REPORT

2009 ASA Statistical Computing and Graphics Data Expo

NAME: CHANG JIA JIE
STUDENT NUMBER: 200548740

Executive Summary

The 2009 ASA Statistical Computing and Graphics Data Expo consisted of flight arrival and departure details for all commercial flights on major carriers within the USA, from October 1987 to April 2008. This is a large dataset; there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed. The complete dataset along with supplementary information and variable descriptions can be downloaded from the Harvard Dataverse at https://doi.org/10.7910/DVN/HG7NV7.

The purpose of this report targets to analyze two consecutive years dataset, namely 2002 and 2003 to answer the following questions:

1. When is the best time of day, day of the week, and time of year to fly to minimise delays?

2. Do older planes suffer more delays?

3. How does the number of people flying between different locations change over time?

4. Can you detect cascading failures as delays in one airport create delays in others?

5. Use the available variables to construct a model that predicts delays.

In this report, bar chart of delay percentage by period is conducted to discover the chances of departure delay in a specific time period. Secondly, correlation test between plane age and arrival delay is performed to determine whether older planes encounter more arrival delay. In addition, bar chart is conducted to visualize delay percentage by age. Next, a table is computed to show the top 10 frequently travel locations in 2002 and the variations in 2003. At the same time, line chart is plotted to visualize how does the number of top 10 frequently travel locations change over one year time. Furthermore, a detector function, that takes 6 arguments which are year, month, day, start_time, end_time and airport IATA code is conducted to detect cascading failures. Lastly, logistic model is conducted to predict airplane arrival delay from year, day of week, age, distance, air time, departure delay, time of day and season of the year.

All questions are answered using R and Python. In addition to the report, R and Python code are provided separately in RMarkdown and Jupyter notebooks respectively. Each report includes the all steps taken starting from raw data up to the answer for each question. Furthermore, any databases set up, data wrangling, cleaning operations carried out, and any modelling decisions made are also clearly described in each structured report. Each report includes relevant graphics and tables as part of the answer.

Table of Contents

1 Introduction

This report aims to address five specific questions using R and Python programming languages. The dataset provided by the Harvard Dataverse will be used, which contains flight arrival and departure data for commercial flights on major carriers within the USA from 2002 to 2003. Additionally, supplementary data pertaining to airports and planes will be employed for analysis. As per the Federal Aviation Administration (FAA), a flight is categorized as delayed if it arrives 15 minutes after its scheduled time. In this report, a delay will be defined as an arrival delay when the actual arrival time of a flight is 15 minutes or more behind its scheduled arrival time.

2 Data Preprocessing

2.1 Import data

Read 2002.csv as df1 and 2003.csv as df2. Then, combine df1 and df2 by rows into df.

Read airports.csv as airports

Read plane-data.csv as planedata

2.2 Data Wrangling and Data Cleaning

Remove 6 empty columns, namely 'CancellationCode', 'Diverted', 'CarrierDelay', 'WeatherDelay', 'NASDelay', 'SecurityDelay', and 'LateAircraftDelay'.

Remove 6 unused columns, namely 'UniqueCarrier', 'FlightNum', 'TaxiIn', 'TaxiOut', 'Cancelled'.

Remove rows that contain empty values.

Create new variables, ‘aDelay’ and ‘dDelay’, which are set to 1 if delay duration is greater or equals to 15, else 0.

3 Question 1:

3.1 Data Wrangling

Create a new variable, ‘time’ that split ‘DepTime’ into 4 parts, which are '0000-0600', '0600-1200', '1200-1800', '1800-2400'.

Create ‘date’ variable, which is a datetime object from ‘year’, ‘Month’, ‘DayofMonth’ in America/Los_Angeles timezone.

Create variable ‘season’, that classify date into Spring, Autumn, Summer and Winter.

3.2 When is the best time of day, day of the week, and time of year to fly to minimise delays?

Bar chart of delay percentage by period is conducted to discover the minimum possibility of departure delay in that time period. From this, customer can determine the best time to fly.

Note that delay percentage here will not add up to 100% as it is calculated by dividing the total delay by the total flights in that particular period. For example, delay percentage of Monday = total delay on Monday / total flights on Monday. [NOT: total delay on Monday / total delay of the week]. Hence, it can be interpreted as the likelihood of encountering a flight delay if someone takes a flight during that time period.

