

# Analysis of London House Prices

Zervaan Borok

---

## Table of Contents

<b>Abstract .....</b>	<b>2</b>
<b>Introduction .....</b>	<b>2</b>
<b>Exploratory Analysis.....</b>	<b>2</b>
Price .....	3
Distance to Nearest Station .....	4
Number of Habitable Rooms .....	5
Total Floor Area.....	6
Correlation Plot.....	6
<b>Model Construction .....</b>	<b>7</b>
LASSO.....	7
Train/Test Split .....	8
Linear Model .....	8
Machine Learning Models .....	8
Tuning Parameters .....	9
Ensemble Model .....	9
<b>Model Evaluation.....</b>	<b>10</b>
Model Performance & Sensitivity Analysis.....	10
Model Selection .....	10
<b>Investment Allocations .....</b>	<b>11</b>
<b>The Elizabeth Line.....</b>	<b>11</b>
<b>Conclusion &amp; Remarks .....</b>	<b>11</b>
Variable Importance .....	12
Assumptions .....	13
Comments .....	13
<b>Reference List .....</b>	<b>15</b>
<b>Appendix 1 .....</b>	<b>17</b>

# Abstract

This report aims to assess the predictive power of various machine learning algorithms on London House Price data and to ultimately use the best model to select 200 properties to invest in. Two data sets were used in this analysis: a training data set consisting of 13,998 observations of 37 variables and an out of sample data set consisting of 1,999 observations of 37 variables. Five standalone models were constructed for this analysis and are as follows:

- Linear Regression (LR)
- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Random Forest (RF)
- Gradient Boosting Machine (GBM)

A sixth ensemble model was also created by stacking the Linear Regression, K-Nearest Neighbors, Random Forest, and Gradient Boosting Machine models. The root-mean-squared-error (RMSE) and R-squared ( $R^2$ ) metrics were used to evaluate model performance. The results of the analysis that follows indicate that the ensemble model has the greatest predictive power on London House Price data.

# Introduction

Properties, whether commercial or private, have historically been one of the most widely used physical assets for investment purposes. This is partly because they are relatively safe investments. For example, if a company goes bankrupt, their stock price will fall to zero, leaving investors with absolutely nothing. On the other hand, while real estate is much less liquid and often entails a much higher capital investment than shares of a firm, it is far more unlikely that the value of a property will fall to zero. By default, some cities are much more attractive for real estate investment than others, and London happens to be one such city. This report will investigate the city of London through the lens of real estate investment. Ultimately, a model will be constructed with the aim of selecting the 200 ‘best’ properties to invest in from an out of sample data set.

# Exploratory Analysis

The data used in this report had already been cleaned and so no further cleaning was required prior to starting analysis. To begin, a plot of the distribution of the number of properties by London zone was created. Table 1, shown below, summarizes the results of this plot.

**Table 1**

London Zone ♦	Number of Properties ♦
1	569
2	2502
3	3411
4	2922
5	2458
6	2130
7	6

The table shows that zone 7 only has 6 observations in this data set, so price predictions for this particular subgroup will likely be less accurate than those of the other zones.

## Price

Following this, boxplots of the house prices in London were constructed for each London Zone as shown below in Figure 1.

**Figure 1**



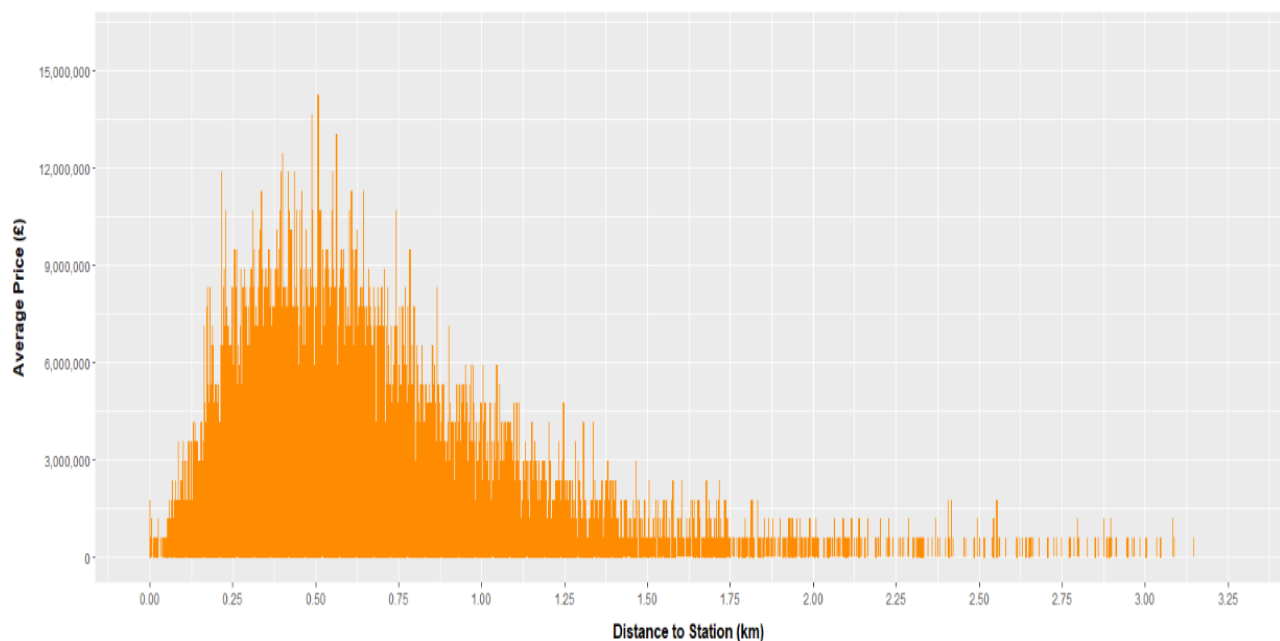
These boxplots illustrate that zones 4, 5, and 6 have roughly the same range of prices. Zone 3 has a larger price range than zone 4; zone 2 has a larger price range than zone 3; and zone 1 has the largest price range of all. To ascertain how the prices in each London zone are distributed, density plots of the prices were constructed for each zone. The density plots revealed that the prices in all London zones are heavily right skewed (Appendix 1, Figure 1).

