

Flight Delay Prediction For Aviation Industry Using Machine Learning

By

H.Haarish Infant Raj

B.Hariharan

N.Arun Kumar

D.Eswaran

❖ INTRODUCTION

Using a machine learning model, we can predict flight arrival delays. over the last twenty years, air travel has been increasingly preferred among travelers, mainly because of its speed and in some cases comfort. This has led to phenomenal growth in air traffic and on the ground. An increase in air traffic growth has also resulted in massive levels of aircraft delays on the ground and in the air. These delays are responsible for large economic and environmental losses. According to, taxi-out operations are responsible for 4,000 tons of hydrocarbons, 8,000 tons of nitrogen oxides and 45,000 tons of carbon monoxide emissions in the United States in 2007. Moreover, the economic impact of flight delays for domestic flights in the US is estimated to be more than \$19 Billion per year to the airlines and over \$41 Billion per year to the national economy In response to growing concerns of fuel emissions and their negative impact on health, there is active research in the aviation industry for finding techniques to predict flight delays accurately in order to optimize flight operations and minimize delays.

Using a machine learning model, we can predict flight arrival delays. The input to our algorithm is rows of feature vector like departure date, departure delay, distance between the two airports, scheduled arrival time etc. We then use decision tree classifier to predict if the flight arrival will be delayed or not. A flight is delayed when difference between scheduled and actual arrival times is greater than 15 minutes. Furthermore, we compare decision tree classifier with logistic regression and a simple neural network for various figures of merit. Finally, it will be integrated to web based application

❖ **PROJECT FLOW:**

- **Define Problem / Problem Understanding**
 - *Specify the business problem*
 - *Business requirements*
 - *Literature Survey*
 - *Social or Business Impact.*
- **Data Collection & Preparation**
 - *Collect the dataset*
 - *Data Preparation*
- **Exploratory Data Analysis**
 - *Descriptive statistical*
 - *Visual Analysis*
- **Model Building**
 - *Training the model in multiple algorithms*
 - *Testing the model*
- **Performance Testing & Hyperparameter Tuning**
 - *Testing model with multiple evaluation metrics*
 - *Comparing model accuracy before & after applying hyperparameter tuning*
- **Model Deployment**
 - *Save the best model*
 - *Integrate with Web Framework*
- **Project Demonstration & Documentation**
 - *Record explanation Video for project end to end solution*
 - *Project Documentation-Step by step project development procedure*

Problem Definition & Understanding

❖ **Business Problem Specification**

Over the past 20 years, air travel has become more popular among travelers. The increase in the growth of air transport has resulted in flight delays both on the ground and in the air. These delays cause huge economic and environmental losses. Precise delays to optimize flight operations and minimize delays. Using a machine learning model, we can predict flight arrival delays. Details like departure date, departure delay, the distance between two airports, scheduled arrival time, etc. Then we use decision tree classification to predict whether the flight arrival will be delayed or not.

❖ **Business Requirements**

- **Prediction Accuracy:** It should predict the delays accurately with the provided data

- Scalability: The system should be able to handle large volumes of data and be scalable
- User-friendly interface: The system should have a user-friendly interface that allows users to easily access and interpret the information provided
- Flexibility: The system should be flexible enough to incorporate changes in airline schedules, airport conditions, and other factors that may affect flight delays
- Security and privacy measures to protect sensitive data
- Reporting and analytics to support decision-making

❖ Literature Survey

Flight delays can cause significant inconvenience to passengers and result in high costs for airlines. The ability to accurately predict flight delays is therefore an important problem in the aviation industry. In recent years, machine learning algorithms have been increasingly used for flight delay prediction.

One of the key challenges in flight delay prediction is the availability of data. In addition to flight information, weather data, and airport data, other sources of information such as social media and news articles have also been used to improve prediction accuracy.

Various machine learning algorithms have been used for flight delay prediction, including decision trees, random forests, support vector machines, and neural networks.

In addition to predicting flight delays, some research has also focused on predicting the length of delays. For example, Huang et al. (2018) proposed a model that predicts both the probability of a delay and the expected length of the delay

❖ Social or Business Impact

Flight Delay Prediction Can be useful for passengers for improving their traveling experience, by informing the accurate Schedule of flight departures and arrivals and delays.

With these Predictions, passengers can plan their travel without missing flights or boarding earlier. This can reduce the stress related to travel for the passengers. And This can also help to reduce the number of Flight Cancellations. By Predicting the Flight Delay, the airlines can improve their Safety Measures and Passengers can plan and board according to the Prediction.

