FLIP ROBO

Flight Price Prediction Project

Submitted by:

Neha

ACKNOWLEDGEMENT

The sources that helped me out in the completion of this project are, I usually take reference from my personal notes, internet sources and some websites and all of these sources helped me out for the completion of this project.

Introduction

Business problem Framing

Nowadays, anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time, airline ticket prices can vary dynamically and significantly for the same flight, even for nearby seats within the same cabin. Customers are seeking to get the lowest price while airlines are trying to keep their overall revenue as high as possible and maximize their profit. Airlines use various kinds of computational techniques to increase their revenue such as demand prediction and price discrimination.

Conceptual Background of the Problem

The concept behind the problem is, the airline industry is considered as one of the most sophisticated industry in using complex pricing strategies.

However, mismatches between available seats and passenger demand usually leads to either the customer paying more or the airlines company losing revenue. Airlines companies are generally equipped with advanced tools and capabilities that enable them to control the pricing process. However, customers are also becoming more strategic with the development of various online tools to compare prices across various airline companies. In addition, competition between airlines makes the task of determining optimal pricing is hard for everyone.

Review of Literature

India is one of the world's fastest growing automobile markets. The last two decades have seen steadily increasing research targeting both customers and airlines. Customer side researches focus on saving money for the customer while airline side studies are aimed at increasing the revenue of the airlines. Conducted researches employ a variety of techniques ranging from statistical techniques such as regression to different kinds of advanced data mining techniques.

From the customer point of view, determining the minimum price or the best time to buy a ticket is the key issue. The conception of "tickets bought in advance are cheaper" is no longer working. It is possible that customers who bought a ticket earlier pay more than those who bought the same ticket later. Moreover, early purchasing implies a risk of commitment to a specific schedule that may need to be changed usually for a fee.

The ticket price may be affected by several factors thus may change continuously. To address this, various studies were conducted to support the customer in determining an optimal ticket purchase time and ticket price prediction.

Most of the studies performed on the customer side focus on the problem of predicting optimal ticket purchase time using statistical methods, predicting the actual ticket price is a more difficult task than predicting an optimal ticket purchase time due to various reasons: absence of enough datasets, external factors influencing ticket prices, dynamic behavior of ticket pricing, competition among airlines, proprietary nature of airlines ticket pricing policies etc.

On the airlines side, the main goal is increasing revenue and maximizing profit, airlines utilize various kinds of pricing strategies to determine optimal ticket prices: long-term pricing policies, yield pricing which describes the impact of production conditions on ticket prices, and dynamic pricing which is mainly associated with dynamic adjustment of ticket prices in response to various influencing factors. Long term-pricing policies and yield pricing are associated with internal working of the specific airline and do not help that much in predicting dynamic fluctuations in price.

On the other hand, dynamic pricing enables a more optimal forecasting of ticket prices based on vibrant factors such as changes in demand and price discrimination, However, dynamic pricing is challenging as it is highly influenced by various factors including internal factors, external factors, competition among airlines and strategic customers. Internal factors consist of features such as historical ticket price data, ticket purchase date and departure date, season, holidays, supply (number of available airlines and flights), fare class, availability of seats, recent market demand and flight distance. External factors include features such as occurrence of some event at the origin or destination city like terrorist attacks, natural disaster (hurricane, earthquake, tsunami, etc.), political instability (protest, strike, coup, resignation), concerts, festivals, conferences, political gatherings and sports events, competitors' promotions, weather conditions and economic activities.

A significant number of research works exits that proposed prediction models for dynamic pricing in airlines which can be classified into two groups: demand prediction. Early prediction of the demand along a given route could help an airline company preplan the flights and determine appropriate pricing for the route. Existing demand prediction models generally try to predict passenger demand for a single flight/route and market share of an individual airline. Price discrimination allows an airline company to categorize customers based on their willingness to pay and thus charge them different prices. Customers could be categorized into different groups based on various criteria such as business vs leisure, tourist vs normal traveler, profession etc. For example, business customers are willing to pay more as compared to leisure customers as they rather focus on service quality than price.

