Analysis and Modeling Domestic Indian Airline Ticket Prices

By Shumel Siraj, Lokesh Bokkisam & Pavani Samala

Introduction

Traveling via airplane continues to be a popular mode of transportation since it is a convenient way to visit both domestic and international destinations. However, this mode of transportation, unfortunately, burdens the consumer with the stressful process of purchasing an airplane ticket that accommodates their traveling needs at a reasonable price. Many consumers can relate to seeing exorbitant airfare costs that are fluctuating hourly, making it difficult to gauge what factors truly influence the price of airfare. Due to this ongoing uncertainty, we were inclined to conduct research regarding the cost of airline tickets and investigate which factors can predict price accurately.

There has been previous work on using machine learning to estimate airfares. For example, a previous project investigated how airline ticket prices fluctuate over time by isolating numerous parameters that might influence price fluctuation and determining their relationship (Papadakis, 2014). Another study 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 (Etzioni et al, 2003).

The purpose of this research is to analyze and examine the relationships between several factors and price, and in the process, develop a predictive modeling framework to predict flight prices. This summary paper will include background information about the dataset, the process of cleaning and preprocessing, exploratory data analysis (EDA), three different predictive models, and an analysis of our results.

Background of the Dataset

The dataset we analyzed was taken from Kaggle and contains data about domestic airlines. This dataset was created by merging information about economy and business class data. The dataset contains 300,153 observations and 11 variables. Each variable is described below:

- **airline:** (Airline Company) The six different airline companies are SpiceJet, AirAsia, Vistara, GO First, Indigo, and Air India.
- **Flight:** (Flight Code) The flight code contains the unique identification information of the plane.
- **Source_city:** (Source City) The five different source cities are Delhi, Mumbi, Bangalore, Hyderabad, and Chennai.
- **departure_time**: (Departure Time) The different departure times are categorized by early morning, morning, afternoon, evening, and night
- **stops:** (Number of Stops) The number of stops is categorized by zero stops, one stop, or two or more stops
- **arrival_time:** (Arrival Time) The different arrival times are categorized by early morning, morning, afternoon, evening, and night
- **destination_city:** (Destination City) The five different destination cities are Delhi, Mumbi, Bangalore, Hyderabad, and Chennai.
- **class**: (Ticket Class) The ticket class options are economy class and business class.
- **duration:** (Duration of Flight) This variable is measured in hours.
- **days_left:** (Days Between Booking and Travel) This variable is measured in days.
- **price:** (Ticket Price) The price of the airline ticket is given in Rupees (Indian currency).

Data Preprocessing

To ensure accurate EDA and modeling results, we checked for inconsistencies within the dataset and preprocessed a few variables. First, the dataset was checked for missing values, null values, and redundancies. Once those inconsistencies were removed, the categorical variables which we will use in our technical analysis, like source city, destination city, arrival time, departure time, and stops, were converted into numeric values.

Summary of Dataset

We have a table highlighting some of the statistics behind our data. The price ranges from 1105 to 12307 rupees. On average, a domestic airline ticket costs 20889 rupees, and the most expensive ticket is 123071 rupees. On average, people booked a flight ticket 12 days in advance, and the average duration of the flight was 12 hours and had an average of 1 stop.

Figure 1. Summary of Dataset Statistics

	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
count	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000
mean	2.202976	1.867814	0.924312	2.699087	2.268316	0.311464	12.221021	26.004751	20889.660523
std	1.683252	1.416183	0.398106	1.351441	1.688644	0.463093	7.191997	13.561004	22697.767366
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.830000	1.000000	1105.000000
25%	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	6.830000	15.000000	4783.000000
50%	2.000000	2.000000	1.000000	3.000000	2.000000	0.000000	11.250000	26.000000	7425.000000
75%	4.000000	3.000000	1.000000	4.000000	4.000000	1.000000	16.170000	38.000000	42521.000000
max	5.000000	5.000000	2.000000	5.000000	5.000000	1.000000	49.830000	49.000000	123071.000000

SMART Question

Our SMART question asks: **Can the price of airline tickets be predicted by the number of stops, duration of flight, and days left until departure, or is it better predicted by adding categorical variables such as departure and arrival locations**.

