Predictive Modelling for London Housing Market Investment

Daniel Thompson

Abstract

This project leverages predictive modelling to guide investment decisions in London's housing market during the post-Brexit and COVID-19 era. It entails data preprocessing, exploratory data analysis (EDA), and the use of predictive algorithms to accurately forecast house prices, culminating in the selection of the top two hundred investment opportunities.

The project begins with meticulous data preprocessing and EDA. This process involves cleaning the dataset, addressing missing values, and transforming categorical variables. EDA, enhanced with descriptive statistics and visualisations, provides deep insights into feature distributions and correlations.

Central to the project are four predictive algorithms: XGBoost, Ridge Regression, Random Forest Regression, and Linear Regression. Each model undergoes thorough training, evaluation, and fine-tuning, assessed using performance metrics such as R-squared. To boost predictive accuracy, a stacking strategy is employed, amalgamating the strengths of individual models into a comprehensive meta-model. The meta-model attained an R-squared of 88.3%, meaning that it could predict house prices using unseen data with a high level of accuracy.

Additionally, the project discusses the expected impact of the Elizabeth Line on property values, proposing a method to assess the effects of proximity to Crossrail stations on house prices. This theoretical approach signifies an interesting direction for future research, especially in evaluating investment opportunities in areas likely to be influenced by this transportation development.

In summary, this project presents a data-driven approach for making informed investment choices in London's real estate market, illustrating the effectiveness of predictive modelling in navigating the complex dynamics of the post-Brexit and COVID-19 era.

Introduction

This data science project is centred on developing predictive models for the London housing market to aid in investment decisions. The analysis is anchored in two pivotal performance metrics: R-Squared and RMSE. R-Squared assesses the proportion of variance in the dependent variable predictable from independent variables, indicating the model's explanatory power. RMSE evaluates the average magnitude of prediction errors, providing insights into prediction precision.

The initial phase encompasses loading and examining two data sets: one with actual house prices and another with asking prices. This stage involves rigorous data preprocessing, including the transformation of character columns into factors for consistency, addressing missing values, and scrutinising the datasets' structures. Subsequently, the datasets are divided into training and testing sets for effective model evaluation and validation.

Further sections of this report are dedicated to detailing the development and performance evaluation of diverse predictive models such as Linear Regression, Decision Tree, XGBoost, Ridge Regression, and Random Forest. The project culminates in applying these models to select the top two hundred properties for investment, a process founded on thorough data analysis and predictive accuracy.

Data Loading and Preprocessing

The initial steps of this project involve crucial data loading and preprocessing stages, essential for laying the groundwork for subsequent analyses. This process ensures data integrity and usability:

1. **Data Importation**: Two key datasets are imported: one containing actual house prices (training data) and the other featuring asking prices (out-of-sample data). Predictions will be made on the out-of-sample data, which is not used in training.

2. **Data Transformation and Consistency**: To ensure consistent data types across datasets, date strings are standardised into Date objects and character columns are converted into categorical factors. This step aids in precise data manipulation and analysis.

3. **Data Examination and Cleaning**: The datasets undergo an initial examination to understand their structure and data types. A thorough check for missing values is conducted, with lesser-used columns like 'address2' and 'town' being removed for reliability.

4. **Data Filtering**: Rows with missing population data are filtered out to enhance dataset robustness.

5. **Data Splitting**: The training data is divided into training and testing sets in a 75:25 ratio, crucial for model training and performance evaluation on unseen data.

6. **Addressing High Cardinality and Consistency Issues**: Variables with high cardinality, such as 'nearest_station' and 'postcode_short', are addressed for efficient analysis. Stations with minimal occurrences are categorised as 'Other', and inconsistencies in 'postcode_short' levels are harmonised across datasets.

