FLIP ROBO

Flight Price Prediction

Submitted by:

Prateek AP

Acknowledgment

Data Credit's: Google Flights

Would like to thank the FlipRobo team, especially Sapna and Kashif for all the extensive assistance through the projects.

INTRODUCTION

Business Problem:

Flight ticket prices are very dynamic in nature and finding an optimal time to purchase the air tickets can be challenging for the travelers, as the buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using the scrapped data and also to suggest the best time to buy a flight ticket by building a regression model to predict the flight prices.

For this project, we have collected data from 12 routes from 4 cities across India. Data collected over 2 months resulting in 1.37 lakh data points each across the Mumbai-Delhi, Mumbai-Bengaluru, Mumbai-Chennai, Bengaluru-Delhi, Bengaluru-Chennai, Bengaluru-Mumbai, Chennai-Delhi, Chennai-Bengaluru, Chennai-Mumbai, Delhi-Mumbai, Delhi-Bengaluru, Delhi-Chennai.

The scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.

Conceptual Background of the Domain Problem

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics known as "revenue management" or "yield management". The cheapest available ticket for a given date gets more or less expensive over time. This usually happens as an attempt to maximize revenue based on -

1. Time of purchase patterns (making sure last-minute purchases are expensive)
2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

So, if we could inform the travelers with the optimal time to buy their flight tickets based on the historic data and also show them various trends in the airline industry, we could help them save money on their travels. This would be a practical implementation of a data analysis, statistics and machine learning

techniques to solve a daily problem faced by travelers. The objectives of the project can broadly be laid down by the following questions -

1. Flight Trends Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time?
2. Best Time to Buy What is the best time to buy so that the consumer can save the most by taking the least risk? So should a passenger wait to buy his ticket, or should he buy as early as possible?
3. Verifying Myths Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

Motivation for the Problem Undertaken:

In today's fast paced world, air transport is one of the most essential modes of transportation. No unlike other modes of the transportations, the flight prices are very dynamic and, in this project, we are trying to solve the problem of dynamism in the flight prices and helping travelers to get the cheapest prices for the tickets considering their needs and forecasting the prices based on those needs so a customer can take an informed decision on selection of his flights

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

On analysis of the scrapped data, Dataset has majority of the features which are categorical and the target variable as continuous values. Which makes it a regression problem.

Data Sources and their formats:

The Data is Scrapped using Selenium from Google Flights.

For this project, we have collected data from 12 routes from 4 cities across India. Data collected over 2 months resulting in 1.37 lakh data points each across the Mumbai-Delhi, Mumbai-Bengaluru, Mumbai-Chennai, Bengaluru-Delhi, Bengaluru-Chennai, Bengaluru-Mumbai, Chennai-Delhi, Chennai-Bengaluru, Chennai-Mumbai, Delhi-Mumbai, Delhi-Bengaluru, Delhi-Chennai.

Data Preprocessing Done:

1. Loading Dataset
2. Dropping duplicate values
3. Feature Engineering, extracting years old from Year
4. Shape of our dataset having 1.37 Lac rows and 14 columns
5. Looking at statistical parameters
6. No Null values in our dataset.
7. Categorical values are labelled using OnHotEncoding
8. Removing Outliers

9. Removing skewness
10. Showing correlation with each other features and dropping City.

Data Inputs- Logic- Output Relationships:

The input data provided, helps to understand the various components of the features responsible for flight pricing. Any change in the individual features may change the target variables.

State the set of assumptions (if any) related to the problem under consideration:

This data is for 4 metropolitan cities in India and logically nobody would take a flight which has a layover internationally, has a high price and takes too long. Under this presumption, we have removed the flights which have layovers internationally.

Hardware and Software Requirements and Tools Used

Minimum Hardware Requirements:

- 8Gb+ of RAM
- 128 GB or more (256 GB recommended)
- 10Gb+ of free hard drive space
- Working keyboard
- Trackpad/Mouse
- Display
- A power Adapter.

Software's and Tool's Used:

- Jupyter Notebook
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Plotly
- Scikit Learn

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods):

The data set contains 1.37 Lac data points with no null values. Since the dependent feature is continuous variable, we understand that this problem is a Regression Problem. On analysis of the dataset, I found outliers in 2 features because all other features are categorical variables which is not suitable for either removing outliers or skewness. The outliers were corrected by IQR method. There were certain features which had no contribution to the Prices based on research, hence those were dropped. After data cleaning and data transformation, data visualization was done to represent data graphically. At last, the most important part was to build model for the data set.

