Tianyuan Deng

APAN5200

Nov 29, 2021

Kaggle Report

Summary

In this project, I used the random forest for the final prediction. Along the way, I also explored linear regression, decision tree and xgboost. Random forest provided me the lowest RMSE. The model could be further improved by using different models or more sophisticate data cleaning techniques.

Explore data

Overview of the dataset

The dataset contains 91 columns, and it is untidy. When I started to analyze the data, I realize: several variables, such as id, name, are unrelated to our prediction therefore they were excluded by intuition; several variables, such as summary and description, they provided certain meaningful information but they are extremely complicate and random, therefore they were excluded as well for simplification purpose; some variables, such as neighborhood and street, they are interrelated and provides similar information, therefore they should be grouped in a way that they do not occur in the final prediction repeatedly. After a brief filter on data, approximately 30 variables are omitted.

Clean missing value

For the remaining variables, I noticed there are many missing values in most of variables. For all the numerical missing values, depends on the meaning of variable, I replaced the missing values (NA) with 0 or mean. Such a replacement seems logical, but for a few variables, the proportion of missing values are high, I realized afterward that this method distorted data distribution and jeopardized the prediction. An alternative way of cleaning missing values was replaced them with random value from the same column: this method proved to be more useful later on.

Test for correlations

After excluding unnecessary variables and cleaned the data, I started to visualize the data and tried analyzing the relation between each variable and price. A few variables, such as accommodate, bathrooms, bedrooms, have obviously relationship by intuition; for other variables, such as minimum nights, review scores, their relationship with price remain unexplored. I used function cor() to test variables correlation with price and I kept all the high correlation variables for further analysis.

Build model

Decision tree

Decision tree can predict with both numerical and categorical value and it provides numerical outcomes. In my opinion it is also the model that easiest explain to people. After I cleaned all the data by replacing with mean or 0, I started with the decision tree model. I input all the high correlation variables which I found in previous steps. I split my dataset into train and test as 0.75:0.25. On my testing data, my first RMSE score was around 82. After I built my first tree, I performed tree control to adjust parameter because decision tree model tended to be overfitting. I

checked different value for min-split, mini bucket, max depth, and cp. I also tried remove variables one by one to see if there is an improvement on my RMSE score. After experiment with different parameters and variable combinations, I tried to use my tree in the scoring dataset. When I performed my tree model into the scoring dataset, an error occurred said 'new levels occurred in zip code and property type'. To proceed with my tree into the scoring data, I deleted these two variables which was a huge mistake.

Forest

Theoretically, a random forest model consists of many decision trees, and it should generate more accurate predictions. I performed the random forest model using the same variables as previously. However, RMSE decreased only by a small amount. It was slightly unexpected, and I questioned about my method of selecting variables. I performed the forward selections on selected variables then passed into the model. Unfortunately, forward selection did not help much about decreasing RMSE. The running time of forest was much longer than tree as well. I decided to shrink the size of forest and I created a 'mini' dataset which only contains 2000 data points for testing purpose. Using a smaller dataset, I found a variable combination with slightly lower RMSE.

Attempted with xgboost

I input the variables from decision tree into xgboost model. Unfortunately, the RMSE score increased. At this point I realized I could not achieve a lower RMSE score not because of model selections, it is more relevant to how do I select my variables. Given the running time of xgboost is long and I can not try all the variables one by one. I created a linear regression model and I used the R-Square value for each variables from that. Since the xgboost model can filter not useful variables, I included all the variables which have significant R-Square value. Another

change I made was when I generate the model for final prediction, I included all data (not just training, testing included as well). I also reflected on my data cleaning methods. Instead of replacing missing value with mean or 0, I replaced the missing values by random value from the same column. In addition, I perform the same data cleaning technique on the scoring data as well before they were used in prediction.

Final step

With new data cleaning methods, I used the new dataset into all models. I found the RMSE is the lowest in random forest model. After I discussed with my friend, I realized zip code is crucial for prediction I should not left out. After including zip code into the random forest, I achieve my lowest RMSE score. Unfortunately, I did not have enough attempt and time to incorporate zip code into xgboost model.