Project 2 London House Prices

Kázmér Nagy-Betegh

Table of Contents

Abstract

The report sets out to develop a model to predict the value of London properties based on features related to their size, location, and build quality. The purpose of the model is to aid investment decisions of London property buyers. The models investigated are Linear regression, Lasso Regression, K-Nearest Neighbours Regression, Gradient Boosting Machine, Tree model, Random Forest Tree, and stacking the mentioned models together which results in the best performance with RMSE of 175508.7 and an Rsquared value of 0.8724963 on the chosen test data. The report discusses the method of tuning the models and presents the best tuning parameters found and the feature selection process followed by a theoretical proposal for modelling the effect of Crossrail's opening on property prices.

Introduction

The London property market has seen significant growth in the past decades, greatly outperforming wage growth within the capital (113% wage increase, 600% property price increase since 1995), making it significantly harder for most of Londoners to own their home. This growth made the London property market attractive to local and foreign investors alike. Since 1995 there has been very few slowdowns in growth. Significant ones were during the 2008 financial crisis, 2016 Brexit referendum and in 2020 Brexit implementations and the Coronavirus pandemics combined effect.

Currently London property prices have experienced their first negative year over year growth in a long time, however prices are still close to historical highs. The uncertainty caused by Brexit and the Pandemic, creates opportunity, and increases risk for investors and home buyers. Prices are somewhat better, which offers buying opportunity, but there is a chance that some properties will depreciate further.

This report presents a price estimation algorithm to allow for better value evaluation of London properties. The algorithm relies on data with features seen in Table 2. The report sets out to present the accuracy of the model, how the features and feature interactions have been chosen and how different regression algorithms performed with the data such as Linear Regression, Lasso Regression, Tree Models, Random Forest Tree, Gradient Boosting Machines, and ensemble methods by stacking the previously mentioned models.

The best performing algorithm was the ensemble model, which was a stacked model of linear regression, lasso regression, KNN, decision tree, gradient boost machine and random forest tree, which subsequently was used for picking 200 properties to invest in from an out of sample dataset.

Predicting London House Prices

Model Optimisation

The available dataset [1] can be considered clean as an inspection using the skimr package [2] package reveals that all features apart from population, address2 and town has a completion rate of 100%, which means there is no missing data. From this we can also see that the data within most features is right skewed, but not strongly enough to require a log transformation. Also, there are no obvious outliers present upon plotting features of the dataset and further inspecting the skimr statistics. Based on this the data is sufficiently clean to be used for model development.

The available dataset is split into testing and training data at a 25% and 75% split by taking random samples for each set using rsample package [3]. The training data then is used for creating models and the testing data to compare the generated model's performance. Throughout all the operations involving randomness the set.seed(123) function is used for comparability and reproducibility.

The first model used is a linear regression model to help with an explanatory analysis of the dataset and feature selection for the upcoming tree-based models. The models explored are Linear Regression, K-Nearest Neighbours, Gradient Boosting Machines, Decision Tree and Random Forest tree, Lasso Linear regression, these algorithms then are stacked to explore if that allows for additional performance gains.

Each model is trained using the caret packages [4] train function using fivefold cross validation and taking advantage of tuneGrid option within the train function to adjust the tuneable parameters for each model. The best parameters are found by adjusting the tuning grid until the best result falls within the limits of the tried values likely indicating that the best combination has been found and no further increase or decrease in the parameters in necessary.

Before the model testing can begin the features to be used and feature interaction need to be decided. This was done using a linear regression model which is great for explanatory analyses, is quick to run and returns a statistical significance for each variable and variable interaction. The model is then built iteratively until R-Squared cannot be increased further and the root mean squared error (RMSE) stops decreasing both in within the train and test datasets. The variable selection here is cross checked with the variable significance results from Boruta analyses [5], which can be see in Figure 1 Appendix C.) along with the variable importance within the linear regression model in Figure 2 Appendix C.).

The features selected can be found in Table 3 within Appendix B.). The Importance of different features by algorithm can be viewed within Appendix C.). Based on these there are some features that carry importance in each model.

Total floor area, district, longitude, and latitude all carry significant weight in each model which makes sense as the size of the apartment significantly influences the price as well as its location. Interactions between number of rooms total floor area and average income are likely to reveal information about the size of the rooms and average number of occupants which helps define the price.

Energy and CO2 ratings also have an effect however interestingly only great energy ratings and very poor ratings are significant. Building onto this idea the energy and CO2 potentials interaction with the current ratings is important as it reveals investment opportunities, for example apartments where the energy rating can be improved significantly by the changing the windows can improve the value of the apartment and this is revealed by the interactions between the windows energy rating and energy potential.