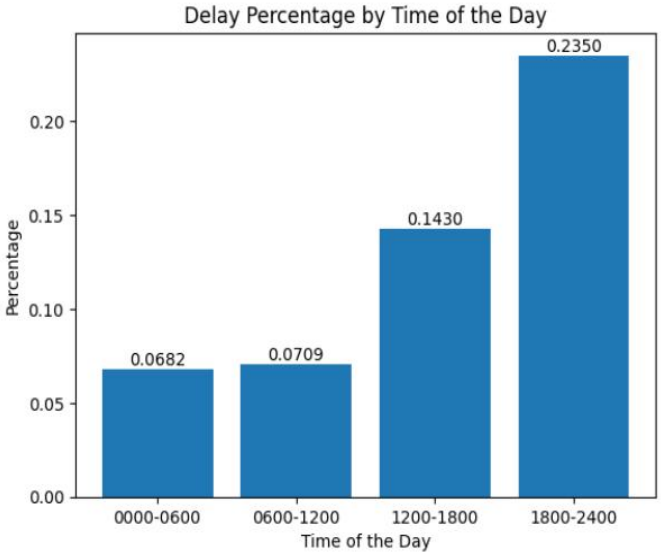


Figure 1: Delay Percentage by Time of the Day

Best time of the day to fly

The ‘Delay Percentage by Time of the Day’ bar chart shows that the best time of a day to fly are 0000-0600 and 0600-1200 which account for the least delay, 6.82% and 7.09%. On the other hand, flights at 1800-2400 suffer relatively high opportunity of delays among the four time period, at 23.5%.

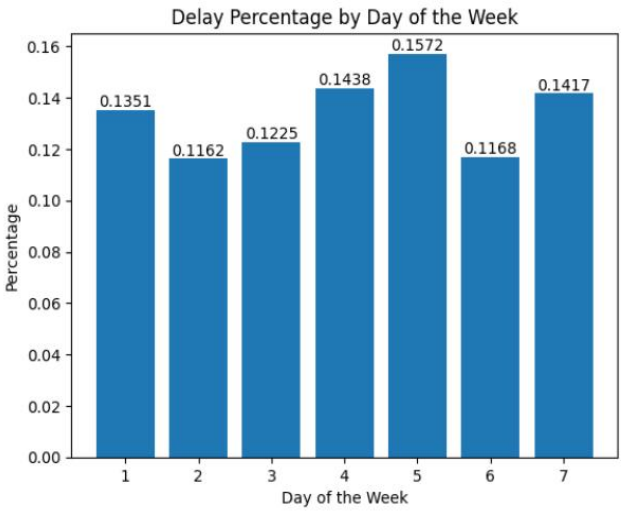


Figure 2: Delay Percentage by Day of the Week

Best day of the week to fly

The 'Delay Percentage by Day of the Week' bar chart shows that the best day of a week to fly are Tuesday and Saturday followed by Wednesday. Tuesday and Saturday have the least delay percentage namely 11.62% and 11.68%, followed by Wednesday 12.25%. Friday is not a recommended day to fly due to it has the highest chance of delay, at 15.72%.

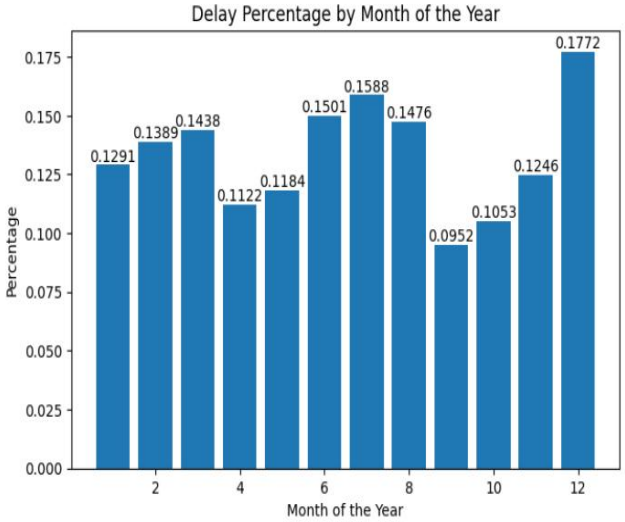


Figure 3: Delay Percentage by Month of the Year

Best month of the year to fly

The 'Delay Percentage by Month of the Year' bar chart shows that the best month to fly are September, October, April and May, which possess the delay percentage of 9.52%, 10.53%, 11.22% and 11.84% respectively. On the other hand, taking flights in December would result in a highest chance of encountering a departure delay, 17.72%.