Testing of Identified Approaches (Algorithms):

The Algorithm's used are as follows:

- Linear Regression
  - ```
    Model                                                    Report:
    RMSE                                                1562.6182617913107
    MAE                                                  1168.658573661111
    r2_score:                                             75.2083159997081
    cv_score                        :                     71.09036296045844
    Difference between r2_score and cv is 4.11795303924967
    ```
- Ridge
  - ```
    Model                                                    Report:
    RMSE                                                1562.6180240429671
    MAE                                                 1168.6575511480285
    r2_score:                                             75.20832354368926
    cv_score                        :                     71.09222060696257
    Difference between r2_score and cv is 4.116102936726691
    ```
- Lasso
  - ```
    Model                                                    Report:
    RMSE                                                1562.5442949862686
    MAE                                                 1168.4617781840836
    r2_score:                                             75.21066298143901
    cv_score                        :                     71.03223459012486
    Difference between r2_score and cv is 4.178428391314156
    ```
- ElasticNet
  - ```
    Model                           Report                           :
    RMSE                                                1766.8069273121673
    MAE                                                 1316.3860752963592
    r2_score:                                             68.30590002618014
    cv_score:                                             43.96825407606805
    Difference between r2_score and cv is 24.33764595011209
    ```
- DecisionTreeRegressor

- Model Report :
  RMSE 938.7551847579989
  MAE 304.7207956189882
  r2_score: 91.05243351678327
  cv_score : 75.02564835648053
  Difference between r2_score and cv is 16.026785160302737
- RandomForestRegressor
  - Model Report:
    RMSE 725.749817272789
    MAE 300.37923105297506
    r2_score: 94.65221292745726
    cv_score : 83.72409435719314
    Difference between r2_score and cv is 10.928118570264118
- GradientBoostingRegressor
  - Model Report :
    RMSE 1320.108826147913
    MAE 888.1811026437481
    r2_score: 82.30625735401263
    cv_score : 79.56942246209962
    Difference between r2_score and cv is 2.7368348919130057
- CatBoostRegressor
  - Model Report:
    RMSE 853.1778822474614
    MAE 501.0997259385046
    r2_score: 92.6094048389248
    cv_score : 89.27207672317789
    Difference between r2_score and cv is 3.3373281157469137

Run and evaluate selected models:

The algorithms used for hyper parameter tuning and fitting train & test dataset are

a. GradientBoostingRegressor
b. CatBoostRegressor
c. Ridge Regressor

Final Model:

GradientBoostingRegressor without HyperParameter Tuning.

Model Report: GradientBoostingRegressor
RMSE 1320.1088261479124
MAE 888.1811026437477
r2_score : 82.30625735401264
cv_score : 79.47464674222854
Difference between r2_score and cv is 2.8316106117841002

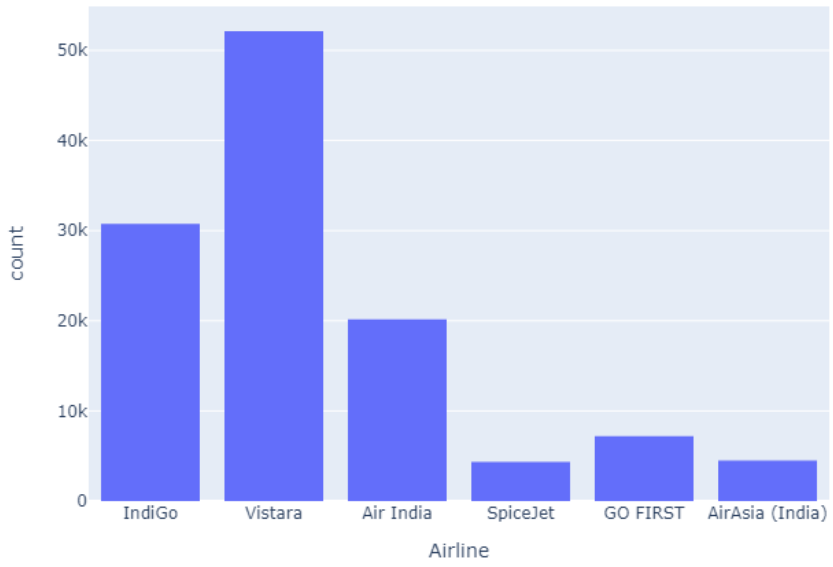
Key Metrics for success in solving problem under consideration

