The Priority List of Statisticians for Boosting Home Value

CS510: Computing for Scientists Final Report

From: Li, Shuo (Olin)

To: Professor Waldrop, Lindsay

Introduction

For many people, a house is not only a residence but also a place where they have been investing throughout their stay. Hence, how to maintain and boost home value during their stay has been a question for many house owners. Generally speaking, there are many commonly known factors that would help increase home values, however, for most people with a limited budget, it is hard to take everything into consideration when they want to boost their home values. Therefore, it is of great significance to learn what should be prioritized during the home improvement with a purpose of value boost. Although home owners are unable to obtain everything they want with a tight budget, they can do the things that really matter and bring up the sale prices of the house by prioritizing the controllable things. To obtain a priority list for house improvement and home value bringing up, statistical methods like linear regression and random forest would be utilized in this project to analyze a Kaggle dataset containing house sale prices of King County, Washington from May 2014 to May 2015. Statistical models would be constructed to find out the most significant house attributes that are related to home prices.

Purpose

The primary objective of this project is to find out a priority list for home improvement that can be helpful for boosting home values from the view of a statistician. Commonly used house attributes would be analyzed and the project would be planned to figure out the most relevant features of a house regarding sale prices. Hopefully, this project could offer some suggestions on house improvement and home value boosting for home investors during their stay.

Background

As the most populous county in Washington with the largest city of the state, Seattle, sitting in the west, King County embraces quite a few large corporations (e.g., Boeing, Microsoft, Amazon). The map below highlights the exact location of King County.

Map of King County, Washington

The strong local economy has been driving the increase in the number of households in this area, which has been creating the demand from the real estate market. The graphic below displays the right-skewed distribution of house sale prices in the time window we analyzed. The house price in King County ranges from *$75,000* to *$7,700,000*, and the average price is slightly above *$540,000*.

Data

A real-world dataset that contains house sale price information and the corresponding house features of King County, Washington from 2014 to 2015 will be used. It is

originated from Kaggle, and can be imported to R from mlr3data package. Basically, there are 21,613 observations along with 19 house features such as the number of bathrooms, bedrooms, floors, and square footage of the house in the original data.

Variables

The full variable dictionary is summarized as below:

id: unique ID of the house

date: the sale date of the house

price: the final sale price of the house

bedrooms: count of bedrooms in the house

bathrooms: count of bathrooms in the house

sqft_living: square footage of the living area in the house

sqft_lot: square footage of the lot for the house

floors: total levels in the house

waterfront: whether the house has a waterfront view. If yes, the value is 1. Otherwise, the value is 0.

view: how many times the house has been viewed

condition: the overall condition of the house

grade: the overall grade given to the housing unit by King County grading system. According to King County Assessor's webpage, this represents the construction quality of improvements. Grades run from grade 1 to 13.

sqft_basement: square footage of the basement

sqft_above: square footage of the house apart from the basement

yr_built: which year the house was built

yr_renovated: which year the house was renovated. If no renovation has been done, the value is 0

zipcode: the zip code for the house address

lat: latitude coordinate of the house location

long: longitude coordinate of the house location

sqft_living15: square footage of the living area in the house measured in 2015

sqft_lot15: square footage of the lot for the house measured in 2015

renovated: whether the house has been renovated. If yes, the value is 1. Otherwise, the value is 0.

basemt: whether the house has basement. If yes, the value is 1. Otherwise, the value is 0.