Answering this question can help people understand which factors determine the price of airline tickets and potentially search for tickets based on these variables.

Normality Test

Before we began EDA, we performed two normality tests to determine whether our dataset followed a normality distribution or not. The two methods we used were the Shapiro-Wilks Statistical Test and Q-Q plots.

Shapiro-Wilks Statistical Test:

The Shapiro-Wilks Test rejects the hypothesis of normality when the p-value is less than 0.05. As you can see from the table below, all the p-values for all the variables are less than 0.05. Therefore, we will reject the null hypothesis and conclude that data is not normally distributed.

Table 1:Shapiro-Wilks Statistical Test Summary

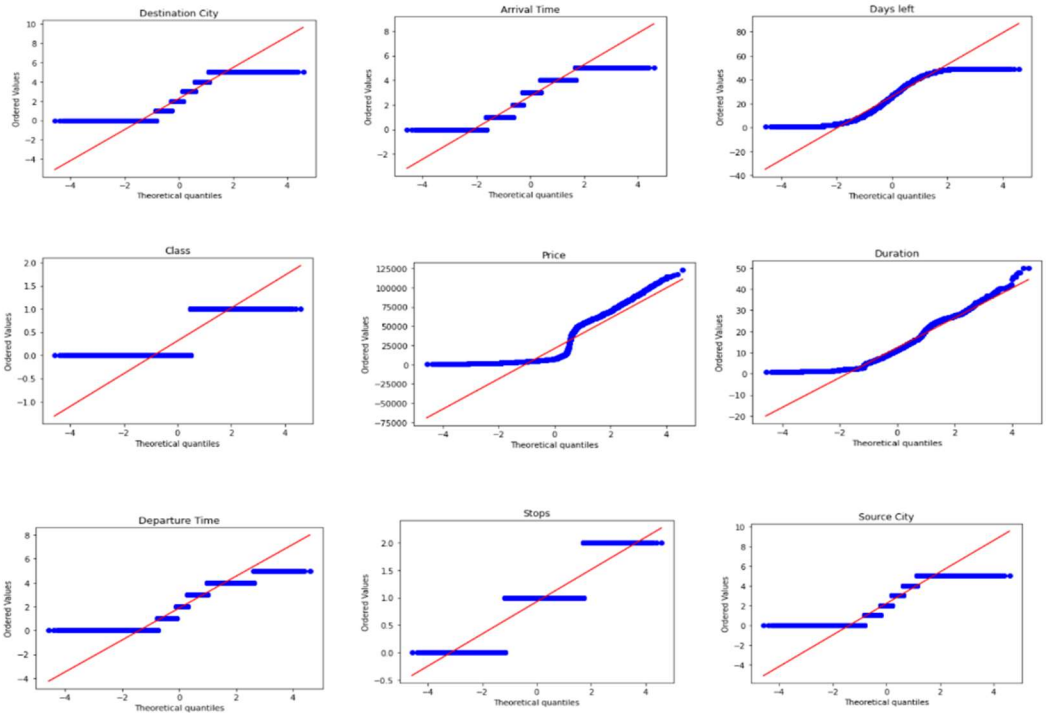
Variables	Statistics	P-Value
Source City	0.903	0.00
Departure Time	0.888	0.00
Stops	0.543	0.00
Arrival Time	0.896	0.00
Destination City	0.905	0.00
Class	0.583	0.00
Duration	0.956	0.00
Days Left	0.959	0.00
Price	0.752	0.00

Plotting Q-Q plots

In addition to the Shapiro-Wilks Test, we plotted Q-Q plots for all the variables. As shown in the plots below, destination city, the source city, class, arrival time, and departure time all have horizontal lines, which indicate the categorical nature of these variables. Also, the price and duration plots have data points that deviate from the diagonal line, showing that the dataset is not normally distributed. As the Shapiro-Wilk test showed us, the Q-Q plots again tell us that the data is not normally distributed.

Although our data is not normally distributed, we decided it was not necessary to normalize it since the variables did not span across extremely different ranges.

