London Business School

Stacking Ensemble Learning for London Housing Price Prediction

Ding Linli

Abstract—In the following, we will analyze the housing price data obtained from Publicly available HM Land Registry's Price Paid Data to identify the features that affect the housing price in London from 2008 to 2020. Moreover, we propose a stacking ensemble model, where the predictions are generated by stacking four base learning models consisting of a simple linear regression (LR), a tree model (CART), k-nearest neighbors (KNN), and bayesian regularized neural networks (BRNN). The prediction is then stacked using LR.

1. Introduction

Housing in the United Kingdom represents the largest non-financial asset class in the UK. The overall net value passed the £5 trillion in 2014. To guide investors to make wise decisions in this lucrative market, housing prediction is of high value. An accurate prediction can generate great value for investors and sellers. Machine learning (ML) is powerful in predicting asset values and is being used throughout global business institutions. Therefore, building a robust ML model is very critical to the success of property investors and sellers.

2. Data Explanation and Exploration

This section, we explore the dataset using multiple exploratory techniques to uncover underlying patterns and select features.

2.1 Dataset

Data is publicly available (https://www.gov.uk/government/collections/price-paid-data) at HM Land Registry's Price Paid Data to identify the features that affect the housing price in London from 2008 to 2020.

Price Paid Data	
ID	A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded.
Price	Sale price stated on the transfer deed.
Date	Date when the sale was completed, as stated on the transfer deed.
Postcode	This is the postcode used at the time of the original transaction.
postcode_short	First part of postcode
Property Type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other Note that: - we only record the above categories to describe property type, we do not separately identify bungalows. - end-of-terrace properties are included in the Terraced category above. - 'Other' is only valid where the transaction relates to a property type that is not covered by existing values.
Whether_old_new	Indicates the age of the property and applies to all price paid transactions, residential and non-residential. Y = a newly built property, N = an established residential building
freehold_or_leasehold	Relates to the tenure: F = Freehold, L= Leasehold etc. Note that HM Land Registry does not record leases of 7 years or less in the Price Paid Dataset.
Adress 1	
Adress 2	
Adress 3	
town	
local_aut	
county	
district	
Energy Performance of Buildings Data	
total_floor_area	The total useful floor area is the total of all enclosed spaces measured to the internal face of the external walls
number_habitable_rooms	Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar; and also a non-separated conservatory
current_energy_rating	Current energy rating converted into a linear 'A to G' rating (where A is the most energy efficient and G is the least energy efficient)
co2_emissions_current	CO_2 emissions per year in tonnes/year
co2_emissions_potential	Estimated value in Tonnes per Year of the total CO_2 emissions produced by the Property in 12 month period.
energy_consumption_current	Current estimated total energy consumption for the property in a 12 month period (kWh/m2). Displayed on EPC as the current primary energy use per square metre of floor area
energy_consumption_potential	Estimated potential total energy consumption for the Property in a 12 month period. Value is Kilowatt Hours per Square Metre (kWh/m^2)
windows_energy_eff	Energy efficiency rating. One of: very good; good; average; poor; very poor. On actual energy certificate shown as one to five star rating.
tenure	Describes the tenure type of the property. One of: Owner-occupied; Rented (social); Rented (private).
Postcode data	
population	Population of the area covered by the Postcode (from the 2011 census)
households	Number of Households in the area covered by the Postcode (from the 2011 census)
latitude	Latitude of centroid of the Postcode for this row in decimal format i.e. 51.50205
longitude	Longditude of centroid of the Postcode for this row in decimal format. i.e -0.07864. Negative values are those to the West of the zero (Greenwich) meridian
altitude	Height above sea level measured in Metres
london_zone	Transport for London (TfL) Travel Zone indicator (London area only)
nearest_station	The nearest train station to the postcode. For London, also includes Underground and tram stops
distance_to_station	The distance in kilometres to the nearest station from the postcode
water_company	The name of the water company responsible for this postcode
average_income	Average household income of the MSOA that the postcode is located in
Tube Information	
type_of_closest_station	If the nearest station has a tube line ="tube, if it does not have a tube line but a light rail (DLR, Overground)="light rail",otherwise ="rail"
num_tube_lines	Number of tube lines that use the closest station
num_rail_lines	Number of rail lines that use the closest station
num_light_rail_lines	Number of light rail lines that use the closest station

Table 1

The given training dataset has 38 columns in total, including 36 features, 1 label (price), and 1 index (ID). Among the 36 features, there are 18 categorical variables and 18 numerical variables. From initial data inspection, we can see that some variables have a significantly low complete rate as shown in Table 2 and therefore will not be used for training the model.

	complete rate
address2	22.8%
town	4.39%

Table 2

With an inspection of the out-of-sample set, some variables have 0 complete rate as exhibited in Table 3. Therefore, when training the model, these will not be included, since they are unable to explain the variance of the out-of-sample data.

	complete rate
date	0%
postcode	0%
address1	0%
address2	0%
address3	0%
local_aut	0%
county	0%

Table 3

Additional Features

For the purpose of better prediction, I have added two features into the training and out-of-sample dataset.

