Developing A Flight Delay, Prediction Model Using, Machine Learning

Sparsha S, Sakthidhari B, Sarada S P, Sivatharani V, Sathyapriya R

Student, Guide, Assistant Professor, Knowledge Institute of Technology, Salem.

Abstract: Due to its convenience, luxury, and speed, air travel has become more popular than other modes of transportation. However, because of a number of issues, including erratic weather forecasts, engine problems, and others, there have been major delays as a result of the huge demand for it. These delays cause significant monetary and environmental damage. To enhance flight operations and reduce delays, the model's primary goal is to develop a machine learning model to predict flight delays. Our algorithm would take as inputs the departure date, the separation between the two airports, the scheduled arrival, etc. Additionally, for various figures of merit, we contrast the decision tree classifier with logistic regression and a straightforward neural network. The Random Forest consists of various decision trees that select the suitable attribute for a node starting at the root and separate the data into subsets based on the selected attribute.

Keywords: Random forest, Predicting the flight delay, Map plotting, Referring near by hotels and transports.

1. INTRODUCTION

In the modern world, air traffic control, cargo airlines, and passenger airlines are the three main components of any transportation system. Nations all around the world have attempted to develop different methods over time to enhance the aeroplane transportation system. This has significantly altered how airlines operate. Modern travellers occasionally have trouble from flight delays. About 20% of airline flights are cancelled or delayed annually, costing travellers more than $20 billion in lost time and money.

Airport capacity is sometimes described as being determined by the average aircraft delay. One of the most common problems in the globe is flight delay. It's quite difficult to justify a delay. There are a few uncommon causes of flight delays, such as runway work or heavy traffic, but inclement weather seems to be the most frequent one...

1.1. PURPOSE

To optimise flight operations and reduce delays, the model's primary goal is to estimate flight delays accurately. Flight arrival delays can be predicted using a machine learning algorithm. Our study focused primarily on forecasting flight delays for a certain airport over a specific time frame. To determine the importance of each variable, we employed a regression model; in this case, we used Logistic Regression along with techniques like Support Vector Machine, Decision Tree, and Random Forest. In addition, we used the One-Hot-Encoder method to encode categorical variables.

2. PROBLEM STATEMENT

Air flight has been beneficial because it shortens travel distances. However, because of a number of issues, including erratic weather forecasts, engine problems, and others, there have been major delays as a result of the huge demand for it. These delays cause significant monetary and environmental damage. To enhance flight operations and reduce delays, the model's primary goal is to develop a machine learning model to predict flight delays. Our algorithm would take as inputs the departure date, the separation between the two airports, the scheduled arrival, etc. Additionally, for various figures of merit, we contrast the decision tree classifier with logistic regression and a straightforward neural network.

3. PROPOSED SOLUTION

We want to boost prediction accuracy with the system we've proposed. For prediction, we are utilizing the UK dataset, which offers statistics on arrival and departure, including wheels-off time, departure delay, and taxi-out time per airport. It also includes actual departure time, scheduled departure time, scheduled elapsed time, and scheduled elapsed time. To analyze the result more accurately, we will test and train the data. We will provide them with other means of transportation so they can get where they are going.

We employ machine learning techniques including logistic regression, decision tree classifier, and random forest, which also aid in increasing accuracy and performance. The user can choose from the provided locations on a tab to learn about the direct flights offered from the source on the plotted map. It provides a distinct vision of the locations from the original locations. Users have the opportunity to look for a hotel and restaurant to eat at while waiting.
4. DATA FLOW DIAGRAM (DFD)

Figure 4.1 Data Flow Diagram

The application gets input from the user like flight number, origin, destination, scheduled arrival time, scheduled departure time, days of month, days of week and actual departure time. Then with the help of these inputs, the model predicts whether the flight is delayed or not.

5. SOLUTION ARCHITECTURE

6. TECHNICAL ARCHITECTURE

Figure 6.1 Technical Architecture

7. FEATURES

7.1 FEATURE 1 – ANALYSIS

Flight delays are gradually getting worse, which causes airline businesses to have more trouble making money and to lose customers. Supervised machine learning models were used to forecast aircraft delays in order to tackle this issue. For the prediction, information from flights leaving from JFK airport over the course of a year was gathered in a data set.

Various algorithms are used in the process of predicting the flight delay such as Logistic Regression, K-Nearest Neighbor, Support Vector Machine and Random Forest were trained and tested to complete the binary classification of flight delays. The
evaluation of algorithms was fulfilled by comparing the values of two measures: accuracy and precision. These measures were weighted to adjust the imbalance of the selected data set.

