FLIGHT DELAY PREDICTION

ALEN ROY – 21BDA17
HENDRICK CARDOSO – 21BDA20
ARYA CHANDRAN – 21BDA34
AKSHAY S – 21BDA62

TABLE OF CONTENTS

ABSTRACT

In the current world, where a huge amount of importance is given to every minute that exists on the clock, the occurrence of a delay especially in the travel industry creates a sense of displeasure among the travelers. In this project, our objective is to predict if a certain flight would be delayed by 15 minutes or more, programmed in the R software using certain models such as Logistic Regression and Random Forest.

INTRODUCTION

More than 12,000 flights delayed, hundreds canceled during busy July Fourth weekend. This was one of the headlines published in a renowned article, that one of us had come across when we sat down to decide a topic for this project. The rate of flight cancellations and delays is higher this year than before the pandemic thanks to bad weather and staffing shortages. Some airlines have trimmed their schedules to give themselves a little comfort room.

Airline delays. They are the bane of every travelers existence and anxiety. Airlines won't tell you if your flight is likely to be delayed or not. Delayed flights can cause you to miss a connecting flight or an important business meeting. Thus, it was deemed to be a topic of high relevance, and we decided to explore if airline delays can be predicted with a reasonable degree of accuracy. In this analysis we try to develop a machine learning model that aims to predict if a flight arrival will be delayed by 15 minutes or more.

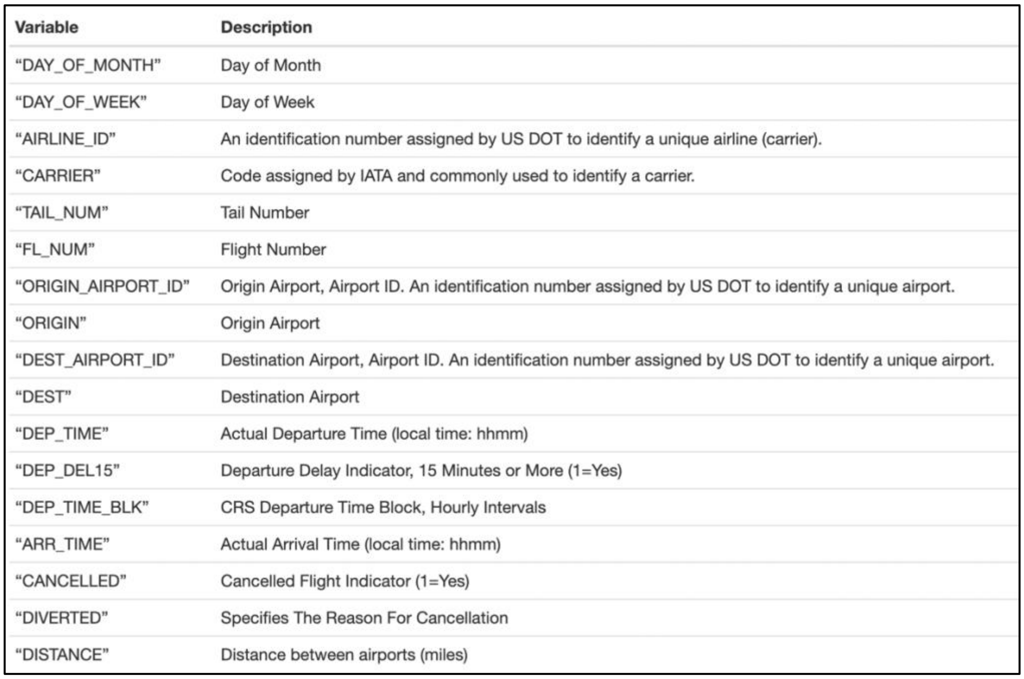
PROBLEM STATEMENT

The main objective of the project is to develop machine learning models that aim to predict if a flight arrival will be delayed by 15 minutes or more based on a chosen set of variables.

METHODOLOGY

Data Collection

The dataset is taken from the Bureau of transportation Statistics, a US government website that keeps the records of all commercial flights in the country. Since it is indeed a huge number, the focus was limited to a single month and January 2022 was chosen with relevance to proximity of current times. The dataset contains 5,63,737 records and 22 columns.