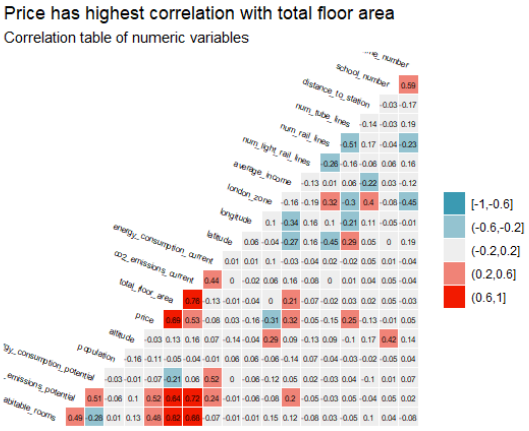
The first one is the number of schools aggregated at borough level in 2016, which is publicly available (https://data.gov.uk/dataset/6b776872-c786-4960-af1d-dab521aa4ab0/london-schools-atlas). This is expected to have a positive effect on housing prices. Our estimation results show that distance to all three school levels including elementary, middle and high schools has a significant impact on house values: the closer a house is to schools, the higher price a house will have. In addition, we find that distance to elementary schools and distance to middle schools play a more important role

than distance to high schools in affecting housing prices (Huang, 2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3096145

The second one is the number of crimes in each London borough in 2020, which is publicly available (https://data.london.gov.uk/dataset/recorded_crime_summary). A high number of crimes can have a negative effect on housing prices, since more dangerous neighbourhoods tend to have less demand from buyers.

In addition to the two features introduced above, a measure of how developed a neighbourhood is can be useful in predicting prices (e.g. parks, hospitals, etc.). The more developed the infrastructures are, the higher the price can be. The other piece of information that can be useful is air and noise pollution. For example, houses situated near railway or sewage channels tend to have lower prices.

2.2 Correlation Table



The correlation matrix between numeric variables is presented. The coefficient ranges between -1 and 1, implying the attributes are perfectly negatively or positively correlated, respectively. And 0 means the pair is perfectly uncorrelated. Attribute ‘price’ has a strong positive relationship with ‘total floor area’, ‘number of habitable rooms’, ‘co2 emissions potential’, ‘co2 emissions current’, ‘average income’ and ‘number of tube lines’. In contrast, the attribute ‘Price’ has a strong negative relationship with attributes ‘energy consumption potential’, ‘London zone’.

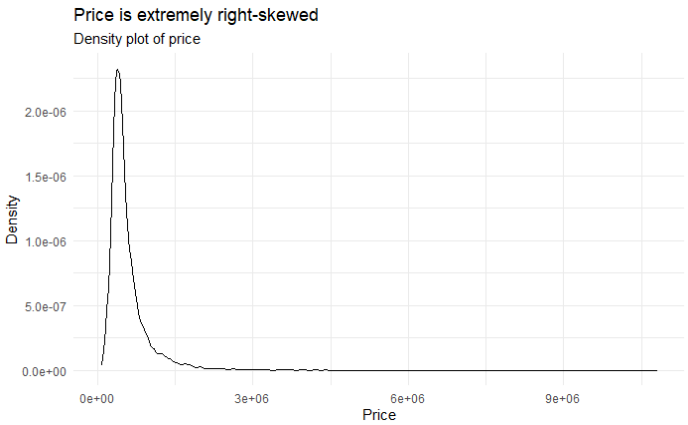
Intuitively, positive correlation between ‘total floor area’ and ‘price’ means the bigger the house, the higher the price. Similarly, the positive correlation between ‘average income’ and price indicates that richer neighborhoods have higher housing prices. The more central the location is, the higher the price, as explained by negative correlation between ‘London zone’ and ‘price’. Since energy efficiency is important for long term investment of a house, the lower the energy consumption, the higher the price. Note that unlike ‘energy consumption current’, ‘co2 emission current’ is positively associated with ‘price’. This is because co2 emission, unlike energy consumption, is not penalized in monetary terms for private building owners and therefore is not a critical factor for investment decisions. And co2 emission highly depends on how big the house is.

