Flight Fare Prediction using ML

The Three Musketeers (Lakshya Kumar, Meghana Kandru, Srijan Srivastava)

Summary

In this project, we predict the flight fare prices using the Linear Regression model. The data was obtained from the Bureau of Transportation Statistics (<https://www.bts.gov/>). The dataset contains information about the flight date, aircraft carrier, origin, destination, distance flown and the ticket price for that journey during the year 2019. We have taken into consideration that the data contains prices for one person per ticket and the predictions would also be for one ticket only.

The highest accuracy score we achieved was 93% using the Linear Regression model after encoding all the categorical features into numerical features and adjusting the data frame for skewness using log transform.

Methodology

The initial data consisted of 0.31 Million row observations and 11 columns of flight data from 2019 where the source or destination was a WA city. After dropping the null values, duplicate values and encoding the categorical features we obtained 0.30 Million row observations and 181 columns.

After the data was prepared, we used bar and box plots to inspect any outliers in the data, histogram to observe skewness in data and cat plot, heatmaps to understand the relationships between various features in the data.

To ensure the bias in the data was not prevalent in our predictions, we took a log transform and adjusted the positive skewness of the data. We partitioned the initial data into training and test data sets. We then used the final “testing” data as comparator against the predicted values to generate accuracy scores like Test MSE, Test MAE, RMSE and R^2 squared values.

We built multiple models with different methodologies but only choose Linear Regression at the end due to speed of calculation and the accuracy of the model. Other models that we built were Ridge Regression, Lasso Regression, Polynomial Regression with degree one and Random Forest Regressor using RandomizedSearchCV.

After the model predictions were near expectations, we stored the model files in a pickle and imported the same in a web application built using Python and Flask framework. The resultant displays the price of the ticket within the web application based on the input parameters of the model. The web app link is: [[USA Flight Fare Prediction Heroku Web App](https://us-flight-fare-prediction.herokuapp.com/)](https://us-flight-fare-prediction.herokuapp.com/) and [[Github Link](https://github.com/srijansrivastava11/us-flight-fare-prediction/tree/srijan)](https://github.com/srijansrivastava11/us-flight-fare-prediction/tree/srijan)

**Coding Steps:**

1. Read and clean data
2. Partition the data in training and testing sets
3. Build a model with LinearRegression() function on training dataset
4. Evaluate out-of-sample accuracy, i.e., cross-validation
5. Predict on final test set and show results on a web application