The Key Metrics used in solving the problem are:

- R2 Score
- Cross-Validation Score
- MSE
- RMSE

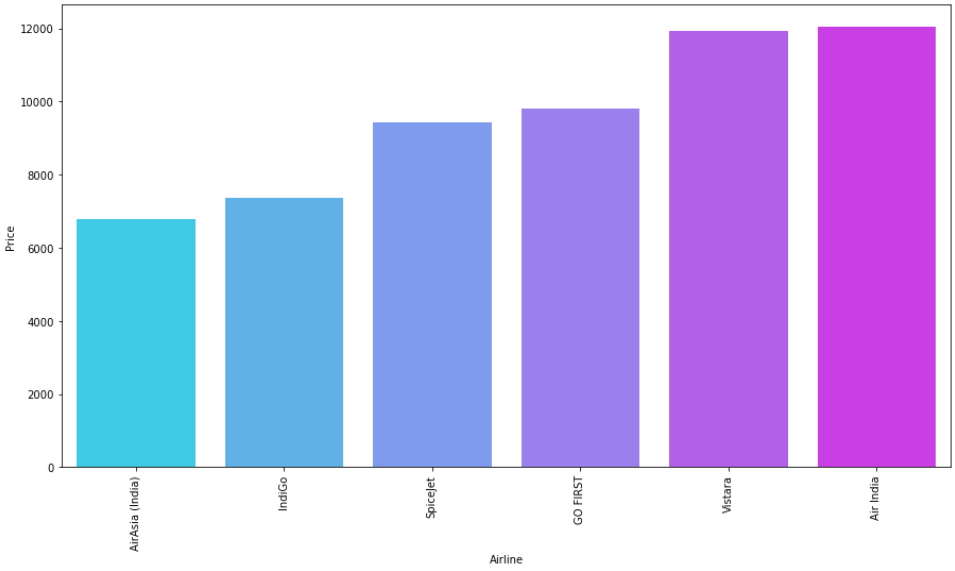
Visualizations:

Airline:



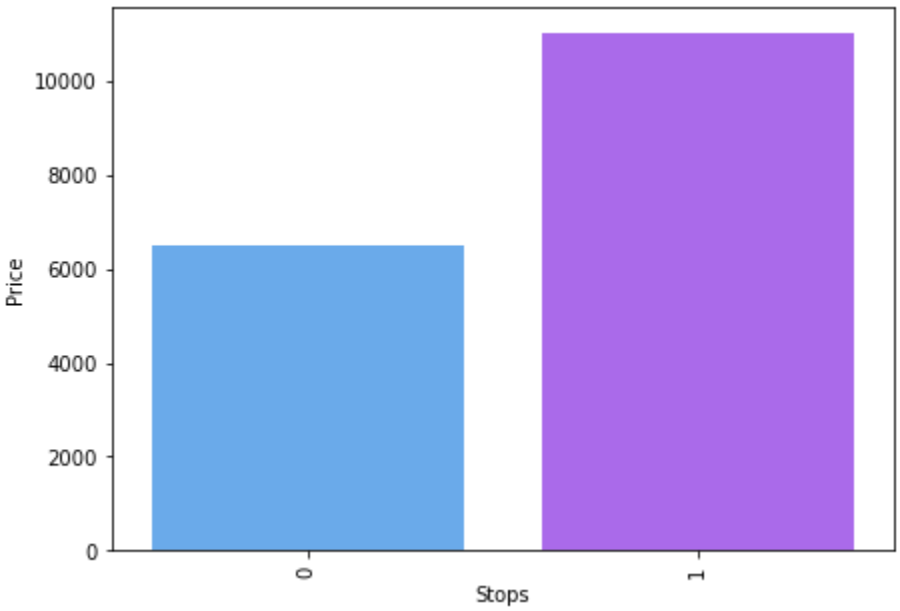
We see that Vistara is the most preferred airlines and Spice Jet the least.