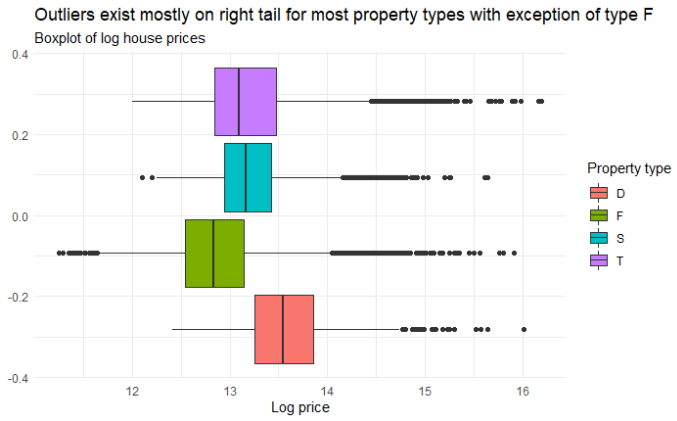
Multicollinearity Problem

Note that among these attributes that are highly correlated with attribute ‘Price’, some pairs of them are highly correlated between themselves (e.g., ‘total floor area’ and ‘number of habitable rooms’). In the following LR model training steps, only one attribute from each pair should be used as input for each model. This is due to the multicollinearity problem, which can negatively impact the result of LR. This is because the key goal of LR is to isolate the relationship between each independent variable (e.g., ‘total floor area’) and the dependent variable (i.e., price). However, when independent variables are correlated (e.g., ‘total floor area’ and ‘co2 emission current’), it means change in ‘total floor area’ shifts ‘co2 emission current’. As a result, the model has difficulties estimating the relationship between each independent variable and the dependent variable. Therefore, only one attribute of each pair will be selected into LR. In the following, I will select ‘total floor area’ as input to avoid multicollinearity.

Other models used in this paper, including CART, KNN and BRNN (De Veaux & Ungar, 1994) are relatively insensitive towards multicollinearity.

2.3 Distribution of House Prices

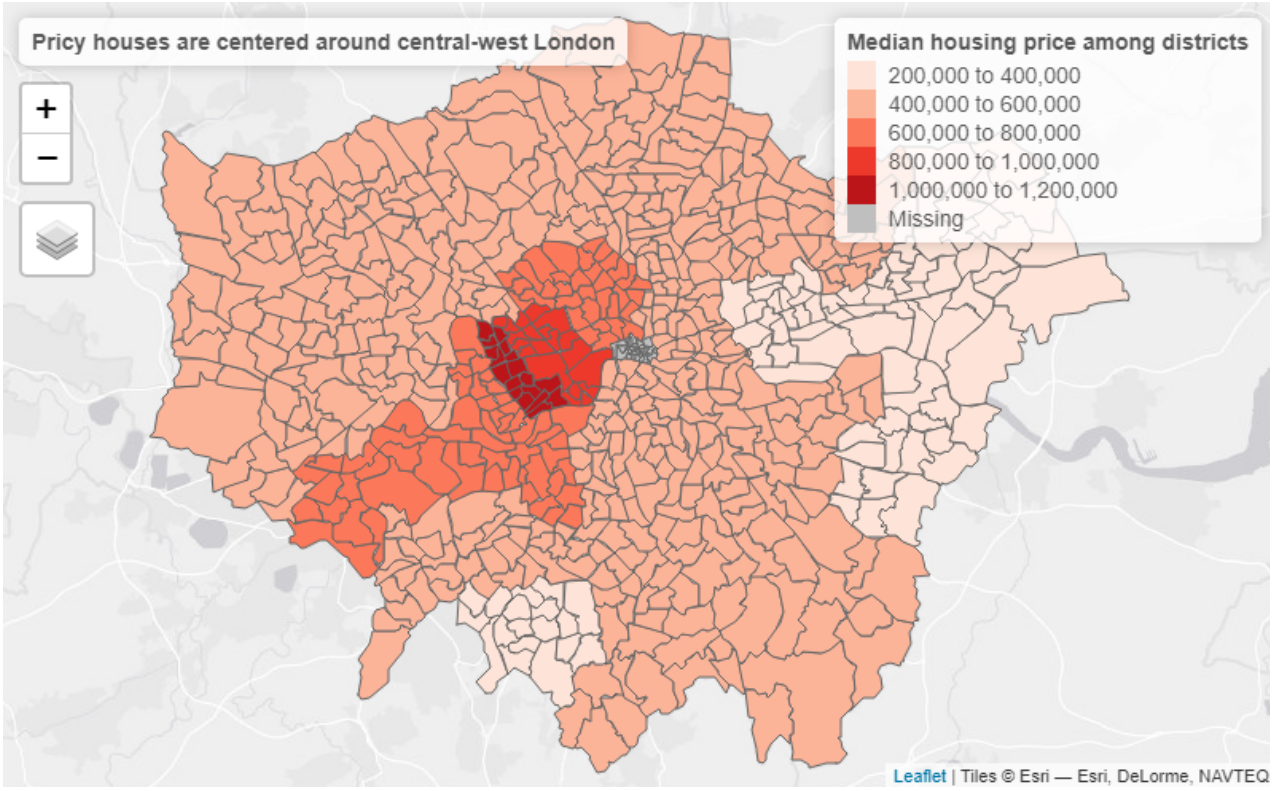




We can see from the density plot that the price is mostly concentrated at low value, with high-value outliers. Therefore, log transformation is used to decrease the impact of high-value outliers, and thus making it easier to visualize the distribution of prices.