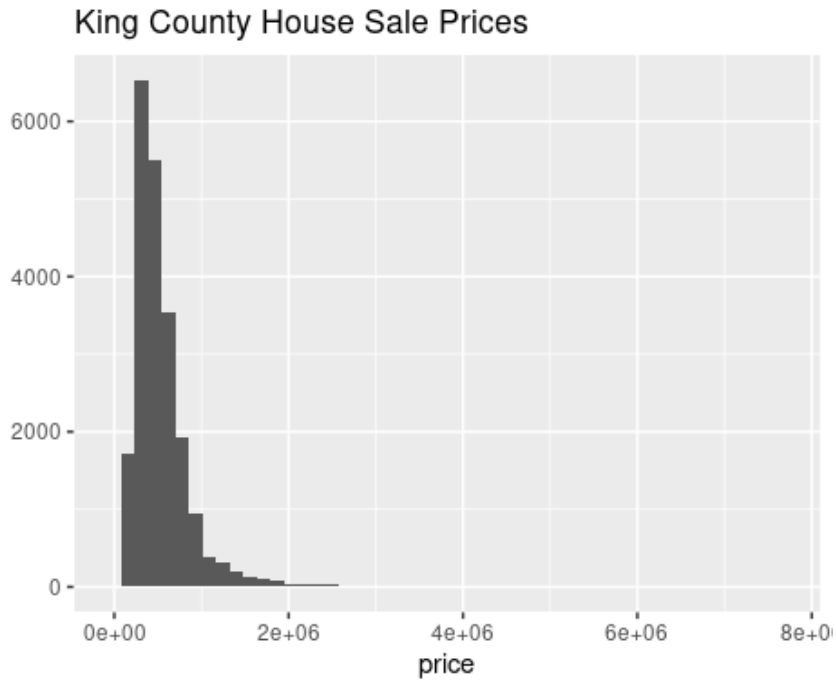
Packages

The first package that should be used in this project is mlr3data, which offers the dataset that this paper are going to analyze. Besides, this project uses the ggplot2 and lattice

packages for the purpose of data visualization and randomForest for random forest models that can be helpful to analyze the effects of the house factors on the house price. Apart from that, the stargazer package is utilized to offer neat and more readable model results of linear regressions. Another important package that can be useful in this project is GGally, in which the ggcorr function can help us obtain the correlation matrix. This report also use the dplyr package to manipulate and modify data frames. After receiving the feedback on the first draft of the project, the project need to add new packages of caret, which is used to tuning random forest models so that the model performs best regarding cross-validation root mean square errors (CV RMSEs) can be found. Additionally, the packages of knitr and kableExtra are also considered to help ones generate a neat and well-formatted final document via R markdown.

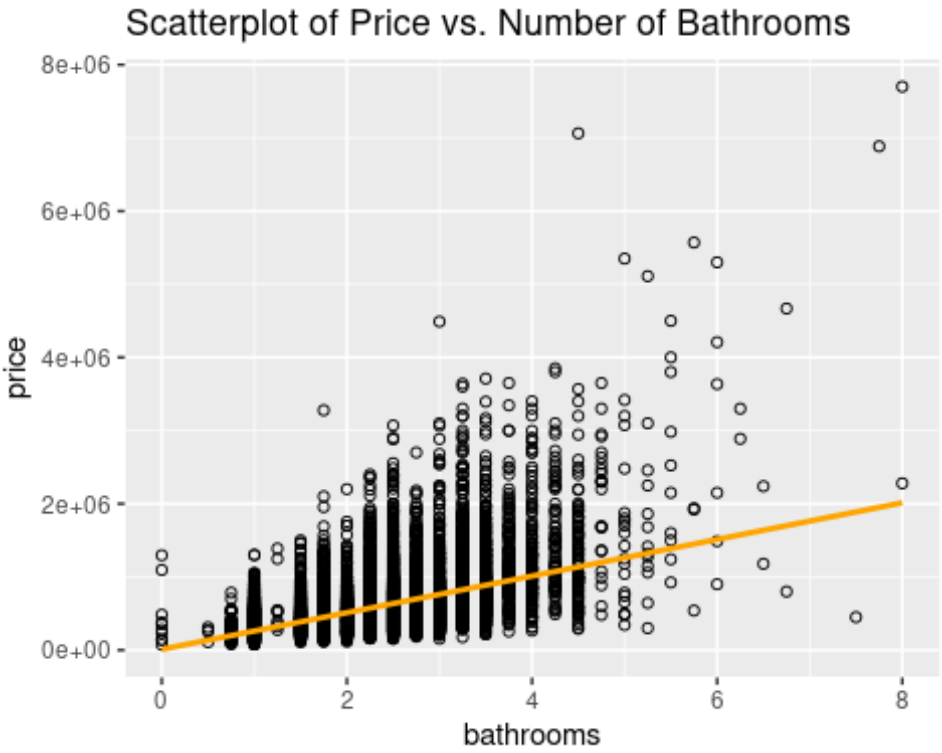
Exploratory Data Analysis (EDA)

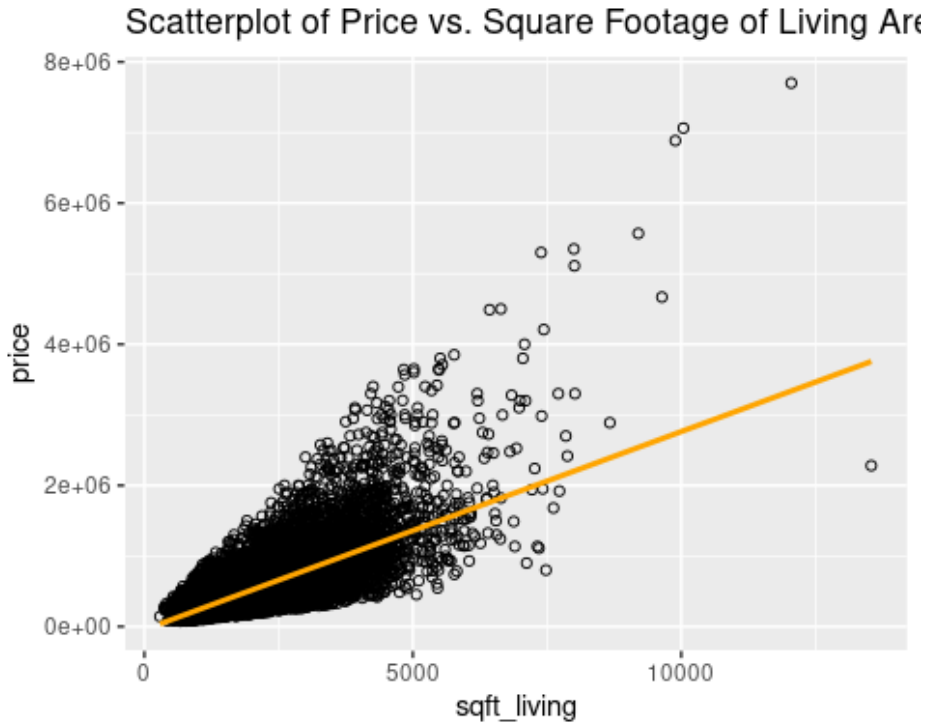
The data exploration begins with the variable of interest price.

