1. **Introduction**

The

Real estate market has always been a popular topic of nationwide concern in UK. Trends in the market are not only of interest to buyers, homeowners analysts and policy makers, they also have potential contribution in economic development. Generally, house price may be determined by a variety of aspects, including the location, property type, investment prospects and supply and demand of the market etc. (Phan, 2018). Therefore, it is necessary to precisely forecast the future housing price using scientific approach for optimising the decision making process. However, there are a vast amount of house price data produced every year, making sense of this huge dataset is absolutely challenging. Since there are numerous factors drive the variation of house prices, which triggers the demand of a forecasting model that can interpret the relationship between the variables and prices of house. Machine Learning (ML) has paved the way in this research context. Such technique can build model to learn from dataset, optimise its parameters and derive data-driven prediction.

Normally ML algorithms are splitted into three categories: supervised, unsupervised and reinforcement. The context house price forecasting is covered by supervised learning and will be suitable for regression model, since the data contains input and output variables. The detailed description of data will be presented in Section of Data. England house transaction data will be set as a study case to explore the prediction model. The goal of this study is to analyze the historical property transaction data to build regression models, then compare the results to obtain the optimal one and find the most influential factors on housing prices.

The code chunk below contains an overview of packages required for the analysis process.

1. **Literature Review**

There are already related studies on housing price prediction. Louati et al. (2022) developed a set of machine learning algorithms to increase the effectiveness of the land price estimation in Riyadh city, which included decision tree, random forest and linear regression. The results indicated that random forest provided the best performance, while the other two had the similar results. Avanijaa and Al (2021) applied XGBoost regression to estimate house price associated with location, neighbourhood and infrastructure and suggested deep learning algorithms for boosting the prediction accuracy, although more data might be required. Phan (2018) presented that support vector machine with mean squared error measurement was a powerful approach when predicting Melbourne house price, while it was reletively difficult to intepret the model and spent more runtime in comparison with other regression including linear regression and regression tree etc. Adetunji et al. (2022) applied random forest model with Boston house price dataset, the predicted outcome had an acceptable error range comparing with actual price.

Based on the previous studies, the house price forecasting in England in this research will apply linear regression firstly to start test with a simple regression model, then more complicated machine learning method will be applied, including random forest and decision tree, to compare and obtain the a relatively better prediction model.

1. **Research question**

According to the investigated dataset and the research process, the research questions are presented: is it possible to predict house price based on property transaction data by using linear regression, random forest and decision tree models? Which regression model can provide better prediction? Which variables can significantly impact the England house price?

1. **Data**

This section will present the data source, data cleaning, and data investigating and pre-processing processes.

The house price data is obtained from UK Government website (GOV.UK., 2022), the most recent data in 2021 will be used for research since it could generate a forecast that would be closer to the present. The dataset contains England and Wales house transaction features which mainly include the unique id, sale price, date of transaction, location information, proper type, building age, tenure and type of Price Paid transaction. It will be cleaned and listed with details later in the subsection.

**4.1 Data Cleaning**

Firstly, before import the dataset, the column headers are added accroding to explanations of the website (GOV.UK., 2022). Then import the dataset and check the column data types.

It can be noticed that there are some postcodes missing in the summary table which should be excluded, since only postcode will be used among the location features which is for converting to coordinate data. And since only England data will be investigated, Wales data should be excluded as well.

Subsequently, from all the available features, the most relevant variables are identified for building the prediction model, and the rest are excluded as they are missing, duplicate or redundant. The variables listed below are those contribute to house price forecasting. It should be noticed that postcode is decided to represent property location rather than the other position information, which is because postcode is succinct and unique, and it will be convert to longitude and latittude in the later process.

