International Journal of Advanced Development in Science and Technology
Volume : 3 Issue : 03 April 2021 www.ijadst.com e-ISSN: 2582-1059

FLIGHT DELAY PREDICTIONS USING SUPERVISED MACHINE LEARNING

[1] Vishaal.S, [2]Adhithyaa Vishal .T,
[1,2] Student, [1, 2] Prince ShriVenkateshwaraPadmavathy Engineering College, Ponmar,
[1] vishaalfeb2000@gmail.com.

ABSTRACT

The primary goal of this project is to predict airline delays caused by various factors. Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. The growth of the aviation sector has made flight delays more common across the world.

They cause inconvenience tithe travellers and incur monetary losses to the airlines. We analysed the various factors responsible for flight delays and applied machine learning models such as Random Forest, XGBoost, KNN, Decision Tree to predict whether a given flight would be delayed or not. Also with certain features we can predict how far the delay is going to be using some regression techniques like Random Forest Regression and Decision Tree Regression.

***Keywords*: flight delay prediction,supervised machine learning ,random forest ,k nearest neighbours ,decision tree,extreme gradient boosting,confusion matrix ,classification and regression techniques.**

I. INTRODUCTION

During the most defining period of human history, where computing has moved from mainframes to PCs to cloud, and now to artificial intelligence. A fundamental sub-area of artificial intelligence has come into notice, called as Machine Learning, which enables computers to get into amide of self-learning without being explicitly programmed.

With the concept of machine learning, we have been able to apply complex mathematical computations to big data iteratively and automatically, that too with efficient speed, this phenomenon has been encompassing momentum over the last several years.

On the other hand, data mining involves data discovery and sorting it among large data sets available to identify the required patterns and establish relationships with the aim of solving problems through data analysis.

Simply combining, machine learning and data mining use the same type of approach and set of algorithms, except the kind of data pre-processing and end prediction varies. By

combining these two core areas to predict and present the most accurate results possible.

A. Supervised Machine Learning

It is a machine learning task where the dataset inputs andoutputs are clearly recognized and already given, then several type of algorithms are trained using labelledexamples. A supervised learning algorithm contains an entire dataset, which is further divided into training and test data; the algorithm examines the training dataset and produces an inferred function, which is then used for mapping new examples. In case of the aviation industry, commercialized aviation is a type of transportation system that is complexly distributed.

It tends to deal with several important resources, demand fluctuations, and various other kinds of stages. Stages are bound to take place at terminal boundaries, runways, airports, and distinguished airspaces that may be susceptible to different kind of delays or cancellations. Summing up, some set of examples include weather conditions, ground delays, air traffic control and several other constraints and unforeseen circumstances that lead to delays and cancellations in the entire aviation industry.

Hence, this becomes an optimal scenario which will allow us to implement a supervised machine learning algorithm to precisely determine and predict the class labels for unrevealed instances. Supervised Learning algorithm here will model relationships and dependencies between the aimed prediction output and the input features, such that I'll be predicting the output values for new data based on the relationships which are learned from the previous data set. Supervised Learning problems can be further categorized into following problems.

- Classification – It is a type problem in which the output variable is an entire category itself, such as "Win" or "Lose", the entire input data is classified into the category variables; it is generally used largely for recommendation problems.
- Regression – It is a type of problem is which the output variable is a real value, such as few raw data values related to something. This is the problem type massively used for prediction analysis, and hence will be used in this project.

B. Classification Methods

Random forest Classification

It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Decision Tree Classification

It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

K Nearest Neighbours Classification

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN is used in statistical estimation and pattern recognition.

XGBoost Classification

Gradient boosting is a method where the new models are created that computes the error in the previous model and then leftover is added to make the final prediction. It uses a gradient descent algorithm that is the reason it is called a "Gradient Boosting Algorithm"

C. Regression Methods

Random Tree Regression

It uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

Decision Tree Regression

A Decision Tree imposes a series of questions to the data, each question narrowing possible values, until the model is trained well to make predictions. It can make a prediction by running through the entire tree, asking true/false questions, until it reaches a leaf node. The final prediction is given by the average of the value of the dependent variable in that leaf node.