## Distance to Nearest Station

Next, a plot of average property price by distance to the nearest train station was created and is shown below in Figure 2.

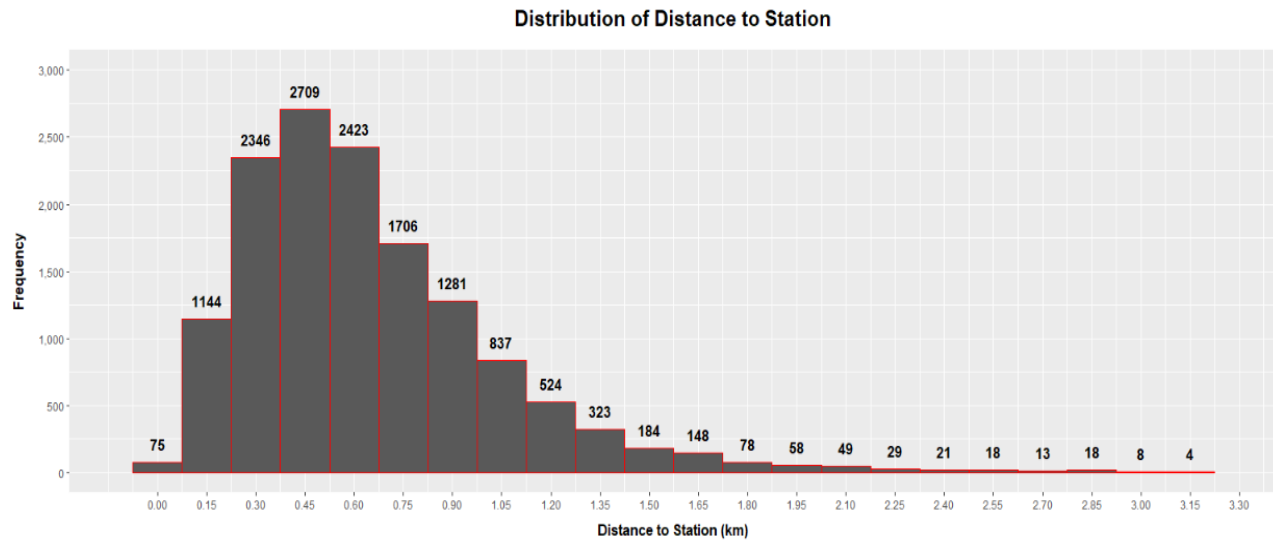
**Figure 2**

**Distribution of Average House Price by Distance to Tube Station**



The plot illustrates that the highest average prices align with properties located within the range of 0.2 to 0.75 miles from the nearest train station. To further investigate the 'distance to station' variable, a histogram detailing the number of properties per distance segment was created and shown below in Figure 3.

**Figure 3**



This histogram clearly illustrates that the vast majority of properties are located within 1.5 miles of the nearest train station and that this data is right skewed.

## Number of Habitable Rooms

Naturally, one expects the number of rooms a property to have a significant impact on price, and so a histogram detailing the frequency of properties by the number of habitable rooms was generated. Table 2, below, summarizes the results of this exercise.