By the way, In Business for airlines, Flight Delay Prediction provides an improvement of their workings Condition and reduces the Amount. By Analyzing the Conditions which are being reasons for flight delays, the airlines Should take essential measures to solve the impacts of the flight delays. By solving the impacts, it can improve the Exact time of departures and arrivals of flights and will be Easier for Passengers. Also, the Flight Delay prediction helps the staff work, Reduces the cost, and Saves time.

Finally, the Flight Delay Prediction Using Machine Learning provides more benefits on Social Impact, by Providing Safety Measures, Increasing Customer Reviews, and making it easier for Airline Working Operations.

Data Collection & Preparation

❖ Data Collection

Data collection or data gathering is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes

In this project, we have used .csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset.

As the dataset is downloaded. Let us read and understand the data properly with the help of some visualization techniques and some analyzing techniques.

We develop a system that predicts a delay in flight departure based on certain parameters. We train our model for forecasting using various attributes of a particular flight, such as arrival performances, flight summaries, origin/destination,

➤ *Import the necessary libraries as shown in the image*

```
import pandas as pd
import numpy as np
import pickle
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import sklearn

from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import RandomizedSearchCV
import imblearn

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, f1_score
```

➤ **Read the data set**

Our dataset format might be in .csv, excel files, .txt, json, etc. We can read the dataset with the help of pandas.

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

In pandas, we have a function called **read_csv()** to read the dataset. As a parameter, we have to give the directory of the CSV file.

❖ **Data Preparation**

As we have understood how the data is, let's pre-process the collected data.

The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

Handling missing values **Handling categorical data**

Let's find the shape of our dataset first. To find the shape of our data, the `df.shape` method is used. To find the data type, **`df.info()`** function is used.

For checking the null values, `df.isnull()` function is used. To sum those null values we use **`sum()`** function

We will fill in the missing values in the numeric data type using the mean value of that particular column and the categorical data type using the most repeated value.

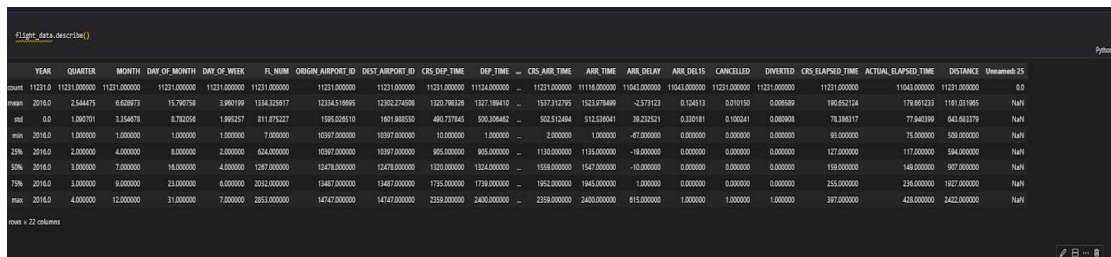
As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding.

To convert the categorical features into numerical features we use encoding techniques. There are several techniques but in our project, we are using manual encoding with the help of list comprehension.

Exploratory Data Analysis

❖ Descriptive statistical

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.



	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_NUM	ORIGIN_AIRPORT_ID	DEST_AIRPORT_ID	CRS_DEP_TIME	DEP_TIME	CRS_ARR_TIME	ARR_TIME	ARR_DELAY	ARR_DEL15	CANCELLED	DIVERTED	CRS_ELAPSED_TIME	ACTUAL_ELAPSED_TIME	DISTANCE	Unnamed: 25
count	11231.0	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	11231.000000	0.0
mean	2016.0	2.544475	6.620773	15.790758	3.890199	1334.325817	1234.518895	1232.274528	1320.788328	1327.186410	1537.312785	1523.578499	-2.573123	0.124813	0.070150	0.006989	190.852124	179.881233	1181.851965	NaN
std	0.0	1.306701	3.354878	8.732058	1.895257	811.875227	1395.628510	1601.888550	480.727845	500.306462	502.512484	512.538041	38.232521	0.320181	0.105241	0.080008	78.338317	77.840239	642.881379	NaN
min	2016.0	1.000000	1.000000	1.000000	1.000000	10397.000000	10397.000000	10397.000000	10.000000	1.000000	2.000000	1.000000	-47.000000	0.000000	0.000000	0.000000	93.000000	75.000000	528.000000	NaN
25%	2016.0	2.000000	4.000000	6.000000	2.000000	624.000000	10397.000000	10397.000000	905.000000	905.000000	1138.000000	1135.000000	-19.000000	0.000000	0.000000	0.000000	127.000000	117.000000	594.000000	NaN
50%	2016.0	3.000000	7.000000	16.000000	4.000000	1287.000000	12478.000000	12478.000000	1320.000000	1324.000000	1558.000000	1547.000000	-10.000000	0.000000	0.000000	0.000000	159.000000	148.000000	907.000000	NaN
75%	2016.0	3.000000	9.000000	23.000000	6.000000	2032.000000	13487.000000	13487.000000	1735.000000	1738.000000	1962.000000	1945.000000	1.000000	0.000000	0.000000	0.000000	255.000000	236.000000	1927.000000	NaN
max	2016.0	4.000000	12.000000	31.000000	7.000000	2051.000000	14747.000000	14747.000000	2359.000000	2400.000000	2358.000000	2403.000000	615.000000	1.000000	1.000000	1.000000	397.000000	428.000000	2422.000000	NaN