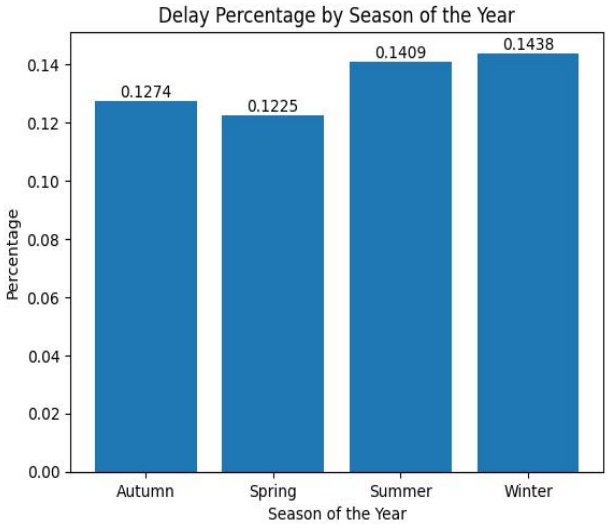


Figure 4: Delay Percentage by Season of the Year

Best season of the year to fly

The 'Delay Percentage by Season of the Year' bar chart shows that the chances of delays among the four seasons are similar with a range of 2.13% only. However, the best season to fly is Spring with the least delay percentage 12.25%. However, is it not recommended to fly during Summer and Winter due to the highest chance of delay.

4 Question 2:

4.1 Data Wrangling and Data Cleaning

Rename 'tailnum' column of planedata to 'TailNum'

Replace blank values with NA

Merge df and planedata by 'TailNum'

Remove nine unused columns 'TailNum', 'tailnum', 'type', 'manufacturer', 'model', 'status', 'aircraft_type', 'engine_type', 'year'

Remove rows that include NA and string 'None'

Convert 'issue_date' column to POSIXct format

Create ‘age’ variable by calculating the difference between 'date' and 'issue_date'

Convert 'age' column to numeric type and years in units

Remove rows where 'age' is negative

Create a new variable, ‘agegroup’ that group ‘age’ into 3 group, namely '0-8', '9-17', and '18-28'

4.2 Do older planes suffer more delays?

Plane age is the difference between flying year and issue date of plane. Age and arrival delay are used to perform a correlation test to determine whether older planes use more elapsed time than the scheduled one, which result in an arrival delay.

```
        Pearson's product-moment correlation

data:  df3$Age and df3$`Percentage of Delay`
t = 0.33461, df = 24, p-value = 0.7408
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3278619  0.4437821
sample estimates:
       cor
0.06814418
```

Figure 5: Pearson's product-moment correlation

Based on the correlation test result, the p-value of 0.7408 indicates that the observed correlation between the two variables is not statistically significant at the 5% significance level. This means that if we assume the null hypothesis of no correlation, there is a 74.08% chance of observing a correlation coefficient as large or larger than 0.092 by chance alone. Hence, we cannot conclude that there is a true relationship between the two variables based on this result.

