**Airline Arrival Delay Prediction**

Table of Contents

[**CHAPTER ONE - INTRODUCTION** 4](#_Toc152773142)

[**1.1 Background** 4](#_Toc152773143)

[**1.2 Research Questions** 6](#_Toc152773144)

[**CHAPTER TWO – LITERATURE REVIEW** 7](#_Toc152773145)

[**2.1 What made us take this data?** 7](#_Toc152773146)

[**2.2 From where did we collect this data?** 8](#_Toc152773147)

[2.3 About the Data 9](#_Toc152773148)

[2.4 Brief explanation of the work 12](#_Toc152773149)

[**2.4.1 Hypothesis or research question 1** 13](#_Toc152773150)

[**2.4.2 Hypothesis or research question 2** 13](#_Toc152773151)

[**2.4.3 Hypothesis or research question 3** 13](#_Toc152773152)

[**2.4.4 Hypothesis or research question 4:** 13](#_Toc152773153)

[**CHAPTER THREE METHODOLOGY: DATA PREPARATION** 14](#_Toc152773154)

[**3.1 Software** 14](#_Toc152773155)

[**3.2 Data Collection** 14](#_Toc152773156)

[**3.2 Data Wrangling** 16](#_Toc152773157)

[**3.3 Data Preprocessing** 17](#_Toc152773158)

[**CHAPTER FOUR METHODOLOGY: EXPLORATORY DATA ANALYSIS** 22](#_Toc152773159)

[**4.1 Data Description** 22](#_Toc152773160)

[**4.2 Missing Data** 23](#_Toc152773161)

[**4.3 Numeric Variable** 23](#_Toc152773162)

[**CHAPTER FIVE: MODELLING** 28](#_Toc152773163)

[**5.1 Hypothesis 1** 28](#_Toc152773164)

[**5.2 Hypothesis 2** 31](#_Toc152773165)

[**5.3 Hypothesis-3** 37](#_Toc152773166)

[**5.4 Hypothesis 4** 42](#_Toc152773167)

[**CHAPTER SIX: MODEL EVALUATION** 47](#_Toc152773168)

[**6.1 Hypothesis 1 - RandomForest Classifier** 47](#_Toc152773169)

[**6.2 Hypothesis 2 - Logistic Regression** 47](#_Toc152773170)

[**6.3 Hypothesis 3 - XGBClassifier** 47](#_Toc152773171)

[**6. 4 Hypothesis 4 - Logistic Regression** 48](#_Toc152773172)

[**CHAPTER SEVEN CONCLUSION** 49](#_Toc152773173)

[**7.1 Summary Discussion** 49](#_Toc152773174)

[**7.2 Applications** 50](#_Toc152773175)

[**7. 3 Limitations and Future Research** 50](#_Toc152773176)

[**7.4 Recommendation** 51](#_Toc152773177)

[**7.5 Conclusion** 52](#_Toc152773178)

[**References** 53](#_Toc152773179)

# **CHAPTER ONE - INTRODUCTION**

## **1.1 Background**

The development of commercial flights represents one of the most compelling episodes that demonstrate how air travel moved from being a privilege reserved for an elite group to becoming a prevalent form of transport experienced by countless people. The aeronautics industry has seen major developments, from the inception of open cockpit flights to the emergence of mighty jet engines. Nevertheless, due to increased air travel, flight delays have emerged as a constant problem to be dealt with not only by airlines but also by travelers. The modern aviation industry now finds it essential to predict and manage airport arrival delays.

Air travel in the start-up commercial flight era was a unique event that only the rich and daring individuals could afford. Since there were discomforts of flying in the early days, including open cockpits and vulnerability, this was still a means of transport that was only for the rich and famous. In spite of these obstacles, air traffic steadily grew over the years until finally, in the fifties, jet airplanes made mass air travel a reality. With this age came great ideas of possibilities, which included an increase in the affordability of air travel for a larger population.

As we move on with time, air travel is now a fundamental aspect of the transport sector that connects people and business organizations all over the world. This sector has seen extraordinary development whereby tens of millions of flights are transporting billions of people yearly. The upsurge has its challenges in scheduling, air traffic, and unpredictable disturbances. Predicting airline arrival delays is of great importance in today's aviation industry. Airline delays cause massive problems for airports and travelers. Airlines encounter high operational costs, schedule disruptions, and unsatisfied customers as a result of delays. However, customers can experience discomfort during travel, fail to connect with trains or lose money and time. On-time performance is a passenger characteristic in the competitive environment of the air transportation industry.

Airline arrival delay prediction is important because of a number of reasons. It helps airlines be in control so that they anticipate any potential delays during the operation and minimize the effect of these delays on schedule and travelers’ experiences. This allows the airlines to identify possible delays before time and take preventive actions like changing the route, rebooking crews, or notifying the customers on time. Being proactive boosts the productivity and positivity of brands. Additionally, forecasting delays in arrivals helps in the efficient use of such resources as aircraft, workforce, and other airport facilities. Airlines must manage resources effectively in order to cope with the complexities of modern air travel and provide smooth journeys for their customers. This is crucial as timely and correct estimations help to build up more reliable information systems in aviation.

The forecast of arrival delays also brings a great benefit to passengers. Passengers are advised when a delay is likely to occur so that they can find other modes of transportation. This will enable them to decide on matters such as connecting flights, ground transport, and other plans that will reduce the pressures of such travel. Moreover, the fact that some airline companies may engage in proactive communication among their clients can go a long way toward reassuring passengers that these carriers are making an effort to be transparent with them (Wang et al., 2022).

The utilization of machine learning models in the proposed study is in line with the industry's quest to employ cutting-edge technologies in making decisions about business operations. With the use of machine learning algorithms, lots of historical as well as real-time data related to fliers' delay can be analyzed to spot some interesting patterns and trends. Such models consider things like departure delays, weather state, and crowding at airports to provide reasonable inferences concerning arrival delays.

Airline arrival delay forecasting is an essential component of modern aviation management. Issues such as airline management and problems that passengers face are considered in it. With the development of the industry, air transport has become more reliable and predictable due to the introduction of effective predictive models (Kiliç et al., 2023). Therefore, this project concerning airline arrival delay prediction helps in this effort by using data-driven measures to boost efficiency within today’s aviation sector.

## **1.2 Research Questions**

Prompt arrival of flights forms an integral component of successful airline operations as well as content customers in today’s aviation world. This study, therefore, has a research question that is centered on the importance of forecasting airline arrival delays. With this growth and more complexities in the aviation sector, it is essential to understand the causal factors of uncertainty. The focus of this study is on using machine learning methods to improve forecasting abilities that would see airlines managing resources efficiently while giving passengers timely information for smooth journeys.

1. Do flights with higher schedule time have more chances of delay or not?
2. Do the flights that are scheduled to fly in the daytime have less chance of delay?
3. Is it true that the airplanes that are scheduled to fly to hub airports have more chances of getting delayed?
4. Do Flights with longer taxi-out times or NASDelay have a higher risk of arriving late due to traffic congestion or a problem at the origin airport?

# **CHAPTER TWO – LITERATURE REVIEW**

## **2.1 What made us take this data?**

These days, the use of airlines has increased so much that most people are more interested in traveling by airplane rather than using road and rail transport. So, the scope for research in this area became popular (Xu et al., 2017). This made us really interested in choosing this as our research area. Previously, people used different machine learning models to predict flight delays. Now, in this paper, we use some machine learning models for predicting the delay in commercial airlines. We used RandomForestClassifier, LogisticRegression Model, and XBGClassifier Model to predict and justify our hypothesis. Besides, this study will give numerical classification for a better understanding of whether the flight is delayed or not.

According to the Airline trends from the Department of Transportation statistics, around 20-25% of commercial airline flights were delayed in 2022. This level of arriving late is simply unacceptable and falls far below passenger expectations. When flights are delayed, it has an adverse impact on the system. Because airlines operate on tightly linked schedules, a single delay can spread when aircraft and crew move out of position for following flights (Kiliç et al., 2023). Passengers experience severe disruption when delays result in missed connections, canceled flights, or late arrivals. Business travelers frequently have limited tolerance for delays.

Flight delays cause passengers to lose productivity, spend additional expenses, and experience general dissatisfaction. According to surveys, on-time performance is a major element influencing airline satisfaction. When delays cause misconnections or cancellations, airlines invest much in passenger re-accommodation. Compensation costs are also high. As a result, enhancing on-time performance remains a primary focus for airlines. It has a quick effect on customer happiness, income, and costs.

## **2.2 From where did we collect this data?**

We collected this Reporting Carrier On-Time Performance data(aviation data) from the Bureau of Transportation Statistics. The Bureau of Transportation Statistics (BTS) is a division of the United States Department of Transportation responsible for collecting, compiling, analyzing, and disseminating transportation data. According to the Bureau of Transportation Statistics (BTS), American Airlines handled 83.1 million systemwide (domestic and international) scheduled service passengers in August 2023. When seasonality is considered, August enplanements are up 1.7% from July but down 4.2% from the all-time high set in January 2020.

