance of thorough preprocessing and the effectiveness of advanced machine learning techniques in producing reliable and interpretable real estate price predictions.

1 Data Source

The dataset used in this project was obtained from Kaggle: "London House Price Data". It was provided in a Parquet format and contained approximately 418,201 property records. Each record included detailed property information such as:

- Full address, postcode, and geographic coordinates (latitude, longitude)
- Property characteristics (number of bedrooms, bathrooms, living rooms, floor area) - Property type and tenure (freehold, leasehold, etc.)
- Energy ratings
- Historical estimates of rental and sale values with confidence levels
- Transaction history including price changes over time

For the purpose of this regression project, the target variable selected was the saleEstimate_currentPrice column, which represents the current estimated market value of each property.

2 Data Cleaning

The following tasks were carried out:

1. Column Reduction: Non-essential columns such as full addresses and redundant sale/rent estimates were removed. This reduced noise and improved computational efficiency.
2. Handling Duplicates: 5,409 duplicate rows were identified and removed, ensuring no data leakage.
3. Missing Values: Missing data were handled according to their proportion.
    - Features with less than 5% missing values (saleEstimate_currentPrice, tenure, propertyType) were dropped since removal at this level has negligible impact on the dataset.
    - For features with 5–20% missing values (bathrooms, bedrooms, floorAreaSqM, livingRooms), values were imputed using grouped medians based on property type and postcode.
    - The feature with more than 20% missing values (currentEnergyRating) was assigned the category "Unknown" to retain records while avoiding potentially misleading imputations.
4. Outlier Treatment:
    - Extreme values were capped using the 0.5th and 99.5th percentiles.
    - Skewed variables, including sale prices and floor areas, were log-transformed to normalize distributions.
5. Feature Engineering:
    - Created a bedroom-to-bathroom ratio to capture density effects.
    - Extracted time-based features (year, month, day of week) from property history dates.
6. Final Dataset: After cleaning, the dataset contained 400,893 entries with 14 relevant features.

3 Exploratory Data Analysis

EDA was used to understand the distribution of features and identify important trends:

- The correlation heatmap revealed strong multicollinearity among the rent and sale estimate features (correlations above 0.9), indicating that they carry highly redundant information. To reduce this redundancy and avoid instability in modeling, columns with very high correlations were dropped, retaining only representative variables. Geographic coordinates (latitude and longitude) showed minimal correlation with other features, while structural attributes such as floor area, bedrooms, and bathrooms exhibited moderate positive correlations with price-related variables.

- The distribution plots showed that property attributes such as bathrooms, bedrooms, and living rooms are highly skewed, with the majority of properties clustered at the lower end (e.g., 1–2 bedrooms, 1 bathroom). Floor area also exhibited a right-skewed distribution, with most homes concentrated between 50–150 m^2 but some extreme cases extending beyond 400 m^2. Similarly, the sale estimate variable displayed a heavy right tail, highlighting the presence of very high-value properties.
- Boxplots confirmed the existence of significant outliers across multiple features, particularly floor area and sale price. Since these extreme values could distort model performance, outlier treatment strategies were applied. Depending on the feature, this involved either capping extreme values within reasonable bounds or applying transformations to reduce skewness.

Overall, the EDA highlighted the importance of dimensionality reduction by addressing multicollinearity and ensuring robust preprocessing through the handling of skewed distributions and outliers.

4 Machine Learning Model

4.1 Data processing

The dataset included both numerical and categorical variables. Numerical features such as latitude, longitude, property attributes (e.g., bathrooms, bedrooms), engineered features (e.g., log-transformed floor area, bedroom-to-bathroom ratio), and temporal variables from transaction history were standardized using StandardScaler to ensure consistent scaling across different ranges.

Categorical features such as tenure, property type, and current energy rating were encoded using One-Hot Encoding to create binary indicator variables. After preprocessing, the processed numerical and categorical features were concatenated to form the final training and testing datasets. The target variable (saleEstimate_currentPrice) was log-transformed to reduce skewness and stabilize variance, making the modeling more robust.

4.2 Model Building

Two regression models were implemented:

- Linear Regression: A baseline model to capture linear relationships between predictors and the log-transformed target variable.
- Extreme Gradient Boosting (XGBoost): A tree-based ensemble method, configured with 100 estimators, a learning rate of 0.1, and maximum tree depth of 6. XGBoost was chosen for its ability to capture non-linear relationships and handle complex feature interactions effectively.

Both models were trained on 80% of the dataset, with 20% reserved for testing.

4.3 Evaluation Metrics

Models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score, with Adjusted R^2 also computed for XGBoost to account for the number of predictors.

Linear Regression achieved an MAE of 0.266, RMSE of 0.354, and R^2 of 0.694, indicating moderate predictive accuracy but limited ability to capture non-linear patterns.

XGBoost significantly outperformed Linear Regression, achieving an MAE of 0.144, RMSE of 0.199, and R^2 of 0.903. The Adjusted R^2 of 0.903 confirmed that the model explained over 90% of the variance in the log-transformed sale price, with minimal penalty for model complexity.

These results highlight that while Linear Regression provides a useful benchmark, XGBoost offers superior performance by capturing complex interactions between property features and sale prices.

5 Conclusion

This project demonstrated the process of cleaning, preparing, and modeling real-world housing data. It emphasized the importance of thorough preprocessing and careful handling of outliers, while also showing the benefits of advanced machine learning techniques. The results highlighted that advanced models can significantly improve property price prediction: while Linear Regression served as a baseline, XGBoost achieved far superior performance (Adjusted R^2 0.90), demonstrating its strength in capturing non-linear relationships.