Airline Vs Price:



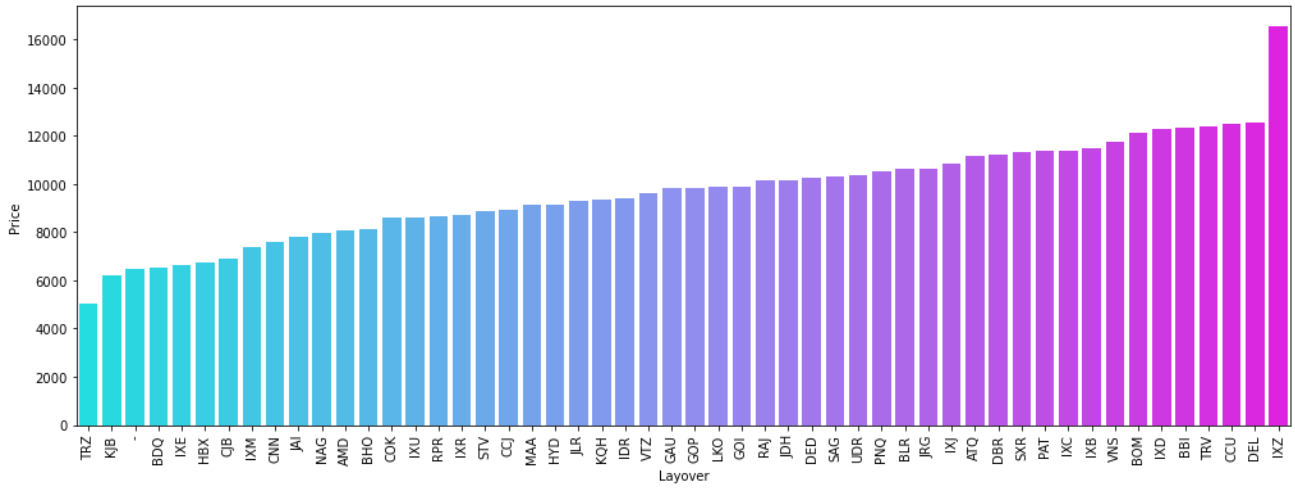
- We also see that AirAsia (India) is the cheapest when it comes to price and Air India is the most Expensive.

Stops Vs Price:



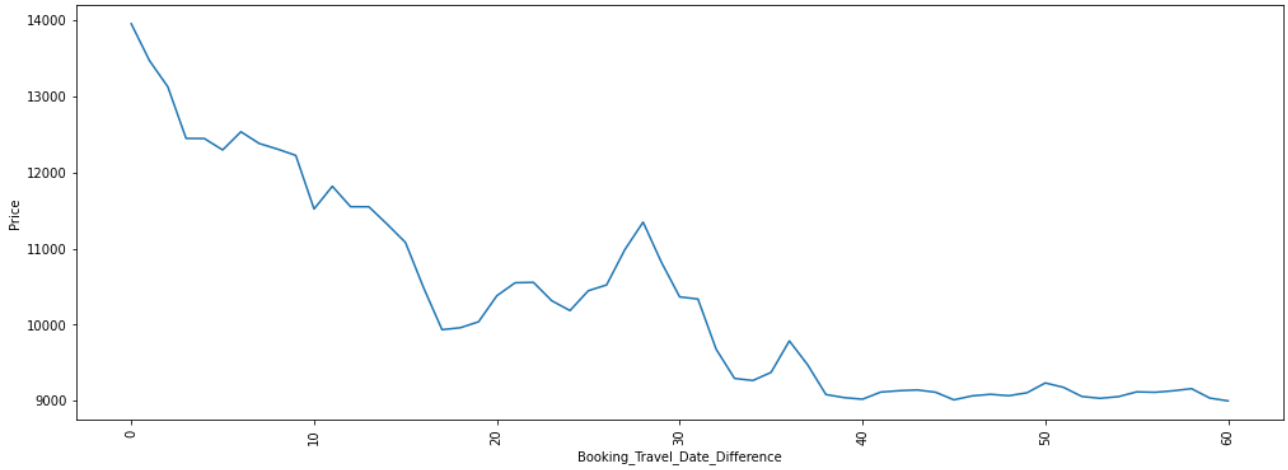
We see that Non-Stop flights are cheaper than flights with a stop.

