Predicting Flight Delays

Anuneet Anand
2018022
anuneet18022@iiitd.ac.in

Mohnish Agrawal
2018053
mohnish18053@iiitd.ac.in

Rhythm Patel
2018083
rhythmkumar18083@iiitd.ac.in

Abstract

The growth of the aviation sector has made flight delays more common across the world. They cause inconvenience to the 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, Logistic Regression, Decision Tree and Naive Bayes to predict whether a given flight would be delayed or not. The XGBoost Classifier performed exceptionally well, giving an accuracy of 0.88 and an AUC of 0.93. ***GitHub:*** *github.com/rhythm-patel/Predicting-Flight-Delays*

1. Introduction

There has been a remarkable expansion in commercial aviation in the past decade, with more people preferring air travel for a fast and comfortable journey. However, flight delays have become quite common across the world with the growth of the aviation sector. Besides inconvenience to the travellers, flight delays have a negative impact on the economy. The airline companies incur substantial monetary losses and observe a fall in their reputation if their flights are delayed often. The unforeseen delays also have a cascading effect on various other sectors. According to a report by the Joint Economic Committee of United States Congress, the total cost of flight delays to the US economy was over $40 billion with $19 billion to the airlines, $12 billion to the passengers and around $10 billion to other industries. The delayed flights also pose certain environmental concerns. Delayed flights consumed an additional 740 million units of jet fuel and released over 7 million metric tonnes of additional Carbon Dioxide [1]. Thus, the prediction of flight delays is a crucial task.

This project aims at analysing factors responsible for flight delays and designing a machine learning model to predict them. We classify a flight as 'Delayed' if the arrival delay of the flight is more than 3 minutes. Prediction of delays can help customer choose best flight for the journey and help airlines to identify flaws in their organisation.

2. Literature Review

A significant amount of research work has been done in the field of air-traffic control and commercial aviation. Many researchers have made attempts to solve this problem and have presented different machine learning approaches.

The researchers at School of Computer, Wuhan Vocational College of Software and Engineering, developed a multiple linear regression model and compared its performance with other models such as Naive-Bayes and C4.5. The delay value predicted was classified into two classes with the flights having a delay of more than 30 minutes being classified as 'Delayed'. The model achieved 79% accuracy, and it was observed that weather was not a significant feature for classification, except in extreme conditions. [2]

The researchers at Nova Information Management School, Portugal used various data-mining techniques along with a Knowledge Discovery Database approach and then trained Decision Tree, Logistic Regression and Multi-Layer Perceptron models. The SMOTE technique gave better results than the Under-sampling technique. The features like distance and month were found to insignificant. The Multi-Layer Perceptron gave an accuracy of 85% and emerged as the best model for prediction. [3]

The researchers at State University of New York at Binghamton, New York and Defense Sciences Institute, Turkish Military Academy, Ankara, introduced DMP-ANN model for prediction of defects by applying it to the system of air traffic control. Traditional ANN had a hard time handling nominal variables and had to convert to 1-of-N encoding, which reduced the performance. Results showed that this new ANN outperformed traditional ANN in terms of error (RMSE) as well as time required for training the data. However, since the number of layers increases with connections, complexity becomes a limitation. The model was used to predict flight delays at JFK airport and gave remarkable results. [4]

3. Dataset

We used the 2015 Flight Delays and Cancellations Dataset, collected and published by U.S. DOT's Bureau of Transportation Statistics, for training and testing our prediction models. The data from the three files were merged into a single dataset. The merged dataset consisted of over 5 million samples and had information about airlines, airports, flight schedules, flight routes, delays and cancellation reasons. Most of the columns had an excellent filling factor, except those related to cancellation reasons.

3.1. Cleaning and Feature Extraction

We selected January month's data for our use, keeping the computational complexity in mind. Since we were concerned with only flight delay prediction, we dropped the columns related to cancellation. Some other irrelevant and redundant columns like Tail Number, Airport Name, Airline Name and Wheels On were dropped. The Date-Time format was corrected, and rows with NaN values were removed. We also removed outliers which had an arrival delay of more than 500 minutes. A dataset of dimensions 456685 rows × 11 columns was obtained with the following features.

