Flight Delay Prediction and Analysis

Stanley Varghese
EMSE
Univeristy of Texas at Dallas
Richardson, Texas
sjv140030@utdallas.edu

Kaden Tran
EMSE
University of Texas at Dallas
Richardson, TX
stt032000@utdallas.edu

Sarah Pelosi
EMSE
University of Texas at Dallas
Richardson, TX
sjp200002@utdallas.edu

Jason Anthraper
EMSE
University of Texas at Dallas
Richardson, TX
jaa200001@utdallas.edu

***Abstract*—The commercial aviation industry plays a crucial role in enabling mobility across the United States. This report outlines the application of a machine learning classification technique to predict flight delay occurrences based on historically available data. This data includes airline, airport, and weather-related information along with corresponding instances of flight delay.**

Keywords—Classification, Supervised Machine Learning, Logistic Regression, Airline, Delay, United States

I. Introduction

The commercial aviation industry within the United States facilitates vital economic activity across industries. An early 2020 report from the Federal Aviation Administration (FAA) estimates that domestic aviation enables over 5 percent of gross domestic product in the U.S. and accounts for the employment of approximately 11 million individuals. This FAA report also outlined that the number of annual passenger flight onboardings across over 5,000 available public airports and 210,000 aircraft is nearing 1 billion annually [1]. Due to the interconnected nature of domestic aviation networks, the accurate identification of potential flight schedule interruptions enables mitigation arrangements which are essential to maintaining continuous operations. Many techniques have been explored within the field of machine learning seeking to gain valuable insights from existing data.

In this report, we outline the application of a supervised machine learning algorithm to predict flight disruptions based on available historical data consisting of airline, airport, and weather-related details along with corresponding instances of flight delays.

II. Background

A. Data Source

Data pertaining to flight departure delays for large domestic airlines is monitored and collected by the Bureau of Transportation Statistics, a component of the United States Department of Transportation (DOT), since mid-2003. Should a flight delay occur, commercial airlines report the cause of delay based on available classifications which include security, National Aviation System, severe weather, airline, and late aircraft arrival. Airline related delays encompass flight postponement resulting from conditions within an air carrier's realm of accountability.

Potential delay causes within the airline category include, but are not limited to, aircraft maintenance, fueling, and luggage loading. National Aviation System (NAS) related postponements result from adjustments made in response to air traffic control, flight traffic, weather, and airport procedures. The severe weather classification involves delays corresponding to forecasted or acutely occurring extreme weather events including hurricanes, tornadoes, and snowstorms. An important distinction regarding weather and how it is reported includes the fact that the severe weather label refers exclusively to circumstances that inhibit flying entirely, whereas all other postponements relating to weather conditions are reported within the NAS grouping. The late arrival category involves flight postponements resulting due to schedule issues from prior flights involving the same scheduled aircraft. Lastly, security broadly includes interruptions due to evacuations, reboarding, security scanning equipment failures, and unusually lengthy security screening area lines [2]. Variations in annual flight delay causes between years 2003 and 2016 are presented in Figure 1 below.

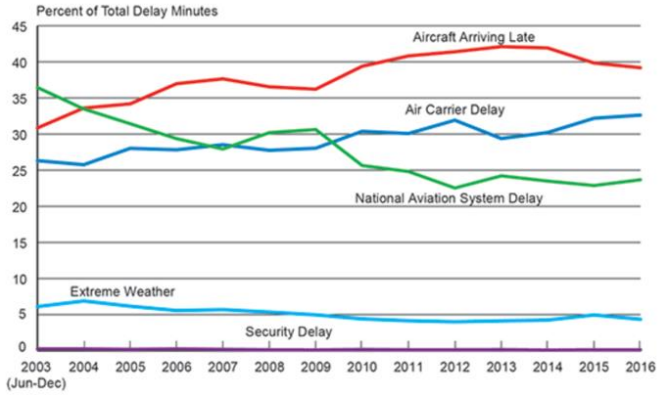

Fig. 1. Delay Cause Per Year [2]

The dataset utilized within this experiment was sourced from the Bureau of Transportation statistics, along with weather information obtained via the National Centers for Environmental Information. Airport weather related features available within the data set include total inches of daily precipitation and snowfall along with inches of snow present at ground level, maximum daily temperature, and maximum daily wind speed. From this dataset, we can see the percentage of flight delay is closed to twenty percent (Fig. 2).

