Flight Delay Prediction

Milestone: Project Report

Group 5
Abhiram Desai
Jwalit Shah

(425-864-3723)
(732-209-2041)

desai.abhi@northeastern.edu
shah.jwa@northeastern.edu

Percentage of effort contributed by Abhiram Desai - 100%
Percentage of effort contributed by Jwalit Shah - 100%

Signature of Abhiram Desai

Signature of Jwalit Shah

Submission Date - 24 June 2022

Problem Setting

In today’s world where air travel is getting costlier day by day, a lot of people who book airline tickets online face a plethora of issues in case a flight gets delayed or canceled due to any reason. Apparently, some of these reasons like delays due to aircraft maintenance, cleaning, baggage loading or fueling are within the airline's control. While others like delays due to weather issues,air traffic volume,security breach or inoperative screening equipment at airports don’t fall into that category. In either case, the customers are the one to suffer as they go through a lot of stress, anxiety and fear as they’re unable to reach their respective destinations in time.

Problem Definition

The primary aim of this analysis was to identify the best classification model along with the predictors, which accurately predict the flights which are delayed at the destination airport and give us an estimation about the delayed time. This estimation in turn, would help customers make better decisions while booking tickets online, if they know in advance about the flights which are likely to get delayed and the duration of delay. The prediction about the flight delay at a given destination airport is specifically for the month of January in upcoming years as the data is for January 2019 only.

Data Source

The flight delay prediction dataset has been taken from kaggle, an online website for data science and machine learning practitioners. This data is collected from the Bureau of Transportation Statistics, Govt. of the USA. This data is open-sourced under U.S. Govt. Works.
(https://www.kaggle.com/divyansh22/flight-delay-prediction)

Data Description

The aforementioned dataset contains 583,986 records and 22 attributes. Out of the 22 attributes, 1 is ‘unnamed’ and does not contain any record. Whereas, “ARR_DEL15” is the target variable, which indicates if a given flight is delayed(1) or not delayed(0) at its destination.

DAY_OF_MONTH	Day of Month
DAY_OF_WEEK	Day of Week starting from Monday
OP_UNIQUE_CARRIER	When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example PA, PA (1), PA (2) etc.
OP_CARRIER_AIRLINE_ID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT Certificate regardless of its code, name or holding company/organization.
OP_CARRIER	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique.
TAIL_NUM	Tail Number
OP_CARRIER_FL_NUM	Flight Number
ORIGIN_AIRPORT_ID	An identification number assigned by US DOT to identify a unique airport.

ORIGIN_AIRPORT_SEQ_ID	An identification number assigned by US DOT to identify a unique airport at a given point of time.
ORIGIN	Origin Airport
DEST_AIRPORT_ID	An identification number assigned by US DOT to identify a unique airport.
DEST_AIRPORT_SEQ_ID	An identification number assigned by US DOT to identify a unique airport at a given point of time.
DEST	Destination Airport
DEP_TIME	Actual Departure Time (local time: hh mm)
DEP_DEL15	Departure Delay Indicator, 15 Minutes or More (1=Yes, 0=No)
DEP_TIME_BLK	Departure Time Block, Hourly Intervals
ARR_TIME	Actual Arrival Time (local time: hh mm)
ARR_DEL15	Arrival Delay Indicator, 15 Minutes or More (1=Yes, 0=No)
CANCELED	Canceled Flight Indicator (1=Yes, 0=No)
DIVERTED	Diverted Flight Indicator (1=Yes, 0=No)
DISTANCE	Distance between airports (miles)

Data Exploration

After cleaning the data, the attributes were examined using exploratory analysis. Data Exploration, facilities the comprehension of insights, which in turn helps in identifying trends and patterns in a straightforward manner. After determining the unique number of airlines, the number of delayed flights as a percentage of total number of flights were plotted as a stacked bar chart.

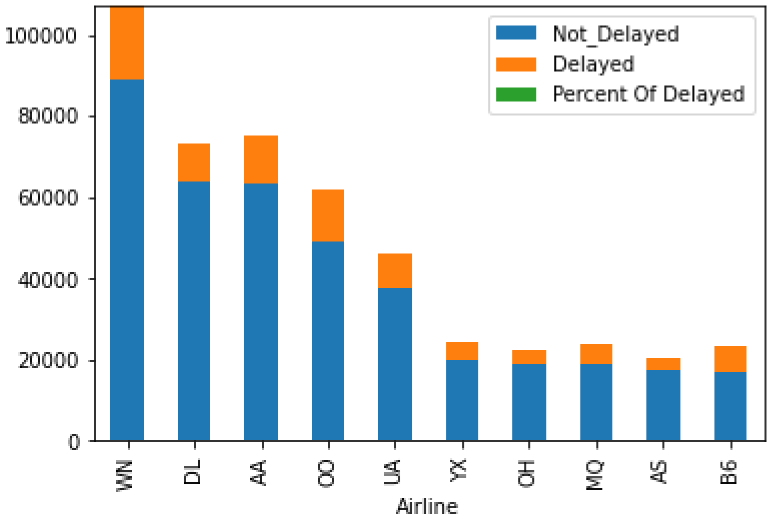


Figure 1- Number of delayed flights v/s Airlines

Similarly, the number of delayed flights, a percentage of flights as a percentage of total number of flights were plotted, segregating them as per their origin airports.

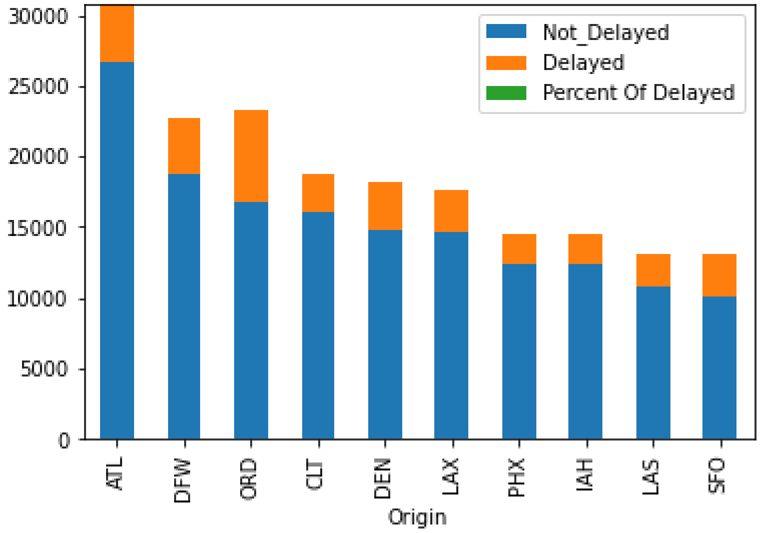


Figure 2- Number of delayed flights v/s Origin Airport

The graph below shows the number of delayed flights classified as per their departure time blocks.

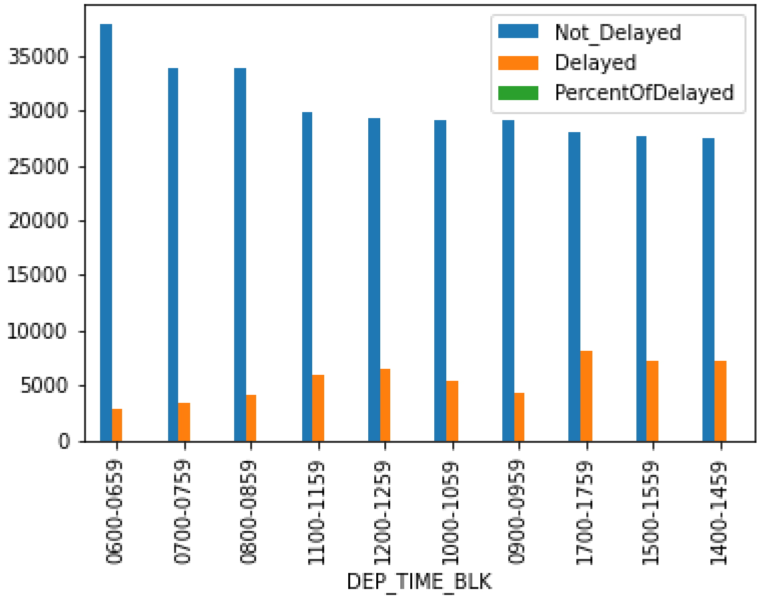


Figure 3- Number of delayed flights v/s departure time blocks

The scatter plot is created in order to check if there is a relationship between distance and delay. From the plot generated, it is imperative that there is not a clear relationship between distance and getting delays.