From the boxplot we can see that different property types have different distributions of log price. We can see that F (flat) has relatively low prices compared to other types. It also has more lower-end outliers, meaning that some flats (e.g., basement flat) have significantly low prices. Type D (detached house), has on average higher prices than other property types. This is intuitively the case in that detached houses are higher class and more luxurious in general. We can also see that S (semi-detached house), has the smallest quantile range.

2.4 Average Housing Price Among Districts

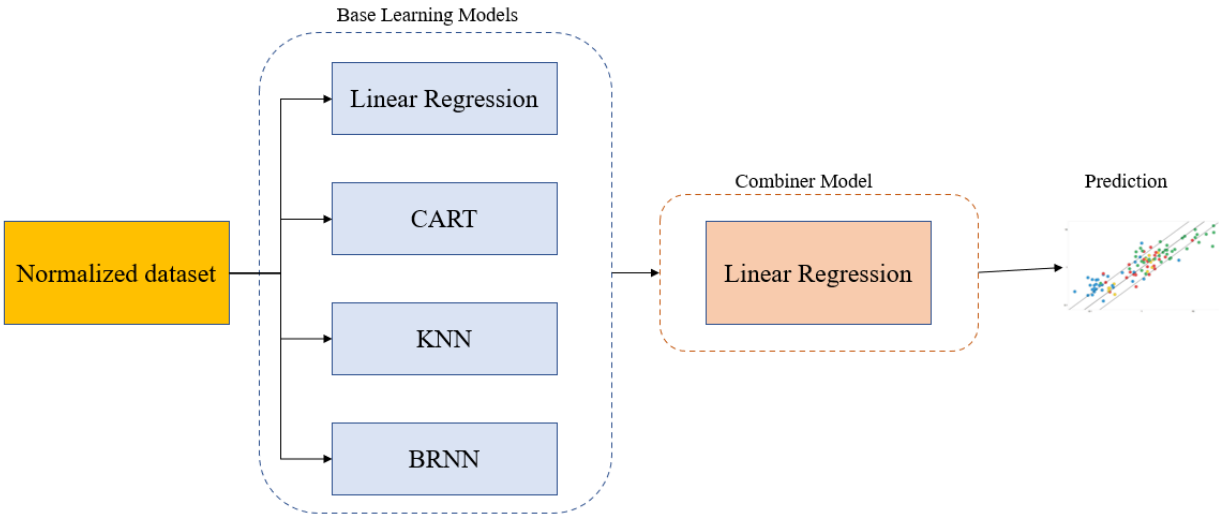


From the graph above we can see that the most pricy houses are concentrated around central and west London. This is probably because west London has more advanced infrastructures, parks and mansions that attract the middle to upper class investors.

Note that our dataset does not include any observations in the district of the City of London. Therefore, the model will not be used to predict prices in the City of London.

3. Methodology

In this section, we will explain the procedures of building the stacked ensemble model which is used to predict housing prices.



- The dataset is first normalized, this is necessary for KNN and BRNN to ensure unbiasedness. Normalization does not affect LR and CART.
- Then split the dataset randomly into 10,442 (75%) training set and 3481 (25%) test set. Training data is used for training the model, while testing data is used for validation.
- Afterwards, different subset of features are utilized to train each of the baseline models. Each subset is selected based on the suitability of each algorithm. For example, total floor area is selected while number habitable rooms is deselected for LR to avoid multicollinearity. To moderate the overfitting issue, we used k-fold cross validation with k=5.
- Once we obtain the results of the training models, a combiner model (LR) is used for stacking the four base learning models.
- Compare the performance of base learning models and stacked ensemble models and choose the optimal stacked ensemble model by looking at the RMSE and Rsquared.
- Predict the out-of-sample house prices using the optimal model.
- Lastly, compare the predicted price and asking price of the out-of-sample houses to choose the most profitable 200 houses.

