Architecture Design

Flight Fare Prediction - System

Project Detail

Project Title	**Flight Fare Prediction - System**
Technology	Business Intelligence
Domain	Aviation
Project Difficulty Level	Intermediate
Programming Language Used	Python
Tools Used	Pycharm, Docker, CiCd

Objective:

The goal of this project is Develop a Model to predict Flight Fare for different Flight. The Model help to predict the fare for different Flight based on different factors.

Benefits:

- Detection of Flight Fare.
- Gives better insight of customers base.
- Helps in easy flow for managing resources.
- Getting some basic idea of the Flight fares. So, before planning the trip will surely help many people save money and time.

Problem Statement:

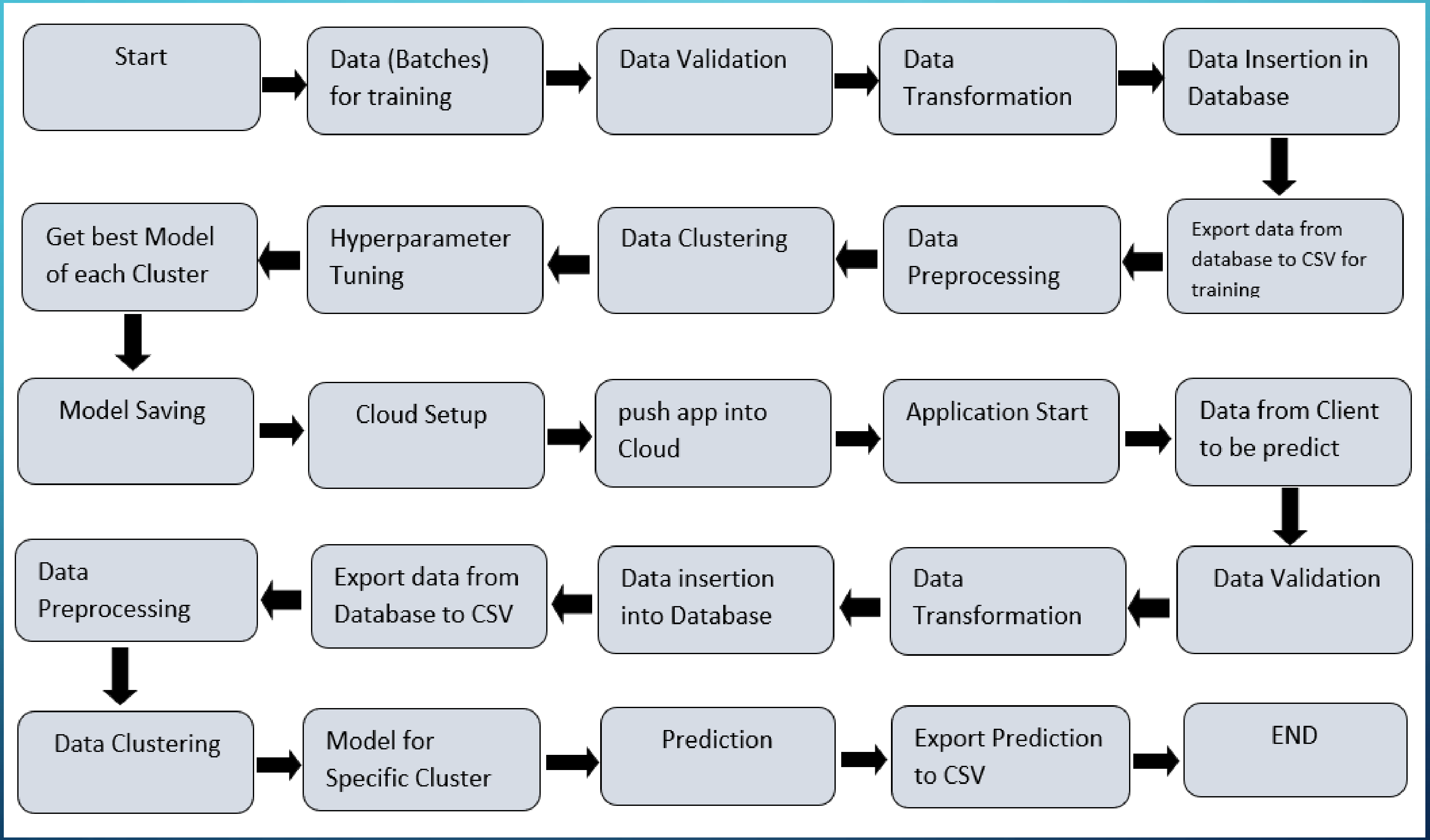
Flight ticket Fare can be something hard to guess, today we might see a Fare, check out the Fare of the same Flight tomorrow, it will be a different. We might have often heard travelers saying that Flight ticket Fare are so unpredictable.

