**Related Work**

There has been previous work on using machine learning to estimate airfares.

Papadakisiii investigated how airline ticket prices fluctuate over time by isolating numerous parameters that might influence price fluctuation and determining their relationship.

Etzioni et al. conducted a 41-day pilot study using 12,000 price observations. Their multi-strategy data mining algorithm (Hamlet) produced a prediction model that might save consumers a significant amount of money on airline tickets.

Rama Murthyii created a model to predict airline ticket costs, with an emphasis on how different factors affect ticket pricing.

Groves and colleagues suggested a technique for predicting the expected minimum price of all flights on a given route. The model was also used to forecast the prices with a variety of target attributes including estimates from a single flight, nonstop only flight, and so on.

**Introduction**

Different airline industries globally have different rules and regulations due to which airline flight pricing strategies have developed into complex structures with sophisticated mathematical models that drive the price strategies of flight tickets if all the factors and characteristics are considered. Traditional factors or variables such as distance, duration, and ticket class (economy and business) play a significant role in dominating the pricing strategy. The previous research on flight prediction uses statistical models such as linear regression and the speculation that the relationship between dependent and independent variables is a linear relationship, which is not true in many cases. The advancement of statistical models and machine learning (ML) makes it possible to let us infer rules and simulate fluctuations in flight costs based on a variety of factors. The difficulties in acquiring access to the data in the former case make replicating the results and expanding the study practically impossible. The issue with the latter is that because there are several sites and people also book tickets from agents and directly from the airport, each online booking site's transaction records represent a small percentage of total ticket sales in the entire market. Hence, the data is skewed, as most of the data is not normally distributed and thus does not reflect the true nature of the entire market. Thus, we took the dataset, which is publicly available on Kaggle. The purpose of this research is to analyze and examine the relationships between several factors that might be playing a significant role in the pricing strategy of airlines and to develop a statistical modeling framework to predict flight prices.

**Why did we choose this topic?**

When we are traveling by flight, we are always confused by changing airline ticket prices. Why are flight prices so exorbitant one minute and reasonable the next? What is the optimal time to book a flight to obtain the best price? When is the best time to get those great airplane ticket offers and discounts? As a result, we have compiled an interesting list of points to consider when purchasing an airline ticket. As a result, we are inclined to conduct research on predicting flight prices because it is relevant to our daily lives.

**An overview of the dataset**

The dataset contains statistics on flights between India's top six metro cities. The dataset contains 300,153 data points and 11 variables.

The different characteristics of the cleaned dataset are described in the following sections.

1. **Airline:** The airline column contains the name of the airline firm. It has six distinct airlines.
2. **Flight:**   The flight code contains information regarding the plane.
3. **Source City:** The location from which the flight departs. It has six unique cities.
4. **Departure Time**: The derived categorical feature was created by grouping time periods into bins. It stores departure time information and has six unique time labels.
5. **Stops:** A three-valued categorical feature that counts the number of stops made between the origin and destination cities.
6. **Arrival Time:** This is a derived categorical feature created by grouping time periods into bins. It stores arrival time information and has six unique time labels.
7. **Destination City:** The location where the flight will land. It has six unique cities.
8. **Class**: A categorical feature containing seat class information; it has two distinct values: business and economy.
9. **Duration:** the total number of hours required to travel between cities.
10. **Days Left:** Calculated by subtracting the trip date from the booking date.
11. **Price:** Target variable stores ticket price information.

Table

Description automatically generated

**Data Preprocessing**

Data preprocessing is the process of converting raw data into understandable and usable forms. The datasets that we collected from Kaggle were the raw data for us, and the dataset can be characterized by incompleteness, inconsistencies, lack of behavior, and trends while containing errors. Thus, preprocessing is necessary to standardize and resolve all the characteristics mentioned earlier. During the data preprocessing stage, impossible data combinations such as arrival time and departure time (early morning, morning, night, and late-night) were handled. Missing values, null values, and redundancies were also addressed. Proper headings were given to column names, and we converted all the categorical variables into numeric, which we are going to use in the analysis and modeling. After preprocessing, the dataset is more reliable and relevant.

Table

Description automatically generated with medium confidence

**Normality Test**

The normal distribution, often known as “the Gaussian distribution”, that represents the distribution of values for a variable. It is symmetric distribution in which the majority of observations cluster around the center peak, and the probability for values further from the mean tapers down equally in both directions, with fewer outliers at the extremes of the data range.

Chart, histogram

Description automatically generated

Normality tests are used to determine if data is derived from a Gaussian distribution or whether a variable or sample has a normal distribution. We have two methods for determining the normality of data.

1. Shapiro-Wilks Statistical Test.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Statistics** | **P-Value** |
| **Source City** | 0.903 | 0.00 |
| **Departure Time** | 0.888 | 0.00 |
| **Stops** | 0.543 | 0.00 |
| **Arrival Time** | 0.896 | 0.00 |
| **Destination City** | 0.905 | 0.00 |
| **Class** | 0.583 | 0.00 |
| **Duration** | 0.956 | 0.00 |
| **Days Left** | 0.959 | 0.00 |
| **Price** | 0.752 | 0.00 |

**Result:** since all the p-values for all the variables are less than 0.05. Therefore, we will reject the null hypothesis. And thus, we can say that data is not normally distributed.

1. Plotting Q-Q plots.
2. Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generated
3. Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generatedChart, line chart

   Description automatically generatedChart, line chart

   Description automatically generatedChart, line chart

   Description automatically generated
4. re
5. re

**Result:** We plotted Q-Q plots for all variables to be clearer and to confirm the normality of our data. As we can see, there are horizontal lines on some of the plots, such as the destination city, the source city, class, arrival time, and departure time. These horizontal lines show the categorical nature of the variables. The points follow a strongly nonlinear pattern in the price and duration plots.

Thus, we can say that the data is not normally distributed with both the Shapiro-Wilk test and the Q-Q plot.

**References**

Etzioni, O., Tuchinda, R., Knoblock, C. A., & Yates, A. (2003, August). To buy or not to buy: mining airfare data to minimize ticket purchase price. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 119-128).

Rama-Murthy, K. (2006). *Modeling of United States Airline Fares--Using the Official Airline Guide (OAG) and Airline Origin and Destination Survey (DB1B)* (Doctoral dissertation, Virginia Tech).

Papadakis, M. (2014). Predicting Airfare Prices.

Groves, W., & Gini, M. (2011). A regression model for predicting optimal purchase timing for airline tickets.