3.1 Base Learning Models

LR

The first base learning model is LR. This model assumes a linear relationship between each independent variable and dependent variable. The advantage of this model is that it can most effectively capture the linear relationships (e.g., total floor area and price). It is highly capable of predicting continuous variables and is simple and transparent. However, it does not work well with non-linear relationships (e.g., longitude and price). It also is not sufficiently capable of making use of categorical variables.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -1.341e+07  1.545e+06  -8.682  < 2e-16 ***
latitude                       2.697e+05  3.003e+04   8.980  < 2e-16 ***
longitude                     -2.956e+05  1.495e+04 -19.771  < 2e-16 ***
london_zone                   -6.437e+05  2.746e+04 -23.440  < 2e-16 ***
total_floor_area               2.557e+02  1.128e+02   2.266   0.0235 *
crime_number                  -6.781e+00  4.699e-01 -14.430  < 2e-16 ***
school_number                  4.748e+02  1.185e+02   4.008 6.17e-05 ***
distance_to_station           -4.736e+04  6.300e+03  -7.518 6.03e-14 ***
population                    -3.827e+02  5.480e+01  -6.984 3.05e-12 ***
property_typeF                -1.632e+05  1.279e+04 -12.763  < 2e-16 ***
property_typeS                -1.201e+05  1.144e+04 -10.493  < 2e-16 ***
property_typeT                -1.431e+05  1.151e+04 -12.435  < 2e-16 ***
`london_zone:total_floor_area` 1.951e+04  2.322e+02  84.037  < 2e-16 ***
```

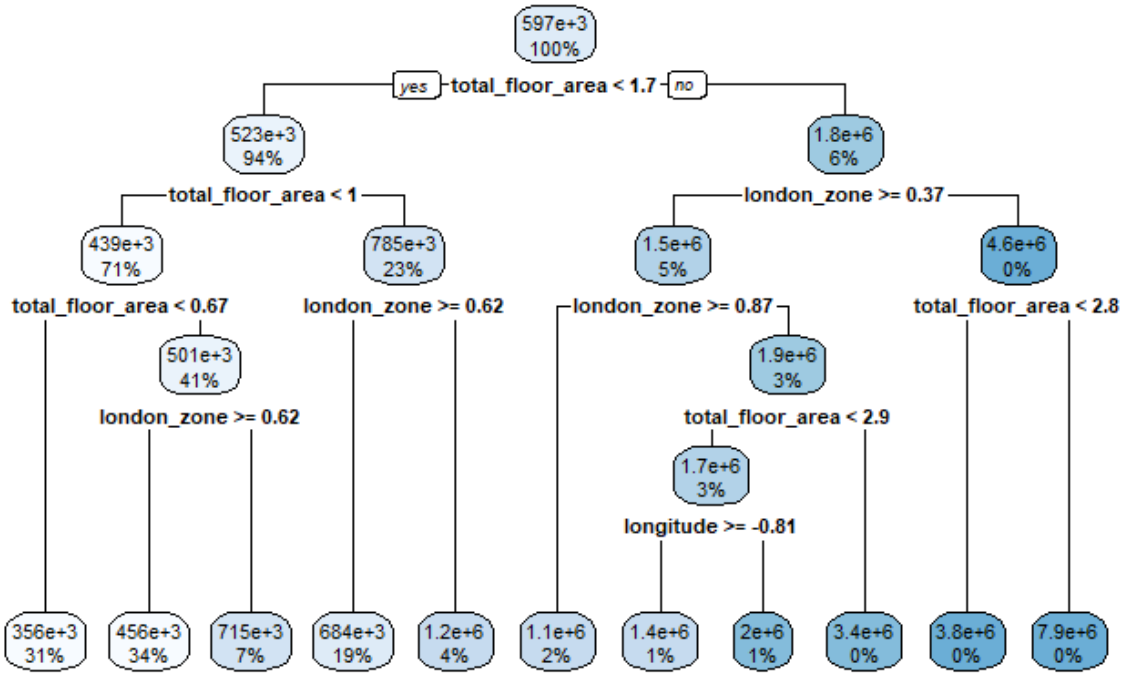
The model calculated the coefficients for each of the variables. For example, the coefficient for ‘total floor area’ is 255.7. This number means that with 1 square feet increase in the house, there will be a £255.7 increase in price.

I added the interaction term ‘total floor area*london zone’ for the purpose of explaining more variance in the data. Note that the London zone here is transformed as ‘1/London zone’. Therefore, the higher the value of this attribute, the more prime the location. This is based on the assumption that if the house size is big and based on central location, the effect on price is more significant than house size or London zone alone. Looking at the coefficient for this interaction term, we can see that in London zone 1, if total floor area is increased for 1 square feet, the price will go up £19,510.

CART

The second base learning model is classification and regression trees (CART). This model is transparent in terms of easy to follow the algorithm steps. It is also highly capable of making use of categorical variables. However, it is less capable of predicting continuous variables as it can only predict the value ranges of an observation. It is also computationally expensive and requires a longer time to train a model.

