Predicting Airline Ticket Prices

Using Linear Regression Models

Juri Alsayigh

10/13/21

Machine Learning

Predicting Airline Ticket Prices

Abstract

The goal of this project was to use linear regression models to predict Airline tickets to Riyadh during the Riyadh entertainment season 2021. The purpose of this model is to help our client to provide the best tickets for their customers during Riyadh Season 2021. the data worked on was obtained using Web Scraping methods on Google Flights. After exploring, cleaning, and analyzing the data, plots and graphs were plotted to visualize the correlation relationships between features in order to choose the best features possible , later on a simple linear regression model was validated on the validation set which came out with a curve instead of a linear line , in order to correct this a polynomial regression model was used instead , After refining the model, a Prezi presentation was made to communicate the prediction accuracy and results.

Wego Saudi Arabia

Wego is an award-winning meta-search engine; they're the largest online travel marketplace in the Middle East, North Africa & Asia Pacific. Wego's core markets are in the Asia Pacific, Middle East, and India. Wego travel app currently ranks within the top five of the most popular apps in the iOS and Play store travel category for Middle Eastern region. Wego partners up with various smartphone manufacturers (Samsung being a prime example) to have their app come preinstalled. Wego provides its services through Android/IOS apps and online website.

Each month Wego sends flight and hotel booking referrals worth US$1.5B to its travel partners. In addition to partnerships with hotels, airlines and OTAs, Wego continuously collaborates with tourism boards of various countries, such as Singapore, South Korea, Jordan, Macao, and many others, to increase travel demands to and from said countries.

Today, Wego is used by millions of people every month — people who travel for adventure, for work, for family and for many other reasons.

Design

This project data which is web scraped from Google flights reflects the flight ticket information including the location from and to, the booking date and flight date, departure time, arrival time, flight duration, airlines, and ticket price. Other information was derived from the previously mentioned is the weekday and the booking period (from booking date to flight date). The flight price prediction model will enable Wego to provide their clients of flight ticket prices close enough with the actual prices during Riyadh entertainment season.

Data

The uncleaned dataset contained around 121202 tickets with 8 columns for each. which are: date of booking, date of travel, from, to, airline, departure Time, arrival Time, the duration, and price of ticket. After adding dummy data and calculated fields (booking days and weekday) and doing some feature engineering and removing duplicates and outliers and cleaning the whole dataset the dataset turned onto 18481 rows and 14 columns, that later on was checked for correlation which turned the focus on 6 features which are the most highly correlated with the target variable price.

Algorithms

1. Web scrapped the data from google flights of Saudi local cities (Jeddah, Dammam, Jazan, Medina) to Riyadh in Riyadh season period which is from October 2021 to March 2022 using Selenium web scraping method.
2. Upload the web scrapped data into a csv file called tickets.
3. Downloaded the tickets data csv file into Jupiter notebook then Loaded the file into a pandas Data Frame.
4. Explored and got to know the data while understanding how to use it and which attributes should be considered as features and prepare to be used in the prediction model.
5. Cleaned the data and filtered it removing outliers and duplicates.
6. Converting categorical features to binary dummy variables during EDA.
7. Checking for correlation and normalizing the data taking the highly correlated features with the target and removing some features that are highly correlated with each other.
8. Splitting the data into sets: train 60%, validation20%, and test 20% in chronologically sorted order.
9. Fitting the Linear regression 6 models on train data using different features and comparing the score of each. Ending up with the best score of R^2: 0.9114945339772216 and Adjusted R Square: 0.9114705812123981, the Intercept: 1.031231756552298 and the following features: 'bookingDays', 'depaHour', 'durationMin', 'airline_Flynas', 'from_jeddah' with their coefficients as:

    - bookingDays : -0.25
    - depaHour : -0.11
    - durationMin : -0.16
    - airline_Flynas : -2.11
    - from_jeddah : -0.66

10. the results ended up in a kind of a curve shape; so, a polynomial regression model was compared with the linear regression model on the validation set, where the polynomial regression model has a better R^2 score of 0.9.
11. test on the test set to get a final evaluation of expected generalization performance.

Tools

- selenium web scrapping
- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting
- Sklearn for machine learning
- Statsmodel for statistics description

Communication

In addition to the slides and visuals presented, everything is submitted through the submission form.