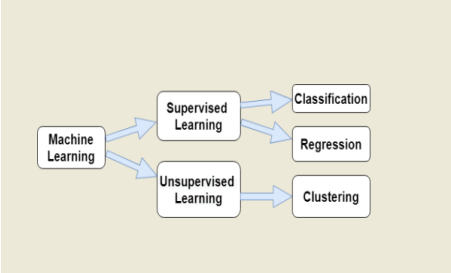

Fig.1: Overview and classification of Machine Learning

II. RELATED WORKS

Flight Delays has become a common and complex phenomenon, it occurs due to the problems at the originairport, at the destination-airport, any ground reasons or a combination of these entire factors can also give rise to delays. Delays are also being regarded as caused due to specific airlines. Even if it is complex, it is still measurable with decent accuracy And with respect to the schedule an on-time performance of airlines, their generally exists some pattern of flight delay (Wu, 2005)[4]. The results obtained from this project, Airline Delay Predictions using Supervised Machine Learning, it can help to better understand the phenomenon and up to a very large extent.

In 2013, it was estimated that approx. 36% of flights were delayed by more than five minutes in Europe, 32% of flights delayed by more than 15 minutes in the US, and 16% off lights were cancelled or sobered delays greater than 30-40minutes in Brazil[1]. Hence, it indicates how important this indicator is and how it acts no matter how wide the scale of airline meshes exists.

Furthermore, coming to the Indian scenario, in 2017,according to the reports by the Directorate General of Civil Aviation (DGCA), between January and April, close to 5.12lakh domestic passengers in India faced issues due to airline companies denying boarding, as well as flight cancellations and delays [2]. Airline companies had to pay the passengers compensations of over Rs. 25 core for various inconveniences during the first four months of this year. Hence, the prediction analysis retrieved from this project can contribute in the form of a prototype in helping to identify operational variables that contribute to delays in any country scenario.

(Allan et al., 2001)[3]analysed delays at NYC Airports from September `96 through August`00, with the aim offending out some major causes of delay occurred during the first year of an Integrated Terminal Weather System (orbits) use and delays occurred with ITWS in operation that were “avoidable” if in case weather conditions would have been improved. The methodology used in the study has considered some major causes of delays (for example, convective weather inside and outside the terminal area, and high winds), and these causes were generally neglected impervious studies of capacity constrained airports such as Newark International Airport (EWR). The research concluded that the usual methods of assessing delays only inters of Instrument Meteorological Conditions (IMC),Visual Meteorological Conditions (VMC) and the respective airport capacities is way more simplified than required for determining the type of air traffic management investments that in the best ways reduces the possible "avoidable” delays.

(Hansen and Hsiao, 2005)[5] analysed the rise in flight delay in the United States domestic system by estimating an econometric model of average daily delay that combines the effects of arrival queuing, terminal weather conditions,seasonal effects, and secular effects (such as a half).The results suggested that even after controlling these factors altogether, the delays decreased gradually from 2000through mid-2003, but the trend reversed drastically thereafter.

(Rosen, 2002)[6] measured the rate of change in flight timings that resulted due to infrastructure-constant changes in passenger demand. Results indicated that as the ratio of demand to fix infrastructure increased, the

delays increased proportionately, which resulted in proper decrease in yearaverage flight times by approx. 7 minutes after the rapid decrease in the fall'01. The flight time differences between the airlines in the data sample were small, though the United Airlines had lesser average flight times in the winter quarter than America West, which is considered even smaller airline.

Over the past couple of years, various analytical models and simulation methods have been used to analyse flight delay, including deterministic queuing models, neural networks, econometric models etc. Although it is evident that the analysis on delays carried is either on macroscopic or microscopic data over a period of couple of days and this has happened because of the huge data of flights every day. Hence, the predictions led to less accurate results or relapse in the trend among the results.