❖ Visual analysis

Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

1. Univariate Analysis

- ✓ univariate analysis is understanding the data with a single feature

2. Bivariate Analysis

- ✓ . Bivariate analysis is a statistical method examining how two different things are related

3. multivariate Analysis

- ✓ multivariate analysis is to find the relation between multiple features.

Model Building

❖ **Training the model in multiple algorithms**

Now our data is cleaned and it's time to build the model. We can train our data on different algorithms. For this project we are applying four classification algorithms. The best model is saved based on its performance.

Decision tree model

A function named `decisionTree` is created and train and test data are passed as the parameters. Inside the function, `DecisionTreeClassifier` algorithm is initialized and training data is passed to the model with the `.fit()` function. Test data is predicted with `.predict()` function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.

We are going to use `x_train` and `y_train` obtained above in `train_test_split` section to train our **Decision Tree Classifier** model. We're using the `fit` method and passing the parameters

Random Forest model

A function named `random Forest` is created and train and test data are passed as the parameters. Inside the function, `Random Forest Classifier` algorithm is initialized and training data is passed to the model with `.fit()` function. Test data is predicted with `.predict()` function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.

ANN model

Building and training an Artificial Neural Network (ANN) using the Keras library with TensorFlow as the backend. The ANN is initialized as an instance of the `Sequential` class, which is a linear stack of layers. Then, the input layer and two hidden layers are added to the model using the `Dense` class, where the number of units and activation function are specified. The output layer is also added using the `Dense` class with a sigmoid activation function. The model is then compiled with the Adam optimizer, binary cross-entropy loss function, and accuracy metric. Finally, the model is fit to the training data with a batch size of 100, 20% validation split, and 100 epochs.

❖ **Testing the model**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the

functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Performance Testing & Hyperparameter Tuning

❖ Testing model with multiple evaluation metrics

Multiple evaluation metrics means evaluating the model's performance on a test set using different performance measures. This can provide a more comprehensive understanding of the model's strengths and weaknesses. We are using evaluation metrics for classification tasks including accuracy, precision, recall, support and F1-score.

❖ Comparing model accuracy before & after applying hyperparameter tuning

Valuating performance of the model From sklearn, cross_val_score is used to evaluate the score of the model. On the parameters, we have given rf (model name), x, y, cv (as 5 folds). Our model is performing well. So, we are saving the model by pickle.dump().

Model Deployment

❖ Save the best model

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its weights and configuration. This can be useful in avoiding the need to retrain the model every time it is needed and also to be able to use it in the future.

```
import pickle
```

```
pickle.dump(rfc,open('flightRFCmodel.pkl','wb'))
```

❖ Integrate with Web Framework

In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.

This section has the following tasks

- **Building HTML Pages**
- **Building server side script**
- **Run the web application**

❖ Building html pages

In this project we have created two html pages

Index.html
Result.html

Index.html

For
example:

```
<head>
<title>Registration Page</title>

<link rel="stylesheet" href="{{url_for('static', filename='/main.css')}}" type="text/css">
</head>
<body>
  <h1>Predict your flight delay chances</h1>

  <form action="/predict" method="POST">
    <label>Flight Number</label>
    <input type="number" name="flightnumber" id="flightnumber"
placeholder="Flight Number">

    <label>Month of Travel</label>
    <input type="number" name="month" id="month" placeholder="Month of
Travel">

    <label>Travel day of Month</label>
```

