**Housing Price Analysis in the UK**

Table of Contents

[Introduction 3](#_Toc163616136)

[Data and EDA 4](#_Toc163616137)

[Brief Explanation of Data Set 4](#_Toc163616138)

[A Brief EDA: Features, Labels and Correlations 5](#_Toc163616139)

[Data Preprocessing 15](#_Toc163616140)

[Feature Creation 16](#_Toc163616141)

[Data Encoding and Scaling 16](#_Toc163616142)

[Machine learning Analysis 17](#_Toc163616143)

[Model Choice 17](#_Toc163616144)

[Model Justification 18](#_Toc163616145)

[Model Explanation, Specification and Results 19](#_Toc163616146)

[Linear Regression Model 20](#_Toc163616147)

[Discussions and Conclusions 23](#_Toc163616148)

[Results 23](#_Toc163616149)

[Recommendations 24](#_Toc163616150)

[References 27](#_Toc163616151)

[Appendices 28](#_Toc163616152)

# Introduction

The real estate sector contributes a significant portion of a nation’s economic growth. However, the soaring housing prices have become have lowered the hopes for individuals who want to rent or own a home.High house prices are one of the key factors that have created a ‘perfect storm’ and growing unlikelihood of home ownership in the UK (Butler, 2023). This has made it more difficult for individuals to get on the property ladder. Therefore, there is a need to investigate the factors that affect house prices to improve house affordability. During the past decade, there has been a rapid increase in the number of studies on intertemporal changes in real estate property prices. Such an upsurge is attributed to the larger availability of all-embracing micro-level data sets, developments in modeling techniques, and extended business applications. Geerts et al., (2023) adds that predicting house price is a difficult problem since real estate valuations is affected by many factors including economic factors such as inflation, interest rates, mortgage rates and physical factors such as location, physical characteristics, and neighborhood. However, data science techniques such as descriptive, time series and inferential analysis and other recent modelling techniques such as ML have contributed to the rise in the study of house prices (Engerstam, 2023). Understanding the factors that drive property values in the ever-changing and complex UK property market can help one make informed decisions about buying, selling, or investing in residential property.

Analytics has played a key role in solving the housing problem globally. Various real estate players use analytical tools such as regression algorithms to predict the estimated price of a house. Moreover, Microsoft Power BI, python, SQL and other analytical tools have been used to predict house prices and sales volume of properties. Therefore, analytics play a crucial role in understanding the behavior of real estate market and predicting house prices and sales volume, understanding the relationship between house prices and its determinants, and observing trends. For instance, Power BI data analysis has been used to compare housing market trends in the UK through various statistics. The two periods show that a fall in interest rate results to a fall in house prices and vice versa. Similarly, in 1991, at 11% interest the average UK house prices fell by 1.43%. A further interest fell to 8.38% and 5.63% in 1993 and 1990 respectively led to decline in house prices by 3.98% and 1.69% respectively (Simoneli, 2023). Data analysis tools such as descriptive a show that the house prices trend line goes up the same way the interest rates are being released (Simoneli, 2023). Schonlau & Zou (2020) explain that exploratory data analysis (EDA) and correlation analysis techniques are effective in predicting house price, depending upon all the other variables. Statistical analysis helps property investors and consumers to highlight the trends and patterns in the housing sector. This paper focuses on predicting the house prices using random forest ML model.

Data and EDA **(450 words)**

## Brief Explanation of Data Set

The data set was obtained from Kaggle.com and contains key variables that show the factors that affect house prices in the UK. The dataset comprises of 13 features that are mainly divided into numeric and categorical features. The numeric features are price, area, bedrooms, bathrooms, parking and stories. The categorical features in the dataset include main road, guestroom, basement, hot water heating, air conditioning, area, furnishing status and main road. The price is the target variable while the predictor variables will be the numeric and categorical features.

The dataset does not have missing values in any of the columns as proven by the codes in appendix. In terms of data types, the data can be categorized as quantitative and qualitative. The quantitative ones are discrete values as they have numeric features while the qualitative are nominal with “yes or no categorical features.

## A Brief EDA: Features, Labels and Correlations

EDA is effective in predicting house prices in the UK based on the independent variables. The histogram shows that majority of the houses are priced between £2,000,000 and £8,000,000.



The histogram shows a plot of house price frequency against the price level, where the general shape of the distribution points to a Poisson-like distribution that is a unimodal distribution, which is skewed to the right. The median price for the distribution is between the 0.3 and 0.4 price levels. The graph shows that there are more occurrences of houses in the lower price ranges than in the higher price ranges.



The histogram above shows that there are more houses in the lower area distribution beginning from the 2,000 to 4,000 square feet levels than in the higher house area distribution from 14,000 to 16,000 square feet. This is likely to be because smaller houses are generally cheaper to build and buy.