- **Categorical Features:** Airline, Origin, Destination
- **Numerical Features:** Distance, Taxi Out, Departure Delay, Day of Week, Arrival Delay
- **Date/Time Features:** Date, Scheduled Departure, Scheduled Arrival

3.2. Exploratory Data Analysis

We identified the busiest airports and air traffic shares of different airlines and visualised the mean delays in airlines on different days of the week. It was observed that most flights were delayed on Sunday and Monday. Alaska Airlines Inc. and Delta Air Lines Inc. were the best performing airlines whereas Frontier Airlines Inc. and American Eagle Airlines Inc. were often delayed.

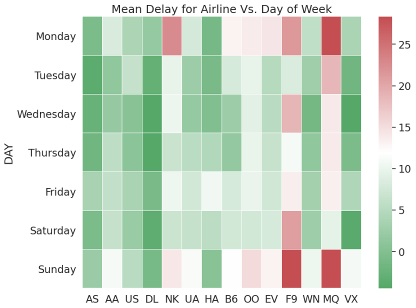

Figure 1. Heat Map of Mean Airline Delays Vs. Days

It was found that the mean delays of flight dropped with an increase in route distance. However, flights travelling distances over 3000Km showed a spike in delays.

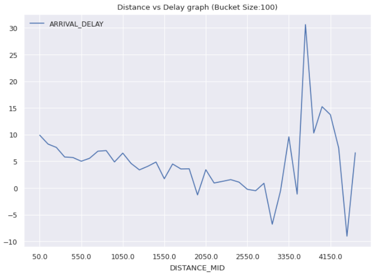

Figure 2. Distance Vs. Delay

We also conducted a statistical study of different numerical features in the dataset and derived important insights about their distributions. It was found that almost 68% of the flights had no delay or a delay of less than 3 minutes.

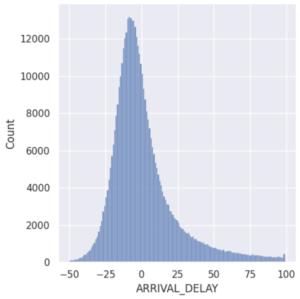

Figure 3. Distribution Of Arrival Delay

3.3. Pre-Processing

We had to pre-process the data before training and evaluating different models on it. The Categorical features had to be converted into numerical forms so that they can be interpreted by the machine learning models.

- One-Hot Encoding was used to handle the Airline feature [14 possible values].
- Label Encoding was used to handle Origin and Destination features [300+ possible values].
- Date and Time were expressed using sine and cosine values of their individual attributes to incorporate their cyclic nature.
- Keeping 3 minutes as a threshold, the samples with an arrival delay less than 3 minutes were assigned class 0 whereas the remaining samples were assigned class 1.

The features were scaled using Standard Scaler. We obtained the numpy arrays X and Y with dimensions $(456855, 32)$ and $(456855,)$ respectively.

4. Methodology

To validate the performance of our models, we performed a train-val-test split of 70:10:20. We trained Random Forest Classifier, XGBoost Classifier, Naive Bayes Classifier, Decision Tree and Logistic Regression, from the Sklearn library, on the training set. Grid Search CV was used to find optimal hyper-parameters for the models. Appropriate graphs and metrics were generated for the analysis and performance of the different models were compared.

Gaussian Naive Bayes Classifier

Gaussian Naive Bayes Classifier is a supervised probabilistic machine learning model for continuous data, based on the Bayes Theorem. It assumes independence among the input features and calculates posterior probabilities of different classes to make predictions. The Gaussian conditional probability is as given in (1).

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (1)$$

Logistic Regression

Logistic Regression uses the sigmoid function to transform the output into a probability value which can be used to classify the output. The hypothesis function for logistic regression is given in (2).

$$h(\theta) = \frac{1}{1 + e^{-\theta^T x}} \quad (2)$$

Decision Tree

A decision tree has a structure similar to flowchart in which each internal node is a test on a feature and each leaf node is a class label. Information gain is used to select the feature to split on at each step while building the tree. The information gain can be calculated using entropy loss (3) or gini impurity index (4). Decision tree is easy to understand and can handle both numerical and categorical data.

$$I_H = -\Sigma p_i log_2(p_i) \quad (3)$$

$$I_G = 1 - \Sigma p_i^2 \quad (4)$$

XGBoost Classifier

XGBoost algorithm is an ensemble learning algorithm which uses Boosting. XGBoost Classifier creates a number of Decision Trees like Random Forest Classifier. However, unlike Random Forest, the trees are trained sequentially such that each new model corrects the errors of the previous one.

