Flight Fare Prediction

High Level Document (HLLD)

2021 June Prashanth Kumar G V

Document Version Control

Month	Version	Description	Author
July 21	V.0	Initial HLD	Prashanth Kumar GV
July 21	V.0	Initial HLD	Prashanth Kumar GV

Contents

ABSTRACT

Airline companies use complex algorithms to calculate the flight prices for given various factors present at the particular time. These methods take financial, marketing and various social factors into account to predict flight fares.

Nowadays the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices changes dynamically due to different factors. That's why we try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help the customers to predict future flight prices and plan their journey accordingly.

1. INTRODUCTION

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:

o Security

o Reliability

o Maintainability

o Portability

o Reusability

o Application compatibility

o Resource utilization

o Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.2 Definitions

Airline : Name of the airlines are operating.

Date_of_Journey : Travelling date.

Source : Starting travel point.

Destination : Ending travel point.

Route: Travel route (Via)

Dep_Time : Departure time of the flight.

Arrival_Time : Arrival time of the flight.

Duration :

$$TravelDuration=(Departuretime-Arrivaltime)$$

- Time taken by flight to reach destination point called Duration

Total_Stops : Number of stops of flight to destination point.

Additional_Info : Additional information of the flight, flight routs, time etc.

Price : Price /Fare of the flight.

2. GENERAL DESCRIPTION

2.1 Problem statement:

To predict flight fare charges of various airlines according to the 2019 dataset.

2.2 Data Overview:

It contains the train.xlsx excel format file contains the data related Airlines prices based on the various parameters as mentioned in the 1.2 Definitions.

- Downloaded the dataset from https://www.kaggle.com/nikhilmittal/filght-fare-prediction-mh
- train.xlsx

2.3 Problem type :

Continuous changing the prices /fare of the flights depends on the various parameters as mentioned in 1.2 Definitions. Price is the ground truth variable and independent variables for this dataset based on those parameters price is decided, so price is not fixed with one variable it is continuously changing.

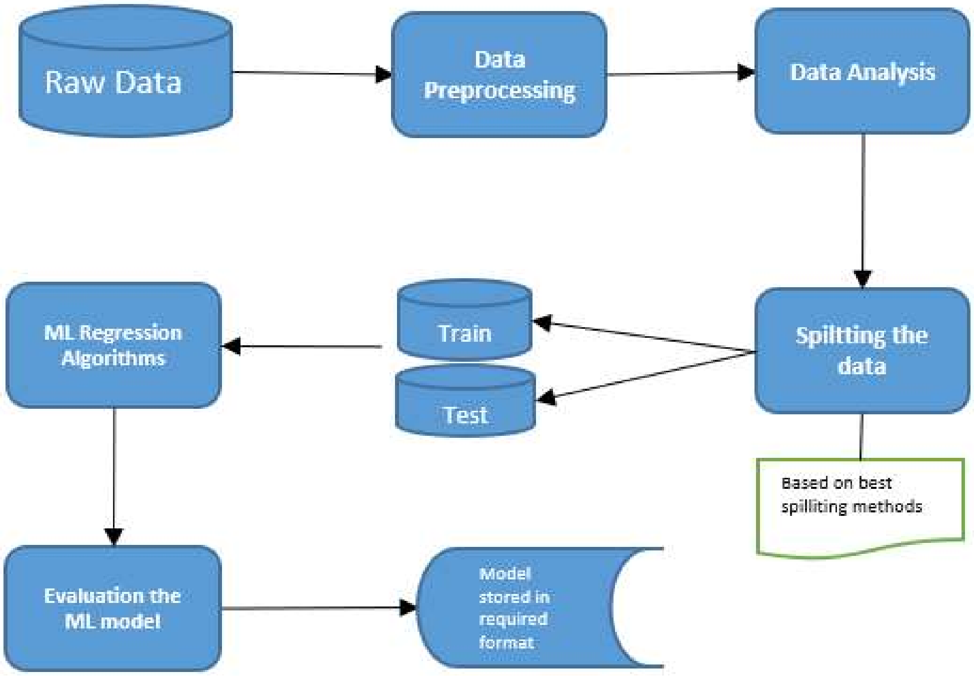
Hence this problem is **Regression Problem**.

2.4 Used Tools :

1. Jupyter Notebook used.
2. Python is used to execute the results.
3. For visualization of the plots, Matplotlib, Seaborn are used.
4. Numpy and Pandas are used.
5. Excel is used to retrieve, insert, delete, and update the database.
6. Feature engineering is performed to the dataset to extract new features for useful insights.
7. Sklearn modules used.
8. Building the machine learning models (Supervised machine learning algorithms).
9. Performance matrics are implemented to find the best model to predicting.

