

Final Project

Lauren Walker and Shanaya Piramal

Introduction

Project Motivation and Revelance

The data set we selected for our final project is the “Flights” data set. We got this data set from Kaggle, and it contains 3000153 unique observations, each representing a flight from one of India’s 6 metro cities. In order to collect the data, researchers made use of the Octoparse scraping tool to extract data from the website “Ease My Trip”. Data was collected in two parts: one for economy class tickets and another for business class tickets. The data was collected over a period of 50 days, from February 11th to March 31st, 2022.

While Covid-19 took a toll on a lot of industries, one of the industries that it took a major toll on was the Airline industry. As a result, flight prices have significantly increased (<https://www.nerdwallet.com/article/travel/travel-price-tracker>). Both of us have had a passion for travelling and seeing new places. However, as college students, we understand that travelling is a big cost. Hence, exploring how to strategically book tickets, depending on the number of stops, arrival/destination time, and airline is a topic that interested us.

As travel becomes a significant part of our lives again, this project has practical applications for individuals, businesses, and policymakers and can contribute to a better understanding of air travel pricing dynamics.

Data

Our data has 300153 observations and 12 variables. Each row in our dataset represents a particular flight. Each of the variables represent attributes of the flight. The variables are explained below:

Table 1: Variables Explained

| Variable | Explanation |
|----------------|--|
| Airline | The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines. |
| Flight | Flight stores information regarding the plane’s flight code. |
| Source City | City from which the flight takes off. |
| Departure Time | This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels. |
| Stops | A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities. |
| Arrival Time | This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time. |
| Dest City | City where the flight will land. It is a categorical feature having 6 unique cities. |
| Class | A categorical feature that contains information on seat class; it has two distinct values: Business and Economy. |
| Duration | A continuous feature that displays the overall amount of time it takes to travel between cities in hours. |
| Days Left | This is a derived characteristic that is calculated by subtracting the trip date by the booking date. |
| Price | Target variable stores information of the ticket price in INR. |

<https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

Research Question and Hypothesis

After doing some research, we found out that the typical price of a domestic flight in India is ~ 7,000 INR (As per SkyScanner, in May 2023). Keeping in mind our motivation for this project, the main research question that we will explore through our final project is: “For flights that are departing out of cities, to what extent do the source destination, duration, departure time, and number of days before booking the flight influence the predictive odds of a flight being priced above 7,000 INR?”

Our formal null hypothesis and alternative hypothesis based on this research question are as follows:

Null Hypothesis H_0 : None of the aforementioned factors will be predictive of the odds that the flight price is above 7,000 INR.

Alternative Hypothesis H_1 : At least one of the aforementioned factors will be predictive of the odds that the flight price is above 7,000 INR.

Data Cleaning

We first checked our data for missing values and duplicates. After looking through our data, we realized that we had none of the aforementioned problems. However, we still felt the need to modify certain variables to make our modelling process smoother and conduct an effective EDA.