The model uses binary decisions to classify the range of prices. The depth of this tree model is controlled by complexity parameter (CP). This parameter penalizes high complexity. With higher CP, the depth of the tree is lower. In this paper, CP=0.01, so that the depth is controlled at 6 for model transparency and avoid overfitting. In this regression tree, each leaf is calculated as the mean of the observation group. For example, the first leaf means that if total floor area < $\mu+0.67\sigma$, there are 31% of the houses with average price=£356,000. Some of the leaves on the far right represent the largest houses (e.g., total floor area > $\mu+2.8\sigma$ in the prime location (e.g., london zone < $\mu+0.37\sigma$).



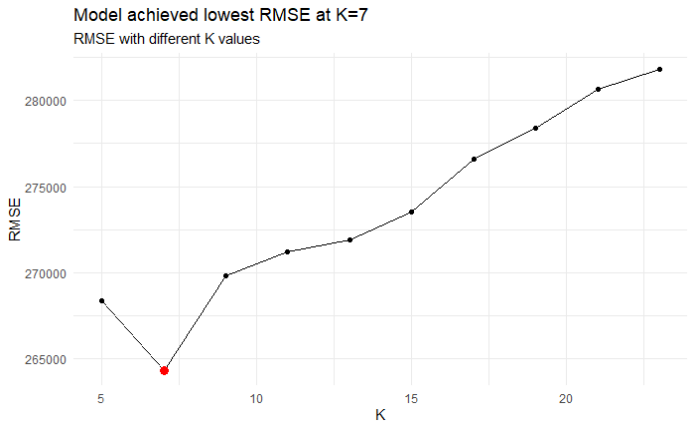
We can see from the result that the RMSE reached 294609.8, which is 1.1% higher than the RMSE obtained from the LR model. This is the case probably because the tree model predicts the price range instead of the exact price as LR does. Therefore the error is higher.

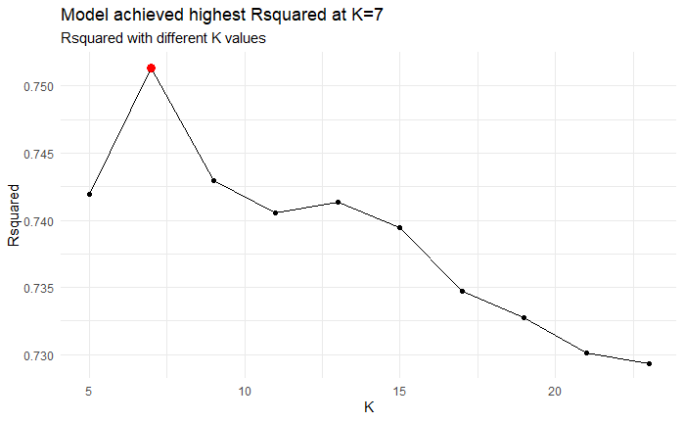
Rsquared reached 69.0%, which is 2.5% higher than that of LR model. This is the case probably because many relationships between independent variables and dependent variable are non-linear, and LR fails to detect the non-linear relationship. Tree model is able to capture the non-linear relationship. Therefore, Rsquared is higher than LR model.

KNN

The third base learning model is k-nearest neighbors (KNN). KNN assumes that similar outcome exist in close proximity. Unlike LR and CART, KNN requires data input to be normalized. This is because nearest neighbors are selected from the shortest euclidean distance between pairs of data observations. This will result in variables with bigger range (e.g., average income $36,000-£85,200) to be much more informative compared to variables with smaller range (e.g., London zone 1-7). The algorithm will essentially rely on variables with bigger ranges. Therefore, in the following analysis, numeric variables are preprocessed to be normalized. Although KNN can produce good accuracy on testing dataset, the biggest downside of this algorithm is that it is computationally expensive.

To find the best number of k, we calculated the RMSE and Rsquared for k = 5, 7, …21, 23. The performance of KNN is below. We found that k = 7 to be the optimal parameter.

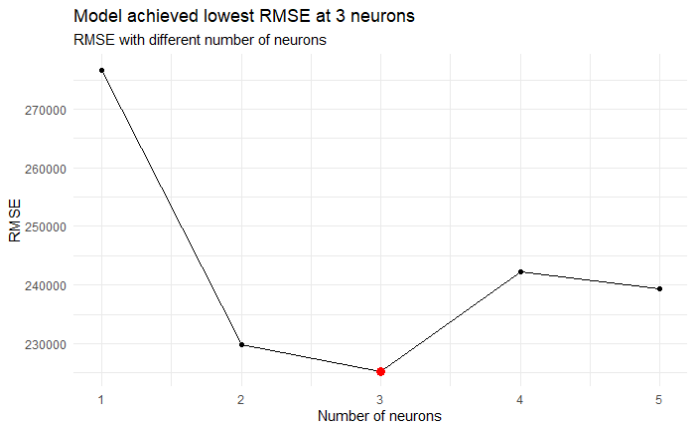


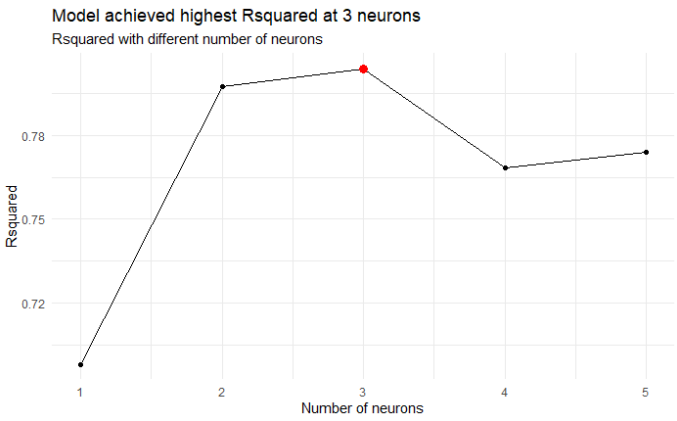


BRNN

The fourth base learning model is bayesian regularized neural networks (BRNN). BRNN is suitable for capturing complex nonlinear relationships between independent and dependent variables (e.g., longitude and price). It is also capable of detecting variable interactions. However, BRNN's disadvantage is being a "black box", i.e., not transparent. It is also computationally expensive.

To find the best number of neurons, we calculated the RMSE and Rsquared for neurons = 1, 2, 3, 4, 5. The performance of KNN is below. We found that neurons = 3 to be the optimal parameter.



