INTRODUCTION

Flight is an essential mode of transportation in this century, allowing people to travel across far distances in a short amount of time. Several industries have been blooming along with airline industries, and tourism is one of the key players.Also, despite the growth of aviation industries, operational inefficiencies still need to be addressed, and one of the prominent ones is flight schedule delays.

Several factors lead to flight delays. The Bureau of Transportation Statistics (BTS) of the United States of America has grouped the delay factors into five categories: air carrier, extreme weather, previous late flight, security, and others (BTS, 2021).

MOTIVATION OF WORK

Flight delays are significant concerns in aviation industries, leading to revenue loss, fuel loss, and customer dissatisfaction. It creates fear among passengers taking a connecting flight, whereby the delay from the first flight could potentially cause them to miss the subsequent flight. Therefore, this scenario is a factor of motivation for this study. With a reliable method to predict flight delays, the event mentioned in the previous context could either be prevented or better managed.

PROBLEM STATEMENT

The ability to predict a delay in flight can be helpful for all parties, including airlines and passengers. This study explores the method of predicting flight delay by classifying a specific flight as either delay or no delay. From the initial review, the flight delay dataset is skewed. It is expected since most airlines usually have more non-delayed flights than delayed ones. Hence, this study compares different methods to deal with an imbalanced dataset by training a flight delay prediction model.

OBJECTIVES

The objectives of this study are:

1. To develop machine learning models that classify flight outcomes (either delayed or not delayed) with selected features.
2. To evaluate the performance of different machine learning models.

DATA SOURCES

Scraped data from different URLS and merged them into one. Have cleaned the dataset as well. For scraping following functions have been performed.

1) Top_search

This function will receive a URL containing Tel Aviv airport's data, and the function returns the top 10 countries to which the most flights from Israel have departed.

2) scraping_top:

This function will get a URL containing a page of a country, all the flights that have left it in the last few days, and future

flights. We pull the information from flights from previous days with a limit up to that day excluding so that we can get information about the landing. After we have received the information and entered it into the data frame, we will go through it and for each flight, we will take the departure destination, concatenate it to the URL appropriately, and send it to the weather_departures function.

3) weather_departures:

This function retrieves weather information from a specified URL, the URL it receives forwards it to the page of the country from which the flight departed and takes the data from the country's weather. and returns them

4) delays:

This function will receive a URL that contains the information for all flights that can match the flight number contained in the URL, the function will find the correct flight, according to the date and time of the flight and will return the take-off and landing times according to the schedule and also the times that actually happened

For cleaning following functions have been performed

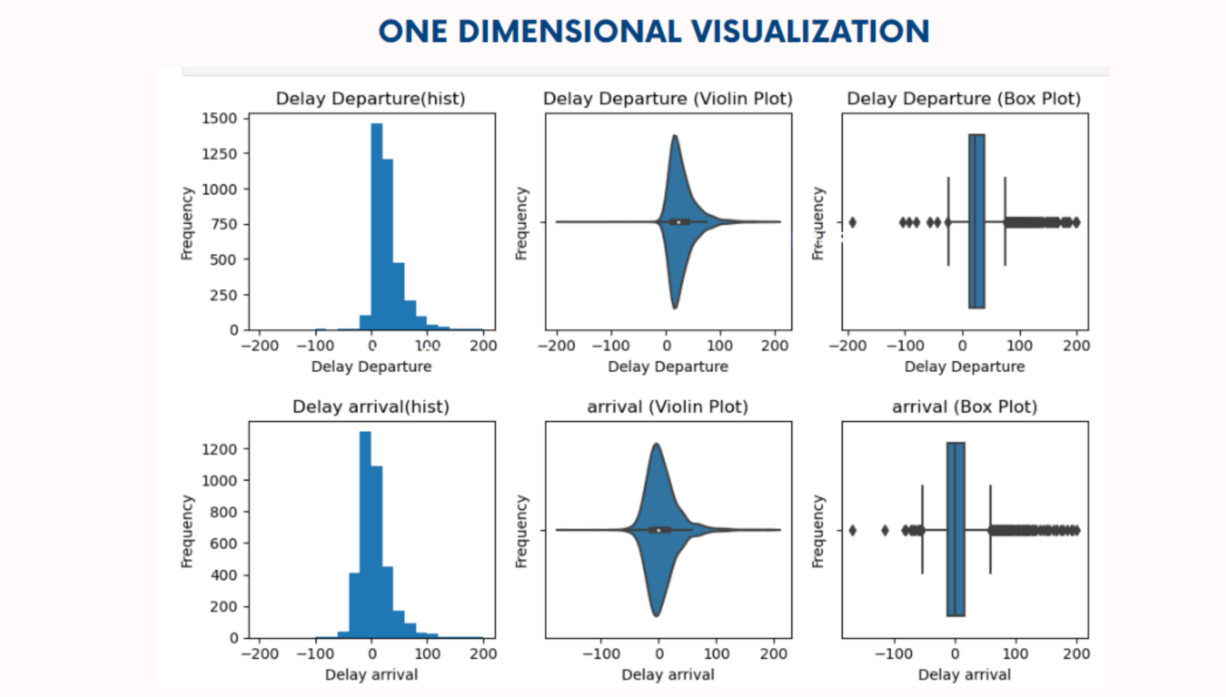
1. In cases where we have an empty cell in the columns -scheduled_departures, FLIGHT, TEMPERRATURE_departures.
2. In cases where the character "-" indicates an empty cell in the columns - actual_outputs, temperature_outputs.
3. In cases where we have the string "calm" and not a number in the column - WIND_departures
4. In cases where we have the string "variable" and not a number in the column - DIRECTION_departures