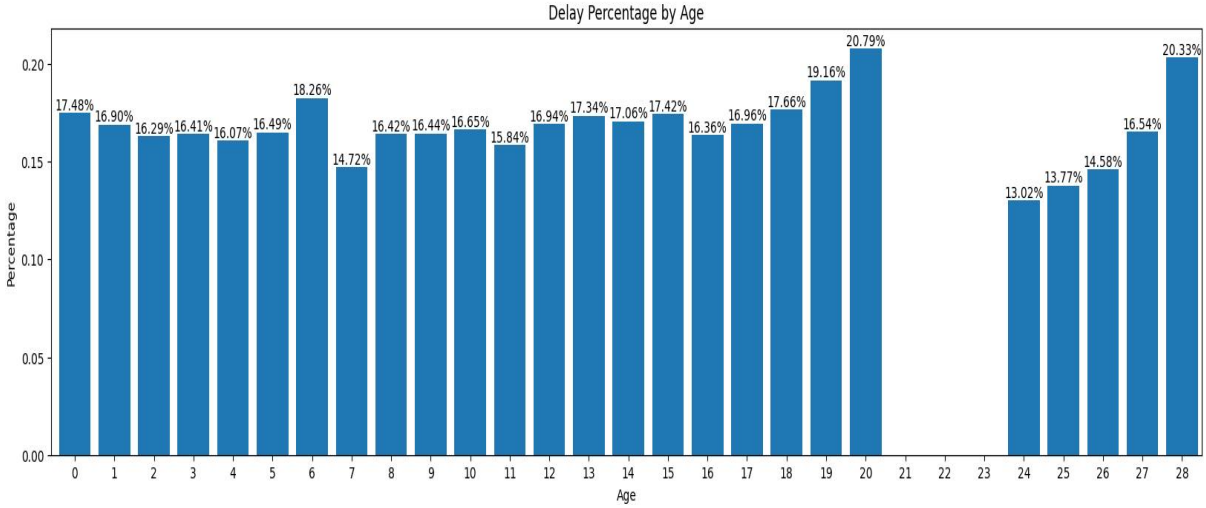


Figure 6: Delay Percentage by Age

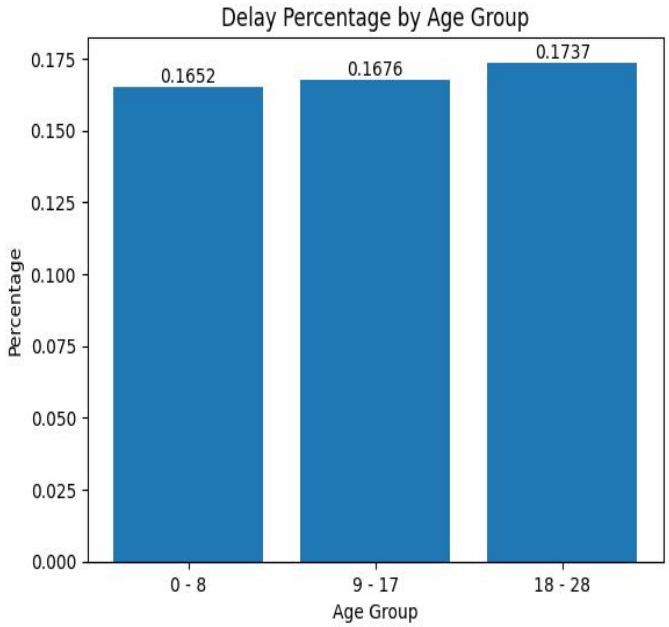


Figure 7: Delay Percentage by Age Group

The 'Delay Percentage by Age' bar chart above shows that the highest delay percentage are 20.79% at age 20 and 20.33% at age 28. However, according to the 'Delay Percentage by Age Group' bar chart, it is clear that the delay percentages among three age groups are almost the same with only 0.85% difference between the highest and the lowest.

Hence, both data visualization and correlation test result indicate that the statement of older planes suffer more delay is not true.

5 Question 3:

5.1 Data wrangling and data cleaning

Create a new column 'location' by concatenating 'Origin' and 'Dest'

Calculate the top 10 frequently traveled locations in 2002 named top_10_values_1

Calculate the frequency of the same 10 locations in 2003 named top_10_values_2

Convert top_10_values_1 and top_10_values_2 into data frames and merge them by rows

Reshape the data frame with 2002 and 2003 as columns to have better comparison

Join new data frame with airports based on Origin and Dest columns to achieve the full name of airports

Remove unused columns

5.2 How does the number of people flying between different locations change over time?

A table is computed to show the top 10 frequently travel locations in 2002 and the variations in 2003. From the table, number of people flying between different locations can be determined. Then, plot a line chart to visualize how does the number of top 10 frequently travel locations change over one year time.

	Location	2002	2003	Origin	Dest
1	LAX-LAS	12733	11262	Los Angeles International	McCarran International
2	LAS-LAX	12526	11177	McCarran International	Los Angeles International
3	MSP-ORD	12007	12377	Minneapolis-St Paul Intl	Chicago O'Hare International
4	ORD-MSP	11907	12254	Chicago O'Hare International	Minneapolis-St Paul Intl
5	BOS-LGA	11539	12662	Gen Edw L Logan Intl	LaGuardia
6	LGA-BOS	11537	12568	LaGuardia	Gen Edw L Logan Intl
7	LAS-PHX	11143	10345	McCarran International	Phoenix Sky Harbor International
8	PHX-LAS	11091	10219	Phoenix Sky Harbor International	McCarran International
9	PHX-LAX	11066	10498	Phoenix Sky Harbor International	Los Angeles International
10	LAX-PHX	11027	10518	Los Angeles International	Phoenix Sky Harbor International

Figure 8: Table of top 10 frequently travel locations in 2002

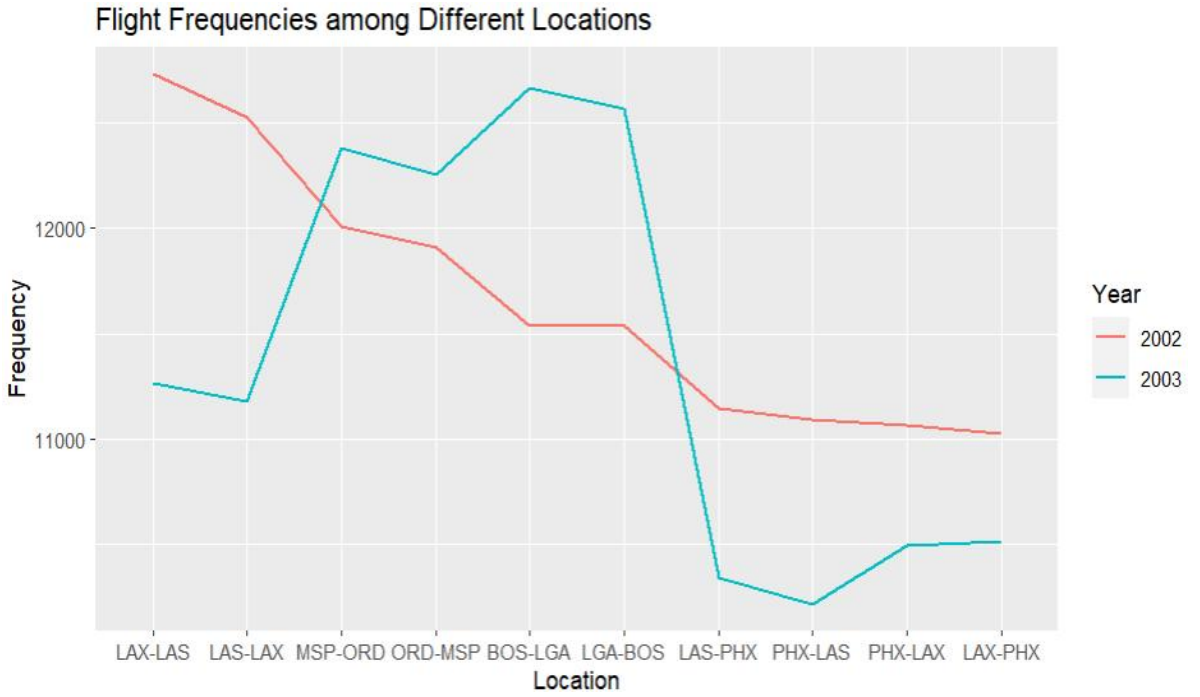


Figure 9: Flight Frequencies among top 10 frequently traveled locations in 2002

According to the 'Flight Frequencies among top 10 frequently traveled locations in 2002' line chart, six of the top 10 frequently traveled locations experienced a decrease after one year. Especially, the number of flights between the top two locations in 2002, Los Angeles International Airport and McCarran International Airport, experienced a major decrease. In addition, the remaining four decreases were also flights to and from these airports, highlighting the significant impact of this decrease on air travel to and from this region. However, the number of flights between Minneapolis-ST Paul Intl airport and Chicago O'Hare airport International increased massively in 2003.

6 Question 4:

6.1 Can you detect cascading failures as delays in one airport create delays in others?

Yes, a detector function is conducted to detect cascading failures as delays in one airport create delays in others. The function takes 6 arguments which are year, month, day, start_time, end_time and airport IATA code. To ensure a certain accuracy, recommended time interval is less than 2 hours. There are two levels of detection in the function.

Function:

```
detector <- function(year, month, day, start_time, end_time, airport)
```

Level 1: Examine the arrival delay at the selected airport within the selected date and time interval.

Level 2: If there is matched record in the flight datasets, the function will further examine the departure delay at the same airport within the same date and time interval.

Example:

Case1: There is arrival delay(s) at the selected airport within the selected date and time interval which causes cascading failures.

Input: detector(2002, 1, 31, 1500, 1600, 'CLT')

Output: There was delay in Charlotte/Douglas International airport from time 1500 to 1600

which has created cascading delays in Piedmont Triad International airport

which has created cascading delays in Myrtle Beach International airport

which has created cascading delays in Gen Edw L Logan Intl airport

which has created cascading delays in Philadelphia Intl airport

Case2: There is arrival delay(s) at the selected airport within the selected date and time interval but does not cause cascading failures.

Input: detector(2003, 2, 22, 2300, 2400, 'BOS')

Output: There was delay in Gen Edw L Logan Intl airport from time 2300 to 2400 but it does not create any cascading delay in other airports.