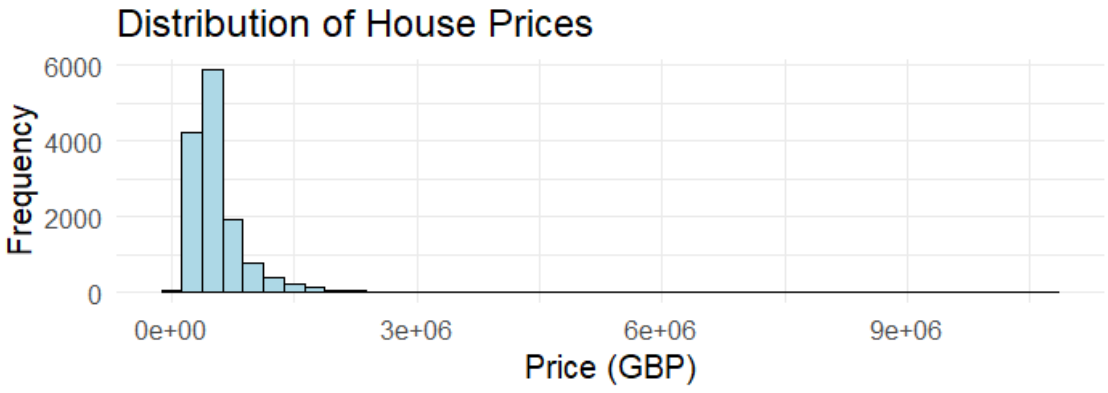
Exploratory Data Analysis (EDA)

Visualisations

The EDA process begins with visualising key aspects of the housing market data to understand underlying patterns and anomalies.

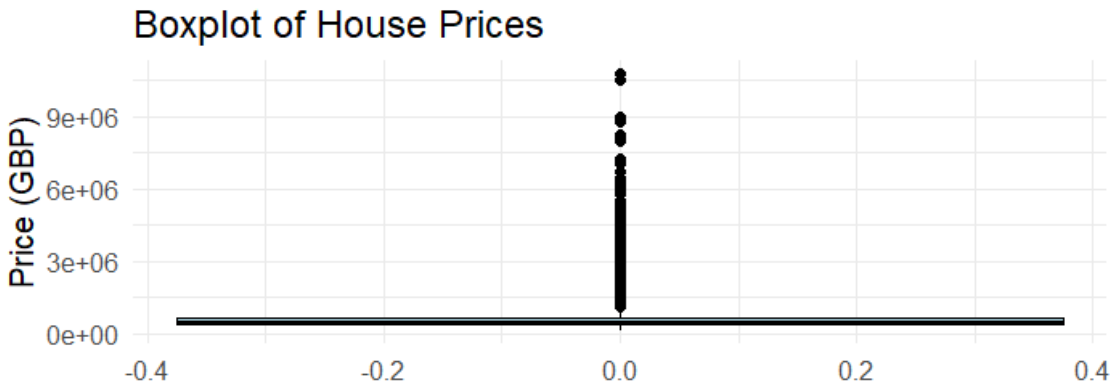
Histogram of House Prices:

- The histogram reveals a positive skew in the distribution of house prices, with a concentration between £250,000 and £750,000.



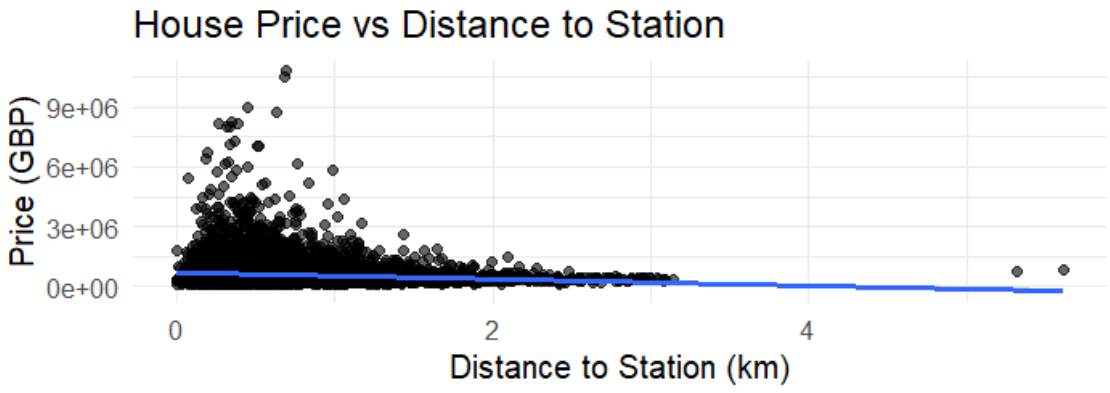
Boxplot for House Prices:

- A boxplot uncovers numerous outliers, indicating significant variations in house prices.

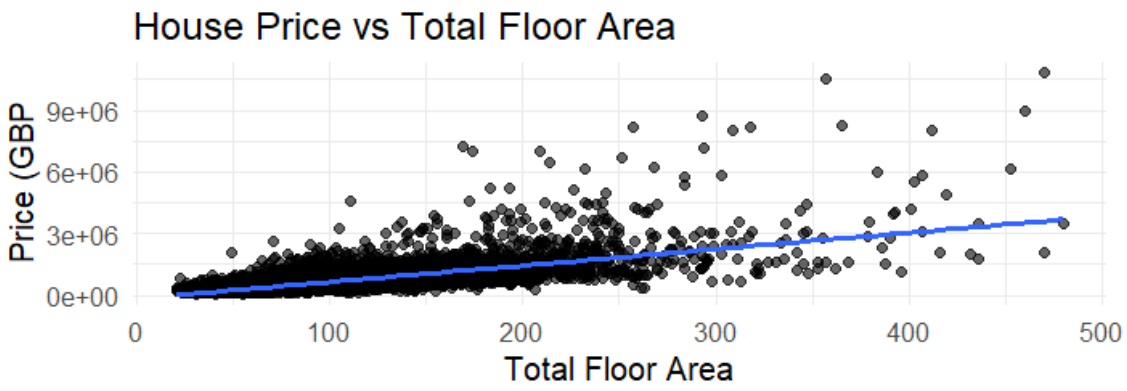


Scatter Plots for Price Correlations:

- **Price vs. Distance to Station:** Indicates a negative correlation; prices tend to decrease as the distance to the nearest station increases.

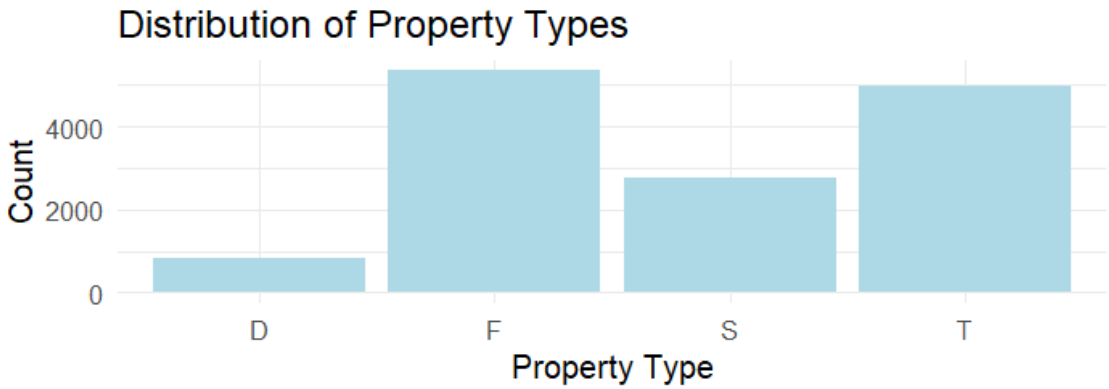


- **Price vs. Total Floor Area:** Shows a positive correlation with increasing variation, suggesting possible interaction with other variables.