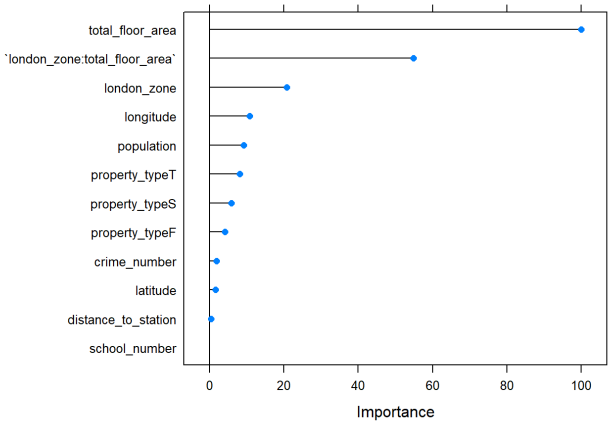


Feature importance

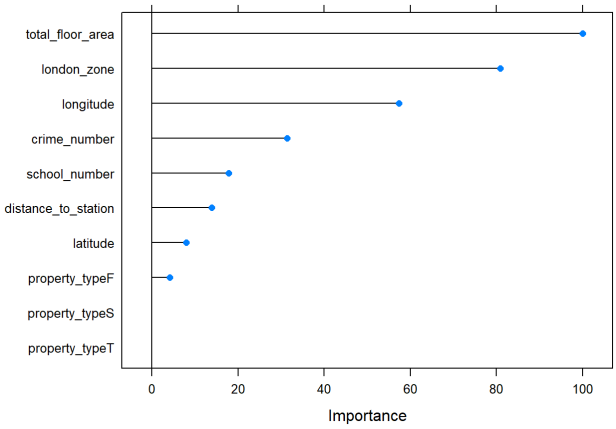
Feature importance of all four base learning algorithms are exhibited below. We can see that in general, most features have different levels of contributions to different models. KNN and CART have relatively the same selection of important features and ordering of feature importance, while LR and BRNN are different.

Location-wise, total floor area and London zone, being two of the dominant features for all four models, explains the most variance compared to other features. This is the case intuitively in that the bigger the house, the higher the price. We can also see that 'distance to stations' is generally powerful for all four models since commute is a critical consideration for homeowners.

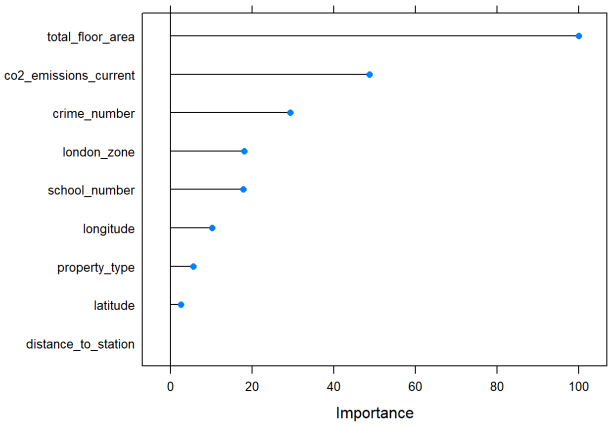
Demographic-wise, ‘crime number’ plays an important role in CART, KNN and BRNN, while contributing significantly less in LR (excluded because of a p-value > 0.05). Attribute ‘average income’ is important only for BRNN, while for other algorithms it is not as important and therefore has been excluded.



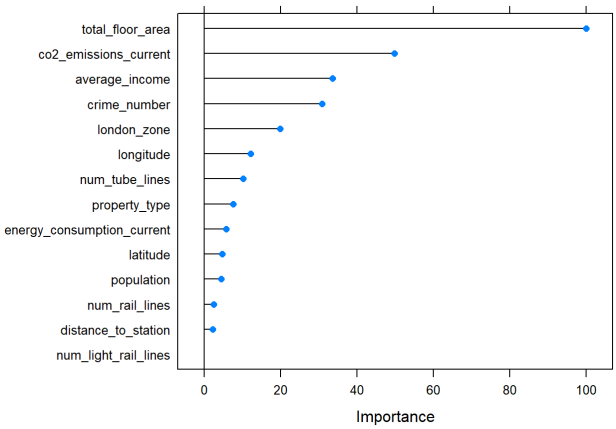
LR



CART



KNN



BRNN

3.2 Combiner Model

The four models trained above are then stacked using LR. After the training set has been predicted, we take the predicted prices as four features, and the original price column as the label to train the combiner model using LR.

3.3 Stacked Ensemble Model

With the test data set, we calculated the performance of the base learning models as well as the stacked ensemble model. Based on the table below, we can see that compared to individual models, the stacked ensemble model achieves an R-squared of 81.4%, significantly outperforming LR, CART and KNN, while slightly outperforming BRNN.

	RMSE	R-squared
LR	291132.9	66.6%
CART	298042.6	65.2%
KNN	264300.5	75.1%
BRNN	225238.2	80.4%
Stacked Ensemble Model	217699.6	81.3%

Finally, to select the 200 most profitable houses, we will compare the predicted price from stacked ensemble model and the asking price of the out-of-sample dataset.

The profitability is calculated as below.

Profitability = (predict price - asking price) / asking price

We then ordered the houses based on profitability, and selected the highest 200 as recommendation for investment.

4. Further Research

4.1 CART and KNN

From the above analysis, we see that KNN and CART may explain similar variances of the data. Therefore, one of the algorithms can be deleted to improve computational efficiency. However, further context and requirement should be given to decide which model should be dropped. While KNN significantly outperforms CART, CART consumes less computational resources. Further investigation is needed to see whether performance or resource efficiency is more important.