Cleaned Data Link: flight_Data

DATA VISUALIZATION AND EDA

Visualized the data for two main features flight delay for both the arrival and departure and have plotted them for different visualization types:

1. Visualization for a single variable

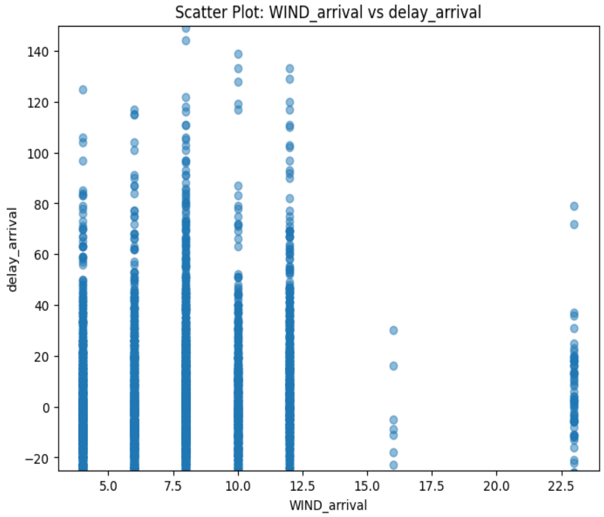


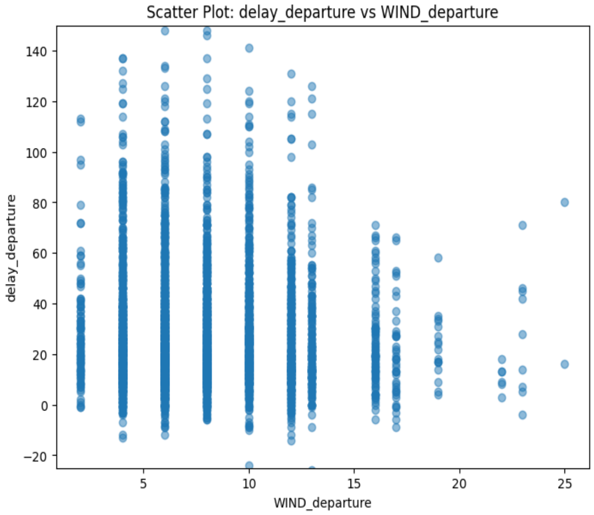
ONE DIMENSIONAL VISUALIZATION

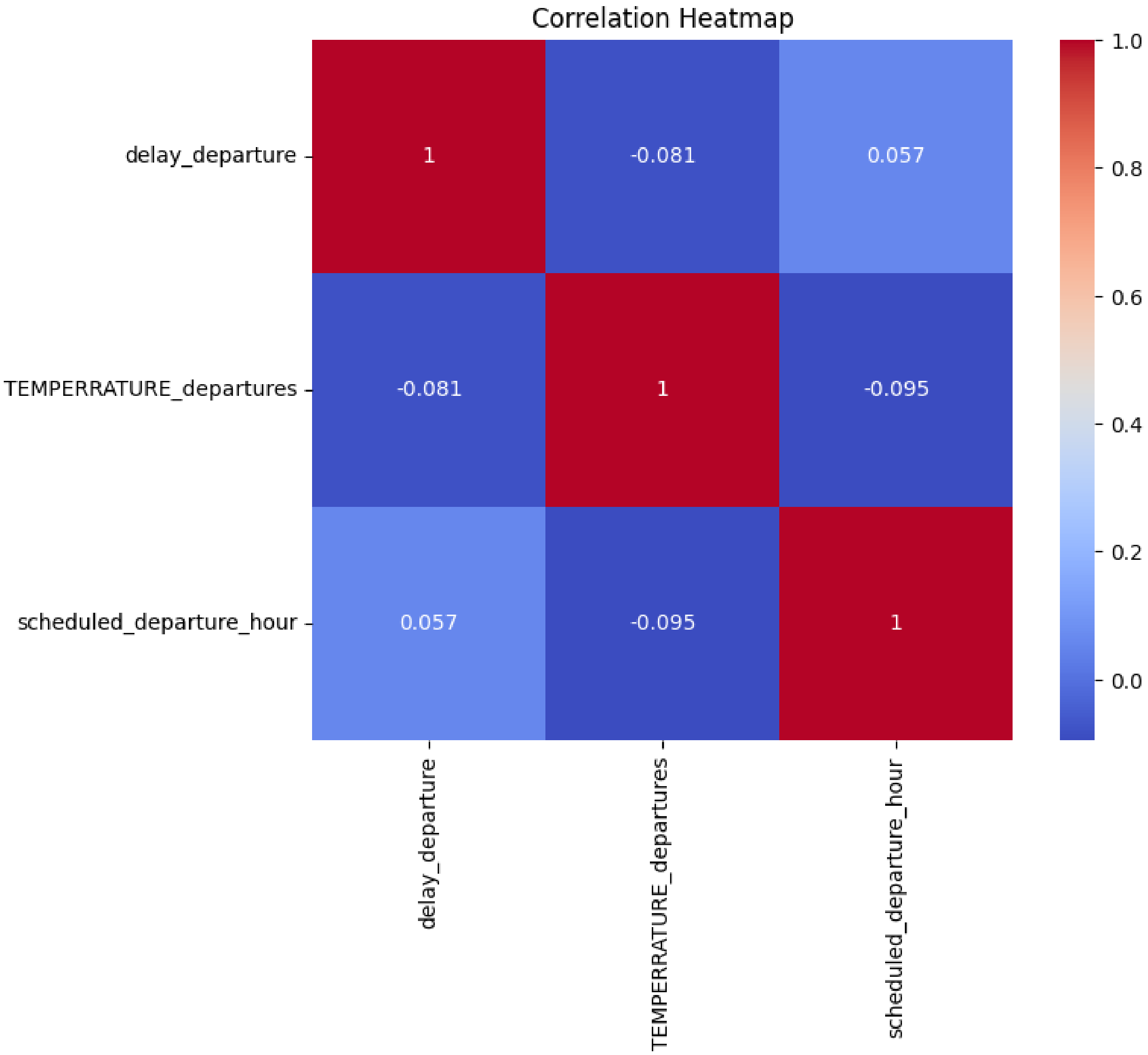
2. Two-dimensional visualization

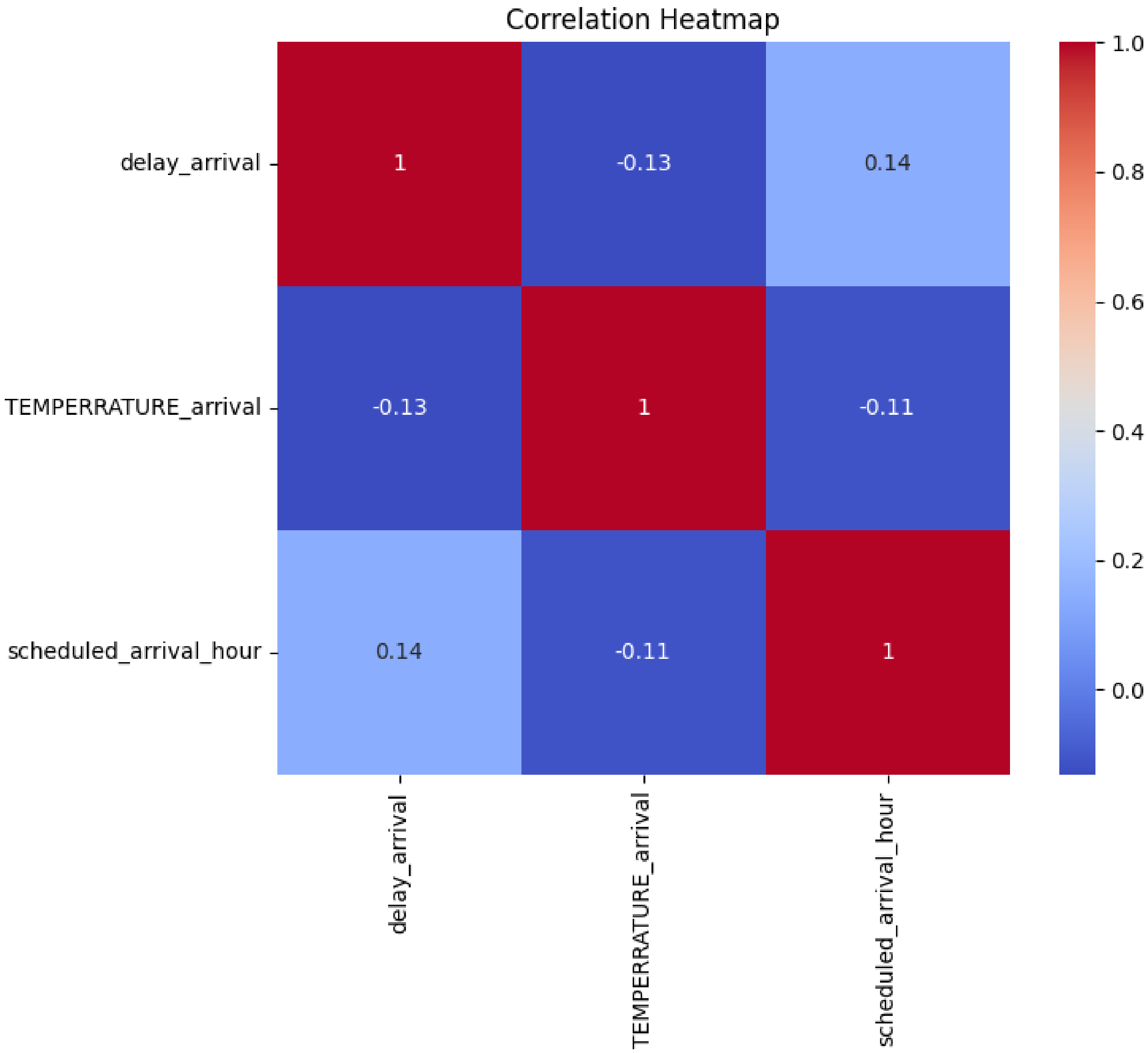
   In this First I have plotted Scatter plots between:

   a) Delay Arrival and Wind Arrival
   b) Delay Departure and Wind Departure
   c) Correlation Heatmap for all the delay variables
   d) Correlation Heatmap for all the arrival variables

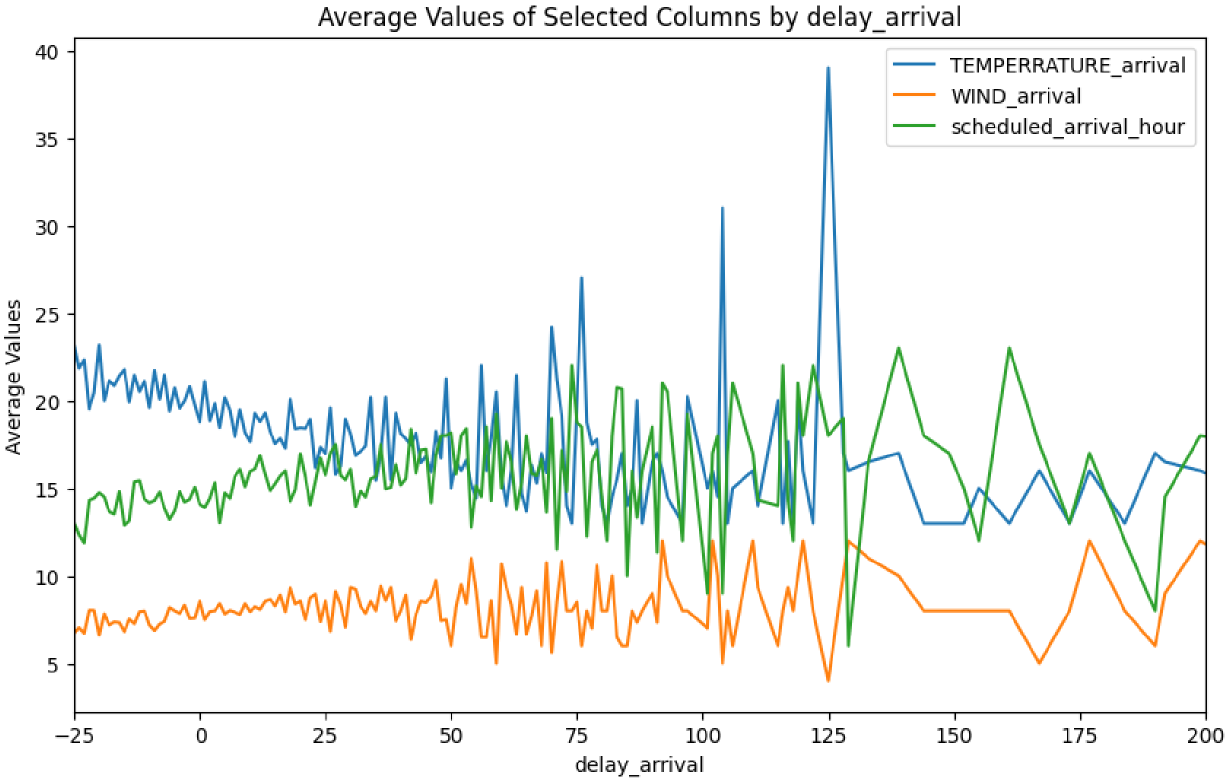

Scatter Plot: WIND_arrival vs delay_arrival

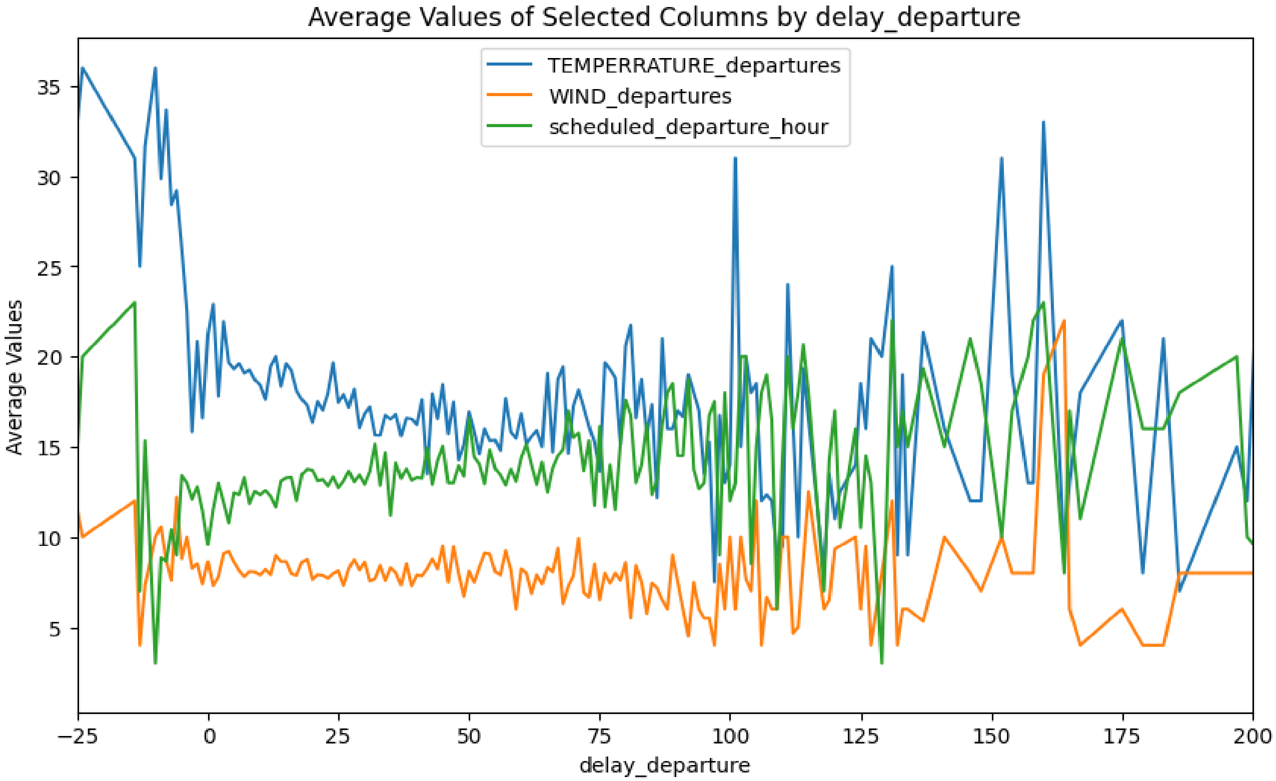

