Nhi Le

Regression Analysis on Flight Price Prediction

Air India Flight Price Prediction departing from Mumbai with morning departure time

Western Governors University

November 17, 2023

**Research Question**

At an online travel company, they specialize in offering budget-friendly flight tickets. Various factors influence the price of these tickets, and stakeholders are keen on understanding the correlation between the number of days left (calculated by subtracting the trip date from the booking date), seat class, and the flight price of Air India flights with morning departure times. The research question driving this analysis is: How are the price of Air India flights affected by the number of days left and seat class? Through this analysis, the company will be able to develop a regression model that can predict flight price based on days left and seat class variable. By accurately predicting flight prices and considering the practical significance of the regression model, the company intends to provide its customers with valuable suggestions on the best deals available.

Here are the hypotheses:

* Null hypothesis (H0): Days left, and seat class do not have a statistically significant relationship with flight price.
* Alternate Hypothesis(H1): Days left, and seat class have statistically significant relationship with flight price.

To examine if there is a statistically significant relationship between the independent variables and the dependent variable, the overall F value of the model and the corresponding p-value will be analyzed. If any of the p-values is less than 0.05, the null hypothesis can be rejected (Zach, 2021). This rejection implies that the independent variables exhibit a significant relationship with the dependent variable.

**Data Collection**

I will need to collect data containing flight booking options from the website with at least three variables: days left until the trip, seat class, and prices. However, I will not need to gather the data as I have found an existing data set that includes all the information required for the analysis. The dataset is publicly available on Kaggle.com with no restrictions, titled 'Flight Price Prediction,' last updated two years ago. The zip folder comprises three CSV files, and I will be utilizing 'Clean\_Dataset.csv.' The dataset was published by Shubham Bathwal, who collected the data using the Octoparse scraping tool on the Easymytrip.com website. It is in the Kaggle public domain, and there are no restrictions on its usage. The information was gathered for the period from February 11th to March 31st, 2022 (Bathwal, 2022). The dataset consists of 12 columns and 300,153 rows. After filtering the dataset, only 17,768 rows will be used in this analysis. The variables for this analysis include price (a continuous variable), days\_left (a continuous variable), and class (a categorical variable).

The advantage of using this dataset in data collection step is its cleanliness and readiness for use. This eliminates the need to spend a substantial amount of time in web scraping and cleaning the dataset. However, this also poses a disadvantage since I did not directly conduct the data gathering. Consequently, I am not entirely certain about the quality of the dataset. However, I can later delve into a more thorough examination to better understand the dataset, perform any necessary cleaning, and prepare accordingly.

The dataset is excellent but one challenge I encountered was redundancy. The data set has many variables and values are not needed in this analysis. I will explain the process of filtering it to obtain the dataset required for this analysis in the Data Extraction and Preparation phase. To overcome this challenge, I will need to understand the dataset by using Python functions to gather information about the types, basic statistics, column/row counts, etc. Subsequently, I will filter the dataset to include only the information necessary for this analysis.

**Data Extraction and Preparation**

In this analysis, I will utilize Python. Given my basic knowledge of Python, using Python for this analysis will prove beneficial. Python's syntax is simple and straightforward. Moreover, Python has a lot of libraries and packages that can assist me in constructing my predictive model. I can employ Python for tasks ranging from data cleaning using NumPy and Pandas libraries to data visualization using Matplotlib and Seaborn, and finally, to machine learning with Scikit-learn (Pruciak, n.d.). Furthermore, Python's longevity, open-source nature, and a widespread global IT community enhance its appeal. I can find support and solutions quickly for my code if I need it, which helps me to complete this analysis effectively and accurately.

The disadvantage of using Python is that Python is slow at runtime. Since Python is a high-level programming language, Python's execution relies on an interpreter rather than a compiler. Consequently, Python code is processed line by line, leading to a runtime slower than that of C/C++ or Java (GeeksforGeeks, 2023).

To acquire the data set, I went to this link <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction> and downloaded the zip folder. Then I extracted all the files to the desired folder. The file I will be using for this analysis is Clean\_Dataset.csv.

My data preparation goal is to detect missing data, duplicate data, and outliers, then decide to treat them with appropriate methods:

Data preparation steps:

* Import dataset Clean\_Dataset.csv into Jupyter Notebook.
* Get information (column names, data types), and statistical details (count, min, max, mean, std, percentile) of the dataset.
* Filter the data set to retrieve entries with the airline ‘Air\_India’ and departure\_time in ‘Morning’.
* Detect duplicates and delete the duplicated records if there are any.
* Find missing data and impute missing data with meaningful measures of central tendency (mean, median, or mode).
* Find outliers and treat them by removing them, retaining them, excluding them, or imputing them with the median.
* Run univariate and bivariate visualizations to see the spread of data.
* Drop variables that will not be needed for the analysis.
* Create dummy variables for categorical variables.
* Encode categorical values to numerical values: For the class variable, it has 2 unique values Economy and Business. DummyEconomy is 1 when class is Economy, else it is 0.
* Spot-check the statistical details of the dataset to make sure categorical values are encoded correctly.
* Drop the categorical variable from the data set.
* Extract the prepared dataset as CSV file named ‘df\_prepared.csv’.

