Analysis and Prediction of Flight Price

Project By - Aditya Aryan

Overview

- About
- Business Value
- Exploratory Data Analysis
- Feature Engineering
- Model
- Results
- Conclusion

About

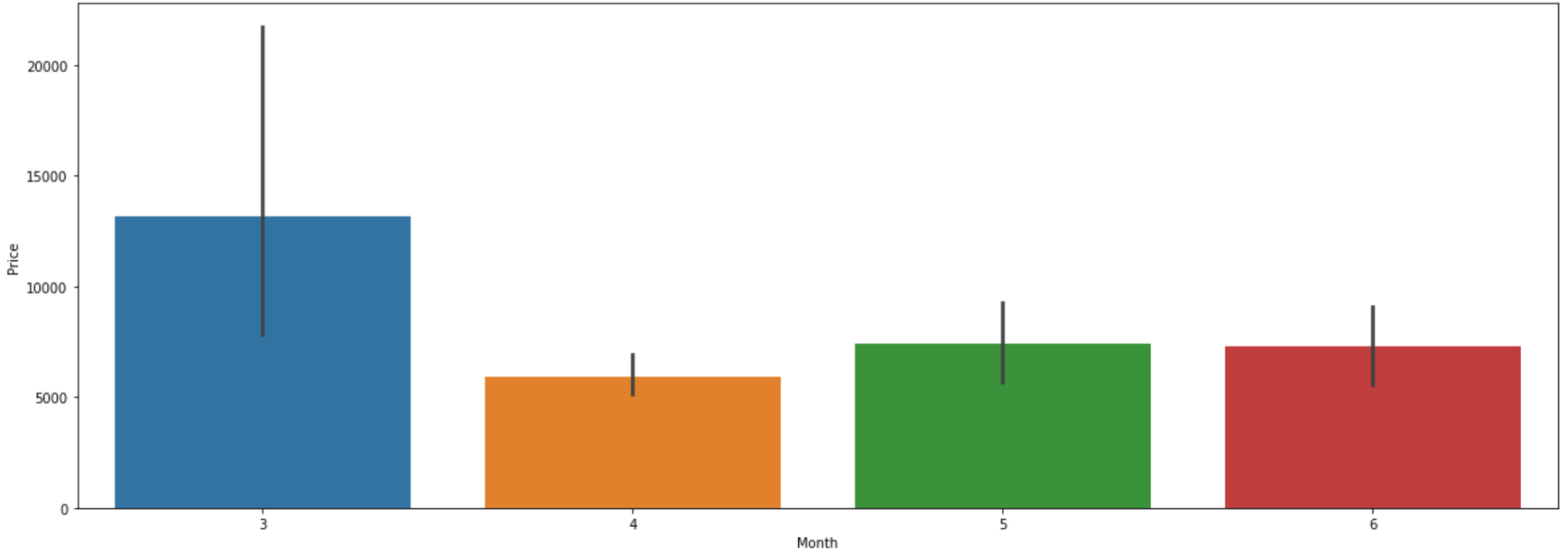
Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, and it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable. As data scientists, we are going to prove that given the correct data anything can be predicted. Here you will be provided with prices of flight tickets for various airlines for the year 2019 and between multiple cities. Size of training set: 10683 records. Our goal here is to create a machine-learning model for predicting flight ticket prices.