Layover Vs Price:



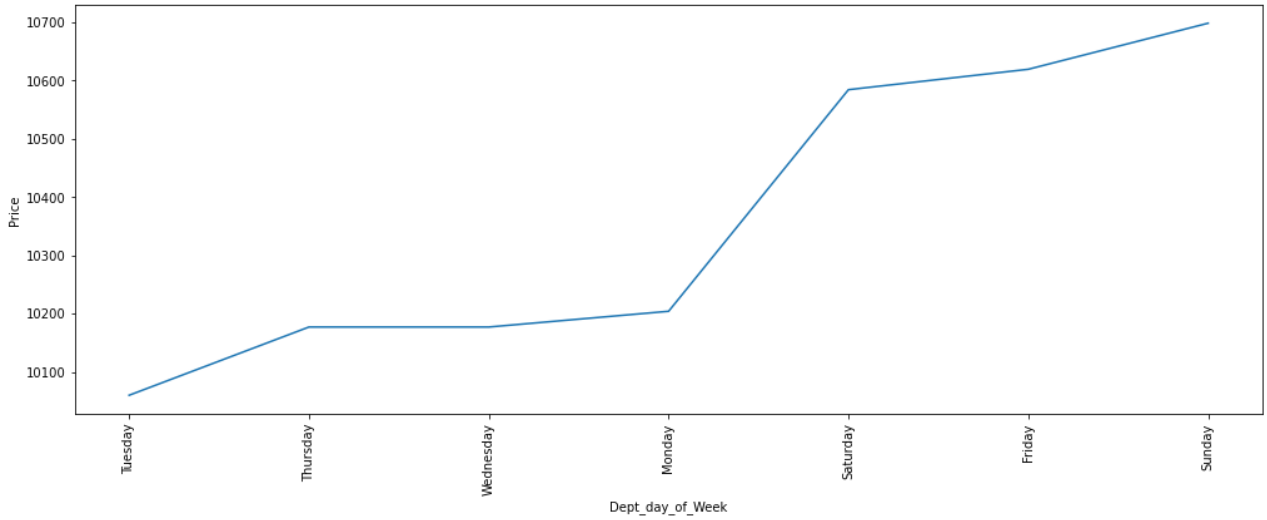
We see flights with layover at TRZ and KJB are the cheapest followed by non-stop flights.

Booking Date, Travel date difference Vs Price:



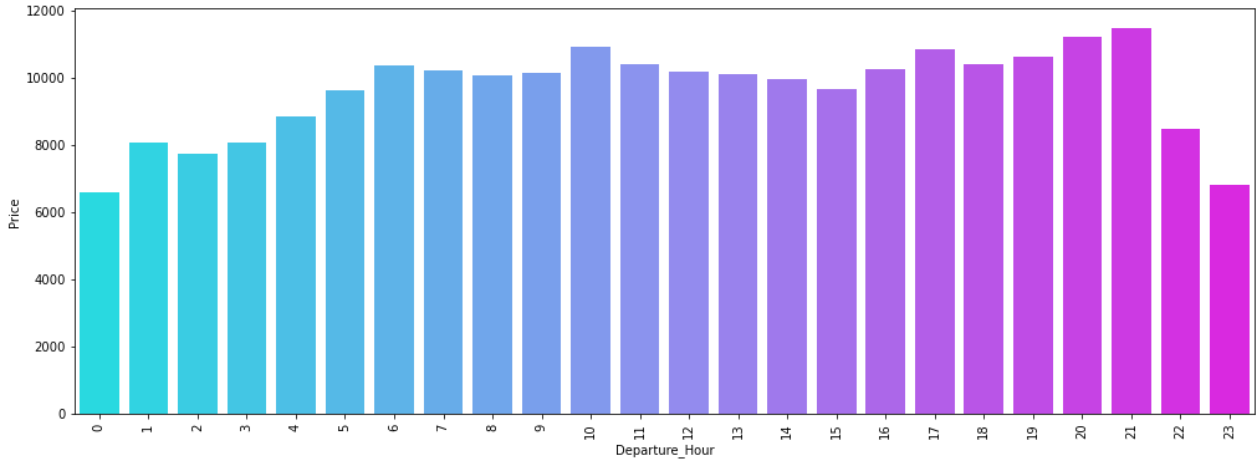
We see that as the Booking Date vs Travel Date Difference increases, the prices decrease.