**Table 2**

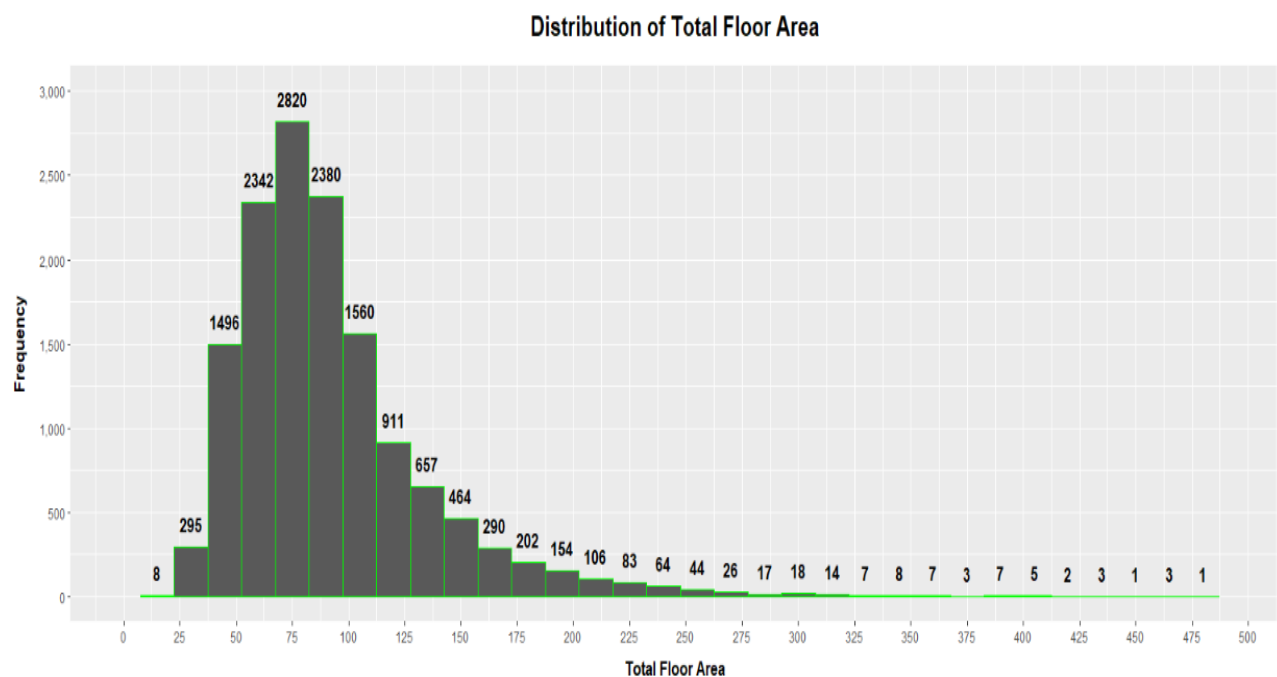
Number of Habitable Rooms	Number of Properties
1	134
2	1476
3	3486
4	2988
5	3007
6	1528
7	816
8	353
9	118
10	56
11	18
12	15
13	2
14	1

From the table, one can calculate that roughly 89.19% of properties within this data set have between two and six habitable rooms, and so this distribution is also right skewed.

## Total Floor Area

The final variable selected for visualization was the 'total floor area' as again, this characteristic ought to have a significant impact on property price. A histogram detailing the number of properties per total floor area segment was generated and is shown below in Figure 4.

**Figure 4**

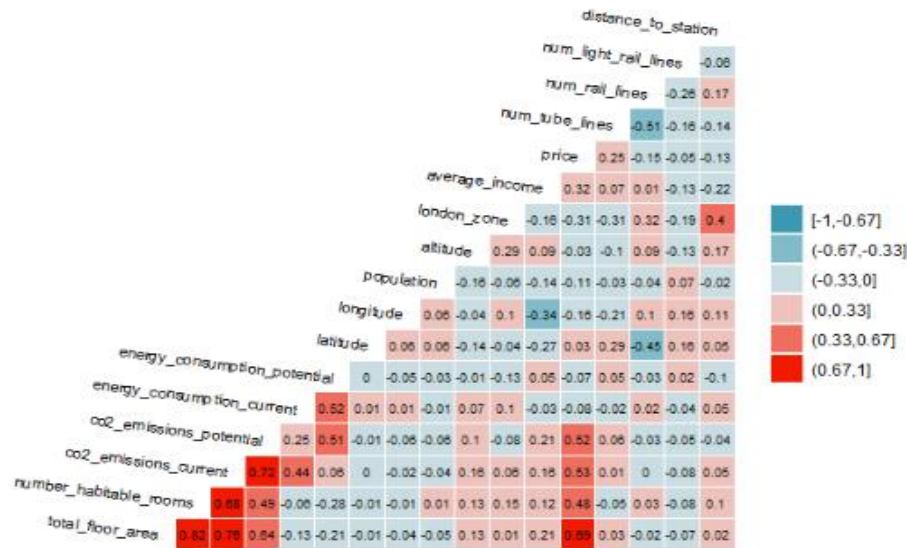


Again, it is evident from the plot that this data is quite heavily right skewed and that the vast majority of properties have a total floor area of less than 200 square meters.

