Singapore Airbnb rental price (from Kaggle Competition)

Anitha Veeramani

July 13th, 2020

Machine Learning Engineer Nanodegree Capstone Proposal

- ***Domain Background:***

  Airbnb is an internet marketplace for short-term home and apartment rentals. It allows the owner to rent out home for a week, or rent out empty bedroom. One challenge that Airbnb hosts face is determining the optimal nightly rent price. Hosts with multiple listings are more likely to be running a business, are unlikely to be living in the property, and in violation of shortest term rental laws designed to protect residential housing.

  As in **SINGAPORE** — Short-term rentals offered by platforms such as **Airbnb** will remain **illegal**, and later the rule was implemented were rental is allowed only minimum stay of three months to private residential properties, hence personally would like to explore Airbnb Singapore data to check the violation and hotspot rental areas in Singapore after analyse the dataset

- ***Problem Statement***

  The purpose of this task is to predict the rental price of Singapore Airbnb based on the data provided in the kaggle dataset and analyse what are the key features that affect the price of the listing. According to AirBnB official website information, AirBnB reservation price is based on following costs

  - Costs determined by the host:
    - Nightly price: Nightly rate decided by the host;
    - Cleaning fee: One-time fee charged by some hosts to cover the cost of cleaning their space;
    - Extra guest fees: One-time fee charged by some hosts to cover other costs related to using their space;
  - Costs determined by Airbnb: Airbnb service fee

- ***Datasets and Inputs***

  The Singapore Airbnb dataset used for this project comes from Insideairbnb.com and was collected on 28 August 2019 according to the website. This shows that the data set contains 7907 entries and 15 columns (not including the id). Some rows contain missing information for certain columns like last_review and reviews_per_month there are only 5149 entries.

Some of the more important features this project will look into are the following:

- accommodates: the number of guests the rental can accommodate
- bedrooms: number of bedrooms included in the rental
- bathrooms: number of bathrooms included in the rental
- beds: number of beds included in the rental
- price: nightly price for the rental
- number_of_reviews: number of reviews that previous guests have left

To predict the rental price, first find high level trends and correlation between other features with property prices are analysed using heatmap. Use machine learning for further analysis.

- ***Solution Statement***

The solution will be the predicted price value of the listing given in the test dataset. First data is cleaned, normalised and important features were extracted after detailed data analysis. Then split the data into train and test dataset. For training models train more than one algorithm and finally select the best model for this problem and fine tune parameters to get best accuracy. apply machine learning methods to see which features in the dataset influence the price the most. In order to do this we will train two popular models (**Linear regressor**/ RF algorithm and **XGBoost regressor)** based on decision trees and feature importance variable.

- ***Benchmark Model***

For this problem, the benchmark model will be XGBoost and Linear regression model. Using Linear Regression, I was able to predict price based on a cleaned dataset, with a score of ~0.62 on both training and test datasets

- ***Evaluation Metrics***

Prediction results are evaluated on the log loss between the predicted values and the ground truth. Accuracy based on r squared is calculated. A solution with low root-mean-squared error (RMSE) based on cross-validation that can be reproduced

- ***Project Design***

Generate multiple models after cleaning the data, split the dataset as train and test set, build the model and compare the accuracy for each model on the same test set. Evaluate the model performance and select the best model that can predict the prize for the given test dataset based on AUC performance metrics. Identify the features such as minimum stay, price and number of reviews have been used to estimate the occupancy rate and the income per month for each listing.

- ***Reference***

1. https://www.kaggle.com/aleksandradeis/airbnb-seattle-reservation-prices-analysis

2. https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a
3. **AirBnB official website information**
4. https://en.wikipedia.org/wiki/XGBoost