London Housing Market Investment Recommendations

Rufei Wang

# Executive Summary

Building an estimation engine to help guide investment decisions on London properties.

* Explored, built, and compared various models. Combined the optimal ones to achieve a stacking model that performs better than any individual model performs on its own.
* The stacking model is the best-performing algorithm, due to its highest explanatory power.
* The model can be recreated, tuned, and improved by other data scientists for different usages, such as adding new features to estimate the effect Elizabeth Line has on London housing price.
* The model can also be used to help make investment decisions in the London property market.
* 200 property investment recommendations were made using the model, which can be accessed via the [link](https://github.com/wrufei/AM04_Individual_Assignment).

## 1. Introduction

London housing price has been on a consistently rising trend in the past 20 years, rising well above the general inflation level, despite two major drops in between 1990 & 1992 and 2007 & 2010. The depreciation of Pound Sterling had attracted foreign investors into the UK property market, which could boost up demand and in turn result in a further rise in housing price. In London in particular,

the opening of Elizabeth Line can also potentially affect the housing price. Further, there are several government initiatives implemented to slow down the rise in property prices.

Given these conflicting forces at play in the London property market, I endeavoured to construct an estimation engine to help guide investment decisions in this market and give recommendations using the findings from data exploration in combination with the results given by the estimation algorithms.

I would be using the HM Land Registry’s Price Paid Data, Energy Performance Certificate (EPC) database and the public transport information to do data exploration and algorithms training.

I would firstly explain my findings from the initial exploratory data analysis (EDA), then I would elaborate on the models that I built. Next, I would explain the ensemble methods I chose and used. I would also investigate into how the estimation engine can be used to help make investment decisions in neighborhoods that will be served by the Elizabeth Line. Finally, I will conclude by giving recommendations and reflecting on limitations.

The report is structured in the following way:

- Exploratory Data Analysis

- Individual Models (LM, Lasso, Tree)

- Ensemble Method (Random Forest and Stacking)

- Elizabeth Line

- Recommendations & Limitations

- References

## 2. Exploratory Data Analysis

To explore and examine the data, I would draw scatter plots to investigate correlations, line graphs to investigate the change of pricing over time, histograms to explore the distributions of different prices and ggcorr table to investigate correlations among different variable pairs of interest.

Firstly, I explored how housing prices changed by different types of properties. We can see that there had been fluctuations among all types of properties. Both the mean and median prices fluctuated the most for detached properties.

Chart

Description automatically generated

Chart

Description automatically generated

Next, I explored the distribution of the housing prices using histogram. It suggests that the price distribution was negatively skewed. There were fewer properties that were in the extremely expensive price category.

Chart, histogram

Description automatically generated

Next, I drew the ggpcorr table to investigate correlations among different continuous variables. We can learn how strong the correlations are between different continuous variable pairs. For the purpose of modeling, we will be looking at the variables correlations with price. From the table, I found a few strong correlations with price, which are interesting to investigate further into. Positive correlations in order of strength: total\_floor\_area, C02\_emission\_current, C02\_emission\_potential, number\_habitable\_rooms, and average\_income. Negative correlations in order of strength: london\_zone and distance\_to\_station. We also need to keep in mind the other variable pairs that have strong correlations (excluding the ones I listed) such as C02\_emission\_potential and energy\_consumption\_potential because we would want to avoid including two strongly correlated variables in our linear regression building later on to avoid multicollinearity.

Chart, scatter chart

Description automatically generated

Next, I explored further the correlation between total\_floor\_area and price because that had the highest correlation from the correlation table we drew previously. I also included the scatterplot of number of habitable rooms in the same graph. There are strong correlations between total floor area and price as well as number of habitable rooms and price. Also notably, the number of habitable rooms and total floor area variables are also strongly positively correlated. We want to include only one of them as features for our model to avoid multicollinearity.

Chart, scatter chart

Description automatically generated

Finally, I want to explore the correlation between London zones and price as well as distance to station and price respectively. I learnt that there is weak negative correlation for both pairs. Prices are higher in central London and the closer the properties are to the station, the higher the price tend to be.

Chart, line chart, box and whisker chart

Description automatically generated

Chart

Description automatically generated

## 3. Individual Models

I built three models, a simple linear regression model, a Lasso regression model, and a tree model. I split the data into testing and training datasets. I compared the three predict functions to test the performance of the model in testing data.