The distance to the closest station from the apartment carries no importance or low importance in most models, the reason it was included was to be able to assess its usefulness for modelling the effect Crossrail opening on property prices and which model could help in that based on the available features.

Interestingly the water company that supplies the flat carries useful information, as it likely indicates the year the property was built and is tied to location.

Another surprising finding was that the variable indicating if the property is old or new didn’t carry much importance in any of the models which is likely because the properties of a home defined by its age are captures by the energy, CO2, and window ratings along with the water company information already.

While each tested model might have shown some marginal improvements from removing the features unimportant within the given model from its formula, the chosen features give a good representation of the data and were kept consistent across the compared models to allow the comparison to the stacked ensemble method as stacking models built on different feature sets it not possible.

Please find the performance comparison of models in Table 1. The best tune parameters for the random forest tree might change slightly from run to run as the set.seed() feature doesn't seem to remove all the randomness of the caret train algorithm for random forests.

Table 1 Regression Model Performances

Model Performance on Test Data

model_name	RMSE	Rsquared	Best Tune Parameters
Linear Regression Model default	440579.3	0.1776749	
Linear Regression model optimised	257835.8	0.7196096	
Tree Model	225265.9	0.7866105	cp = 0.00004
Random Forest Tree	175720.6	0.8698560	.mtry = 19 .splitrule = "variance" .min.node.size = 3

Gradient Boosting Machine	184213.5	0.8607442	n.trees = 1500
			interaction.depth = 20
			shrinkage = 0.1
			n.minobsinnode = 6

KNN	224306.1	0.7921005	K = 7

Model Performance on Test Data			
Lasso Regression	257522.1	0.7201801	Fraction = 0.8578947
Stacked models from above	**175508.7**	**0.8724963**	**method = “glm”**

Best Model

The best performing model was the stacked ensemble model consisting of all the models tested and built before. For model comparison, the same testing data was used for each model and RMSE was the main metric of comparison. The stacked model has the lowest RMSE and highest R-Squared when applied to the test dataset. The component performance of each model within the stacked model can be viewed in Figure 7 Appendix D.).

Choosing 200 investments

The criteria for evaluation of the investment choices are the average percentage return. Which means after predicting the prices within the out of sample dataset the investment decision will be made based on the percentage difference between the predicted price and asking price. This should result in investments with the highest probability of positive return.

Only selecting the properties that result in the highest sum in the difference between predicted and average price, might lead to a higher profit prediction, but this would mean taking on more risk because in case of high value properties the difference in percentage might be low while the money difference seems high. This would be a misguided approach as the model is not 100% accurate and lower return percentages are more likely to fall within the error margin of the model meaning that the predicted profitability might be an error.

Estimating the Effect of Elizabeth Line (Crossrail)

The effect of the Elizabeth line could already be estimated using the current models but not to a great degree. Within the Lasso, KNN, Random Forest tree models the distance to station carries importance but is outweighed significantly by other features.

Using these models would allow to capture some of the effect of the opening Crossrail stations close to properties. As the distance to station feature of these properties would change.

However, to improve the estimation of Crossrail's effect some new features could be engineered within the data. The proposed features would be the average journey time and number of stations to reach significant London locations. Calculate the number of stops and time to reach Oxford Circus, Kings Cross, Regents Park, Canary Warf, City of London and Major airports such as London Gatwick, London Heathrow, London Luton, London City. These features should allow for capturing how well connected a location is thanks to the station within its vicinity. Based on these locations it could be assessed how quickly occupants of each property are able to reach the office, and leisure activities in London (exact destinations might change based on testing the proposed ones are based on initial intuition).

The value of transport links in London is manifested by how well it connects property occupants to both useful and desirable destinations within London. To create a predictive model based on these features the previously mentioned algorithms need to be investigated. In the decision of choosing the right model the variable importance of the new features and distance to station should play a factor as well along with the RMSE and R-Squared values of the model.

To create the mentioned features a London tube dataset could be used (such as this [6]) or making use of the Transport for London journey planner API [7].

Conclusions

The report presents the evaluation of multiple regression models for predicting London property prices within the given dataset. It presents how the features of interest were found, for the models and how well each model performs.

It also presents the best or close to best tuning parameters for each model found using the tuneGrid option within the train() function of the caret package [4].