The distribution below is a barchart of the distribution of the furnishingstatus variable, plotted by evaluating count versus the furnishing status of the house in question. The plot shows that semi-furnished houses are the highest in occurrence followed by the unfurnished houses, and lastly followed by furnished houses. Most of the houses have 3 bedrooms. Only a few houses have 1 bedroom and 6 bedrooms. Through this metric, one can conclude that the UK housing market needs to enhance the status of furnishing in the available houses in the market as most seem to fall in the semi-furnsihed and unfurnished categories. 

Many of the houses do not have guestrooms, basement, hot water heating, air conditioning and prefab area. On mainroad, many of the houses are located near mainroads.

On furnishing status, most of the houses are semi-furnished followed by unfurnished and then furnished. However, the difference between number of houses withoin the 3 furnishing statuses is minimal.

On parking area, most of the houses lack parking area. A few have a parking space for 1 and 2 cars.

The distribution above represents the preferred area, which is plotted against the count of such areas in the housing market. The data distribution demonstrates that most individuals prefer 

On stories, most of the houses have 1 and 2 stories. Houses with 2 stories are slightly more than those with 1 stories.



On bathrooms, the plot shows that many houses have 1 bathrooms followed by 2 bathrooms. Only a few houses have more than 2 bathrooms.



The scatterplot of price and area above shows the possibility of linear relationship between price and area. However, there are many data points that are scattered and this shows the linear relationship may not be a strong one.



The box plot above suggests that there is a positive correlation between bedrooms and house price. That is, there is a higher median price with an increase in the count of bedrooms.



The scatter plot above suggests that there exists a positive relationship between the bedroom area and price. However, the relationship is not linearly perfect as the data points are scattered.



The plot above shows that many houses have 3 bedrooms followed by 4 bedrooms. Only a few houses have 1 and 6 bedrooms.This shows there is a demand for family houses.

# Data Preprocessing

Data preprocessing is the iterative process for transforming raw data to reduce noise and make the data ready for data analysis (Yoo, Ramirez, & Liuzz, 2014). Some of the steps during data processing include cleaning the data, data transformation, data integration and data reduction. In the previous EDA section, the dataset was found not be having missing values, duplicates, or inconsistencies hence no need for data cleaning. Since the dataset is clean, other preprocessing steps can be done on the data before it is passed to the machine learning.

## Feature Creation

Machine learning performs better when trained using quality features. Feature engineering is among the ways of ensuring that the model trains using quality data as it entails creating new features from the existing features. The first feature engineering step done for this research is feature creation. The feature created for this analysis is is "price\_per\_sqft" which is arrived at by dividing the price by the area. This feature is important as it offers insights to the model on the cost of house per square feet.

## Data Encoding and Scaling

Another feature engineering step is data encoding and scaling. Encoding the categorical variables is key step in data transformation process as it ensures that the data is in a format that the machine learning models can understand (Cerda, Varoquaux, & Kégel, 2018). There are various data encoding techniques such as one-hot encoding, label encoding, get\_dummies and binary encoding among others. These techniques convert categorical data into numerical format that the algorithm will understand. For this research, the technique used is get\_dummies. This approach changes the categorical variables into binary vectors with one element taking the value 1 and the rest taking the value 0. The categorical variables in the dataset that needed encoding are 'mainroad', 'guestroom', 'basement', 'hotwaterheating', 'airconditioning', 'prefarea', 'furnishingstatus'. Encoding using the get\_dummies python function changes the elements in these variables to numerical format.

The get\_dummies function takes parameters including the dataset, the columns to be encoded, a condition indicating that the first level of the categorical variable will be deleted and finally a parameter indicating that the datatype of the resulting encoded dummy variables should be integer. The ‘drop\_first=True’ parameter is important as it ensures that multicollinearity issues are avoided. Multicollinearity arises in the data when one dummy variable can predict the rest of the variables and when this occurs, the model will not give reliable predictions.

After encoding the data, the next step is scaling. Data scaling is a critical step in data processing as it ensures that the all the data is of the same scale. The commonly used method for scaling data is through the use of the `StandardScaler` function module from scikit-learn. This method was relevant as it improves the performance of the model by standardizing the range of values, avoiding potential outliers after the encoding steps above. Usually, the outcome of the data scaled using StandardScaler is a range between 0 and 1.

**Data splitting**

The final step in the data preprocessing step is data splitting. In this step the dataset is split into subsets of data which are used in mode training, validation and evaluation. This step is possibly through the ‘train\_test\_split’ function of the Sklearn library. Splitting the data leads to two new subsets of the data: train and test data. The train data is used to train the model while the test data is used to check the performance of the model on the new data. Yoo, Ramirez, & Liuzz, (2014) cite that proper data splitting is important in ensuring that resulting machine learning model is reliable and generalizable.