4.2 Effect of Elizabeth Line (Crossrail)

First, I would filter the houses that are near Elizabeth Line as observations with the range of only 2020 data, let's call this dataset crossrail_dataset. Since we would assume that we can find the distance from each property to the closest Crossrail station, this can be done by setting a radius (e.g., 1 mile), and then selecting the houses within the range of radius of each station along the Elizabeth Line as illustrated below.

In this case, we will select House X and House Y, and deselect House Z.

Next, we calculate the reduction in distance to the nearest station from each house after opening the Elizabeth Line, with the following illustration as an example.

Reduction in distance = distance to nearest station_old - distance to nearest station_new

In this case, the reduction in distance is equal to 1-0.5 = +o.5 mile.

Next, I would filter the houses that are near Jubilee Line as observations with the same procedures as mentioned above, with the year range from 1979 to 1999, let's call this dataset jubilee_dataset. Since Jubilee Line's development is the most recent and with similar pace as Crossrail, it can be used to estimate price elasticity against reduction in distance.

We will train a model to calculate the annual rate of price growth, with independent variables of both postcode and reduction in distance for jubilee_dataset. Assume that price growth rate remains the same from 2020 to 2021. Once we obtain the annual growth rate, by both postcode and reduction in distance, we apply the growth rate to the crossrail_dataset to arrive at predicted prices as of 2021.

Finally, we can compare the asking price and predicted price to see if the house is profitable or not.

5. Conclusion

We propose the stacked ensemble model for price prediction. After data exploration, we found that 'total floor area', 'average income' and 'number of tube lines' have high positive correlation with 'price', whereas 'energy consumption potential', 'London zone' have a negative correlation with 'price'. Additionally, we engineered 'number of schools' and 'number of crimes' as features into the dataset.

In order to build an optimal model, we selected LR, CART, KNN and BRNN as base learning models, capturing different variances of the dataset. Our result shows that BRNN significantly outperforms other algorithms, while CART and KNN may explain relatively the same variances. Finally, LR is used as a combiner to stack the base learning models. The final stacked ensemble model gives us the best performance with R-squared of 81.4%.

References

De Veaux, R.D. and Ungar, L.H., 1994. Multicollinearity: A tale of two nonparametric regressions. In Selecting models from data (pp. 393-402). Springer, New York, NY.

Huang, P. (2018). Impact of distance to school on housing price: Evidence from a quantile regression. The Empirical Economics Letters, 17(2), 149-156.