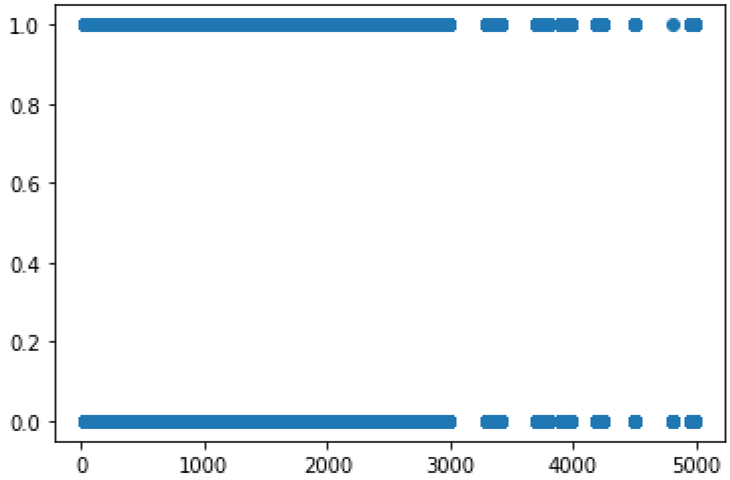


Figure 4- departure delay v/s distance

The number of flights delayed, are grouped as per the day of the week. From the graph below, it can be said that most flights are delayed on the 4th day of the week.

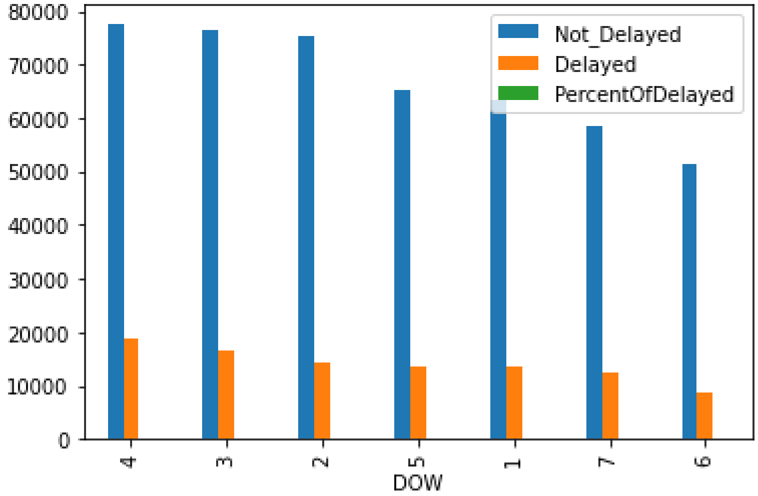


Figure 5- Number of delayed flights v/s day of the week

The plots below show the number of delayed flights in the descending order for their respective origin and destination airports. As per these plots,the Atlanta airport has the most number of delayed plots, both in terms of departure delay as well as arrival delay.

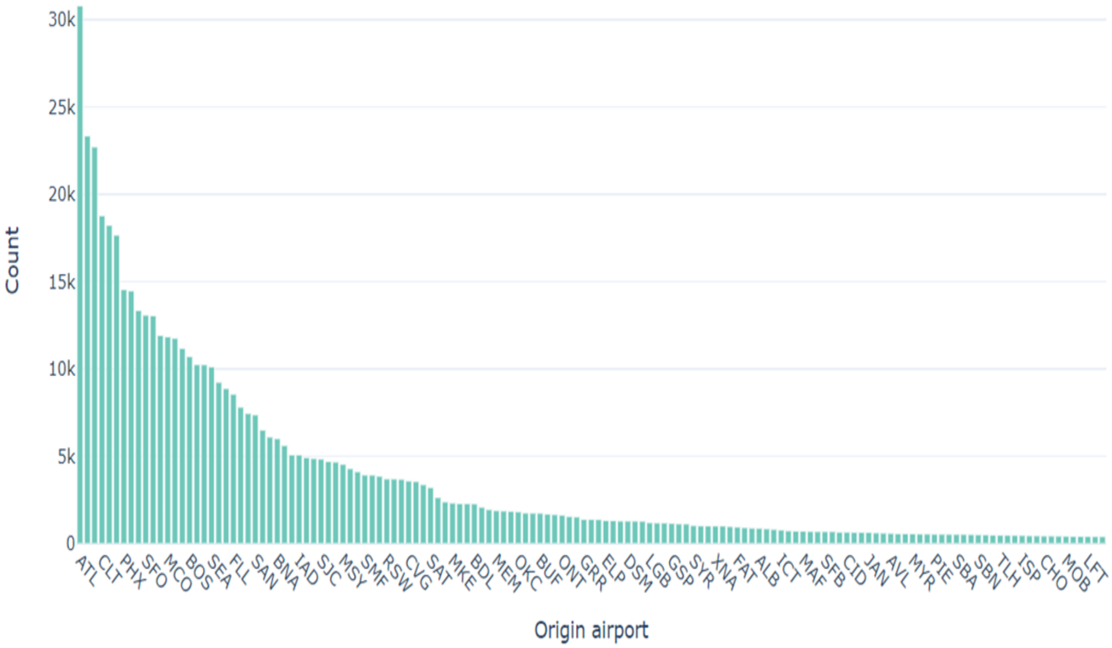


Figure 6- Number of delayed flights (departure) v/s Origin Airports

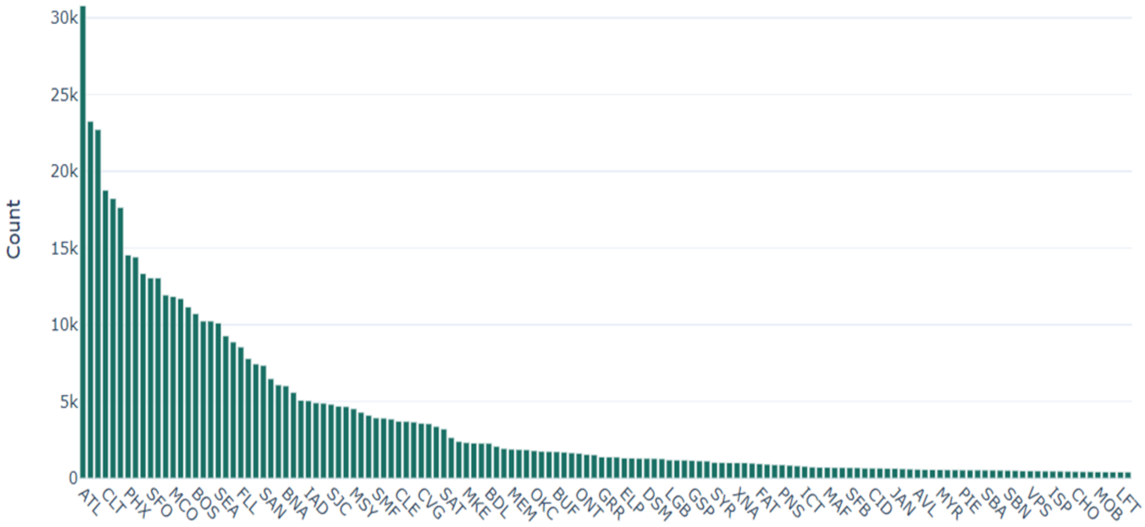


Figure 7- Number of delayed flights (arrival) v/s Destination Airports

The donut charts represent the top 10 airlines with the most number of delayed flights with respect to arrival and departure time.