Scatter Plot: delay_departure vs WIND_departure

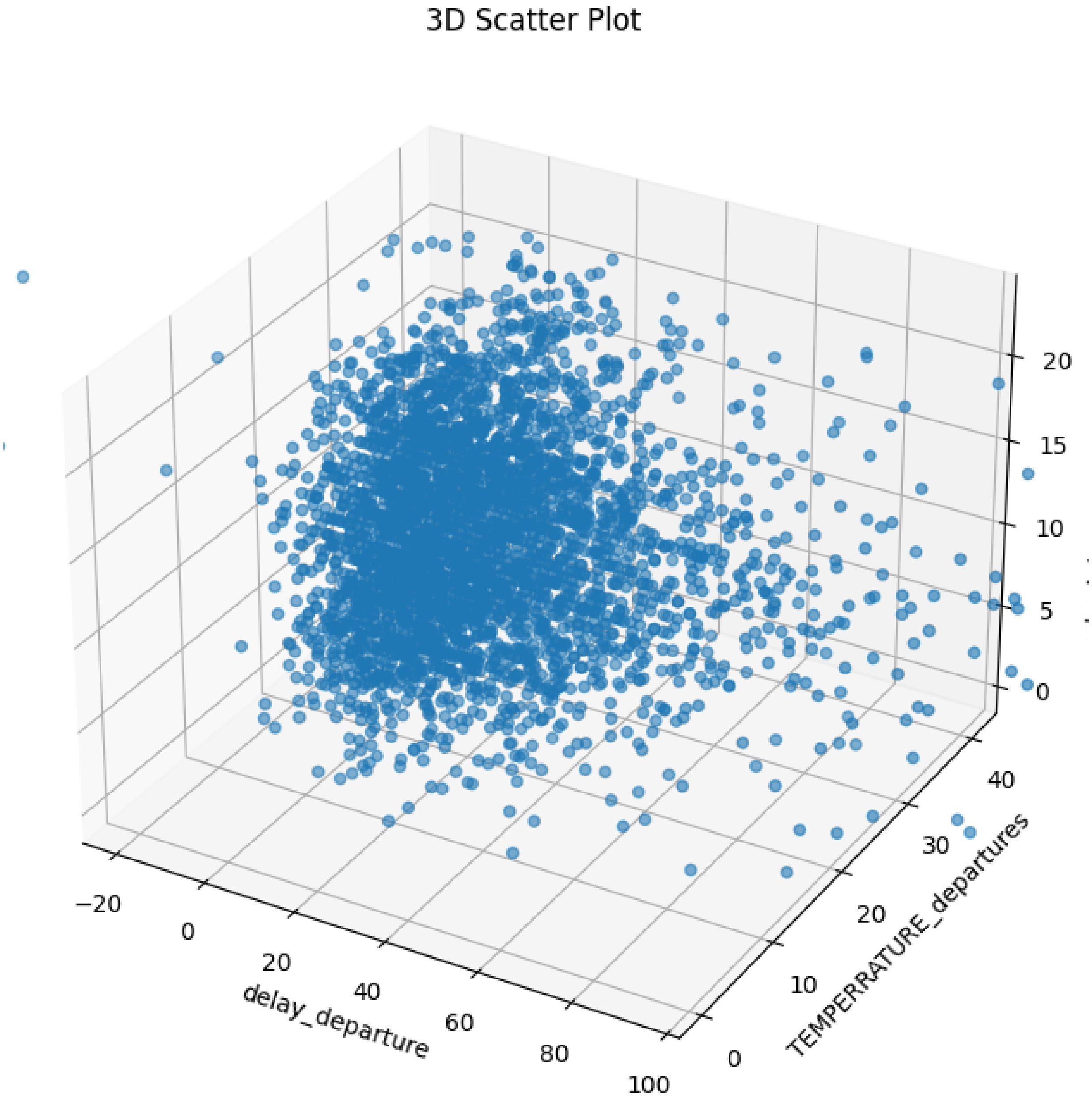

Correlation Heatmap


Correlation Heatmap
delay_arrival
TEMPERRATURE_arrival
scheduled_arrival_hour
1
-0.13
0.14
-0.13
1
-0.11
0.14
-0.11
1
delay_arrival
TEMPERRATURE_arrival
scheduled_arrival_hour
