

Predict Flight Delays using Supervised Machine Learning Technique



Project Guide: Mr.S.Vinu M.E.,(Ph.D)
Assistant Professor
Department of Computer Science and Engineering

Presented By:
Saiprasaad K (312317104146)
Rohith Vigneshwar D (312317104143)

Abstract

- In this project, we analyzed the various factors responsible for flight delays and applied machine learning models 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.
- We also added a recommendation feature in which given a source and destination, we would list flights which are recommended to travel. Also we can know the percentage of Delay and Not delayed of a particular journey by entering Source , Destination and the name of Airlines.

Project Objectives

1. Given certain features, we would predict that whether a flight would be delayed or not.
2. Also we can predict, how far the delay would be in minutes.
3. We can also predict the percentage of delayed and non-delayed by giving the source, Destination and name of the Airlines.
4. Another feature was added to rank airlines and to recommend which airlines to prefer for a journey.



Existing System

The existing system deals with Time series forecasting of data which sees the data year wise and country wise. The existing system uses a approach of articulation point which is assumed to be having the greatest delay. And using arima modelling, the system has seen the change in the delay over years, concluding the delay of which airport is the delay most likely to occur in the country.

Disadvantages:

The existing system deals with year-wise timeseries forecasting and uses the approach of clustered networks, where the delay is forecasted only by airport wise. They are harder to explain and to interpret coefficients requires stationary series with constant autocorrelation

Proposed System

In the proposed system, we use classification algorithms like KNN Classification, XGB Classifier , Random Forest Classifier and Decision Tree Classifier to predict whether the flight would be delayed or not. Out of these algorithms XGB Classifier gives best accuracy using this algorithm we can predict whether the flight arriving in the destination will have delay or not. Also regression techniques are performed to predict arrival delays of how much delay is to occur. A feature of ranking and recommendation are also added for better travel.

Advantages:

- Takes several features into account
- Has predictions which helps both for airline management system and passengers
- Recommendations are added to suggest flights for their journey

EXISTING SYSTEM	PROPOSED SYSTEM
Deals with Time-Series forecasting of data only by year	Deals with supervised machine learning solution involving various parameters
Uses a approach of articulation point which is assumed to be having the greatest delay.	No such approach of articulation point.
Country wise forecasting.	Origin , Destination and Airlines wise forecasting
Works only on small datasets	Works also on large datasets
Just shows the trend of delay over years.	Shows delayed or not , along with how much delay is to occur.
No recommendations of flights for journey is made.	Flights are ranked with score based on their delays and speed , recommendations are done.
Do not show the ratio of flights delayed and non delayed.	With the predicted value a pie chart is represented to know how far the flight is to get delayed.

System Requirements

Hardware Requirements:

- a) Processor -I3,I5,I7
- b) RAM -4GB
- c) Hard Disk -250GB

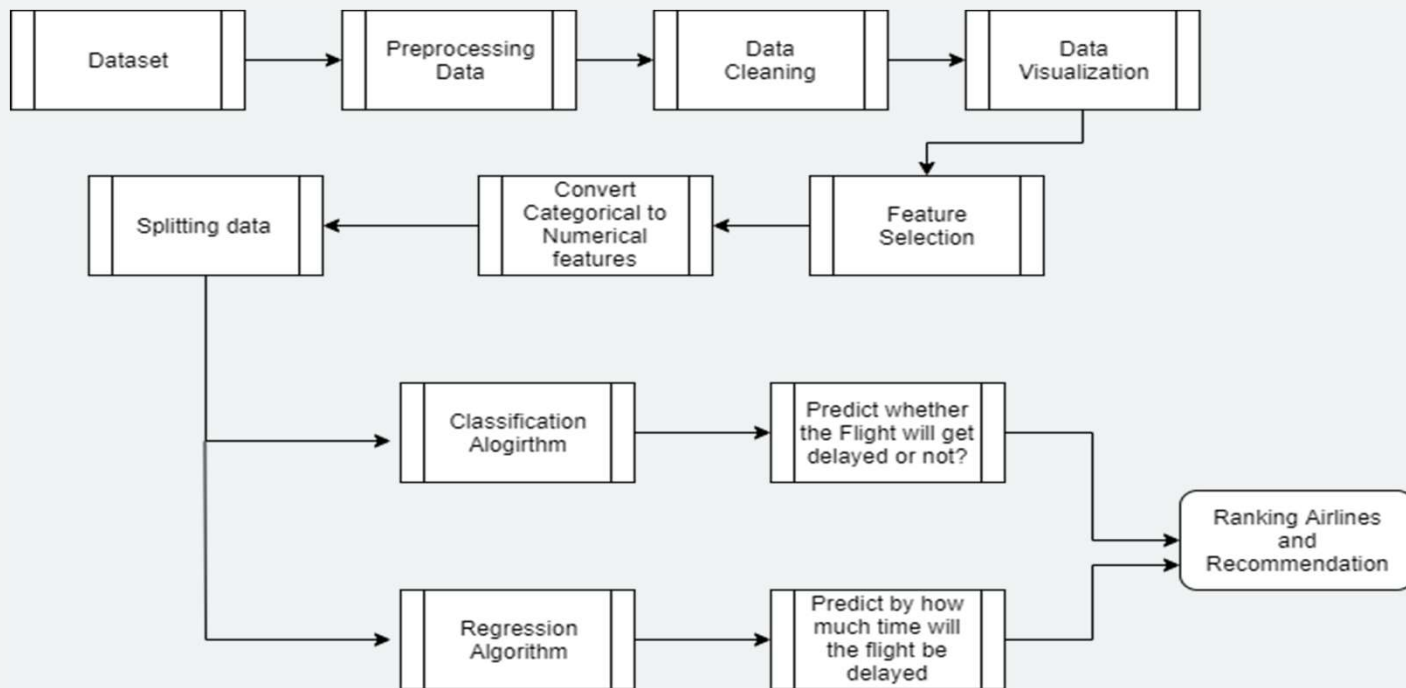
Software Requirements:

- a) Python version 3.6 and above
- b) Google colab software or Jupyter Notebook

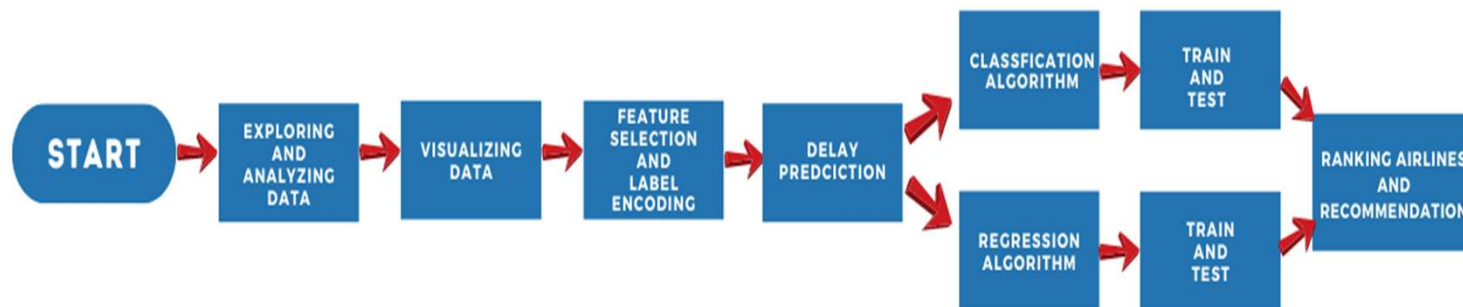
Packages

- ❖ Pandas
- ❖ Numpy
- ❖ Matplotlib
- ❖ Xgboost
- ❖ Sklearn
- ❖ Flask

System Design



Architecture Diagram



Flow Chart Diagram

System Design

- ❑ This project deals with predicting the flight delays using supervised machine learning techniques. The dataset which is obtained from Kaggle is first being imported and preprocessed. The processed data is cleaned to handle missing values and is visually represented in the form of graphs, charts and heatmaps.
- ❑ The data with selected features is now fit into classification and regression algorithm for training. The trained data is then tested to get accuracy of models of both classification and regression techniques .The model with best accuracy is used for prediction. The model is further used to get recommendation to fly from a origin to destination, where airlines which is most preferable is represented as output.
- ❑ Also one can see the ratio of delayed to not-delayed given their origin, destination along with airlines.

Modules

- ☐ Exploring And Analyzing Data
- ☐ Feature Selection And Label Encoding
- ☐ Predicting Flight Delays
- ☐ Ranking And Recommendation Of Flights
- ☐ Deploying Using Django

Exploring And Analyzing Data

Importing Dataset:

The dataset was located from Kaggle, this dataset was collected from U.S department of transportation. The dataset tracks the performance of domestic flights within the united states.