Customer side models generally utilize restricted features extracted from historical ticket price data, ticket purchase date and departure date. In a similar way, airlines side models are also developed based on limited internal factors such as seasonality, holidays, supply (number of available airlines and flights), fare class, availability of

seats, recent market demand, flight distance and competitive moves by other airlines etc. However, ticket prices and passenger demand can also be affected by many of the dynamic external factors mentioned earlier. Even though the attributes used by earlier researchers play a significant role in predicting ticket pricing/demand, the incorporation of these external factors could also lead to a better result.

Airlines side models represent studies targeting profit gained by airlines and OTAs.

Motivation for the Problem Undertaken

My only objective behind this project is to use data

Analytics and predict the flight prices and help the customers to get ticket prices at good price.

Thus, we can accordingly manipulate the strategy of the firm that will yield high returns.

Analytical Problem Framing

Mathematical/Analytical Modeling of the Problem

Mathematical /Analytical functions and methods are used such as df.shape () to check out the number of rows and columns in the dataset. df.dtypes() is used to check out the data type of the columns we have in the dataset. df.isnull() is used to check whether null values are present or not, we have also checked it by using Heatmap from seaborn. And the other one is checking out the statistical summary of the dataset by using df.describe() method.

We got lots of fresh information from statistical summary is like:

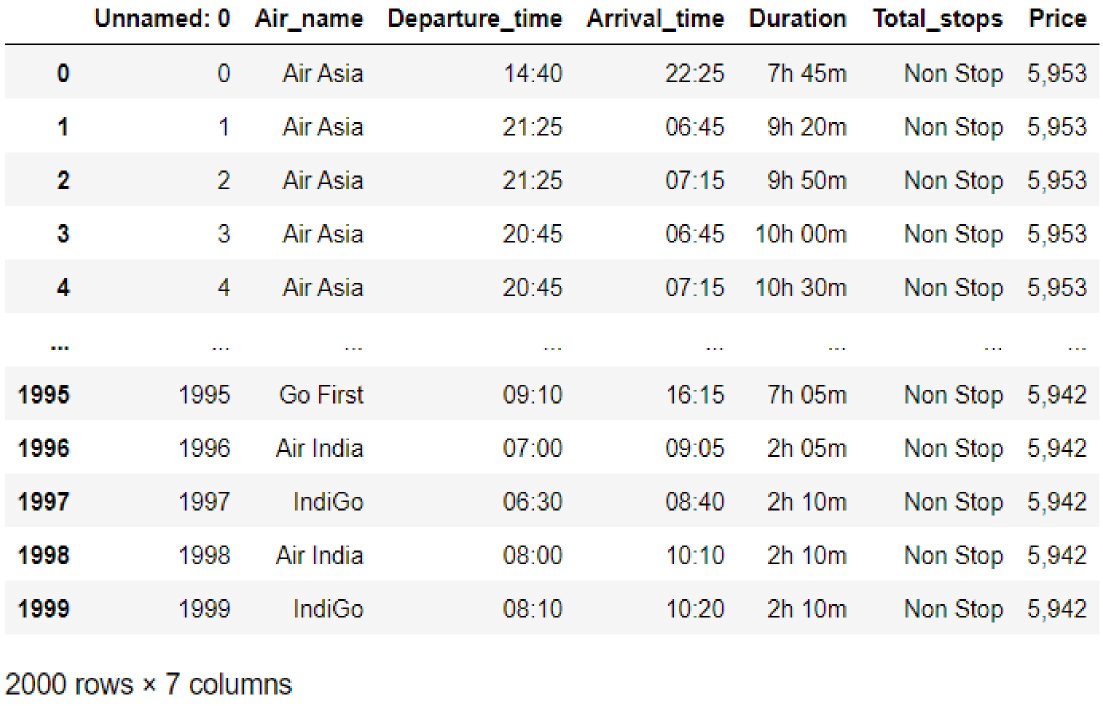
1. We can see that in the count function that all the columns have same values, which clearly means no null values are present in the columns.

2. The difference in mean and median is almost similar.
3.From the count we can see that we have no missing value.

4.We can see that we have small difference in 75% percentile and max in columns like departure time ,arrival_time, duration.

