Capstone Project - 1 Report

Project Title :

“Fly High and Pay Low!”

Project Mentor : Vikas

What is the Problem to be solved?

There are many instances where lots of people say that the Flight Prices are unpredictable. It is also observed that Flight prices may vary based on the distance, time, stops, etc.. So the problem that I would be solving is the Prediction of Flight Prices.

Who is this project beneficial for?

This project is user beneficial as in for the public. People who fly frequently, people going on a holiday can also be benefited from it price prediction. It is also beneficial for different companies who have to fly their employees or clients.

What is the data used for the project? And where I received the data from?

So the data consists of all flight records in India for a span of 6 months and I have extracted the data from Machine Hack website. This is a multivariate data consisting of origin, destination, departure time, arrival time, price, etc.

Techniques that I would use to solve the problem:

1. Data Wrangling - I will observe the data for all the missing values, incorrect entries and based on the observation I will clean the data for further use.
2. Exploratory Data Analysis - I would plot the relation between different parameters in the dataset and based on that I can find the trends for storytelling. 3. Inferential Statistics - This is another important aspect based on which we can find out some statistical relations in the dataset. I will be applying different testing methods in Inferential Statistics. 4. Machine Learning Model - After all the initial work on the dataset it would be time to build a Machine Learning Model to find the Prediction of the Flight ticket Prices. I may use Linear Regression, Random Forest regression.

What is the methods I will use to deliver the project?

I will be using the following tools to work on the project :-

- Jupyter Notebook for the major part of the project.
- I will be using Python version 3.x for writing the code.
- I will use Google doc for submission on the in process project reports.
- Powerpoint for Data Storytelling presentation.

References:

- Springboard Curriculum material
- DataCamp courses
- Google, Stackoverflow, Github, etc.
- And any other Open Resources.

Data Wrangling Report:

Data Source:

The data is obtained from a www.machinehack.com and was one of the hackathons to participate which was to predict flight prices. The website holds many such different hackathons for Data Science enthusiasts to encourage them to participate in it and enhance one's knowledge. There is a certain reward for top 3 participants of the hackathons.

Initial Data Observations:

The data is split into two datasets: train and test. The train dataset has a shape of (10683,11) and the test dataset is of the shape (2671,10) with Price column dropped. The features of the train dataset are ['Airline', 'Date_of_Journey', 'Source', 'Destination', 'Route', 'Dep_Time', 'Arrival_Time', 'Duration', 'Total_Stops', 'Additional_Info', 'Price']. The column types of the dataset are as follows: **Column Name Type**

Airline object Date_of_Journey object
Source object Destination object
Route object Dep_Time object
Arrival_Time object Duration object
Total_Stops object Additional_Info
object Price int64

Data Wrangling process:

The initial step in the Data Wrangling process was to change the data types of the columns into their appropriate format. The columns that needed data types conversion were Date_of_Journey, Dep_Time, Arrival_Time, Duration, Total_Stops. I converted the Date_of_Journey, Dep_Time, Arrival_Time from object data type to appropriate datetime format and also merged the Date_of_Journey with Dep_Time and set it as the index named Date_Time_Dep where it shows the Date and Time of the departure of

each flight in the dataset. Next, I converted the Duration column into integer signifying the Duration of the flight journey time in Minutes. Finally the Total_Stops column had to be converted to an integer by replacing ‘non-stop’, ‘1 stop’, ‘2 stops’, ‘3 stops’, ‘4 stops’ with 0, 1, 2, 3, 4 respectively.
The data types of the columns after formatting them into appropriate types is as follows:

Column Name Type

Date_Time_Dep Index Airline object
Date_of_Journey datetime64[ns]
Source object Destination object Route
object Dep_Time object Arrival_Time
object Duration int32 Total_Stops int32
Additional_Info object Price int64

Missing Value Treatment:

The columns of Route and Total_Stops each had a single missing value. Since the missing values were not in great number compared to the shape of the dataset I just removed the entire row with missing values. The shape of the dataset after removing the missing values is (10682,11).

Outlier Treatment:

Based on the initial observation there were some extreme values in the Price column but after further investigating the reason for those values were the Class of the Airline. The Jet Airways Business Airline had higher Price compared to other Airlines. This extreme value of Price cannot be considered as an outlier since it will predict the Price of luxury airlines in the test dataset. So the dataset had some extreme values for some features but those couldn’t be considered as outliers for the prediction to make using the test dataset.

Now the dataset is mostly clean and ready for further process of Exploratory Data Analysis, Data Visualization and Machine Learning modelling.

Exploratory Data Analysis & Data Storytelling:

Variables that are Significant in Explaining Project Questions:

How the Price will vary with different variables in consideration?

Flight Price varies due to many factors and each factor should be taken into consideration. The variables that affect the Flight Price are:

- Date of Departure: The Departure Date matters in the change of Price because of many reasons. During March and April months the Price will be higher because of more demand for tickets since families frequently travel to their natives and for holidays during that period.
- Time of Departure: Also, the time of day is important in Price variation. Early morning flights are usually of higher side of price range because most of Business flyers prefer early flights for their travel. Also late night flights are less in demand because of low travellers.
- Total Stops: The layover between flights are another factor in Price change. The more Stops in between flights accordingly the Price will also increase.
- Class of Airlines: The Class of Airlines is very important to check the variations in Price. Business and Premium classes will charge more amount for tickets compared to Economy class.

Correlation between pairs of Variables:

What does a pair of Variables suggest when compared with each other?

- Total Stops and Duration: These two variables are interrelated with each other in a precise manner. As the total number of stops of a certain flight increases so does the duration between the Source and Arrival Destination increases. If the flight has 4 stops obviously the duration of the flights will be on the higher side than the flight with 1 stop. There is a positive correlation of 0.74 between Total Stops and Duration
- Total Stops and Price: As the number of stops increases for a flight the Price increases. The positive correlation of 0.604 can be observed between Total Stops

and Price.