After comparing the model performances on the same test dataset, it is revealed that the best performing model for this dataset is an ensemble method created by stacking, linear regression, Random Forest Tree, Tree, Gradient Boosting Machine, K-Nearest Neighbours and Lasso models, resulting in a RMSE of **175508.7** and an Rsquared value of **0.8724963.**

Within the "*Estimating the Effect of Elizabeth Line (Crossrail)"* section a feature engineering method is proposed to improve the estimation of the effect of Crossrail on property prices that also might improve model performance for general price estimation as well. The proposed features to add were the journey time and number of stops to London areas of interest.

References

[1] UK Goverment, ‘HM Land Registry: Price Paid Data’, *HM Land Registry: Price Paid Data*. https://www.gov.uk/government/collections/price-paid-data

[2] E. Waring, M. Quinn, A. McNamara, E. A. de la Rubia, H. Zhu, and S. Ellis, *skimr: Compact and Flexible Summaries of Data*. 2021. [Online]. Available: https://CRAN.R-project.org/package=skimr

[3] J. Silge, F. Chow, M. Kuhn, and H. Wickham, *rsample: General Resampling Infrastructure*. 2021. [Online]. Available: https://CRAN.R-project.org/package=rsample

[4] M. Kuhn, *caret: Classification and Regression Training*. 2021. [Online]. Available: https://CRAN.R-project.org/package=caret

[5] M. B. Kursa and W. R. Rudnicki, ‘Feature Selection with the Boruta Package’, *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.

[6] ‘Tubemaps Github repo’. https://github.com/nicola/tubemaps

[7] ‘Transport for London Unified API’. https://api.tfl.gov.uk

Appendix

A.)

Table 2 Data Dictionary

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.

Price Paid Data

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

Price Paid Data

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

Price Paid Data

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

B.)

Table 3 Selected Features and Feature Interactions for the model

Features Selected for Models
average_income: london_zone
district
water_company
property_type
freehold_or_leasehold
latitude*longitude*total_floor_area*altitude
average_income:number_habitable_rooms:total_floor_area
I(average_income^3)
energy_consumption_current
energy_consumption_potential
windows_energy_eff
co2_emissions_current:co2_emissions_potential
current_energy_rating
distance_to_station

Features Selected for Models
type_of_closest_station
num_rail_lines
num_light_rail_lines

C.)