Data Sources and their Formats

We have scraped the data from multiple flight ticket booking websites and saved the data in a csv file. Then we have imported the dataset, we have 2000 rows and 7 columns in our dataset we can see the snapshot of the dataset below:



	Unnamed: 0	Air_name	Departure_time	Arrival_time	Duration	Total_stops	Price
0	0	Air Asia	14:40	22:25	7h 45m	Non Stop	5,953
1	1	Air Asia	21:25	06:45	9h 20m	Non Stop	5,953
2	2	Air Asia	21:25	07:15	9h 50m	Non Stop	5,953
3	3	Air Asia	20:45	06:45	10h 00m	Non Stop	5,953
4	4	Air Asia	20:45	07:15	10h 30m	Non Stop	5,953
...	...	...	...	...	...	...	...
1995	1995	Go First	09:10	16:15	7h 05m	Non Stop	5,942
1996	1996	Air India	07:00	09:05	2h 05m	Non Stop	5,942
1997	1997	IndiGo	06:30	08:40	2h 10m	Non Stop	5,942
1998	1998	Air India	08:00	10:10	2h 10m	Non Stop	5,942
1999	1999	IndiGo	08:10	10:20	2h 10m	Non Stop	5,942

2000 rows × 7 columns

Data Preprocessing Done

Following are the steps followed for data cleaning:

Missing Values if we have missing values then treat them by using suitable method, but we have no missing data in our dataset.

Label Encoding, we have used label encoding on the columns which were in string format.

Removing the outliers, we have few outliers present in the dataset, we can remove these outliers by using zscore or IQR, but as we have few outliers, we can ignore them.

Removing the highly correlated columns, we have from the correlation matrix that few columns are highly negatively correlated, thus we have dropped those columns.

Removing Skewness, we have observed that small skewness is present in many columns, So no need to remove the skewness.

Data Inputs-Logic–Output Relationships

The input columns of the data frame express the features affecting the flight ticket price so that by the help of these features we can predict the price of the ticket. The price of the ticket completely depends on the features.

Hardware and Software Requirements and Tools used

The hardware device used is only the pc. In software devices so many applications and libraries are used to for the completion of the project. I have used Jupyter software for using jupyter notebook. The libraries used are Selenium, pandas, NumPy, matplotlib, seaborn these libraries are used for analysis purpose. And from sklearn power transform is used for removing skewness.

Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**

Methods are used such as df.shape () to check out the number of rows and columns in the dataset. df.dtypes() is used to check out the data type of the columns we have in the dataset. df.isnull() is used to check whether null values are present or not, we have also checked it by using Heatmap from seaborn. And the other one is checking out the statistical summary of the dataset by using df.describe() method. We got lots of fresh information from statistical summary is like:

1. We can see that in the count function that all the columns have same values, which clearly means no null values are present in the columns.

2. The difference in mean and median is almost similar.
3.From the count we can see that we have no missing value.

4.We can see that we have small difference in 75%

percentile and max in columns like departure time ,arrival_time, duration.

- **Testing of identified Approaches**

We have used different regression models listed below:

LinearRegression()

DecisionTreeRegressor()

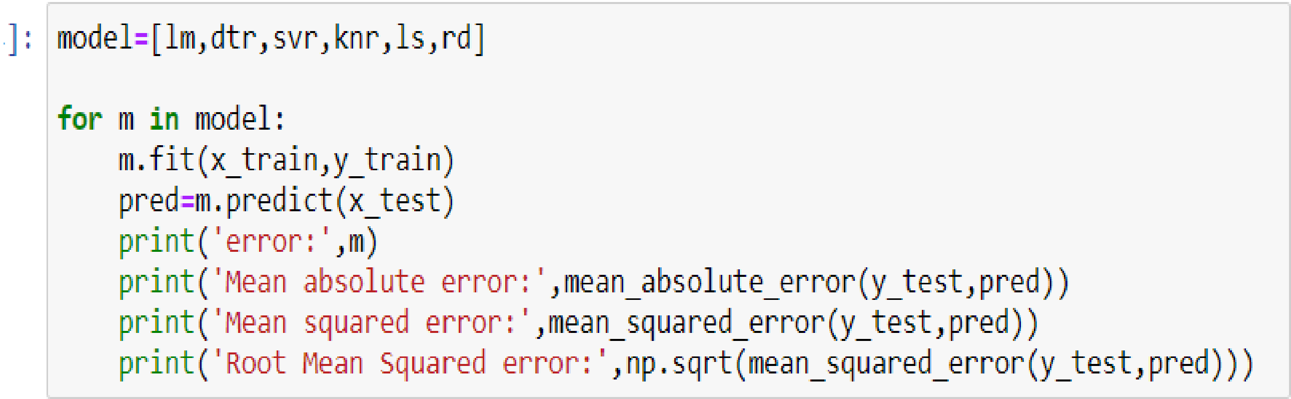
SVR()

