Yichang Chen, Yuanqi Gai, Rui Han, Jiahe Hou, Dingran Lu

NYC Airbnb Operations

1. Motivation and goal

Airbnb stands out as a popular choice in tourism and residential market in recent years, becoming the main competitor of hotels and regular rentals. Meanwhile, debating on whether Airbnb hosts only do short-term renting with Airbnb itself claiming its hosts just occasionally rent the homes in which they live becomes more intense. Since many cities or states have legislation addressing this issue, we are initially curious about whether Airbnb's claim is valid. Besides, it is meaningful to explore the overall customer satisfaction and to figure out which area people choose the most, which can give both hosts and customers better guidance location-wise. Furthermore, we think price prediction can be helpful for customers to see the reasonability of the price of a listing and whether it is overpriced, as well as helping new hosts to set price properly. Having these questions, we decide to focus on Airbnb in Manhattan, NYC. Inside Airbnb (http://insideairbnb.com/about.html) provides a public data source about city-wise Airbnb's listings. In this project, we use both reviews and listings dataset in NYC to achieve our goal.

2. Technical aspect

Data visualization:

In this section, we mainly use seaborn, matplotlib and folium packages to perform data visualization. Heat map gives a general visualization of listings in Manhattan. To answer the question about whether houses and apartments are only rent to short-term customers, the histogram for the number of listings with respect to days available in a year gives us a clear idea, where we conclude that there are a high number of listings available for the whole year, which is inconsistent with what Airbnb claims. The distribution of listing price grouped by the number of bedrooms provides a general picture about the relationship between the number of bedrooms and the listing price, which is what we think initially as a major factor affecting the price.

Sentiment analysis:

To prepare for sentiment analysis, we divide Manhattan into seven districts and focus only on reviews made in 2018. We merge the listing and review dataset together by the column *listing_id* because one listing may have multiple comments. For each district, we combine the comments, strip the comments to list, and keep only English word with no punctuations by using re, nltk.corpus, string, unidecode packages. We categorize sentiments as trust, sadness, fear, positive, negative, surprise, anger, joy, disgust, anticipation using NRC emotion dataset. Net sentiment for each district is calculated by the sum of all positive sentiment frequencies less the sum of all negative sentiment frequencies. Besides, we apply word cloud on comments in each district, deleting short words, hostnames since they frequently appear and has no meaning, and other meaningless words with a high frequency such as "apartment," "place," "host," "great," "good," and so on.

Price predicting models:

- Regression

In this part, we predict the combinational effect of over 20 possible factors on housing price by multivariate regression, polynomial regression and k-NN algorithm in sckit-learn to help Airbnb new hosts set leasing price properly.

Firstly, we use single linear regression to predict the relationship between price and number of bedrooms. After drawing the scatter plot, we can briefly see that these two variables are clearly positive related. Later, we try to add more variables such as the number of beds, bedrooms, accommodates, room type, location, cancellation policy score and so on to conduct the multiple regression as well as polynomial regression for higher accuracy. To unify different types of variable for regression, we add dummy variables for locations and set value from one to six based on different levels of cancellation policy from strict to free. Pearson correlation matrix and 3D plots are also plotted to eliminate the correlation of similar variables, thus deleting the number of beds and bedrooms given the higher correlation between price and number of accommodates.

Finally, we perform the k-NN regression by setting n_neighbors from 5 to 30, 5 per step to predict the numerical price based on the similarity measure. By comparing the above 4 models together, we can figure out that the best prediction model is the second degree polynomial model and the adjusted R-squared of test data is 0.484.

- Decision Tree and Random forest:

We first treat the price as a continuous variable and train the data with regressor tree. By plotting the R-squared value of training data and testing data, we find the best depth to be 5, which returns high testing R-squared as well as avoiding overfitting. The R-squared value for testing data is 0.4275.

Then we divide the price into 21 intervals, from 0 to 2000 USD, with a step of 100 USD and assigning each interval with values from 1 to 21. We first consider the intervals as discrete ordinary variables and use regressor tree again. With a depth of 6, the R-squared value for the testing data is 0.4204. Next, we consider the intervals as 21 unique classes and do the classifier tree to predict the price interval. In this case, with the best depth of 5, the accuracy of the model is 0.573. All three models tell us that the most important feature is the cancelation score.

In order to put all features into consideration and return a more fair result, we do random forest for all three tree models. For the regressor tree of continuous price, the hyperparameters include max_depth: 8, min_samples_leaf: 4, min_samples_split: 4, n_estimators: 50 and R-squared value: 0.4658. For the regressor tree of discrete price interval, the hyperparameters include max_depth: 8, min_samples_leaf: 4, min_samples_split: 4, n_estimators: 50 and R-squared value: 0.462. The random forest for classifier tree has parameter of max_depth: 8, min_samples_leaf: 4, min_samples_split: 4, n_estimators: 90 and accuracy: 0.5834. The random forest of classifier tree gives us the most accurate prediction.