Departure Day Vs Price:



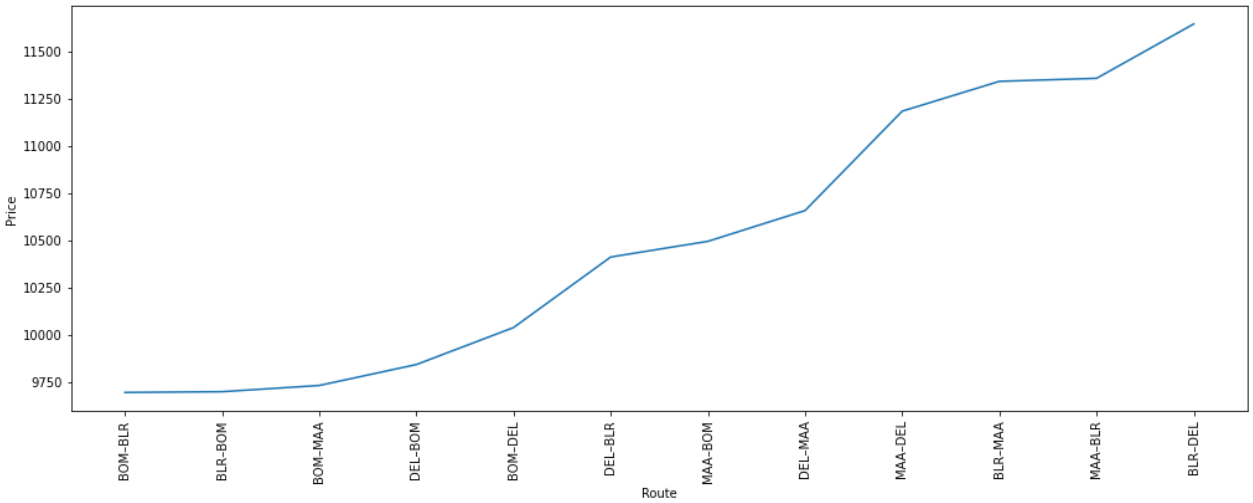
We see the flight prices are the cheapest during midweek and the most expensive on weekends.

Departure Hour Vs Price:



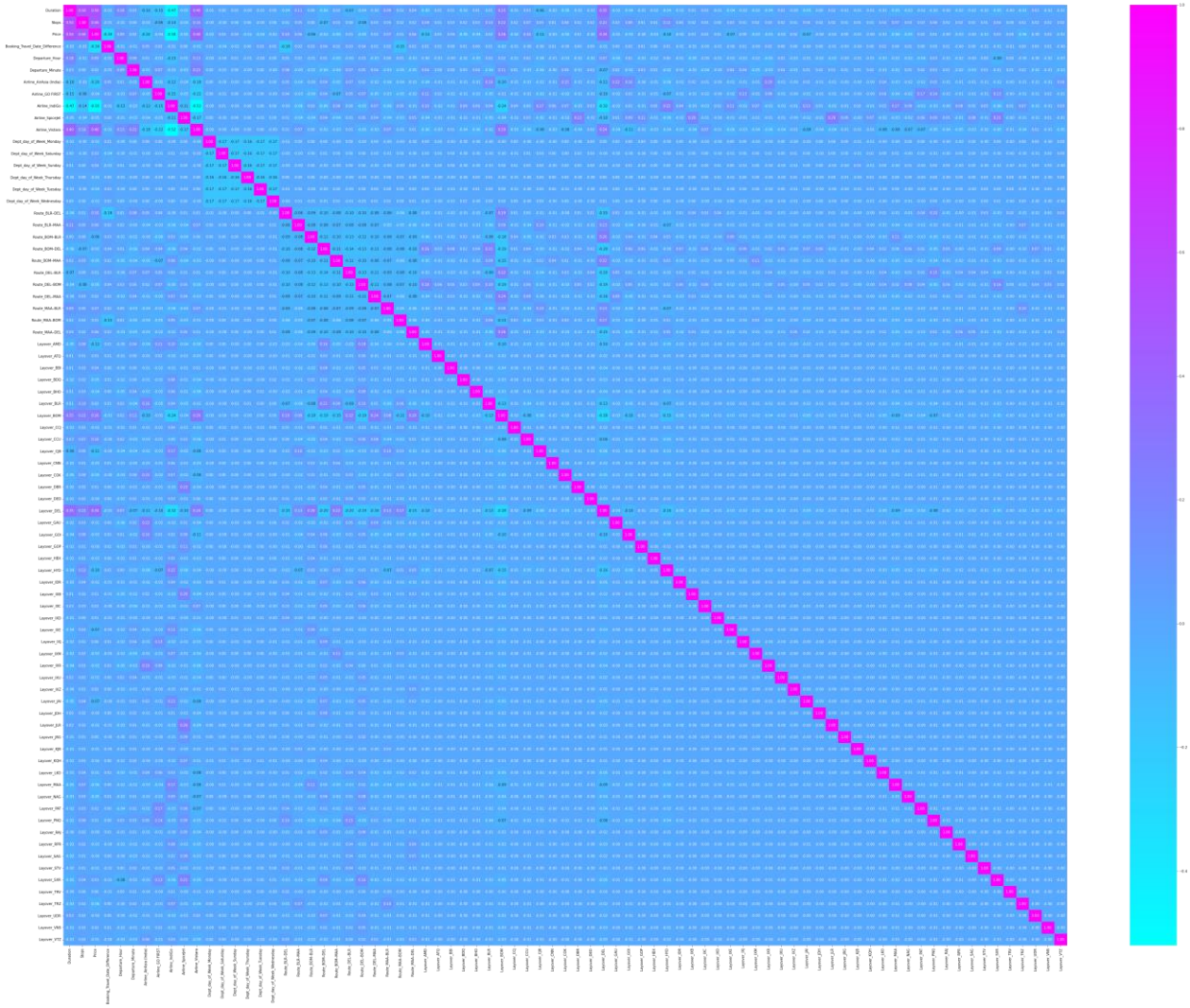
We see that the flight prices are generally higher for the flights between 4pm to 9 pm.