Figure 2. Q-Q Plots



EDA

Correlation Matrix:

The correlation matrix below shows us how each variable is correlated with price. We can see that the class variable is highly correlated with price since its correlation value was 0.937. Duration and the number of stops are also correlated with each other with a value of 0.468. Even though correlation does not mean causation, the correlation map suggests that we may incorporate class, duration, and the number of stops somewhere in our analysis and investigate them further.

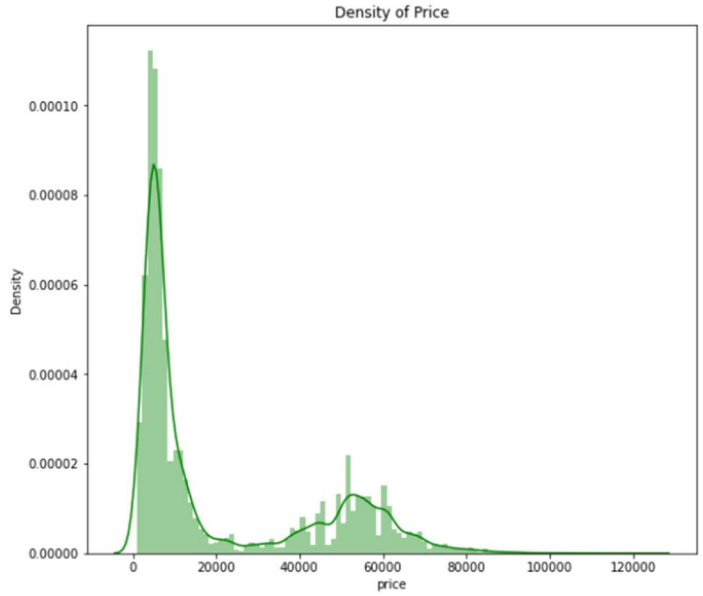
Figure 2. Correlation Map

	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
source_city	1.000000	0.002259	0.050644	0.028616	-0.205550	-0.000888	0.056980	0.010491	0.013490
departure_time	0.002259	1.000000	-0.068986	-0.079679	0.024507	0.030956	0.132773	-0.000222	0.020948
stops	0.050644	-0.068986	1.000000	0.046436	0.109122	0.001027	0.468059	-0.008540	0.119648
arrival_time	0.028616	-0.079679	0.046436	1.000000	-0.085398	-0.022473	-0.123949	-0.000700	-0.001019
destination_city	-0.205550	0.024507	0.109122	-0.085398	1.000000	0.007707	0.125406	0.000016	0.019641
class	-0.000888	0.030956	0.001027	-0.022473	0.007707	1.000000	0.138710	-0.013039	0.937860
duration	0.056980	0.132773	0.468059	-0.123949	0.125406	0.138710	1.000000	-0.039157	0.204222
days_left	0.010491	-0.000222	-0.008540	-0.000700	0.000016	-0.013039	-0.039157	1.000000	-0.091949
price	0.013490	0.020948	0.119648	-0.001019	0.019641	0.937860	0.204222	-0.091949	1.000000

Price Distribution:

To have a basic understanding of how price is distributed in our dataset, we created a density plot. The plot below shows that price is skewed to the left.

Figure 3. Density Curve of Price



Box Plots:

To understand whether the name of an airline company influenced the price, we made a box plot (Figure 4) to compare the prices. The box plot shows that SpiceJet, AirAsia, GO First, and Indigo all have smaller price ranges than Vistara and Air India. We decided to further investigate this observation and created a second boxplot that included class as a parameter. From the second boxplot (Figure 5), we learned that while all the airline companies offer economy class tickets, only Vistara and Air India offer business class tickets.

