**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 travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

**Introduction:**

“Once you have tasted flight, you will forever walk the earth with your eyes turned skyward, for there you have been, and there you will always long to return.”

Most of us wish to travel through air. But the flight price is the deciding factor, not only for travellers but also for airline companies as they always try to keep the price balanced based on the facilities, demand, ease of access and so on.

In this article, we will be evaluating the flight fare price prediction using various Regression and classification algorithms. The flight price depends on the features such as:

* Airline
* Date of Journey
* Source
* Destination
* Route
* Dep Time
* Arrival Time
* Duration
* Total Stops
* Additional Info

About the dataset:

* Airline: There were 12 airline companies, offering various price ranges
* Date of Journey: The dataset contains the details from March to June 2019
* Source: This feature represents source location of the flight
* Destination: This feature represents destination location of the flight
* Route: This feature represents the routes covered during the flight
* Dep Time: This feature represents Departure time of the flight
* Arrival Time: This feature represents Arrival time of the flight
* Duration: This feature represents Duration time of the flight
* Total Stops: This feature represents total stops during the flight
* Additional Info: This feature represents extra features during the flight

As per the dataset, we are provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

**EXPLORATORY DATA ANALYSIS:**

Before we start data analysis or run the data through a machine learning algorithm, we must clean your data and make sure it is in a suitable form. Further, it is essential to know any recurring patterns and significant correlations that might be present in your data. The process of getting to know your data in depth is called Exploratory Data Analysis.

Exploratory Data Analysis is an integral part of working with data. In this tutorial titled ‘All the ins and outs of exploratory data analysis,’ you will explore how to perform exploratory data analysis on different data types

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations

Steps Involved in Exploratory Data Analysis

1. Data Collection
2. Data Cleaning
3. Univariate Analysis
4. Bivariate Analysis

**Data Collection:**

* The first step is to import the required libraries
* Read the csv files of both training and testing files merge and convert it to single dataframe df

After extracting the merged dataframe, we have the following conclusions:

* Apart from price all features are categorical
* Price feature is continuous
* Training dataset contains (10683 rows and 11 columns)
* Testing dataset contains (2671 rows and 10 columns)
* Total dataset contains (13354 rows and 11 columns)
* Date of Journey feature needs to be changed into day, weekday, month
* Duration has to be converted into minutes
* Arrival and departure time has to converted into hours for convenience
* Transform the dataset features Date of Journey, Source, Destination, Arrival & Departure features

We have 12 airline companies with 5 flight chartings such as:

* Banglore to New Delhi
* Kolkata to Banglore
* Delhi to Cochin
* Chennai to Kolkata
* Banglore to Delhi
* Mumbai to Hyderabad

Bangalore to New Delhi and Bangalore to Delhi were merged as one and there were 5 destinations.

Since we are having the train and test dataset, we are merging both datasets.

Merging the Source and Destination features into a single feature

Splitting the Date of Journey into Date, Weekday, Month and year

**Data Cleaning:**

In the modified dataframe df3, we find there are some null values in Route, Total Stops, Price. Replace those null values with mode.

Some of the features, are transformed, deleted based on the requirements

* Source & Destination is clubbed into Flight Charting
* Arrival Time & Dep Time is converted to hours as it is better for visualization
* Route – Since stops convey the same meaning
* Date is transformed to Day, weekday, month, year is removed since it happens in the same year.

**Univariate Analysis & Bivariate analysis:**

In Univariate Analysis, we analyse data of just one variable. A variable in our dataset refers to a single feature/ column.

Here we use Line plots, Box plot, Scatter plot, Violin plot. In this plot we compare each feature with the target feature (Price).

From the plots we come to the following conclusions:

Airline:

* Flight price on March 1 was highest when compared to other days
* Trujet is the cheapest with only one trip (Mumbai to Hyderabad)
* Jet airways business is the costliest flight
* Jet airways have more flights and trips
* Multiple carrier flights have only trip from Delhi to Cochin

Duration:

* Chennai to Kolkata have the shortest duration because they are non stop
* Delhi to Cochin take more time that too more time as the no of stops increases

Total Stops:

* Banglore to Delhi has 4 stops
* Mumbai to Hyderabad is non-stop
* Most flight has 2 stops

Additional Info:

* Most flights above ticket price of 40000 are of business class
* Bangalore to Delhi have different additional info.