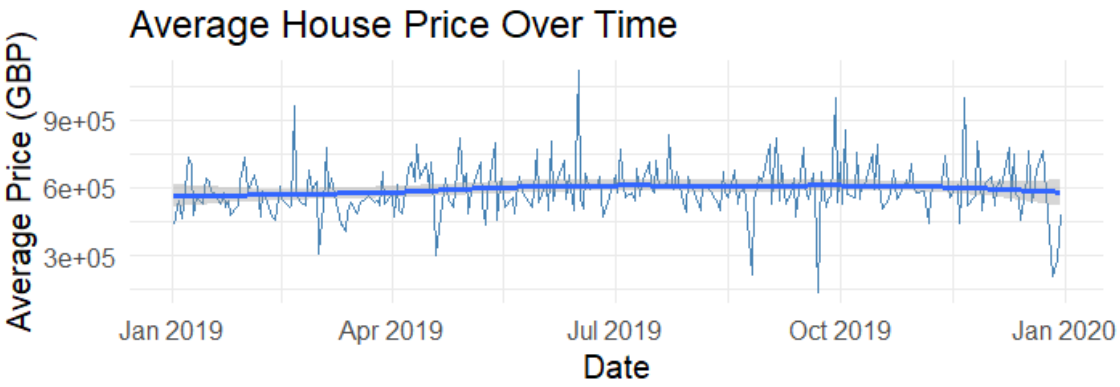
Bar Plot for Property Types:

- This plot shows the distribution of property types, highlighting Flats and Terraced houses as the most common in London.



Time Series Analysis of Average House Price:

- A time series plot reveals a steady trend in average house prices over time, although this raises questions about the model's robustness in more volatile markets.



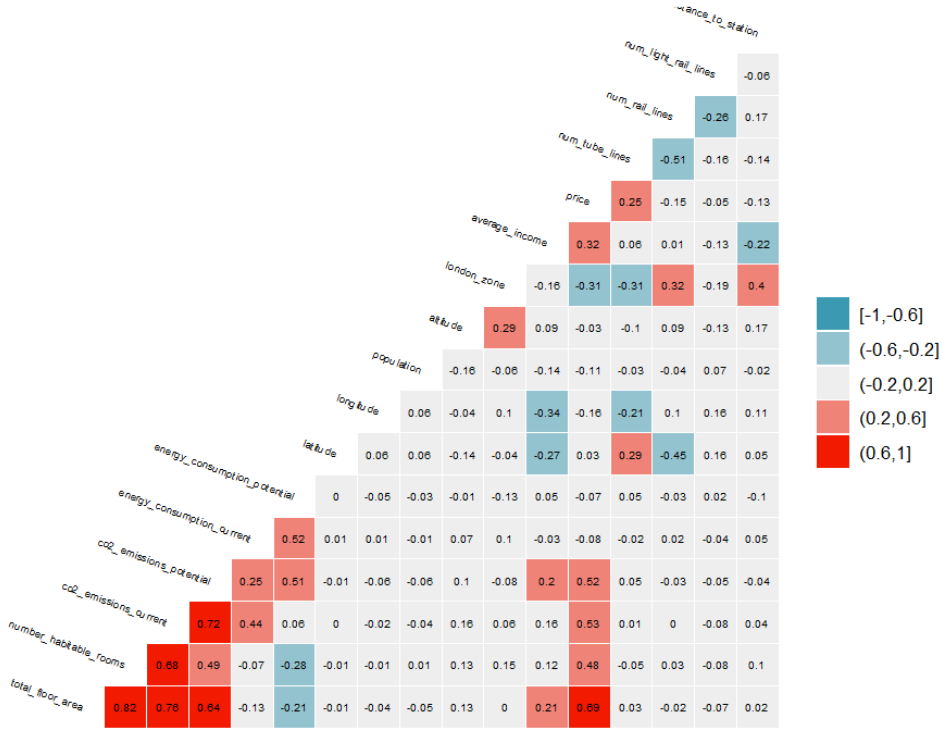
Observations and Decisions

- **Handling Outliers**: The abundance of outliers in the house prices leads to the decision to retain them for a more comprehensive model that can predict extreme values effectively.