## Correlation Plot

Lastly, a correlation plot of the numeric explanatory variables was constructed and is illustrated below in Figure 5.

Figure 5



The results of the correlation matrix are rather encouraging as for the most part, the variables within this data set are mildly to moderately correlated. A few are strongly correlated, such as 'number habitable rooms' and 'total floor area', but there are no correlations with an absolute correlation coefficient magnitude greater than 0.85. This means there is a lower risk of over-fitting any given model, which results in greater model accuracy on unseen data. The following section discusses the construction of the machine learning algorithms that were used to model this data set.

## Model Construction

### LASSO

Instead of repeatedly re-running a linear or logistic regression to determine which explanatory variables are most useful in predicting the response variable, one can use LASSO (least absolute shrinkage and selection operator) regression. In R, the LASSO function is designed to be fed the response variable and all explanatory variables which *might* be useful for prediction. The LASSO regression will then shrink the coefficient of any given explanatory variable to 0 if said variable does not add to the predictive power of the model<sup>1</sup> (Glen, 2015). The response variable fed into the model used for this report was the logarithm of the property price as the distribution of property prices within this data set is far from normal. The results from the LASSO regression showed that none of the explanatory variables' coefficients had been shrunk to 0. Thus, variable selection was conducted via a linear regression (LR).

## Train/Test Split

In this analysis, two separate data sets were used: the ‘training data set’ and the ‘out of sample data set’. Before training any models intended to be used for prediction, the ‘training data set’ was split into training and testing sets. The training set was allocated 75% of the observations in the ‘training data set’ while the testing set was allocated 25% of the observations in the ‘training data set’. Thus, in the analysis that follows, each model was trained on the training set and then tested on both the testing set and the ‘out of sample data set’.

## Linear Model

The initial linear regression was constructed using the logarithm of the property price as the response variable and all available explanatory variables as predictors. By using the p-values of the resulting coefficient estimates, the full model was reduced to one which contained the following explanatory variables:

**Table 3**

Variable <chr>	Definition <chr>
distance to station	The distance in kilometers to the nearest station from the postcode
property type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other
longitude	Longitude of centroid of the Postcode in decimal format
total floor area	Total of all enclosed spaces measured to the internal face of the external walls in square meters
number habitable rooms	Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar
london zone	Transport for London (TFL) Travel Zone Indicator
type of closest station	The type of closest train station (i.e. tube, DLR, or overground)
num tube lines	Number of tube lines that use the closest station
num light rail lines	Number of light rail lines that use the closest station
average income	Average household income of the MSOA that the postcode is located in
district	The district the property is in (i.e. Kensington, Chelsea, etc)
windows energy eff	Energy efficiency rating of windows with values of very poor, poor, average, good, very good
co2 emissions potential	Estimated value in Tonnes per Year of the total CO2 emissions produced by the property in 12 month period
co2 emissions current	CO2 emissions per year in tonnes/year
energy consumption potential	Estimated potential total energy consumption for the property in a 12 month in kilowatt hours per square meter

## Machine Learning Models

Four other models were constructed in addition to the linear regression:

- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Random Forest (RF)
- Gradient Boosting Machine (GBM)

Each of these models was passed the same list of explanatory variables used in the linear model (Table 3). Additionally, all of these models, except the decision tree model, also used the logarithm of the property price as the response variable. The input data was also centered and scaled (i.e. normalized) before being passed to



each of these four models. In general, logarithmic transformations and normalization techniques are both appropriate to implement when working with data that deviates significantly from the Normal Distribution. These transformations are merely linear and so no loss of information is incurred by utilizing them. Additionally, cross validation with 10 data points was used in the training process of each of these models.

## Tuning Parameters

The only available tuning parameter for the LR model was the ‘intercept’, which was set to “True”. The available tuning parameter for the DT and KNN models was the same, but the corresponding values were different. The tuning parameter in question was the ‘tune length’, which tells the train function how many different parameter values it should try. For the DT model, the optimal tune length was 35. The optimal tune length for the KNN model was 10 with  $k = 7$  (optimal number of clusters). The RF model also had the ‘tune length’ parameter available and its optimal value under this model was also 10. Additionally, the RF model also had a tuning parameter called ‘importance’, which is the method used to determine variable importance. The chosen value for this tuning parameter was ‘permutation’, which means the function will leave one variable out and fit the model again. Three other tuning parameters were also available for the RF model and are defined below in Table 4.

**Table 4**

Parameter <chr>	Definition <chr>
mtry	Number of randomly chosen variables to do a split each time
splitrule	Purity measure
minimum.node.size	Minimum size allowed for a leaf

The optimal values for ‘mtry’, ‘splitrule’, and ‘minimum node size’ were 29, “extra trees”, and 5, respectively. Finally, the GBM model also had the ‘tune length’ parameter available and its optimal value was 35 under this model. There were four other tuning parameters available as defined in table 5 below.