Key officials in BTS: **Ms. Patricia S. Hu** is the Director of the Bureau of Transportation Statistics. Ms. Hu has substantial statistical experience, a thorough understanding of transportation, and a strong research background. Ms. Hu joined the Department of Transportation in 2011 after spending 29 years at the Oak Ridge National Laboratory, where she was the head of the Center for Transportation Analysis from 2001 to 2009. They have created and directed initiatives focusing on transportation survey methodologies and data quality, transportation analysis and model creation, and visualization-based transportation decision-making tools. Much of her work has been published in top transportation and applied statistics publications, and she has given numerous presentations at conferences.

**Dr. Rolf R. Schmitt** is the deputy director of the Bureau of Transportation. He is a nationally recognized transportation policy specialist, as well as a statistician who develops statistics to help transportation choices. Since 1977, he has worked for BTS, the Department of Transportation's National Transportation Policy Team, the Federal Highway Administration, and the National Transportation Policy Study Commission. Between 1992 and 2000, he was instrumental in the establishment of BTS, and he returned as Deputy Director in 2012.

BTS has been assigned to compiling unbiased transportation data, statistics, and information for policymakers across all modes of transportation. It collects statistics on aviation, roadways, maritime transportation, pipelines, public transportation, trains, and other transportation modes. BTS provides complete data on flights, passengers, finances, performance indicators, delays, problems, and more for aviation. BTS issues reference manuals, statistical compendiums monthly and annually, the National Transportation Atlas Database, and other reports. The data is made publicly available for examination and download via the BTS website and online data tools.

We selected a dataset that contains about 600000 observations with 110 variables. We chose this dataset because it has more values in it, and we can train our models more to predict the values accurately. We use test and train-based strategies for building the models.

## 2.3 About the Data

This part will explain the Data set that we picked in this paper. Our dataset had different variables and values. Below are the some of the variables and values,

**Dates and Times:**

FlightDate

CRSDepTime

DepTime

CRSArrTime

ArrTime

**FirstDepTimeDuration/Time Intervals:**

CRSElapsedTime

ActualElapsedTime

TaxiOut

TaxiIn

**AirTimeAirport IDs:**

OriginAirportID

**DestAirportIDAirline Codes:**

Reporting

Airline

**IATA\_CODE\_Reporting\_AirlineAircraft Tail Number:**

Tail Number

**Flight Number:**

Flight\_Number\_Reporting

**AirlineDelays:**

DepDelay

ArrDelay

CarrierDelay

WeatherDelay

NASDelay

SecurityDelay

**LateAircraftDelayLocation Attributes:**

OriginCityName

OriginState

DestCityName

DestStateDistance

**DistanceCategorical Attributes:**

Month

DayOfWeek

**CancellationCodeBinary Variables:**

ArrDel15

DepDel15

Cancelled

Diverted

## 2.4 Brief explanation of the work

Here, our goal is to develop a classification model to predict whether a flight will arrive on time (ArrDel15 = 0) or delayed (ArrDel15 = 1). Here, (ArrDel15 = 0) and (ArrDel15 = 1) are the attributes in the dataset that will be used to identify whether it is delayed or arrived on time. Also, attributes like scheduled and actual departure time, arrival time, delays, airtime, Distance, and more will help.

The target variable: ArrDel15 (indicating an arrival delay of 15 minutes or more as (1), or if it is not, it is (0). So, the predictor variables could be CRSDepTime, DepDelay, AirTime, Distance, OriginAirportID, DestAirportID, CarrierDelay, WeatherDelay, NASDelay, and other delay types.

**2.4.1 Hypothesis or research question 1:** Flights with longer scheduled airtime will have a higher likelihood of arrival delays due to a greater chance of operational issues.

Here, the Predictor variable to justify the hypothesis: CRSElapsedTime, Distance

**2.4.2 Hypothesis or research question 2:** Flights departing early in the day have a lower delay risk due to less propagation of delays.

Predictor Variable: CRSDepTime

**2.4.3 Hypothesis or research question 3:** Flights to hub airports have a higher delay risk due to congestion.

We chose "DestAirportID" to satisfy this research question.

**2.4.4 Hypothesis or research question 4:** Flights with longer taxi-out times or NASDelay have a higher likelihood of arrival delays due to traffic congestion or issues at the origin airport.

Predictor Variables: TaxiOut, NASDelay

# **CHAPTER THREE METHODOLOGY: DATA PREPARATION**

## **3.1 Software**

We used Python to write the modules. Python is a powerful language; unlike R, Python is a complete language that can be used for research and development. Python provides a lot of built-in modules and libraries. By using those libraries and modules, we can solve the problem in different ways. We used different modules and libraries from Python to fulfill our requirements. One of the modules that we used is "pandas ."It is one of the main libraries in Python.

Finally, we used the Python 3.9.0 version to write the ML models. It is the latest stable version that is available on the market at present. It is recommended to use this version for quick processing and better results. As I mentioned earlier, we also used Pandas in our models. Pandas are the most popular tool that is available in Python by default. It is an inbuilt Python package which is developed by "Wes McKinney."

We used Pandas for cleaning our data and for data Processing as well. Moreover, we also used it for data visualization. Apart from these two, we also used "Seaborn." Seaborn is a good visualization library that is available in Python and used for visualizing statistical graphics plotting.

## **3.2 Data Collection**

We need to analyze the data across multiple dimensions to answer various questions to improve the flight experience for travelers and not impact the margins of the airlines like:

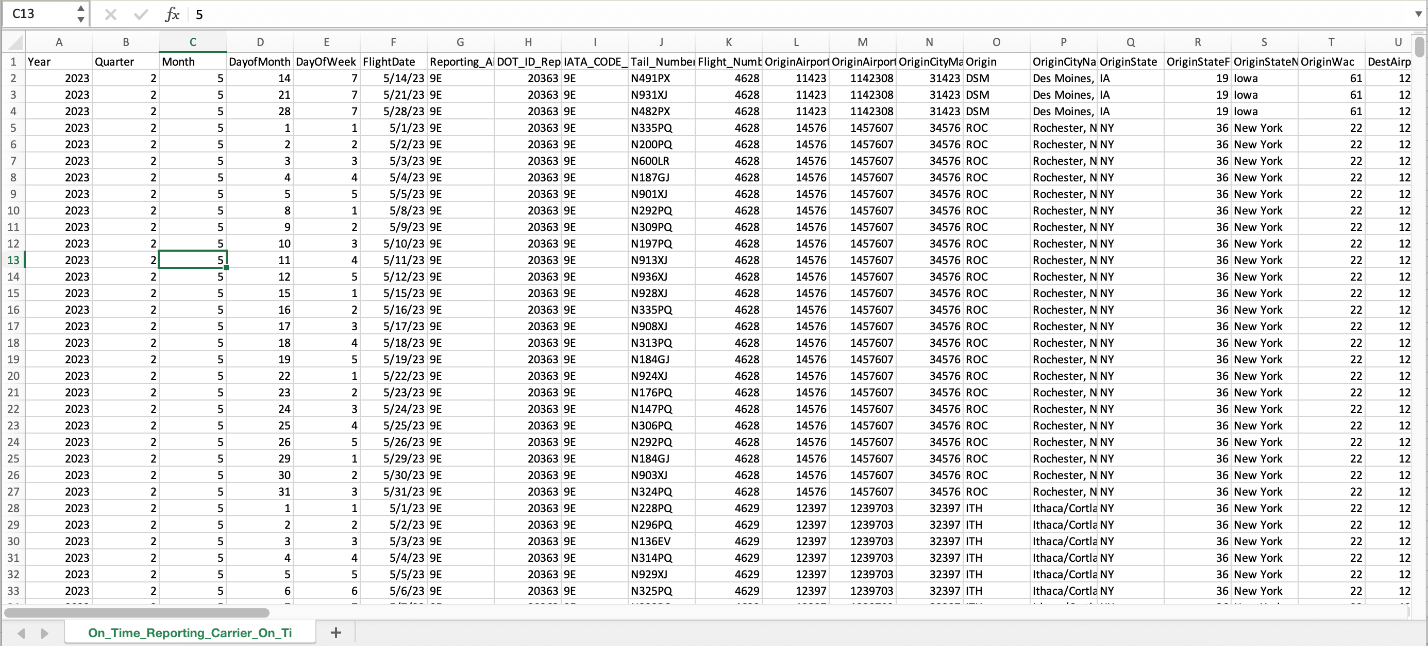
a. What are the factors that are associated with flight delays?

b. Is it possible to predict the flight delay using these variables?

c. Are the number of flight delays increasing per flight over some time?

At a high level, the features are classified and grouped as below.

* scheduled time
* departure time
* arrival time
* delays
* airtime
* distance
* taxi in
* tax out
* wheels on
* wheels off
* flight number
* origin airport
* destination airport
* origin city
* destination city
* diverted
* has delay
* time of the year Year/quarter/month/day of the month/day of the week
* weather



This is the sample Dataset that we collected from the BTS website, Which is a popular website for flight data. This is a state bureau website. The Categories of the data are available in 110 different features or variables. The dataset contains data from 1987 to the present (2023). We used Pandas' inbuilt Python module, which is a .csv extension to data frames, for importing the dataset. We imported the pandas as "import pandas as pd" in our code and then read the .csv file into the code; after reading that, the pandas converted the CSV file into Data frames, which is understood by the ML model easily (Pophale et al., 2023).