Business Value

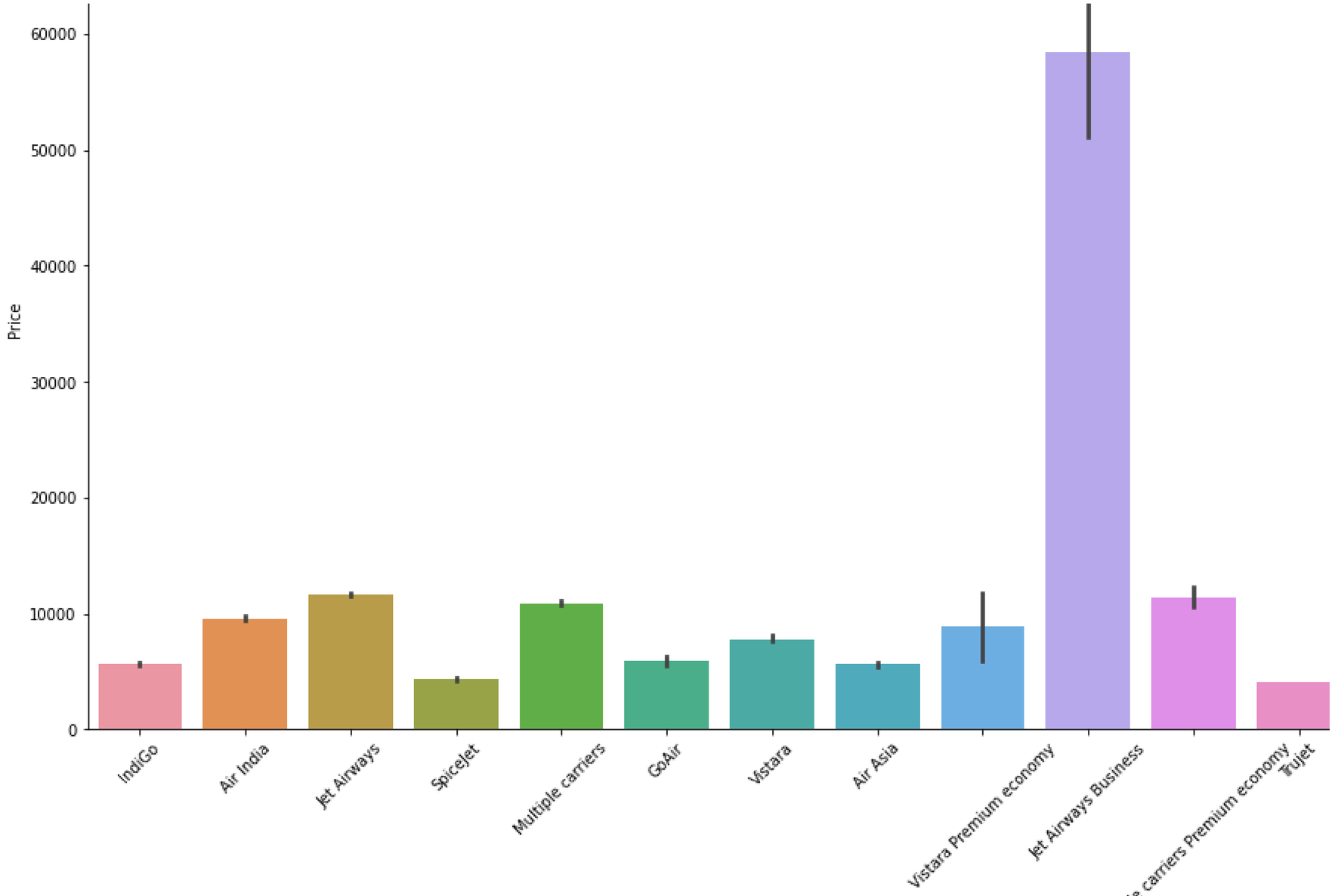
The flight ticket price in India is based on the demand and supply model with few restrictions on pricing from regulatory bodies. It is often perceived as unpredictable and, a recent dynamic pricing scheme added to the confusion. The objective is to create a machine learning model for predicting the flight price, based on historical data, which can be used for reference prices for customers as well as airline service providers