Route Vs Price:



We see those flights operating on Mumbai- Bangalore route is the cheapest and Bangalore-Delhi the most expensive.

Correlation Matrix:



We see that the features are both positively and negatively correlated with the Target Variable.

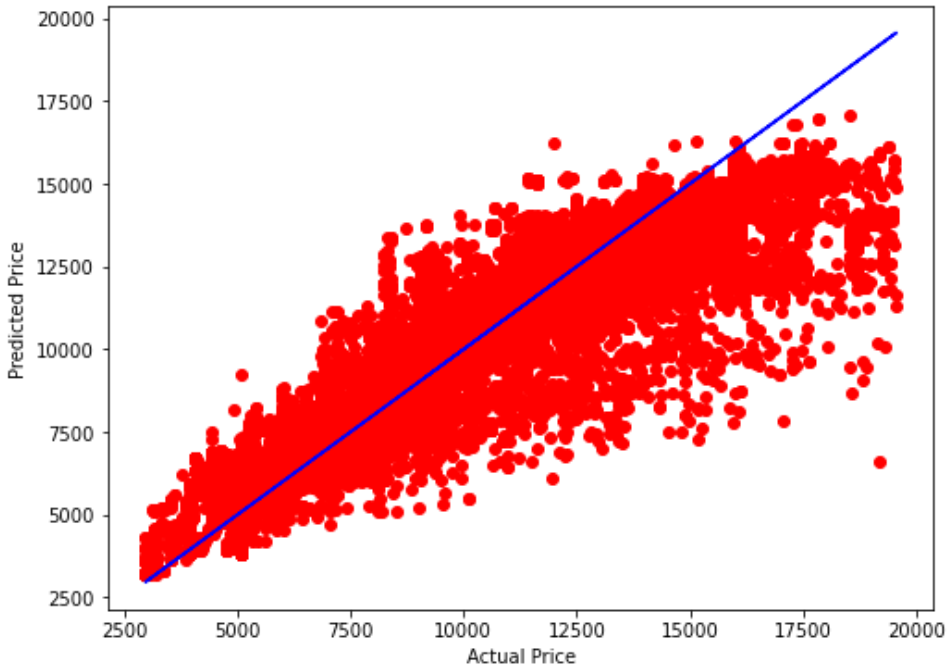
- `Duration` has the Highest Positive Correlation.
- `Airline_Indigo` Highest Negative Correlation