So here, obtaining the airline on-time performance data set from the U.S. DOT Bureau of Transportation Statistics (BTS) website, and the linear and polynomial regression models to be used along with regularization technique in machine learning is far better to identify the delay pattern. In this project, studies on airport delay and individual airlines delay behavior analysis arecarried out, using linear regression model, polynomial regression models, and regularization. The performances of the models are tested using various metrics, e.g., CVMethod, MSE/RMSE Scores, etc.

This project will be able to complete several objectives like the statistical description of airlines, temporal variability of delays, the relation of delays with the origin airports, estimating geographically the flights from each airport, etc., along with the main prediction analysis.

III.REGRESSION ANALYSIS MODELLING

A. Overview of the Dataset

The dataset has been taken from a reliable online available government agency website that provides the air traffic delay statistics in the United States. The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. BTS compiles daily data for the benefit of the customers or for any data analysts.

Table 1: Description of the attributes involved in the dataset

Attributes	Descriptions of Attributes
YEAR, MONTH, DAY, DAY_OF_WEEK	dates of the flight
AIRLINES	It is the IATA Code

	toidentify unique airlines
ORIGIN_AIRPORT and DESTINATION_AIRPORT	Code attributed by IATA to identify the airports
SCHEDULED_DEPARTURE and SCHEDULED_ARRIVAL	scheduled times of take-off and landing
DEPARTURE_TIME and ARRIVAL_TIME	real times at which take-off and landing took place
DEPARTURE_DELAY and ARRIVAL_DELAY	difference (in minutes) between planned and real times
DISTANCE	

B. Data Exploration and Visualization

Data cleaning is the critical initial step in evaluating the dataset for final analysis. With the enormous amount of data available, databases are prone to have noisy, missing and inconsistent data. The data in this project is obtained formats source, which has varying kinds of 31 variables involved, and may not be compatible with the format in which we require the data to use in Python. Data Cleaning helps in removing noisy data, and removing inconsistencies.

	IATA_CODE	AIRLINE
0	UA	United Airlines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways Inc.
5	OO	Skywest Airlines Inc.
6	AS	Alaskha Airlines Inc.
7	NK	Spirit Airlines
8	WN	Southwest Airlines Co
9	DL	Delta Airlines Inc.
10	EV	Altantic Southeast Airlines.
11	HA	Hawaiian Airlines Inc.
12	MQ	American Eagle Airlines Inc.
13	VX	Virgin America

Fig.2: All the airlines in the dataset associated with particular IATA carrier codes.

```
flightsinfo_modified = flightsinfo1.dropna(subset = ['AIR_SYSTEM_DELAY','SECURITY_DELAY','AIRLINE_DELAY','LATE_AIRCRAFT_DELAY','WEATHER_DELAY'])
flightsinfo_modified = flightsinfo_modified.drop(['YEAR','MONTH','DAY','DAY_OF_WEEK','TAIL_NUMBER','SCHEDULED_DEPARTURE','DEPARTURE_TIME','SCHEDULED_TIME',
                'SCHEDULED_ARRIVAL','ARRIVAL_TIME','DIVERTED','CANCELLED','CANCELLATION_REASON','FLIGHT_NUMBER','WHEELS_OFF',
                'WHEELS_ON','AIR_TIME'],axis = 1)
```

```
flightsinfo_modified.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15408 entries, 27 to 49997
Data columns (total 14 columns):
#   Column              Non-Null Count  Dtype
---  ------              --------------  -----
0   AIRLINE             15408 non-null  object
1   ORIGIN_AIRPORT      15408 non-null  object
2   DESTINATION_AIRPORT 15408 non-null  object
3   DEPARTURE_DELAY     15408 non-null  float64
4   TAXI_OUT            15408 non-null  float64
5   ELAPSED_TIME        15408 non-null  float64
6   DISTANCE            15408 non-null  int64
7   TAXI_IN             15408 non-null  float64
8   ARRIVAL_DELAY       15408 non-null  float64
9   AIR_SYSTEM_DELAY    15408 non-null  float64
10  SECURITY_DELAY      15408 non-null  float64
11  AIRLINE_DELAY       15408 non-null  float64
12  LATE_AIRCRAFT_DELAY 15408 non-null  float64
13  WEATHER_DELAY       15408 non-null  float64
dtypes: float64(10), int64(1), object(3)
memory usage: 1.8+ MB
```

```
flight_Delays = flightsinfo_modified
```

```
flightsinfo1 = flightsinfo1.drop(['CANCELLATION_REASON','AIR_SYSTEM_DELAY','SECURITY_DELAY','AIRLINE_DELAY',
                'LATE_AIRCRAFT_DELAY','WEATHER_DELAY'],axis = 1)
```