* **Price**: sale price of the properties (*dependent variable*)
* **Date of Transfer**: date when the sale is finished, Year/Month/Day\_Time (*independent variable*)
* **Postcode**: postcode of th properties (*independent variable*)
* **Property type**: D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other (*independent variable*)
* **Building age**: the age of the property, Y = a newly built property, N = an established residential building (*independent variable*)
* **Tenure**: tenure of the properties, F = Freehold, L= Leasehold (*independent variable*)

Although the features have been refined, the data is still extremely large, which is about 900k rows. Using the entire dataset would be computationally challenging and significantly affect the program running and training time, and could even force stop the running program due to the limited memory of docker. Therefore, to improve the time and training efficiency, only 20 thousand sets of data are selected and would be enough for the model training and building. The data with na values will be excluded first, then the rows will be randomly selected to avoid data bias and inequity.

### 4.2 Data Investigating and Pre-processing

In order to apply the data for establishing prediction model, pre-processing is required.

Firstly, the transaction date should be converted from datetime data to numeric data. To simplify this, it is represented by the month of 2021.

Secondly, the postcode data is replaced by corresponding latitude and longitude data to capture the geographic differences in property price more effectively and conform with the reqirement of model building.

The coordinate data is obtained from Geonames.org. (2022). As the data is too large, it was manually simplified to have only postcode, latitude and longitude columns, and was splitted into two parts and stored in my github repo.

Ultimately, the rest binary categorical variables are converted to dummy variables. For building age, 0 represents 'N' (old), 1 represents 'Y' (new); for tenure, 0 represents 'L' (leasehold), 1 represents 'F' (freehold); for property type, each category creates a dummy variable respectively.

After data conversion, the summary statistics of both dependent variable and independent variables are displayed below.

Now check the NA values again among the dataset.

As there is no NA values, now histograms are used to investigate the distribution of all the variables.

The strange situation is noticed here: the price graph almost displays only one bar. The reason might be that most of the house price gathering at a lower section, the outliers are too tiny to view on such a limited distribution. Therefore, the plot method is changed and attempt log transform to compress the outliers and make a normal distribution.

The outcome is quite good after transformation. Therefore, the log price will be used to replace price as the dependent variable.

1. **Methodology**

This section will separate the dataset into train, validation and test data subsets, multicollinearity will be checked to investigate the correlation between variables. Then models using linear regression with VIF, random forest and decision tree with hyperparameter tuning will be presented respectively.

### 5.1 Set Train, Validation and Test Subsets

Firstly,the dataset is splitted into train, validation and test subsets according to 70-15-15 split. The split is randomly preocessed to avoid selection bias. Random state is set here to ensure reproducibility.

### 5.2 Multicollinearity Checking

Subsequently, multicollinearity checking is conducted here by using correlation matrix, in order to examine whether there are highly correlated variables. This method is only for viewing the rough correlation rather than removing variables. It can be seen from the matrix that property type D and S might have relatively higher correlation with tenure. Accurate calculation will be conducted by VIF method.

### 5.3 Linear Regression with VIF

For model establishing, it should be started with the simplest linear regression. In this subsection, the Variance Inflation Factor (VIF) method is firstly used to deal with the multicolinearity between the independent variables. The higher the VIF value, the more it needs to be removed. The code for using VIF method are below:

After removing the highly correlated variable, multiple linear regression can be conducted.

Then check the intercept value and coefficients.

Next, R2 score will be generated to explore the model performance.

It could be seen from the summary that R2 is 0.318, which is a quite low value. It indicates that only 31.8% of the variance in house price prediction can be explained by the model, and the model is not well fitted with high bias. In addition, as the 'Month' variable has a p-value higher than 0.05, it is not a significant variable. The others are significant.

The test and validation data should produce the same results.

It can be noticed that the three sets generate similiar values with slight difference. The linear regression with VIF has high bias for the house price prediction.

### 5.5 Random Forest

This section will establish a random forest model to fit the dataset.

It shows the R2 of test data is much lower than the training data, which reveals overfitting issue, however, it also shows lower bias than linear regression models.

As the RF model has too many trees and leaves, visualization will not be helpful. Instead, the Permutation Feature Importance (PFI) is introduced to explore the feature importance.

The top features are latitude, longitude and property type D. The model demonstrates these variables are the most important factors when predicting the house price.