Interpretation of the Results

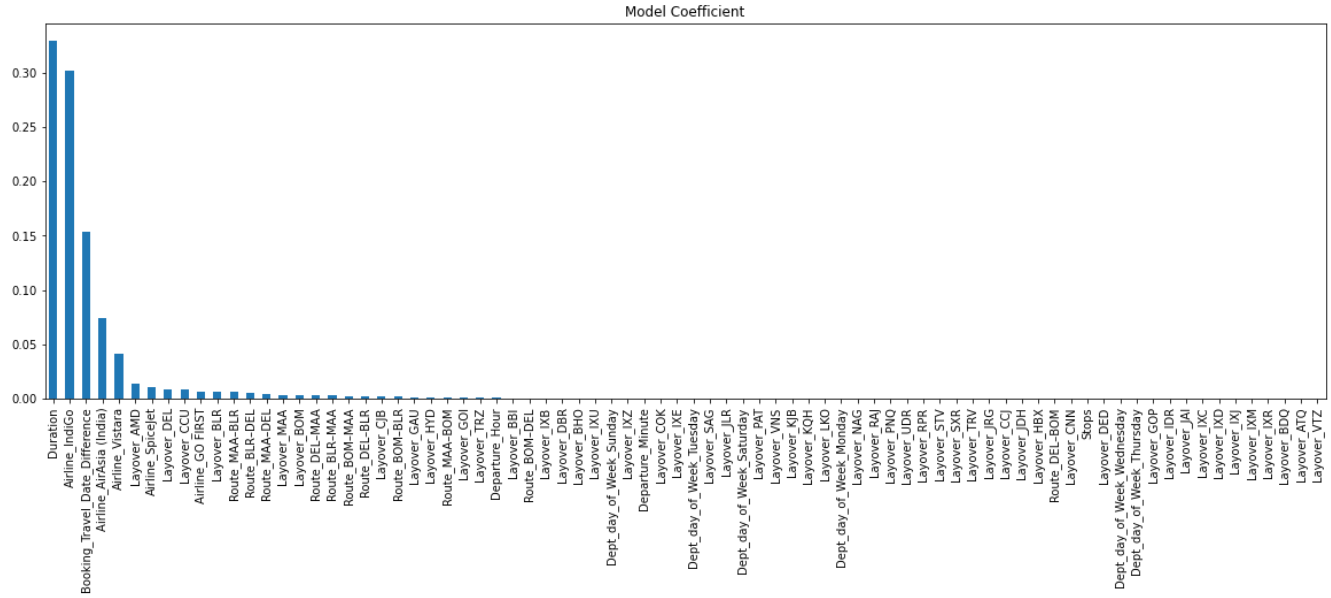
We have achieved an accuracy of `82` % which is pretty good with a difference in cross-validation of `2.83`.

	Actual	Prediction	% Difference
98688	6979	6851.396164	1.828397
99545	6034	5820.618866	3.536313
115057	6013	6395.168231	-6.355700
17114	7399	7158.668547	3.248161
41022	14023	11495.480143	18.024102
48371	10696	12186.124531	-13.931606
118061	10563	10692.331035	-1.224378
87417	6874	7185.816224	-4.536169
36884	13039	11336.288884	13.058602
64959	11685	11246.330306	3.754127

Plotting the results:



The Features that played the most significant role in determining the price are:



CONCLUSION

Key Findings and Conclusions of the Study

- The farther the booking dates and travel dates, the cheaper the prices.
- The Flights Prices are relatively higher on Fridays, Saturdays and Sundays.
- People looking for cheapest prices should check Air Asia as their option.
- Whereas Vistara is the most preferred airline even though its prices are higher.

Learning Outcomes of the Study in respect of Data Science

The dataset was not having so much outliers. The only challenge was we needed more data to analyze and train the machine. Data we scrapped was Raw and misplaced, data cleaning was very important to get proper prediction. The most Important factors determining the Price were Duration, Airline and Difference between the booking and travel date. Gradient boosting algorithm was used as it gave the best results.

Limitations of this work and Scope for Future Work

The Actual prices and predicted prices are having some difference which may lead to underfitting. This may be due to the training of model with less specifications or features. So, in future if we need better performance model, we would need to scrape most of the specifications so that the model can give better outcomes.