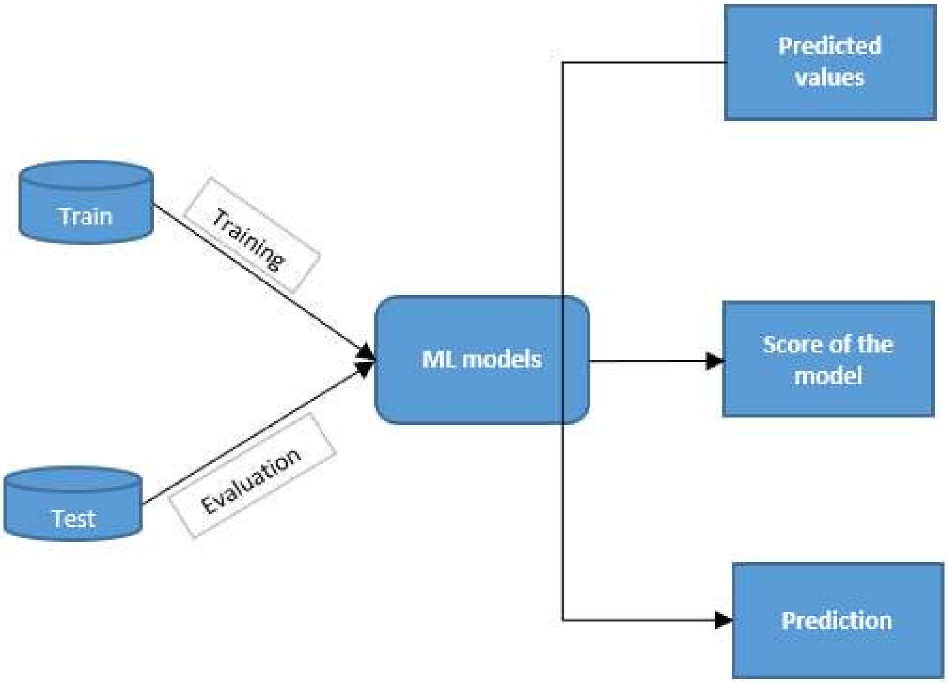
3. DESIGN

3.1 Process Flow:

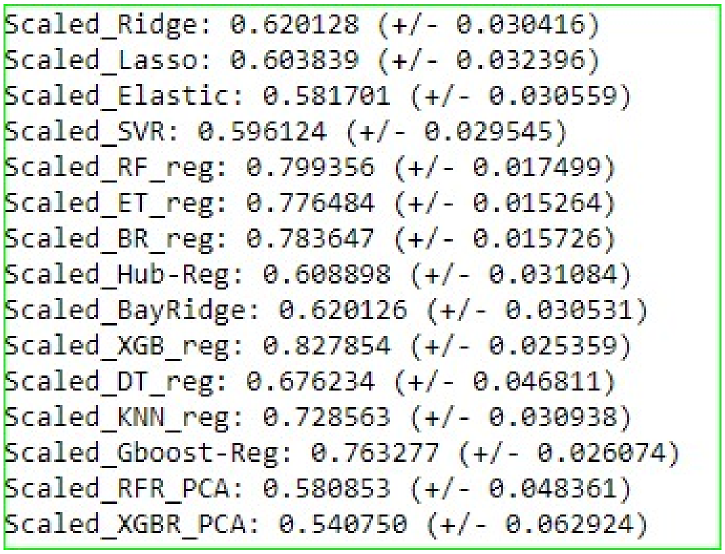


1. Data is collected from Kaggle.com, raw data is in excel format.
2. Importing the data by using pandas' data frames into the python notebook to perfume the various operations and building useful insights.
3. Data visualization by using the matplotlib, seaborn libraries for analysing the data and exploring the data in different perspectives. And also, we find the useful insights related to data.
4. Data Analysis for relation between the different features which finds the relations between the one feature to another feature to find the ground truth predictable.
5. After analysing the data by visualization and various statistical parameters and methods we splitting the data into train and test because first we train the data to ML model then we test the Model to test data.
6. Evaluating the model by performance metrics to pick the best model accurately predicting the truth variables based on the train set.
7. Store the model into standard format like JSON, PICKLE, SQL etc for the further process like building the web APIs or model deployments and testing.

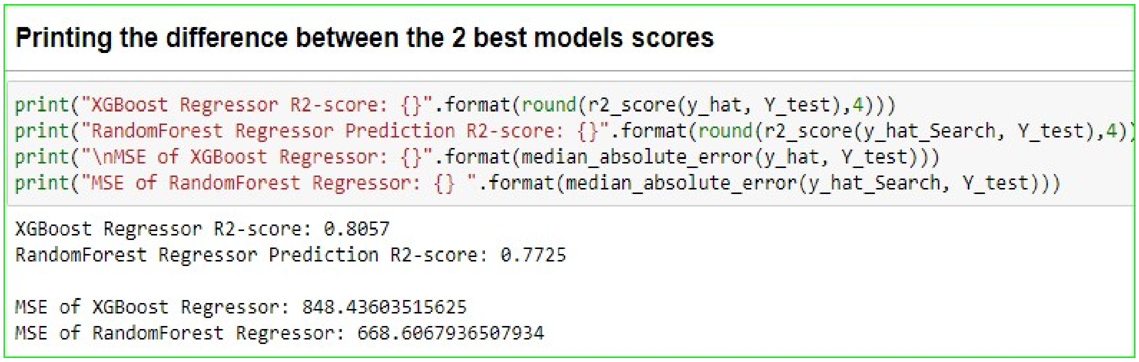
3. 2 Performance and Evaluation :