Weekday:

* Most flights on Friday and next on Sunday (Travellers go their homes on weekend and return before Monday
* Friday’s tickets are costly

Month:

* Most travellers travelled during March month and least during April
* March months tickets are costly
* April month’s tickets are cheap

Date:

* March 1 ticket price is the highest
* 3rd April ticket price is the cheapest

Arrival Hour:

* Most flights arrive at 19:00 hrs
* Flights arrive at 5:00 hrs are costly
* Flights arriving at 2:00 hrs are cheap

Dep Hour:

* Most flights depart at 9:00 hrs
* Flight departing at 1:00 hrs is cheap
* Flight departure at 3,8,16,18 hrs is costly

After visualization, categorical features are converted to continuous features.

**Correlation:**

Correlation is used to check how one or more variables are related to each other. From these variables can be input data features which have been used to predict our target variable.

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives the idea about the degree of the relationship of the two variables. It’s a bi-variate analysis measure which describes the association between different variables. In most of the business it’s useful to express one subject in terms of its relationship with others.

Check correlation and plot heatmap from heat map we find that none of the features are highly correlated. So, we can keep all features.

**Removing the outliers:**

Regarding the outliers, we used two methods, IQR rule and Z-Score

Z-Score is essential to know how many standard deviations away is my actual value from the mean value based on the actual data, you can define the threshold value for the z score to classify a point as an outlier or not in the current scheme of things.

By Z-Score we removed some of the outliers, and the new shape was (13243, 11) and boxplot was also plotted.

Histogram was plot to check the skewness and the distribution

**Training:**

Now the model is cleaned, we are training the model for that we are splitting the dataframe into train and test data

Once the splitting is done, the data is scaled using standard scaler.

Testing:

The model is tested using regression and classification algorithms

The regression models used are:

* Linear Regression
* Gradient Boosting Regressor
* AdaBoost Regressor
* Decision Tree Regressor
* KNeighbors Regressor,
* Extra Trees Regressor,
* Random Forest Regressor

We are taking all these models and looping it from random state 42 to 95 to find the best r2 score,

The training data is 67% and the testing data is 33%

From these iterations we find that the random forest regressor gives the highest accuracy of 87% in the 74th random state.

**Hypertuning**:

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

For the Random Forest regressor, the param grid contains n\_estimators, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf.

These parameters are hypertuned using randomized search cross validation, from this the best parameter and estimator is found.

Using the results from cross validation, we fit Random Forest model using the best model with the best parameter and estimator

From this we can find the following results:

R2 Score : 87.29045154629833

Mean Absolute Error: 113.60667921555623

Mean squared Error: 42662.14334362899

Root Mean Absolute Error: 10.658643404090233

Similarly we test the model using classification algorithms such as Decision Tree Classifier & K KNeighbors Classifier

We are taking all these models and looping it from random state 42 to 95 to find the best r2 score,

The training data is 67% and the testing data is 33%

From these iterations we find that the Decision Tree Classifier gives the highest accuracy of 51% in the 86th random state.

Hypertuning:

For the Random Forest regressor, the param grid contains n\_estimators, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf.

These parameters are hypertuned using Grid search cross validation, from this the best parameter and estimator is found.

Using the results from cross validation, we fit the Decision Tree Classifier model using the best model with the best parameter and estimator

From this we can find the following results:

Accuracy Score : 45.91605218377765

From this we find that regression algorithm of Random Forest regression gives the maximum r2 score of 86%

**Feature importance:**

Feature importance refers to the process of assigning a score to input features based on how important they are at predicting a target variable.

There are many types of feature importance scores, although most examples include statistical correlation scores, coefficients are calculated as part of linear regressor models, decision tree model, and permutation importance scores.

Feature importance scores play an important role in a predictive modeling project, including providing insight into data, insight into the model, and provide the dimensionality reduction and feature selection that can improve the effectiveness and efficiency of a predictive model on the dataset.

Using this model, we will be predicting the test dataset

From the test, we found that Random Forest Regressor give good results:

In Random Forest Regressor, we find the most important features using feature importance options

From this we find that the target variable is mostly dependent upon Duration, Airline, Day, Total Stops.

The purpose of pipeline is to consolidate several steps that can be cross-validated together while setting different parameters. For this, it enables setting parameters of the various steps using their names and the parameter name separated by a '\_\_'. A step’s estimator may be replaced entirely by setting the parameter with its name to another estimator, or a transformer removed by setting it to 'passthrough' or None.

From the pipeline,we get a r2 score of 66.5 with 5 features

**Conclusion:**

The flight price dataset is collected, cleaned, outliers are removed using Z-score, scaled and test model is prepared. Using this data, it is spitted into train and test data and run with regressor and classifier algorithms

Random Forest Regressor algorithms gives an r2 score of 90 after Hypertuning. This dataset can also be predicted by using pipeline consolidating standard scaling and regression algorithms of 5 features which gives an r2 scores of 65

From this Random Forest Regressor model, we can predict the Test data and the flight price is predicted.