1.0
0.8
0.6
0.4
0.2
0.0

3. Multidimensional visualization






3D Scatter Plot
delay_departure
−20
0
20
40
60
80
100
TEMPERRATURE_departures
0
10
20
30
40
0
5
10
15
20

MODELS EMPLOYED

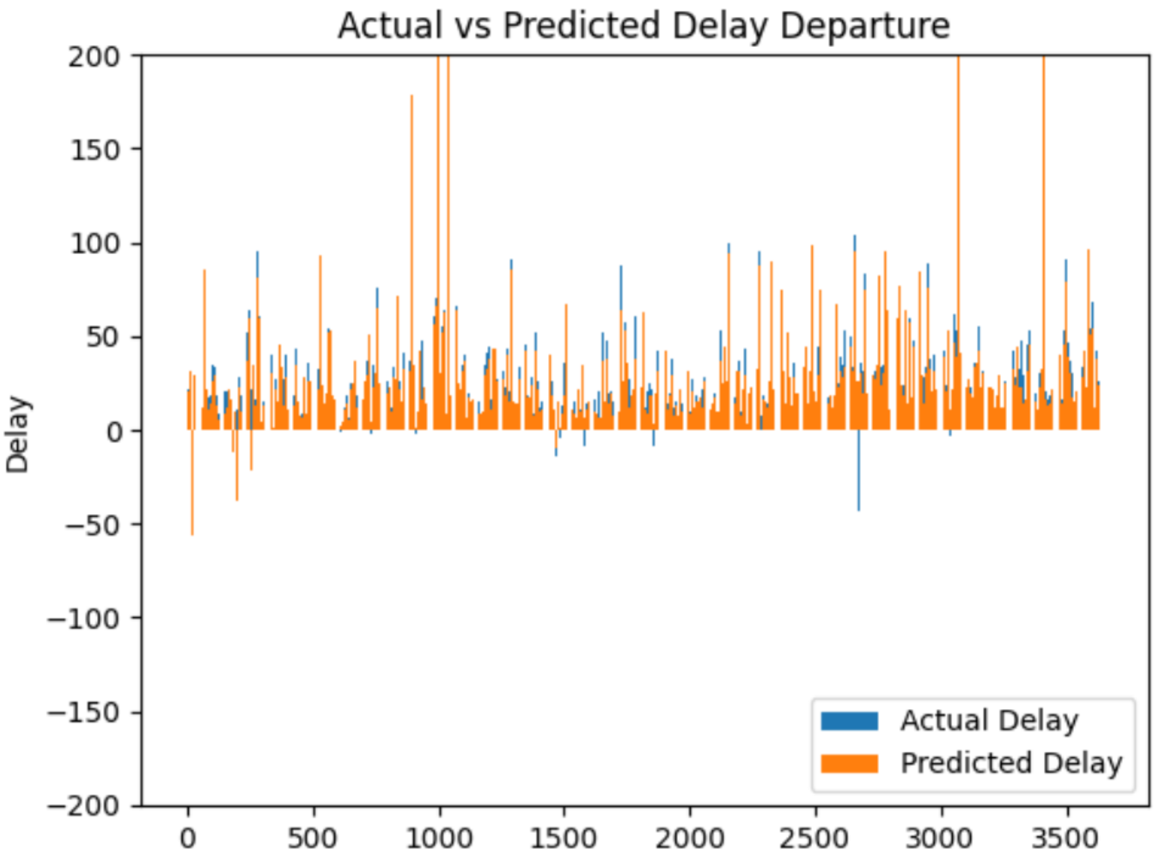
The following table shows all the models with their accuracies