Figure 4. Boxplot Comparing Prices Between Airlines

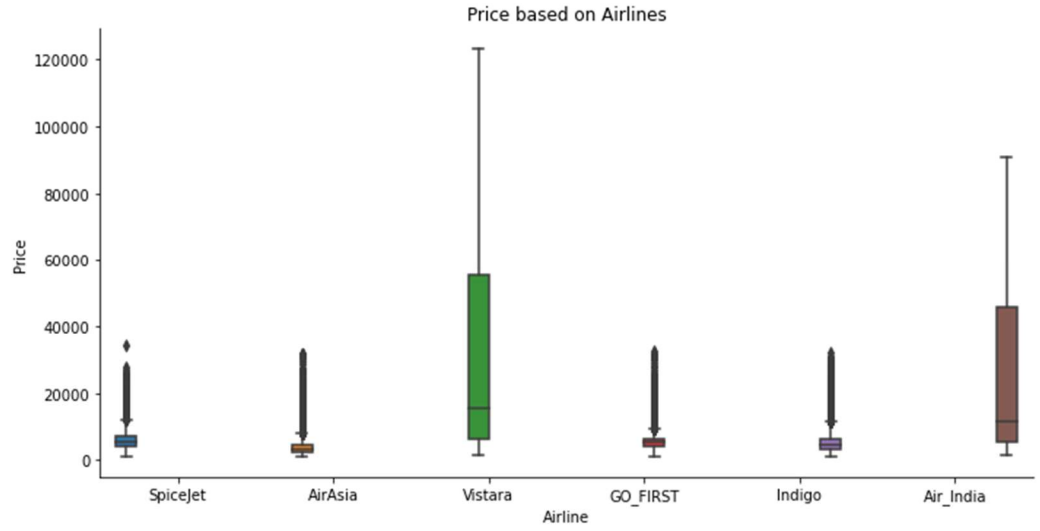
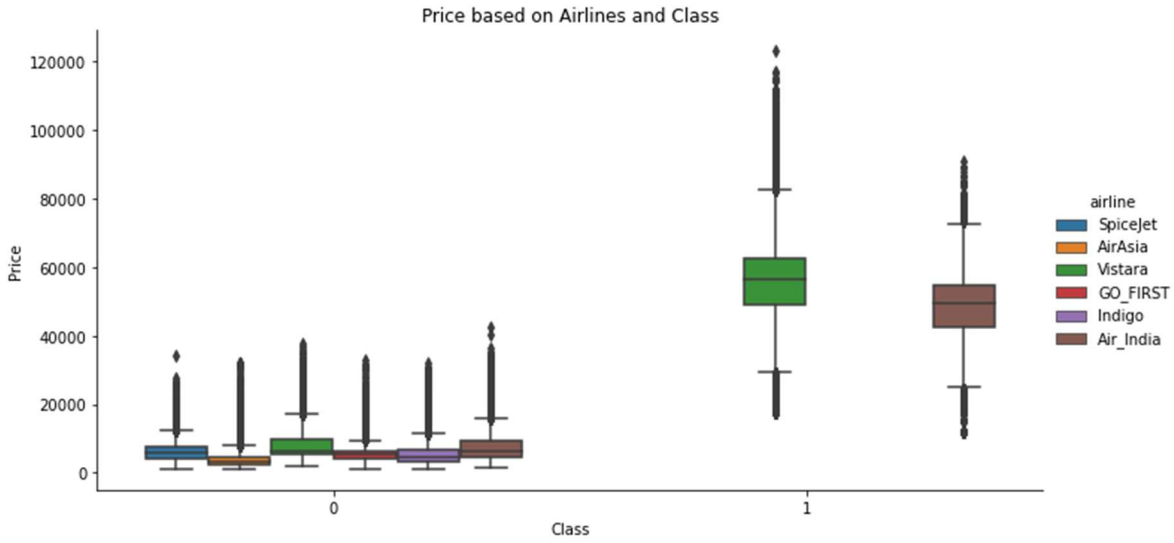


Figure 5. Boxplot Comparing Prices Between Airlines and Class



The following two boxplots compare the source and destination cities against price because we wanted to see if location heavily influenced price. We can see that for both the source city and destination city, Kolkata has the widest range of prices and Delhi has the smallest range of prices. Delhi also has the most outliers for both the source and destination city.

