Predicting Flight Delay Causation in United States using Logistic Regression

By: Huzaifa Gul

<https://github.com/huzaifagul/Predicting_Flight_Delay.git>

# Abstract

Airline are committed to providing the highest quality services and on-time performance for passengers. As such flight delays lead to negative impacts, mainly economical for customers, airline industries and airport authorities. For passengers, delay arrivals is associated with increased stress, missed connections, inconvenience and potential expenses for food and hotel stays. While even small improvements are impactful, they are difficult to achieve. Using the publicly available Bureau of Transportation Statistics dataset for on-time reporting of carrier performance from Jan-2020 to March 2021 (https://www.transtats.bts.gov/DL\_SelectFields.asp?gnoyr\_VQ=FGJ) with logistic regression techniques to construct an analytical model for analysis of flight attributes (such as origin, destination, carrier, delay in minutes and reason for delay). Additional models will be created to determine the most likely cause of a flight delay and to predict the approximate length of the delay. The tools to be used for the project will include Jupyter Notebook, Python language and Python libraries such as numpy, panda, matplotlib, seaborn, map, etc. This prediction will helpful for giving detail analysis of the performance of individual airlines, airports and making a well-assessed decision.

# Introduction

The purpose of this project is to look at the approaches used to build models for predicting flight delays that occur due to many reasons by using Bureau of Transportation Statistics dataset for on-time reporting of carrier performance from Jan-2020 to March 2021. The information can be used in creation of new policy and procedures within the airline industry. This project will focus on answering questions such as: What is the most common Origin and Destination? Which carrier has the most delays? Which city has the most delays? What are the reasons for the delays? These questions and answers will assist in the creation of a better artificial intelligence model in booking flights for customers and answering questions.

This project will provide in-depth predictive analysis on causes of airline delays and the delay time resulting in revenue growth and customer satisfaction.

# Literature Review

Flight delay predictions are a key component in the airline business. Prediction of flight delays is often conducted on a single or airport route. Gui and colleagues' (2020) article assessed the influences of flight delays through a random forest-based model in a generalized flight delay prediction task design. The dataset used automatic dependent surveillance-broadcast messages received, pre-processed, integrated with information on weather conditions, flight schedule and airport information in a classification and regression task. Gui and colleagues (2020) found that long short-term memory can obtain aviation sequence data. However, overfitting problems continue to occur. Therefore, the prediction of individual flight delays is significant enough for binary classification tasks of 90.2 % using the forest-based model.

Predicting flight delays is considered the most difficult in aviation control. A study by Ding (2017) assessed arrival flights prediction using a multiple linear regression algorithm, departure delay and route distance in predicting delays compared to Naïve-Bays and C4.5 approaches. Through a realistic dataset of domestic airports found that the prediction model was signed up to 80%. Therefore, using an ordinal classification task and a multiple linear regression model can predict the delay of flight arrivals. The information gathered allows the civil aviation industry to improve flight delay prediction accurately and the decision-making ability in the relevant departments and to use computer simulation technology that predicts early warning.

Yu and colleagues (2019) used a multifactor approach to track flight delays' inner patterns through Beijing International Airport data. Using specific micro influential factors, a novel prediction model based on a deep belief network and a support vector regression predicts delays' accuracy. Yu and colleagues (2019) found that the method of prediction can large datasets and assess key factors influencing delays. Through this, connected airlines and airports can work collaboratively in reducing inaccurate delay predictions. Therefore, assessing and using microelements such as air route and crowded degree of the airport and side macro factors such as weather, seasonal effect, air traffic control, and delay propagation give a more accurate assessment of delayed flights.