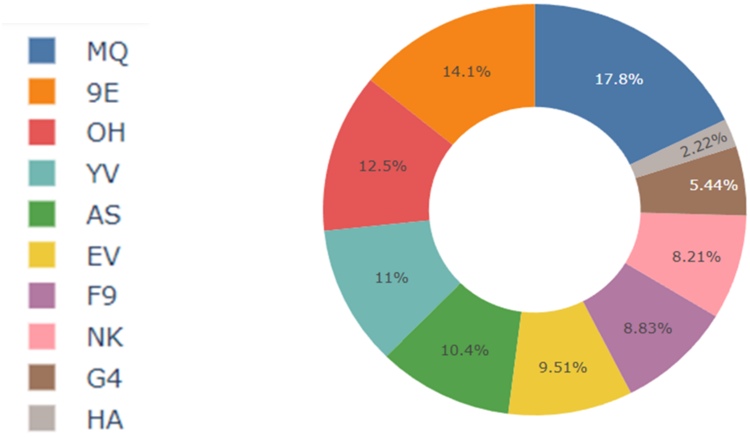


Figure 8 - Top 10 airlines with most delayed flights (departure)

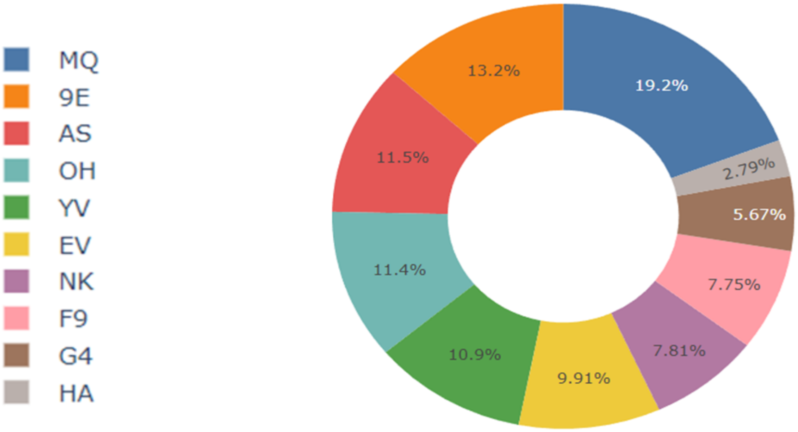


Figure 9- Top 10 airlines with most delayed flights (arrival)

The pie charts represent the top 10 airports with the most number of delayed flights with respect to arrival and departure time.

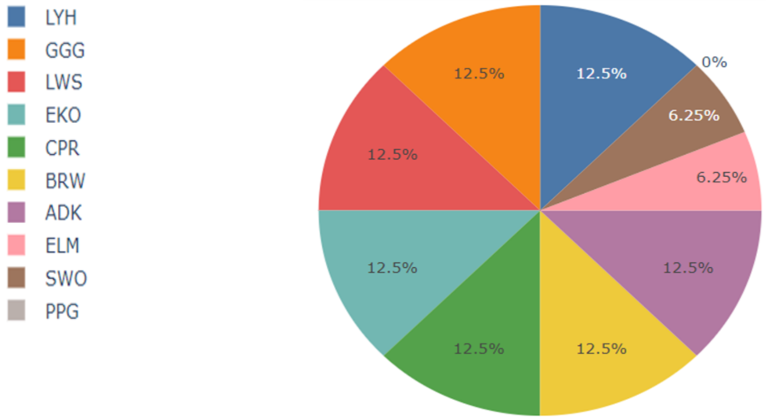


Figure 10- Top 10 airports with most delayed flights (departure)

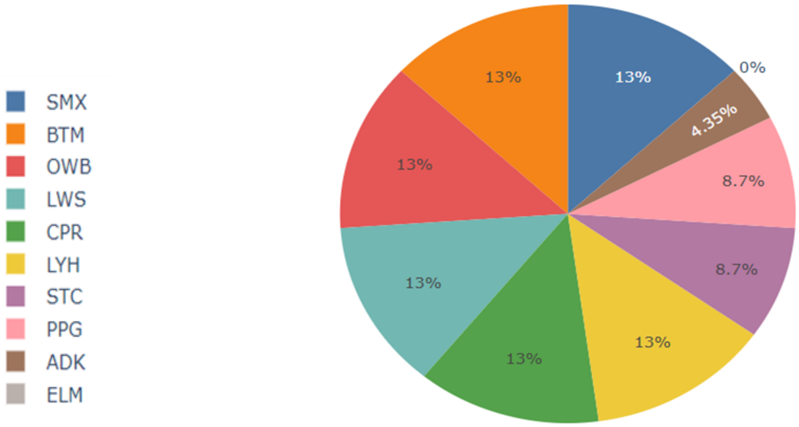


Figure 11- Top 10 airports with most delayed flights (arrival)

The concentration of total number of delayed as well as non delayed flights corresponding to arrival delay and departure delay is depicted in the visualization below.

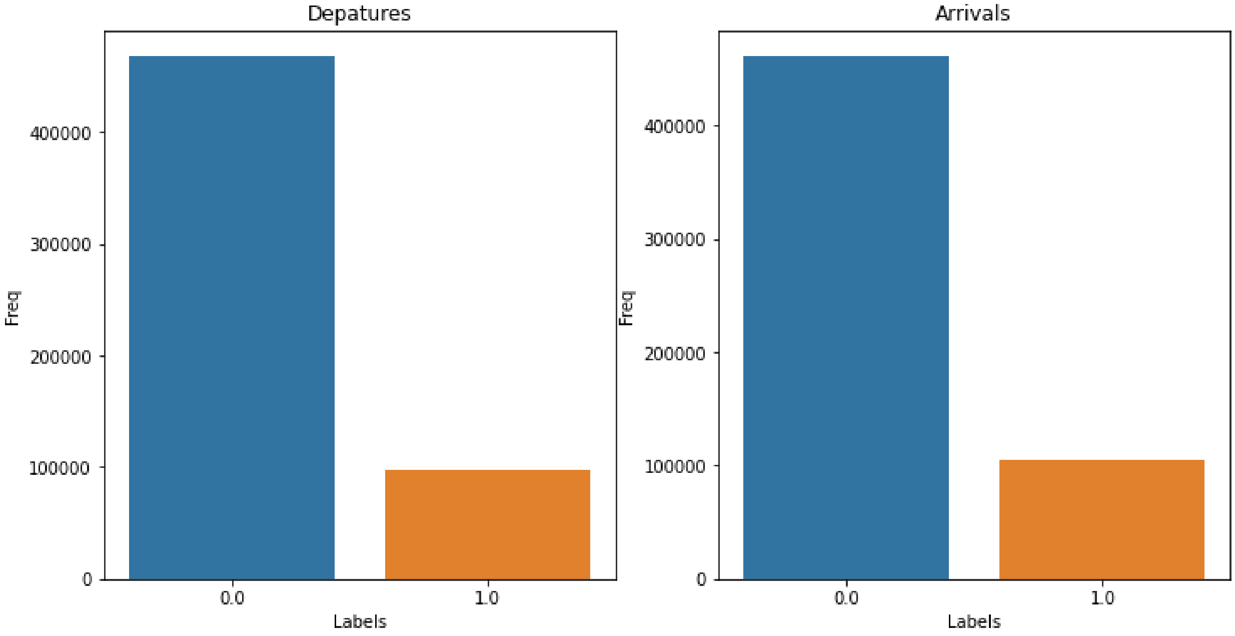


Figure 12- No.of delayed and non delayed flights with respect to arrival and departure

The histogram represents the frequency distribution of all the relevant attributes.