Figure 5. Boxplot Comparing Prices Between Airlines and Class

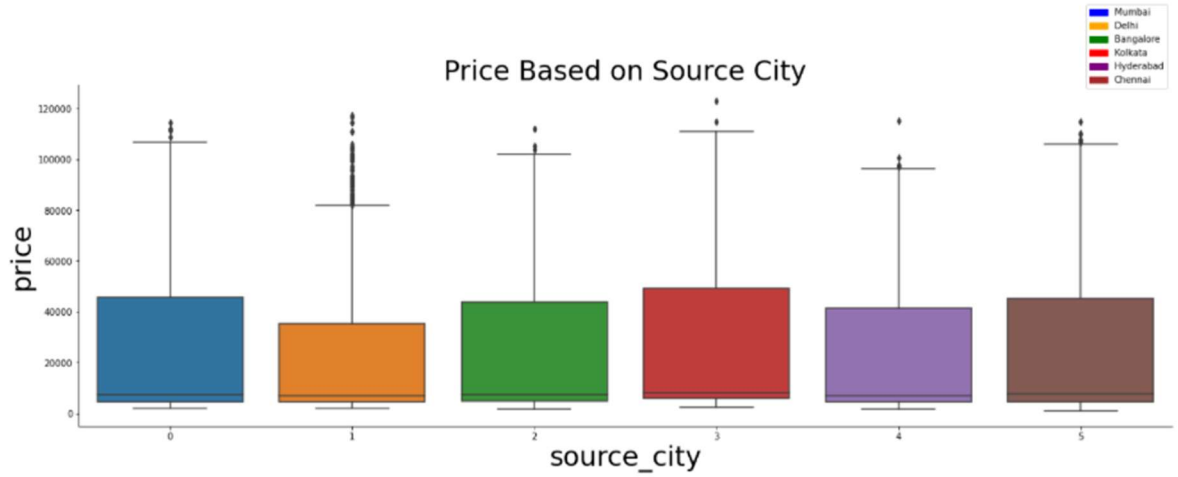
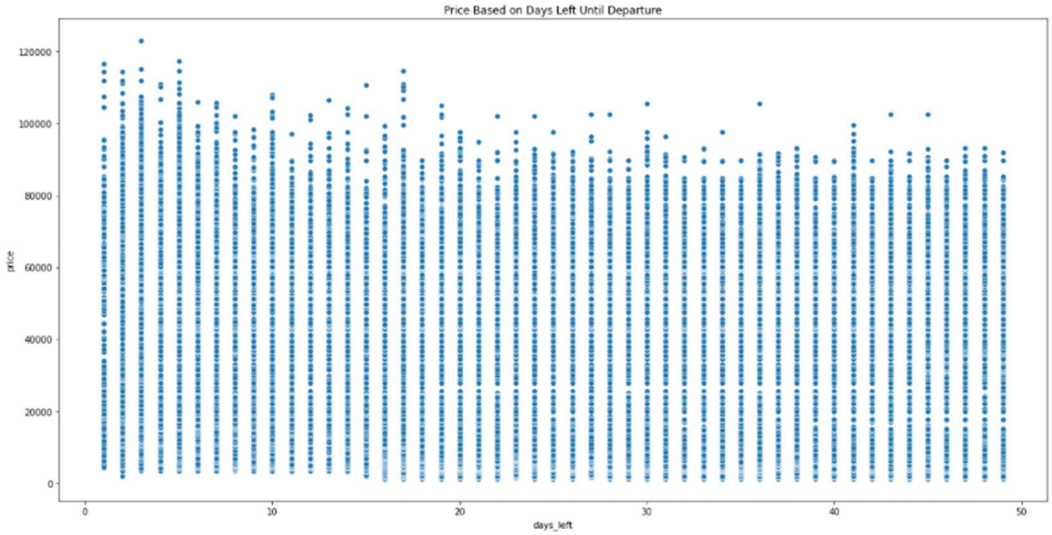


Figure 6. Boxplot Comparing Prices Between Airlines and Class

Scatterplot:

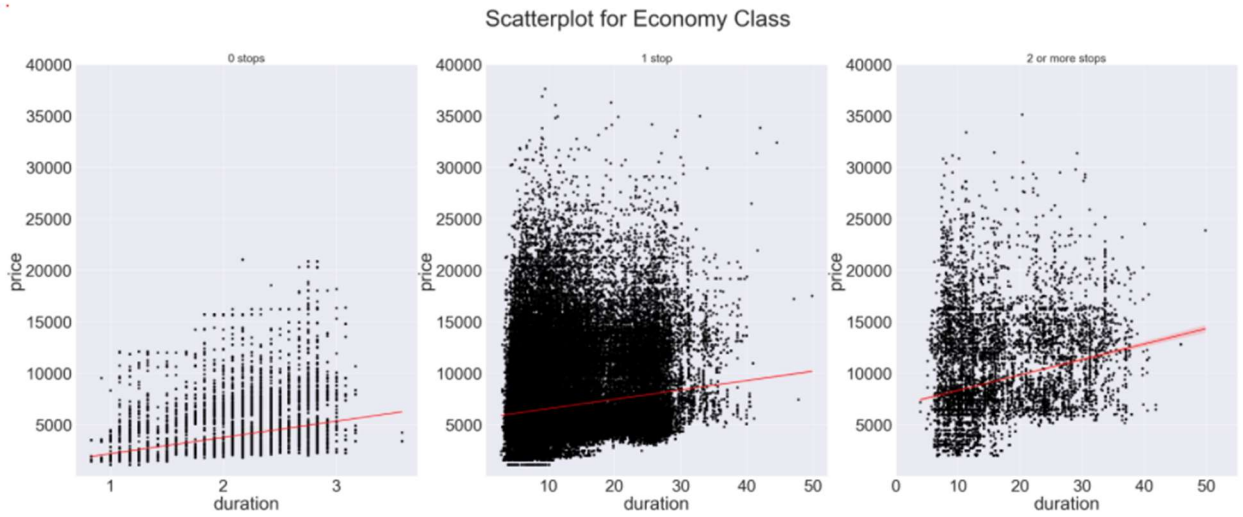
Another factor we analyzed was the days left until departure. The scatterplot below shows it plotted against price. Observing this plot here shows that the price is slightly higher when there are fewer days remaining until departure.

Figure 7. Scatterplot Comparing Prices and Days Left Until Departure



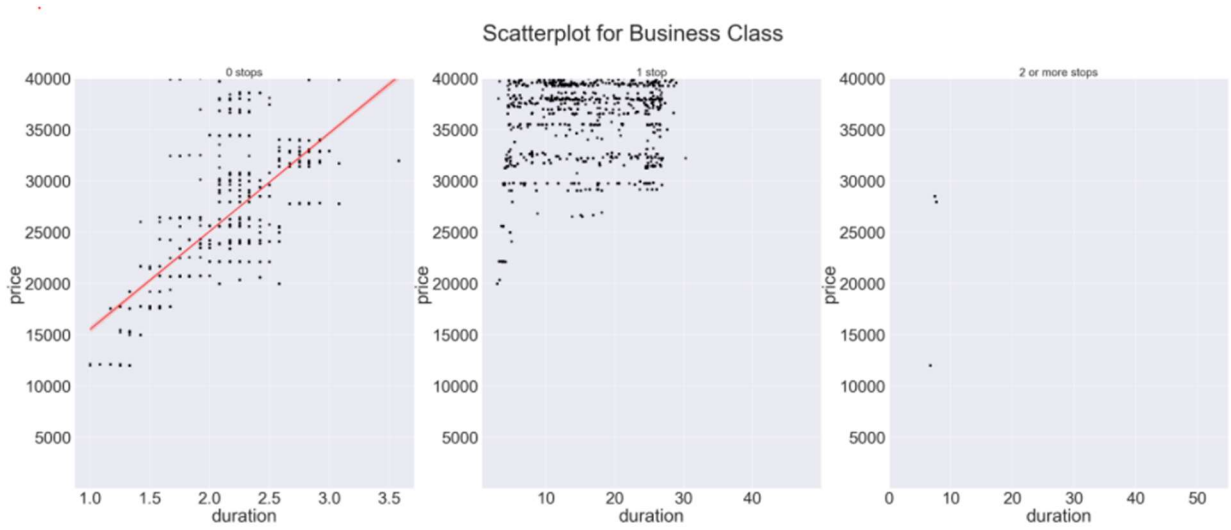
Since the correlation map showed us that stops and duration were correlated, we wanted to see how those two variables related to price. The following scatterplot shows the economy class ticket prices against flight duration for zero stops, one stop, and 2 or more stops. From general knowledge, we would conclude that airplane tickets are less expensive for longer air travel time and routes with more stops. However, the scatter plots show the converse and the prices increase with the number of stops.

