Estimating rental price for lodging, homestay or tourist houses in New York City

Capstone Project Proposal-1

By: Muluemebet G Ayalew
Jan, 2020

Introduction

The New York city Airbnb open data is chosen to be analyzed for the first capstone project. The data is provided by Airbnb. Airbnb is an online marketplace for booking and/or listing lodging, homestay or tourist houses.[1] The data is accessible from the web, Inside Airbnb, and contains information about host, location, room type, price, review and availability.[2] The data contains roughly 50,000 rows and 16 columns.

The primary aim of this project is to predict the price of a new listing using the features given in the publicly available Airbnb data as well as other relevant features extracted from publicly available datasets. The results of the project benefits all three stakeholders including Airbnb, hosts and guests. Airbnb enables to recommend an appropriate price estimation for the hosts and therefore increase its revenue by increasing the number of hosts that use Airbnb's service. On the other hand, the price estimation can save hosts' time and effort to setup listing price. More importantly, if the estimated price is released publicly, hosts can attract more guests by adding some discount and gain more money. Guests can also make informed decisions when estimation is public.

The approach to address this problem is outlined in the next sections.

Research Questions

The objectives of the project is to make a robust model that can predict the price of the rent. To achieve this, the following research questions will be addressed:

---

[1] *2020, "What is Airbnb and how does it work?"* Airbnb. Accessed Jan 30, 2020. https://www.airbnb.com/help/article/2503/what-is-airbnb-and-how-does-it-work

[2] Dec, 2019, "*Get The Data"*, Inside Airbnb, Accessed Jan 30, 2020. http://insideairbnb.com/get-the-data.html

- How are the predictor variables related to price ?
- Are the predictor variables correlated?
- Which variables are relevant to predict the price?
- Are there any other variables other than in the dataset that can affect rental price? What are they? Where can they be found?
- Which method is better for price prediction for the dataset at hand and why?
- Does the model perform well for the test data? What metrics used to measure the performance of the model

Methodology

The data will be explored to observe relationships between variables, outliers, missing values and summary statistics. Then, data cleaning, feature engineering, transformation or feature scaling will be performed as needed. To increase the accuracy of the prediction, information from publicly available datasets will also be included. For example, distance from the nearest subway and distance from main attractions will be computed and added to the original dataset.

Once the data is prepared, regression methods such as linear regression, decision tree and random forest will be used to fit the training dataset and then to test for the test dataset, split during data preparation. Then, a method accurately predicts listing price will be used to make the final model. The most commonly used python libraries such as pandas, numpy, matplotlib, seaborn, scikit-learn will be used to explore, prepare, model and test the data. At the end, the code, visual stories and report will be delivered.