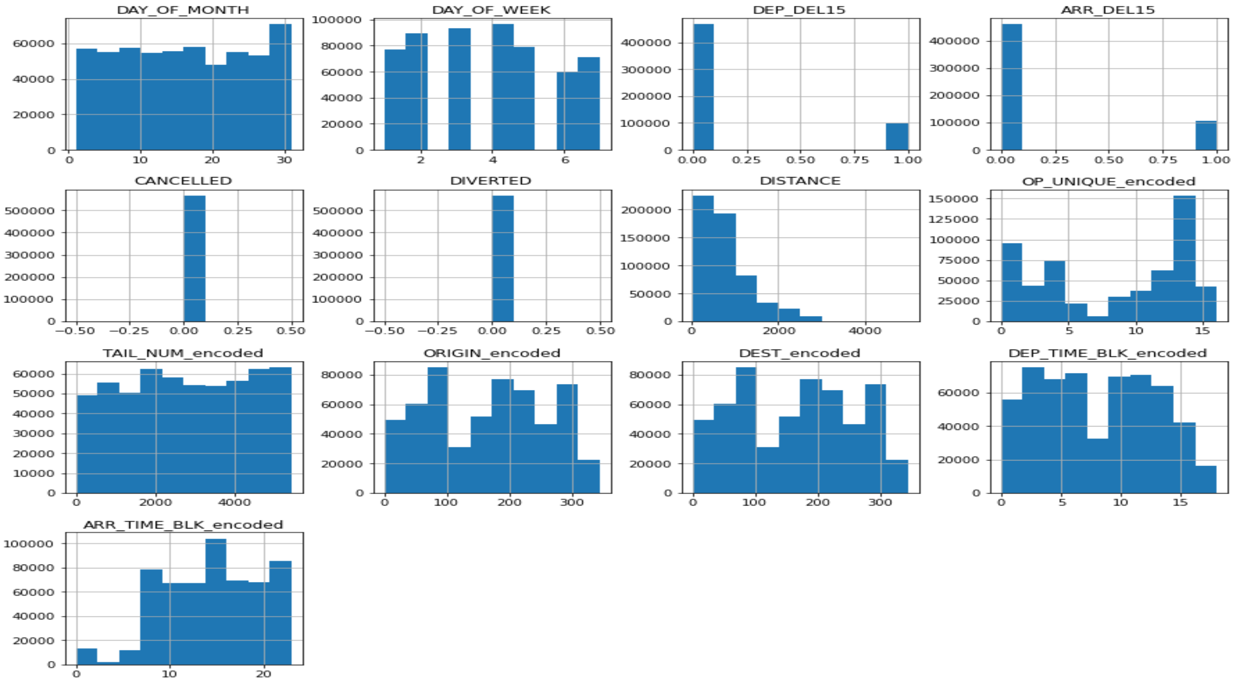


Figure 12- Histogram for frequency distribution of variables

After standardizing the cleaned data, the correlation analysis was done by generating the heat map, which determines the relation among various attributes. The correlation values near -1 and 1 depict a strong correlation, whereas any correlation value near 0, depicts a weak correlation.

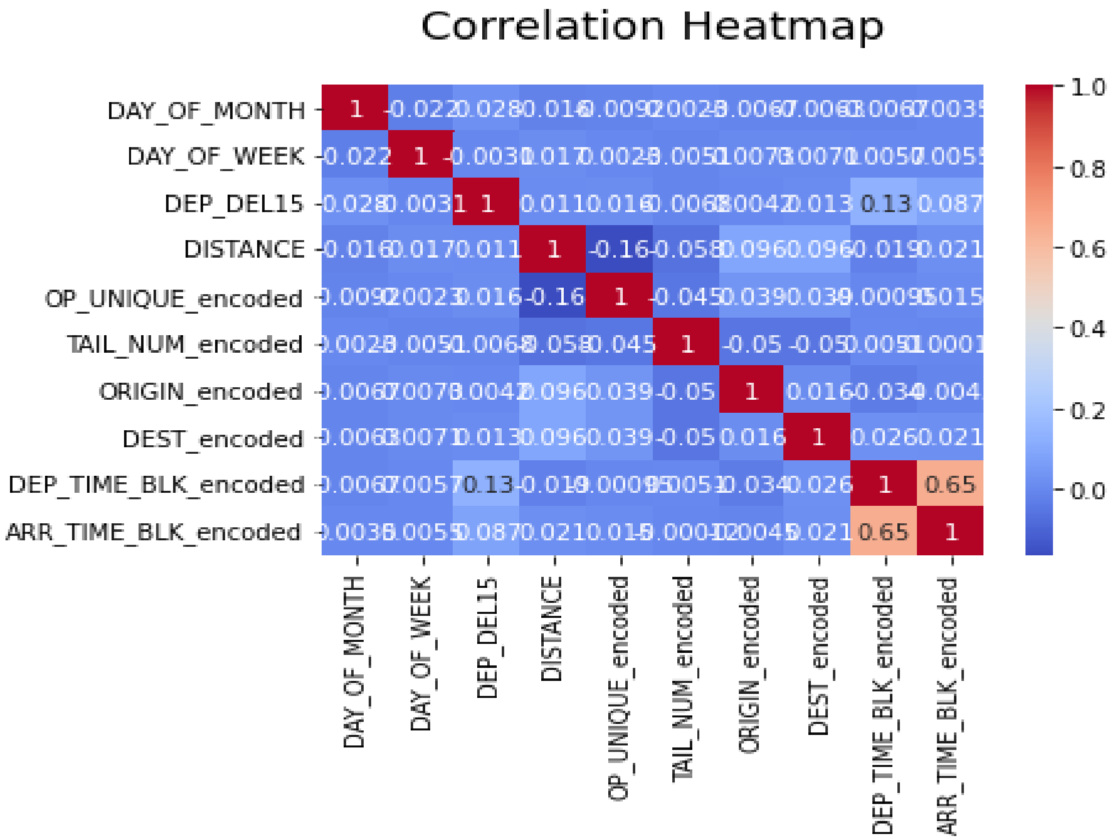


Figure 13- Heatmap used for correlation analysis of variables

Principal Component Analysis (PCA)

After standardizing the data, Principal Component Analysis was applied on the standardized data. However, during the course of the project it was observed that the subsequently implemented models resulted in good accuracy scores without the implementation of PCA. As a result, PCA was not applied on our data while generating the models.

Data Mining Tasks

a. Data Understanding

The initial dataset had 583,986 rows and 22 columns, out of which “ARR_DEL15” is the target variable. “ARR_DEL15” is determined as the target variable of our dataset due to the fact that reaching the destination on time is the most pertinent thing for a passenger. By analyzing the data types of the variables, it was noted that numeric as well as categorical data types were present. Most of the categorical variables including the target variable are binary in nature, which means their values are either ‘0’ or ‘1’.After carefully going through, examining and understanding the dataset, it was observed that a particular flight traveling from an origin to a destination can be identified by more than one attributes like “OP_CARRIER_AIRLINE_ID”, “ OP_CARRIER”, “ORIGIN_AIRPORT_ID”, “ORIGIN_AIRPORT_SEQ_ID”, “DEST_AIRPORT_ID”, “DEST_AIRPORT_SEQ_ID”, “OP_CARRIER_FL_NUM” and “OP_UNIQUE_CARRIER”.

b. Data Preprocessing

Initially, the missing values for each of the attributes were calculated. As variables are categorical, the records with missing values were dropped. Moreover, the redundant attributes were dropped as a part of data cleaning process, as one variable is enough to identify the flight which travels from one place to another. Apart from this, the pertinent attributes were converted from binary categorical to numerical data type, using Label encoding. These steps brought down our original dataset to 565963 rows and 13 variables.

Data Mining Models/Methods

After splitting the standardized data into 80% training data and 20% test data along with random state set to zero in order to reduce variability, a total of 5 models were built using the training data. The following models were implemented:-

1. **Decision Tree Classifier**

   A non parametric tree-resembling model that creates splits on predictors such that homogeneity increases after each split. The split segregates records into records into sub-groups, resulting in easily interpretable logical rules.

   Advantages:-

   - They are robust to outliers and thus, do not require Normalization or standardization of predictor variables.
   - They can directly handle missing values and do not require imputing them.
   - Variable subset selection occurs automatically in each split.

   Disadvantages:-

   - They are sensitive to changes in data, and any change can result in different splits.
   - As the splits occur on one predictor at a time, instead of combinations of predictors, they might miss recognizing relationships between predictors.
   - They are expensive to grow, as multiple sorting operations are carried out in computing all possible splits on every variable.