**Table 5**

Parameter <chr>	Definition <chr>
n.trees	Number of iterations
interaction.depth	Complexity of tree
shrinkage	Learning rate, i.e. how quickly the algorithm adapts
n.minobsinnode	The minimum number of training set samples in a node to stop splitting

The optimal values of ‘n.trees’, ‘interaction.depth’, ‘shrinkage’, and ‘n.minobsinnode’ were found to be 150, 6, 0.075, and 10, respectively.

## Ensemble Model

In addition to the standalone models described above, an ensemble model combining KNN, RF, LR, and GBM was also constructed. Initially, this stacked model contained all five standalone models; however, the model summary indicated that the p-value associated with the DT model was not statistically significant, and so it was dropped from the stacked model. The respective optimal tuning parameter values for each of these

four individual models were included in the ensembled model. The stacked model was trained on the same training data set that the standalone models were trained on. Stacking multiple models often results in greater predictive power, but it should be noted that this is not always the case. The next section is concerned with evaluating the predictive power of the five standalone models and the ensemble model.

## Model Evaluation

### Model Performance & Sensitivity Analysis

Model performance and sensitivity analysis were evaluated using both the testing data set and out of sample data set. Two metrics were used to assess how well each model performed: Root Mean Square Error (RMSE) and R-squared ( $R^2$ ). These metrics are defined as follows:

- RMSE: Reports the average distance between the values predicted by the model and the actual values contained in the data set<sup>2</sup> (Bobbitt, 2021).
- $R^2$ : Represents the proportion of variation in the predictor variable that is accounted for by the explanatory variables<sup>3</sup> (Fernando, 2021).

Table 6, shown below, illustrates the performance of each of the models on both the testing data set and out of sample data set. In the table, the columns 'RSME\_Test' and 'Rsquare\_Test' contain the values for these two metrics based on the testing data set. The columns 'RSME\_OOS' and 'Rsquare\_OOS' contain the values for these two metrics based on the out of sample data set. Lastly, the columns 'Average\_RSME' and 'Average\_Rsquare' contain the average values for these two metrics across the testing data set and out of sample data set.

Table 6

Model	RMSE_Test	Rsquare_Test	RMSE_OOS	Rsquare_OOS	Average_RMSE	Average_Rsquare
Linear Regression	273054.3	0.7554537	277399.0	0.8365919	275226.6	0.7960228
Decision Tree	242886.2	0.7963774	353176.9	0.7570202	298031.6	0.7766988
K Nearest Neighbor	273568.0	0.7942476	337737.7	0.8019112	305652.9	0.7980794
Random Forest	213409.5	0.8740774	338038.9	0.8021586	275724.2	0.8381180
Gradient Boosting Machine	214256.9	0.8557589	318024.4	0.8161633	266140.6	0.8359611
Stacked Model	191632.7	0.8816287	295707.8	0.8336424	243670.2	0.8576356

### Model Selection

When using RMSE as a metric with which to compare models, the model with the **lowest** RMSE is considered the best model. On the other hand, when using  $R^2$  as a metric for model comparison, the model with the **highest**  $R^2$  is considered the best model. Therefore, out of the models listed in the table above, the best model is the one with the lowest average RMSE and highest average  $R^2$ . Thus, the best model is the ensemble model, and this was the model used to select the properties to invest in.

## Investment Allocations

In the training and testing data sets, which are mutually exclusive subsets of the original training data set, the actual price of the property is given. However, in the out of sample data set, only the asking price is given. The objective of this report was to utilize the highest performing model to predict the actual price of the properties in the out of sample data set. If the actual value of the property (i.e., the value the model is trying to predict) is greater than the asking price, then an investor will get a positive return on investment (ROI). More specifically, the task was to identify the top 200 properties to invest in out of 1,999 total properties in the out of sample data set. These top 200 properties correspond to the 200 properties with the highest ROI. The formula used for calculating the ROI is as follows:

$$ROI = \frac{(\text{actual price} - \text{asking price})}{\text{asking price}}$$

Using the ensemble model, the expected total raw profit from these 200 selected properties was equal to £48,994,725.48 and the average expected ROI across all 200 selected properties was 63.64%.

## The Elizabeth Line

From the exploratory analysis conducted at the beginning of this report, it is evident that properties close in proximity to the nearest train station are more expensive than those that are not. In theory, one could pass the ensemble model a data set containing the asking price of the Crossrail properties and the same explanatory variables listed in Table 3. However, before passing this new data set to the model, the column corresponding to 'distance to station' would need to be updated to reflect the distances to the new Crossrail stations instead of the current ones. Once this data has been passed to the model, it will generate predicted prices for each property using the updated values for the 'distance to station' variable. Using the same ROI formula given in the 'Investment Allocations' section, one could then obtain the expected ROI for each property based on the difference of the predicted price and the current asking price. Alternatively, if the actual current value of each property is available, one could use this value instead of the asking price value to calculate expected ROI.