## 3.1 Linear Model (model2\_lm)

Graphical user interface, text, application, email

Description automatically generated

In this model, I included 13 variables as explanatory variables. To find out which variables to include, I used the ggcorr table I previously constructed and selected the ones that were strongly correlated with price first. I then applied log for the variables that are highly skewed, for example London\_zone. I checked whether that would increase the R-squared and it did. I then applied inverse square for average\_income and co2\_emission\_current because they have stronger effect on prices. Next, I included the weaker correlated variables such as property\_types and tenure etc. To arrive at this model, I also tried different combinations of features and I used the one that had the highest R-squared. This is because higher R-squared essentially means that the explanatory variables in that model have higher explaining power for price (the variable we are interested in). I also excluded variables such as energy\_consumption\_current and number\_habitable\_rooms in my model to avoid multicollinearity.

I used the predict function to test the performance of the model in testing data. The performance and quality of model2\_lm model improved significantly from the default OLS model. R-squared increased from 16% to 69%.

## 3.2 Lasso Regression (model4\_lasso)

Next, I constructed a Lasso regression model. Firstly, to determine the optimal lamba, I performed k-fold cross-validation and identify the lambda value that produces the lowest test mean squared error (MSE). We set alpha equals 0, which is suggesting Lasso regression. I then used the final lasso regression model to make predictions on prices.

It is interesting to note that the performance of this algorithm is poorer than both the model2\_lm and model3\_tree. The R square is lowered at 60%. Although this is still performing much better than the original model model1\_lm.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## 3.3 Tree Model (model3\_tree)

I fitted the tree model using the same subset of features I used for model2\_lm. To compare the performance of the linear regression model with the tree model, the tree model performs better, improving R square from 68% to 74%. This is because the automatic tuning feature of tree model allows for non-linearity correlations whereas the linear model I used only accounts for linear correlations.

Graphical user interface, text, application

Description automatically generated

## 4. Ensemble Method

The reason why we want to use ensemble methods is that they allow us to combine multiple learning algorithms to obtain better predictive performance. Here I used two ensemble methods: random forest tree and stacking.

## 4.1 Random Forest

The reason why I chose to use random forest is that it overcomes the potential problems with ‘bagging’ which is that it tends to generate too much correlation. Random Forest doesn’t use similar data and the same set of predictors in each iteration. Further, it can be used improve the tree model (model3\_tree). The common problem of using a simple tree model is over-fitting. Using random forest tree is a good alternative. Random forest tree model allows using a sub-sample and a subset of features each and every time, to avoid dominate features.

I explored mtry = 5,10 and found that 5 is the optimal parameter I should use. The RMSE and R squared evidenced an improvement from the previous simple tree model. With regards to the feature importance, it is very similar to what we found in the simple tree model.A picture containing text

Description automatically generated

## 4.2 Stacking

Using the stacking method, I combined model2\_lm, model4\_lasso and model5\_rft. From the result, we can see that stacking performs than all the other algorithms perform on their own. The RMSE is reduced to 217500 and R-squared improved to 85.85%.

Graphical user interface, text, application, email

Description automatically generated

## 5. Elizabeth Line

Assuming that we can measure the distance of each property to Crossrail. We can add in this feature to our existing model. If this addition improves the predicting power of the model (higher R-squared), then we can use the updated algorithms to make better informed property investment decisions, considering the impacts of Crossrailson property price.

However, it is key to note that the effects of Crossrail may not be significant on property price. That could be because most of the Elizabeth Line stations are quite far out of the central London and the impact it has on property price is minimal. In order to isolate the effect of Crossrail on property price in London, we need to consider controlling for other variables, such as london zone.

## 6. Conclusions & Limitations

Having explored and compared the 6 models (model1\_lm, model2\_lm, model3\_tree, model4\_lasso, model5\_rft and lm\_ensemble). We conclude that using the stacking method (lm\_ensemble) gives the best prediction performance and quality, with the RMSE reduced to 217500 and R-squared improved to 85.85%. This estimation engine can be used by investors as reference and additional guidance. For instance, it can be tuned with new features to estimate the effect Elizabeth Line has on London housing price. Being created in a rmd. File, the model can also be recreated, tuned, and improved by other data scientists for different usages.

There are limitations with regards to this lm\_ensemble model using stacking. The first one being that it could be complex to understand, especially for readers who are not data experts. In this project, I explored 5 models, however, in a real-world scenario, more models should be built and compared and that will add on complexity to interpret. If the model idea cannot be understood by the reader/users/business managers, then it is very unlikely to be applied. The second limitation is with regards to time and computing power consumption. Ensemble methods such as stacking costs more to create, train, and deploy. The ROI of an ensemble approach should be considered carefully in a real word business case.

## 7. References

Data Sources:

[HM Land Registry’s Price Paid Data](https://www.gov.uk/government/collections/price-paid-data);

Energy Performance Certificate (EPC);

[Data Dictionary](https://learning.london.edu/courses/8133/files/folder/Course%20Materials/Session%2010/Assessment?preview=2697575)

GitHub Link: <https://github.com/wrufei/AM04_Individual_Assignment>