Variable	Description
“DAY_OF_MONTH”	Day of Month
“DAY_OF_WEEK”	Day of Week
“AIRLINE_ID”	An identification number assigned by US DOT to identify a unique airline (carrier).
“CARRIER”	Code assigned by IATA and commonly used to identify a carrier.
“TAIL_NUM”	Tail Number
“FL_NUM”	Flight Number
“ORIGIN_AIRPORT_ID”	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
“ORIGIN”	Origin Airport
“DEST_AIRPORT_ID”	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
“DEST”	Destination Airport
“DEP_TIME”	Actual Departure Time (local time: hhmm)
“DEP_DEL15”	Departure Delay Indicator, 15 Minutes or More (1=Yes)
“DEP_TIME_BLK”	CRS Departure Time Block, Hourly Intervals
“ARR_TIME”	Actual Arrival Time (local time: hhmm)
“CANCELLED”	Cancelled Flight Indicator (1=Yes)
“DIVERTED”	Specifies The Reason For Cancellation
“DISTANCE”	Distance between airports (miles)

A few of the columns in the dataset can be seen in the above table. For the purpose of this study we consider “ARR_DEL15” which indicates if a particular flight is delayed by 15 minutes or more, as our response variable which is a binary classifier containing values of 0 and 1 indicating if there is no delay or there exists a 15+ minute delay respectively.

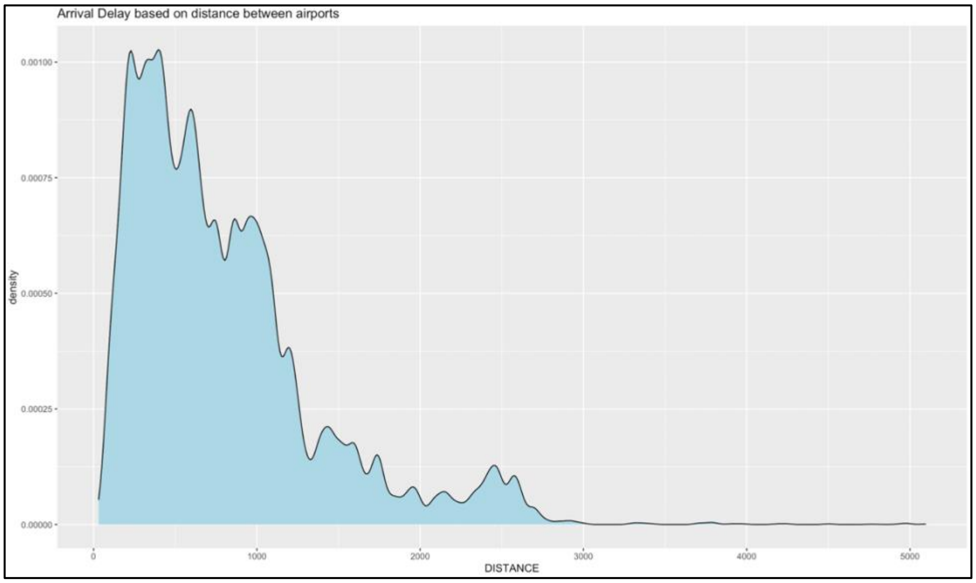
Data Preparation

The data set used is that of all flights originating and ending in the United States in January 2022. The data is automatically downloaded as a CSV file from the Department of Transportation Website. The downloaded dataset included 550,000+ rows of data. As this data was way too large for R to handle and run a regression model and the random forest algorithm, the dataset was minimized to include the 10 busiest airports in the US by total passenger traffic.