## Conclusion & Remarks

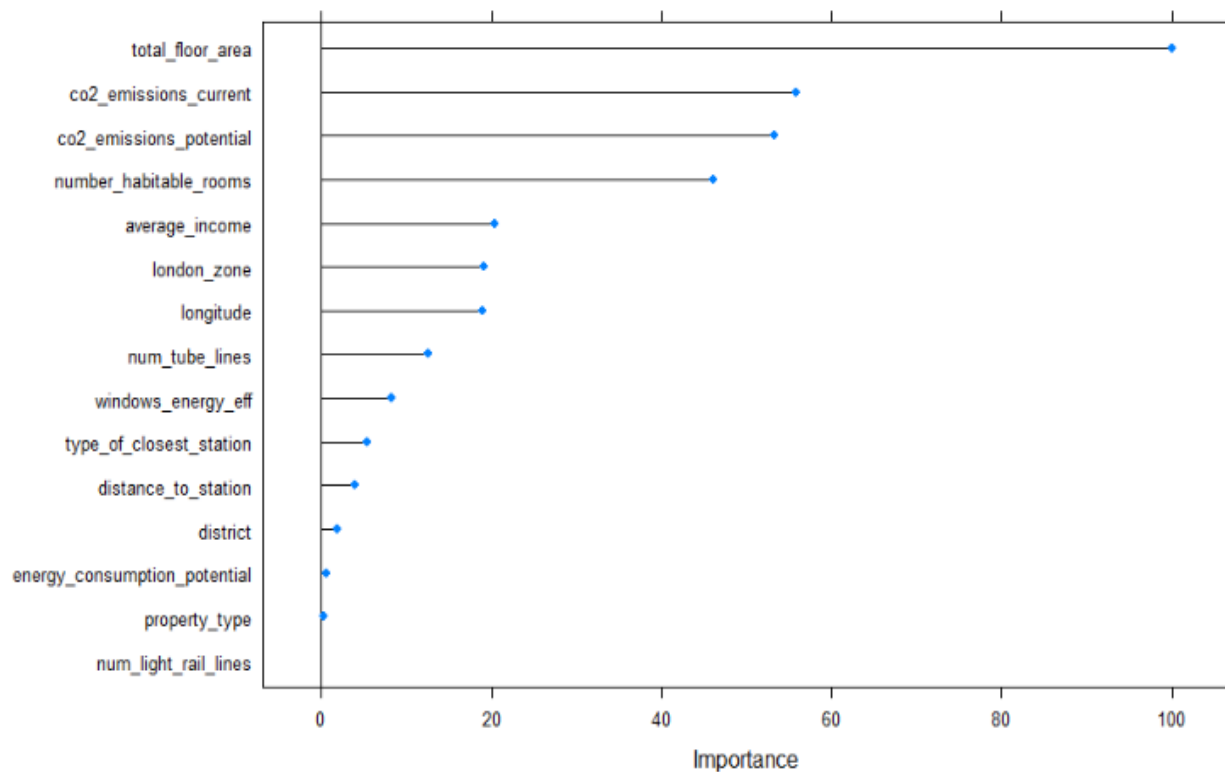
The results of this analysis indicate that out of the models explored in this report, the ensemble model performs the best on these data sets. This is not surprising as the inherent purpose of model stacking is to increase overall predictive power on unseen data. While the predictive power of the ensemble model is quite good according to the evaluation metrics used, there are still some extra features that could potentially increase the predictive power of this model. Obtaining information regarding the price of each property over the last 10 years would likely increase predictive power as historical prices tend to be important for predicting future prices. Additionally, a feature describing whether a property has been interiorly renovated within the last 10

years might prove useful because updating the interior of a property tends to increase property value regardless of whether the building it resides in is of old or new construction.

## Variable Importance

The variable importance plot in Figure 6 (below) identifies the variables that were most important for predicting property prices.

**Figure 6**



Intuitively, it makes sense that these are the most important variables because many people use these characteristics to determine whether they will buy the property. Additionally, the closer one gets to the center of nearly any city, the more expensive property becomes, and typically, the larger a property is, the more expensive it is. Ease of access to public transport is another factor which will naturally impact price, especially in a city like London. Finally, the more environmentally friendly a property is, the better, as pollution has become a considerable global issue.