#### Hyperparameter Optimization Using Cross-validation

As there is still an overfitting problem, cross validation is applied, as it is a robust method to tune the model hyperparameters through optimizing model performance with the development dataset.

The optimised score varies slightly with the number of trees. It shows that the model is robust enough. It looks like the training and development scores do not change significantly with the number of trees, which means that the model is quite robust to this number of trees in the random forest.

The variation between final model and the orginal one is slight as the model is robust enough.

**5.6 GBDT and XGBoost**

Ultimately, decision tree model will be built to prediction of house price.

The R2 of both data indicate slightly lower values than random forest but larger than linear regression models, and there is also a similiar value difference between training data and testing data as random forest. These implies the relatively higher variance and lower bias.

Next, PFI is also applied to investigate the feature importance here.

The order of the top features are the same as random forest, which also proves they are the critical variables.

#### Hyperparameter Optimization Using Holdout Validation

To optimise the model, holdout validation is applied instead of cross validation. Although the later one is robust, the program is cut off and restart due to the large data and docker memory limi. Therefore, holdout validation is used.

The score after improvement is better than the original one. Now check the final scores.

The final R2 values are quite close to that generated by random forest, which indicates the robustness of the model.

1. **Results and Discussion**

As linear regression, Random Forest and Decision Tree have all been built, the code below will collate the model results to compare the performance.

From the summary above, it can be investigated that linear regression model is not fitted for this prediction as the R2 is the smallest. Although the difference between training data and testing data is tiny which reveals low variance, the high bias results lead to a non-robust model. To improve its performance, more relevant features such as room type, house area and neighborhood infrastructure etc. should be involved for analysis.

The random forest and decision tree models are higher than the linear regression, and produced the similar results. The slight difference can be ignored since there is a variation among each time of running. The R2 values of training data are normally higher because of overfitting, and the larger R2 comparing with linear regression indicates relatively lower bias. however, since there is still a difference between training data and testing data, high variance is a problem. More training data can be added for improvement.

Another potential reason for the regression results is the outliers, which normally contains extremely small number of high or low values. The outliers of dependent variable are processed by log method to gain a normally distributed plot. However, outliers from independent variables were not considered in the data manipulation process. It could be checked through box plot and removed the abnormal values.

The most important variables impact England house price are gained: latitude, longitude and detached property. It could be understandable that location generally is the main driver of house price, the closer the area is to the convenience of living or economically developed areas, the higher the price tends to be. This is critical to forecast price in a certain district. Besides, detached house is the other important feature of price, this type of property tends to have more building area and living facilities, the target clients are mostly wealthy groups with certain assets. Therefore, there is a strong correlation between detached house and price variation.

There are also some limitations during the research. Firstly, the the number of variables is finite in the dataset, which might lead to high bias of the model. Secondly, although more training data can reduce model variance, the dataset is too large and exceeds the docker memory limit, resulting in my notebook died after attempt running. Therefore, the cross validation cannot apply twice, and nueral network cannot apply to this prediction program as insufficient dataset. Subsequently, the property price variation is relevant to the economic trends and has a significant time-varying. Appropriate prediction may need more robust data and more complex deep learning models to train the data

1. **Conclusion**

To conclude, housing price involves multiple interests and it is necessary to accurately predict the value. There are various factors affecting the price tendency, several studies have already applied machine learning algorithms to predict house price. This research aims to explore an appropriate regression model for England house price prediction and find out the impact factors. The research uses historical property transaction data in 2021 for manipulation. The data is cleaned and pre-processed with only valuable variables and limited dataset left, then the data is split into training, validation and testing datasets for model fitting. There are three regression models being applied: linear regression with VIF, random foreest and decision tree. The results indicate that linear regression is bad fitted for the context, while random forest and decision tree provide same prediction with relatively high variance after hyperparameter tuning. Improvement is needed to add more training data. The most important factors of house price investigated from the research are latitude, longitude and detached property.There are also some limitations of the research is presented, including finite dataset, limited features, docker memory limit and limited regression models.

**Bibliography**

*Price Paid Data* (no date) *GOV.UK*. Available at: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads (Accessed: 17 April 2022).