Pricing Recommendations for Airbnb Hosts in NYC

Group Xpecial
Lili Yan(ly2459), Xinyi Zhou(xz2771),
Jiakai Liang(jl5243), Yijian Pang(yp2496)

I. Introduction

Airbnb is one of the most popular global online marketplaces to provide hospitality services to both guests and hosts. It not only provides opportunities for hosts to determine price and other details for their listings, but also offer guests a place to rate and write comments for each listing.

We recognize that there are many data analysis projects for Airbnb on website on a customer perspective. So the innovative part of our project would be analyzing the data from a host perspective and put ourselves in their shoes.

In this project, we try to give pricing recommendations for hosts to get high ratings in terms of customer preference given two scenarios: the host is considering about getting a new house or apartment and put it on Airbnb (Group A), or the host has a specific listing already (Group B).

II. Data Processing

We get the listing information, including host response time, room type, amenities, price, rating score, and other features in NYC from "Inside Airbnb" (http://insideairbnb.com/get-the-data.html) which is a platform providing Airbnb listing and review datasets. Based on the dataset, we performed data processing and feature engineering on the raw data and successfully convert 106 variables into 42 useful variables.

We also do a correlation analysis to see if these variables we get have strong correlations. Based on the results, we drop six features that have more than 70% correlation with others. This step also improves the results when we put data into regression models.

III. Exploratory Data Analysis

We make choropleth map to visualize the number of listings in different areas, and we picture the mean price of different areas in order to have a more straightforward comparison.

We use boxplot to visualize price comparisons of different property types and unit types, in order to give people in Group A a feasible suggestion on what kind of property type and unit type they should consider (ignoring the initial cost of purchasing a new property) in order to have a high-priced listing.

IV. Prediction

Based on correlation analysis, we selected 42 independent variables as features. At first we apply linear regression model on price value, but R^2 is less than 0.6, indicating there is no significant linear relationships between features and price value. Therefore, we separate price into ten intervals as the dependent variable. For Group A, since there are no previous ratings on new listings, we apply Linear Regression, Random Forest Regression Tree, Gradient Boosting Regressor, and Light GBM on price range using features without rating. For Group B, we apply these models using features with rating, in order to provide an updated price range prediction. The result shows that random forest performs best due to the highest R^2 and the smallest Mean Square Error.

After performing price range prediction, in order to further analyze the important features to improve listings. We divide data into 4 groups by labeling the top 25% highest price with 0, lowest price with 1, highest rating with 0, and lowest rating with 1. Finally we got 4 groups of listing with different labels, including “high price high rating” with label “00”, “low price low rating” with label “11” and etc. In order to find the most important features, we do the decision tree classifier and test the accuracy of classifier model by confusion matrix.

Since the accuracy of decision tree classifier is relatively high, we select 5 important features with top 5 weights. Then we draw 2 radar graphs in order to compare “high price high rating” to “low price low rating” and “low price high rating”. We could provide suggestions for hosts to make improvement on listings based on feature difference.

V. Text Mining and Sentiment Analysis

From the website, we also get review dataset, including comments from guests for each listing. We separate reviews to 4 groups by matching listing id in review dataset to id in “high price high rating”, “high price low rating”, “low price high rating”, “low price low rating” of listing dataset. For “high price high rating”, We draw word cloud graphs based on summary gathered in listing dataset and reviews in review dataset. Furthermore, we performed sentimental analysis on reviews in these 4 groups and generated a graph to show the percentage of positive emotions and negative emotions in all words.

VI. Interesting things in this project

1. We do not simply give predictions on price once, but instead we have a “dynamic” system on price predictions -- We can first give predictions when we don’t have any previous data, and as long as we have enough rating data, we can update our predictions by introducing rating as an additional feature
2. Based on our research, most previous data analysis for Airbnb uses Linear Regression directly on price prediction, but we separate price into 10 intervals so that we can apply more advanced models such as tree-based algorithms.

VII. Conclusion

For people in Group A, we have following 3 suggestions:

1. Location: in midtown or lower town Manhattan or brooklyn area that is near to Manhattan
2. Property types: loft or serviced apartment
3. Unit types: 2b2b, 3b2b, etc. (The ratio of number of bathroom to number of bedroom should be greater than or equal to ½)

For people in Group B, we can give them suggested prices of their listings based on our models before and after having rating data.

1. If they want to get a higher price and rating, they have to be careful about accommodates, number of available days, distance to Time Square and whether it’s entire house/apt or not.
2. Also, they can include “beautiful”, “neighborhood”, “home”, “spacious”, “great location”, etc. in summary to make their listings more attractive.