Exploratory Data Analysis

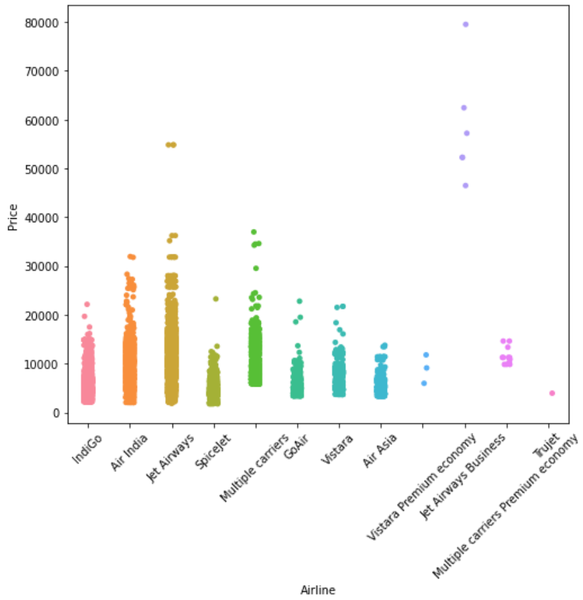
The sum of fare is maximum towards the month-March, which can be because of Holi vacations in the month of March for schools/colleges, hence most families are also generally going for vacations around this time. The count of flights is lowest in April as schools, and colleges have exams and offices are busy as it is the end of quarter 1. From May onwards, it starts to increase.

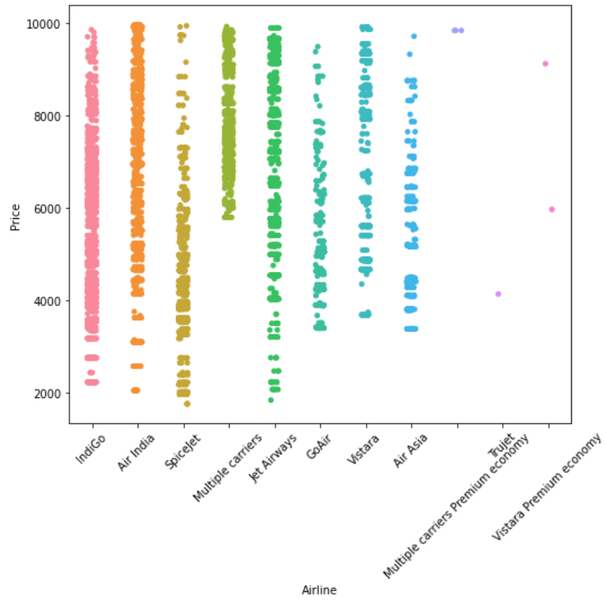


There are slight differences between each companies on this graph, Spice Jet and Truejet seems to have the cheapest flights when Jet Airways Business and Vistara Premium are more expensive. However it looks like Jet Airways business tickets are a far more expensive than the business classes.