# Dataset

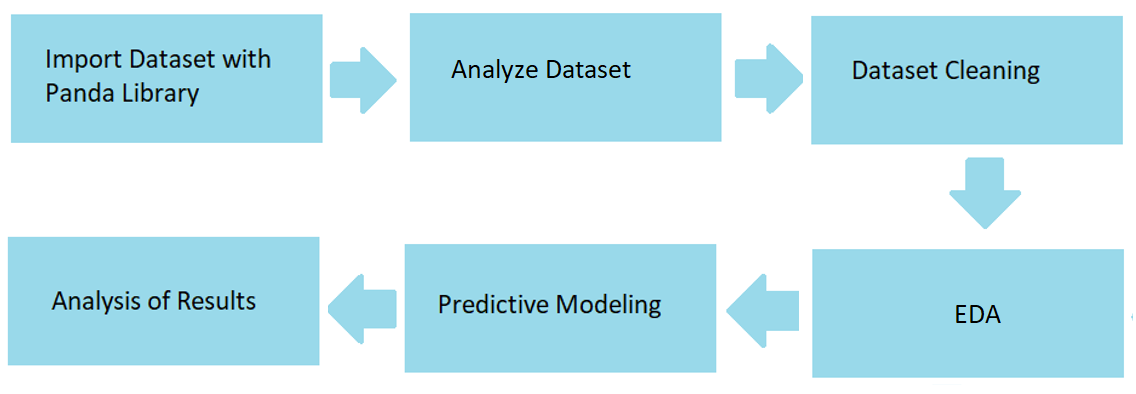
The dataset from Bureau of Transportation Statistics for the on-time reporting of carrier performance between Jan-2020 to March 2021 was merged to a single file from a set of 13 csv files provided by Bureau of Transportation Statistics which resulted in 28 columns and 5826726 rows to be used for analysis.

The columns to be used for analysis include:

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| Month | Integer | Month of Journey |
| Origin | Object | Origin Airport |
| OriginCityName | Object | Origin Airport, City Name |
| Dest | Object | Destination Airport |
| DestCityName | Object | Destination Airport, City Name |
| DepDelayMinutes | Float | Difference in minutes between scheduled and actual departure time. Early departure set to 0. |
| Cancelled | Float | Cancelled Flight Indicator (1=Yes) |
| Diverted | Float | Diverted Flight Indicator (1=Yes) |
| Carrier Delay | Float | Carrier Delay, in Minutes |
| Weather Delay | Float | Weather Delay, in Minutes |
| NAS Delay | Float | National Air System Delay, in Minutes |
| Security Delay | Float | Security Delay, in Minutes |
| Late Aircraft Delay | Float | Late Aircraft Delay, in Minutes |

Preliminary analysis of the data included importing and cleaning the data for use then methods of panda library such as describe, info and columns were used. Graphs were created to show the initial data set conditions and use cases.

# Approach



Step 1: The panda library was used to import and analyze the dataset and the dataset was imported as a csv file.

Step 2: Using the head method of the panda library, the columns of the dataset were shown.

Step 3: Using the info method, we can see the different datatypes within the dataset for easy manipulation

Step 4: The first analysis of the dataset was done by the describe method to comprehend the count, mean, standard deviation, min and max

Step 5: The next step was to clean the dataset which was done by clearing the empty rows and null values. To better understand if there was delay, a new object column was added with the text ‘yes’ or ‘no’. The dataset had months as numeric values and for easy readability, those numeric values were converted to acronyms of the months

Step 6: Developing the Figures as part of the result for further analysis.

Step 7: Using Linear regression for predictive analysis of current data.

Step 8: Formulation and interpretation of findings

# Exploratory Analysis

To understand the causation of delays and weather a flight will be delayed, a exploratory analysis performed. There is a correlation of 0.977 between arrival delays and departure delays which means flight that is delayed at take-off is more likely to arrive later than scheduled. The correlation is not 1 due to a flight being on time even if the take-off was delayed.

The dataset includes flight data over 14 months which were not delayed as shown in Figure 1.1 however, majority of the flights were significantly delayed. In Figure 1.2, most of the delays are within the months of Jan, Feb and March and to further understand what causes these delays, we need to compare the different types of delays.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 1.1 – Comparison between the number of Figure 1.2 – Comparison between the number of

of times a flight was delayed versus of times a flight was delayed versus

on time. on time over the course of 12 months.

In the next step we plot figures to visualize the difference in delays caused by Weather, NAS, Security and Late Aircraft Delay vs Delays caused by the Carrier. As shown in Figure 2.1, Carrier delay was the most common cause of delay next to late aircraft delays and security caused the least amount of delays as confirmed by Figure 2.2.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure 2.1 – Comparison of number of times delayed over 12 months between Carrier and Weather, NAS, Security and Late Aircraft Delays.