   Implementation:- A grid search was performed to obtain the best model using max_depth=6, impurity method= ‘gini’, with an accuracy of 87.7%.

2. Naive Bayes

One of the key assumptions of Naive Bayes Models is that it assumes to be conditionally independent given the target class. It scales linearly with the predictors and is robust to outside noise.

Advantages:-

- Naive Bayes is simpler and quicker compared to the Exact Bayes.
- The risk of overfitting noisy data is less.
- It is highly scalable.

Disadvantages:-

- Although Naive Bayes can manage with small data, it requires a large number of records to obtain reliable parameter estimates.
- Compared to the Exact Bayes, Naive Bayes retains the order of propensity but fails to generate the accurate propensity.

Implementation:- The Naive Bayes model was implemented on our data, which provided us with an accuracy 91.74%.

3. Multi-layer Perceptron Classifier (MLP)

A multilayer perceptron (MLP) is a fully connected class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation).

Advantages:-

- This allows for probability-based predictions or classification of items into multiple labels.
- Capability to learn non-linear models.
- Capability to learn models in real-time (on-line learning).

Disadvantages:-

- MLP with hidden layers have a non-convex loss function where there exists more than one local minimum.
- MLP requires tuning a number of hyperparameters such as the number of hidden neurons, layers, and iterations.
- MLP is sensitive to feature scaling.

Implementation:- The Multi-layer Perceptron (MLP) model was implemented on our data, which provided us with an accuracy 91.95%.

4. **Logistic Regression**

A parametric model that generates the output as the estimates of the probabilities of belonging to each class, and uses a threshold cutoff on the probabilities for classifying into either of the classes. The outcome variable called logit, can be modeled as a linear function of the predictors.

Advantages:-

- Easily understandable and offers an intuitive explanation of predictors.
- Computationally fast and cheap to classify large samples of new data.

Disadvantages:-

- Cannot be used to solve nonlinear problems.
- Requires predictor variables to be linearly associated with log odds.
- Sensitive to outliers.

Implementation:- The logistic regression model fits our data well, and it results in an accuracy of 91.74%.

5. **Random Forest Classifier**

The Random Forest is an ensemble method, which uses a collection of a large number of decision trees to make predictions instead of individual models. The individual models must make predictions which are independent of each other and each individual model should be better than a random classifier.

Advantages:-

- Helpful in reducing overfitting in decision trees and drastically improving accuracy.
- Efficient in handling missing values and robust to outliers.
- Works well with large input training sets.

Disadvantages:-

- Results from a random forest cannot be displayed in a tree-like diagram, thereby losing interpretability.
- Higher Computation time and complexity during training as multiple trees are generated.

Implementation:- The random forest classifier provides us with an accuracy of 91.85%.

Performance Evaluation

1. Decision Tree Classifier

A grid-search was done to obtain the best model using max_depth=6, impurity method= 'gini', with an Accuracy of 87.7%. The sensitivity (recall) value of 92.99% indicates that it was able to correctly classify the True Positives, or the arrival delays. The specificity value of 68.17% indicates that it was able to correctly classify 68.17% of the True Negatives, or the legitimate instances. The precision value of 80% indicates that it was able to correctly classify 80% of the positive observations, as true positive occurrences. The F1-score of 80% indicates low False Positives and low False Negatives, hence correctly identifying arrival delays.

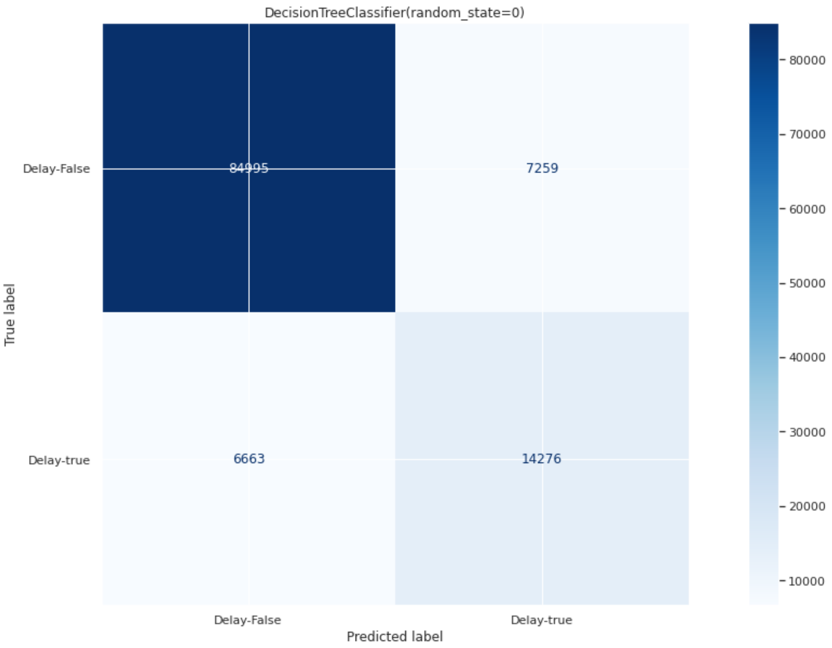
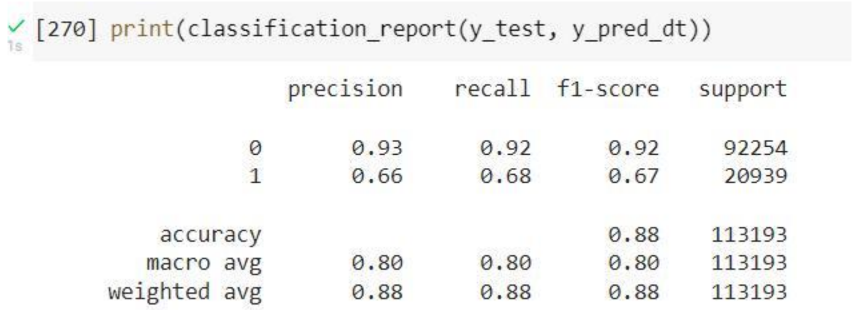


Figure 14 - Confusion Matrix of the Decision Tree Classifier



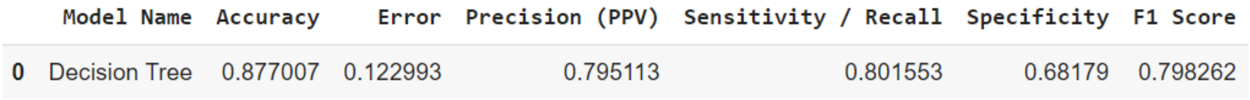
```
[270] print(classification_report(y_test, y_pred_dt))

              precision    recall  f1-score   support

           0       0.93      0.92      0.92     92254
           1       0.66      0.68      0.67     20939

    accuracy                           0.88    113193
   macro avg       0.80      0.80      0.80    113193
weighted avg       0.88      0.88      0.88    113193
```

Table 1 - Classification Report of the Decision Tree Classifier



	Model Name	Accuracy	Error	Precision (PPV)	Sensitivity / Recall	Specificity	F1 Score
0	Decision Tree	0.877007	0.122993	0.795113	0.801553	0.68179	0.798262

Table 2 - Evaluation Metrics of Decision Tree Classifier

2. Naive Bayes Model

A grid-search was done to obtain the best model with an Accuracy of 91.75%. The sensitivity (recall) value of 85% indicates that it was able to correctly classify the True Positives, or the arrival delays. The specificity value of 74.44% indicates that it was able to correctly classify 74.44% of the True Negatives, or the legitimate instances. The precision value of 86.95 % indicates that it was able to correctly classify 86.95 % of the positive observations, as true positive occurrences. The F1-score of 85.96% indicates low False Positives and low False Negatives, hence correctly identifying arrival delays.