The main aim of this project is to create an AI solution to predict Flight Fare. So, before planning the trip it will surely help many people to save money and time.

Data Sharing Agreement :

- Sample file name (eg: FlightPrice_20062021_101010)
- Length of date stamp(8 digits)
- Length of time stamp(6 digits)
- Number of Columns
- Column names
- Column data type

Architecture



Start

Data (Batches) for training

Data Validation

Data Transformation

Data Insertion in Database

Export data from database to CSV for training

Data Preprocessing

Data Clustering

Hyperparameter Tuning

Get best Model of each Cluster

Model Saving

Cloud Setup

push app into Cloud

Application Start

Data from Client to be predict

Data Validation

Data Transformation

Data insertion into Database

Export data from Database to CSV

Data Preprocessing

Data Clustering

Model for Specific Cluster

Prediction

Export Prediction to CSV

END

Data Validation :

- Name Validation - Validation of files name as per the MDM. We have created a regex pattern for name validation. It check file name as well as date format and time format if these requirements are satisfied, we move such files to "Good_Data" else "Bad_Data“ directory.
- Number of Columns – Validation of columns number present in the files, and if it doesn't match then the file is moved to "Bad_Data“ directory.
- Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad_Data“ directory.

Dataset Information:

Airline: Airline provide services.

Date_of_Journey: Passenger Journey Date date from Source location.

Source: Place where Flight will start journey.

Destination: Passenger Destination.

Route: Route follow by the Flight to travel from source to destination.

Departure Time: when Flight will start journey.

Arrival Time: Time when Flight reaches the destination.

Duration: Time taken to travel from source to destination.

Total_stop: Number of stops in the journey.

Additional info: Extra facility provide by the Airline organization.

Price: Traveling Cost.

Airline: Airline is an another important parameter because depend on this parameter identify how many times Airline provide services as well as Airline organization provide service. Depend on this services Model can build a relation to find Fare.

Date_of_Journey: This feature can identify that have any special occasion on that date so the Flight Travel Fare will be calculate depend on this special occasion. Similarly if that particular date have no occasion then corresponding Fare will be calculated.

Source: This parameter identify source place. Flight provide service for some busy city for this characteristic Flight Fare may be high in.

Destination: This parameter identify destination place. Flight provide service for some busy city for this characteristic Flight Fare may be high in sometime.

Route: Follow the route in travel. Sometime between source and destination have long distance for that reason Flight take stoppage. So stoppage will be high Flight fare should be going to less.

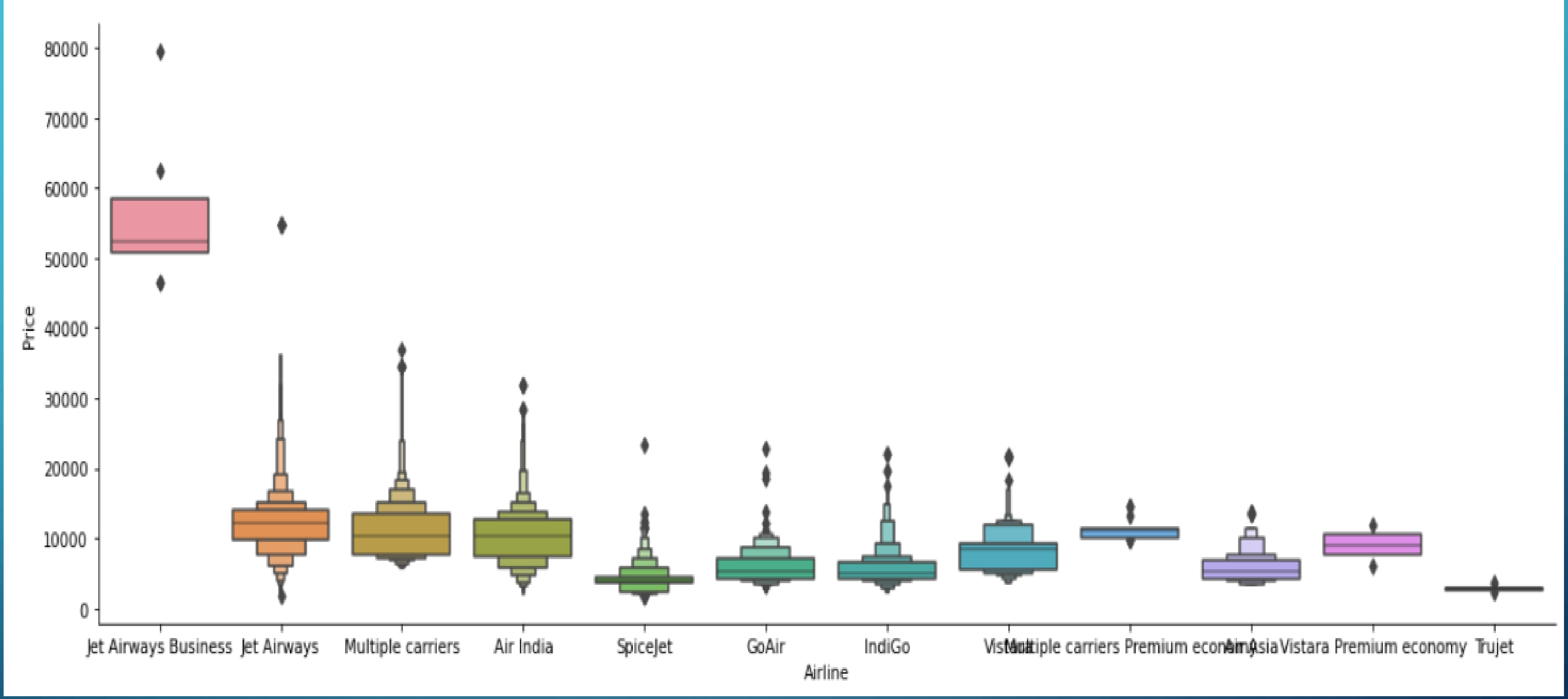
Departure Time: Depend on the departure time Flight Fare may be increase or decrease. As in office time Fare is high, in morning and night also Flight Fare is less. So This feature also important to predict the Flight Fare.

Arrival Time: Depend on the arrival time Flight Fare may be increase or decrease. As in office time Fare is high, in morning and night also Flight Fare is less. So This feature also important to predict the Flight Fare.

Duration: This feature determine time taken in entire journey from source to destination. If duration is high Fare may be high and vice versa.