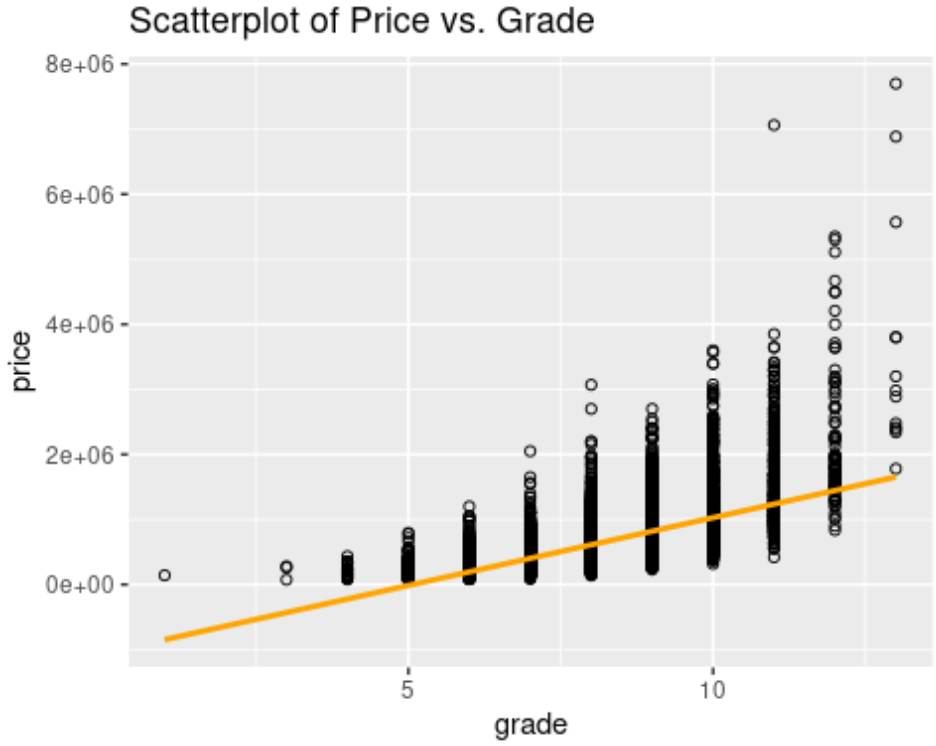


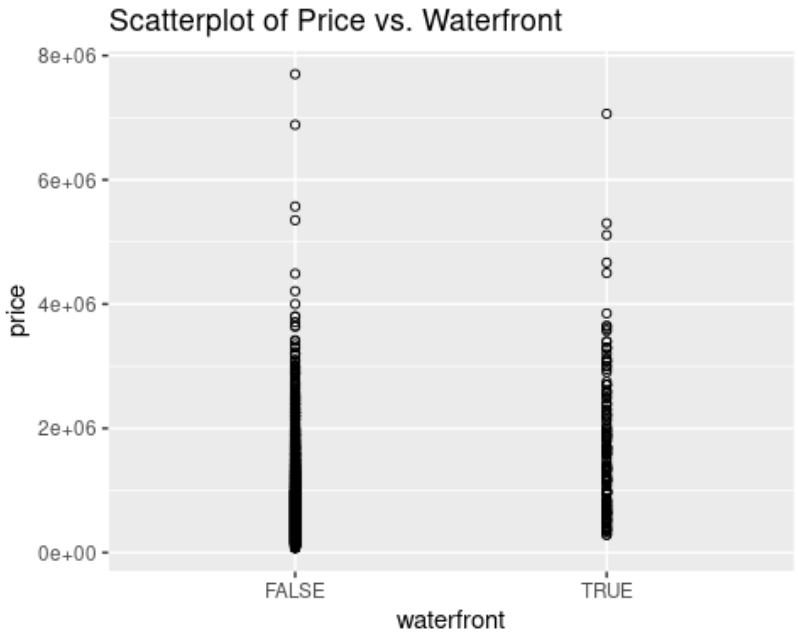
It is clear that the distribution of the home price is positively skewed with a long right tail, which implies that some houses are expected to have higehr values than others.

This project explores the relationship between the features of the house and the home prices by plotting price with each feature. Ones are able to find that some features like bathrooms, sqft_living, grade and waterfront have relatively stronger relationships with price than others.