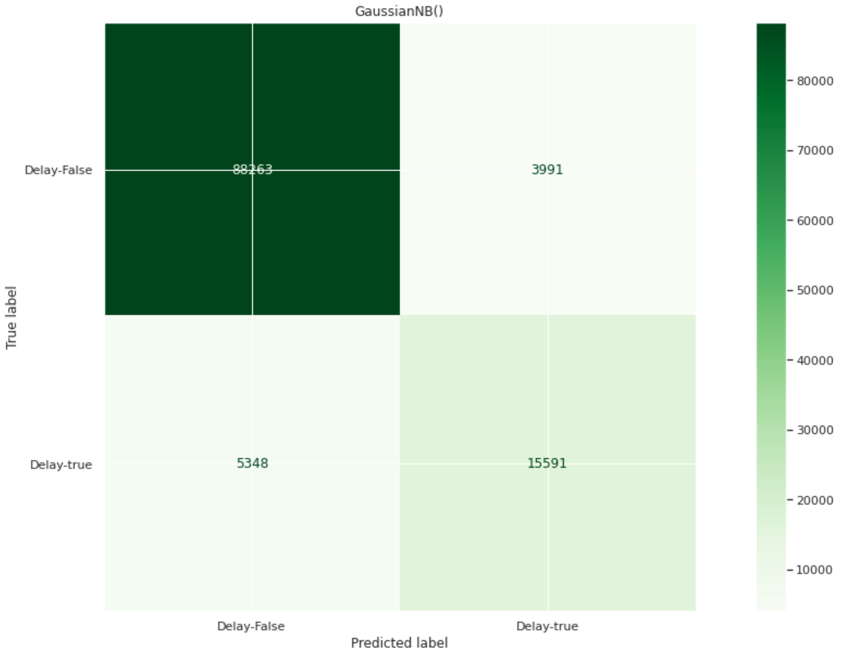
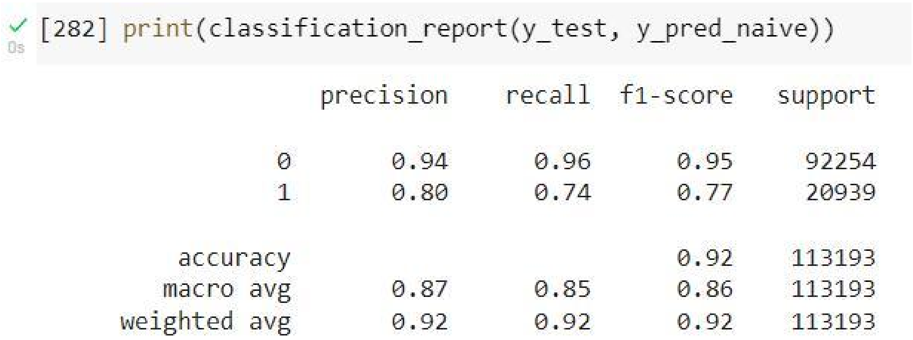


Figure 15 - Confusion Matrix of the Naive Bayes Model



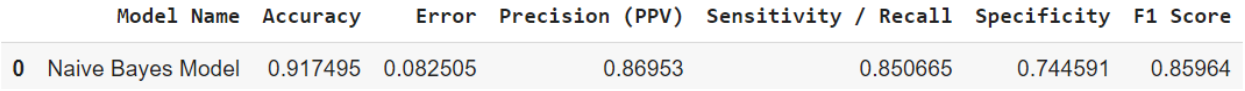
```
[282] print(classification_report(y_test, y_pred_naive))

              precision    recall  f1-score   support

           0       0.94      0.96      0.95     92254
           1       0.80      0.74      0.77     20939

    accuracy                           0.92    113193
   macro avg       0.87      0.85      0.86    113193
weighted avg       0.92      0.92      0.92    113193
```

Table 3 - Classification Report of the Naive Bayes Model



	Model Name	Accuracy	Error	Precision (PPV)	Sensitivity / Recall	Specificity	F1 Score
0	Naive Bayes Model	0.917495	0.082505	0.86953	0.850665	0.744591	0.85964

Table 4 - Evaluation Metrics of Naive Bayes Model

3. Multi- Layer Perceptron (MLP) Classifier

A grid-search was done to obtain the best model with an Accuracy of 91.92%. The sensitivity (recall) value of 84% indicates that it was able to correctly classify the True Positives, or the arrival delays. The specificity value of 71.45% indicates that it was able to correctly classify 71.45% of the True Negatives, or the legitimate instances. The precision value of 88.12 % indicates that it was able to correctly classify 88.12 % of the positive observations, as true positive occurrences. The F1-score of 85.86% indicates low False Positives and low False Negatives, hence correctly identifying arrival delays.

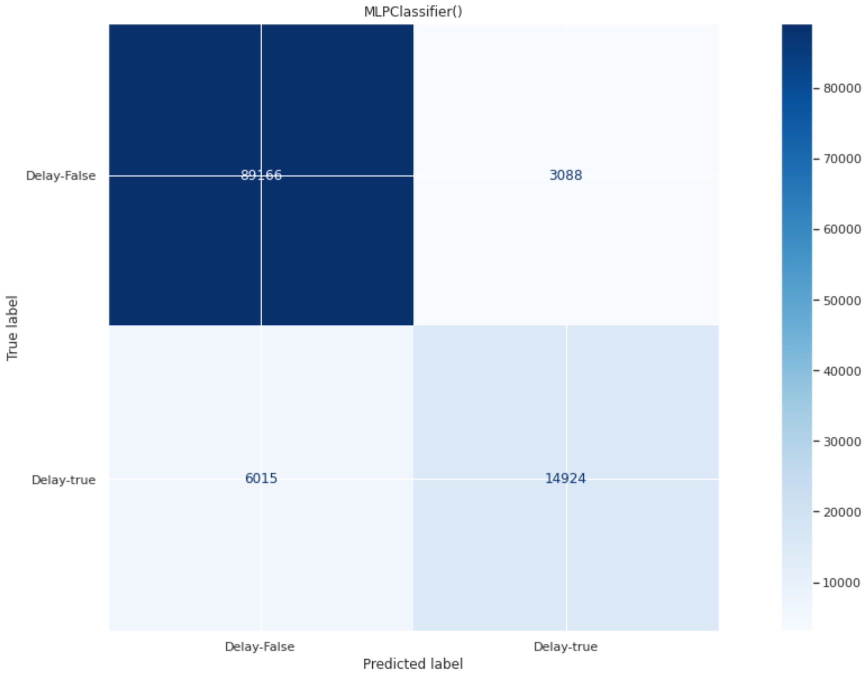
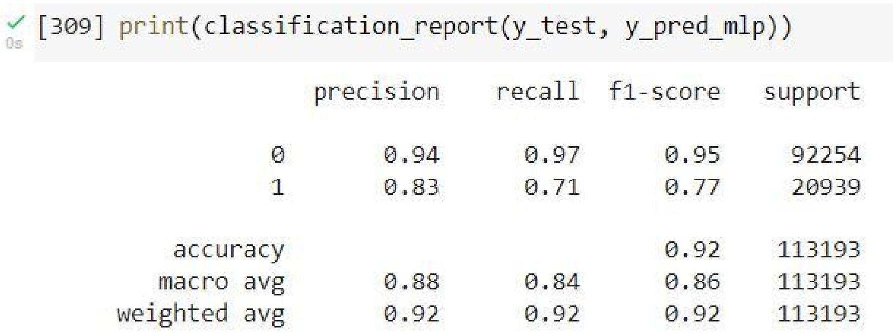


Figure 16 - Confusion Matrix of the MLP Classifier model



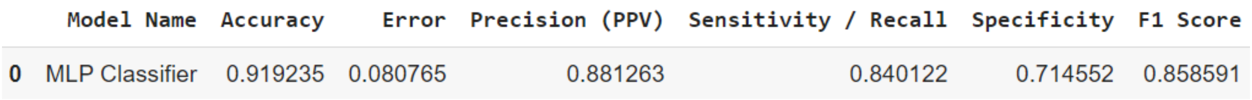
```
[309] print(classification_report(y_test, y_pred_mlp))

              precision    recall  f1-score   support

           0       0.94      0.97      0.95     92254
           1       0.83      0.71      0.77     20939

    accuracy                           0.92    113193
   macro avg       0.88      0.84      0.86    113193
weighted avg       0.92      0.92      0.92    113193
```

Table 5 - Classification Report of the MLP Classifier Model



	Model Name	Accuracy	Error	Precision (PPV)	Sensitivity / Recall	Specificity	F1 Score
0	MLP Classifier	0.919235	0.080765	0.881263	0.840122	0.714552	0.858591

Table 6 - Evaluation metrics of the MLP Classifier Model

4. Logistic Regression Model

A grid-search was done to obtain the best model with an Accuracy of 91.75%. The sensitivity (recall) value of 85% indicates that it was able to correctly classify the True Positives, or the arrival delays. The specificity value of 74.44% indicates that it was able to correctly classify 74.44% of the True Negatives, or the legitimate instances. The precision value of 86.95 % indicates that it was able to correctly classify 86.95 % of the positive observations, as true positive occurrences. The F1-score of 85.96% indicates low False Positives and low False Negatives, hence correctly identifying arrival delays.