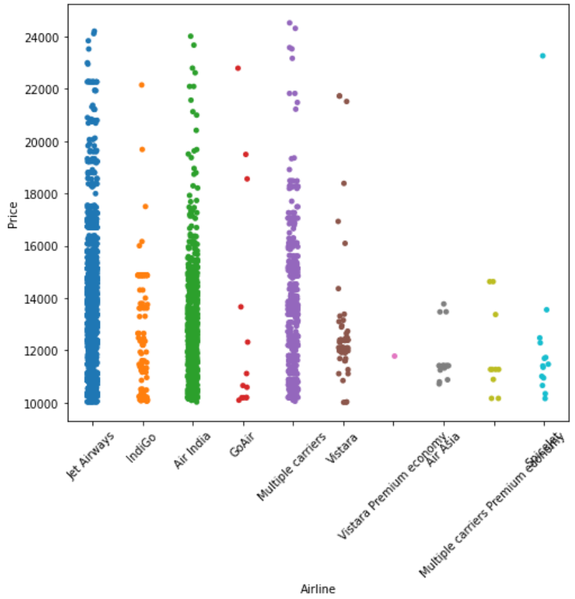


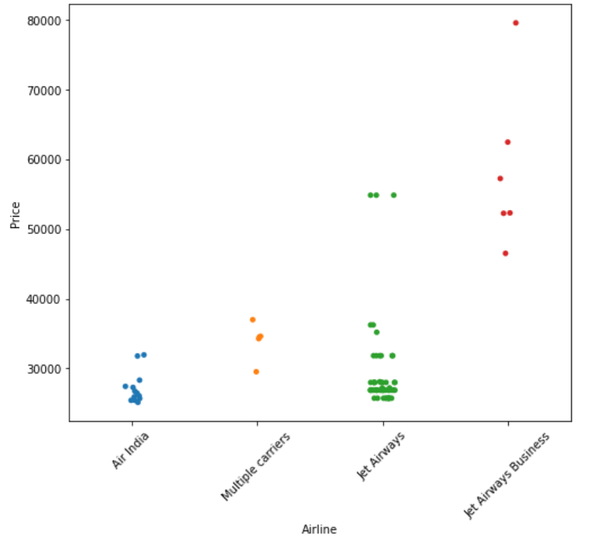
To visualize the price distribution for airlines, I created three price bins: economy (>5000 and <10000), premium economy flights(>10000 and <=25000), and first-class (>25,000).





For economy, the price is heavily distributed towards Indigo, Spice Jet, Air India, Jet Airways