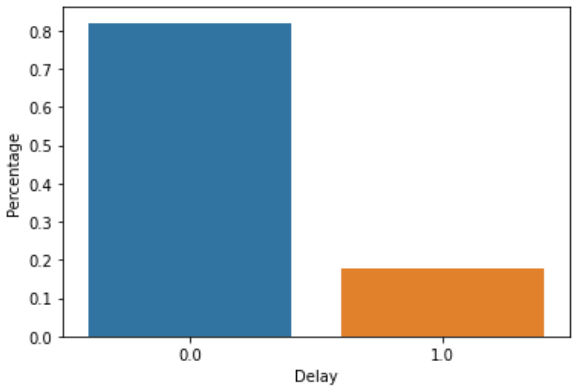

Fig. 2. Occurance of Flight Delays

B. *Preprocessing Steps*

Data preprocessing techniques were applied to the dataset before it was supplied to the developed logistic regression algorithm. Features containing a significant number of null entries were dropped from the dataset as this data might cause undesired results. A correlation matrix was generated via the seaborn visualization library to identify and remove features with high levels of correlation to one another. This relationship is also referred to as multicollinearity and negatively impacts the accuracy of logistic regression models. In such cases, only one feature among those highly correlated was selected to preserve the performance of our model.

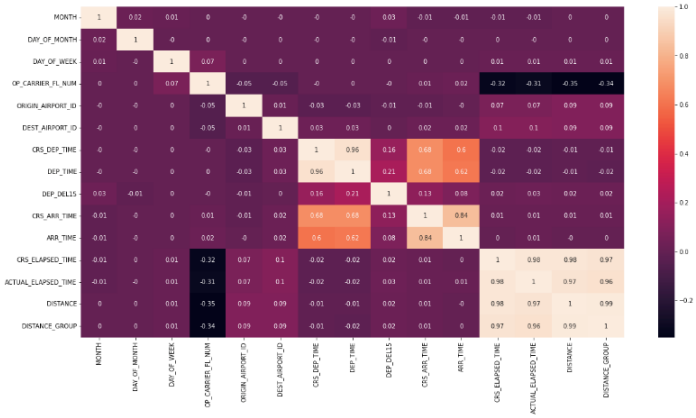

Fig. 3. Correlation Matrix of Utilized Features

III. Study of Technique

A. *Logistic Regression*

Logistic regression is a linear classification technique used when the dependent variable of a dataset is categorical. Some example use cases are the classification of cancer patients, or the probability and confidence of spam or valid email messages. The binary logistic regression model has the loss function:

$$L(\mathbf{w}; \mathbf{x}, y) := \log(1 + \exp(-y\mathbf{w}^T\mathbf{x})) \quad (1)$$

Given a new data point **x**, the model applies the logistic function:

$$f(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

where $z = \mathbf{w}^T\mathbf{x}$. The most common threshold is if $f(\mathbf{w}^T\mathbf{x}) > 0.5$, the outcome is positive; otherwise the outcome is negative [5].

When choosing the logistic regression model, we make a series of assumptions about the dataset:

1) The output variable is binary, meaning the output must fit into a classification of delayed (0) or on-time (1).

2) There is a linear relationship between gthe independent and dependent variables.

3) The data is normalized, meaning that extreme outliers are removed.

4) To avoid overfitting, highly correlated inputs are removed.

B. *Rating Parameters*

To help us analyze the success of the implemented logistic regression model, the output of the model is represented by a confusion matrix. A confusion matrix is a performance measurement for a machine learning classification problem where the output can be two or more classes [7]. The confusion matrix is useful for visualizing and comparing actual values and prediction values. Each quadrant represents True Positive, True Negative, False Positive, and False Negative, as shown in Figure 4.

	Predicted class (y=0)	Predicted class (y=1)
Actual class (y=0)	TN	FP
Actual class (y=1)	FN	TP

Fig. 4. Confusion Matrix

From the confusion matrix, we can calculate the accuracy of each run of the logistic regression model. Accuracy is defined as:

$$Accuracy = \frac{TP}{TP+FP} \quad (3)$$

We will use accuracy to compare the reliability and success of each model prediction.

IV. Results and Analysis

A. *Results of Logistic Regression Model*

In this experiment, we preprocessed a large dataset pertaining to flight departure delays for large domestic airlines and implemented a logistic regression algorithm to analyze this data and create a prediction model for potential flight schedule interruptions. The results of this experiment are represented by the confusion matrixes shown in Figure 5.

To test our model, we performed multiple runs while varying the number of iterations and threshold value. The learning rate was kept constant at 0.000001 for all the runs.

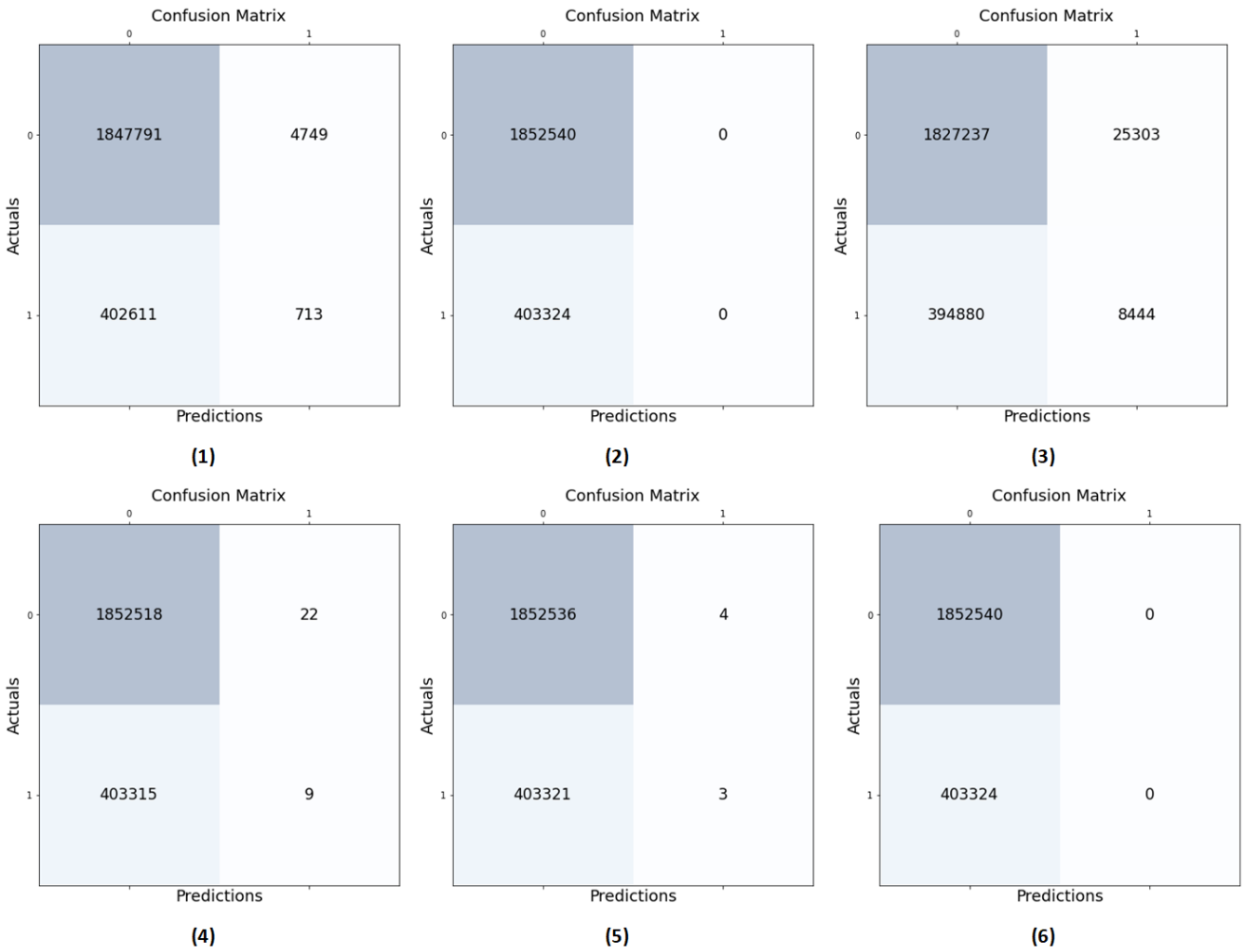

Fig. 5. Results of Logistic Regression Model

We found that our model is greatly dependent on the threshold value. When threshold equals 0.5, which is common for binary classification, the predicted values are all negative (same result as run #2). Both False Positive and True Positive values equal zero. We were not able to calculate the optimal threshold value via a function in our experiments. Our threshold values were determined by the trial-and-error method. The highest accuracy this model was able to achieve is 82.1%.

TABLE I. LOGISTIC REGRESSION MODEL RESULTS

Run	***Iterations***	***Threshold value***	***Accuracy (%)***
1	5	1.4788975056432135e-74	81.9
2	5	3.257488532207521e-70	82.1
3	200	2.1948785080142992e-72	81.4
4	200	7.175095973164411e-66	82.1
5	200	5.301718666092324e-65	82.1
6	300	5.301718666092324e-65	82.1

B. Comparison with Existing Libraries

In order to gauge the accuracy of the logistic regression model implementation, the results were compared against an existing public linear methods library. The same preprocessed flight departure delay dataset was run through the scikit-learn built in Logistic Regression [5]. The confusion matrix for this result is shown in Figure 6. The accuracy of this model was 86.6%, which is slightly higher than the best accuracy achieved from the implemented model.

