ANLY 506-91

Course Project Proposal

New York City Airbnb Data

Prepared by:

Srikanth Goli

Ramaswamy Iyer

Abinaya Karunagaran

Abisheik Mani

Shuang Tao

Pradeep Paladugula

Introduction

Airbnb is an online ecommerce company serving the industry of lodging and hospitality where people can list, discover and book unique rooms around the world. Whether travellers want an apartment for a night, a castle for a week, or a villa for a month, San Francisco based Airbnb began operations in 2008 and currently has thousands of employees across the globe supporting property rentals. The company does not own any real estate but simply acts as a broker and makes money in the form of commissions from each of these rentals.

1. What is the problem you are trying to solve or question you are trying to answer?

The data set contains 48,000 Airbnb listings in New York City. It also contains other metrics such as information about hosts, geographical location, type of rental, availability, reviews and price. The goal of this project is to predict the price of Airbnb rentals in New York City. Link to the data: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

2. What work do you plan to do in the project?

After getting a high-level overview of the New York City Airbnb dataset,

We intend to use some basic pre-processing steps (such as data accuracy checks, handling missing data, identification & potential exclusion of irrelevant variables that add no value etc.) to clean and preparation of data for exploratory data analysis. The subsequent steps will include feature extraction & correlation between different features including but not limited to price distribution of listings by neighbourhood group, by type of listing within each group, based on top hosts, reviews etc. In a nutshell, it will look at how the price of the listings compare with the median price across and within neighbourhood groups. Also analyze the total listings and price of listings correlated with room types within each group (supported by relevant visualizations techniques such as density plots, geographic heat map etc.) Finally, we will be testing the regression models for price prediction based on different attributes.

3. Which algorithms/techniques/models do you plan to use/develop? Be as specific as you can?

Basically, we decide to use Multiple linear regression models as our primary model development, to predict the room price in New York City. We will be focusing on the data that we have in the dataset for the independent variables, which includes neighborhood group, room type, price, minimum nights, number of reviewers/reviewers per month, room availability in a year and calculated host listing count. Between those the neighborhood group and room type are the dummy variable. The “room price” is an independent variable.

We will pre-process the data by performing exploratory data analysis using data visualization tools like (ggplot, hierarchical clustering) for our dataset. After pre-processing the data, we will use PCA and backward variable selection techniques to choose the independent variables. After the model is developed, we will pick another year of data to analyze the performance of the model.

4. How will you evaluate what you’ve done?

To evaluate what we have done, we will divide the data set to train set and test set, using the test set to test the model, checking the test data set score and the cross validation score to see how many responses we can explain by the statistics of imbalance we picked from the data set.

As for specific evaluation metrics, we use the R-Square method to evaluate and then from the R-Square output we determine how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). The equation is as follows:

$$R - Square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})\text{^}2}{\sum(Y_{actual} - Y_{mean})\text{^}2}$$

After we run a value of R-square, which is always between 0 and 1, we will be able to evaluate our models. We can also make use of Adjusted R-Square, which considers the degree of freedom of our model, and would represent a modification of R-Square, for evaluation of our work. R-square has a disadvantage: it only increases/remains at the similar level when X (independent variables)

are added. However, the modified metric, Adjusted R-Square, only increases when X (independent variables) are enhancing the model. Thus, we should also use Adjusted R-Square especially when we need to judge whether new variables are contributing to the accuracy of our model. The equation is as follows:

$$R^2 \text{ adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

5. What do you expect to submit/accomplish by the end of the project?

We are expecting to learn various exploratory data analysis techniques which will help in drilling down the data to discover useful insights. Furthermore, through the application of statistical methodologies such as regression modal. We get to hands on experience in using these techniques, explore and analyze the data to draw a final thesis. R code in markdown format and a brief project document is what the we are planning to submit at the end of this project.