**Flight Price Prediction**

**Problem Definition**

The variation in prices of flight tickets has always been very confusing for customers and is very difficult to guess. The Airlines companies implement dynamic strategies to assign pricing for air fare tickets to increase demand for their seats and maximize their revenue ,Hence it becomes difficult for consumers to buy tickets at minimum price .The closely connected terms with their Airlines prices such as commercial ,financial ,social factors and marketing is under consideration while practicing these dynamic Strategies .The airlines companies are trying their best to keep their revenue high and increase their profit .The Traveller often find the flight prices unpredictable as the flight prices tomorrow will not be the same as the same flight today. The system is complicated because each flight has limited number seats to be sold .In case the demand of air tickets is high ,Than the price will increase and on the other hand if the seats are left unsold than the cost of the air tickets might decrease as it represents a loss of revenue .To solve this problem of predicting flight prices, Machine Learning is great idea to learn from historical data of the past flight prices and build logic on given data. We will use Linear Regression which will help us in predicting the flight prices on the basis of certain factors which will involve data extracting, Data Analysing and Data interpretation.

Data information

The prices of flight tickets for various airlines are between the months of march and June of 2019 and between various cities. We have two date sets i.e. Train data and Test data. The size of **Training Set** is 10,683 records which consists of both categorical and numeric data. Some special Characters are also seen with in data to which we will apply data transformation before using it on the Model.

The **Features** considered initially for each flight are :-

Air lines :The name of the airline.

Date of Journey :The date of the journey.

Source : The source From which the service begins

Destination: The Destination where the service ends.

Route: The route taken by the flight to reach the destination.

Deep Time : The time when the journey starts from the source.

Arrival time: Time of arrival at the destination.

Duration : Total duration of the flight.

Total Stop: Total stop between the source of destination.

Additional Info.: Information about the flight.

Price: The price of the Ticket.

The Size of **Testing** **Set**  is 2671 records. The testing data is similar to the training data, except for the “price” column which will be predicted using the model.

Data Analysis

From numerous sources the data was collected. The flight ticket detailed information is retrieved from an online data source (Github.com).We took out this data from the website which is in the form of csv record .The file consists of the information with input features and its target variable required for analyzing data. We have retrieved additional features from the existing variables to get more accuracy in the results Features such as “Arrival Time” , “Arrival Date”, “Arrival Month” , “Day” , “Month”& “Year” are generated to make analysis of data

Cleaning And Data Preparation

All the accumulated data needs a great deal of work. So after information gathering ,All the irrelevant Data “features” Such as “Arrival Date” , “Arrival Month” ,Duplicate features And invalid qualities (Features Such “Additional information”& “Year”)are deleted .As the data set contains missing values in variables such as “Route” and “Total Stops”, these values are resolved using SimpleImputer function .One of the necessary steps during data preparation is data transformation ,Hence the data present in string format is transformed to float data type using encoding techniques such as OrdinalEncorder().For example ,”Airline” is a string data type and not an integer. This is done to construct a final data frame from the original data and make a proper analysis of data .In all ML projects , this is the important and time consuming procedure.

Data analysis using visualizations

Data preparation is done by breaking down the information, understanding patterns and than applying different ML algorithm in this case we divide our data into three different data frame on the basis of its type such as nominal ,ordinal or continuous types of data.

Visualizations of Nominal Data

We use count plot for nominal/categorical data as it gives the frequency of the columns. The following observations where made while analyzing data.

∆ The maximum the flights flying are from jet airways , indigo And Air India And lest are from True jet and Vistara Premium economy.

∆Most of the flights are flying from Delhi And Lest From Chennai.

∆Majority of the flights are landing at the Cochin Airport and lest are landing at the Kolkata Airport.

∆The majority flights have only one stop in between the journey and the most of them are also Non-stop but very few flights have 3 and 4 stops.

∆The flights have journey mostly in the month of June , May ,and March.

Visualization of Continuous Type of Values

We use displot and scatter plot graphs to understand the continuous type of data as distplot displays the density of values and scatter plot helps us find relations between the variables.

∆The data is broadly scattered in all columns excepts for the price columns.

∆The data is the price columns is right skewed but it is a target variable.