Result.html

```
<!DOCTYPE html>
<html lang="en" >
<head>
  <meta charset="UTF-8">
  <title>Result Page</title>
  <link rel="stylesheet" href="{{ url_for('static', filename='/main.css') }}">
</head><body>
  <h1>Flight Delay Prediction</h1> <p>Your flight : {{flightnumber}} <br> Status :
  {{prediction}} delayed </p></body>
</html>
```

❖ Building server side script

Import the libraries

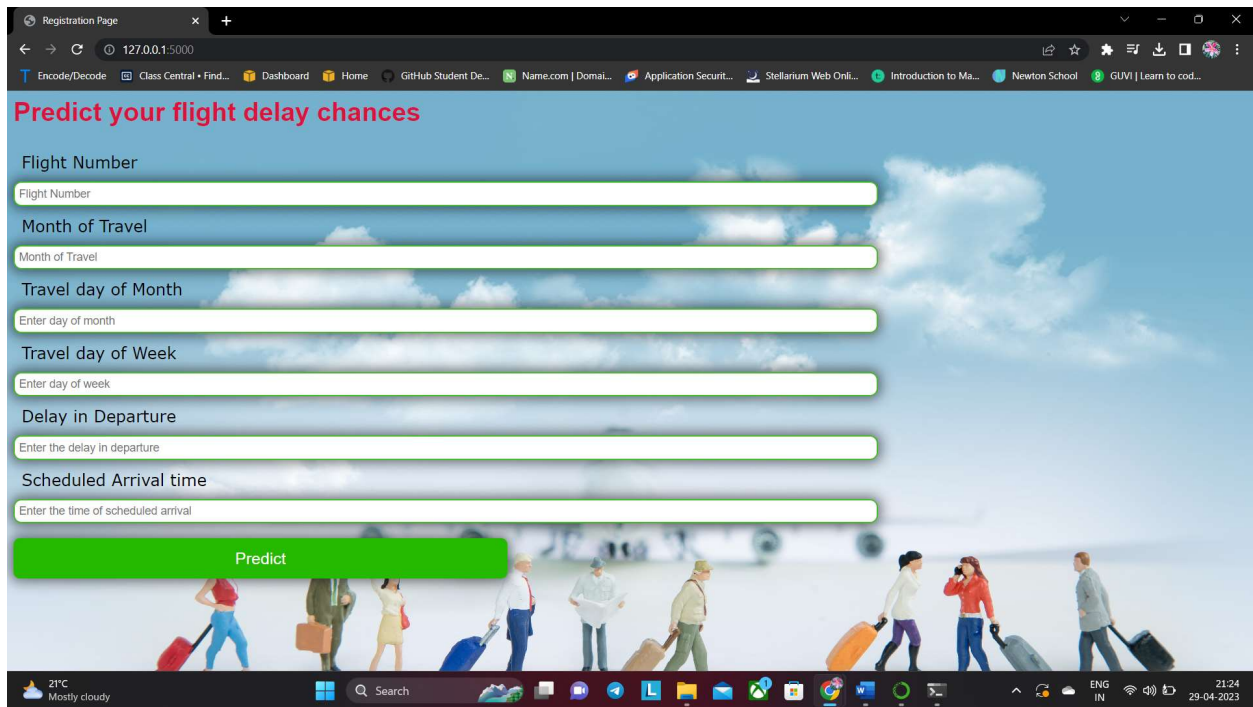
```
from flask import Flask,request,render_template
import pickle
from sklearn.preprocessing import StandardScaler
```

Load the saved model. Importing the flask module in the project is mandatory. An object of Flask class is our WSGI application. Flask constructor takes the name of the current module (__name__) as argument.

Main Function:

```
if __name__=='__main__':
    app.run(debug=True)
```

Output



The screenshot shows a web browser window with the title "Registration Page" and the address bar displaying "127.0.0.1:5000". The browser's address bar and tabs are visible at the top. The main content area has a blue header with the text "Predict your flight delay chances" in red. Below the header is a form with the following fields:

- Flight Number:
- Month of Travel:
- Travel day of Month:
- Travel day of Week:
- Delay in Departure:
- Scheduled Arrival time:

Below the form is a green button labeled "Predict". The background of the form is a blurred image of an airport tarmac with people walking and a plane in the distance. The Windows taskbar is visible at the bottom of the browser window, showing the date and time as 21:24 on 29-04-2023.

Flight Delay Prediction

Your flight : 1799

Status : Won't get delayed