## **3.2 Data Wrangling**

Data Wrangling is a process of cleaning and removing the errors from the data set. Because of the high availability of data, it is important to do data analysis (Thiagarajan et al., 2017). One more use case for Data Wrangling is that it will convert the raw data into more usable data.

This is the piece of code we used for Data Wrangling.

drop\_cols = data.columns[data.isna().all()].tolist()

print('cols to be dropped: start')

print(drop\_cols)

print('cols to be dropped: end')

clean\_data = data.drop(columns=drop\_cols)

print(clean\_data.describe())

Coming to the explanation part of this code, the first line of the code creates a column where all the values in those columns are naN. The second line in the code prints a message indicating the start of the list of the columns to be dropped. The next line prints the list of the columns where all values are NAN. The fourth line prints the message indicating the end of the list of columns to be dropped. The fifth line is where the actual data processing will start, and it cleans the data. In the next line, the final cleaned data will be printed.

## **3.3 Data Preprocessing**

Data preprocessing involves five different steps,

Data cleansing

Data reduction

Data transformation

Data Enrichment

Data validation

list\_of\_cols = [

'Year,' 'Quarter,' 'Month,' 'DayofMonth,' 'DayOfWeek,'

'DOT\_ID\_Reporting\_Airline', 'Flight\_Number\_Reporting\_Airline',

'OriginAirportID,' 'OriginCityMarketID,'

'OriginStateFips,' 'OriginWac,' 'DestAirportID,' 'DestAirportSeqID,'

'DestCityMarketID', 'DestStateFips', 'DestWac', 'CRSDepTime',

'DepTime,' 'DepDelay,' 'DepDelayMinutes,' 'DepDel15',

'DepartureDelayGroups,' 'TaxiOut,' 'WheelsOff,' 'WheelsOn,'

'TaxiIn', 'CRSArrTime', 'ArrTime', 'ArrDelay',

'ArrDelayMinutes,' 'ArrDel15', 'ArrivalDelayGroups,'

'Diverted,' 'CRSElapsedTime,' 'ActualElapsedTime,' 'AirTime,' 'Flights,'

'Distance,' 'DistanceGroup,' 'CarrierDelay,' 'WeatherDelay,'

'NASDelay,' 'SecurityDelay,' 'LateAircraftDelay,' 'FirstDepTime'

]

columns\_with\_mean\_gt\_0 = clean\_data[[\*list\_of\_cols]]

print(columns\_with\_mean\_gt\_0.corrwith(columns\_with\_mean\_gt\_0['ArrDel15']))

# Added new features on checking correlation with ArrDel15

# DepDelay 0.497804

# DepDelayMinutes 0.481266

# DepDel15 0.740944

# DepartureDelayGroups 0.680244

# ArrDelay 0.562191

# ArrDelayMinutes 0.512619

# ArrDel15 1.000000

# ArrivalDelayGroups 0.756383

# TaxiOut 0.202383

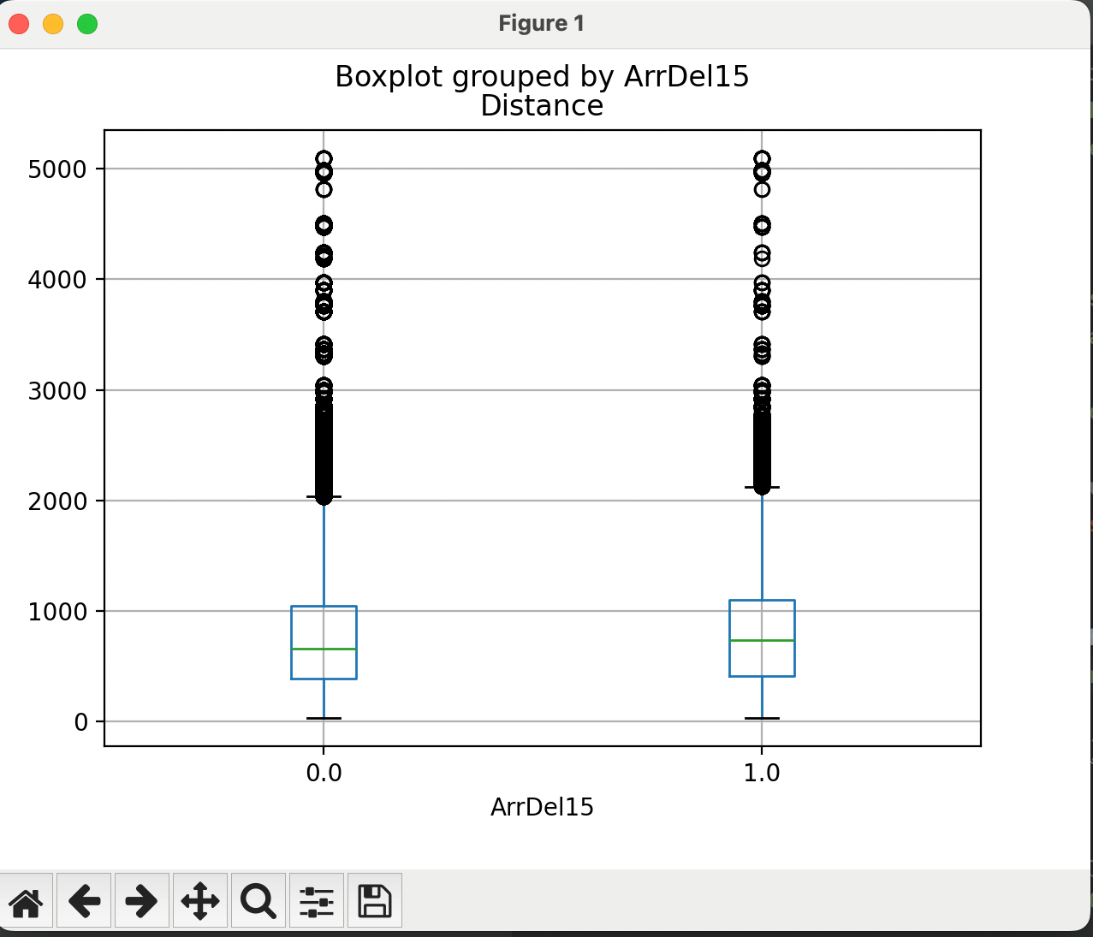
# WheelsOff 0.203919

# WheelsOn 0.089645

The above piece of code will do the data preprocessing and give the final usable data as data frames.

**Figure 1**

*Boxtplot Grouped by ArrDel15*



# **CHAPTER FOUR METHODOLOGY: EXPLORATORY DATA ANALYSIS**

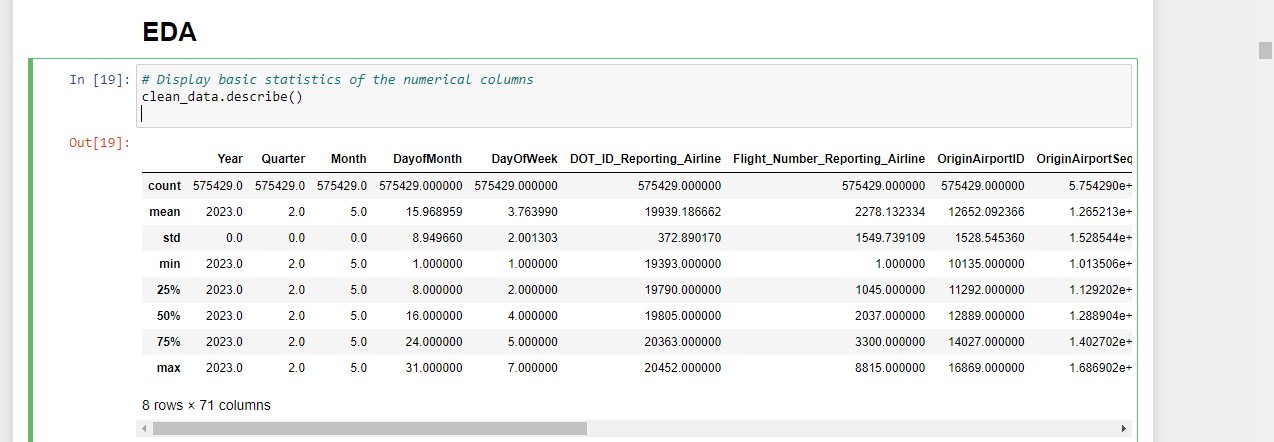
EDA means exploratory Data Analysis. Data scientists use it to analyze the data, investigate the data sets, and summarize their main characteristics.

## **4.1 Data Description**

This data description shows that there are some numbers for columns of a dataset on commercial flights. It covers a number of attributes, some of which are flight details, airport information, and times. The key statistics involve mean, standard deviation, minimum, maximum, and quartile values for respective columns. Various other columns like Div2AirportSeqID and Div3LongestGTime seem to hold data gaps with empty spaces labeled "NaN," which could be seen as indicating some holes or missing data in certain specified rooms. The above summary captures the gist of this dataset, pinpointing some areas that can guide data exploration and cleansing.