Figure 8. Scatterplot Comparing Prices Between Flight Duration and number of Stops for Economy Class



The final scatterplot below shows the business class ticket prices against the duration of flight for each stop. Although there are not many data points, we again see that the price for airplane tickets that have 1 stop is more expensive than airplane tickets that have zero stops.

Figure 9. Scatterplot Comparing Prices Between Flight Duration and number of Stops for Business Class



Modeling

To further our understanding of our data and investigate which variables best predict the cost of airline tickets, we used three different regression models: linear regression, decision trees, and knn regression. To obtain accurate results, we divided our data based on ticket class and built models separately for each class. This procedure was prioritized since we learned that four of the airlines in our dataset did not offer business class tickets. For the decision trees and knn models particularly, the optimal model-specific parameters (min_depth, min_sample_leaf, random_state, and neighbors) were chosen by running a few iterations and seeing which combination was yielding balanced results.

Also, in order to benchmark and compare which variables predicted price effectively, we first built each model only using three variables which were the number of stops, duration of flight, and the days left until take off. The next set of models was built using the same three variables and adding the source city and destination city. To guide our analysis, we recorded the R^2 values along with key values such as p-values for linear regression, mean squared error (MSE) for decision trees, and score values for knn regression.

Summaries and specific detail of each model are provided below:

Table 2. Summary of Linear Regression Models

Linear Regression				
	economy		business	
	p-value	R^2	p-value	R^2
stops, duration, days left	all<0.05	0.432	all<0.05	0.385
stops, duration, days left, source_city, destination_city	11 p-values > 0.05 2 p-values = nan	0.454	14 p-values > 0.05 2 p-values = nan	0.474

Table 3. Summary of Decision Tree Regression Models

Decision Tree Regression				
	economy		business	
	MSE	R^2	MSE	R^2
stops, duration, days left	~5,714,276	0.59	~93,013,576	0.452
stops, duration, days left, source_city, destination_city	~5,405,555	0.6166	~85,186,687	0.497

* please note that the following function parameters were used since they were yielding an adequate balance between high R^2 and low MSE value: min_depth=8, min_sample_leaf=1, and random_state=50

Table 4. Summary of KNN Regression Models

KNN Regression				
	economy		business	
	score	R^2	score	R^2
stops, duration, days left	0.034	0.216	0.0211	0.122
stops, duration, days left, source_city, destination_city	0.178	0.329	0.195	0.212

* please note that the following function parameter was used since it was yielding an adequate balance between high R^2 , high score values, and short runtime: neighbors=100

Results

As previously mentioned, the R^2 values and p-values, MSE, and score values were used to holistically analyze the performance of the three models respectively.

Linear regression:
As shown in Table 1, the p-values for the first set of linear regression models, which only used the variables stops, duration, and days left, were all less than 0.05 for both classes. The R^2 for the economy models was 0.432 and the R^2 value for the business model was 0.385. After including the source and destination city into the linear regression models, the R^2 values increased for both models; however, 11 interaction terms were greater than 0.05 for the economy model and 14 interaction terms were greater than 0.05 for the business model. At best, linear regression was only explaining 45% of the economy class data and 47% of the business class data.