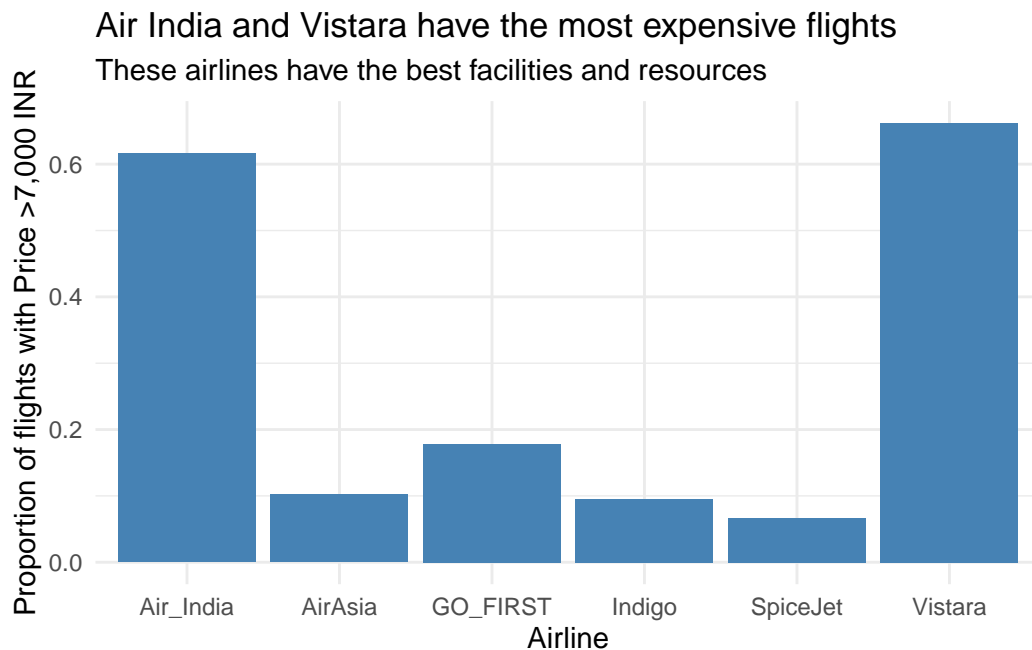
Since Mumbai and Delhi are the two biggest Hubs in India, we decided to focus our research project with these two cities as our primary source and destination.

This then reduces our number of observations to 30,098 flights.

Our main change was to our main variable of interest. We converted it to a binary variable, 1 for if it was greater than 7,000 INR and 0 if it was less than 7,000 INR.

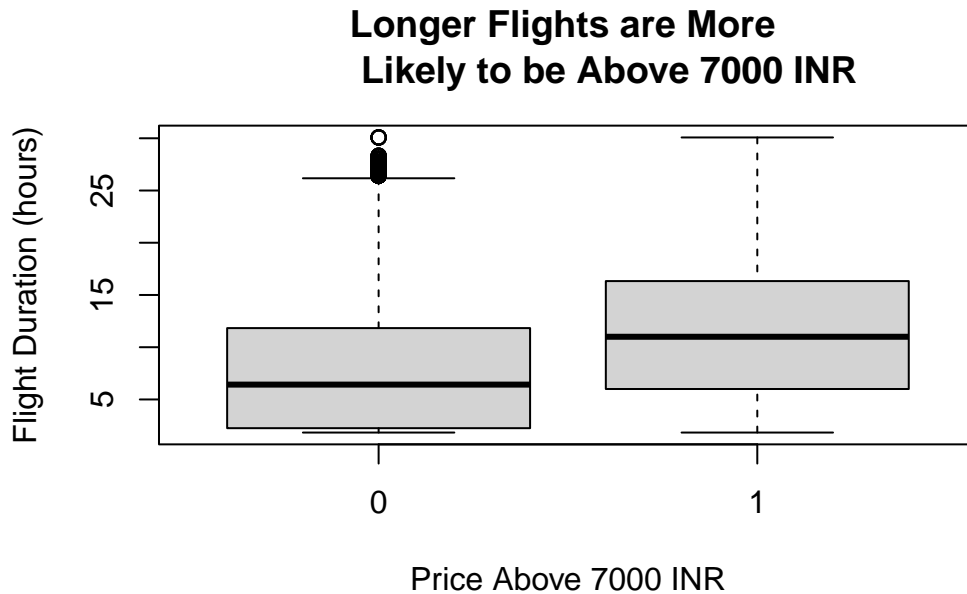
Exploratory Data Analysis

Before we explore our research question, we are going to analyze summary statistics and other relevant relationships between our main response variable of interest, Ticket Price, and other variables in the data set such as duration, airline, class and days left. By understanding the patterns behind these trends, we will be able to effectively explore our research question.