**Figure 2**

*EDA* *Screenshot*

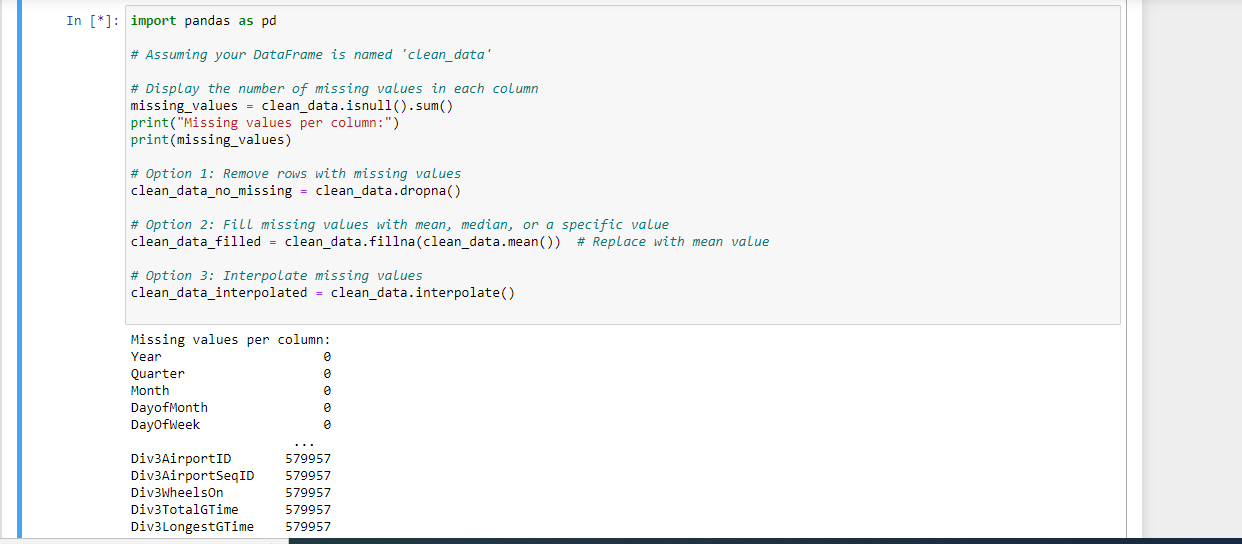


## **4.2 Missing Data**

Missing data is present in the data. They need to be cleaned.

**Figure 3**

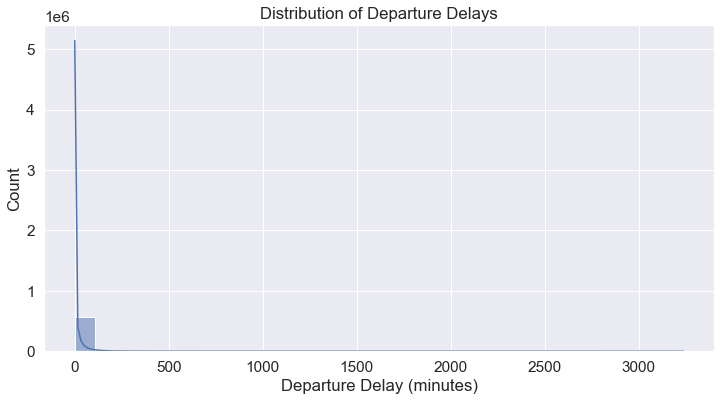
*Missing data*



## **4.3 Numeric Variable**

**Figure 4**

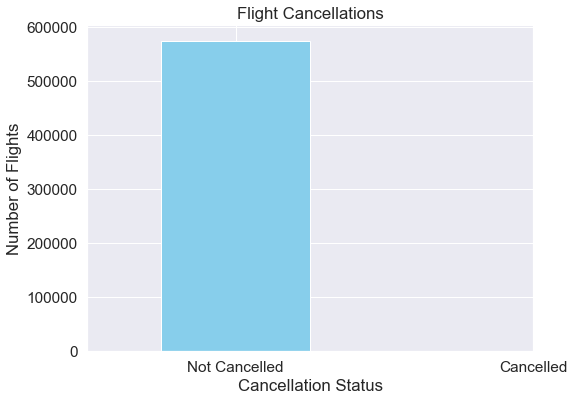
*Distributing of Departure Delays*



The figures show that the delays were always zero.

***Figure 5***

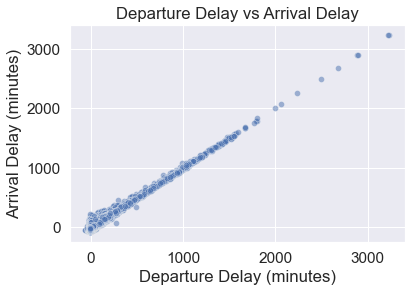
*Flight cancellations*



The figure shows that all the flights were not canceled.

**Figure 6**

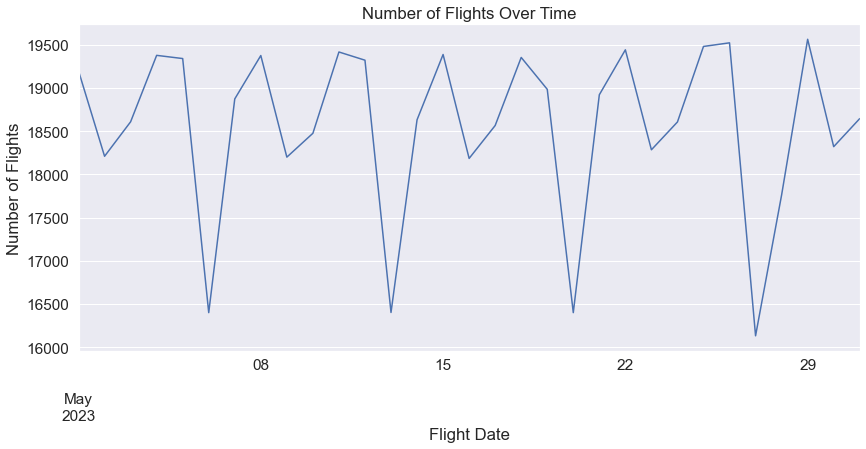
*Departure Delay vs Arrival Delay*



The two have a linear relation.

**Figure 7**

*Number of flights over time.*



There is a trend in the number of flights over time.

# **CHAPTER FIVE: MODELLING**

## **5.1 Hypothesis 1**

Model: We used The RandomForestClassifier Model to justify this hypothesis against the below

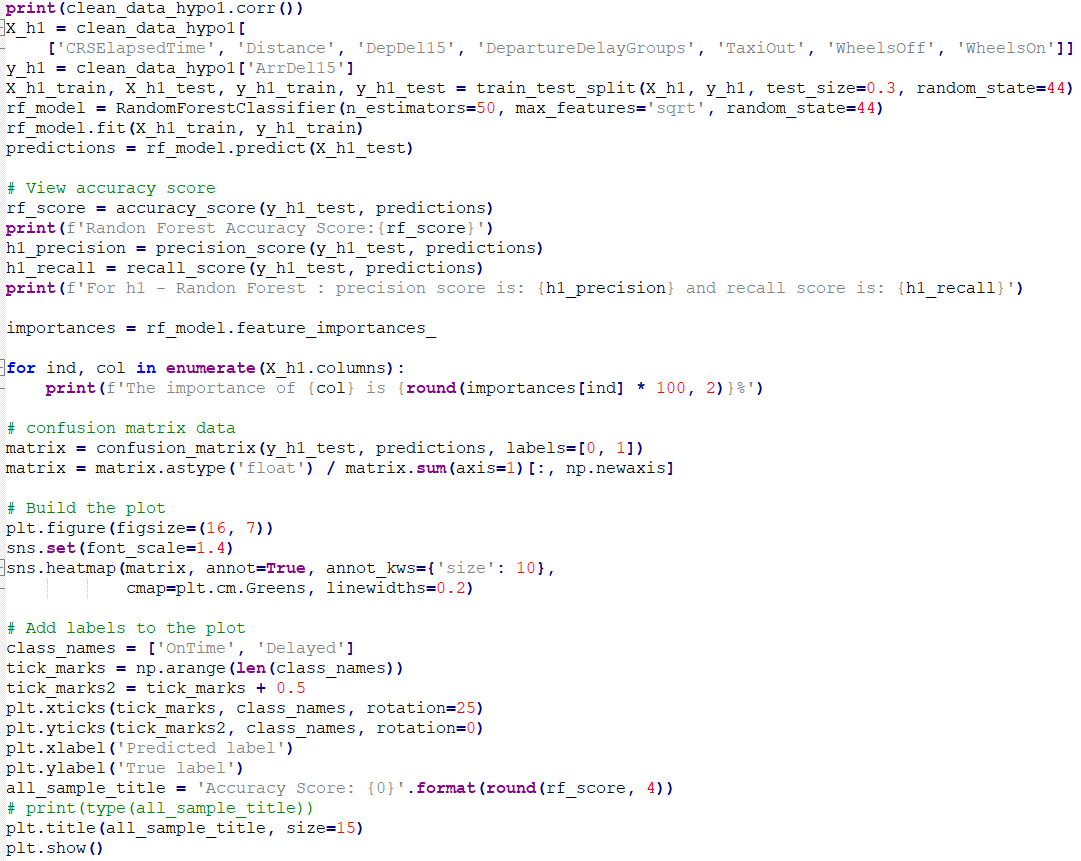
Inputs Variables: 'CRSElapsedTime,' 'Distance,' 'DepDel15', 'DepartureDelayGroups,' 'TaxiOut,' 'WheelsOff,' 'WheelsOn'

Output Target: ArrDel15

The training and testing data was split in the 70:30 ratio, and for the model, these values were chosen: n\_estimators=50, max\_features='sqrt', random\_state=44

**Figure 8**

*Codes*



The above picture shows the code side of the RandomForest Model. Leo Breiman and Adele Cutler developed this RandomForest Model. This model predicts the value by creating different decision trees. We input 70 percent of our data as our train data and 30 percent as our test data. Furthermore, we got some output values, which are given below.