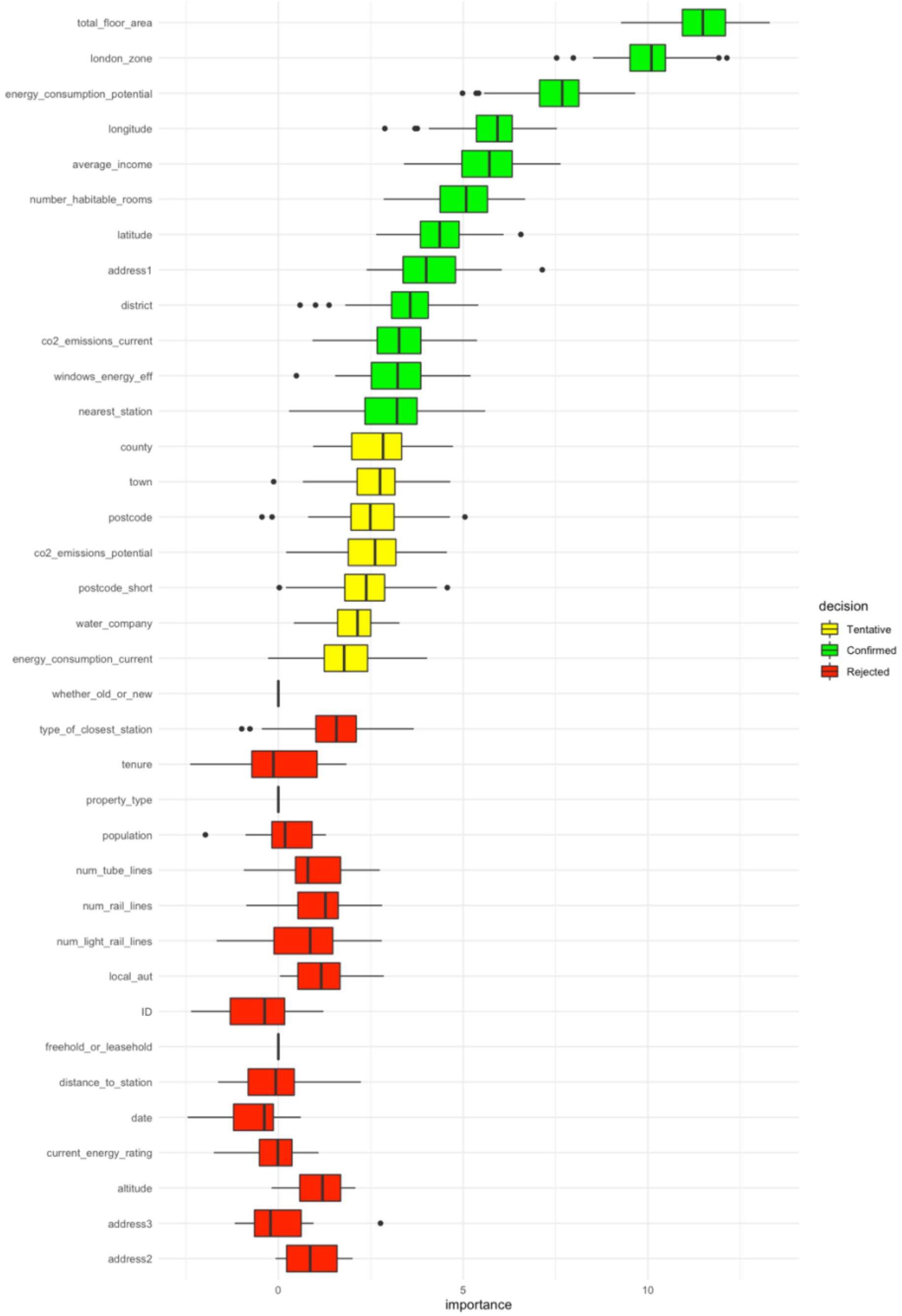

Figure 1 Boruta Analyses of Variable Importance

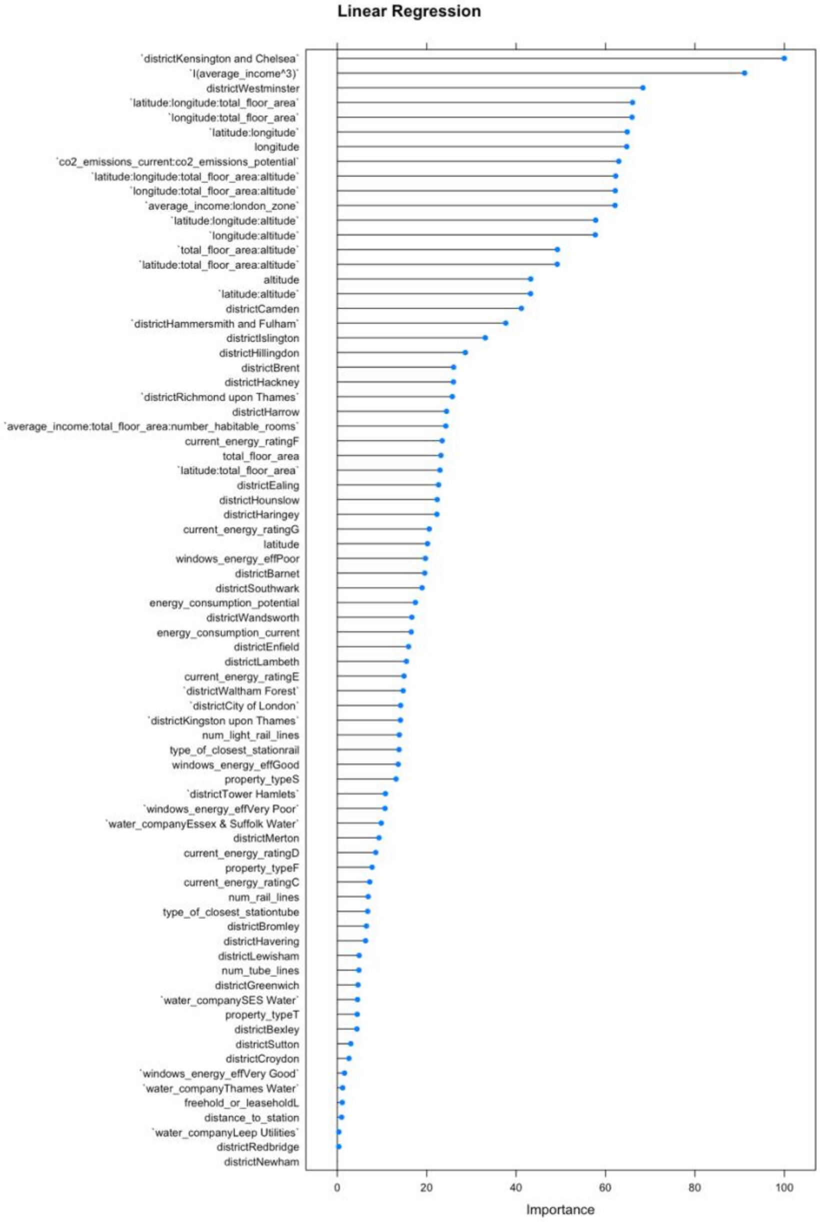