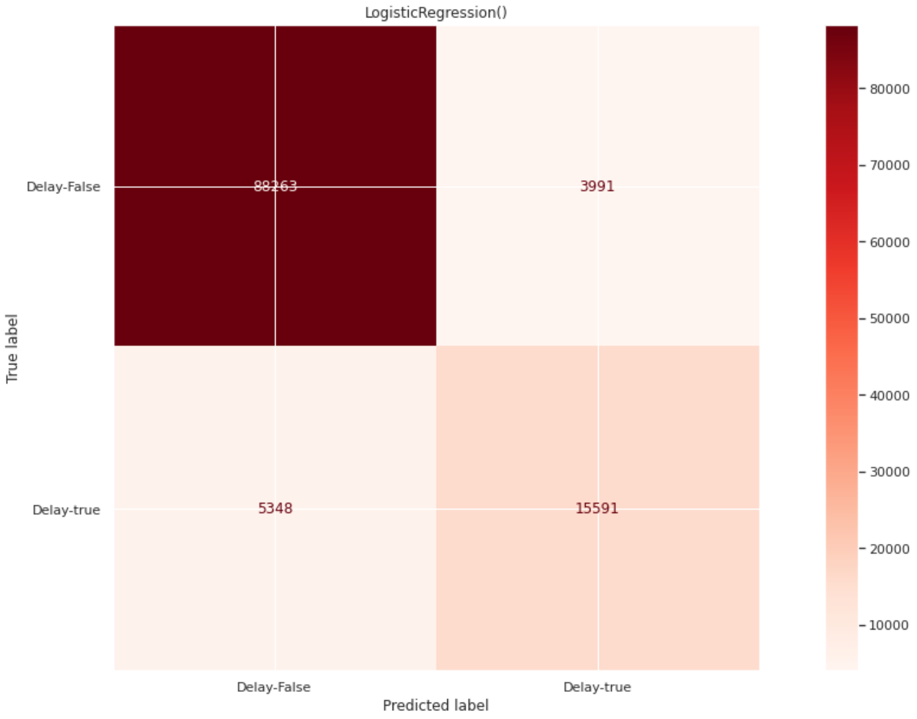
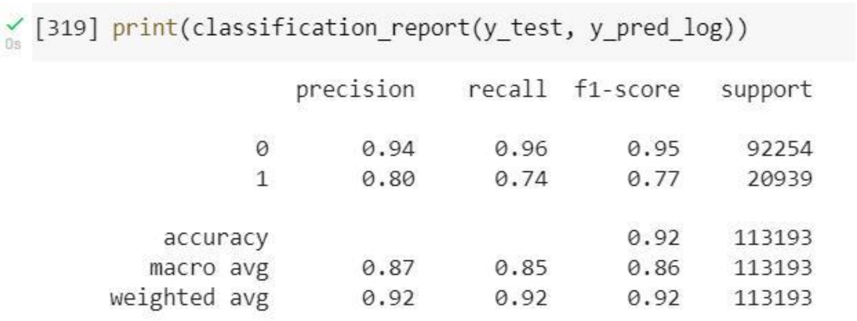


Figure 17 - Confusion Matrix of the Logistic Regression model



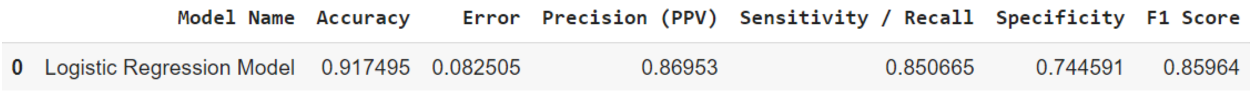
```
[319] print(classification_report(y_test, y_pred_log))

              precision    recall  f1-score   support

           0       0.94      0.96      0.95     92254
           1       0.80      0.74      0.77     20939

    accuracy                           0.92    113193
   macro avg       0.87      0.85      0.86    113193
weighted avg       0.92      0.92      0.92    113193
```

Table 7 - Classification Report of the Logistic Regression Model



	Model Name	Accuracy	Error	Precision (PPV)	Sensitivity / Recall	Specificity	F1 Score
0	Logistic Regression Model	0.917495	0.082505	0.86953	0.850665	0.744591	0.85964

Table 8 - Evaluation Metrics of the Logistic Regression Model

5. Random Forest Classifier

A grid-search was done to obtain the best model with an Accuracy of 91.9%. The sensitivity (recall) value of 84.13% indicates that it was able to correctly classify the True Positives, or the arrival delays. The specificity value of 71.8% indicates that it was able to correctly classify 71.8% of the True Negatives, or the legitimate instances. The precision value of about 88 % indicates that it was able to correctly classify 88 % of the positive observations, as true positive occurrences. The F1-score of 85.86% indicates low False Positives and low False Negatives, hence correctly identifying arrival delays.

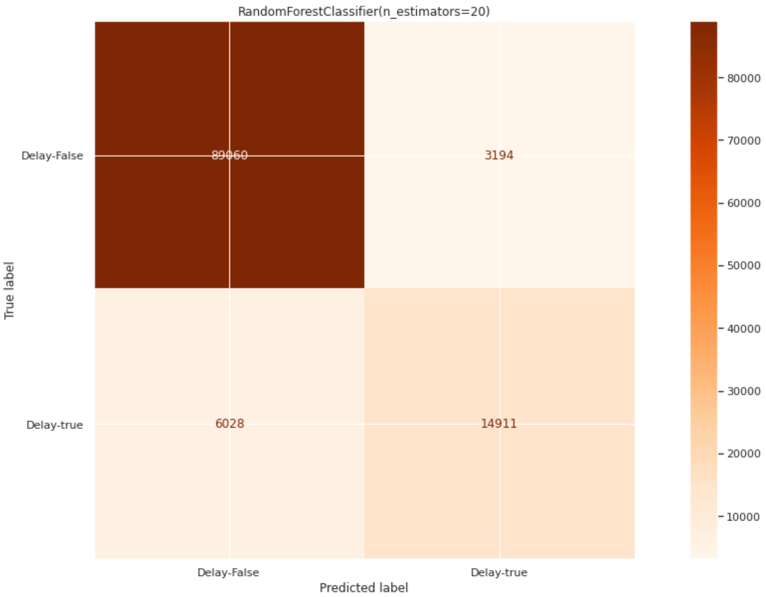
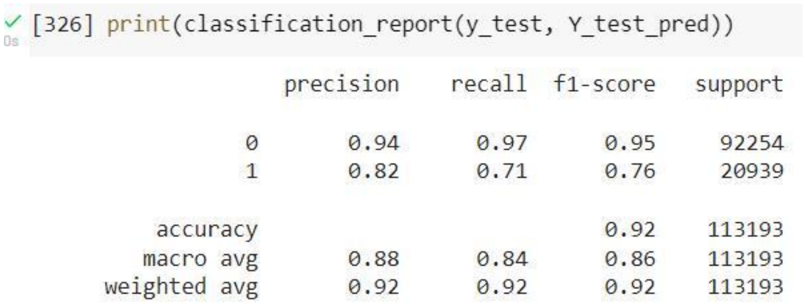