National airlines such as Vistara and Air India have the highest proportion of flights that we are considering expensive. The more commercial airlines have cheaper options. I found this quite interesting, because I would assume the national airlines would be cheaper and more

affordable to all. We also see a great difference between the proportion of expensive flights within the airlines. This suggests that the airline chosen might be predictive of the odds of a flight being overpriced.



Another important factor to consider is the duration of the flights. Based on the box plot, we observe that the median flight duration as well as the middle 50% of flight durations for flights priced above 7,000 INR is about 5 hours more than the median flight duration and middle 50% of the data for flights priced 7,000 INR or below. Moreover, we notice that the box plot for flights costing 7000 or less INR is more right skewed than the box plot for flights costing more than 7000 INR, suggesting that more flight duration lie on the lower end of flight durations compared to flights costing more than 7000 INR. These findings indicate that flight durations are longer for flights costing more than 7000 INR than that of flights costing 7000 or less INR. These results were not surprising to us because we assumed that flights that lasted longer would cost more.

Methodology

Our modeling Process

In order to answer our research question - for flights that are departing out of cities, to what extent do the source destination, duration, departure time, and number of days before booking

the flight influence the predictive odds of a flight being priced above 7,000 INR? - we created a logistic regression model.

Variable Selection

After using our domain knowledge, through our research and understanding of the airline industry, we decided that it would make the most sense to focus our research on understanding the relationship between destination city, duration, departure time, and days left. We believed that most of these factors are the factors that people take into consideration while booking flights. Most people cannot decide their source city.

We also considered adding an interaction term, however, we realized that in this case we decided that with our selected variables, there is a low chance that one of the variables depends on the other.

Conclusion

Results

Hence our linear model is as follows: Our linear model is as follows:

$$Price \sim Destinationcity + Duration + DepartureTime + DaysLeft$$

Table 2: Results of our Model

| term | estimate | std.error | statistic | p.value |
|-----------------------------|-------------|------------|-------------|-----------|
| (Intercept) | 0.2375921 | 0.0433832 | 5.4765982 | 0.0000000 |
| destination_cityMumbai | 0.0847402 | 0.0249825 | 3.3919849 | 0.0006939 |
| duration | 0.0737530 | 0.0017932 | 41.1290295 | 0.0000000 |
| departure_timeEarly_Morning | 0.1492581 | 0.0398304 | 3.7473408 | 0.0001787 |
| departure_timeEvening | -0.2264743 | 0.0398642 | -5.6811516 | 0.0000000 |
| departure_timeLate_Night | -15.0598922 | 90.5077536 | -0.1663934 | 0.8678474 |
| departure_timeMorning | 0.0829194 | 0.0394727 | 2.1006777 | 0.0356693 |
| departure_timeNight | -0.3664974 | 0.0464043 | -7.8979201 | 0.0000000 |
| days_left | -0.0385056 | 0.0009341 | -41.2232408 | 0.0000000 |

Our null hypothesis is that there is none of the aforementioned factors will be predictive of the odds that the flight price is above 7,000 INR, holding the other predictors constant. Our

alternative hypothesis is that at least one of the aforementioned factors will be predictive of the odds that the flight price is above 7,000 INR, holding the other predictors constant.