Fig.3:Dropping unused attributes and values

Data Visualization

It is a way to graphically represent the data especially when the data is numeral. It allows us to recognize new patterns and relations with the data. Here in this project we use bar chart , pie chart and scatter plot for

International Journal of Advanced Development in Science and Technology
Volume : 3 Issue : 03 April 2021 www.ijadst.com e-ISSN: 2582-1059

identifying trends and clusters.AHeat map is visually represented to find the correlation ship between different attributes.

The dataset contains small percentage of missing values for certain columns and these values are dropped as they make up a very small portion of the dataset.

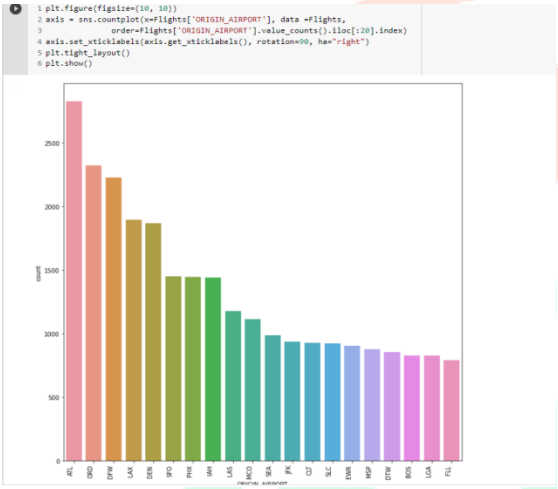

Fig.4:A plot between Origin City and the Number of flights

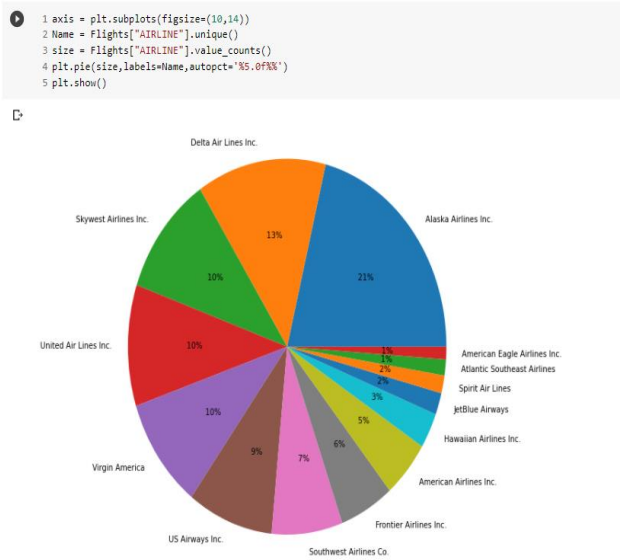

Fig.5:Pie chart with % of flights per company

The above diagram depicts the percentage of flights travelled by each airlines.