Figure 2 Linear Regression Variable Importance Scaled

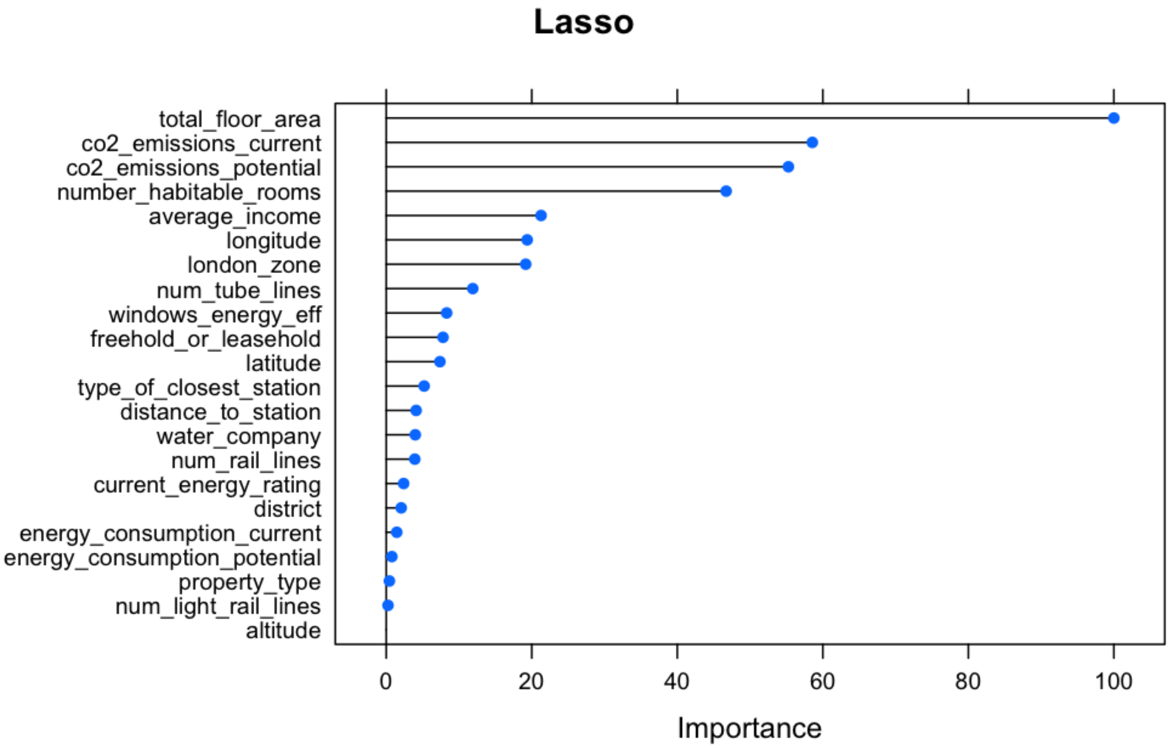

Figure 3 Lasso regression variable importance Scaled

