Prediction of Airbnb Stays Prices

Abstract

The goal of this project is to apply Linear Regression on data scraped from Airbnb website, to predict the price of a stay based on the features of that stay (number of beds, rating, etc.). I worked with data that I have scraped from Airbnb.com using Selenium to go through the pages of the website and scrape data using BeautifulSoup. After that, the data was combined into a single Pandas data frame, in order to achieve the goal of this project. After cleaning the data, I started to explore the dataset and split the data into train/validation, and test data. Next, Linear Regression model was used to train the data and to achieve the goal of this project. Finally, I communicated my work using a PowerPoint presentation to show my findings.

Design

By training the data using Leaner regression models, we could predict the price of a certain stay, based on the features of that stay (number of beds, rating, etc.).

Data

Sample size is 6 major cities (New York, Los Angeles, Chicago, San Diego, Houston, and Philadelphia) that are located in the United States. The dataset has 1080 rows and 10 columns.

Algorithms

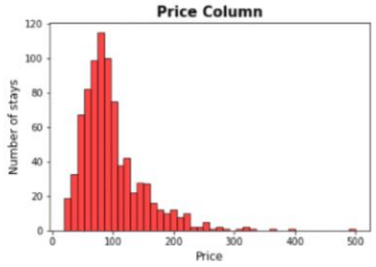
The methodology used in this project is: Problem understanding, data validation, Data exploration, data visualization, feature engineering, training and modeling the data.

Tools

- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting
- Dython for visualizing the correlation between nominal and continuous data.
- BeautifulSoup and Selenium to go through the webpages and scrape the data.

Communication

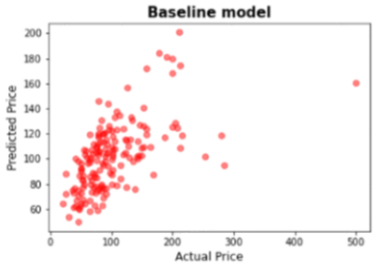
EDA



- The data has some outliers, but they were kept as they had meaning.

airbnb

Baseline Model



- The R2 score of training data is 0.2280, and for the validation data is 0.3327.
- The model is not very accurate. Also, the R2 score of validation data is slightly higher than training data, which means that the model is not robust enough.

airbnb

Models used

Model	Validation R2 Score
Linear Regression	0.199
Linear Regression with log	0.244
Polynomial degree 2	0.211
Ridge	0.197

airbnb