- **Variable Correlation and Implications**: The analysis of scatter plots and property types provides valuable insights into factors influencing house prices. For instance, proximity to stations and floor area emerge as significant variables, affecting the desirability and value of properties.

- **Market Trend Analysis**: The time series analysis offers a glimpse into the overall market trends, crucial for understanding long-term patterns and potential market shifts.

Correlation Matrix



Key Findings from the Correlation Matrix

- **Moderate to Strong Correlations:** The analysis revealed several noteworthy correlations. Most prominently, 'total_floor_area' demonstrated a strong correlation with house prices, indicating a significant relationship between the size of a property and its value. This correlation, quantified at 0.69, suggests that as the total floor area increases, so does the price.

- **Variables with Medium Correlation:** Other variables like 'co2_emissions_potential', 'co2_emissions_current', and 'number of habitable rooms' showed a medium degree of correlation. These findings are intuitive, as they relate to the environmental impact and living space of the properties, factors increasingly valued in modern housing markets.

- **Variables with Weak Correlation:** Interestingly, factors such as 'number of tube lines accessible', 'average income', and 'London zone' demonstrated weaker correlations. While these factors are less directly tied to property prices, they offer valuable contextual information that could influence pricing in less direct ways.

Feature Engineering Based on Correlations

- **Focus on 'Total Floor Area':** Given its strong correlation with price, 'total_floor_area' is chosen for feature engineering. An interaction term between 'total_floor_area' and 'postcode_short' is introduced, hypothesising that the value per square meter may vary across different postcodes.

- **Polynomial Term for Non-Linear Relationships:** To capture potential non-linear relationships, a polynomial term for 'total_floor_area' is added. This step aims to refine the model's accuracy in predicting prices, particularly for properties with unusually large or small floor areas.

The correlation analysis has been instrumental in identifying key variables for further exploration and model enhancement. The insights derived will guide the next stages of our analysis, particularly in model building and feature engineering. The subsequent sections of the report will build upon the foundation laid by this correlation analysis.

Further Preprocessing

During the data preparation phase, it was discovered that several variables present in the training dataset were absent in the out-of-sample dataset, including postcode, address1, address3, local_aut, county, and date. For the sake of consistency and comparability in dataset analysis and model development, these variables were removed from both the training and testing datasets.

Building and Implementing the Model

Linear Regression Model

A Linear Regression model was built to predict house prices in London. The model was constructed with careful consideration of feature engineering, interaction terms, and polynomial variables to enhance its predictive power.

Feature Engineering

- *Interaction Term (total_floor_area:postcode_short):* An interaction term was introduced to capture the relationship between the total floor area of a property and its location, as indicated by the postcode. This interaction term allowed the model to account for variations in property values across different postcodes, acknowledging that the impact of floor area on price can vary by location.
- *Polynomial Term (poly(total_floor_area, 2)):* A polynomial term for total floor area was included to account for potential non-linear relationships between floor area and price. This was important because larger properties in the real estate market may not exhibit a simple linear correlation with price. The polynomial term allowed the model to capture more complex patterns.

Model Training

- The Linear Regression model was trained using the `train` function with the method specified as "lm" (Linear Model).
- Cross-validation with five folds was used to ensure robustness and avoid overfitting.
- The training data included features such as property type, total floor area, CO2 emissions potential, energy consumption potential, and more.

Model Evaluation

- The performance of the Linear Regression model was evaluated using metrics such as Root Mean Square Error (RMSE) and R-squared (R^2) on a test dataset.
- An R-squared value of approximately 84.1% was achieved, indicating that the model effectively explained 84.1% of the variance in house prices. This high R-squared value demonstrated the model's ability to capture key factors influencing London's house prices.

Decision Tree Model

A decision tree model was also built with the primary goal of identifying the most influential variables affecting house prices. The `rpart` library was utilised to build the decision tree model, with the target variable being 'price.'

Key steps in developing the tree model included:

- Creating a basic decision tree.