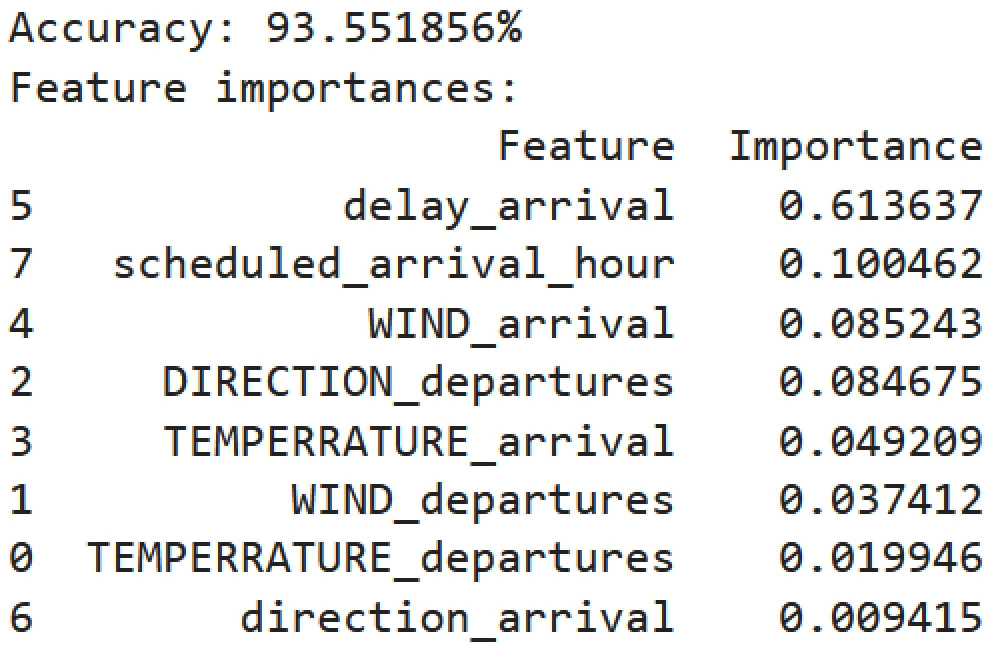
	Accuracy	
MODELS	Delay Departure	Delay Arrival
Linear Regression	71.83	56.96
Random Forest	77.65	84.36
Gradient Boost	84.33	86.59

Based on the above three algorithms Gradient Boost has shown promising results compared to other two, so have tried to improve the Gradient Boost model further

