Predicting House Prices

Vivian van Oosten

# Abstract

House prices are determined through many factors related to the house characteristics itself, the location and the economy as a whole. Because house prices are expected to increase, they make a relatively safe investment. In this paper, we aim to predict the price of a set of houses in London so as to invest in the ones that are prices below the price we would expect. We analyse the houses sold in London in 2019, combining information on the house, the location and the energy efficiency. Several machine learning models are combined in a final model to most accurately predict the house prices of the houses we are considering investing in. Comparing our predicted price with the asking price, we select the 200 houses for which we would expect the highest percentage return on investment. Based on our model, we expect a return of 64%.

# Introduction

We have been provided 2000 houses in London as options to invest in. The goal of this report, is the recommend a specific 200 houses to invest in. We will determine the price we would expect the house to be sold for based on past houses. The sales data in London from 2019, as found on the HM Land Registry’s Price Paid Data (https://www.gov.uk/government/collections/price-paid-data), will be used to train a Machine Learning model to predict the price of a house based on its characteristics. We have combined this dataset with the Energy Performance Certificate data, which contains further information on the size and energy efficiency of the houses in our dataset. Finally, we add more information on the postcode our houses reside in. The predicted house prices are taken as an indication of what the market would likely pay for the house. If that is higher than the current asking price, the investment is expected to be profitable. The rate of return is calculated with the predicted price being the return and the asking price the initial investment. The top 200 houses are recommended for investment.

# Data exploration

We explore the data using various plotting methods, of which the most interesting can be found in this report. First, we assess the distribution of house prices. We see that we have prices from 77 thousand up to 108 million pounds. To clarify the visualization and improve our predictions, we take the logarithm of the price instead. In the histogram, we see that the house price distribution is skewed to the right.

Second, we consider square footage as an important consideration in house prices.

Third, the location is taken into account.

Lastly, we consider the correlation matrix to determine the variables we will be considering in our machine learning model.

Graphical user interface, application, table, Excel

Description automatically generatedWith the dataset in the appropriate format, we set out to understand the distribution of the data with the goal of identifying any outliers. We plot the distribution of total time spent watching (Figure 1). The graph clearly identifies that the distribution is right skewed, which will affect our choice of the clustering method in subsequent parts of the project.

**Fig. 1: Distribution of customers by total watching time**

# Feature selection

We ran a simple linear regression on all of our relevant variables and started our feature selection process with the top 15 variables. As some such as district are categorical variables, we have included the variable if at least 1 category fell in this list.

* Average income
* London zone
* Districts
* Total floor area
* Windows energy effiency
* Property type
* Number of habitable rooms
* CO2 emissions potential
* Freehold or leasehold
* Water company
* Energy consumption potential
* Distance to station
* Altitude
* Tenure
* Current energy rating

Then, we further narrow down the selection.

The geographical variables (zone and latitude/longitude) can be used to estimate centrality of the property, which affects the price. The total floor area, type of property and current CO2 emissions provide information on the quality of the house itself, while the average income gives information on the specific neighborhood. The distance to station and number of tube lines show the connectivity. Interestingly, the water company also seems to have a relationship with the house price.

# Prediction model

We now turn to performing predictive machine learning on our dataset. We have used four distinct methods: a linear regression, a simple tree, a lasso regression and k nearest neighbors.

* **Linear regression**: Predicting prices based on the ordinary least squares algorithm
* **Tree**: Predicting prices by iteratively determining the best way to split the dataset into two and
* **Lasso regression**: Predicting prices with a linear regression combined with a penalty term for the size of the weights
* **K-NN**: Predicting prices by finding the closest houses (not geographically, but in terms of all variables), and averaging their price

INSERT TABLE TO WITH RMSE/R2 TO COMPARE THE DIFFERENT MODELS AGAINST EACH OTHER

### Linear Regression

We train a simple linear regression to test our more complicated future models against.

### Selecting best investments

Having aggregated and combined our final model, we have chosen a K-NN model based on the RMSE and R2 of our models. The final stacked model had a worse RMSE and R2 than our K-NN model.

