Airbnb Price Prediction

Domain background

Over the years, many companies in the lodging industry Airbnb and Zillow have released many open data sources and are analyzed by many users with machine learning technics. Many of the related projects are posted on platforms like GitHub[1] , Medium[2] and even written into papers[3]. However, related publishments so far only includes features like numerical or categorical features to do the prediction. But in fact, people also focus on some brief text descriptions, reviews, and pictures to decide their accommodations. Thus, the goal is to try to include this additional information to see if the predictive models will perform better.

[1] Milind Deore, “AirBnBPriceOptimizer”, https://github.com/milinddeore/AirBnBPriceOptimizer/, 2019

[2] Laura Lewis, “Predicting Airbnb prices with machine learning and deep learning”, https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6, 2019

[3] Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, Hoormazd Rezaei, “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis”, https://arxiv.org/abs/1907.12665, 2019

Problem Statement

Since both the house providers and the customers sometimes have a hard time knowing whether the price is attractive from customers and also be profitable for the hosts at the same time. The problem I am trying to solve is about making a more accurate prediction on the price according to information that the users used more often when selecting accommodations but rarely taken into consideration when making price predictions. The goal is to provide a better price for both users and hosts. Also, the project will also include dashboards pointing out important features to help hosts to increase their business and user experience.

Datasets and inputs

The datasets will be obtained from http://insideairbnb.com/get-the-data.html and that is the only open-source provided by Airbnb. The datasets include information such as house type, house description, and the corresponding price, etc, which I believe is directly related to the problem I am trying to solve.

The datasets will be downloaded through the URL above, and the name of the files are all called “listing.csv”. Since there will be over 50 files in total, I will implement Python web scrapping skills to download the files. The reason I am only using “listing.csv” is that it contains all the information that Airbnb can provide. Any reduced version of the dataset is not helpful for the project’s goal of searching for additional useful features to build the model

Solution statement

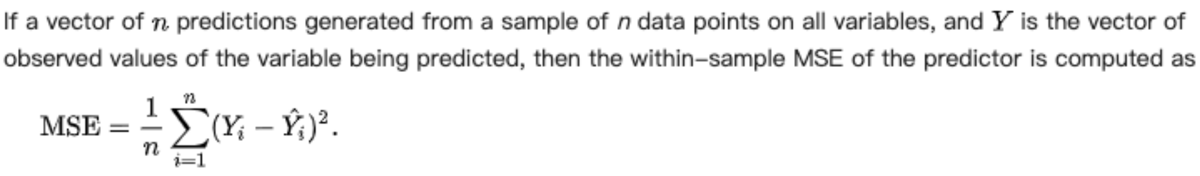
The solution is to provide a predicted price for each query using the machine learning models. The first thing is to include features similar to previous researches. Second, I will use text data as one of the important inputs. The way I will handle the text data may include tokenization, stemming, stop-words removing, and vectorization. The project may include image information if applicable. Finally, I will treat all the above data as input and train the machine learning model.

Benchmark model

I will choose the benchmark models from several famous posts from Github such as AirBnBPriceOptimizer created by Milind Deore and Predicting Airbnb prices with machine learning and deep learning by Laura Lewis. Both of the projects selected some simple features as their final model, and which are perfect samples for me to do the comparison.

Evaluation metrics

Currently, the evaluation metrics will be the mean square error or the root mean square error depending on the actual work. The formula is shown below, Y hat is the predictive results from the model created by various of features such as house type, size, amenities, and etc. I will select the model with the smallest MSE.



If a vector of n predictions generated from a sample of n data points on all variables, and Y is the vector of observed values of the variable being predicted, then the within-sample MSE of the predictor is computed as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2.$$

Project design

The project will be completed through several phases:

1. **Problem understanding**

   Understand the business problem and find a suitable data source to solve the problem.

2. **Data collection**

   Using Airbnb data from InsideAirbnb.com and only select data across the US. Due to a large number of datasets, I will prefer collecting the URLs first through web scrapping for convenience and future maintenance. The tasks will be done by Python “requests” and “beautifulsoup” package.

3. **Data cleaning and data exploration analysis**

   Look at the distribution and the missing values of the data to check which features I interested in are available for the projects. Also, do some exploring data analysis such as aggregation or data visualization to find insights. This pr ocess is not just for data understanding but also a way to evaluate if the model is performing on the right track. Some of the works here may be implemented into visual dashboard for greater user experience.

4. **Train machine learning model**

   First, train the benchmark model according to previous researches to see first cut performance. Then, train on traditional machine learnin g technics with new features included. Third, train the model through deep learning algorithms to enhance model performance as possible. I will use mean square error (MSE) as evaluating metric for the model and hyperparameter tuning will also included in this process.

5. **Model deployment and result presentation**

   The result will be displayed on a web application with different selecting bars for users to filter their ideal accommodation and a price given by my machine learning model. The web application also includes some visualization for some important features.

   Currently, I will select Flask and Bootstrap as the major frameworks for the project and use D3 for the dashboard creation. The machine learning model is provided through an API created by AWS Lambda function and its APIGateway. Ideally, users will receive realtime results once they send their housing information.

6. **Improvement**

   The future may include showing the predictive results on a map like a chart so that the user can have a very quick understanding of the house price immediately. Also, building a recommendation system may be another way to enhance the user experience and give hosts some guide to improving their business.