∆In the day column, the maximum number of flights are flying between the date 3 and 7

**EDA Concluding Remarks**

After performing all the transformation integration and cleaning of data , We get all the relevant variables and significant information required for building an ML model. We end Up having 11 variables and 10,683 records in data set.The final data det consists of important features used for analysis are :

* Airlines
* Source
* Destination
* Route
* Deep Time
* Arrival Time
* Duration
* Total Stops
* Price
* Months
* Day

Concluding Observation Remarks

* The standard deviation in the “Route” , “Deep Time” , “Arrival Time” , “Duration” and “Price” column is two high which means that the values in this columns are largely scattered and are not near to the mean value .
* The standard deviation of other columns is to high which shows us a normal distribution of data and less chances of having skewness.
* The value in target variable (“Price”)has its minimum price at 1759 and maximum price at 79512.The range is to high.
* The most negatively correlated column is that of the “Total Stops” which means more the number of stops less the flight price.
* The most positively correlated variable is “Route”.
* The variables “Route”, “Arrival Time”, “Source”, “Month”, “Deep Time” and positively correlated with the target variable and variables “Airline” , “Destination”, “Duration” , “Day” and “Total Stops” are negatively correlated.

Pre-processing pipeline

Skewness correlation

* As the range for skewness is threshold +/-0.5,we did not see much skewness in our data. Only variables “Airline” ,”Destination” , “Price” and “Month” showed skewness wherein all the input features are of the object data type and “Price” is the target variable .Hence ,there is no skewness is not found in data.

Outliers Detection And Cleaning

* Statistical methods are used and factor analysis is done to detect outliers.

In this case we use the z- score function for detecting outliers. The outliers were observed Only in the “Price” column which is the Target Variable and cannot be detected as an outlier. Therefore ,there are no outlier in the data.

Normalization

In order to build and train ML model ,Behave to standardize the data and get the values within a particular range for the model to understand data. We have used Standard Scaler() technique for normalization bcauase the values ranges are high in the data. Standard Scaler function will get all the values in the datasets within the range 0 to 1.This will help the ML algorithms to learn data better.Since the target variable is of the contious type of values,we use regression algorithms. In this case,we have used Linear Regression.

**Building Machine Learning Models**

* Building a model that will help measure the performance of a better and more refined algorithms is the major goal here. We have used different Regression and Ensemble Technique to compare and check which algorithm gives better performance and stack them all at the and to see how the model giving prediction.

Linear Regression

Linear Regression is a supervised machine learning algorithms that performs Regression tasks. Simple Linear Regression analysis is used to identify the correlation between two continuous variables . Prediction error is minimum when we find the best fit line for the given data using Linear Regression Algorithm.

The error rate the model after using Linear Regression was Found to be-

Mean absolute error (MAE):2.81%

Mean squared error (MSE):1.36%

Root Mean Squared error (RMSE):3.69%

A cross -validation technique was applied to all the samples and the mean performance of the model is produced.

The best fit line of the model was seen covering almost all the data points which is an illustration of getting the best accuracy and indicates that the model has studied all the points in the data and there are no chances of having over fitting and under fitting issues.

Elastic Net Regression

ElasticNet Regression is a regularization regression technique which uses the best of both lasso and ridge regression models by learning from there drawbacks to better the regularization of statistical methods. Hence we have applied regularization technique was put in use with the help of GridSearchCV.

AdaBoost Regressor

AdaBoost Algorithm is a boosting algorithm that is used as an ensemble Technique AdaBoost is a short form of adaptive Boosting and is used to boost the performance of ML Algorithm . It helps in combining multiple weak trainees and merge their performance to form one strong act. We do this to achieve a model with higher stability overcoming all the issues within the model by decreasing the variance of a single measure and combine several measures from variant models.

**Concluding Remark**

* The r2 score achieved for Linear regression is 100%.
* The accuracy achieved after applying cross validation and ElasticNet

Regularization regression technique comes out to be 99.99%.

* Applying Ensemble Techniques to this regression Models in order to achieve stability in the performance of the model, we get r2 score at 98.21%and CV val score atv 97.66%.
* Following the same procedures for the testing file as done for the training file ( the complete EDA process), we have used the best saved model of the training file to predict the analysis of the testing file.
* Majority of the predicted outputs were found to be similar to the actual outputs at 98% accuracy.