## Assumptions

Assumptions were only made for the LR and KNN models as assumptions do not apply for the DT, RF, and GBM models. For the LR model, the assumptions are as follows:

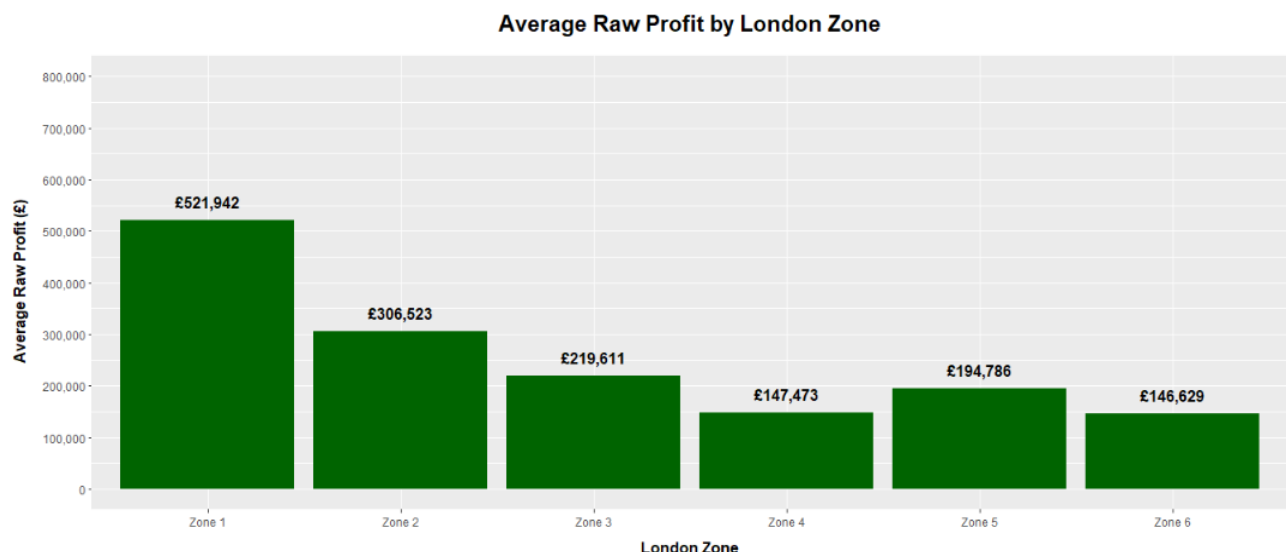
- The data has a linear relationship.
- The variables are normally distributed.
- There is little to no multicollinearity between explanatory variables.
- There is no auto-correlation present.
- The variables exhibit homoscedasticity<sup>4</sup> (Statistics Solutions, 2021).

For the KNN model, the only assumption made was that the data lie in a metric feature space - i.e., there is some notion of measurable distance between observations<sup>5</sup> (Thirumuruganathan, 2010).

## Comments

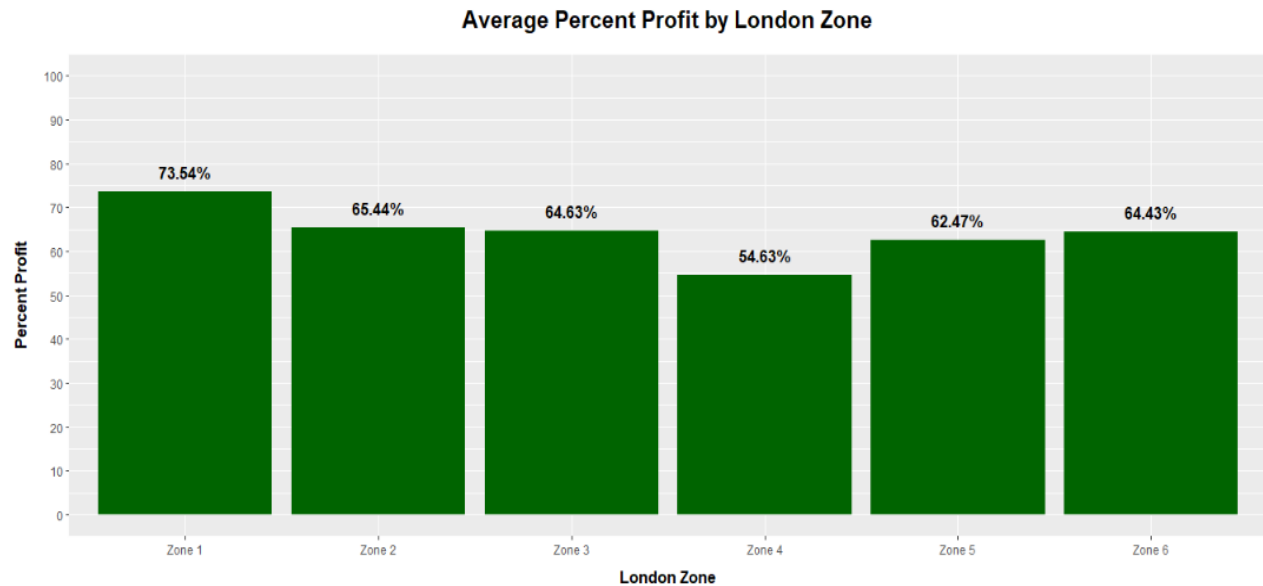
It should be noted that despite the expected ROI of 63.64%, the total amount of capital required to acquire these select 200 properties amounts to £90,669,000.00 (Appendix 1, Figure 2). The only entities likely to invest this much capital at once are institutions and perhaps select very high net worth individuals such as Jeff Bezos or Elon Musk. The summary statistics of the boxplot in Figure 1 (Appendix 1, Figure 3) revealed that the median property prices in zones one and two are markedly higher than those in zones three through seven. Thus, one would expect the raw profits from properties in zones one and two to be markedly higher than those in zones three through seven. Figure 7, below, illustrates the average raw profits by zone for the 200 properties selected by the model and confirms this suspicion.