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated



A screen shot of a computer

Description automatically generated

**Analysis**

Multiple linear regression (MLR) will be used to analyze the data. MLR is an appropriate technique because given that flight price is a continuous variable. MLR is convenient in this predictive model using multiple independent variables, instead of just a single linear regression. In summary, employing MLR will aid in addressing the research question, enabling executive leaders to predict flight prices and offer customers recommendations for optimal deals.

MLR is a supervised machine learning algorithm calculating the strength of the linear relationship between the independent variables and the dependent variable.The four assumptions of a multiple linear regression model (MLR):

* There is a linear relationship between an explanatory variable and a response variable.
* Among explanatory variables, they are not too highly correlated with one another.
* yi observations are selected independently and randomly from the population.
* Residuals normally need to be distributed with a mean of 0 (Hayes, 2023).

The advantage of MLR is its ease of implementation and interpretation of output coefficients. Through that, the model will be able to illustrate the relationship between the independent variables and the dependent variable. MLR’s disadvantage is that it assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. Therefore, it might oversimplify the real-world problems (GeeksforGeeks, 2023).

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Once the model has been created, I would verify if my model met all the 4 assumptions. Firstly, I used heat maps to see the relationships between the independent variables and the dependent variables. I need to see whether those independent variables having correlations with the dependent variable more than 0.05 or less than -0.05 Secondly, I checked multicollinearity by calculating VIF (Variance Inflation Factor) if those variables having VIF less than 5. Thirdly, I created residual plots to verify independence of observations, homoscedasticity, and normality of residuals.

A screenshot of a computer screen

Description automatically generated

The heat map indicates there is correlation between days\_left, DummyEconomy and price, since the absolute values are greater than 0.05.

A screenshot of a computer code

Description automatically generated

The VIF result 2.015739 shows that there is not a high correlation between days\_left and Dummy Economy.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

In the residual plots, the points are scattered randomly around the 0 line. The observations are independent and random. The spreads of residuals are not biased to one side, so the model does not violate heteroscedasticity. It is rare to obtain a perfect model, so the result is acceptable.

**Data Summary and Implications**

The multiple linear regression equation including 2 independent variables:

Price = 50095.39 – 144.73 \* days\_left – 38983.56 \* DummyEconomy

The coefficients suggest that for every 1 unit of:

* Days\_left (days until the departure date): price will decrease 144.73 units.
* DummyEconomy (seat class is economy): price will decrease 38983.56 units.

F-statistic is 67,230. R-squared/ Adjusted R-squared is 88.30%, meaning the model can explain 88.30% of the variance. P-values for days\_left and DummyEconomy are both 0.000.

With p-values very close to 0.000, we can reject the null hypothesis. Therefore, we can conclude that the model is statistically practical. "DummyEonomy" and "days\_left" have a statistically significant relationship with flight “price”. The results imply that for flights departing from Mumbai in the morning, the earlier customers buy the flight ticket until the departure date, the lower the flight price can get to. Moreover, when customers purchase a ticket in Economy class, the flight price is lower than a ticket in Business class. The limitation of the analysis is that the dataset has a relatively short period of time, with data collected for only 50 days. If we had data spanning at least a year, we could track and identify more accurate trends.

Here are some recommendations for the company course of action. Firstly, for customers who want Business class flight tickets, the company should suggest customers book their flights as early as possible. Secondly, for customers intending to book flights close to the departure date, Economy class tickets are recommended for cheaper prices. Lastly, according to the data set, the options to book the flight range from 1 to 49 days in advance. The company can suggest customers book their flights at least 30 days before the departure date. The best deals they can offer customers are flights in Economy class booked 49 days in advance.

For future studies of the data set, I would like to add more variables in the regression model to explore whether there are any other variables having statistically significant correlations with the flight price and if we can create a better model. Secondly, we can employ clustering techniques on the data set to see the characteristics of the flight price. K-means clustering is a good technique to consider to group flight prices into groups based on some variables such as days\_left and duration.

**References**

Zach. (2021, May 18). *Understanding the null hypothesis for linear regression*. Statology. https://www.statology.org/null-hypothesis-for-linear-regression/

Bathwal, S. (2022, February 25). *Flight price prediction*. Kaggle. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Pruciak, M. (n.d.). *Why is python a great choice for data analysis?*. Agile Software Development Agency in Europe. https://www.ideamotive.co/blog/python-for-data-analysis

GeeksforGeeks. (2023, November 2). *Disadvantages of python*. GeeksforGeeks. https://www.geeksforgeeks.org/disadvantages-of-python/

GeeksforGeeks. (2023a, March 30). *ML - advantages and disadvantages of linear regression*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/

Hayes, A. (2023, May 9). *Multiple linear regression (MLR) definition, formula, and example*. Investopedia. https://www.investopedia.com/terms/m/mlr.asp