**PREDICTING THE FLIGHT TICKET PRICE USING MACHINE LEARNING**

**INTRODUCTION**

Air fares can be difficult to predict, we could see a price today, search the price of the very same flight tomorrow, it's going to be a new story. People would have heard commuters say that plane ticket prices are so volatile. As data scientists, we must illustrate that with the right information, something can be predicted

All of us are going to be on holiday this summer for a well-deserved break. It 's possible that you spent hours on the internet investigating flight offers, trying to sort out an air fare pricing scheme that seems totally random.

Costs tend to change without explanation, and lengthy flights are not necessarily more costly than shorter flights. These seem random, though is the competitive pricing of airlines, using a technique called airline revenue management.

It operates in real time with one goal of increase sales. Changes are based by an algorithm that calculates prices by using data such as previous tickets, available space, aggregate demand for all of these route options, and the likelihood of providing some tickets later.

**PROBLEM STATEMENT**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable

Airline companies are now using nuanced techniques and approaches to dynamically allocate air fare rates. Such techniques take into account a range of environmental, promotional, business and socioeconomic aspects which are closely linked to final air fare prices. Given the complex nature of the valuation methods applied by the airlines, it is very hard for some people to buy an airplane ticket at the best value, as the price changes dynamically.

In this blog, we will know how to prepare the data and build a machine learning algorithm with a simple regression approach to the Estimate Flight Ticket Price We will use different regression algorithms such as Linear Regression and Random Forest Regressorr to forecast plane ticket fares

**THE DATASET**

This dataset is taken from kaggle and aimed at predicting the dynamkcprices of tickets. The dataset has data collected from different sources and includes the following features.

Airline(which shows the name of the airline ),Date of Journey(which shoes the date of the journey),Source( from where the journey begins),Destination(where the journey ends),Route(The path taken by the plane to reach the destination), Dep\_Time(departure time),Arrival\_Time(Time of arrival),Duration(Length in time of the flight),Total Stops (Total stoppages or layovers between the source and destination),Additional\_Info(Extra information about the flight) and finally Price(The price of the ticket).

**DATA PRE-PROCESSING PIPELINE**

Following steps have been used to develop a machine learning model for this dataset :

1.Data Cleaning

2.Data Visualization

3.Data Pre-processing

4.Feature Engineering

5.Choosing the model

6.Fitting the model

7.Conclusion and results

**DATA CLEANING**

The dataset is divided into training dataset and testing dataset . It can be observed that the training dataset has no null values while the testing dataset has two columns and attributes in which there are null values which are

1.Routes

2.Total Stops

Since the count of null values in very low and hence insignificant , we will just drop the null values and proceed with data visualization.

**DATA VISUALIZATION**

This step deals with summarizing the dataset in the form of graphs,chats and figures with an aim of finding out patterns and trends between different attributes

First we used a countplot which is a plotting algorithm in seaborn to plot the count of different airlines resent in the dataset and can observe that jet airways had the highest count and this shows that it might as well be the most populer amongst consumers as majority of the Airlines in the dataset were of jet airways.

Similarly , we can see that the most opted destination was Cochin and the least opted was Kolkata.

Now in order to visualize data based on different months of the year , a feature engineering tehnique is used in the form of datetime where the date is converted into a datetime format and can be split based daily,monthly and yearly data which is very helpful to analyse short term and long term trends

So firstly by plotting a monthly countplot of the number of flights , we can observe that most number of flights were booked for the month of May . This can be attributed to the holiday season where most people like to take a holiday to different destinations .

**DATA PREPROCESSING**

Now to further classify the date into hours and minutes (both for arrival time and departure time) , we have again used the datetime function. This is important because the airline industry is very competitive and minutes can cause huge fluctuations in prices and demand.

As a result of creating different columns based on Minutes , Hours , we no longer need to departure and arrival time column.

For the duration column i have used loops , and len function to create a list so that i can split the string from the numeric values to preprocess data. Here i have used a string strip function. This has given rise to two new columns which are duration hours and duration minutes . This will help us analyze the data better

**FEATURE ENGINEERING**

Throughout this stage, we prefer working on the given dataset and do some restructuring, such as making multiple column frames, cleaning messy data when it can be used in our Ml algorithm. This move is very critical since you need to constantly make adjustments to the high forecast score

Now we are left with the 'stops' column where we need to convert it into numeric i have used 'if' and 'elif' function to get a numeric value and remove the common string which was 'stops'. This category is a mixture of numerical and a categorical variable, such as '1 stop.' So we just have to have the details of the numbers in this section, so we break that and take the descriptions of the number, just adjust the 'non stop' to '0 stop' and change the column to an integer form.

Now for the Airline, Source and Destination column, i have used get\_dummies to get more columns and and refined the data based on further classifications so that the result could be precise.

Note: All the above mentioned steps have been repeated with test dataset.

**THE MODEL**

The aim of this phase is to create a model that supports us as the basis for measuring the output of a stronger and more tailored model. We use various Regression Techniques and compare them to see which model provides better results

The whole data is divided into X and Y where X has all the variables except the target which is 'Price' and Y only has one feature which is 'Price'.

I have used three models here namely:

1.Linear Regression-RMSE-3000

2.Random Forest Regressor-2200

3.Decision Tree Regressor- 3220

All the three models were fitted to the processed dataset and the following evaluation metrics were found out namely:

1.MAE- Mean absolute error

2.MSE-Mean Squared Error

3.RMSE-Root mean squared error

4.Score of accuracy of the model

Between the models , based on the accuracy score and minimum deviations between predicted and actual output in the form of RMSE , we have chosen Random Forest Regressor for the model

Now to further classify the date into hours and minutes (both for arrival time and departure time) , we have again used the datetime function. This is important because the airline industry is very competitive and minutes can cause huge fluctuations in prices and demand.

As a result of creating different columns based on Minutes , Hours , we no longer need to departure and arrival time column.

For the duration column i have used loops , and len function to create a list so that i can split the string from the numeric values to preprocess data. Here i have used a string strip function. This has given rise to two new columns which are duration hours and duration minutes . This will help us analyze the data better

**CONCLUDING REMARKS**

The results of the model show the ML Models are a satisfactory method to forecast airline tickets costs. Many key issues for forecasting air fares are data Set and collection of features into which we drew a few Practical findings. According to tests we have concluded, which features affect the air fare forecast at most. In addition to the features chosen, there are other features available This could increase the precision of the forecast. This, in the future, Research may be applied to forecast air travel prices.

It can be noticed that the most commonly travelled flight was jet airways and the most common flight length was 2 hours.

Feature Selection and engineering is the most crucial thought in this type of issue. You will see how we've treated numeric and categorical data and how we've developed a different Machine learning framework on the very same sample group. We also review the RMSE score of each method so that we can explain how it should be done in our test dataset. Finally, you could also help develop the Model by tuning the multiple metrics used in the model.