FIN 510: Final Project

EXECUTIVE SUMMARY

Due: Upload to Compass by 11:59pm on Friday, December 3, 2021

Your Team Name (be creative): Big Data Survivors

Select whether this is an individual or group submission. **No more than 3 members per group.** Beyond the fact that all group members may submit the same answers, each submission must be separate work.

☐ Individual Submission
☐ Group Submission. Group member names: _Jason Fu, Ivan Abarca, Aizhan Kassym

Case Overview

A large portion of the local government's revenue is derived from property taxes and the revenue generated from these taxes serve important functions in the community. It is the revenue generated from taxes that are used to improve and maintain infrastructure, fund schools, police, firefighters, and pay pensions among other public good programs and projects. As a result, it is imperative that property taxes are fairly collected. The Cook County Assessor's Office (CCAO) is in charge of valuing over 1.8 million properties, equating to around $15.58 billion in tax revenue. However, CCAO had a history of incorrectly valuing property (and thus collecting the wrong amount of taxes) so severely that the public demanded transparency and reform that eventually came in the form of a newly elected assessor. As Fritz Kaegi took office, he made all of CCAO's methodology and data public. Our objective is to use the public data to predict the value of a home as close to its actual value as possible. To do this, we would have to utilize the historic_property_data.csv to train models that predict the value of homes listed in predict_property_data.csv with a low MSE.

Methodology

As historic_property_data.csv contained 50,000 entries with 63 variables, our first step was to trim the data. To do this, we used linear regression to determine which variables were statistically significant and only use the statistically significant variables in our models. It is important to note that most variables are categories, characters, or logical values, instead of numerical and therefore factor() was applied to those variables so the models did not interpret those numbers as if they had numerical relationships. For the code to run more efficiently, several geographic indicators that appeared less than 500 times (1% of the dataset) were merged into a single group called "distinct". In an attempt to make the code run more efficiently, we thought of reducing the number of variables so that only those variables with a J Column of "TRUE" in the codebook.csv would be included. However, upon preliminary testing, it was found that excluding variables with FALSE in that column actually increased MSE quite significantly.

Originally, we planned on training four different types of models: stepwise regression, lasso regression, random forest, and boosting. The idea was to use the model with the lowest MSE to predict home value in predict_property_data.csv. However, due to the number of variables we kept in our models, stepwise regression became prohibitively resource consuming (4.5 hours), which combined with the fact it never had the lowest MSE even during our reduced variable test codes, we have decided not to pursue it in our final code. Random Forest ended up being the model with the lowest MSE, which is most likely due to the fact that random forests are able to find non-linear relationships along with being low bias and only moderate in variance.

In our random forest code, we have decided to use a mtry of 8. Mtry is the number of features to consider at each split point and is supposed to be the square root of the number of variables. Because we had around 63 variables, we chose an mtry of 8. The result is an MSE of 12384452558, which is much lower than that of boosting's 19308120278.