KNeighborsRegressor()

Lasso(alpha=0.0001)

Ridge(alpha=0.0001)

Run and Evaluate Selected models



```
]: model=[lm,dtr,svr,knr,ls,rd]

   for m in model:
       m.fit(x_train,y_train)
       pred=m.predict(x_test)
       print('error:',m)
       print('Mean absolute error:',mean_absolute_error(y_test,pred))
       print('Mean squared error:',mean_squared_error(y_test,pred))
       print('Root Mean Squared error:',np.sqrt(mean_squared_error(y_test,pred)))
```

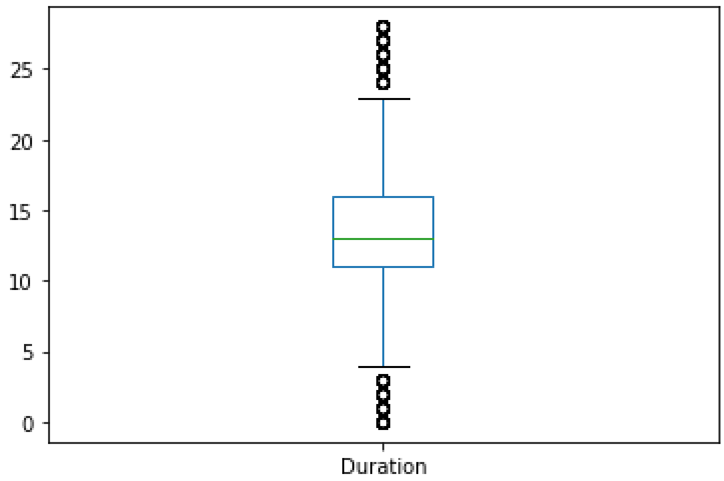
These are the models used to evaluate.

- **Key Metrics for success in solving problem under consideration.**

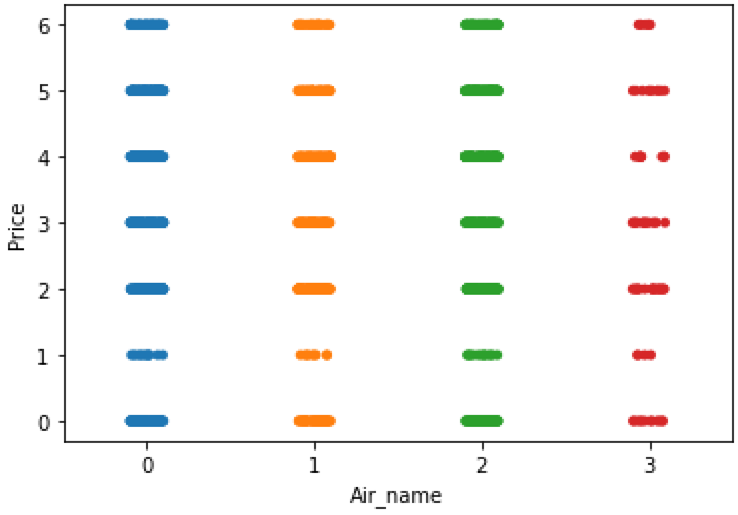
It's a regression problem thus we have used r2_score, mean absolute error mean squared error, and root means squared error as the evaluation metrics.

- **Visualization**

Using Boxplot, we have used a boxplot for each of the columns. Here we have a boxplot of duration column in the plot, we can see that we have many outliers as some of the duration are very long and some of are very short.

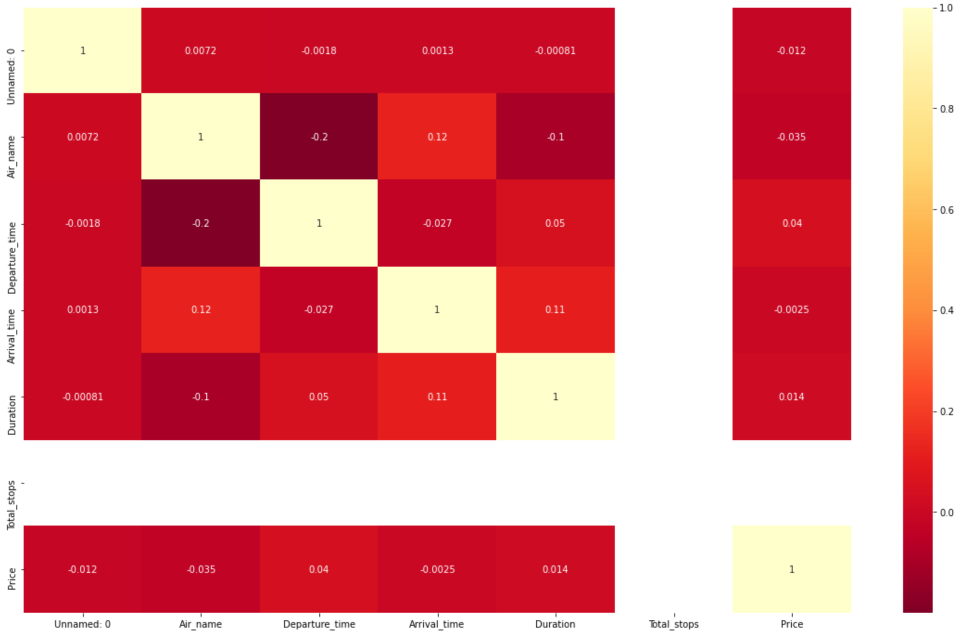


Using Strip plot We have used a strip plot for bivariate analysis to see the relation of each column with the target column. Here is a strip plot of the column Air_name. Here in the column, we can see the price range of each type of airline.



Multivariate Analysis

In the correlation matrix, we can see the relation of each column with all the other columns. We can see the correlation matrix in the plot. We have also used a pair plot to see the graphical relation of each column with the other columns.



Key Observation:

Now we can clearly indentify the correlation of independent variable with the target variable"Price".

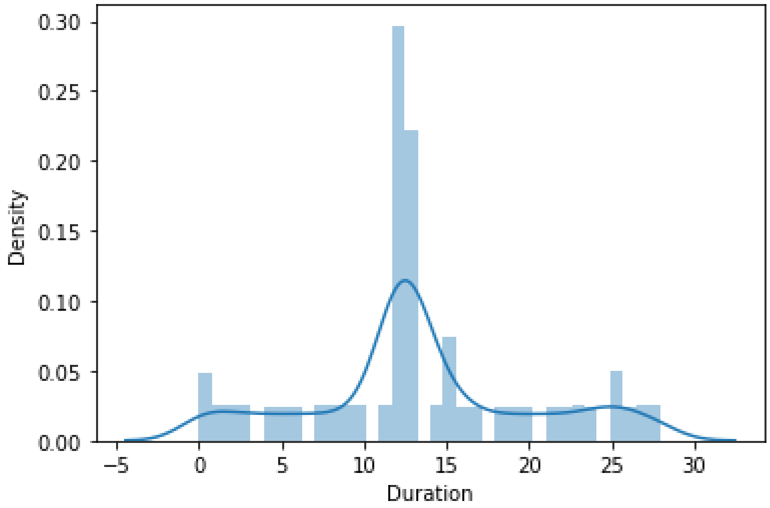
Light shades are highly positively correlated.

Air_name is highly negatively correlated with price.

Another column is slightly positively correlated with price column.

Skewness

Here we have checked the skewness of each column of the dataset. Here we have a skewness plot of the column duration, Here also we can see the skewness in skewenss in the first and last curve.



Till here visualization of the data is done

Interpretation of the Results

From the complete analysis of flight ticket price prediction, we are having a trained machine learning model by using which we can make the prediction of flight ticket prices.

Conclusion

The conclusion comes out to be we have complete Machine learning model for flight price evaluation for the customers who has facing problems in the certain changes of flight price.

This will help out the industry facing problem in this aspect. Thus the business problem in this aspect is resolved.