GRADIENT BOOST IMPROVISED

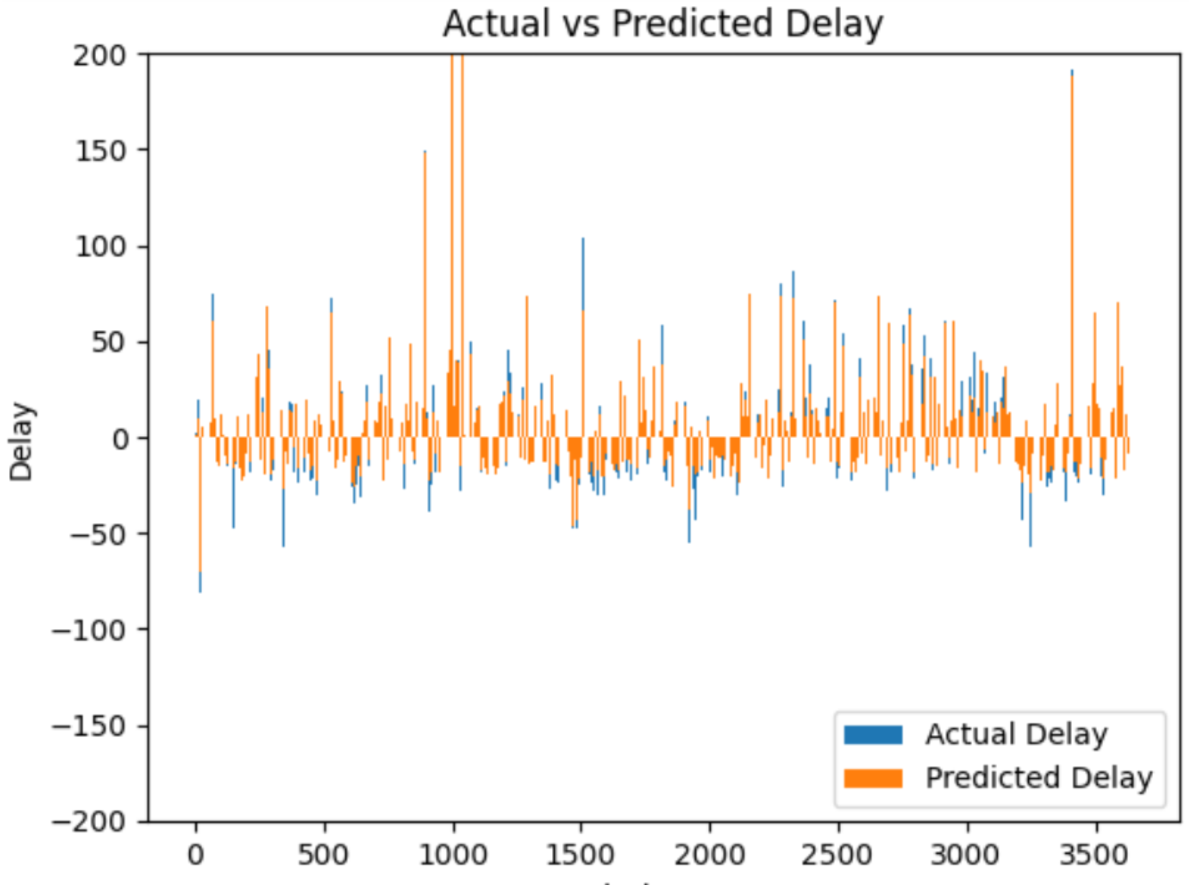
1. Delay Departure

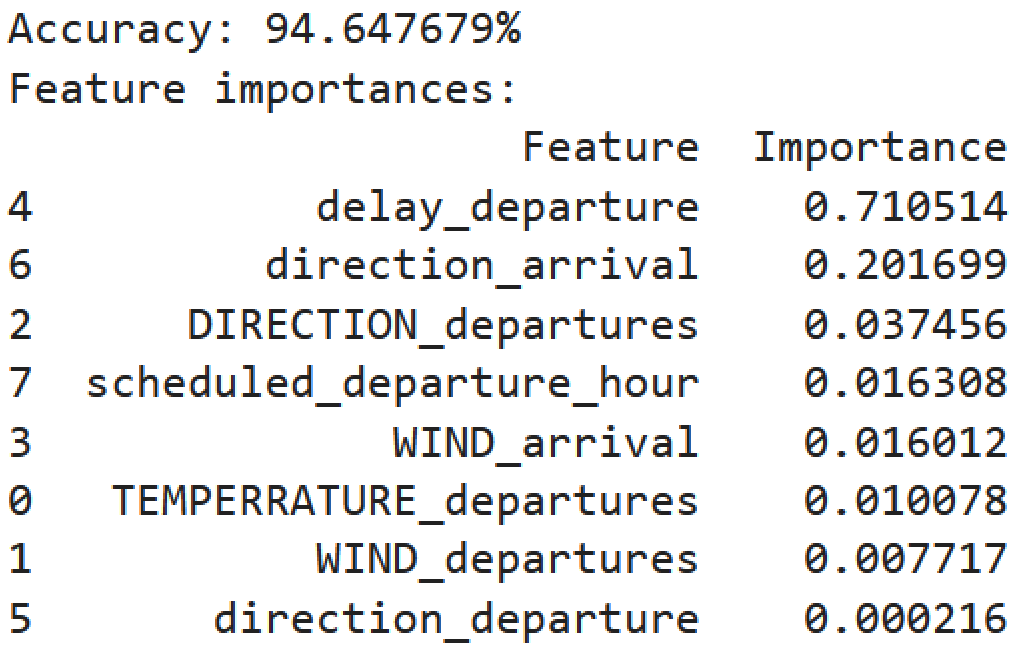




```
Accuracy: 93.551856%
Feature importances:
                     Feature  Importance
5              delay_arrival    0.613637
7     scheduled_arrival_hour    0.100462
4               WIND_arrival    0.085243
2       DIRECTION_departures    0.084675
3       TEMPERRATURE_arrival    0.049209
1            WIND_departures    0.037412
0   TEMPERRATURE_departures    0.019946
6          direction_arrival    0.009415
```

2. **Delay Arrival**





```
Accuracy: 94.647679%
Feature importances:
                    Feature  Importance
4           delay_departure    0.710514
6         direction_arrival    0.201699
2      DIRECTION_departures    0.037456
7  scheduled_departure_hour    0.016308
3              WIND_arrival    0.016012
0    TEMPERRATURE_departures    0.010078
1           WIND_departures    0.007717
5       direction_departure    0.000216
```

CODE: **EDA_Flight_Prediction_1.ipynb**