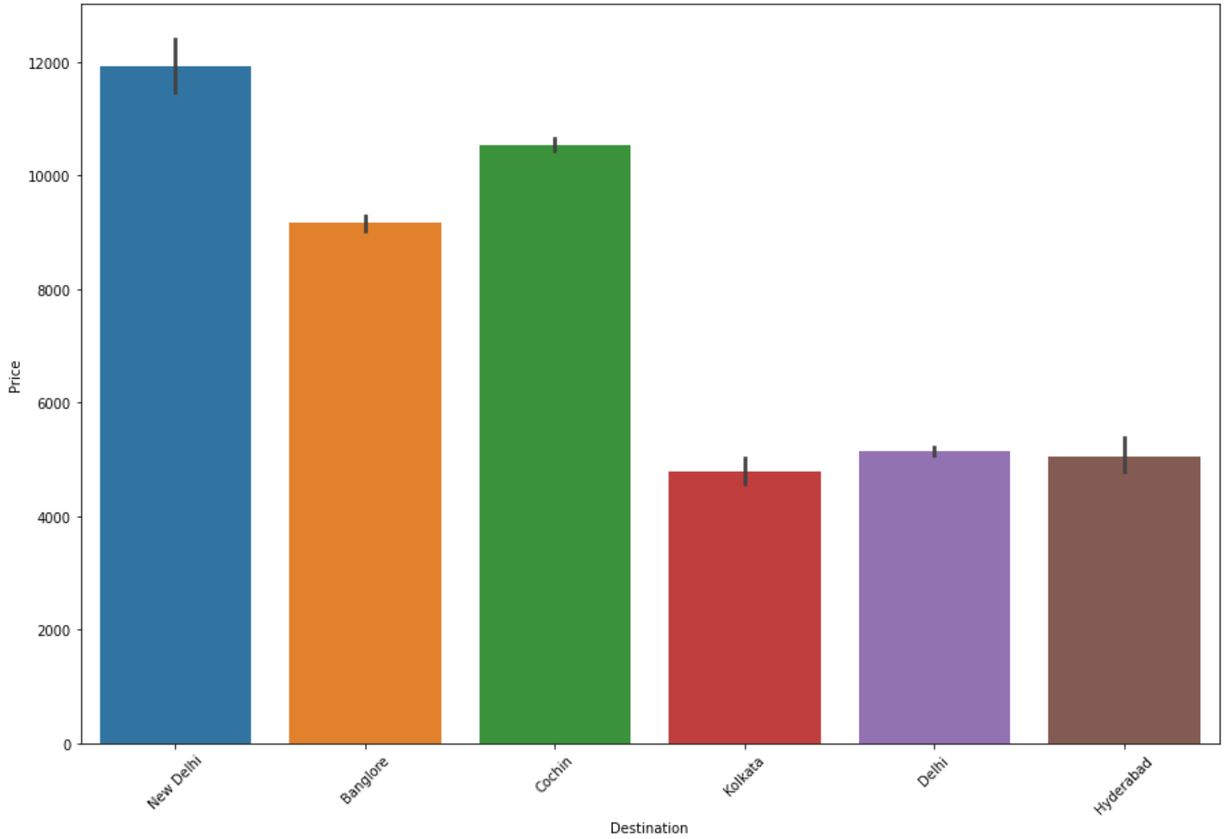


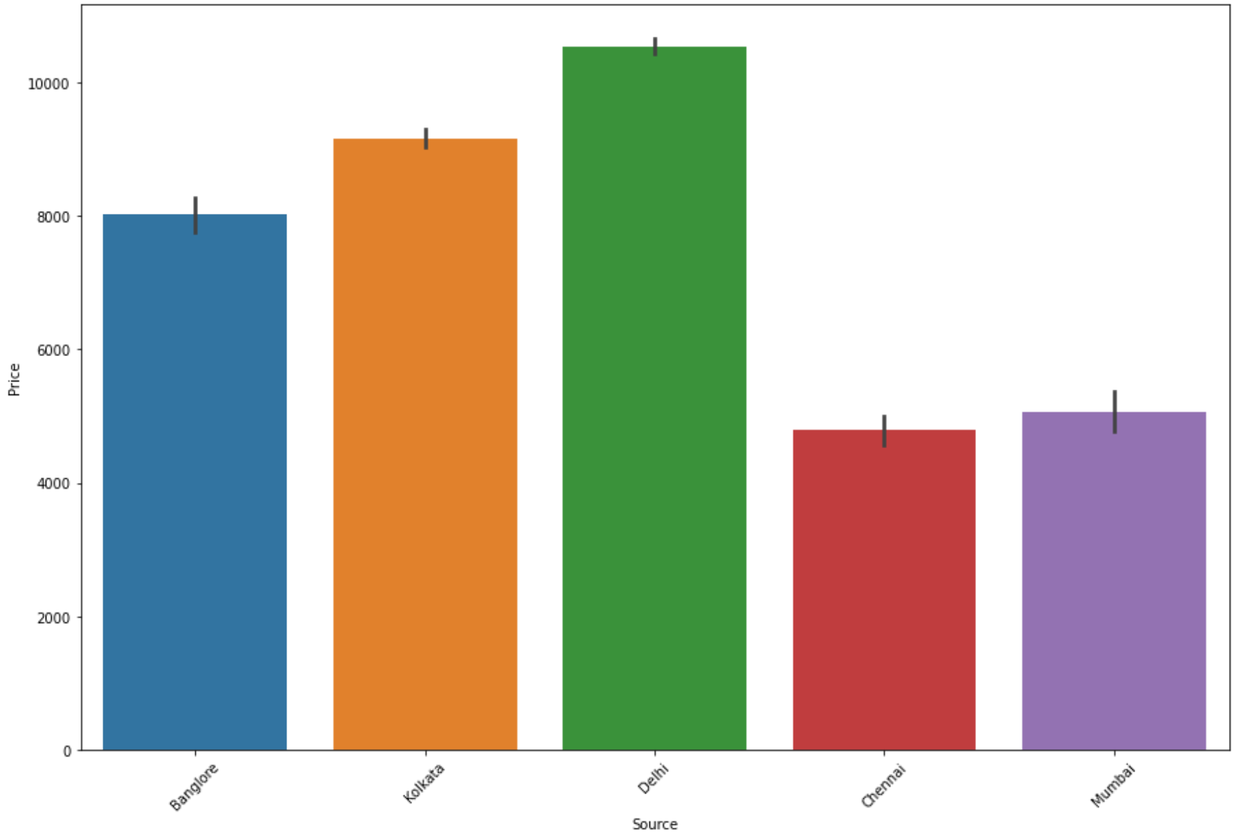


For premium economy flights, the price is heavily distributed towards Air India, Jet Airways and Multiple Carriers.