# Machine learning Analysis

## Model Choice

The researcher chose the random forest machine learning model as the main model for the analysis of housing prices in the United Kingdom. A random forest model is a tree-based model- where tree-based models focus on periodically dividing the dataset into two parts, then evaluate them according to a specific criterion until a specific condition is achieved (Schonlau & Zou, 2020). While most tree-based methods utilize this approach- it is prone to over-fitting, such as in the case of decision trees, but random forests overcome this problem by considering a subset of the dataset, building numerous trees on this set, and averaging the predictions from these numerous trees. As a result, random forests perform significantly better on prediction tasks, while accurately estimating the error rates in the process. According to research by Schonlau and Zou (2020), a comparison of the random forest machine learning method against both logistic regression and linear regression models demonstrated a superior performance of the random forest method against the latter two models, thus demonstrating its effectiveness in prediction tasks. In the comparison against logistic regression, random forests proved superior for predicting credit card defaults by providing a lower error rate than the logistic regression model. As a result, this research has adopted the random forest method as the primary model for this analysis.

## Model Justification

The random forest is the most suitable approach for this specific case as it is well-suited to handle complex relationships between housing factors which may be overlooked by simpler models such as linear regression. Since housing is a multivariable issue, this is crucial as it enables researchers to identify previously unidentified thus opening up avenues for enhanced research on the issue. Biau (2012) cites that random forest models are robust to irrelevant features, where through feature selection, they can reduce the significance of such features enabling a more accurate picture of the issue at hand. Additionally, random forests are unaffected by overfitting as they utilize random dataset subsets to prevent overfitting, ensuring an accurate model that performs well (Biau, 2012). Random forest models also provide feature importance enabling enhanced interpretability since one can visualize the features that are strongly correlated to the predictions.

## Model Explanation, Specification and Results

This section is dedicated to the setting up and training of a machine learning model for predicting house prices. One begins with the importing of libraries as shown below:

 `from sklearn.model\_selection import train\_test\_split`- this line imports the `train\_test\_split` function from scikit-learn's model selection module. The function is used to split the data into training and testing sets. In this section- `from sklearn.ensemble import RandomForestRegressor`- the line imports the `RandomForestRegressor` class from scikit-learn's ensemble module, where the class is used to create a random forest regression model. This line - `from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score`- imports three different metrics for evaluating the performance of a regression model: the `mean\_absolute\_error (MAE)`which calculates the average absolute difference between predicted and actual values; the `mean\_squared\_error (MSE)’ which calculates the average squared difference between predicted and actual values; and `r2\_score`which calculates the coefficient of determination, a measure of how well the model explains the variance in the target variable. In preparing data for training, the following lines are utilized:

`X = df\_house.drop(columns = ['price'])`- creates a new data frame, `X`, that contains all the features (columns) from `df\_house` except for the "price" column. The "price" column is the target variable to be predicted.

`y = df\_house['price']`- creates a new variable, `y`, that contains just the "price" column, which is the target variable we want to predicted. The next step involves splitting data into training and testing sets as shown below:

`X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)`: This line splits the data into training and testing sets using the `train\_test\_split` function.

The `test\_size=0.2` specifies that 20% of the data will be used for testing, and the remaining 80% will be used for training the model.

The `random\_state=42` sets a seed for the random number generator, ensuring reproducibility if you run the code again. This makes sure the split between training and testing data is always the same.

The next step involves creating and training the model:

The `RF = RandomForestRegressor (n\_estimators = 100, random\_state =42)`- creates a random forest regression model using the `RandomForestRegressor` class.

The `n\_estimators = 100` specifies that the model will use 100 decision trees. One can experiment with different numbers of trees to improve performance.

The `random\_state=42` sets the same random seed as before, again for reproducibility. The `RF.fit(X\_train, y\_train)`- trains the model using the training data (`X\_train` and `y\_train`). The model learns the relationship between the features in `X\_train` and the target variable `y\_train`.

The code sets up and trains a random forest model to predict house prices. The code is on Appendix 3.

## Linear Regression Model

This section extends the model evaluation section to analyze the feature importance in the random forest model. The first step is analyzing feature importance:

The line- feature\_importance = pd.DataFrame({'Feature': X\_train.columns, 'Importance': RF.feature\_importances\_}) creates a pandas DataFrame named feature\_importance. It uses the column names of X\_train (features) as the 'Feature' column in the DataFrame. Further, it extracts the feature importance scores from the trained model (RF.feature\_importances\_) and stores them in the 'Importance' column. The feature importance scores indicate how much each feature contributes to the model's predictions.

The line- feature\_importance = feature\_importance.sort\_values(by='Importance', ascending=False) sorts the DataFrame feature\_importance by the 'Importance' column in descending order. This ensures the features with the highest importance scores are at the top.