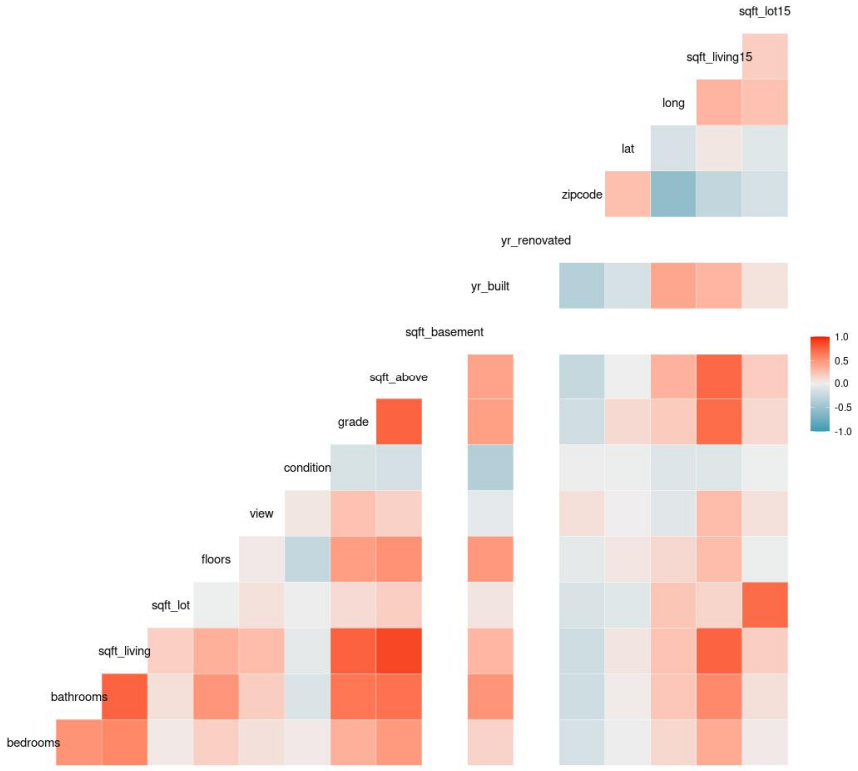



Scatterplot of Price vs. Square Footage of Living Are
8e+06
6e+06
4e+06
2e+06
0e+00
price
0
5000
10000
sqft_living


Scatterplot of Price vs. Grade
8e+06
6e+06
4e+06
2e+06
0e+00
price
5
10
grade



This project also includes the correlation matrix that reflects how variables are correlated with each other.



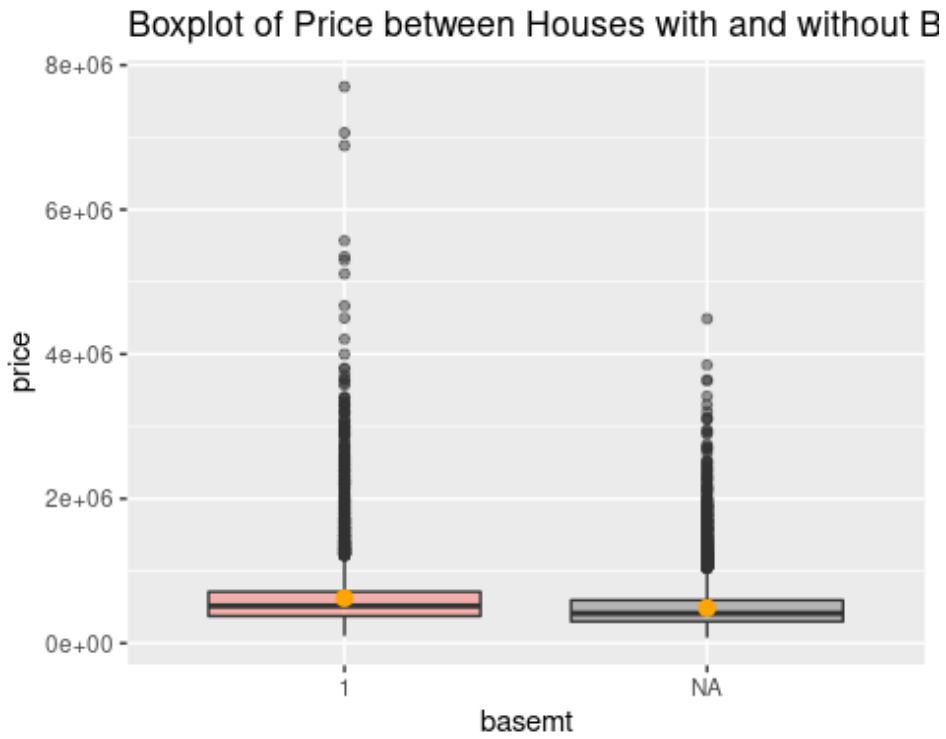
People can notice that sqft_living and sqft_above are highly correlated with a correlation of 0.8765966. This makes a lot of sense because most of living area is usually above the basement. The univariate correlation between sqft_living and price (0.7020351) is higher than that between sqft_above and price (0.6055673). Similarly, sqft_living and sqft_living15 are highly correlated with a correlation of 0.7564203. The univariate correlation between sqft_living and price (0.7020351) is higher than that between sqft_living15 and price (0.5853789). The variables that are not likely to affect the house price are considered to be removed later.