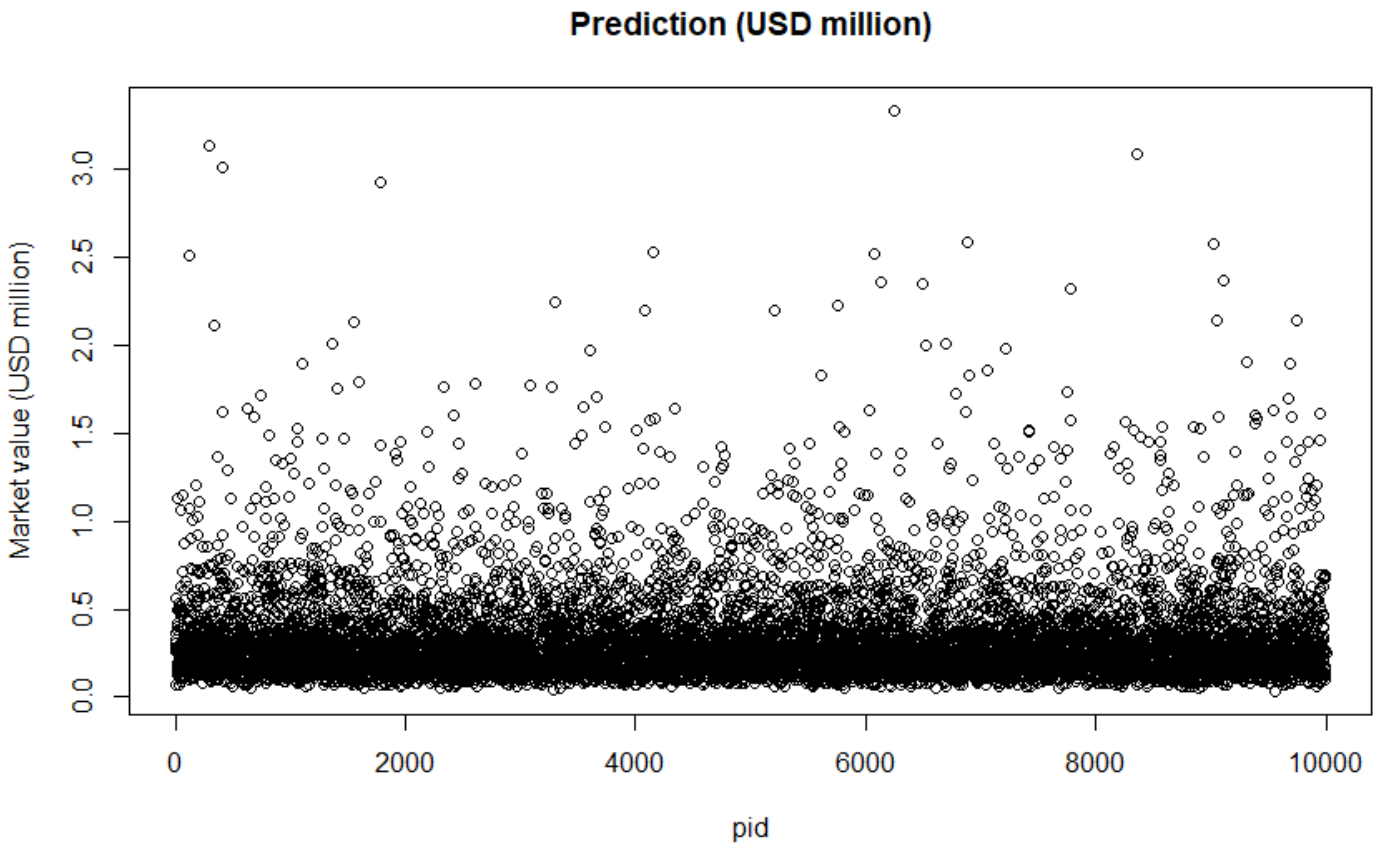
Conclusion

Our summary statistics show a min of 36455.3, a max of 3329663.3, a mean of 325433, and a median of 252315.1. The quartiles are 175100.7 for the 1st quartile and 374530.1 for the 3rd quartile. These figures all seem to be in line with our knowledge of the housing market and house prices in general, which gives us more confidence in our model's accuracy. With more computing resources, we would recommend more trees in the random forest model but not too many because random forest models have a tendency to overfit if too many trees are used. Our assessed_value.csv file includes 10,000 sequential PIDs, each with a non-negative and distinct assessed value, thus matching the requirements of the project. It appears that the relationships between the variables in their predictive power has non-linear properties, which is why the random forest model was far superior in terms of MSE.

	X..pred_result.assessed_value
nobs	10000.0
NAs	0.0
Minimum	36455.3
Maximum	3329663.3
1. Quartile	175100.7
3. Quartile	374530.1
Mean	325433.0
Median	252316.1
Sum	3254329537.8
SE Mean	2587.5
LCL Mean	320360.9
UCL Mean	330505.0
Variance	66952958293.8
Stdev	258752.7
Skewness	3.3
Kurtosis	18.2

Appendix

Graph 1: PIDs and their market value



Graph 2: PIDs and their market value distribution