Random Forest Classifier

Random Forest algorithm is an ensemble learning algorithm which uses Bagging. The Random Forest Classifier creates a number of Decision Trees which are trained on bootstrap samples of training set by randomly choosing a set of attributes for each split. Each Decision Tree in the forest independently predicts the class of a test sample. The votes from all the trees in the forest are aggregated to decide the class of the test sample.

5. Results

Gaussian Naive Bayes Classifier

Gaussian Naive Bayes Classifier was used and an accuracy 0.820 was obtained. This was a considerably good result to start with.

Logistic Regression

Logistic Regression with default parameters was also tried on the given data and an accuracy of 0.864 was achieved.

Decision Tree

Decision Tree with default parameters was tried on the given data and an accuracy of 0.806 was achieved.

XGBoost Classifier

The XGBoost Classifier gave a base accuracy of 0.881. It was further tuned by testing different values for Hyper-parameters class_weight, learning_rate, max_depth and n_estimators using Grid-Search CV.

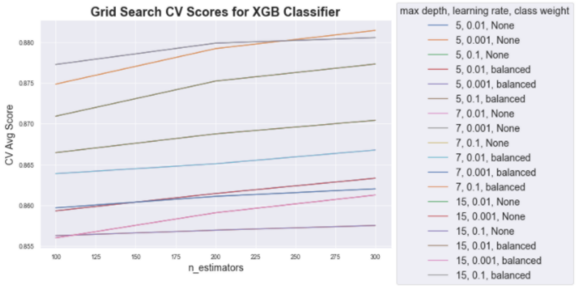

Figure 4. Tuning XGB Classifier

The best accuracy obtained after tuning was 0.884.

Optimal Hyper-Parameters

- class_weight: None
- learning_rate: 0.1
- max_depth: 7
- n_estimators: 300

Classification Report	Predicted class 0				.	Predicted class 1				
	LR	GNB	DT	XGB	RF	LR	GNB	DT	XGB	RF
True Class 0	54439	51967	48808	54878	54719	3372	5844	9003	2933	3092
True Class 1	8913	10430	8360	8231	8338	23734	22217	24287	24416	24259
Precision	0.86	0.83	0.85	0.87	0.86	0.88	0.79	0.72	0.89	0.88
Recall	0.94	0.89	0.84	0.95	0.94	0.73	0.68	0.74	0.75	0.74
F1-score	0.90	0.85	0.84	0.90	0.89	0.79	0.73	0.73	0.81	0.80

Random Forest Classifier

The Random Forest Classifier gave a base accuracy of 0.873. It was further tuned by testing different values for Hyper-parameters class_weight, criterion and n_estimators using Grid-Search CV.

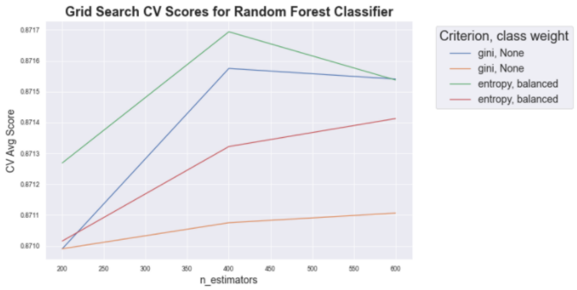

Figure 5. Tuning Random Forest Classifier

The best accuracy obtained after tuning was 0.875.

Optimal Hyper-Parameters

- class_weight: None
- criterion: entropy
- n_estimators: 400

5.1. Analysis

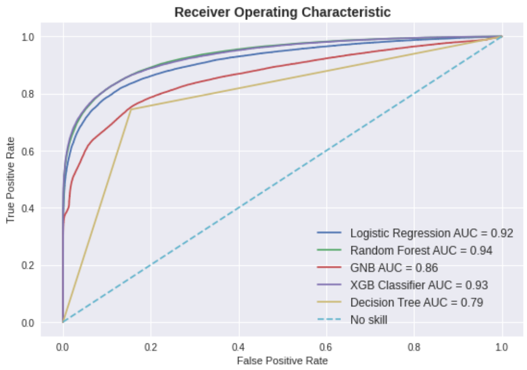

Figure 6. Receiver Operating Curves for Random Forest Classifier, Decision Tree, XGBoost Classifier, Naive Bayes Classifier and Logistic Regression

From the Receiver Operating Curves curves given in Figure 6, we note that Area Under Curve score was highest for XGBoost Classifier and Random Forest Classifier and lowest for Decision Tree.