The next step - visualizing feature importance proceeds as follows: libraries like matplotlib (plt) and seaborn (sns) for data visualization. The plt.figure(figsize=(10, 6)) creates a figure for the plot with a specific size (10 inches width and 6 inches height). The sns.barplot(x='Importance', y='Feature', data=feature\_importance) creates a bar chart using seaborn. The x-axis represents the 'Importance' scores while the y-axis represents the 'Feature' names. The data for the plot is provided by the feature\_importance DataFrame. The function plt.title('Feature Importance') sets the title of the plot. The function plt.xlabel('Importance') and plt.ylabel('Feature') labels the x and y axes of the plot. The function plt.show() displays the generated bar chart.

****

****

The house price (independent variable) has a positive correlation with all the variables namely: area, bedrooms, bathrooms, stories, parking, and price per square feet. Thus, an increase in those variables such as number of bedrooms and parking space, results to an increase in the prices of the housing properties.

# Discussions and Conclusions

## Results

The results of the random forest model as shown below:

Train MAE: 0.04163872920905745, Validation MAE: 0.15135952520195808

Train MSE: 0.006902841320271715, Validation MSE: 0.09403135645844926

Train R-squared: 0.9921821222048643, Validation R-squared: 0.9350351789656476

While the results of the linear regression model are as shown below:

Linear Regression MSE: 0.168154659394538

Linear Regression R-squared: 0.88382452657177

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** |  | **Linear Regression** |
|  | Train | Validation |  |
| **RMSE** | 0.0069028 | 0.0940313 | 0.1681547 |
| **R-Squared Value** | 0.9921821 | 0.9350351 | 0.8838245 |

The results above are important as RMSE values enable researchers to evaluate how accurate a model is in predicting the value of a variable, while the R-squared value shows how accurate the features can explain the variation in the target variable. In this case, when looking at the quality of the model’s fit, we check the R-squared value which should be closer to 1. Here, the random forest has the upper hand against the linear regression. For RMSE, the most effective model should have the lowest value thus indicating low errors in the model. The random forest model has superior RMSE values as they are lower than the linear regression model.

## Recommendations

In this case, the random forest has better ability to predict housing prices in the UK. In regards to features that better predict housing prices, the table below shows the various feature importance plotted against each other. It is evident that area and price per square feet are the most significant predictors of housing prices, followed by number of bathrooms, availability of air conditioning, and parking space availability. All other variables have negligible effects upon predicting the housing prices in UK.



Understanding the key factors that increase real estate property prices is necessary to solve housing problems. Based on the statistical analysis results, the location, number of bathrooms, variations in square feet affect house prices. Therefore, it is recommendable for a buyer to check on the area offered in the listed houses in the market, as well as the prices per square feet before making renting or purchase decisions. The second factor property owners or consumers should focus on are the number of bathrooms, air conditioning availability, and parking space availability. Property owners should not increase their prices significantly based on such amenities. On the other hand, renters should consider renting or buying houses in areas that have shared parking spaces or have street or public parking to avoid paying high rents. Regarding business process changes, I would recommend a change in ad-hoc decision-making and a pivot towards data-driven decision-making, seeing as it provides superior insights. This analysis may bring about a change of about 200% in housing uptake as housing sales firms will better target customers according to the insights generated herein. Changes will include integrating data-driven insight generation into housing market decision-making to improve sales rates, which would mean employees would need to upskill themselves in data analysis related skills. However, organizational support will ensure this change proceeds smoothly to prevent resistance as well as ensure a positive transition.

These models have limitations including assumptions such as high variance, sensitivity to outliers, dependence on the data used for training, sensitivity to scaling and a tendency to overfit models. However careful data preprocessing and cleaning can eliminate most of these problems. Further, one may use a hybrid model to ensure accuracy where one combines multiple models to predict the target variable.

# References

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.

Butler, P. (2023). England’s Broken Housing System is Now A Problem No Council Can-Avoid https://www.theguardian.com/society/2023/oct/30/englands-broken-housing-system-is-now-a-problem-no-council-can-avoid#:~:text=High%20house%20prices%2C%20soaring%20rents%2C%20housing%20benefit%20cuts%2C%20a,have%20created%20a%20perfect%20storm.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8), 1477-1494.

Engerstam, S. (2023). Fundamental determinants of house prices in advanced economies: A systematic literature review.

Geerts, M., & De Weerdt, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. ISPRS International Journal of Geo-Information, 12(5), 200.

Li, X. (2022). Prediction and analysis of housing price based on the generalized linear regression model. Computational Intelligence and Neuroscience, 2022.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.

Simoneli, T. (2023). UK House Prices: 1990-2025 Power BI Data Analysis. <https://www.linkedin.com/pulse/uk-house-prices-1990-2025-power-bi-data-analysis-thiago-simoneli>

# Appendices

Appendix 1:



Appendix 2:

