
# Flight Price Prediction

**Submitted by:**

**Muktikanta Sahoo**

**ACKNOWLEDGMENT**

I would like to express my deep and sincere gratitude to my Project supervisor, Ms. Sapna Verma, Project co-ordinator, Flip Robo Technologies for giving me the opportunity to do research and providing invaluable guidance throughout this project. His dynamism, vision, sincerity and motivation have deeply inspired me.

## INTRODUCTION

- **Business Problem Framing**

  Nowadays, the number of people using flights has increased significantly.
  Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Flight price prediction is a challenging task since the factors involved in pricing dynamically change over time and make the price fluctuate.
  The cheapest available ticket on a given flight gets more and 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 to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

- **Conceptual Background of the Domain Problem**

  The price of an airline ticket is affected by several factors, such as flight distance, purchasing time, fuel price, etc. Each carrier has its own proprietary rules and algorithms to set the price accordingly.

  Flight ticket data is not well organized and ready for direct analysis, collecting and processing those data always requires a great deal of effort. Recent advance in Artificial Intelligence (AI) and Machine Learning (ML) makes it possible to infer such rules and model the price variation. It can also help customers to predict future flight prices and plan their journey accordingly.

- **Review of Literature**

  The flight ticket buying system is to purchase a ticket many days prior to flight take-off to stay away from the effect of the most extreme charge. Mostly, aviation routes don't agree this procedure. Plane organizations may diminish the cost at the time, they need to build the market and at the time when the tickets are less accessible.

  They may maximize the costs. So, the cost may rely upon different factors. To foresee the costs this venture uses AI to exhibit the ways of flight tickets after some time. All organizations have the privilege and opportunity to change its ticket costs at any time. Explorer can set aside cash by booking a ticket at the least costs. People who had travelled by flight frequently are aware of price fluctuations.

  The airlines use complex policies of Revenue Management for execution of distinctive evaluating systems. The evaluating system as a result changes the charge depending on time, season, and festive days to change the header or footer on successive pages.

The aim of the airways is to earn profit whereas the customer searches for the minimum rate. Customers usually try to buy the ticket well in advance of departure date to avoid hike in airfare as date comes closer. But this is not the fact. The customer may wind up by giving more than they ought to for the same seat.

- **Motivation for the Problem Undertaken**

  It is hard for the client to buy an air ticket at the most reduced cost. For this, few procedures are explored to determine time and date to grab air tickets with minimum fare rate. Many of these systems are utilizing the modern computerized system known as Machine Learning.

  Someone who purchase flight tickets frequently would be able to predict the right time to procure a ticket to obtain the best deal. Many airlines change ticket prices for their revenue management. The airline may increase the prices when the demand is to be expected to increase the capacity.

  To estimate the minimum airfare, data for a specific air route has been collected including the features like departure time, arrival time and airways over a specific period. Features are extracted from the collected data to apply Machine Learning (ML) models. We use the machine learning regression methods to predict the prices at the given time.

## Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

  In this project we have used different inbuilt python methods to check the statistics of the data.
  To understand the different datatypes of the attributes I have used 'dtype' method, to check if there are any null values present in the dataset, I have used 'isnull().sum()' method.{It is also provided in dataset there are no null values}.
  As there were less number of numerical attributes earlier, I have not used 'describe()' method, But the describe() method returns description of the data in the DataFrame.

  If the DataFrame contains numerical data, the description contains this information for each column: count - The number of not-empty values. mean - The average (mean) value. std - The standard deviation. min-The minimum value. max- The maximum value of attribute.

  The info() method prints information about the DataFrame. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values). Note: the info() method actually prints the info

- **Data Sources and their formats**

  The accumulation of information is the most significant part of this venture. The different wellsprings of the information on various sites are utilized to prepare the models. Sites provide data about the numerous courses, times, flights, and charge.

  I have used Selenium tool to gather the data from different websites. I have used 'Easemytrip.com' and scrape the various details like name of airlines, arrival time, departure time, duration, Source location, Destination location, Date of departure, Number of stops, and collected the data in form of csv format. All the attributes collected are in categorical format which need to be treated except the date of departure.

  There are 10181 rows and 10 columns which I have scraped from this website. All the attributes are categorical except departure date, departure month.

- **Data Pre-processing Done**

  All the gathered information required a great deal of work, so after the accumulation of information, it should have been perfect and be ready as indicated by the model prerequisites.
  Different statistical methods and logics in python clean and set up the information. For instance, the price was character type, not a number. Additional features are created to get more accurate results.

  As Vistara, Indigo, Air-India, Air-Asia, Go-First, SpiceJet has very less flights we need to take care of this data.
  Sources have same locations but mentioned in different ways, we changed the format.
  Destination has same locations but mentioned in different ways, we corrected the format.
  Total number of stops have the information but mentioned in different ways we have changed the format.

  Based on the hours we split the data into early morning flights, morning flights or evening/Night flights.

- **Data Inputs- Logic- Output Relationships**

  Indigo flights and Vistara constitute nearly 50% of total flights compared to other flights used.
  The hospitality of the flight, pricing ranges may be considered for this high number of usages.
  The number of flights is distributed normally except Weekends or holidays over the period.

We can see there are peaks formed in all the 1 months of observations marked in different colours. All the spikes seen are due to weekends (Saturdays & Sundays). The increase in price of tickets is about 10% of the regular price range.
We can expect the price is gradually higher then regular days.
As per the observations, No matter of date of booking, we can expect pricings are to be higher during the weekends, one month prior of the expected date of journey can reduce the price of tickets during weekdays.

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

  The following figure shows the correlation between top 8 attributes and target variable.

  1. I assume that dataset gathered from website is complete random sample which is representative of the population.
  2. There is minimal or no multicollinearity among the independent variables.

- **Hardware and Software Requirements and Tools Used**

  Following are the recommended hardware requirement to build and run machine learning model.

  i. 11th generation (Intel Core i5 processor)
  ii. 8GB RAM /500GB SSD (recommended)

  We have used following software and tools for the machine learning model.

  ANACONDA

  Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows.

I have used built in Data science libraries like pandas, NumPy, Visualization libraries like matplotlib and seaborn. Jupyter Notebook, a shareable notebook that combines live code, visualizations, and text.

Machine learning libraries like scikit-learn for data pre-processing, model selection, model evaluation, SciPy for standardizing& normalizing the data.

## Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**
  The dataset provided has huge volume of the data which did not have any null values, but there were outliers present in the 'price' attribute of dataset, unless outlier treatment there is possibility of our machine learning model overfitting the data or

increase the variability in the data. Keeping the data loss into concern, Z Score method is implemented to reduce the outliers.
The attributes which are having less correlation with target variable have been dropped and removed based on the inference learned from heatmaps and bar plot.

- **Testing of Identified Approaches (Algorithms)**
  To feed the dataset to model, the independent and dependent variables are to be split and the independent attributes are standardized using 'StandardScaler' library.
  Now the data obtained is clean and is having least multicollinearity, independent and balanced data. 'train_test_split' function in model selection is used for splitting data arrays into two subsets: for training data and for testing data.
  We train the model using the training set and then apply the model to the test set.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

I have chosen 5 regression machine learning algorithms which is suitable for the pre-processed and cleaned data to train and test the dataset

i. Linear Regression

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.

ii. Random Forest Regressor

A random forest is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

iii. Gradient Boosting Regressor

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Iv. XGBoost Regressor

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning.

V. LASSO

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model

Vi- Deciscion Tree Regression

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application. It is a tree-structured classifier with three types of nodes.

- **Run and Evaluate selected models**

```
random_forest_grid = RandomForestRegressor(max_depth=12, max_features='auto', min_samples_leaf=3, min_samples_split=12
random_forest_grid.fit(X_train, y_train)
y_pred = random_forest_grid.predict(X_test)

print("******************Results********************")
print('The r2 score is:', r2_score(y_test, y_pred))
print('The mean absolute error', mean_absolute_error(y_test, y_pred))
print('The mean squared error', mean_squared_error(y_test, y_pred))
print('root mean square error', math.sqrt(mean_squared_error(y_test, y_pred)))
cv = cross_val_score(random_forest_grid, X,y,cv=5)
print('The cross validation score', cv.mean())

print("\n*****************xxxxxxxxxx*********************")

#graph
plt.figure(figsize=(8,8), facecolor='w')
plt.scatter(y_test, y_pred,alpha=0.8,color='blue')
plt.plot(y_test, y_test, color='green')
plt.title('Plot of Actual value vs Predicted value')
plt.xlabel("y_test")
plt.ylabel("y_pred")
plt.show()
```

```
******************Results********************
The r2 score is: 0.08846987656291883
The mean absolute error 799.6039134449677
The mean squared error 1383592.5746356
root mean square error 1176.2621198676763
The cross validation score -0.053183571515969016
```

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

  An key/evaluation metric quantifies the performance of a predictive model This typically. Following are the metrics I have used to evaluate the model performance. I have used R2 score, mean absolute error, mean squared error and root mean squared error.

- **Visualizations**

  We have used matplotlib and seaborn to interpret the relationship, we have plotted the graph using histogram to know how the data is distributed and box plot is used to check the outliers present and how is variance spread around the mean of the data.
  Indigo flights and Vistara constitutes nearly 50% of total flights compared to other flights used.
  The number of flights are distributed normally except Weekends or holidays over the period.
  The price range was varied in between 1000 to 15000 on average.

- **Interpretation of the Results**

  We have trained several models above for the dataset we had prepared, and we got different results for different algorithm. For the selected test data set , output of the model is plotted across the test dataset. Graph shows the comparative study of original values and predicted results. By the analysis of the results obtained from the algorithm such as Linear Regression, Random Forest, Decision Tree, Gradient

Boosting, XG Boosting, Lasso gives the predicted values of the fare to purchase the flight ticket at the right time.

Random Forest Regressor gives 88.5% accuracy.

## CONCLUSION

- **Key Findings and Conclusions of the Study**
  To evaluate the conventional algorithm, a dataset is built for various routes in India and studied a trend of price variation for the period of limited days. Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost. Data is collected from the websites which sell the flight tickets so only limited information can be accessed. The values of R-squared obtained from the algorithm give the accuracy of the model.

- **Learning Outcomes of the Study in respect of Data Science**
  1. This project has demonstrated the importance of having large dataset for training and testing the machine learning model.
  2. Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.
  3. Through different powerful tools of visualization, we were able to analyse and interpret different hidden insights about the data.
  4. Build Models, since it was a supervised regressiion problem, I built 6 models to evaluate performance of each of them: a. Linear Regression b. Random Forest c. Decision Tree d. XGboost regressor , e. Lasso f. Gradient Boosting.

- **Limitations of this work and Scope for Future Work**

In the future, if more data could be accessed such as the current an availability of seats, the predicted results will be more accurate. One can focus on collection of real time customer-oriented data which can be useful for EDA. And more inference can be provided based on the analysis.