Data Modification

Based on above findings, dataset is modified by introducing two new binary features:

- renovated: Equal to 1 if the house have been renovated and 0 otherwise
- basemt: Equal to 1 if a house has basement and 0 otherwise.

This project also excludes the useless information from the original set and remain the variables that are most relevant to the house price. Moreover, the variables of sqft_above and sqft_living15 are dropped in the further analysis to avoid the issue of multicollinearity.



However, as ones noticed in the boxplot of basemt, this new variable is not likely to have significant influence on the home prices. Therefore, this is also a needless variable to make predictive model later.

Combined with the variable modification investigation with the previous EDA part, this project then creates a new dataset that contains the necessary predictors only, and summarizes the final dataset with bedrooms, bathrooms, floors, etc.

Modeling

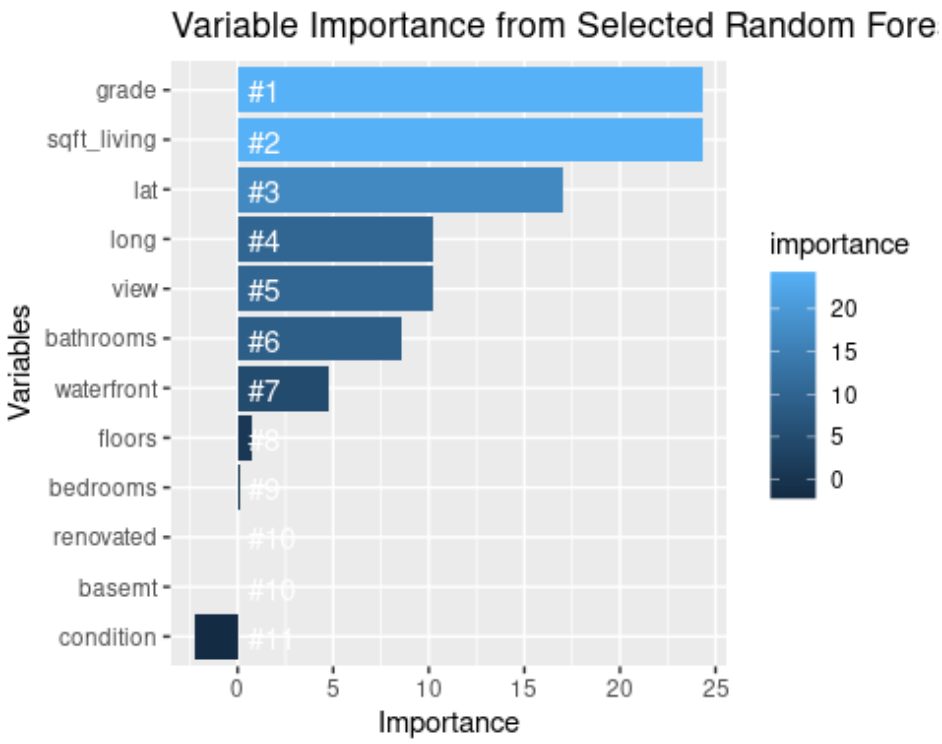
Before fitting the models, this project partitions the dataset in which 70% are used as train sets and 30% are used as test sets. In order to make our results consistent every time ones run the data, a seed of 913 is set to remove the sampling randomness. Then the project uses the training set house_trn to tune the models and reporting the cross-validated

errors measured as RMSEs. Here, this project uses a 5-fold cross validation. The latter part of the project constructs the functions for the modeling procedure.

Random Forest Model

In the first place, this project trains random forest models using all the predictors in the house dataset with price as response variable. The default tuning parameters chosen by the caret package would be used.

The best random forest model with scaling uses 7 for mtry, meaning that 7 features will be randomly chosen every time a tree is grown. The number of trees is 500 by default. People can find that about 88% of the variation in prices can be explained by this model.



