APAN 5200: Applied Analytics Frameworks and Methods I

Kaggle Report

Columbia University

Haoyu Wang

Initial Exploration

In the Kaggle competition of predicting rent Price for Airbnb in New York City, tasked with conducting the best prediction model, I had worked with **90** independent variables and **41322** observations.

Data exploratory and data cleaning was the essential part in my work and cleaning different types of data required a lot of work. By filtering out variable which were clearly not relevant to the dependent variable by personal experience was and human sense such as id, name which had no meanings to predict rent price was the first step of my journey in this competition. After I removed all the irrelevant variables, I started to build a thorough understanding of the data. I started to look at the distribution of the dependent variable, price. I removed the outliers in price which had zero values. And then, it was time to begin analyzing the relationships between price and all the independent variables.

I implemented several data cleaning methods to clean up the variables with NAs, blanks, and outliers. Also, some of variables contained different data formats such as zipcode, so I also did data tiding for such columns. Firstly, I installed visdat package and studied the missing value. I replaced the NAs in the host_total_listings_cout column into 0, I replaced the NAs in the bed column into 0, I replaced the NAs in the reviews_per_monthcolumn into 0, but I replaced the NAs in the square_feet column into median, and I replaced the zipcode which contained characters were not equal to 5 to Other. After the initial data cleaning for some numeric variables, I implied skimr package's summarytools function to examine missing values and outliers. Since there were some missing values in the Boolean type columns, I replaced missing values in host_is_superhost to TRUE, I replaced missing values in host_has_profile_pic to TRUE, I replaced missing values in host_identity_verified to TRUE, and I replaced the missing values in zipcode column to Other. Lastly, I checked the numbers of the missing value of the selected and leaned analysing data set. The result showed 0 missing values.

For numeric variables, I isolated the numerical variables from the whole data set
I implemented several feature selection methods. I used VIF to visualize the variables to identify the multicollinearity among those variables, and I found availability_30, availability_60, and availability_90 had serious multicollinearity. I also applied the best selection and it showed that 18 numeric variables should be used among the 20 numeric variables, which indicated that the availability_30 should be taken off for the best selection for the numerical variables. After that, I tried forward selection, backward selection, and stepwise variable select to filter out the numerical variables for the best combination. As a result, all those methods showed a same result that the availability_30 made the model worse. Therefore, I took off the availability_30 column for the dependant variables.

For categorical variables, I realized that property type in analysis data contained two factors that scoring data did not have. Therefore, I replaced the propter type of Lighthouse and Timeshare in analysis data to Other.

Models and Feature Selection

I have tried several feature selection methods such as subset selection including Best Subset Selection, Forward Selection, Backward Selection, and Stepwise Variable Selection for numeric variables. In addition, I used Rondom Forest and Xgboost model to predict the rent price.

Before I used any feature selection method, I manfully picked up the numerical variables that I thought was relevant to the result, and I used linear regression to make the predication. However, it only gave me around 92 RMSE. For best subset selection, the best subset selection is very common and popular, so I wanted to get the best combination of the numerical variables. According to cp, bic, and adjr2, it all showed that availability_30 should be in the best performing model. For forward selection, backward selection, and stepwise variable selection, I just wanted to test if there would be a different result compared to the best subset selection. However, they all gave me the same result that availability_30 should not be included. After I selected the best performing model for numerical variables, I used linear regression to predict the model. But the result was not good enough, so I continued to find a better model. I observed the categorial variables and filter out those were not relevant. After I cleaned up and reorganized the categorial model, I use random forest to predict the rent price using both numerical and categorical data. The RMSE from the random forest method only returned me 78.83 RMSE, so I realized that the model still could be improved. I changed the ntree several times to see if there would be a change, but the RMSE only decreased to around 74. At this moment, I thought the model should be changed if I want to see a big improvement on RMSE. Therefore, I decided to use XGBoost model to predict the rent price since XGboost is an optimized boosting libaray. However, XGBoost only accepted numerical type data, so I made matrix for those categorical columns for both analysis data and scoring data to convert them into dummy variables. Then, I applied the XGBoost model to make the predication. Finally, by adjusting the nround, I got my best result which that RMSE is 65.58 for the Public Score and 59.33 for the Private Score. In conclusion, **XGBoost was my best performing model for this competition.**

Model Comparison

Model from previous section	**Public Score**	**Private Score**
Linear Regression 1	92.668	87.33889
Subset Selection	91.93412	85.86453
Tree	74.90603	70.22604
Random Forest 1	74.26522	69.69814
Random Forest 2	74.0706	69.69814
Xgboost	**65.58107**	**59.333**

Discussion

I spent most of my time on data exploration and data cleaning in this competition. My best performance model was XGBoost and brought me 59.33 for RMSE. XGBoost is highly efficient, flexible, and portable, so it worked the best in my work compared to other models. For my opinion, XGBoost was the top performing model for the 'clean-up' data by what I have learned so far.

Future directions

If I had more time on this competition, I would like to spend more time on exploring the categorical variables since categorical variables are more complicated and unorganized. For example, for the amenity's column, I would like to divide each amenity into different columns to made prediction. I believe that exploring categorical deeply can benefit the prediction result.