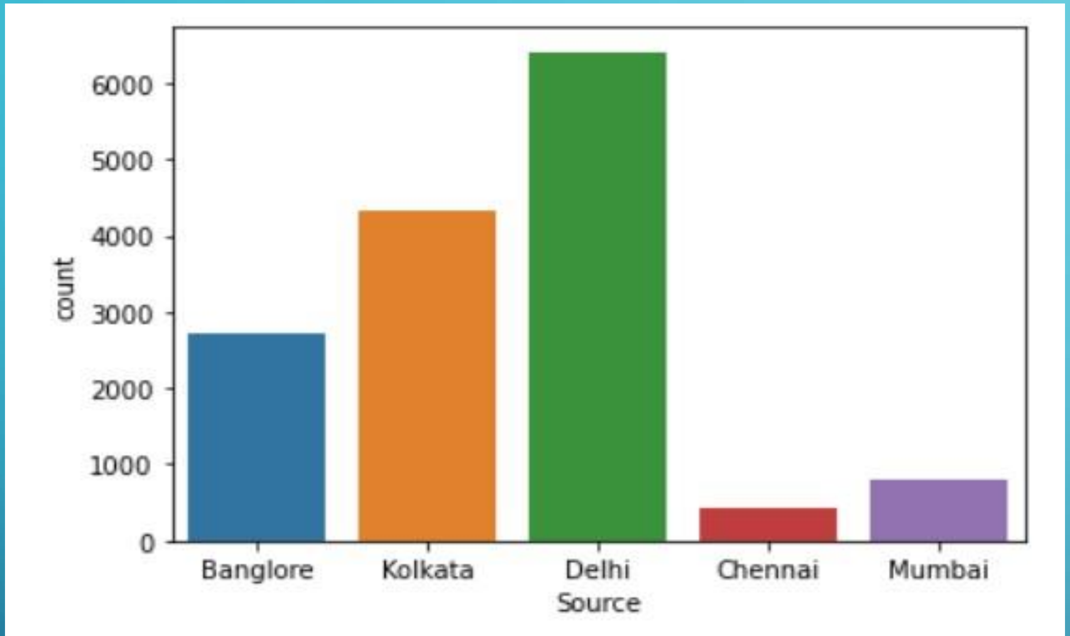
Additional info: Extra facilities provide by the Flight organization. Depend on the facility Flight Fare also going to be increase or decrease.