Some of the important observations for this model are:

Accuracy Score:0.9455537597970214

the precision score is 0.9003978138825891, and

the recall score is 0.7827913892300731

The importance of CRSElapsedTime is 6.31%

The importance of Distance is 6.87%

The importance of DepDel15 is 21.23%

The importance of DepartureDelayGroups is 40.06%

The importance of TaxiOut is 8.74%

The importance of WheelsOff is 8.23%

The importance of WheelsOn is 8.55%

Interestingly, the identified input features CRSElapsedTime and Distance had the least importance among all the input features.

Import matplotlib. pyplot as plt

importances = rf\_model.feature\_importances\_

feature\_names = X\_h1.columns

# Sort the features by importance

indices = importances.argsort()

# Plotting the bar chart

plt.figure(figsize=(12, 8))

plt.barh(range(len(indices)), importances[indices], align='center')

plt.sticks(range(len(indices)), [feature\_names[i] for i in indices])

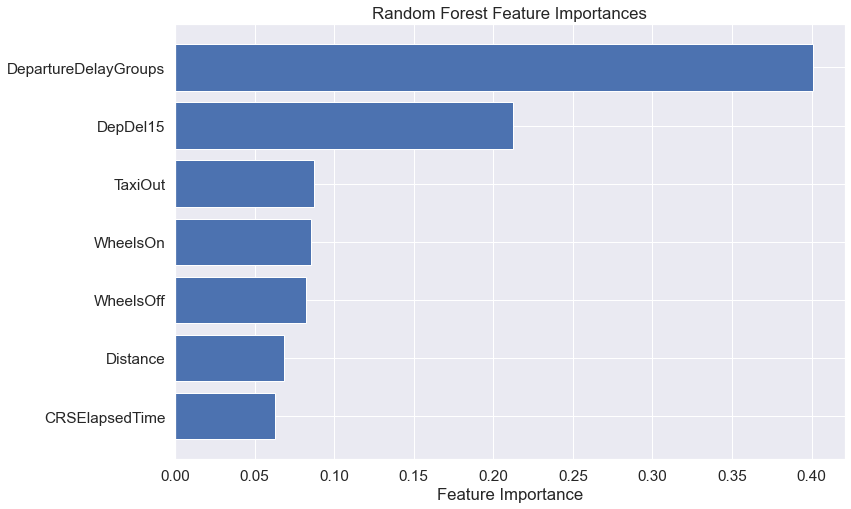
plt.xlabel('Feature Importance')

plt.title('Random Forest Feature Importances')

plt.show()

**Figure 9**

*Random forest importance*



This bar chart shows the imported features from the Random Forest model. The importance of those features was placed into each bar. The result of this finding is that it has the highest importance among them for flight delays. Conversely, CRSElapsed time contributes less towards determining delays since it is illustrated with a short bar. A glance at this chart helps us appreciate that there are key factors influencing the forecasting of flight delays whose details can be used in the analysis and refinement of model inferences.

## **5.2 Hypothesis 2**

Model: The LogisticRegression Model was chosen to run against the below. Inputs features: CRSDepTime, 'Distance,' 'DepDel15', 'DepartureDelayGroups,' 'TaxiOut,' 'WheelsOff,' 'WheelsOn.'

Output Target: ArrDel15

The training and testing data was split in the 70:30 ratio, and for the model, these values were chosen: explicitly set max\_iter = 1000

The above screenshot will explain the logistic regression model. It is a Classification algorithm that is used to predict the probability of certain classes based on some dependent variables. After running the code, we got some observations, as shown below.

Some of the important observations for this model are:

The importance of CRSDepTime is -0.01%

The importance of Distance is 0.06%

The importance of DepDel15 is 0.05%

The importance of DepartureDelayGroups is 18.93%

The importance of TaxiOut is 2.03%

The importance of WheelsOff is -0.01%

The importance of WheelsOn is -0.01%

The accuracy score is 0.9385213376663248

the precision score is 0.8888463745934571, and the recall score is 0.750856551813304

Interestingly, the identified input feature, CRSDepTime, had the least importance among all the input features.

# Plotting the grouped bar chart

fig, ax = plt.subplots(figsize=(10, 6))

bar\_width = 0.35

# Plot bars for each class

bar1 = ax.bar(np.arange(len(class\_names)), cm[:, 0], bar\_width, label='Class 0', alpha=0.8)

bar2 = ax.bar(np.arange(len(class\_names)) + bar\_width, cm[:, 1], bar\_width, label='Class 1', alpha=0.8)

# Add labels, title, and legend

ax.set\_xlabel('True label')

ax.set\_ylabel('Normalized Counts')

ax.set\_title('Confusion Matrix')

ax.set\_xticks(np.arange(len(class\_names)) + bar\_width / 2)

ax.set\_xticklabels(class\_names)

ax.legend()

# Display the values on top of the bars

for bar in [bar1, bar2]:

for rent in bar:

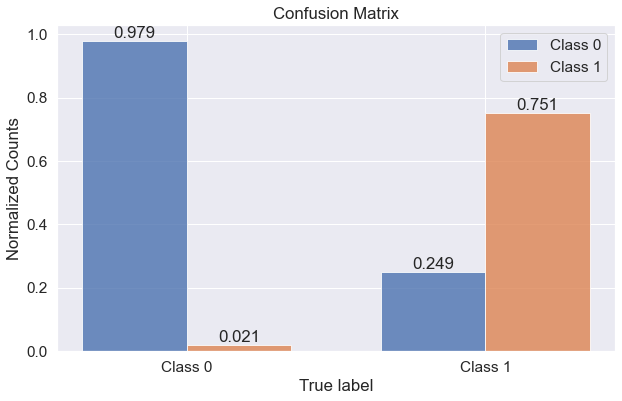
Height = rect.get\_height()

ax.text(rect.get\_x() + rect.get\_width() / 2, height, f'{height:.3f}', ha='center', va='bottom')

plt.show()

**Figure 10**

*Confusion matrix*



Normally, a classification model is shown by using the grouped bar chart as a confusion matrix. It shows the proportion of correct predictions (the diagonal blocks) in each category with regard to mistakes that result in classification errors (the off-diagonals). A graph illustrates how well the model forecasts each class, thus giving an easy overview.

Hypothesis-3:

Model: The XGBClassifier Model was chosen to run against the below

Inputs features: 'DestAirportID,' 'Distance,' 'DepDel15', 'DepartureDelayGroups,' 'TaxiOut,' 'WheelsOff,' 'WheelsOn,' 'ArrDel15.'