### Elizabeth Line

Our current model uses distance from station as one of the variables to determine the price of a property. With our linear and lasso regression models, we can use the weights to interpret the effect. We notice there is a negative relationship, so if the distance to station decreases, the price should increase. This means that adding more stations, as the Elizabeth line is doing, should decrease the price of properties next to those stations. We need to use our model to predict the price with the current distance to station and with the distance to station when the Elizabeth line is added to the tube-network.

Then, using a similar method as we used for the selection of our top 200 investments, we compare the price including the Elizabeth line to the current asking price of the houses. If the asking price falls outside of the confidence interval of our expected price, we can assume that there is space for investment. Specifically, given a parametric model such as a linear or lasso regression, we could estimate how much closer the nearest station would be after the Elizabeth line opens and multiply that by the weight for the distance to station. To properly assess the effect, we would also need to assess the impact on the number of tube lines variable. Given these are new stations, the number would now be 1, so we need to estimate the difference from before and either add or subtract that effect from the distance effect.

# Discussion

A 67% return on investment is quite significant, given an expected return of 8% when investing in the S&P 500. Unfortunately, our model still has an RMSE of 250 thousand, which is only a slightly less than our average return of 279710 pounds per house in our chosen 200 houses. We can still make decisions based on the predictions, but we need to keep in mind that this error is relatively large.

Part of the reason why our error is relatively high is due to the high skewed-ness of the data. Prices are distributed between (XXX) and (XXX), which we could simplify by using the logarithm of the prices. However, using the logarithm would reduce the interpretability of the weights and the error. It would also increase the error for higher priced properties, since a difference in the logarithm of 0.24 (the RMSE for our final model in the case of logarithmic prices) would be more for higher prices. When higher initial investments (higher prices) have a higher error, we also increase the risk we would be taking with our investment. Depending on the preferences of the investor, we could choose to go with a riskier approach. For the purposes of this investment, we assume a certain risk-averseness, especially for larger investments, and choose not to use the logarithm of the prices.

Our model is also limited by the data we have. While house-prices are certainly partly determined by the characteristics of the house, other variables will have an effect as well. Firstly, the stock market influences house prices. It can be interpreted as an alternative investment to buying real estate, and it shows the current state of the economy. In a recession, both house prices and stock prices would fall, making the stock price a useful proxy for economic conditions. Secondly, the interest rates determine the demand for housing. Interest rates are the price of borrowing money, and therefore an added cost for anyone wanting to buy a home as an investment or a living space. Lastly, inflation in general will affect house prices. A higher inflation is generally either driven by a house price increase or will cause it, since housing is a large percentage of the money we spend.

While all the mentioned variables will influence the overall level of the house prices, individual houses bought at the exact same time can still largely be distinguished from each other through individual characteristics because the macro-economic status of the market will be the same at the time of purchase. Only if we are considering buying real estate now versus later do we need to consider variables such as that, and in that case, we do not need to know the individual properties of the houses. Instead, we can focus on aggregate measures.

Finally, our investment choices were in the end made solely based on the percentage return they are projected to offer. While that is an important element of any investment, there are more considerations to be made. A time-element would need to be introduced to properly assess the investments. When would we expect to sell each house and how does that change the current time-value of the investment? The investment size is also not capped using this method, we have simply selected the top 200 properties to invest in without considering the size of the investment. If funds are limited, we would need to make a different selection. This investment choice also does not take into account the sustainability of properties. Investing in houses with a better energy label could improve the rating provided by ESG rating agencies, which could be vital to the reputation of our client. In general, it would be important to have a conversation with our investors to determine their exact requirements and goals with this investment. This could further inform the fine-tuning of the model and the recommendations.

Figure 6 compares the size of clusters obtained using different clustering method. It is visible that all algorithms produce clusters of similar size. Figure 7 shows the differences between cluster centres

# Conclusion

Based on the RMSE and R2, we have selected the K Nearest Neighbors algorithm as our final model. Due to the local nature of the algorithm, it will better predict the tail of high priced properties present in our data than a linear model. The algorithm also mimics the way realtors and sellers determine the price for their new property: by looking at similar properties sold recently.

SUMMARY OF THE RESULTS ? + recommendations for the future