Kaggle Report: Predicting Airbnb Rental Price

**Summary:**

This project aims to predict Airbnb rental prices in New York using data on renters, property, and reviews. The dataset contains 38513 observations on 91 variables, and the result was evaluated by RMSE. This report illustrates the overall process of data cleaning, variable selection, and model building, hence indicates where I could improve in the future.

**Brief Exploration:**

The analysis data has 38513 rows and 91 columns. The dependent variable is price, and the independent variables vary from character, factor, and numeric variables. After a brief exploration, I found that those variables could be divided into five categories:

Location information: neighborhood, neighborhood group, etc.

House information: amenities, room type, beds, etc.

Host information: host response rate, host listings, etc.

Reviews: number of reviews, review scores, etc.

Other: weekly price, notes, etc.

The variables within the same category mentioned above may have a higher correlation and result in multicollinearity, so I decided to draw the correlation matrix to see the relationship among different variables. Then I imported the dataset into RStudio for further exploration.

**Further Exploration and Cleaning:**

Since we are using the scoring dataset to predict, we should perform the same transformation to scoring data as we did to analysis data. Therefore, I combine the two datasets into one set to make the cleaning process easier. Then I first drop the rows with price = 0, which are useless, and some other variables with too many N/A or not so important based on manual selection. Then after testing the correlation, I dropped bed from the dataset since it has a high correlation with bedrooms and accommodate, and could be replaced by bedrooms only.

After that, by looking at the table and smooth ggplot, I found out that after about 6 bedrooms, the observations were limited, and the impact on price was not stable, so I encoded all bedrooms greater than 6 to 6. Similarly, I did the same transformation for the other 5 variables like property type, bathrooms, and guests. I consider this step has a positive impact on my result since it reduces the impact of some extreme values on the coefficients, which can make my prediction more accurate.

For variables involving text like notes and summary, I counted the length of them based on my hypothesis that with longer descriptions, hosts consider this house more valuable. Hence, the price would be higher. However, only 2 out of 6 variables are significant through the variable selection, with coefficients really close to 0. I consider this a failed step, and I should have found a more appropriate way to deal with these texts.

Besides, I converted the date of first/last review and host\_since to numeric variables, which the number represents the count of days till now. Also, I divided the original amenities column into multiple new columns containing dummy values, like Gym, Shampoo, Microwave, and variables with less than 1000 observations that were dropped here. I believe these two steps benefit me a lot since most of them play a significant role in the final model, especially those amenities.

After all these processes, I imputed missing value with the median and saved it in another excel sheet for further use. However, there were another two missteps: first, the column of price was imputed as well, since I created an empty column of the price for scoring data. If it was imputed, I couldn’t separate the data later using price=N/A. Then I delete the price for scoring data manually in the output excel. Also, the cleaning process of amenities was not that successful. Though I thought I tried to select unique columns, there were still duplicate columns with an underscore, like “toilet” and “\_toilet.” Then I also delete those duplicate columns with underscores manually in the new file.

**Feature Selection and Modeling:**

After cleaning the data, I began to select variables for the modeling process. I planned to use both stepwise and lasso to select variables. However, since I divided new columns for amenities, there were over 100 variables in total, and the stepwise method took too much time to run and didn’t perform well. Therefore, I use only Lasso to deal with feature selection. Finally, 54 variables were selected for the final use, and about half of them were amenities, like "Elevator," "Dryer," and "Doorman."

Then I created new columns for dummy variables, and split the dataset to analysis and scoring data again. Further, I divided the analysis data into train and test set to help me choose the best model. After testing many times, I found that the linear regression model and lasso regression model were unable to further improve prediction accuracy after a certain point (around 65 to 70). Also, it was too time-consuming to run the random forest with a large number of trees. Therefore, the Gradient Boosting model was chosen to be my final modeling method.

Through the modeling process, I have tried different combinations of variables to build the model, as well as various combinations of parameters. The parameters in the code of appendix were not what for my best model but were the only one left now, since I have tried too many times and couldn’t save every single model. The result shows that the price of Airbnb in New York was primarily influenced by amenities, host services, and review scores. What confused me was when I use all my variables in the analysis set instead of those selected variables by mistake; the results seem to be better. Then I would explore the reason behind this in the future.

**Discussion:**

Three steps contributed the most to my results. The first is combining levels of some variables after a certain threshold, like combining all bedrooms over 6 to exactly 6. It reduces the impact of some extreme values on the coefficients, which can make my prediction more accurate. Also, the conversion of date to the count of days, like how long did the host joined Airbnb, also played a significant role in my final model. What’s more, the most important part was to divide amenities into separate columns, and nearly half of the selected variables were amenities.

While there also exists many failed and missteps. For example, the count of length on the text didn’t seem that useful. Also, the imputation using the median and cleaning process of amenities were not that successful, so that I had to do some cleaning work manually. The method of trying different parameters or running random forest models was too time-consuming. Also, there still exists some overfitting problem among the final models.

If I had to do this over, I would first find a better way to deal with the text, like count the appearance of specific keywords or perform sentiment analysis on the text. Also, I didn’t use the information of location that much, and I do consider zip code and neighborhood are important variables that worth further manipulation. Besides, the imputation using the median was not that accurate when facing too many missing values. I would try to impute the missing values of variables like weekly price using predictions from other variables to see if this would work better.