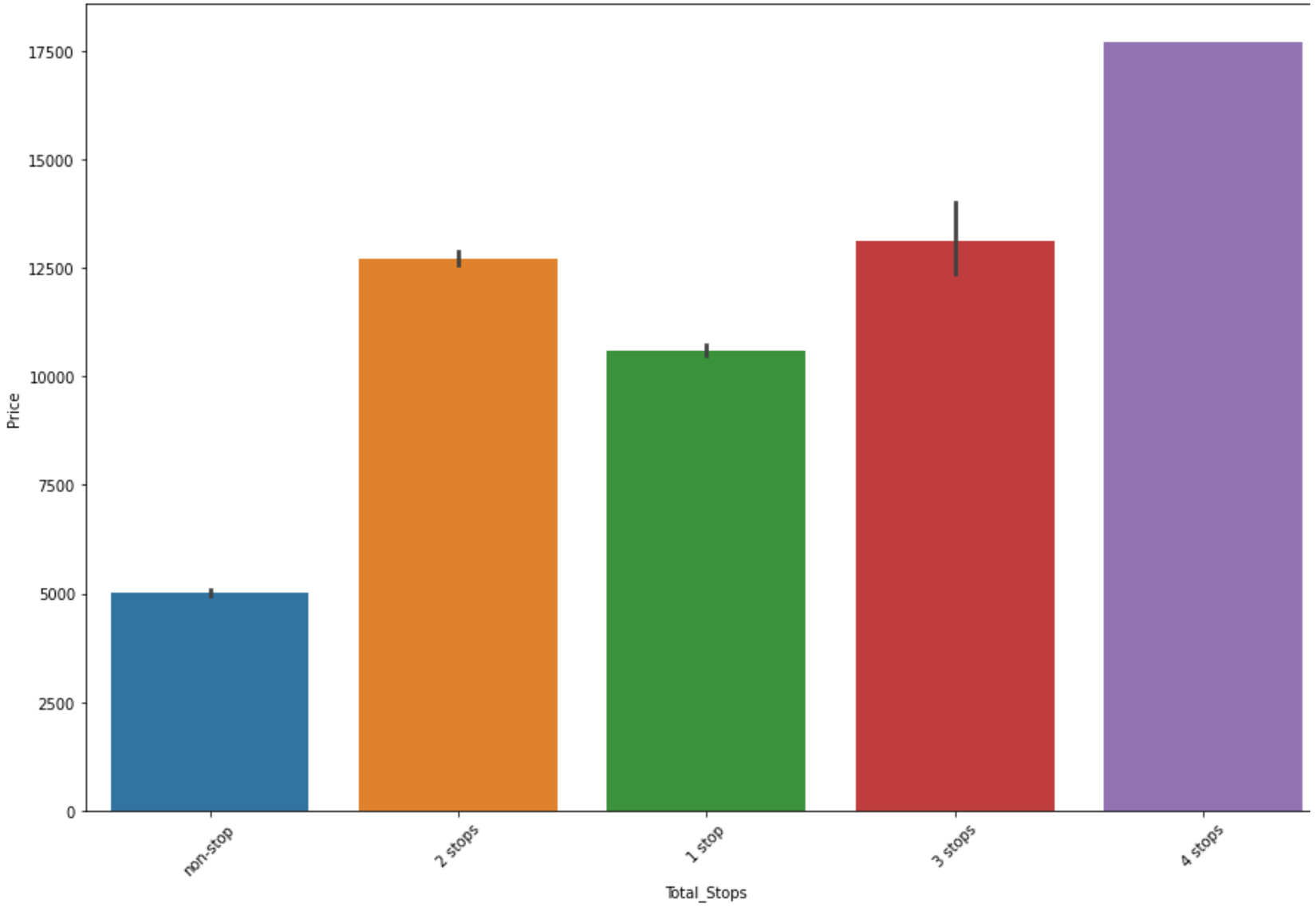
Air India, Jet Airways, Jet Airways Business, and Multiple Carriers offer a few first-class flights.

As we can see from the bar chart that the airfare price range in Delhi & New Delhi is the maximum, this can be due to: Jet fuel prices in Delhi increased in the year 2018 by 26.4%, it is also the National Capital, the political seat of power and a highly visited place for vacations(same for Bangalore & cochin)The same reasoning can be given for higher price range in Delhi as the source of the flight.