Case3: There isn't any arrival delay(s) at the selected airport within the selected date and time interval thus no cascading failures.

Input: detector(2002, 1, 10, 2200, 2300, 'PHX')

Output: There was no delay in Phoenix Sky Harbor International airport from time 2200 to 2300. So, it did not cause any cascading delay in other airports.

7 Question 5:

7.1 Data wrangling and data cleaning

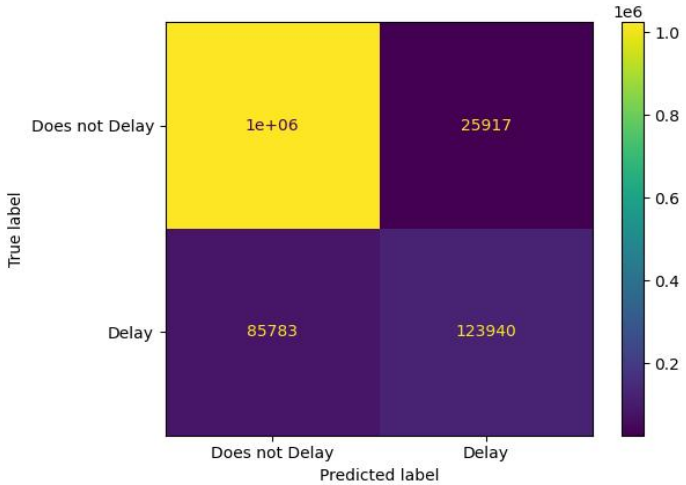
Convert these seven variables, namely aDelay, Year, Month, DayOfWeek, dDelay, time, season into factor type

Create dummy variables for the six variables, namely Year, Month, DayOfWeek, dDelay, time, season

7.2 Use the available variables to construct a model that predicts delays.

A logistic model is conducted to predict airplane arrival delay. The model takes 9 arguments as inputs which are year, day of week, age, distance, air time, departure delay, time of day and season of the year. It outputs 0 or 1, indicating 'Not Delayed' and 'Delayed' respectively. The dataset is split into 70% for trainset and 30% for testset. The model is trained based on the trainset data and uses testset to examine model's accuracy.

Here, is the result:



The confusion matrix shows that the model has correctly predicted 1,025,041 samples as Does Not Delay (true negatives) and 123,940 samples as Delay (true positives), but has incorrectly predicted 85,783 samples as Does Not Delay (false negatives) and 25,917 samples as Delay (false positives).

Figure 10: Confusion Matrix

```
[[1025041   25917]
 [  85783  123940]]
Accuracy: 0.9113970941102467
Specificity: 0.975339642497607
Sensitivity: 0.5909699937536655
F1 Score:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95   1050958
           1       0.83      0.59      0.69    209723

    accuracy                           0.91   1260681
   macro avg       0.87      0.78      0.82   1260681
weighted avg       0.91      0.91      0.91   1260681

AUC: 0.7831548181256363
```

Figure 11: Model Result

The model achieves an accuracy of 91.14%, which means that 91.14% of flights were correctly classified as delayed or not delayed. The precision of the model (specificity) is 97.53%, indicating that out of all the flights that were predicted to be delayed, 97.53% actually were delayed. The recall of the model (sensitivity) is only 59.1%, indicating that out of all the delayed flights in the dataset, the model is able to correctly identified 59.1% of them. The F1 score of the model is 0.69, which is a weighted average of the precision and recall for classifying delay. This indicates that the model has a moderate performance in correctly predicting the delay.

Finally, the AUC score of the model was 78.32%, indicating that the model is quite good at distinguishing between delayed and not delayed flights. However, there is still room for improvement in the model's performance. Overall, while the model is able to achieve high accuracy and specificity, it is less accurate in predicting the positive class, as evidenced by the low sensitivity and F1 score. Further analysis and optimization of the model may be necessary to improve its performance on this class.

8 Appendices

Image used:

Cover Page

<https://api.hub.jhu.edu/factory/sites/default/files/styles/full_width/public/travel-assistance-hub_2.jpg>

References

Flight cancellation and delay (2023) Wikipedia. Wikimedia Foundation.

Available at: <https://en.wikipedia.org/wiki/Flight_cancellation_and_delay#:~:text=A%20flight%20delay%20is%20when,later%20than%20its%20scheduled%20time.>

[Accessed: March 22, 2023]

Dates and times of seasons 1995 - 2025 (no date) Dates and Times of Seasons.

Available at:

<https://www.glib.com/season_dates.html>

[Accessed: March 19, 2023]