Fig.6:Correlation between variables using Heat map

C. Feature Selection

Not all the columns are not really needed for the prediction of delays, so the unneeded features are dropped and few features necessary were kept for the prediction purposes. Some of the features are in the form of a string these were converted to number values using Label encoder and were assigned number beginning with zero, this is done so that the dataset is more machine learning friendly as models tend to not perform well with strings as features. To predict the flight delay an extra feature is to be added in the data called 'Is Delayed' which is actually a binary value indicating

0 = Flight Not delayed

1 = Flight Delayed

International Journal of Advanced Development in Science and Technology
Volume : 3 Issue : 03 April 2021 www.ijadst.com e-ISSN: 2582-1059

```
[266] 1 from sklearn.preprocessing import StandardScaler
      2 from sklearn.model_selection import train_test_split
      3 from sklearn.preprocessing import LabelEncoder

[267] 1 le = LabelEncoder()
```

Applying Label encoder to convert text values to numbers

```
[268] 1 Flights1['AIRLINE']= le.fit_transform(Flights1['AIRLINE'])
      2 Flights1['ORIGIN_AIRPORT'] = le.fit_transform(Flights1['ORIGIN_AIRPORT'])
      3 Flights1['DESTINATION_AIRPORT'] = le.fit_transform(Flights1['DESTINATION_AIRPORT'])

[270] 1 Flights1 = Flights1.drop(['Scheduled_Departure','Scheduled_Arrival','Actual_Arrival','Actual_Departure','ELAPSED_TIME','TAXI_IN','TAXI_OUT',
```

Creating a new feature which has value only as 0 or 1 depending on it it is delayed or not

```
[271] 1 Flights1['Is_Delayed'] = np.where(Flights1['ARRIVAL_DELAY']<=0, 0,1)

[272] 1 Flights1
```

	AIRLINE	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DISTANCE	DEPARTURE_DELAY	ARRIVAL_DELAY	SCHEDULED_TIME	AIR_TIME	Is_Delayed
0	0	15	265	1448	-11.0	-22.0	205	169.0	0
1	0	15	265	1448	-4.0	-14.0	204	173.0	0
2	0	15	265	1448	-15.0	-35.0	218	170.0	0
3	0	15	265	1448	-11.0	-12.0	200	176.0	0
4	0	15	265	1448	-8.0	-14.0	205	179.0	0
...	...	...	...	...	...	...	...	...	...
48688	3	187	229	69	7.0	10.0	35	24.0	1
48689	3	187	229	69	-4.0	-10.0	35	21.0	0
48690	3	187	229	69	57.0	51.0	35	17.0	1
48691	3	187	229	69	176.0	175.0	35	21.0	1