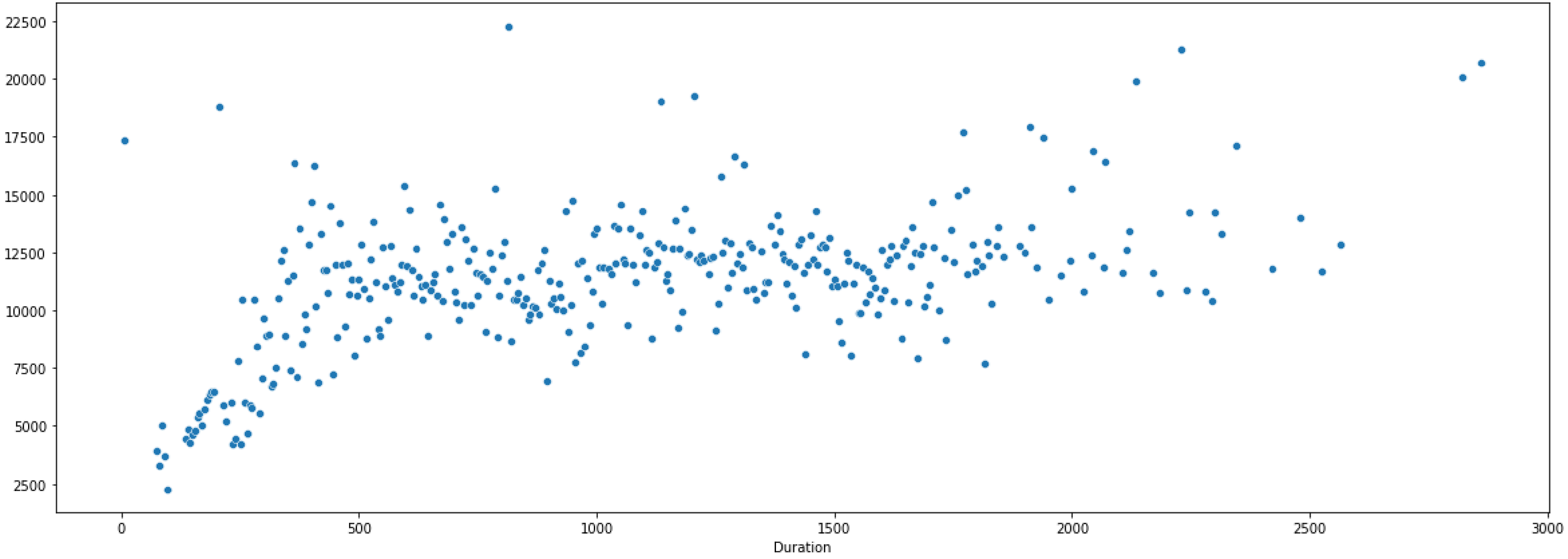




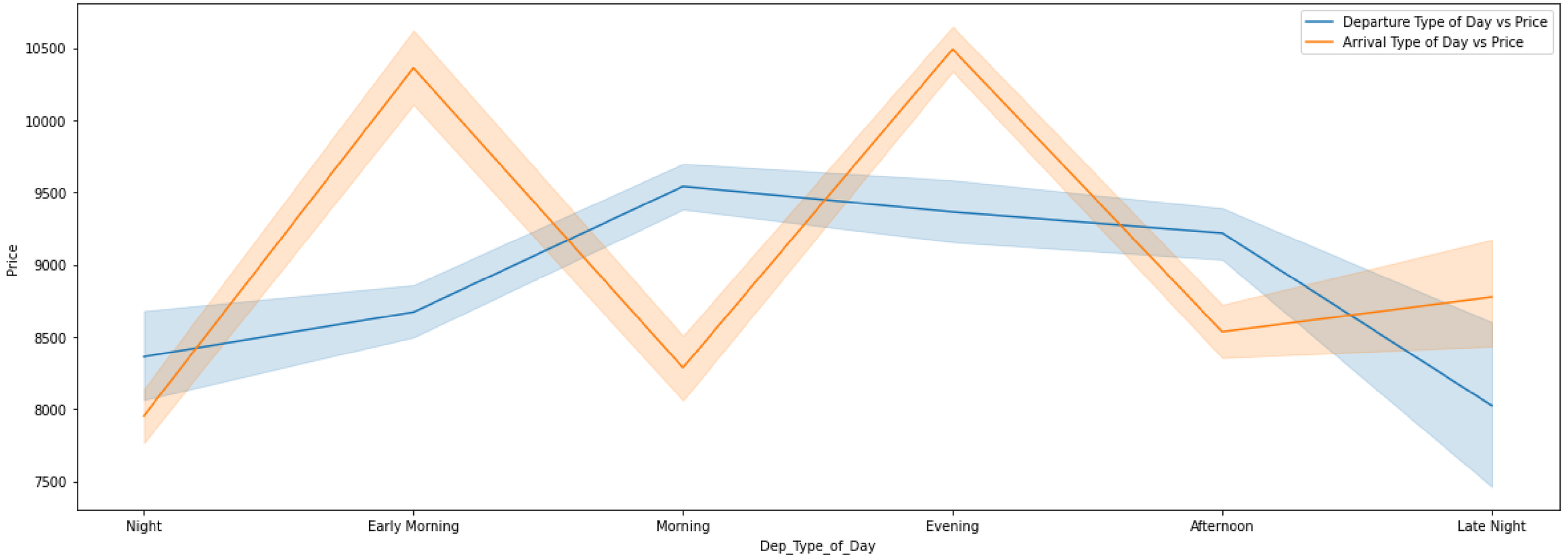
As a direct/non-stop flight is accounting for the fare of only one flight for a trip, its average fare is the least. As the no. of stops/layovers increase, the fare price goes up accounting for no. of flights and due to other resources being used up for the same.



We know that duration plays a major role in affecting air ticket prices but we see no such pattern here, as there must be there are other significant factors affecting airfare like the type of airline, the destination of the flight, date of the journey of flight(higher if collides with a public holiday)



Morning flights are always cheaper and so are midnight flight prices. Evening flights and Early Morning flight fares are expensive due to more demand and are the most convenient time to travel for most people.



Feature Engineering

Within the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. The next step is data pre-processing where we observed that almost all of the information was present in string format. Data from each feature is extracted i.e., day and month is extracted from the date of the journey in integer format, and hours and minutes are extracted from the time of departure. Features like source and destination need to be converted into values as they were of categorical type. For this One Hot Encoding and Label Encoding techniques are used to convert categorical data. Flight Prices are affected by holidays and the distance. I downloaded India holiday data from Kaggle and for Latitude and Longitude, I used the Indian Cities Database