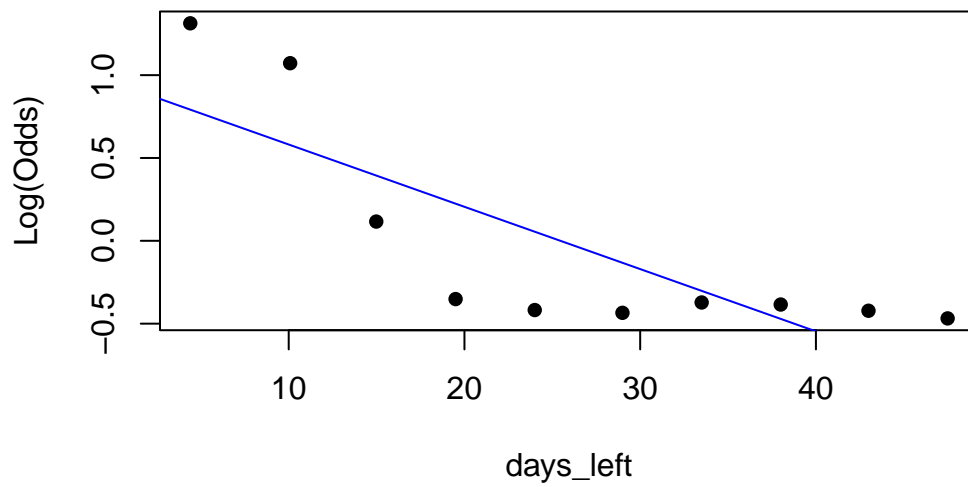
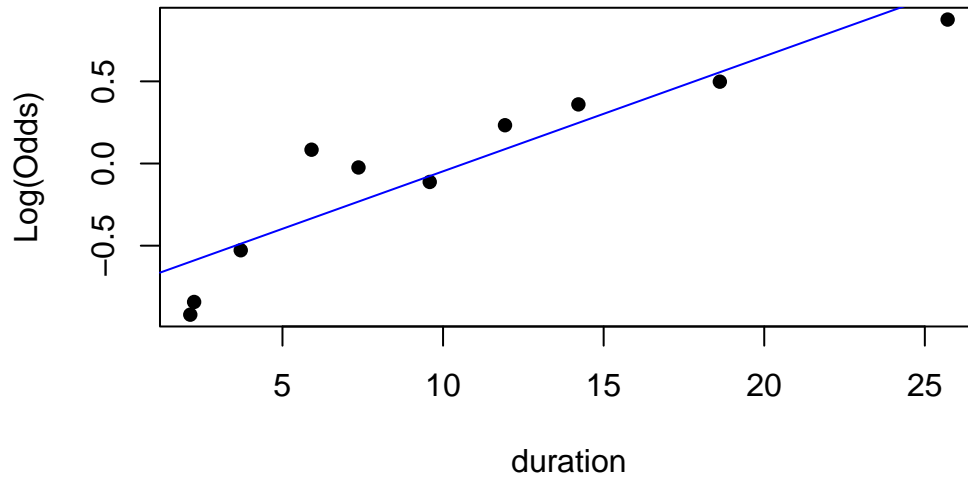
For duration, the z-statistic is -41.223, which follows a standard normal distribution under the null hypothesis. Our significance level is 0.05 and because our p-value $< 2e-16 < 0.05$, we reject the null hypothesis. Holding all other predictors constant, there is evidence to suggest that duration will be predictive of the odds that the flight price is above 7,000 INR. Holding all other predictors constant, for one hour increase in flight duration, the odds of the flight costing more than 7000 INR is predicted to be multiplied by $\exp(7.375e-02) = 1.001$.

For the number days before booking the flight, the z-statistic is 41.129, which follows a standard normal distribution under the null hypothesis. Our significance level is 0.05 and because our p-value $< 2e-16 < 0.05$, we reject the null hypothesis. Holding all other predictors constant, there is evidence to suggest that the number of days before booking the flight will be predictive of the odds that the flight price is above 7,000 INR. Holding all other predictors constant, for one hour increase in the number of days before booking the flight, the odds of the flight costing more than 7000 INR is predicted to be multiplied by $\exp(9.341e-04) = 1.077$.

