Project Airfare – Predicting Airline Prices

Bianca Orozco

Project Design

I web scraped Expedia.com flight information to predict airline prices. Web scraping for this project was done using Selenium and a Google Chrome web driver. My sample pool included major airports located in Atlanta, Los Angeles, Chicago, Dallas, Denver, New York, San Francisco, and Seattle; specifically, all one way, one passenger flights from any combination of the aforementioned cities. Once I was done web scraping, I had about 3,100 data points. The features I was able to use included 5 numerical features, 4 categorical features, and 1 dependent variable (prices).

My initial goal was to gain insights as to how features such as time of day and airline affected the cost of a one way flight. However, due to time constraints, I decided I would focus on creating a model that could predict prices.

Tools

Jupyter Notebook
Web Scraping:

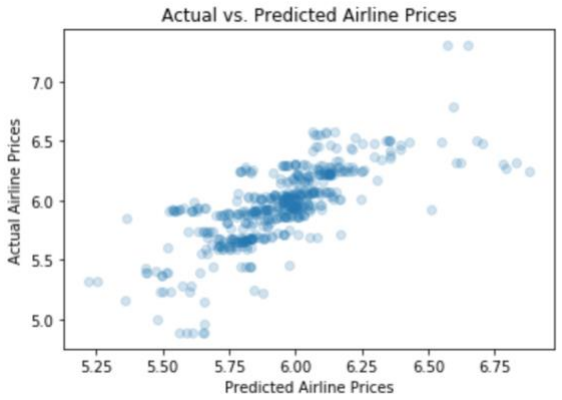
- Selenium
- Pandas
- urllib

Data Analysis:

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Sklearn

Data

After cleaning the data and performing EDA, I was able to see the Prices data was scewed left. I used a log transformation to normalize the data. Using a Linear Regression model, split my data 60% for training, 20% for validation, and 20% for testing. With the final test of my model, I received an R Squared value of 0.548 and MAE of approximately $116.



What Would I Do Differently

I would like to have spent less time coding my web scraping and more time feature engineering and understanding my analysis and betas. This would allow me to look deeper into my features and see which have the greatest effect on prices. I would also perform a K-Fold CV to improve my model.