The Priority List of Statisticians for Boosting Home Value

CS 510 COMPUTING FOR SCIENTISTS

Li, Shuo (Olin)

DECEMBER 13, 2021

RELEASE: https://github.com/Olin1314/CS510_Midterm.git

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.

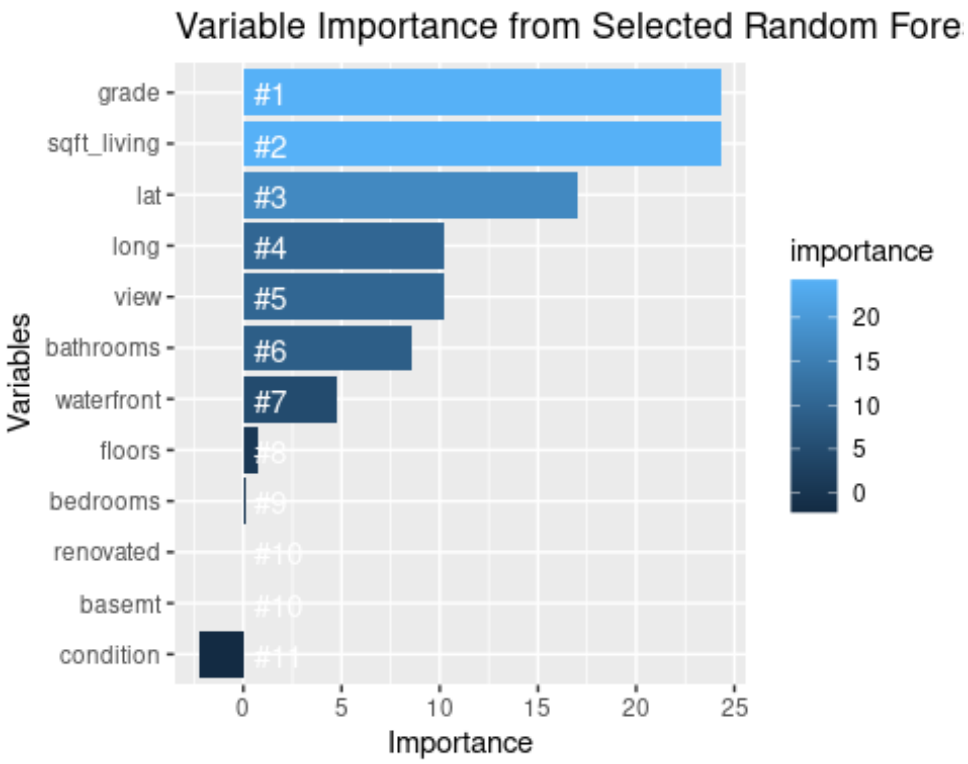
Purpose

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.

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.

Random Forest 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.

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].