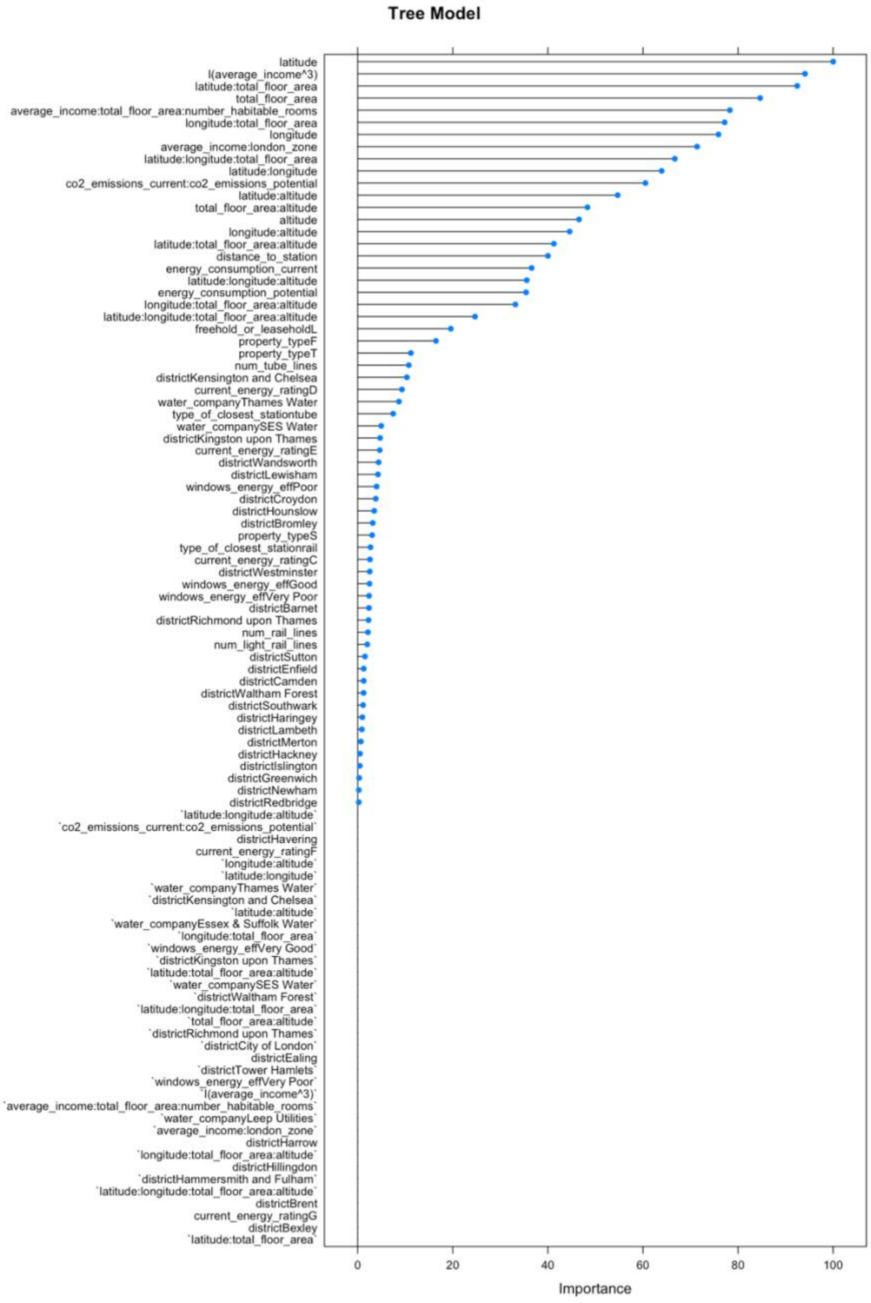

Figure 4 Tree model variable importance Scaled

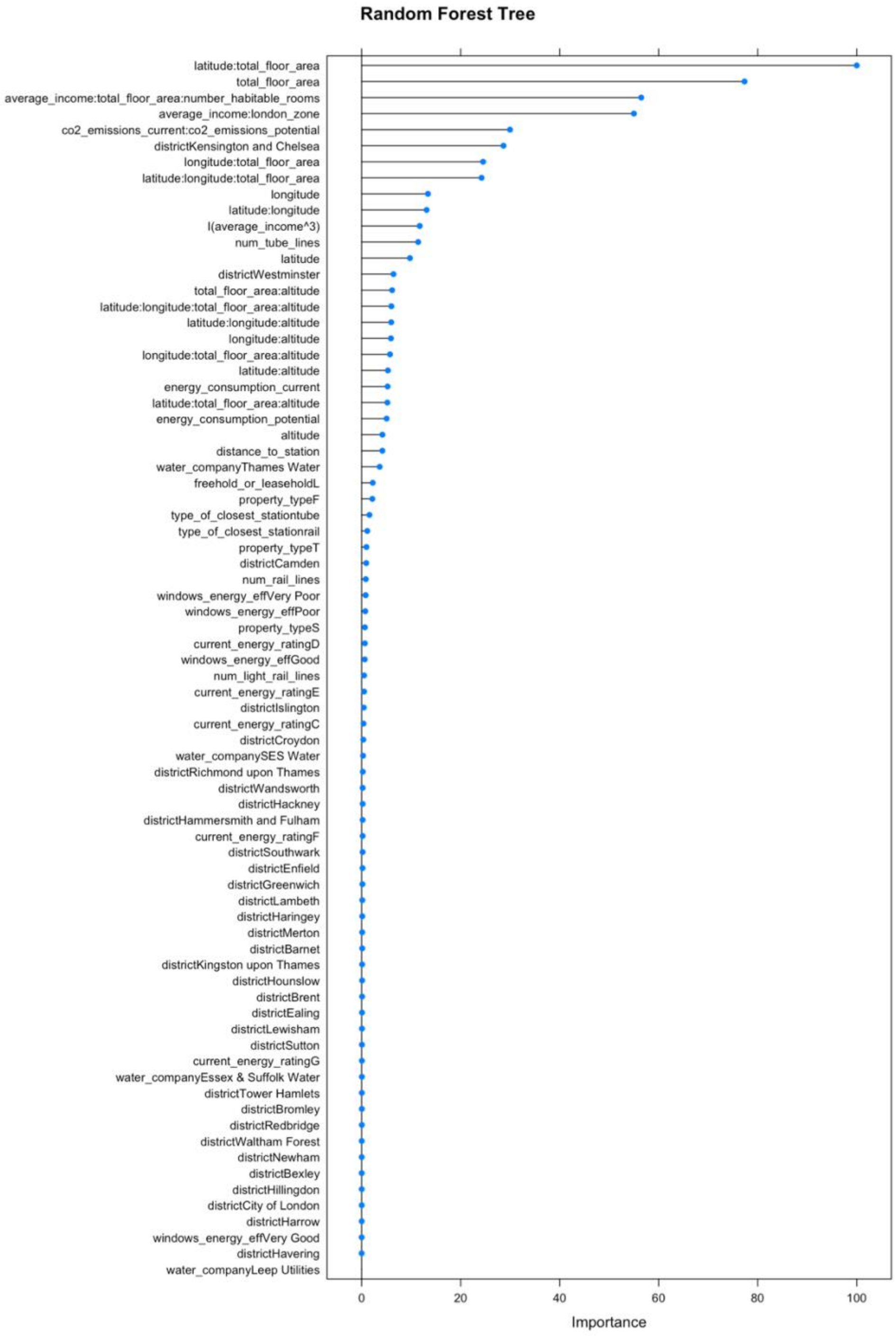