Once the logistic regression model was analyzed, we wanted to compare our chosen model with other prediction models considered for this dataset. Utilizing the scikit-learn libraries again, the same dataset was run through a decision tree classifier and a random forest classifier. The results for all models are summarized in Table 2. The decision tree and random forest classifiers were significantly more accurate for this dataset, reaching over 98% accuracy for predicting flight delays. While this is an impressive value, the decision tree and random forest models are potentially overfitted if the significance of the inputs are not considered.

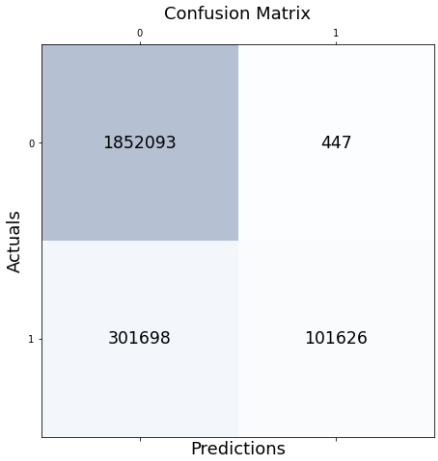

Fig. 6. Results of scikit-learn Logistic Regression

TABLE II. RESULTS COMPARISON

Model	***Accuracy (%)***
Logistic Regression (implemented)	82.1
Logistic Regression (scikit-learn)	86.6
Decision Trees	98.3
Random Forest	98.6

V. CONCLUSIONS AND FUTURE WORK

A. Conclusions

Although certain runs within our experiments were able to obtain accuracy scores over 80%, a significant number of false positives and negatives were still regularly observed. Comparisons performed against the built in Logistic Regression class from the scikit-learn library indicates that our True Positive outputs are much lower. We can implement and tune additional parameters in our model to improve accuracy. Further experimentation with available machine learning classification algorithms such as decision trees, random forest, and ensemble tactics may improve model performance across different performance measures. Additionally, several interesting data insights can be further explored within the dataset utilized in our studies. Future experiments could seek to identify the most reliable airlines for each available airport.

B. Future Work: Decision Trees

Machine learning techniques involving decision trees derive an internal strategy to obtain a conclusion based on presented data. These model's reasoning (learned function) can be represented visually using a tree like structure depicting decision split conditions within internal nodes and resulting conclusions within leaf nodes. The process of classifying a data instance involves starting from the root node and traversing towards a leaf node based on results of testing instance data against attribute conditions specified by intermediate nodes. As outlined by Tom Mitchell, decision tree-based machine learning methodologies work well when applied to problem spaces containing discrete target values and features having a small defined set of potential values [3].

This description fits our problem statement well as the data set contains a discrete binary value indicating if a delay occurred for a flight instance and many features contained categorical values. Exploratory experiments involving the application of decision trees to flight delay prediction were performed, using the scikit-learn.ml library, to see if more optimal results could be obtained than observed via logistic regression. For future work, we hope to explore various decision tree algorithm implantations and subsequently compare findings against logistic regression.

C. Future Work: Random Forest

Machine learning techniques that employ ensemble learning seek enhanced predictive performance by simultaneously joining forecasts generated from several models. One such model, referred to as a random forest, consists of a collection of individual decision tree classifiers, each of which is trained on a different portion of training data. Predictions are then generated by all individual trees and final classification is assigned based on the class with the highest count [4]. Another future work item to potentially explore further involves applying a random forest approach along with other ensemble methods to our problem domain.

REFERENCES

[1] United States Department of Transportation. (2020, January). *The Economic Impact of Civil Aviation on the U.S. Economy*. Federal Aviation Administration. Retrieved November 20, 2021, from https://www.faa.gov/about/plans_reports/media/2020_jan_economic_impact_report.pdf.

[2] *Understanding the reporting of causes of flight delays and cancellations*. Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics. (n.d.). Retrieved November 20, 2021, from https://www.bts.dot.gov/explore-topics-and-geography/topics/understanding-reporting-causes-flight-delays-and-cancellations-0.

[3] Mitchell, T. M. (1997). *Machine learning*. MacGraw-Hill.

[4] Géron, A. (2020). Hands-on machine learning with scikit-learn, Keras, and tensorflow: Concepts, tools, and techniques to build Intelligent Systems. O'Reilly.

[5] "sklearn.linear_model.LogisticRegression"[Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[6] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017, pp. 1-5, doi: 10.1109/ICCCI.2017.8117734.

[7] S. Narkhede, "Understanding Confusion Matrix," 09-May-2018. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.