Detail Project Report (DPR)

Flight Fare Prediction

An airplane flying in the sky

Description automatically generated

Project Member: Rohit Atnoor

Mentor: Ineuron

Latest Version: 1.7

Last Data of Revision: 19/09/2023

**Objective**:

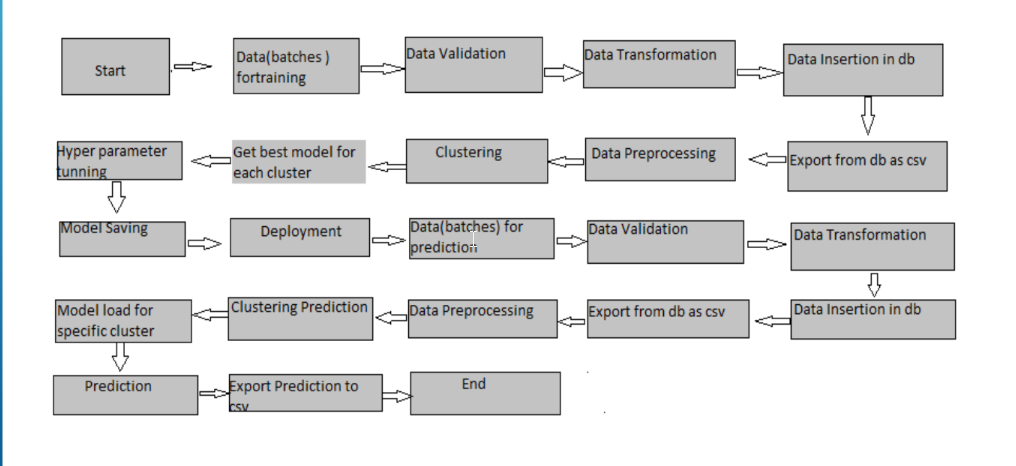
To develop a predictive model for Flight Fare Prediction. The Model will Deliver the user an approximate Flight Fare from any Source Station to any Destination station.

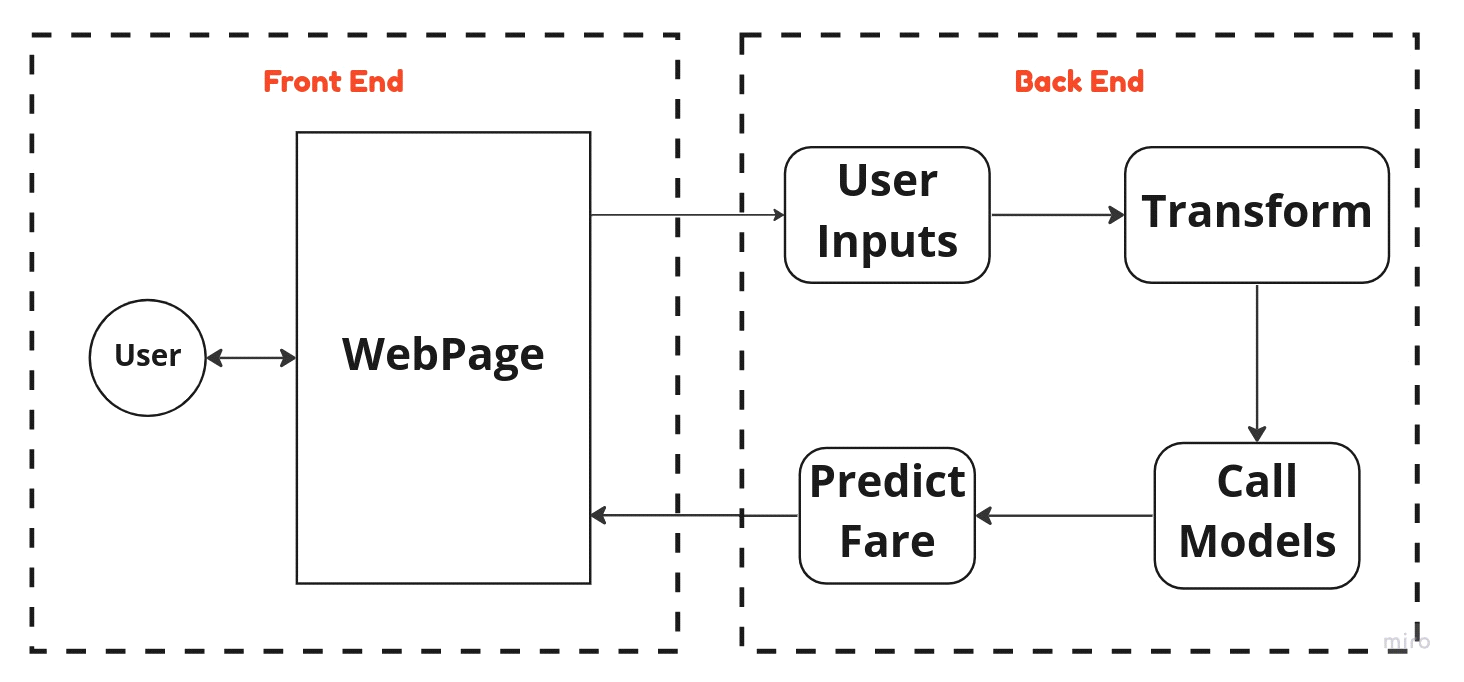
**Benefits**:

User can get an approx idea about the flight fare without visiting any booking website signup.

Users can compare the price of different Airlines for any specific Route and choose the best or affordable Airline

**Architecture**

****



**Data Insertion:**

1. Insert data into database (MySQL) and load and save the data in feature store under Artifact Directory.

**Data Validation**

1. Validation of number of features or columns present in new dataset.
2. Validation of each required feature present or not.
3. Distribution of numerical features are same or not.
4. Null value percentage for each column should be less than critical value already given.

**Data Transformation:**

1. Extraction of day and month from date features
2. extraction and calculation of total duration in minutes.
3. Label encoding or One Hot Encoding for categorical features.

**Model Training:**

1. The accumulated data from db is exported in csv format for model training.

2. Data Preprocessing

* Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc.
* Check for null values in the columns. If present impute the null values.
* Encode the categorical values with numeric values.

3. After completing FE, useful features are selected. Seperate grouped data are created from the main dataframe for individual Airline.

4. Model are created mainly using 3 Algorithms -

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor

5. Best model is selected on the basis of model performance and accuracy.

**Model Selection**

After the clusters are created, we find the best model for each cluster. By using 2

algorithms “SVM”. For each cluster both the hyper tunned

algorithms are used. We calculate the AUC scores for both models and select the

model with the best score. Similarly, the model is selected for each cluster. All

the models for every cluster are saved for use in prediction

**Prediction**:

* The testing files are shared in the batches and we perform the same Validation
* operations ,data transformation and data insertion on them.
* The accumulated data from db is exported in csv format for prediction
* We perform data pre-processing techniques on it.
* Once the prediction is done for all the clusters. The predictions are saved in csv format and shared.