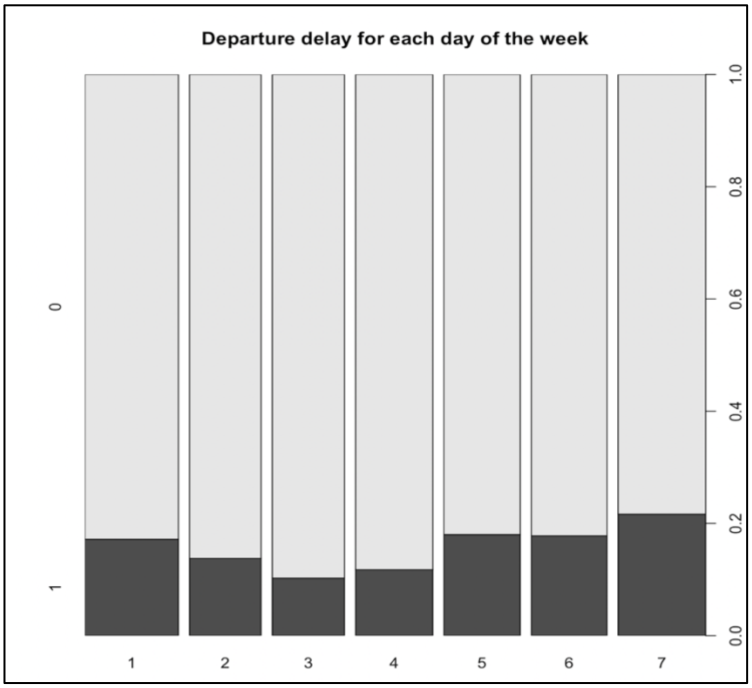
Primarily, we removed the records which had NA or missing values. Some of the columns are useless to our data analysis, so we NULL them out. Also, changing of the data class of the filtered data to enable data processing and running algorithms was implemented.

Exploratory Data Analysis

Exploratory Data Analysis is performed in order to obtain certain trends or patterns that can be obtained from the data in order to draw valid conclusions. A range of variables were taken and compared among themselves using different plots present in the R libraries. A few of the plots obtained with interesting observations are discussed below.



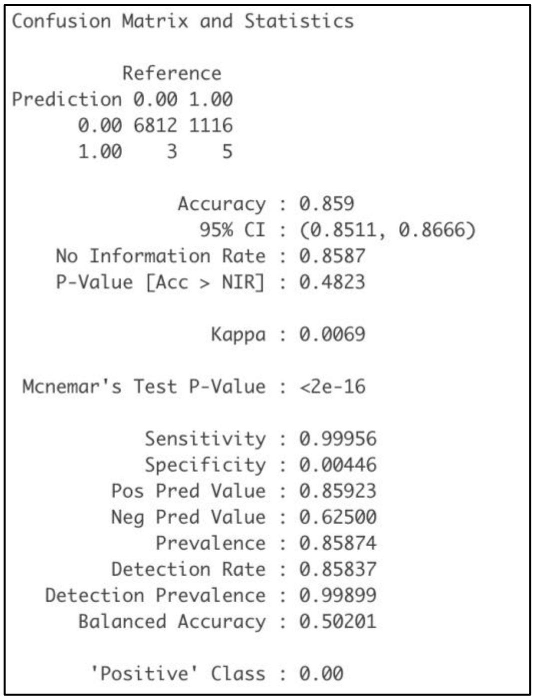
The above density talks about arrival delay based on distance. We can infer that even though there is a delay during departure, it is almost certain that with the increase of travel distance (especially above 3000 miles), the delay on arrival is likely to be negligible, as it is possible to recover time during the journey.



The spine plot discusses about the departure delays for each day of the week. On the Y- axis, 0 indicates no-delay and 1 indicates there is a delay during departure. The values of 1-7 on the x-axis represent the respective days of the week. Sunday(7) is the day with aircrafts having the most delay, and Wednesday(3) seems to be the least. A major reason for this congestion could be the negligence or less preference of the passengers to travel during the weekdays due to work related commitments, which in turn causes less flights to be in operation and therefore a considerably less chance for a delay to occur.

LOGISTIC REGRESSION MODEL

For our dataset, we have built a Logistic Regression Model as the response variable is a binary classifier which can hold only two values 0 and 1. Logistic regression has the ability to provide both the probabilities and classify them using continuous and discrete data.