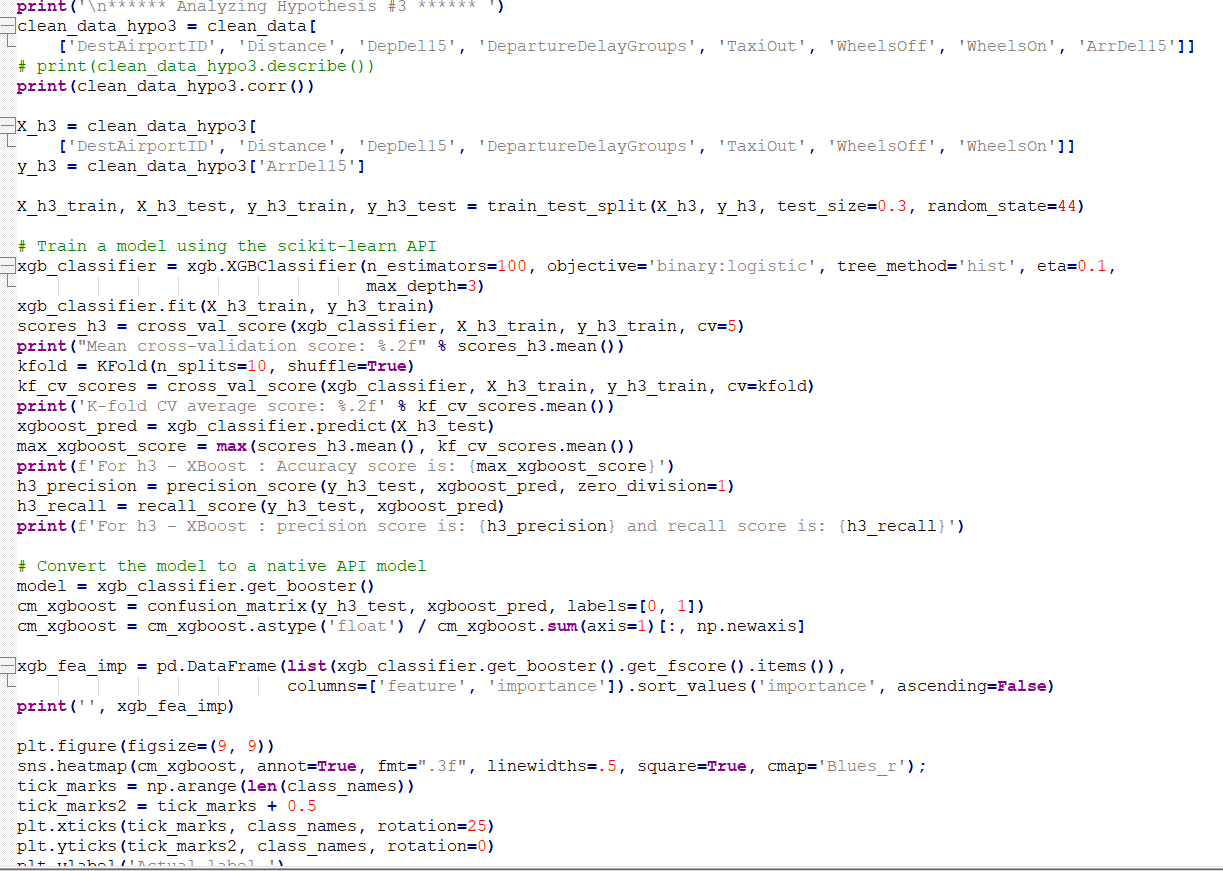
Output Target: ArrDel15

The training and testing data was split in the 70:30 ratio, and for the model, these values were chosen: n\_estimators=100, objective='binary: logistic', tree\_method='hist,' eta=0.1,

max\_depth=3

**Figure 11**

*Training and Testing model*



Some of the important observations for this model are:

Mean cross-validation score: 0.94

K-fold CV average score: 0.94

For h3 - XBoost: The accuracy score is: 0.9393272095332671

For h3 - XBoost, the precision score is 0.9071083343156902, and the recall score is 0.7461697588725839

feature importance

3 DepartureDelayGroups 179.0

4 TaxiOut 147.0

1 Distance 143.0

6 WheelsOn 76.0

5 WheelsOff 53.0

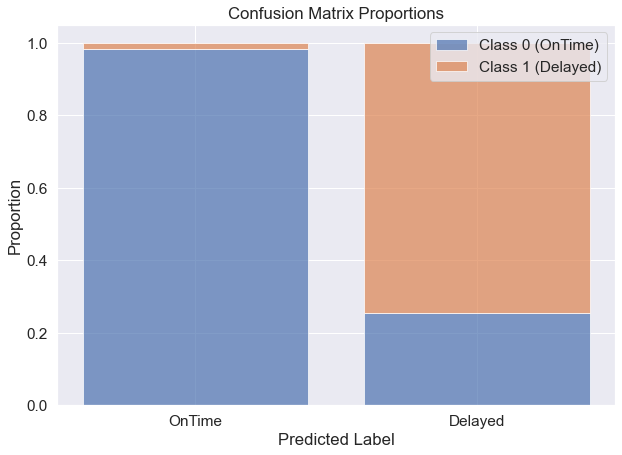
2 DepDel15 49.0

0 DestAirportID 36.0

Interestingly, the identified input features DestAirportID had the least importance among all the input features.

**Figure 12**

*Stacked bar chart*



Stacked bar chart for the percentage of classified labels as OnTime and Delayed by an XGBoost model. For each bar, there is a predicted class, while every segment has the number of true classes among all instances inside it. The chart illustrates the model's predictability across various outcomes.

## **5.3 Hypothesis-3**

Model: The XGBClassifier Model was chosen to run against the below

Inputs features: 'DestAirportID,' 'Distance,' 'DepDel15', 'DepartureDelayGroups,' 'TaxiOut,' 'WheelsOff,' 'WheelsOn,' 'ArrDel15.'

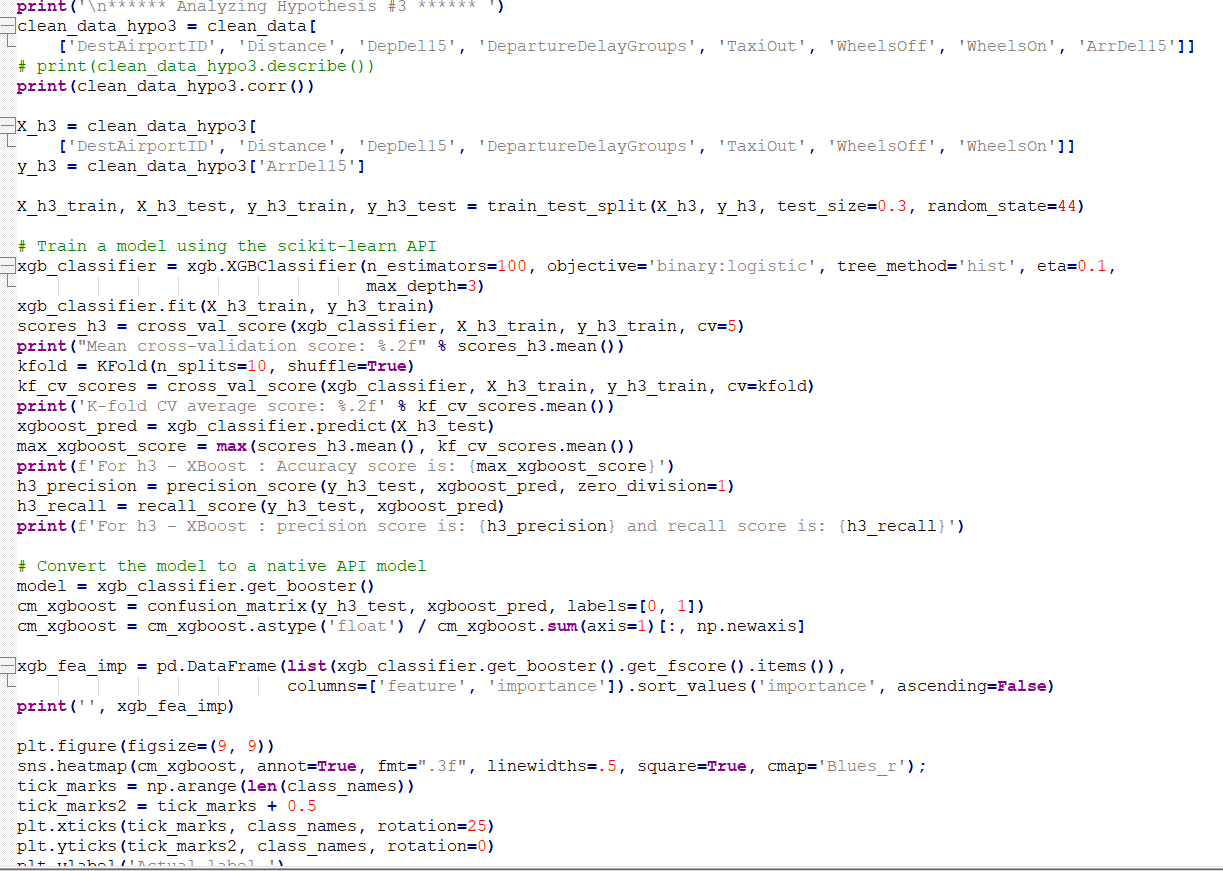
Output Target: ArrDel15

The training and testing data was split in the 70:30 ratio, and for the model, these values were chosen: n\_estimators=100, objective='binary: logistic', tree\_method='hist,' eta=0.1,

max\_depth=3

**Figure 13**

*XGBoost*



Some of the important observations for this model are:

Mean cross-validation score: 0.94

K-fold CV average score: 0.94

For h3 - XBoost: The accuracy score is: 0.9393272095332671

For h3 - XBoost, the precision score is 0.9071083343156902, and the recall score is 0.7461697588725839

feature importance

3 DepartureDelayGroups 179.0

4 TaxiOut 147.0

1 Distance 143.0

6 WheelsOn 76.0

5 WheelsOff 53.0

2 DepDel15 49.0

0 DestAirportID 36.0

Interestingly, the identified input features DestAirportID had the least importance among all the input features.