Figure 5 Random Forest Tree model variable importance Scaled

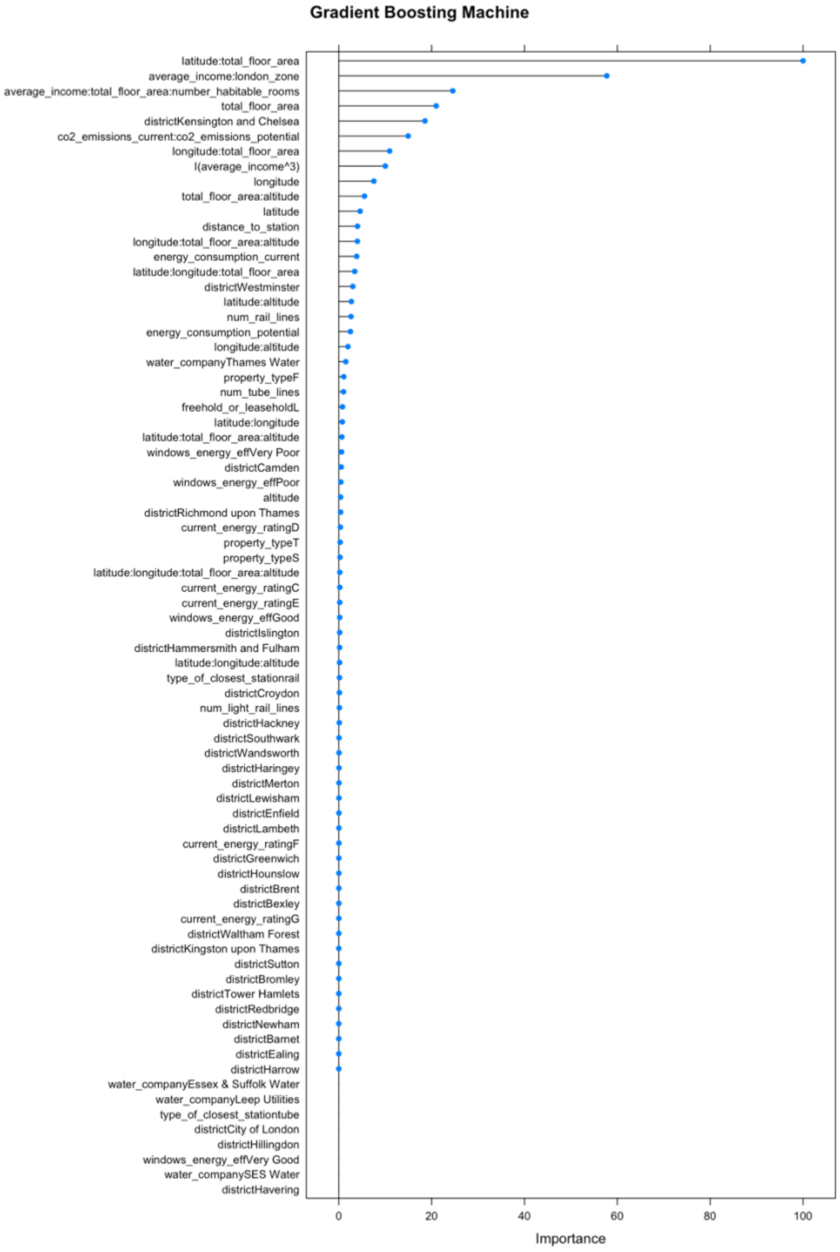