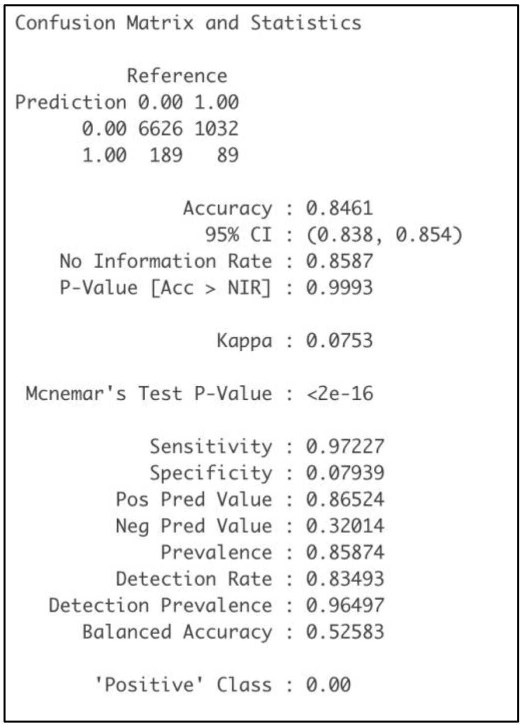


```
Confusion Matrix and Statistics

          Reference
Prediction 0.00 1.00
      0.00 6812 1116
      1.00    3    5

               Accuracy : 0.859
                 95% CI : (0.8511, 0.8666)
    No Information Rate : 0.8587
    P-Value [Acc > NIR] : 0.4823

                  Kappa : 0.0069

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99956
            Specificity : 0.00446
         Pos Pred Value : 0.85923
         Neg Pred Value : 0.62500
             Prevalence : 0.85874
         Detection Rate : 0.85837
   Detection Prevalence : 0.99899
      Balanced Accuracy : 0.50201

       'Positive' Class : 0.00
```

An obtained accuracy of 0.859 shows that the model is highly accurate which is an important parameter especially needed while predicting an output.
Sensitivity is the measure of how the model predicts no delay when there is no delay. The obtained value of 0.99956 shows that the model is extremely sensitive.
Specificity is the measure of model's ability to predict delay when there is a delay. The obtained value of 0.00446 infers that the specificity of the model is particularly low, which means that the model is not very accurate at predicting when there is a delay.

RANDOM FOREST MODEL

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.



```
Confusion Matrix and Statistics

          Reference
Prediction 0.00 1.00
      0.00 6626 1032
      1.00  189   89

               Accuracy : 0.8461
                 95% CI : (0.838, 0.854)
    No Information Rate : 0.8587
    P-Value [Acc > NIR] : 0.9993

                  Kappa : 0.0753

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.97227
            Specificity : 0.07939
         Pos Pred Value : 0.86524
         Neg Pred Value : 0.32014
             Prevalence : 0.85874
         Detection Rate : 0.83493
   Detection Prevalence : 0.96497
      Balanced Accuracy : 0.52583

       'Positive' Class : 0.00
```

Accuracy of the model is measured at 0.8461 which is also quite high, but minutely lower than the logistic regression model (measured at 0.859)
The sensitivity for the model was calculated to be 0.97227, which is also extremely high, but like accuracy it is also found to be lower than logistic regression model (calculated at 0.99956).
The specificity of the random forest model is found to be 0.07939, which is also particularly low, but unlike the other parameters it is observed to be considerably higher than logistic regression model (found to be 0.00446).

CONCLUSION AND FUTURE WORK

After working with the two models simultaneously on the same problem statement and comparing both of them we obtained two interesting results

- Logistic regression model provides a fast and accurate prediction of flights that will not be delayed.

- Random forest model performs four times better at predicting when a flight would be delayed than the logistic model.

The patterns identified using Data exploration methods were validated using the logistic regression model. The model can be made better by adding factors such as weather. This can be done through the API of a weather service. Also it was decided not to segregate by month as the computation was already intensive with so much data to process. Segregating by months according to season will give better insights as well. Other machine learning models can also be used to check for better accuracy and validation.

REFERENCES AND LINKS

- The dataset for a particular month can be obtained from Bureau of Transportation Statistics : https://transtats.bts.gov/DL_SelectFields.asp
- https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
- Link to the GitHub Repository containing codes in R: https://github.com/AryaChandran1999/Flight-Delay-Prediction