● Source and Price: There is also a certain correlation between Source of the flight and the corresponding Price of the flight ticket. Certain cities have a higher demand of flights to particular destination so accordingly the Price of the ticket from that Source city will be higher.

● It can be seen that around 80% of the flight prices are in between **5000 to 16000**. The peak price for most of the flights is **4000 to 6000**. Also there are some flights having a price above 20000 but those cannot be considered as an outlier since it can be a different class of an Airline.

● Also it is clear that the flight prices are from a range of **1759 to 79512** having a mean of **9087**. Again the maximum Price of **79512** is because of the class of the Airline. The higher Price suggests that the Airline is providing a Luxury or Business Class service for their passengers. But most of the Price of the flight tickets lie in the range of **5000 to 12000**.

● The flights that have 1 stop are highest in number followed by flights with 0 stop and then 2 stops respectively. This trend shows that in India most of the flights are connecting flights with **1 stop** having a frequency of **52.66%** followed by **32.68%** flights with **0 stops** and **14.23%** flights with **2 stops**. Also there are very less number of flights with 3 stops and one single flight having 4 stops.

● The highest number of flights are of **Jet Airways** with a share of approxiamately **36%**. This suggests that **1/3rd** proportion of the flights in India are operated by **Jet Airways**. These can be direct or flights with 1 or 2 stops. The next top Airline in India is **IndiGo** with **19.22%** very closely follows the **Air India** airlines with **16.4%** flights in the market. The other airlines economical airlines such as **SpiceJet**, **Vistara**, **Air Asia**, **GoAir** shows that they don't have much connectivity in India.

● The highest number of flights are of **Duration** within **100-200 minutes** that is **1.40 to 3.20 hours**. Also there are certain number of flights between **300 to 1000 minutes**. It show that people mostly prefer flights with less Duration but also a lot of the population takes connecting flight to reach smaller places around India.

● It can be observed that flights taking off from **Bangalore** have certain extreme ticket price suggesting that most **Luxury Class Airlines** flights fly from **Bangalore** and closely followed is **Delhi**. The other places that are **Kolkata**, **Chennai** and **Mumbai** seem to have **Economical Flights** flying out these cities.

● Also there is an unusual trend seen here that **1 Short Layover** and **2 Long Layover** have approximately similar price range. This can be because of many different reasons that during **2 Long layover** there would be multiple carriers of

Airlines included. You cannot exactly tell the reason behind this trend.

● There are certain flights with less duration having more price and certain other flights with more duration and less price. But most of the Price is concentrated from **300 to 1600 minutes** duration of the flights. There is a slight increasing trend of **Price** with increase in **Duration**.

● The **Scatter Plot** clearly shows that the frequency of flights is the most from **5:30 Am to 11 AM** and again from **4:30 PM to 10 PM**. It obviously gives a clear trend that people mostly would prefer early morning to afternoon flights and again the evening flights. These times most of the people who are travelling for work can be seen on the flights. Also there are very few **Domestic Flights** operating in India during the **Midnight**. This suggests that there are very less flyers in the night times and most of the people prefer Daytime flights.

● **Jet Airways** tops the list with most number of flights taking off from **Delhi** which is then followed by **Air India** and **IndiGo**. **Kolkata** has the 2nd highest number of flights of **Jet Airways** taking off from there with other few airlines **Air India**, **IndiGo** with less flights than **Jet Airways**. **Bangalore** again has a decent number of flights of **Jet Airways** and **IndiGo**. From **Chennai** it can be seen that there are very low or nil number of flights of **Air Asia**, **GoAir** and **Jet Airways** airlines.

Machine Learning:

The general machine learning framework is outlined below:

1. **Prediction Engineering:** State the business need, translate into a machine learning problem, and generate labeled examples from a dataset. 2. **Feature Engineering:** Extract predictor variables — features — from the raw data for each of the labels. 3. **Modeling:** Train a machine learning model on the features, tune for the business need, and validate predictions before deploying to new data.

Type of Machine Learning:

The dataset comes under Supervised Learning since it has labelled data. Basically supervised learning is a learning in which we teach or train the machine using data

which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data. Supervised learning classified into two categories of algorithms:

● **Classification**: A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.

● **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”. In this case our Supervised Learning is a Regression algorithm where we have to predict the Prices of the Airline tickets.

The Process of Prediction Engineering:

Prediction engineering requires guidance both from the business viewpoint to figure out the right problem to solve as well as from the data scientist to determine how to translate the business needs into a machine learning problem. The inputs to prediction engineering are the *parameters* which define the prediction problem for the business requirement , and the historical dataset for finding examples of what we want to predict.

Feature Engineering:

Feature engineering, the second step in the machine learning pipeline, takes in the label times from the first step — prediction engineering — and a raw dataset that needs to be refined. Feature engineering means building features for each label while filtering the data used for the feature based on the label’s cutoff time to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

The Machine Learning Modeling Process:

The outputs of prediction and feature engineering are a set of label times, historical examples of what we want to predict, and features, predictor variables used to train a model to predict the label. The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business needs, and validating it on holdout data.

Machine Learning Algorithms:

The Algorithms used to build a model and predict the outcomes are:

1. Linear Regression: Since it is a regression problem the first model which we use to fit and predict will be Linear Regression. The main features which we can relate using Linear Regression are duration and Price in the dataset. Duration being an independent feature and Price is a dependent feature.

2. We will regularize the above regression using Lasso, Ridge, ElasticNet along with GridsearchCV for hyperparameter tuning for regularization techniques

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