- Determining the top twenty most important features from the dataset. These features included factors like 'total_floor_area,' 'nearest_station,' and 'postcode_short.'
- Building a refined tree model by training it with the selected features, including interaction terms and polynomial elements.

When compared to the linear regression model, the decision tree model showed lower performance, with an R-squared value of 62.3%, whereas the linear model achieved an R-squared of 84.1%. This suggests that the linear model was better at interpreting the data. The decision tree model, while insightful, struggled with the complexity of the data.

XGBoost Model

XGBoost, acknowledged for its efficiency, flexibility, and high performance, especially in structured data scenarios, was utilised in this project. It falls into the category of ensemble methods known as boosting, where models are constructed sequentially, with each model aiming to rectify the errors of its predecessor. XGBoost is distinguished by its capacity to handle a wide range of data types, its resistance to overfitting, and its effective management of missing data and high-dimensional spaces.

Model Construction and Fine-Tuning

In this project, XGBoost was selected due to its established track record in accurately predicting outcomes in complex datasets, such as the London housing market. The model was customised to the nuances of the data through fine-tuning capabilities, including parameters like maximum depth, and learning rate. Extensive fine-tuning was conducted through trial and error to optimise performance.

Results

Impressive results were obtained with the XGBoost model. An R-squared value of approximately 86.8% was achieved, demonstrating the model's robustness and accuracy in the prediction of house prices. This high level of performance instils confidence in the model's predictions, making it an invaluable tool for the identification of profitable investment opportunities in the market.

Ridge Regression Model

Ridge regression, recognised for its effectiveness in handling models with many predictor variables, plays a crucial role in this project. By introducing a penalty term (lambda) to the cost

function, the impact of less significant variables is minimied, addressing common issues like multicollinearity and overfitting in high-dimensional data.

Model Construction and Fine-Tuning

The model achieved an R-squared value of approximately 85.5%, underscoring its effectiveness. Opting for alpha = 0 (ridge regression) proved to be more effective than a mixed elastic net approach, likely due to its simplicity and direct approach in shrinking coefficients, thereby enhancing model performance.

Balancing Complexity and Predictability

In the context of this project, focused on accurate prediction of London housing prices, ridge regression strikes a balance between complexity and predictability. It ensures robust predictions without overfitting the training data, a critical consideration in dealing with high-dimensional datasets where multicollinearity and overfitting can present significant challenges.

Random Forest Regression Mode

Ensemble Learning for Enhanced Predictions

Following the development of the tree model, the next step in our predictive journey involves the implementation of a random forest regression. Random forest, an ensemble learning method, builds upon the concept of decision trees. It creates a 'forest' of trees where each tree is trained on a random subset of the data, and the final output is determined by averaging the predictions from all trees. This approach inherently reduces variance, preventing the overfitting issues often seen in individual tree models.

Balancing Complexity

In the analysis, addressing the high cardinality of variables like 'nearest_station' presented a significant challenge. While such features could potentially add depth to the model, they also introduce complexity, which may lead to diminished returns in terms of performance. Consequently, the decision was made to exclude 'nearest_station' from the random forest model. This choice aimed to decrease model complexity for improved performance. Conversely, retaining 'postcode_short' proved beneficial as it significantly contributed to a higher R-squared, indicating its importance in predicting house prices. Lastly, the integration of the caret and ranger packages allowed me to effectively model interaction terms between categorical variables and continuous variables, enabling a more nuanced and accurate representation of the complex factors influencing London's housing market.

Superior Performance and Enhanced Accuracy

The Random Forest model, achieving an R-squared of approximately 85.4%, demonstrates superior performance compared to the basic tree model. This improvement is attributed to Random Forest's ability to average multiple decision tree outputs, thereby reducing variance, and mitigating overfitting. This enhanced accuracy is essential for making reliable real estate investment decisions in our project and as a result, the decision tree model will be left out of the stacked meta-model.