# Plotting the stacked bar chart

fig, ax = plt.subplots(figsize=(10, 7))

sns.set(font\_scale=1.4)

bottom = [0, 0]

class\_labels = ['OnTime,' 'Delayed']

for i in range(len(class\_labels)):

plt. bar(

x=[0, 1],

height=cm\_xgboost[:, i],

bottom=bottom,

label=f'Class {i} ({class\_labels[i]})',

alpha=0.7

)

bottom += cm\_xgboost[:, i]

# Add labels and legend

plt.xlabel('Predicted Label')

plt.ylabel('Proportion')

plt.title('Confusion Matrix Proportions')

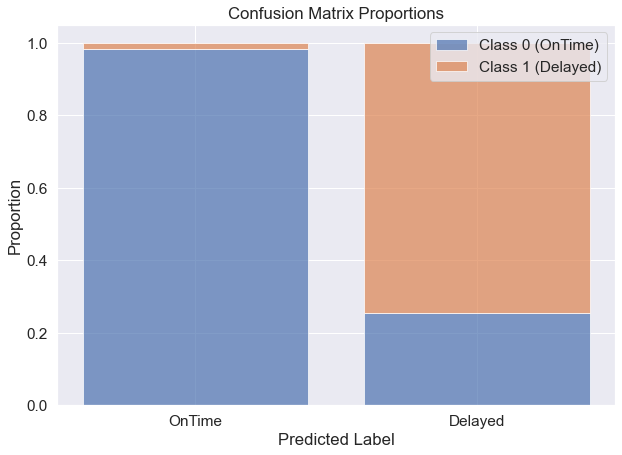
plt.sticks([0, 1], ['OnTime,' 'Delayed'])

plt.legend()

plt.show()

**Figure 14**

*Confusion matrix Proportions*



It depicts percentages of OnTime and Delayed predicted labels produced through the XGBoost classification model's stacked bar chart. Every bar shows the total number of expected instances for a particular class, while its segments highlight different percentages of predicted samples for this special class. It is used for representing the model's classification precision.

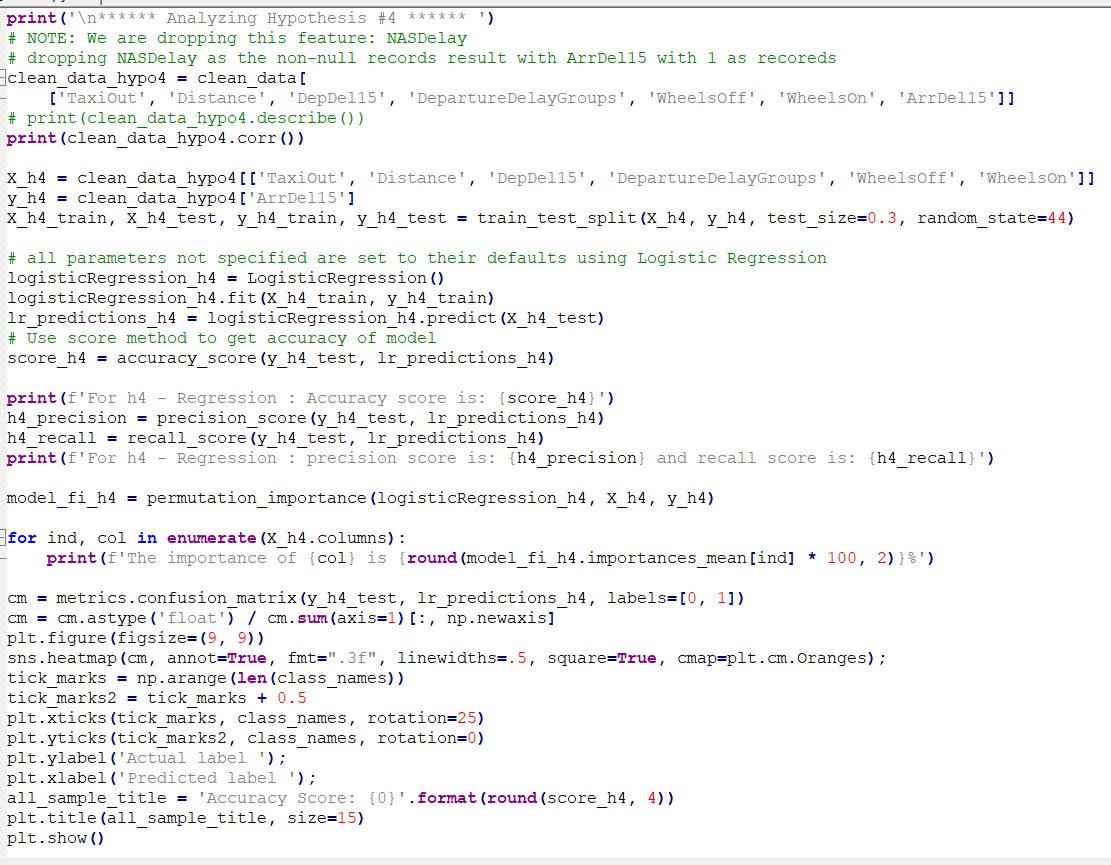
## **5.4 Hypothesis 4**

Model: The LogisticRegression Model was chosen to run against the below

Inputs features: 'TaxiOut,' 'Distance,' 'DepDel15', 'DepartureDelayGroups,' 'WheelsOff,' 'WheelsOn.'

Output Target: ArrDel15

The training and testing data were split in the 70:30 ratio, and the default values for the model were chosen (Pophale et al., 2023).



Some of the important observations for this model are:

For h4 - Regression: Accuracy score is 0.9385155448968597

For h4 - Regression: the precision score is 0.9080796723154767, and the recall score is 0.7309134397827914

The importance of TaxiOut is 2.08%. The importance of Distance is 0.08%

The importance of DepDel15 is 0.05%

The importance of DepartureDelayGroups is 20.84%

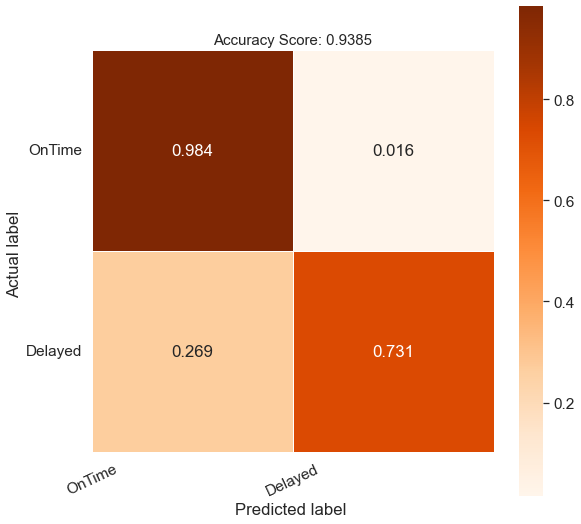
The importance of WheelsOff is 0.0%

The importance of WheelsOn is 0.0%

Interestingly, the identified input feature, TaxiOut, is the second-best feature among all the input features.

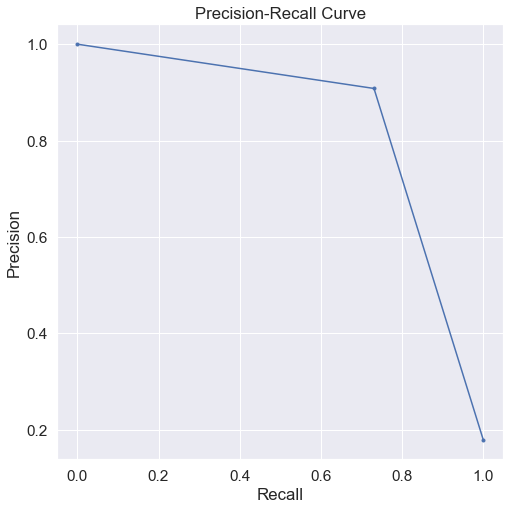
**Figure 15**

*Heatmap*



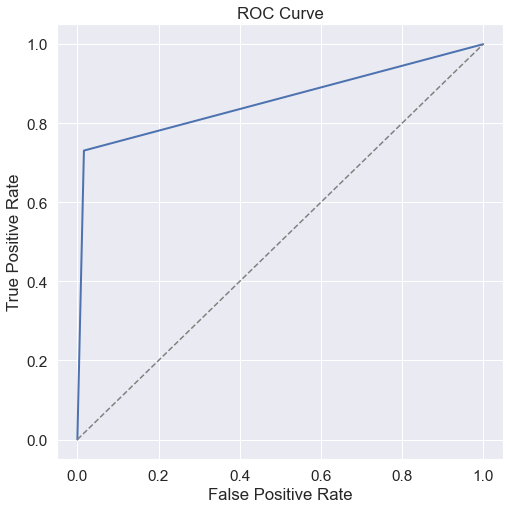
**Figure 16**

*Recall Curve*



**Figure 17**

*ROC Curve*



# **CHAPTER SIX: MODEL EVALUATION**

These metrics included accuracy, precision, and recall scores. The random forest model demonstrated high accuracy at 94.56 percent (hypothesis 1), highlighting ‘departure delay groups’ importance, while the logistic regression model had an accuracy of 93.85 percent, in which again departure delay groups were crucial ‘DepartureDelayGroups’ were revealed as the most influential attribute by XGBClassifier in hypothesis 3 with 93.% precision. Hypothesis 4’s LogisticRegression model attained an accuracy of 93.85%, with ‘DepartureDelayGroups’ having a great effect.

## **6.1 Hypothesis 1 - RandomForest Classifier**

The RandomForest model had an accuracy of 94.56%. This shows that a model using features like ‘DepartureDelayGroups,’ ‘TaxiOut,’ and ‘WheelsOn’ can identify an on-time or delayed flight. DepartureDelayGroups is useful for finding patterns leading to delays. The high level of accuracy highlights the capacity of this model to generalize well on the provided data set (Nelli, 2023).