Figure 6 Gradient Boosting Machine variable importance Scaled

D.)

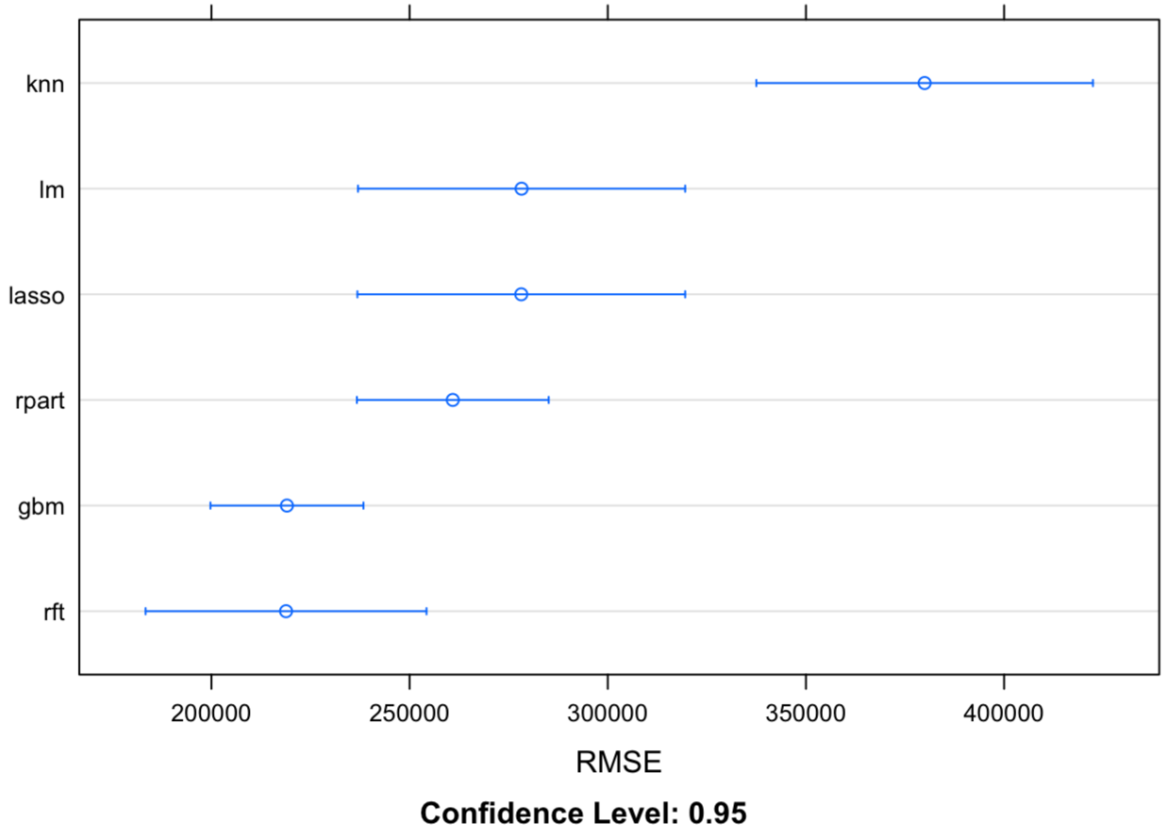

Figure 7 Stacked model component performances