1. We train the model on various supervised regression machine learning algorithms.
2. We got the best score by XGBoost and Random forest algorithms
3. Testing the model by best models as mentioned as above.
4. Hyperparamter tuning the algorithms by Random search.
5. After all, performing the model we got the 0.843 score and R2 error of 0.857.



```
Scaled_Ridge: 0.620128 (+/- 0.030416)
Scaled_Lasso: 0.603839 (+/- 0.032396)
Scaled_Elastic: 0.581701 (+/- 0.030559)
Scaled_SVR: 0.596124 (+/- 0.029545)
Scaled_RF_reg: 0.799356 (+/- 0.017499)
Scaled_ET_reg: 0.776484 (+/- 0.015264)
Scaled_BR_reg: 0.783647 (+/- 0.015726)
Scaled_Hub-Reg: 0.608898 (+/- 0.031084)
Scaled_BayRidge: 0.620126 (+/- 0.030531)
Scaled_XGB_reg: 0.827854 (+/- 0.025359)
Scaled_DT_reg: 0.676234 (+/- 0.046811)
Scaled_KNN_reg: 0.728563 (+/- 0.030938)
Scaled_Gboost-Reg: 0.763277 (+/- 0.026074)
Scaled_RFR_PCA: 0.580853 (+/- 0.048361)
Scaled_XGBR_PCA: 0.540750 (+/- 0.062924)
```

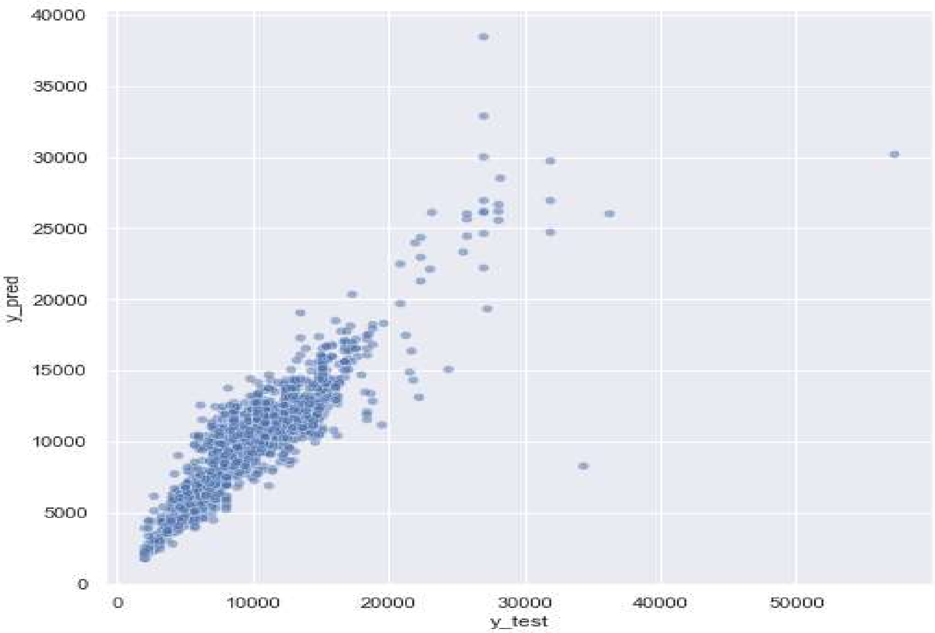


Printing the difference between the 2 best models scores

```
print("XGBoost Regressor R2-score: {}".format(round(r2_score(y_hat, Y_test),4)))
print("RandomForest Regressor Prediction R2-score: {}".format(round(r2_score(y_hat_Search, Y_test),4)
print("\nMSE of XGBoost Regressor: {}".format(median_absolute_error(y_hat, Y_test)))
print("MSE of RandomForest Regressor: {} ".format(median_absolute_error(y_hat_Search, Y_test)))
```

```
XGBoost Regressor R2-score: 0.8057
RandomForest Regressor Prediction R2-score: 0.7725

MSE of XGBoost Regressor: 848.43603515625
MSE of RandomForest Regressor: 668.6067936507934
```



4. CONCLUSION

Flight fare is automatically predicted by choosing the Number of stops, Place of destination & source, travel time and Airline’s operators.

Machine learning model predicting 85.7% accurate prices of the travel.

References

1. Krish Naik youtube Classes
2. Gitub.com/Mandal-21