**Figure 7**



Interestingly, however, the differences in average ROIs between the zones are not nearly as dramatic as the differences in average raw profits, as illustrated in Figure 8 below.

**Figure 8**



From Figure 8, it is clear that the average ROI for zone six is only 9.11% less than that of zone 1. However, the median property price of zone 1 is more than double that of zone 6. Thus, one could argue that in order to minimize capital investment and simultaneously maximize ROI, an investor should create a portfolio predominantly consisting of properties in zone 6 on the condition that the expected ROI of this portfolio is at least 63.64%.

## Reference List

1. Glen, S. (2015) *Lasso Regression: Simple Definition*. Available at: <https://www.statisticshowto.com/lasso-regression/> (Accessed: 21 December 2021).
2. Bobbitt, Z. (2021) *How to Interpret Root Mean Square Error (RMSE)*. Available at: <https://www.statology.org/how-to-interpret-rmse/> (Accessed: 21 December 2021).
3. Fernando, J. (2021) *R-Squared*. Available at: <https://www.investopedia.com/terms/r/r-squared.asp> (Accessed: 21 December 2021).
4. Statistics Solutions (2021) *Assumptions of Linear Regression*. Available at: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/> (Accessed: 21 December 2021).
5. Thirumuruganathan, S. (2010) *A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm*. Available at: <https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/> (Accessed: 21 December 2021).
6. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
7. Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
8. Frank E Harrell Jr (2021). Hmisc: Harrell Miscellaneous. R package version 4.6-0. <https://CRAN.R-project.org/package=Hmisc>
9. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
10. Jeffrey B. Arnold (2021). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.4. <https://CRAN.R-project.org/package=ggthemes>
11. Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. <https://CRAN.R-project.org/package=janitor>
12. Stephen Milborrow (2021). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.1.0. <https://CRAN.R-project.org/package=rpart.plot>
13. Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-88. <https://CRAN.R-project.org/package=caret>
14. Garrett Golemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), 1-25. URL <https://www.jstatsoft.org/v40/i03/>.

15. Ben Hamner and Michael Frasco (2018). Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4. <https://CRAN.R-project.org/package=Metrics>
16. Zachary A. Deane-Mayer and Jared E. Knowles (2019). caretEnsemble: Ensembles of Caret Models. R package version 2.0.1. <https://CRAN.R-project.org/package=caretEnsemble>
17. Julia Silge, Fanny Chow, Max Kuhn and Hadley Wickham (2021). rsample: General Resampling Infrastructure. R package version 0.1.0. <https://CRAN.R-project.org/package=rsample>
18. Yihui Xie, Joe Cheng and Xianying Tan (2021). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.20. <https://CRAN.R-project.org/package=DT>
19. Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2021). GGally: Extension to 'ggplot2'. R package version 2.1.2. <https://CRAN.R-project.org/package=GGally>
20. Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL <https://www.jstatsoft.org/v33/i01/>.
21. Revolution Analytics and Steve Weston (2018). doMC: Foreach Parallel Adaptor for 'parallel'. R package version 1.3.5/r19. <https://R-Forge.R-project.org/projects/domc/>
22. Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2020). gbm: Generalized Boosted Regression Models. R package version 2.1.8. <https://CRAN.R-project.org/package=gbm>
23. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. <https://CRAN.R-project.org/package=dplyr>
24. Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. <https://CRAN.R-project.org/package=scales>



# Appendix 1








Figure 1



Figure 2

	Price Value ↕
Total Acquisition Cost	£90,669,000.00
Total Revenue	£139,663,725.48
Total Raw Profit	£48,994,725.48

**Figure 3**

	Zone 1 	Zone 2 	Zone 3 	Zone 4 	Zone 5 	Zone 6 	Zone 7 
Whisker Minimum	£117,500.00	£96,000.00	£90,000.00	£89,000.00	£80,000.00	£77,000.00	£261,000.00
25th Percentile	£550,000.00	£420,000.00	£375,000.00	£330,000.00	£318,000.00	£327,500.00	£333,000.00
Median	£850,000.00	£580,000.00	£490,000.00	£427,277.50	£416,000.00	£402,250.00	£408,750.00
75th Percentile	£1,540,000.00	£915,000.00	£705,000.00	£555,000.00	£550,000.00	£525,000.00	£516,000.00
Whisker Maximum	£3,000,000.00	£1,655,000.00	£1,200,000.00	£885,000.00	£897,500.00	£820,000.00	£525,000.00