Data Preprocessing

Handling missing values – The dataset contains small percentage of missing values for certain columns like Departure delay, taxi out and so on. These rows containing missing values are dropped as they make up a very small portion of the dataset

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is done to uncover relationships between different variables. A correlation matrix is drawn to measure how strong a relationship is between two variables, using which variables in strong relation to predicting delay is analyzed.

Visualizing Through Graphs

Several graphs and charts are drawn to present the data in a pictorial or graphical format and to analyze data visually. The data in a graphical format allows them to identify new trends and patterns easily. Fig 1 shows the scatter plot of airlines to no of minutes delayed. Overall proportion or percentage of airline companies in the dataset is represented as pie chart in Fig 2



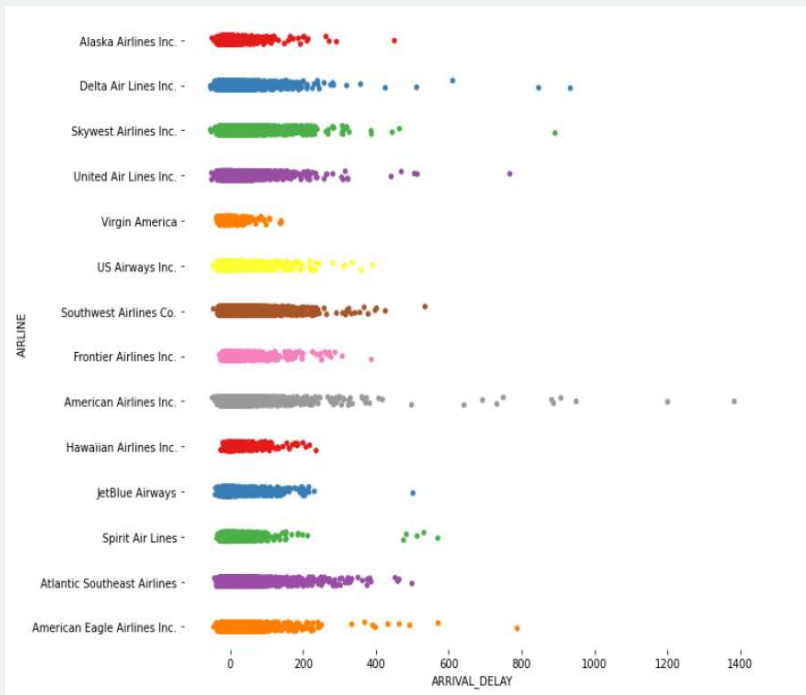


Fig 1 Scatter Plot

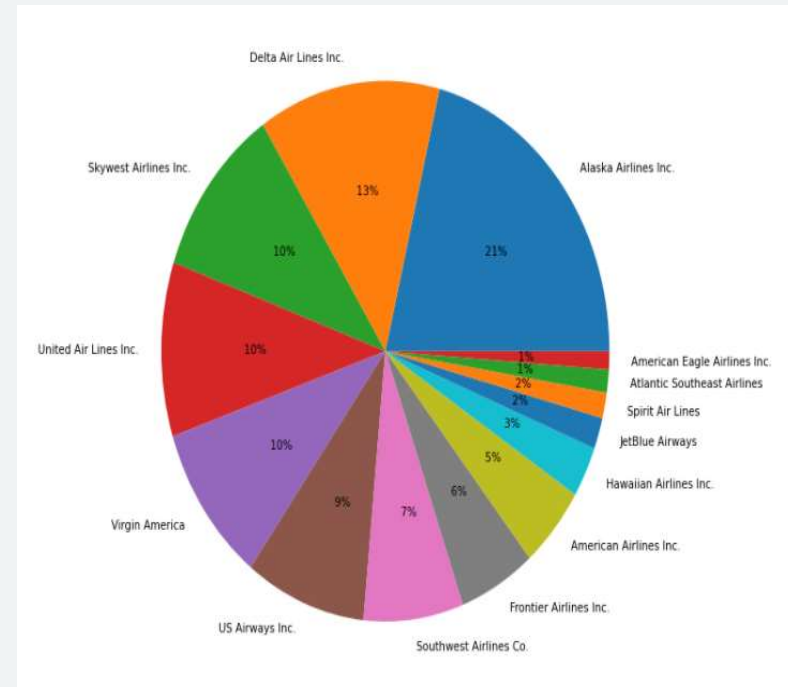


Fig 2 Pie Chart

Feature Selection And Label Encoding

Feature Selection :

The goal of this step is to find best set of features that allows one to build useful models of studied phenomena. So here certain features like Airline, Origin Airport, Destination Airport, Distance, Departure Delay, Scheduled Time, Airtime, Taxi Out are selected for prediction of flight delay

Label Encoding

Prediction can be done only when values in the data are in numbers. There are some categorial features which are to converted into number format. This step is called Label Encoding



Table 1 Label Encoding

For Example, in this Table 1 the airlines name which is in the form of a string is being label encoded into number format which are then used for training

Predicting Flight Delays

- ❑ We have several methods of prediction which some user can get the result of flight delay and by prediction with some more extra features the airline management can get the flight delayed prediction result.
- ❑ Several algorithms like KNN, Decision Tree , Random Forest and XGBoost out of which XGBoost turns to have the best accuracy among all. Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy. At any instant t , the model outcomes are weighed based on the outcomes of previous instant $t-1$.

Predicting Whether or not Flight is Delayed

- ❑ In this module, the primary aim is to predict flight is delayed (1) or not delayed (0) using features like origin, destination , airlines name, distance, airtime, departure-delay, taxi-out, scheduled time.
- ❑ As these features are know only by the flight management system, this is useful for the airline management system to know how far these features can be altered to avoid delays.
- ❑ Classification algorithm is used to predict this delayed or not-delayed status.

Predicting percentage of delay and non-delayed status

- ❑ In this module, the aim is to predict proportion of flight delayed and not delayed only by getting details of the origin, destination , airlines name.
- ❑ The origin airport name, destination airport name and airlines name are got as input from the user, based on which other features are extracted from the dataset and finally the user gets a pie-chart of percentage of delayed to non delayed status.
- ❑ This prediction also uses the XGBoost Classification model for prediction of flight delay.

Predicting delay value in minutes

- ❑ In this module, the aim is to predict by how many minutes the flight will be delayed . This can be beneficial to airline management since they have all the features needed to prediction.
- ❑ This uses regression techniques which explains the changes in criterions in relation to changes in select predictors.
- ❑ Using XGBoost regression technique ,with the help of selected features delay of how much minutes the flight is going to get delayed is predicted.

Ranking and Recommendation

- ❑ Users more often face a problem which is choosing the airlines for their journey. This To rank and recommend flights to the user we use certain features like flight delay, cancelled to operated ratio , speed which is calculated from airtime and distance.
- ❑ For a origin and destination , the airways are grouped by their company and for each airline company the score is calculated accordingly. Then the airlines are sorted according to their scores and recommendations are given to the user, based on which the user can select the airlines to book for his journey.

- ❑ In this Table 2 the source given by user gives origin airport as Hartsfield–Jackson Atlanta International Airport and destination airport as Dallas/Fort Worth International Airport, so for the given source and destination the score is calculated and is displayed for the users in descending order of the scores for his comfortness

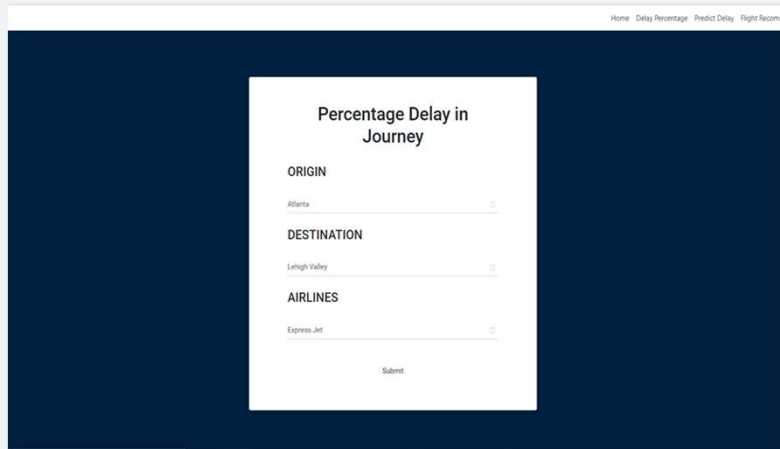
```
DESC_AIRLINE
Delta Air Lines Inc.      251.840153
American Airlines Inc.   163.534996
Spirit Air Lines          34.328611
Atlantic Southeast Airlines 6.222104
Name: SCORE, dtype: float64
```

Table 2 Recommendations

Deployment using Django

- ❑ A webpage using Django is created for user and airline transport system to get delays of flights and recommendations .In flight delay prediction page, user enters the source airport, destination airport and airlines name, so the percentage of delayed to not-delayed is returned to the user as a pie-chart to get an understanding about delays.
- ❑ In another page the airline transport system can get the minutes of delay of airline giving details which are needed to predict the delays. Finally in another page of the website the user gives his source and destination of the journey by which he can get recommendations of the flights which he can prefer to fly for his journey.