Chart, line chart

Description automatically generated

Figure 2.2 – Comparison of Carrier, Weather, NAS, Security and Late Aircraft over 12 month period.

Chart, bar chart

Description automatically generatedPreviously, all the discussion was done on the type of delay and comparison of those delay. On Figure 3.1, we are shown the delay count for each Carrier without any relation to type of delay or weather most of the delay was in departure or arrival. We can perceive from Figure 3.1, Southwest Airlines was the most delayed and Skywest, American Airline and Delta Airline were close behind with over 5000 delayed flight between Jan 2020 and Mar 2021.

Figure 3.1 – Comparison of different Carriers and

the amount of times each carrier was

delayed or on time.

# Results

For the prediction model building, we decided to use built-in functions within the sklearn kit using Logistic Regression which can provide probabilities, can be used with kernel methods, and is faster than neural network and RandomForest which has reduced variance (relative to decision trees).

Since we worked on a large data file, we had to prepare the datasets for different purposes. We had to fix technical issues with the set, primarily adding a new column IS\_Delay which determined wheater a flight was delayed or not using 1 for delayed or 0 for not delayed.

For the sake of simplicity, limited the number of airlines and airports to only include the major ones in the analysis was an option. However, the decision was made not to do that, as it would require to arbitrarily determine limits of the dataset.

The logistic regression created for predicting delays has a 91% accuracy. The data used for the model were Carrier, Weather, NAS, Security and Aircraft delay to train the model if a delay would occur. The precision rate for the model is 96% and the Recall rate of 92%

A picture containing chart

Description automatically generated

Fig 4.1 – Logistic Regression analysis results detailing confusion matrix.

85% (True Negative) There will be no delay and predicted correctly. 14% (False positive) There will be no delay and predicted incorrectly. 7% (False Negative) There will be delay and predicted incorrectly. 92% (True positive) There will be delay and predicted correctly.

The RandomForest created for predicting delays has a 92% accuracy. The data used for the model were Carrier, Weather, NAS, Security and Aircraft delay to train the model if a delay would occur. The precision rate for the model is 96% and the Recall rate of 94%

Chart, waterfall chart, treemap chart

Description automatically generated

Fig 4.2 – RandomForest analysis results detailing confusion matrix, accuracy, precision and recall.

85% (True Negative) There will be no delay and predicted correctly. 15% (False positive) There will be no delay and predicted incorrectly. 5% (False Negative) There will be delay and predicted incorrectly. 94% (True positive) There will be delay and predicted correctly.

Our data set after cleaning and adding a new column performed really well and since both models had a similar score, comparing the model’s performance was not required.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall |
| Logistic Regression | 91% | 96% | 92% |
| RandomForest | 92% | 96% | 94% |

# Conclusions

In this paper, dataset with 557,116 rows detailing delays in flights from 18 carriers departing and arriving throughout the United States between Jan 2020 to Mar 2021. Analysis showed, 448,168 were delayed and 108,948 were on time. The dataset was used to create a model for predicting if a flight would be delayed or on-time. Using Logistic Regression, a model was created with a 92% accuracy rate for predicting if a flight would be delayed. From our dataset, we also determined, most delays were carrier delays, Southwest Airlines was most the delayed and there was a 24.31% chance the flight would be delayed. The dataset can be further analyzed for each carrier and particular cities. Segmenting the dataset will reveal details not available at a general level. All statistical analyses were significant for this dataset. However, one should be aware this is a large data set and therefore easier to acquire statistically significant results.

# References

Bureau of Transportation Statistics (2021). On-Time : Reporting Carrier On-Time Performance from Jan-2020 to March 2021, https://www.transtats.bts.gov/DL\_SelectFields.asp?gnoyr\_VQ=FGJ

Ding, Y. (2017). Predicting flight delay based on multiple linear regression. IOP Conference Series: Earth and Environmental Science, 81, 012198. doi:10.1088/1755-1315/81/1/012198

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2019). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. IEEE Transactions on Vehicular Technology, 1–1. doi:10.1109/tvt.2019.2954094

Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. Transportation Research Part E: Logistics and Transportation Review, 125, 203–221. doi:10.1016/j.tre.2019.03.013