Kaggle project report

Shiyu Hua

This report aims to analyze and predict the rental prices in New York using data from Airbnb on renter, property, and reviews. After loading the “analysis data”, which is the training data for the project, the first step I took was to explore the variables in the dataset. I skimmed the dataset and found out there are 50 non-numeric variables and 41 numeric variables, including the dependent variable “price”. I explored the dependent variable “price” and graphed its distribution using ggplot package. Figure 1 shows the distribution of the price variable. As we know that price cannot be less than zero, so I removed all price less or equal to 0 when prepared the data. I also found that not all variables were complete and a lot of them had missing values. The two numeric variables I pay attention to were the number of beds and the cleaning fee. I dealt with the missing values when prepared the data. My background knowledge about Airbnb showed that the pricing of Airbnb comes from variables such as accommodate, bedroom, bathroom, and amenities. The top 10 amenities guests often search include wifi, air conditioning, heating, washer, TV, kitchen, and etc. (Airbnb, 2020). While in the dataset, the amenities in the property are included in a single column. I did some variable transformation so these amenities variables can be used in analysis.

After exploring the insights from the data, I pre-processed the data. Firstly, I filtered the data to only include properties with price larger than 0. Then, I dealt with missing values in the columns. I assumed all the “NA” value in the cleaning fee to be zero, so I changed all missing value to zero value. Then I used median impute method from the caret package to fill in all the missing values in the numeric variables. These variable transformations went well, but I encountered issue when I tried to extract words from the “amenities” column. I tried to split the words in the “amenities” column so that if a specific word like “Wifi” was in the column, then I can create a new column showing the true or false for that specific amenity. However, at first this method did not work and it showed all false statement. Therefore, I researched online and found another way to do it. The function I used is grepl(). It is a built-in function that searches for matches of a string and return TRUE if the string contains the specific pattern (in this case, a word), otherwise FALSE (RDocumentation). So, I added variables of amenities using this method.

After the data was ready, I first considered using the multi-variables linear regression model. It is commonly in use to predict the value and easy to run. Then I found out that for some categorical variables, not all the factors were significant variables in the linear regression. So, considering not all variables are linearly (either positively or negatively) related to the price, I switched to regression tree analysis and used bagging and random forest method with tuning parameters. Due to the complexity of the advance regression tree analysis, it took a long time to run the codes.

The variables I considered are based on property characteristic, renter information and review scores. For the property characteristics, the variables in use included neighborhood, room type, accommodates, bedrooms, bathrooms, beds, popular amenities, average minimum nights, and availability in 365 and 90 days. For the renter information, I chose the number of listings for each room type and cancellation policy. For the review scores, I included number of reviews and the scores for rating, accuracy, cleanliness, check-in, communication, location and value.

My best analysis technique was the tuned random forest. I used the 31 variables I chose and tuned the hyper-parameters mtry, splitrule and min.node.size with 10-fold cross validation using the caret framework. Then after having the best tune of hyper parameters, I fitted a random forest model using ranger method and generated 500 trees. RMSE value is the performance measure and the lower the RMSE value the better the predictions. This model had the lowest RMSE value in all my models.

I ranked 424 in the Kaggle ranking and it showed that my model prediction was not very powerful in the competition. The top three impactful variables are room type, neighborhood group, and accommodates. To improve my result, I could spend more time and effort in exploring and preparing the data. There are 90 variables in the dataset and I only used about one third of them. I did not fully explore the whole dataset and utilize all the possible variables. I chose to neglect some variables because I did not think of a way to incorporate it. Property type variable was one of the examples, error occurred in predicting in the testing data because there are new property types in the data, so I removed the variable from the analysis. I also could do better with dealing with the missing value, instead of using median impute, I could use bag impute to have more accurate values for variables like square feet. There are too many missing values in the square feet variable and the median impute might not be accurate so I did not add square feet variable in the analysis, while I still think it might be a useful variable. Last but not least, I could try other models such as boosting model and compare the performance of the models.

It is an interesting and meaningful project, more than simply predicting the rental pricing in New York. I not only learned several technical ways to analyze the data and make predictions based on the models, but more importantly I realized the importance of exploring the dataset and pre-preparing the data before any analysis. I learned not to rush to analyze the data and the result of models would not be powerful if no exploring and preparation of data. The analysis technique modeling is not the difficult part of the project, there are limited methods and lots of built-in functions to implement easily. What makes the prediction different is the preparation of data, to explore them so they are powerful during the analysis. For the next time doing the similar project, I will devote more time to explore and prepare the data.

Reference

*The best amenities to offer right now - resource center*. Airbnb. (n.d.). Retrieved December 13, 2021, from https://www.airbnb.com/resources/hosting-homes/a/the-best-amenities-to-offer-right-now-203.

*GREP: Pattern matching and replacement*. RDocumentation. (n.d.). Retrieved December 13, 2021, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/grep.

Appendix

Figure 1 The distribution of variable price in the analysis data