Sample Screenshots



Percentage Delay in Journey

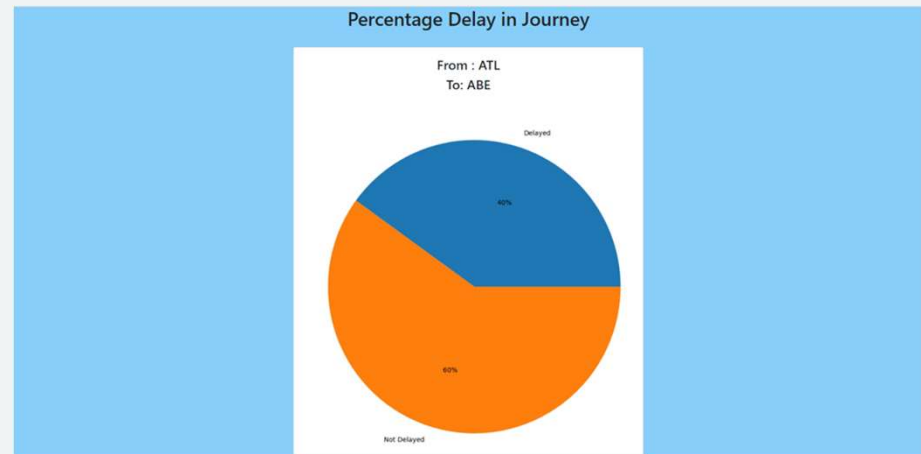
ORIGIN
Atlanta

DESTINATION
Lehigh Valley

AIRLINES
Expressto Jet

Submit

Home Delay Percentage Predict Delay Flight Recomm



Delay and Non Delay Percentage Result Webpage

Sample Screenshots

Home Delay Percentage Predict Delay Flight Recomm

Ranking and Recommendation of Airlines

ORIGIN

Los Angeles

DESTINATION

Boston

Submit

Recommendation of Flights

From : LAX To: BOS

	Airlines	Score
0	American Airlines Inc.	217.754323
1	JetBlue Airways	207.786208
2	Virgin America	143.021080
3	Delta Air Lines Inc.	73.187517
4	United Air Lines Inc.	34.200781

Ranking and Recommendation Webpage

REFERENCE

1. J.E. Aronson, "A Survey Of Dynamic Network Flows," Ann. Opns. Res. 20, 1–66
2. A. Barrat, M. Barthe' Lemy, R. Pastor-Satorras, And A. Vespignani, "The Architecture Of Complex Weighted Networks", Pnas March 16, 2004 101 (11) 3747-3752;
3. Dimitris Bertsimas , Sarah Stock Patterson "The Traffic Flow Management Rerouting Problem In Air Traffic Control: A Dynamic Network Flow Approach"
<https://pubsonline.informs.org/doi/abs/10.1287/trsc.34.3.239.12300>
4. L. R. Ford And D. R. Fulkerson, Flows In Networks, Princeton University Press, Princeton, New Jersey, 1958.
5. L. Macdonald, "Collaborative Decision Making In Aviation," J. Air Traffic Control 40, 12–17 (1998).

6. Pablo Fleurquin , Jose Javier Ramasco And Víctor M. Eguíluz,“ Systemic Delay Propagation In The Us Airport Network” , 10.1038/Srep01159
7. Stanislav Tarasevych; Ivan Ostroumov, “A Light Statistical Method Of Air Traffic Delays Prediction,”. 2020 Ieee 2nd International Conference On System Analysis & Intelligent Computing (Saic)
8. Shuai Li , Yuelel Xu, Mingming Zhu, Shiping Ma, And Hong Tang, “Remote Sensing Airport Detection Based On End-To-End Deep Transferable Convolutional Neural Networks”,Ieee Geoscience And Remote Sensing Letters (Volume: 16, Issue: 10, Oct. 2019)
9. P. Vranas, D. Bertsimas, And A. R. Odoni, “Dynamic Ground-Holding Policies For A Network Of Airports,” Transp. Sci. 28, 275–291 (1994b)
10. P. Vranas, D. Bertsimas, And A. R. Odoni, “The Multiairport Ground-Holding Problem In Air Traffic Control,” Opns. Res. 42, 249–261 (1994a).

Thank You