Fig.7:Feature selection

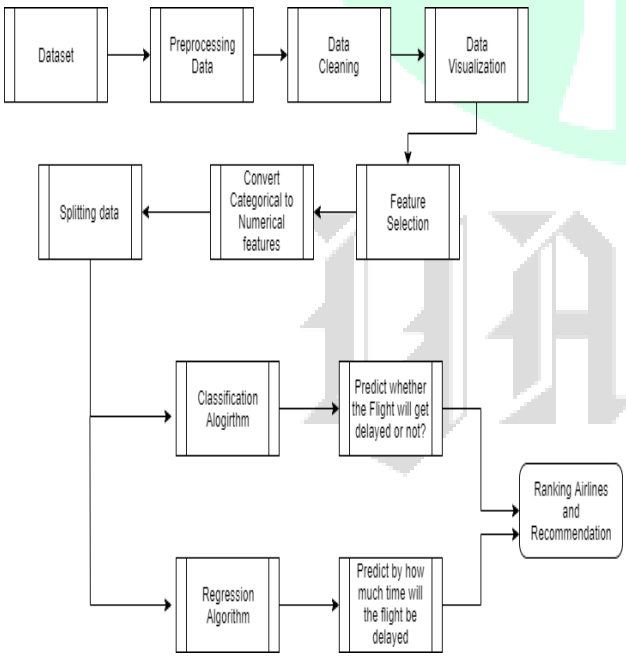

Fig.8: System Architecture Diagram

IV. PERFORMANCE METRIC

Cross Validation Technique and K-Fold Technique

Cross Validation is a very important technique for assessing the performance of machine learning models. It enables us in knowing how a machine learning model would generalize to an independent data set. The model dataset is divided into three sets: Training, test, and validation. The entire set is divided into K-folds or subsets, which is basically applying the K-fold technique, one of the ways of Cross Validation. Then, the K-1 folds are sent for training and the learning is done on it, then themodel's generalization is checked on the test set, which contains just the remaining one fold; and this process godson till the last fold.

MSE

The Mean Squared Error (MSE) is a measure of how close a Fitted line is to the real data points. For every data point on the line, we take the distance vertically from the real point to the corresponding Y value on the curve fitted(which is the error), and square the value. The next step is to carry out the summation of all the squared error values corresponding to all the data points, and, in the case of linear fit, the value we get is divided by the total number of observations minus 2. The squaring is to avoid negative values cancelling the positive values. The quality of the model is assessed by the

Mean Squared Error score we get, the smaller the value, the closer the fit is to the real data and the accurate the machine learning model. MSE can be calculated using the below formula,

$$\frac{1}{N}\sum_{i=1}^{n}(\text{actual values} - \text{predicted values})^2$$

where n=Total no. of attributes or points taken into account.

RMSE

Root Mean Squared Error (RMSE) is another quality that we calculate to measure the accuracy of a model. It is equal to the square root of the mean square error. It is considered as one of the most easily interpreted statistics, as it has the same units as the quantity plotted on the ordinate, which isthe y-axis.RMSE can be calculated using the below formula,

RMSE=sort(sum((predicted_i-actual_i)^2)/total predictions)

V. CONCLUSION

This project and the analysis retrieved are useful not only for passengers point of view, but for every decision maker in the aviation industry. Apart from the financial losses incurred by the industry, flight delay also portray a negative reputation of the airlines, and decreases their reliability. It causes various sustainability issues, for example, increase in fuel consumption and gas emissions. The analysis carried here not only predicts delays based on the previous available data, but also give statistical description of airlines, their rankings based on their on-time performance, and delays with respect to time, showing the peak hours of delay. This project can be used as a prototype by any aviation authority for their benefit, in the Indian Scenario too, it can work as inefficient model or a proper prototype to study delay analysis, based on the real dataset provided. This project has encompassed and showed the importance of Regression Analysis in Machine Learning, Data Mining concepts for efficient data cleaning, Cross Validation technique and Regularization in ML for making proper models and itspredictive analysis.

REFERENCES

[1] ANAC.Agˆenciа Nacional de Aviac¸˜aoCivil.Technical report,http://www.anac.gov.br/, 2017.

[2] Indian Economic Times https://economictimes.indiatimes.com/

[3] MIT, Lexington, Massachusetts, Allan, S.S., S.G. Gaddy, and J.E.Evans, (2001) Delay Causality and Reduction at the New York CityAirports Using Terminal Weather Information.

[4] Wu, C. (2005), Inherent delays and operational reliability of airlineschedules,

Journal of Air Transport Management Volume 11, Issue4(273-282)

[5] Hansen, M., and C. Y. Hsiao (2005), Going South? An EconometricAnalysis of US Airline Flight Delays from 2000 to 2004, Presented atthe 84th Annual Meeting of the Transportation Research Board(TRB), Washington D.C.'05.

[6] Rosen, A. (2002), Flight Delays on US Airlines: The Impact ofCongestion Externalities in Hub and Spoke Networks, Department ofEconomics, Stanford University

[7] Programming in Python 3: A Complete Introduction to the PythonLanguage , By Mark Summerfield.

[8] Prabakaran, N., and R. JagadeeshKannan. "Sustainable life-span ofWSN nodes using participatory devices in pervasive environment."Microsystem Technologies 23.3 (2017): 651-657.