Decision Tree Regression:
Table 2 summarizes the important values for decision trees. The MSE values are approximately 5,414,276 and 93,013,576 for the first set of economy and business models respectively. The R^2 values are 0.59 for the economy decision tree model and 0.452 business decision tree model. After adding the source and destination city to the decision trees, both R^2 values increased to 0.6166 and 0.497 for the economy and business model respectively. The MSE values also decreased for both models, but the business model saw a larger decrease than the economy model. At best, the decision tree models were explaining 61.6% of the economy class data and 49.7% of the business class data.

KNN Regression:

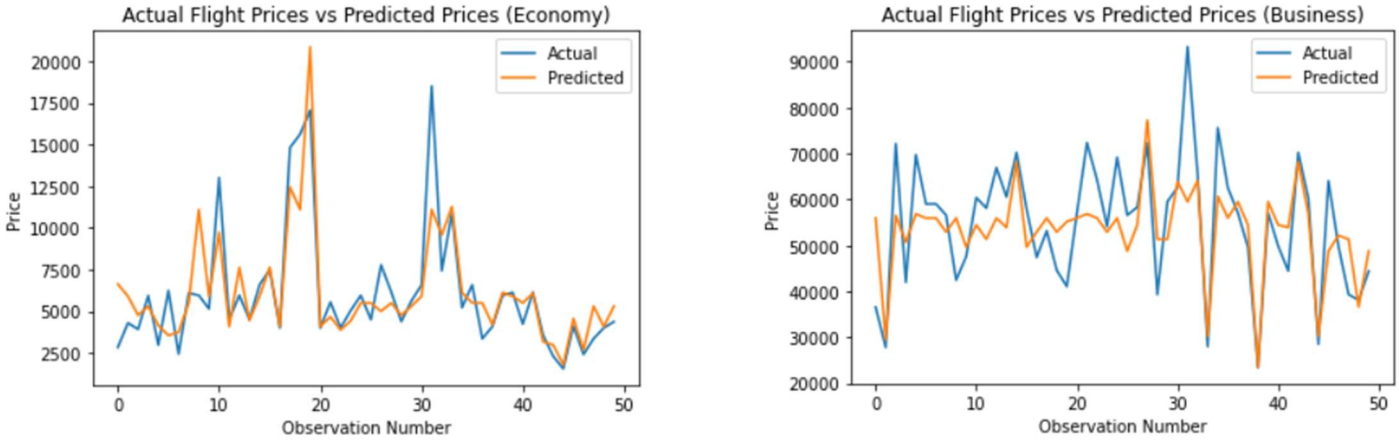
As shown in Table 3, the score values of the first set of KNN regression models were 0.034 and 0.0211for the economy and business model respectively and R^2 values are 0.216 and 0.122 for the economy and business model respectively. After adding source and destination cities to the KNN models, both the score and R^2 values increased marginally for economy class and business class models. The knn models were explaining the least amount of data among the three models.

Analysis

Looking at the results, including the source and destination variables increased the R^2 value for all of the models. KNN had the lowest R^2 values and the decision trees had the highest R^2. In addition to having the highest R^2 value, the decision trees also had a decrease in MSE values in the second set of models. Therefore, when answering the SMART question, the price of airfare for domestic Indian airlines is better predicted using the number of stops, duration of flight, days left until departure, source city and destination city. In addition, the best model to use in conjunction with these variables is a decision tree.

To provide more information on how well the decision tree performed, 50 random observations of the actual and predicted price were plotted below. The predicted and actual prices do overlap in many places, but the models clearly show room for improvement.

Figure 10. Visual Comparison of Actual and Predicted Prices for Economy and Business Class



With our current decision tree models, we can predict the price of other domestic Indian airlines that were not included in our dataset, given that they have similar operation and business strategies. Consumers may potentially use this model to determine which traveling logistics to prioritize based on their traveling budget.

Future Work

As mentioned above the decision tree model has room for improvement. If this work were to continue, a more robust analysis can be performed by including data concerning airport fees, travel distance, and layover locations. Having this data can also segway into building more decision trees and developing a random forest model to better understand the determining factors of airline ticket prices.

References

Borandabak. (2022, February 26). *Flight price-data analysis*. Kaggle. Retrieved April 29, 2022, from https://www.kaggle.com/code/borandabak/flight-price-data-analysis/data

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).

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