Price variation based on Airline

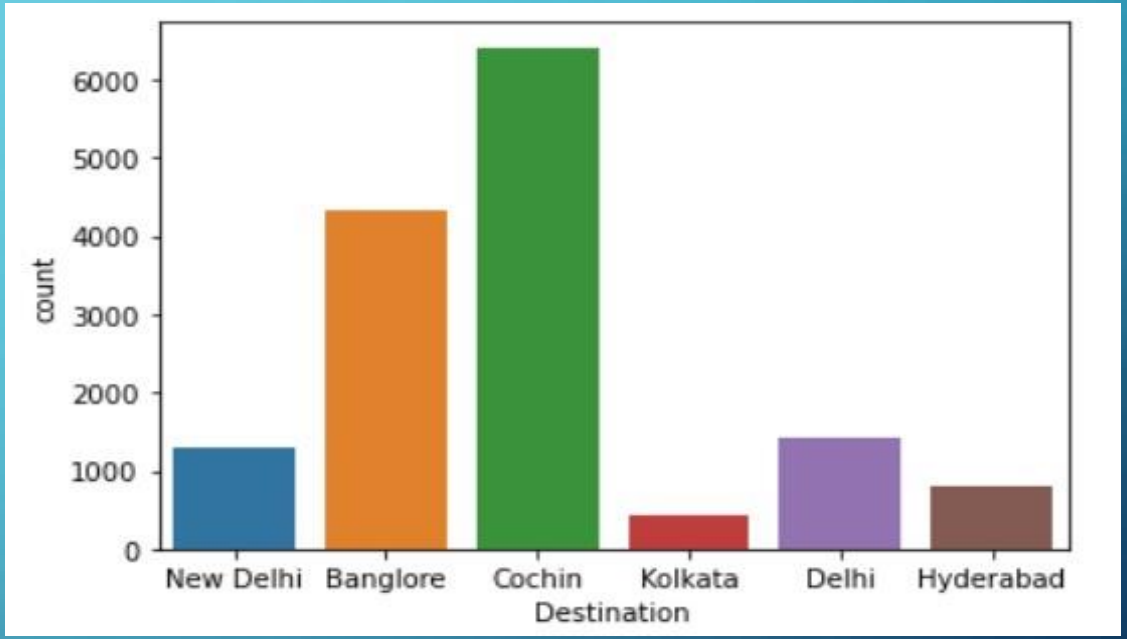


Price based on Airline

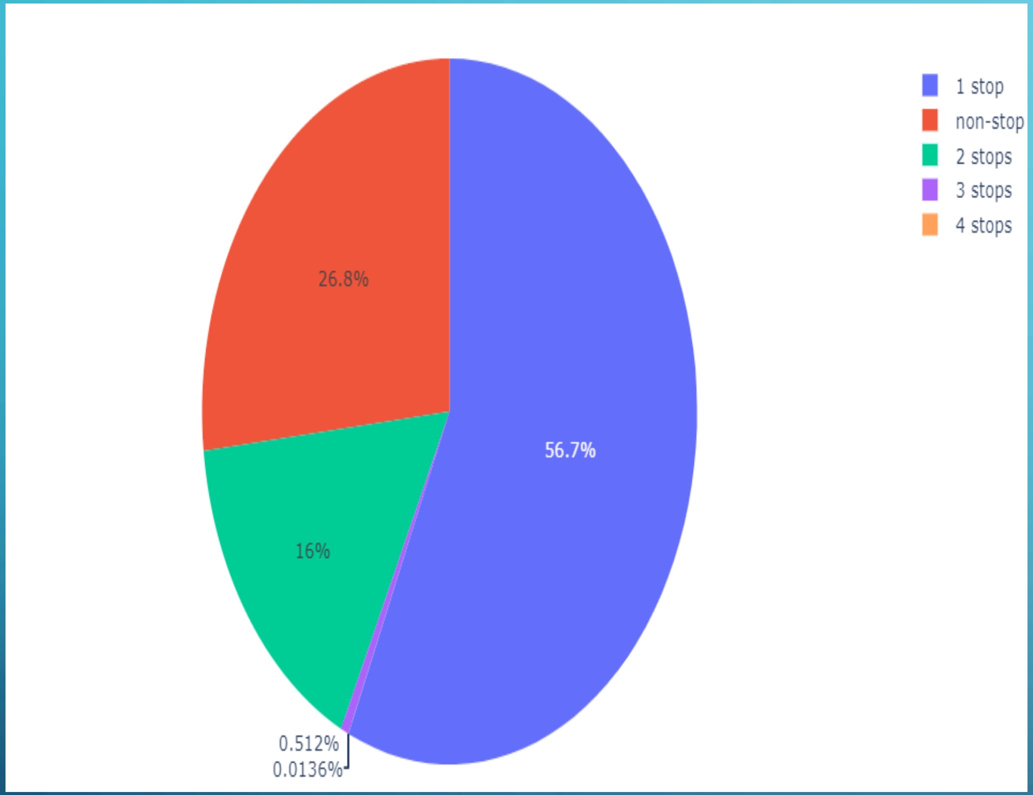
Departure and Destination City



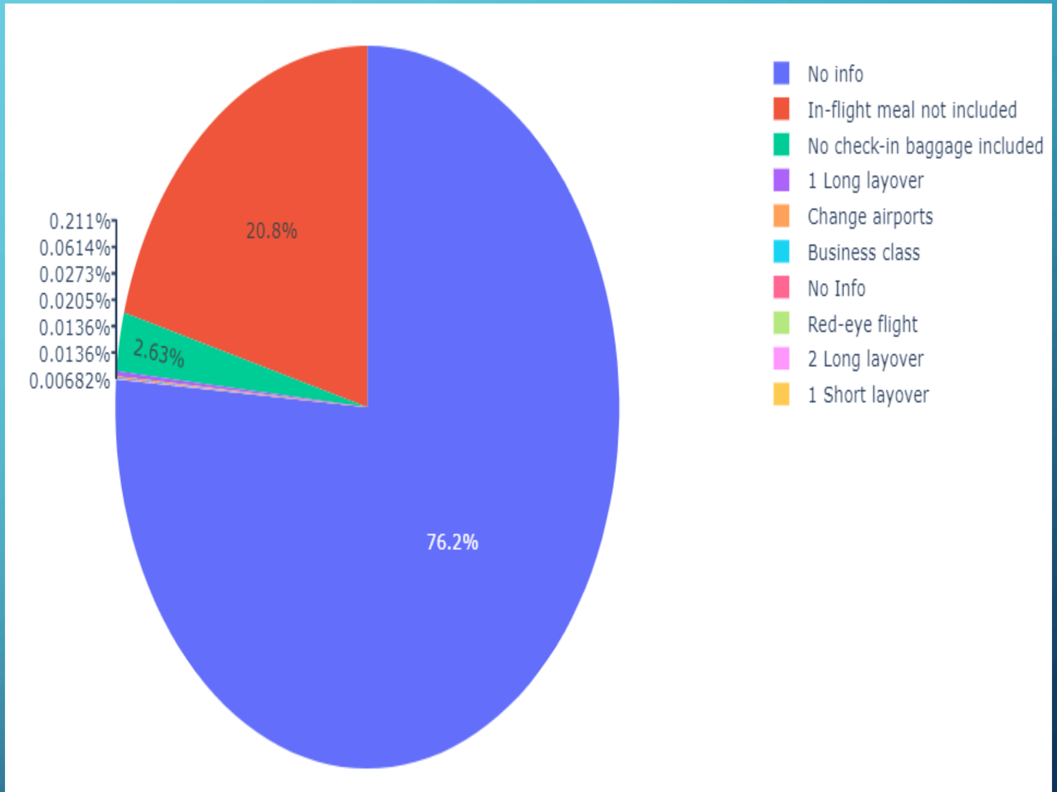
Departure City



Destination City



Total_Stops in Pie-chart format



Additional-info in Pie-chart format

Data Insertion in Database:

- Table creation - Table name “**Good_Data**" is created in the database for inserting the files. If the table is already present then new files are inserted in the same table.
- Insertion of files in the table - All the files in the "Good_Data“ directory are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table.

Model Training:

- Data Export from DB :

  The accumulated data from DB is exported in csv format for Model training.

- Data Preprocessing :

  - Performing EDA to get insight of data like identifying outliers ,Null, duplicates data etc.
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.

- Clustering:
  - KMeans algorithm is used to create clusters for the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using KneeLocator function. The idea behind clustering is to implement different algorithms on the structured data.
  - The Kmeans Model is trained over preprocessed data, and the Model is saved for further use in prediction to find cluster.

- Model Selection:

After the clusters are created, we find the best Model for each cluster, by using 2 algorithms “RandomForestRegressor” and "XGBoost". For each cluster hyperparameter tuning is applying on each algorithms to find the best Model. We calculate the R_square and Adjusted_R_square scores for both Models and select the Model with the best R_square value and Model is selected for each cluster. All the Models for each cluster are saved for use in prediction.

Prediction:

- The testing files are shared in the batches and perform the same data Validation , data transformation and data insertion.
- The accumulated data from DB is exported in csv format for prediction.
- We perform data pre-processing techniques on dataset.
- KMeans Model created during training and save clusters Model to use on the prediction preprocessed data to find cluster.