It turns out that the grade (ranging from 1 to 13), which represents the construction quality, is the most important variable regarding house sale price. For houses only meet

the minimum building standards, their grades are low and ranged from 1 to 3, and their prices are expected to be the lowest on average. For houses that have achieved average performance in terms of construction and design, they are graded as 7, and their averaged prices are expected to be moderate among all houses. As for the houses graded over 12, they are thought to have excellent designs and use the best materials while construction. As a result, their prices are expected to be highest too.

sqft_living, the spacing of the living rooms, turns out to be the next most important feature that is related to home values. The further next significant factor that is highly correlated with the home prices is the location, consistent with the latitude and longitude ranked as third and fourth important variables.

Additive Linear Regression Model

Although, ones learned the effects of location on house price are likely to be significant, this project are not going to interpret them in the multiple linear regression model since location is not a factor that could be changed for home owners after the house being purchased. Hence, in this linear model, only the variables of grade, sqft_living, view, waterfront and bathrooms will be included.

Based on the outputs of R codes, the estimated model is

$$\widehat{logprice} = 0.249grade + 0.0001sqft_living + 0.038view + 0.091bathrooms + 0.183waterfront + 10.976$$

The R-squared of this model is 0.728, which means this linear model containing the most relevant predictors could explain about 72.8% of the total variation in the house prices. The F statistic of this model is 0.314 (p = 0.000), which means the overall model is

statistically significant. In addition, at 5% level of significance, ones should notice all predictors are statistically significant individually.

As noticed, here project uses the log-transformation on the dependent variable because the variable is right skewed based on previous analysis. Hence, the model results imply that one additional point in the house grade is expected to increase the house price by 24.9% on average when other factors are assumed to be the same. What's more, the house with a waterfront is expected to be 18.3% higher in price than the house without a waterfront when other conditions are the same. An additional bathroom in a house is expected to bring up the home value by 9.1%, and with a one-sqft increase in the living room is expected to enhanced the house value by 0.01% on average conditional to all other factors respectively. As for the time of the house being viewed, it is shown that each time the house is viewed, the house value is expected to be boost by 3.8% on average.

Model comparisons

This project also compares the predictive performance of the two models. The reported test rmses for random forest model is 322215.6 and for the linear regression model is 351018.4. That means the best random forest model we found by caret performs better in predicting than the linear model. Therefore, ones are capable of deciding the best random forest model as a better model that can help to predict the house values.

Limitations

The first limitation of this project comes from the fact that some unchangable external factors such as the location of the house and the real estate market such as unexpected

economic crisis could impact the house price unwantedly, even if the priority list is completed by the owner.

Besides, it is quite time-consuming to tune the parameters in the random forest model, which limits attempts of a wider range of possible values in the parameters should be tuned. That means, the random forest model may still be able to be improved if statisticians can take more time and tune more models.

Thirdly, this project only considers two methods in predicting the house price so there may be better models using different methods.

Last but not least, statisticians should notice that the omitted variable biases may still exist, and there are other potential factors that are not included in the data may also affect the home values significantly.

Conclusion

In conclusion, from the view of statisticians, a priority list for boosting home value should contain the following two things:

1. Try to maintain the house in a good condition and add more custom design so that it can be graded higher.

2. Try to expand the space of living room in the house.

If ones want to use the easily collected house features to predict the house prices, a random forest model with 7 features being selected per tree and considers 500 trees at one time is recommended. Although it can be more precise, it can offer a general idea of how valuable the homes would be for the house owners.

References

Andersen, P. K., & Skovgaard, L. T. (2010). *Regression with linear predictors*. New York, NY: Springer.

Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science*, **16**, 199–231.

Breiman, L. (2001b). Random forests. *Machine Learning*, **45**(1), 5– 32.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statisticians. *Biometrical Journa*l, 60(3), 431-449.

Kaggle.com. (2019). Retrieved from https://www.kaggle.com/search?q=house+features+of+King+County%2C+Washington+from+2014+to+2015 [accessed 06 October 2021].