The classification report table shows the combined Confusion Matrix for the different classifiers along with Precision, Recall and F1-score. It helps us to analyse the different models on various parameters like precision, recall and F1-score and find the optimal model for our problem statement.

Random Forest Classifier and XGBoost Classifier reported the best values across all judging parameters.

For our problem statement, we need a model with a high value of Recall so that we can inform about flight delays and the concerned professionals and people can adjust accordingly. Thus, XGBoost with a recall value of 0.75 seems to be the best choice among the five classifiers.

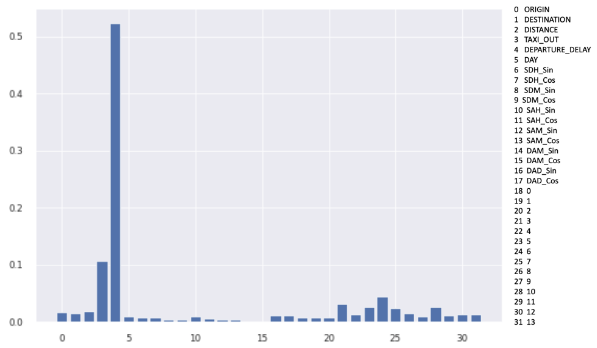

Figure 7. Feature Importance

We calculated feature importance from our XGBoost model and observed some interesting insights. Most important features for predicting flight delays were found to be Departure Delay and Taxi-Out. We observe that distance from origin to destination did not contribute much to the delay.

6. Conclusion

We can conclude that Random Forest Classifier and XGBoost Classifier performed significantly well in predicting flight delays. XGBoost Classifier had a slight edge over Random Forest Classifier and fitted the problem statement better with an accuracy score of 0.88 and AUC score of 0.93.

We hope that our work would contribute to the society and be utilised by the concerned parties. For future prospects, we aim to design an application for the same.

6.1. Member Contribution

- **Anuneet:** Literature survey & data collection, Training different models, Collecting and merging datasets, Statistical analysis, Exploratory Data Analysis, Training different types of models, Code Formatting, Report writing.
- **Mohnish:** Literature survey & data collection, Training different types of models to find the best one for our project, Plotting graphs for data analysis, Optimizing the model through hyper-parameter tuning, Report writing.
- **Rhythm:** Literature survey & data collection, Preprocessing the data, Testing & evaluating different models, Training different types of models, Fine-tuning them, Gaining insights with visualizations, Report writing.

6.2. Learning

In this project, we studied the various factors responsible for flight delays and gained some interesting insights by conducting Exploratory Data Analysis on the dataset. We were successful in pre-processing the dataset and applying machine learning models like Random Forest Classifier, Decision Tree, XGBoost Classifier, Gaussian Naive Bayes Classifier and Logistic Regression to our problem. The project allowed us to go beyond the scope of regular teachings of the classes and learn more about various machine learning algorithms. We developed analytical skills to study the data and come up with appropriate machine learning algorithms which would fit well on the data.

References

[1] Joint Economic Commitee Majority Staff. Your flight has been delayed again. Technical report, Tech. rep, 2008.

[2] Yi Ding. Predicting flight delay based on multiple linear regression. In *IOP Conference Series: Earth and Environmental Science*, volume 81, pages 1–7, 2017.

[3] Roberto Henriques and Inês Feiteira. Predictive modelling: flight delays and associated factors, hartsfield–jackson atlanta international airport. *Procedia computer science*, 138:638–645, 2018.

[4] Sina Khanmohammadi, Salih Tutun, and Yunus Kucuk. A new multilevel input layer artificial neural network for predicting flight delays at jfk airport. *Procedia Computer Science*, 95:237–244, 2016.