- Based on the cluster number respective Machine Learning Model is loaded and is used to predict the data for that cluster.
- Once the prediction is done for all the clusters. Results saved in csv format and shared.

Conclusion:

The Designed Flight Fare Prediction System predict the Fare based on Dataset shared by the user. So that based on the dataset we can develop a Model and Model make the prediction based on various parameter. so the traveler having some basic idea of the Flight fares before planning the trip. It will surely help many people by saving money as well as time.

Q & A

1) What’s the source of data?

The data for training is provided in batches.

2) What was the type of data?

The data was the combination of numerical and Categorical values.

3) What’s the complete flow you followed in this Project?

Refer slide 7th for better Understanding.

4) After the File validation what you do with incompatible file or files which didn’t pass the validation?

Those incompatible files are moved to the Archive directory from the bad data directory.

5) How logs are managed?

We are using different logs as per the steps that we follow in validation and Model building, training like File_validation_Log, Data_Insertion_Log, Database_Log, Model_Training _Log, prediction_Log etc.

6) What techniques we are using for data pre-processing?

- Visualizing relation of independent variables with each other and output variables
- Removing outliers
- Removing Duplicate values.
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Removing unwanted attributes

7) How training was done or what Models were used?

- Before diving the data in training and validation set we performed clustering to divide the data into clusters.
- As per each cluster, data divided into training and validation dataset.
- Algorithms like RandomForestRegressor , XGBoost were used based on the recall final Model was used for each cluster and we saved that Model for Prediction data shared by Client .

8) How Prediction was done?

The testing files are shared by the client .Perform the same life cycle till the data is clustered. Then on the basis of cluster number Model is loaded and perform prediction. At the end we get the accumulated data of predictions and shared in specific location in CSV format.

THANK YOU