7.2 FEATURE 2 – Prediction
Algorithms that allow a computer to analyse data, find probable patterns, and make predictions are referred to as machine learning. Learning algorithms can provide information about how challenging learning is in various contexts.

The Random Forest consists of various decision trees that select the suitable attribute for a node starting at the root and separate the data into subsets based on the selected attribute. It makes use of the bagging method and individual models of decision trees. The trained data are divided into random subsets, and each has its decision tree. The data are given parallel to all trees in the forest, and the class that most trees predicted has the new data.

7.3 FEATURE 3 – Map Plotting
Geopy is a Python library that makes it easy to locate the coordinates of addresses, cities, countries, and landmarks using various geocoding services. With Geopy, you can convert a human-readable location into its corresponding latitude and longitude coordinates, which can be used to plot markers and lines on a map.

First, you need to install the required libraries, namely geopy and pandas. You can install them using pip. Once you have installed the libraries, you can proceed with the code. We create a dataframe of airports and their coordinates, and then create a map using the folium library. You start by creating a “map” object and specifying the initial location and zoom level of the map. We add markers for all airports and then add lines between them to represent direct flights.

7.4 FEATURE 4 – Alternatives
We are going to show a few alternate modes of transport to the customer where the customer or site user can search for alternatives. This is useful for the customer who is looking for a bus or car to travel. They can also look for a hotel or motel to stay in. Created a tab on an HTML page on a card to visualize the alternatives for transport. We are providing recommendations for transportation options for users. The available transportation or accommodation options. This can involve integrating with APIs from travel services such as Uber or Airbnb, or using data sources such as Google Maps to suggest nearby hotels or public transportation options.
8. ALGORITHM USED
8.1 RANDOM FOREST
A popular algorithm for classification and regression issues is the supervised machine learning technique known as random forest. It creates decision trees from various samples, relying on their majority for categorization and average for regression. In order to increase the dataset's predictive accuracy, the Random Forest classifier uses a variety of decision trees on different subsets of the input data.
- The issue of overfitting is resolved because the result is based on majority vote or average.
- It works well even when the data has null or missing values.
- Each decision tree that is produced independently of the others exhibits parallelization.
- It preserves diversity as all traits are not taken into account while building each decision tree, albeit this is not always the case. It is highly stable as the average responses provided by a large number of trees are used.
- It is resistant to the dimensionality curse. The feature space is decreased because each tree does not take into account all the properties.
- Since 30% of the data will always be test data, we don't need to separate the data into train and test.

9. PERFORMANCE METRICS
Performance metrics track a company's actions, behaviour, and overall performance. It measures the necessary data within a range of data that is in the form of measurements. The primary goal to verify is performance, which is measured by this. A classification problem known as the Random Forest identifies the data based on the data's decision trees. The model is trained using the provided dataset. The model gains knowledge from the dataset before branching to several decision trees. The confusion matrix, which takes the form of a table and contains the binary classifiers of the outcome from model prediction, is a crucial evaluation metric for the classification problem. Which explains how the trained model performed.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Metrics</td>
<td>Classification Model: Confusion Matrix – [1840,0,0,407] Accuracy Score- 85% Classification Report – 87%</td>
</tr>
<tr>
<td>2.</td>
<td>Tune the Model</td>
<td>Hyper parameter Tuning – 88% Validation Method – Randomized Search CV</td>
</tr>
</tbody>
</table>

10. CONCLUSION
Because of its speed and, in certain circumstances, comfort, air travel has become more and more popular among travelers during the past 20 years. As a result, in this research we attempt to address such issues by predicting whether an aircraft would be delayed or not using flight datasets. The outcome demonstrates that decision trees have the highest accuracy values, at 100%. Additional features like dining establishments, lodging options, and transportation have been added.

11. FUTURE ENHANCEMENT
In the future, we will track the flight's location and it will be displayed on the map. Notifications will be sent to the user for continuous updating about the flight's status. The flight details will be displayed. If the user gives the origin and destination, the available flights will be displayed. Although weather conditions are the major reasons for flight delays, other unprecedented events such as major calamities, natural or man-made, can cause major delays in flights.
12. SAMPLE OUTPUT

Figure 12.1 Predicting whether flight is delayed or not

Figure 12.2 Result for the flight delayed

Fig. No. 12.3 Result for flight on time

Fig. No. 12.4 Alternatives
Fig. No. 12.5 Map Plotting

13. REFERENCES
7. Yash Guleria, Qing Cai, Sameer Alam “A Multi-Agent Approach for Reactionary Delay Prediction of Flights” 2019