Stacking

Stacking

Integration of XGBoost, Random Forest, Ridge Regression, and Linear Regression The predictive strength of multiple algorithms, namely XGBoost, Random Forest, Ridge Regression, and Linear Regression, is leveraged by the stacking method. Diverse data patterns and relationships are captured through this approach, enhancing the robustness and accuracy of predictions.

The Stacking Process

The stacking process involves generating predictions for training data using each of the base models mentioned earlier. These predictions are then combined to train a Ridge Regression Meta-Model via an elastic net model with alpha set to 0, signifying pure Ridge Regression without Lasso regularisation. The optimal lambda value is selected through cross-validation. The same process is applied to test data to make final predictions using the Ridge Regression Meta-Model.

Enhanced Predictive Performance

The stacking approach results in a final R-squared value of approximately 88.4%, signifying a high level of predictive accuracy. This enhanced capability is invaluable for navigating the complexities of the real estate market and ensuring well-informed investment decisions.

Assessing the Impact of The Elizabeth Line

To assess the potential profit opportunity in investing in neighbourhoods serviced by the Elizabeth Line, the predictive modelling approach can be adapted to evaluate the impact of Crossrail on house prices. The procedure involves the following steps:

1. *Data Integration and Modification:* The first step is the integration of location data of Crossrail stations into the existing housing market dataset. Each property in the dataset is assigned a new variable, 'distance_to_Crossrail_station,' calculated based on the proximity of the property to the nearest Crossrail station.

2. *Exploratory Data Analysis (EDA):* An EDA is conducted to understand the current relationship between 'distance_to_Crossrail_station' and house prices. This analysis helps in identifying trends and patterns, such as whether properties closer to Crossrail stations command higher prices.

3. *Feature Engineering:* Based on the EDA, new interaction terms or polynomial features may be introduced to capture non-linear relationships. For example, an interaction term between 'distance_to_Crossrail_station' and other variables such as 'total_floor_area' might reveal how proximity to a Crossrail station differently impacts properties of varying sizes.

4. *Model Modification and Training:* The predictive models are updated to include 'distance_to_Crossrail_station' as a predictor variable. This modification allows for the assessment of the degree to which proximity to Crossrail stations influences house prices. The models, including Linear Regression, Decision Tree, XGBoost, Ridge Regression, and Random Forest, are retrained with this augmented dataset.

5. *Evaluation of Crossrail Impact:* Using the updated models, the impact of the Elizabeth Line on house prices is evaluated. This involves comparing the predictive performance of the models with and without the 'distance_to_Crossrail_station' variable. A significant improvement in predictive accuracy with the inclusion of this variable would indicate a strong influence of the Elizabeth Line on property values.

6. *Identification of Investment Opportunities:* The final step involves applying the enhanced models to identify potential investment opportunities. Properties in neighborhoods served by the Elizabeth Line that are predicted to have an underpriced value relative to their predicted price post-Crossrail completion are flagged as potential investment opportunities.

7. *Continuous Monitoring and Model Updating:* Given the dynamic nature of real estate markets and ongoing developments in the Crossrail project, the models require continuous monitoring and updating. This ensures that the predictions remain relevant and accurate,

adapting to new data and market trends. This procedure aims to systematically evaluate the impact of the Elizabeth Line on London's housing market and to identify profitable investment opportunities arising from this significant infrastructure development.

Conclusion

Advanced predictive modelling techniques were successfully harnessed to analyse the London housing market, particularly in the post-Brexit and COVID-19 era. Key algorithms, including XGBoost, Ridge Regression, and Random Forest Regression, were utilised, resulting in significant predictive accuracy (88.3% R-squared). Notably, the impact of the Elizabeth Line was evaluated, providing nuanced insights for investment decision-making. The culmination of the project involved the identification of the top two hundred properties for investment, demonstrating the practical application of data science in real estate. Emphasis was placed on continuous model updating and monitoring to adapt to market changes, ensuring the model's relevance and accuracy in a dynamic environment. This project not only highlights the power of data science in real estate investment but also establishes a benchmark for future research and analysis in this field.