from Kaggle and calculated the distance between two cities using the Haversine formula. After that, I apply cosine and sine transformation for cyclical features like an hour, month and day, week, quarter, and minute. Then I calculated the Variance Inflation factor using to find multi-collinearity between variables. There were multiple columns which shows multi-collinearity. I used OLS to check the relation between the variables and their impact on the target variable. I didn't drop any multi-collinear columns as the linear model performed better with those features.

Model Building and Evaluation

I experimented with the basic algorithms and found Extra Tree Regressor, and Random Forest Regressor to give the lowest root mean squared error. After that, I tried XGBoost Regressor and Light Gradient Boosting Regressor. My experimentations revealed that the LightGBM model seemed to perform well. Therefore, I went ahead with the said model to test which of the hyperparameters gave the highest score as per the ground truth

Result

LightGBM

Train Set: Root mean squared error: 600.76, Mean absolute error: 399.28, R Squared: 0.982

Test Set: Root mean squared error: 1428.74, Mean absolute error: 723.85, R Squared: 0.907

Conclusion

This project can result in saving money for inexperienced people by providing them the information related to trends in flight prices and also giving them a predicted value of the price which they use to decide whether to book a ticket now or later. On working with different models, it was found that the LightGBM algorithm gives the best prediction in predicting the output.