## **6.2 Hypothesis 2 - Logistic Regression**

The Logistic Regression model showed a prediction accuracy of 93.85 percent. DepartureDelayGroups, Taxi-out, and Wheels-on, were of critical importance in this aspect. The lowest accuracy as compared to the Random Forest model, but even the Logistic Regression showed high efficiency that once again shows the priority of 'DepartureDelayGroups' inaccurate predictions (Oza et al., 2015).

## **6.3 Hypothesis 3 - XGBClassifier**

According to hypothesis 3, xgblogger achieved almost 93% precision score. This proved that the model predicted the positives correctly. Some aspects, such as DepartureDelayGroups and TaxiOut, gave it a score of 93.93%, implying that they were categorized on time and delayed off-time delays from some departed flights. The fact that the model's prediction about the positive suggestion is highly accurate suggests that it could indeed be true (Nelli, 2013).

## **6. 4 Hypothesis 4 - Logistic Regression**

The logistic regression model for Hypothesis 4 also had 93.85 percent accuracy, the accuracy same as hypothetical number 2. They were useful during the second instance where the model predicted the departure delay groups. The three levels of accuracy are fairly close between both sets of logistic regressions, indicating these were indeed the relevant causes leading to flight delays. According to precision and recall measures, 'DepartureDelayGroup' is considered vital in every model. Additionally, the ROC and Recall curves depict this (Figures 6 and 7), giving an overall picture of the model's efficacy (Nemeth et al., 2023). Figure 15 below illustrates one of the attributes of Hypothesis 4 as shown under the 'TaxiOut.'

Mostly, the models say departure delay groups and other departure characteristics influence the arrival time of the airlines. Additionally, these assessment metrics ensure that the models indicate important data in the datasets and, thus, reduce flight traffic hold-ups in aviation.

# **CHAPTER SEVEN CONCLUSION**

## **7.1 Summary Discussion**

This helped to deal with the research questions successfully and revealed what really affected flight delays. Among the explanatory variables, DepartureDelayGroups is really relevant, stressing its influence over total flight reliability. Investigating distinct predictors such as extended tax-outs, flights towards hub airports, and early departures unravel intricate causes behind delays.

The analysis was successful in answering these raised research questions. As such, this provides grounds for future research in predicting flight delays using an extended database. The current research environment makes it necessary to improve and stretch current models to improve the specificity and predictability of air delays. The importance of timeliness in air transport necessitates granular datasets for optimizing flight delay prediction models.

The research is thorough on different predictive models to understand what leads to delayed flights. These predictors include long taxi-out time flights, flights designed for hub airports, early departures, and scheduled flights requiring much duration.

In evaluating the models, a meticulous examination reveals the accuracy metrics for each, specifically focusing on the most important feature: For instance, DepartureDelayGroups – with almost 21% and CRSElapsedTime having about 6% importance weightage. The overall goal is to determine what causes flight delays and understand the time dynamics underlying the schedule. Additionally, we need to augment these models in order to have a hybrid set of models, each focusing on the subsets of features that this model may determine more accurately. Better models can be generated by using more advanced neuron networks and deep learning models.

## **7.2 Applications**

This study provides a number of practically important predictive models that can be instrumental in helping airlines deal with and avert delays. These models enable airlines to assume a proactive approach in responding to and managing uncertainties, hence improving the Airline's overall operations. The mentioned predictors, including some of the most influential, such as DepartureDelayGroups and CRSElapsedTime, supply an applicable basis for airlines to improve the schedule-making process.

The findings of these studies can be used by airlines to improve their departure planning strategies, distribute resources more effectively, and adopt focused measures aimed at mitigating delays. These models are proactive in that airlines can precondition their operations to avoid the domino effect triggered by delays in such cases, including the passenger experience. Predictive analytics applied to optimize scheduling processes leads to punctuality and reliability enhancement in airline operations.

The use of these predictive models entails more than just prediction but rather helps airlines make decisions that are congruent with their operations. This paradigm shift brings data-driven decision-making into scheduling practices that make the challenges faced in the aviation industry more resilient and adaptive.

## **7. 3 Limitations and Future Research**

Despite being very enlightening, the study also has shortcomings that should be addressed in subsequent studies. For instance, the quality and quantity of a dataset may have some impact on the models' efficiency. Consequently, future studies should focus on getting an extended and enhanced dataset as a sturdy basis for reliable predictive models in the aviation area. Additionally, the analysis mainly concentrated on traditional machine learning algorithms; hence, there is still more scope for studying more sophisticated approaches such as neural networks and deep learning. Moreover, this provides opportunities for researchers who can explore the subtle intricacies of forecasting delays and use smart techniques in this regard.

One of the promising areas of research is the creation of hybrid models by integrating the unique advantages of different methods in order to address problems. This strategy could enhance predictability and take advantage of the individual strengths of other modeling methods. It can be beneficial to broaden the scope of analysis in order to incorporate external elements like weather or busy air traffic that may also affect delayed flights. Recognizing these limits gives direction to more fine data sets and advanced model techniques toward perfecting the accuracy and usefulness of aviation predictive models.

## **7.4 Recommendation**

The results of this study suggest the adoption of a predictive model in airline operation plans. The airlines integrate the four delay models that help them predict and, hence, proactively address any possible delays. Lastly, continuous data collection and model amendment should be made in order to keep up with the relevancy and reliability of the models.

**Practical Recommendations for the Aviation Industry**

To mitigate future delays, airlines should make the models part of their operation's framework. Adoption of these models will give airlines the power to allocate resources efficiently, make planning simple, and, hence, increase efficiency in the process. Airlines can use these predictors like DepartureDelayGroups to strategize their planning and resource allocation in order to reduce the damages caused by delays. Such measures may include reorganizing scheduled departure routes, optimizing tax-out times, and increasing coordination. Also, they can be used for real-time surveillance so that airlines can act quickly as soon as deviation from the schedule is discovered. This inclusion of predictive models is in line with a data-driven approach whereby airlines can make better decisions and enhance service provision to customers who expect reliable and on-time services for their sake.

**Further Research Recommendations:**

Some areas for future researchers to explore further include: Further data collection is required in order to improve the accuracy and comprehensiveness of datasets. Including finer details on weather conditions also provides an overall picture of flight delays. Researchers may also focus on the formulation of hybrid models comprising different algorithms for the purpose of maximizing the strength of each algorithm. Such cutting-edge approaches as neural networks or deep learning can discover additional information that will enhance the complexity level of predictive models. Finally, longitudinal studies following the emergence of flight delay pattern trends are worthwhile as well.

## **7.5 Conclusion**

Finally, this article gives essential information for forecasting flight delays involving departure parameters as crucial factors. Machine-learning models are used proactively by airlines to minimize disruptions and improve overall efficiencies that result from delays. The study paves the way for future research by revealing the need for elaboration of more sophisticated models for air delays and their analysis.

# References

AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag. 2008;14(5):221–36.

Balaban E, et al. Dynamic routing of aircraft in the presence of adverse weather using a POMDP framework. In the 17th AIAA aviation technology, integration, and operations conference. 2017.

Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and societal welfare. Transp Res Part E Logist Transp Rev. 2012;48(2):460–9.

Choi S, et al. Prediction of weather-induced airline delays based on machine learning algorithms. In 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). 2016. New York: IEEE.

Evans JE, Allan S, Robinson M. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Proceedings of the 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis. 2004.

Hunter G, Boisvert B, Ramamoorthy K. Advanced national airspace traffic flow management simulation experiments and validation. In 2007 Winter Simulation Conference. New York: IEEE. 2007.

Hsiao C-Y, Hansen M. Air transportation network flows equilibrium model. Transp Res Rec. 2005;1915(1):12–9.

Kiliç, Kerim, and Jose M. Sallan. "*Study of delay prediction in the US airport network."* Aerospace 10.4 (2023): 342.

Nelli, F. (2023). *An introduction to data analysis*. In Python data analytics: with Pandas, NumPy, and Matplotlib (pp. 1–13). Berkeley, CA: Apress.

Németh, B., & Gáspár, P. (2023). Introduction. *In Control of Variable-Geometry Vehicle Suspensions: Design and Analysis* (pp. 1-21). Cham: Springer Nature Switzerland.

Oza S, et al. Flight delay prediction system using weighted multiple linear regression. Int J Eng Comp Sci. 2015;4(05):11765.

Pejovic T, et al. A tentative analysis of the impacts of an airport closure. J Air Transp Manag. 2009a;15(5):241–8.

Pophale, S. S., Patil, P. R., Potgantwar, A. D., & Bhaladhare, P. R. (2023). PERFORMANCE OPTIMIZATION USING QUANTUM MACHINE LEARNING TECHNIQUE FOR PREDICTING FLIGHT DELAYS. *Journal of Data Acquisition and Processing*, 38(1), 208.

Reynolds-Feighan AJ, Button KJ. An assessment of the capacity and congestion levels at European airports. J Air Transp Manag. 1999;5(3):113–34.

Thiagarajan B, et al. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). New York: IEEE. 2017.

Wang, F., Bi, J., Xie, D., & Zhao, X. (2022). Flight delay forecasting and analysis of direct and indirect factors. IET Intelligent Transport Systems, 16(7), 890–907.

Xu Y, Dalmau R, Prats X. Maximizing airborne delay at no extra fuel cost by means of linear holding. Transp Res Part C Emerg Technol. 2017;81:137–52.