For departure time, compared to afternoon departure which is acting as a the baseline, the z-statistic for flights departing in the early morning, morning, evening, night, and late night are 3.747, 2.101, -5.681, -7.898, and -0.166, respectively, which follows a standard normal distribution under the null hypothesis. Our significance level is 0.05, and p-values of 0.000179, 0.035669, 1.34e-08, and 2.84e-15 for early morning, morning, evening, and night respectively < 0.05 . For late night, p-value = 0.867847 > 0.05 . This suggests that we reject the null hypothesis for early morning, morning, evening, and night departure times and can conclude there is evidence to suggest that those departure times will be predictive of the odds that the flight price is above 7,000 INR while holding all other predictors constant and comparing to afternoon departure time. This also suggests that we fail reject the null hypothesis for late night departure time and can conclude that there is not enough evidence to suggest that a late night departure time will be predictive of the odds that the flight price is above 7,000 INR while holding all other predictors constant and comparing to afternoon departure time. Holding all other predictors constant, comparing to an afternoon departure time, the odds of the flight costing more than 7000 INR is predicted to be multiplied by $\exp(1.493e-01) = 1.161$, $\exp(8.292e-02) = 1.086$, $\exp(-2.265e-01) = 0.797$, $\exp(-3.665e-01) = 0.693$, and $\exp(-1.506e+01) = 2.881$ for early morning departure, morning departure, evening departure, night departure, and late night departure, respectively.

For destination city, compared to Dehli which acts as the baseline, the z-statistic is 3.392, which follows a standard normal distribution under the null hypothesis. Our significance level is 0.05 and because our p-value = 0.000694, we reject the null hypothesis. Holding all other predictors constant, there is evidence to suggest that destination city will be predictive of the odds that the flight price is above 7,000 INR. Holding all other predictors constant, comparing to leaving from Mumbai, the odds of the flight costing more than 7000 INR is predicted to be multiplied by $\exp(2.498e-02) = 1.025$.

Diagnostics



Based on the graphs above, since the points are close to the line, we can say that linearity is fairly assumed for the two continuous variables in our model.

Discussion

Our research question asked for flights that are departing out of cities were in, whether source destination, duration, departure time, and number of days before booking the flight influence the predictive odds of a flight being priced above 7,000 INR. The results of hypothesis testing reveal that source destination, duration, most departure times, and days before booking the flight influence the predictive odds of a flight being priced above 7,000 INR while late night departure time appears to not influence predictive odds. This information is important to consider while booking flights and looking for a less expensive option. For example, to achieve a better flight price, one might opt for flights that leave late at night or book their flight many days in advance.

This study has potential limitations. The data does not reflect general flight trends as the data was representative of flights that left from only select cities in India. This sample bias effects our ability to generalize over overall relationships. Because we did not collect this data, we have no way of correcting this limitation. However, in the future, the data collection could have been expanding to other cities, especially those from a variety of countries.

Furthermore, our data doesn't look at a few important factors that influence flight price. Holidays and other big events also influence flight traffic, hence it would be important to take these factors into consideration while creating a model on the influence of flight prices. It is hard to keep track of this data and perhaps we would need to use mutliple models to keep track of this.

Because we are working with a categorical binary variable, we chose to use logistic regression because logistic regression effectively models categorical and binary variables. The two logistic regression assumptions we considered during our analysis were independence within the observations and linearity between the log-odds of the response and the continuous predictors. In this case, we do not have independence, because knowing information about one flight might tell you information about another flight, especially if the observation corresponds to a recurring flight. In order to correct this failed assumption, we might in the future consider recurring flights or flights with more or less identical characteristics as a single flight. In this case, we do have linearity: the empirical logit plots show linearity for our continuous variables.

In terms of future work, many additional variables are available in the data set but were left out of our statistical analysis and could be included in a future analysis to understand how those other variables might affect flight prices. Additionally, we limited the number of cities used in our model; future work could include exploring flight prices out of some of those other cities. Lastly, one could choose a new research question and outcome variable to discover other relationships relating to flight data.

Sources:

<https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction> <https://www.nerdwallet.com/article/price-tracker> Skyscanner <https://www.skyscanner.co.in/flights/domestic-country-flights/in/cheap-domestic-flights-in-india#:~:text=How%20much%20are%20domestic%20flights,flights%20is%20around%20%E2>