Figure 18 - Confusion Matrix of the Random Forest Classifier model



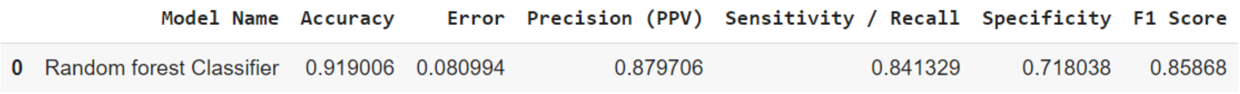
```
[326] print(classification_report(y_test, Y_test_pred))

              precision    recall  f1-score   support

           0       0.94      0.97      0.95     92254
           1       0.82      0.71      0.76     20939

    accuracy                           0.92    113193
   macro avg       0.88      0.84      0.86    113193
weighted avg       0.92      0.92      0.92    113193
```

Table 9 - Classification Report of the Random Forest Classifier Model

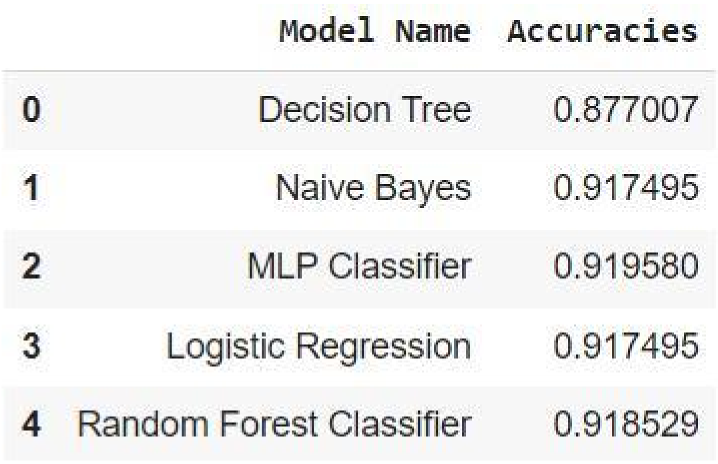


	Model Name	Accuracy	Error	Precision (PPV)	Sensitivity / Recall	Specificity	F1 Score
0	Random forest Classifier	0.919006	0.080994	0.879706	0.841329	0.718038	0.85868

Table 10 - Evaluation Metrics of the Random Forest Classifier Model

Project Results

- The Multi-layer Perceptron Classifier is the best model in terms of accuracy, as it gives an accuracy of 91.95%. It also gives the highest values of Precision,Recall and F-1 score of 0.882,0.839 and 85.8% respectively.
- The Random Forest Classifier is the next best model, which fits our data. It resulted in an accuracy of 91.85% along with Precision,Recall and F-1 score of 0.880,0.838 and 85.7% respectively.
- The Logistic Regression Model and the Naive Bayes are the next best models as they are pretty much similar in terms of accuracies and the other metrics. They result in accuracy of 91.74%.
- The lowest accuracy score obtained was that of Decision Tree Classifier as it resulted in an accuracy of 87.70%. It was also lowest in terms of other metrics like Precision,Recall and F-1 score with values of 0.795,0.801 and 79.8% respectively.



	Model Name	Accuracies
0	Decision Tree	0.877007
1	Naive Bayes	0.917495
2	MLP Classifier	0.919580
3	Logistic Regression	0.917495
4	Random Forest Classifier	0.918529

Table 11 - Accuracies of all the models

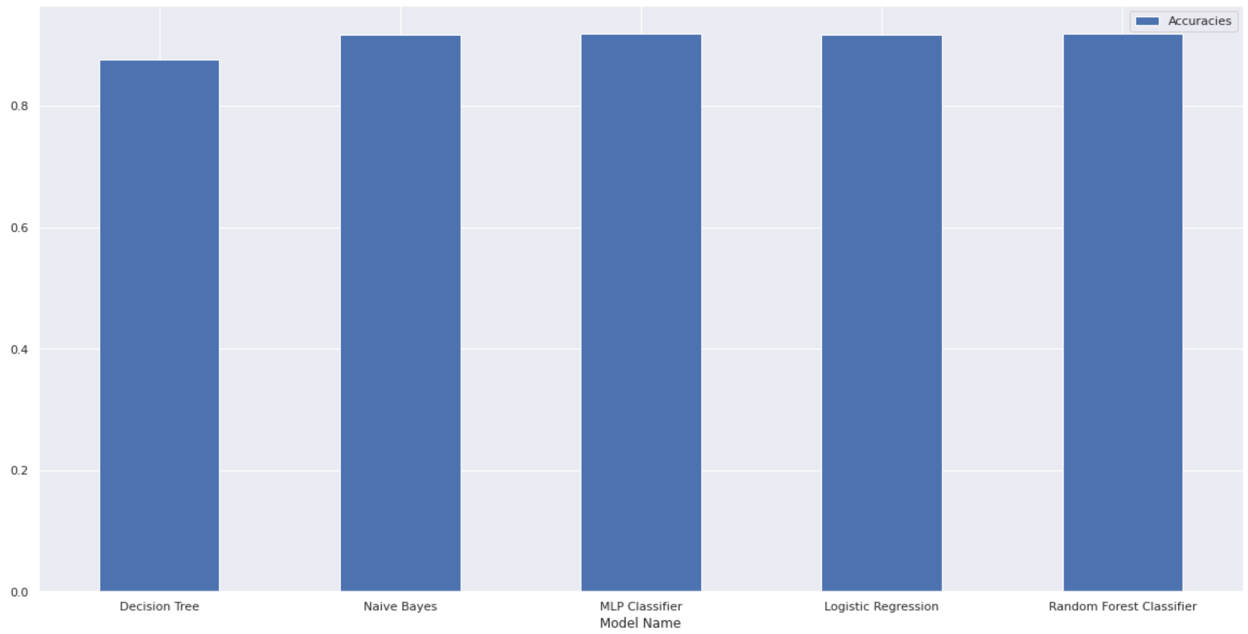
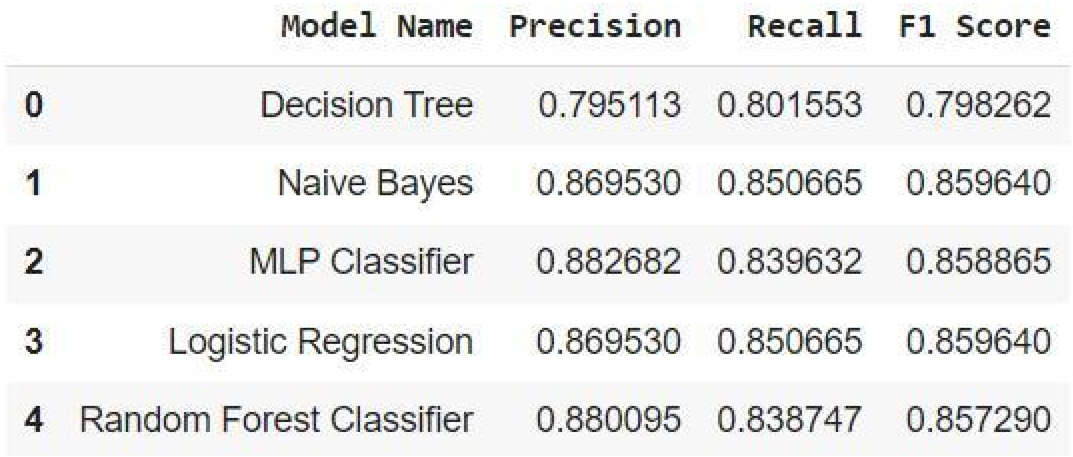


Figure 19 - Bar plot comparing the accuracies of the five models



	Model Name	Precision	Recall	F1 Score
0	Decision Tree	0.795113	0.801553	0.798262
1	Naive Bayes	0.869530	0.850665	0.859640
2	MLP Classifier	0.882682	0.839632	0.858865
3	Logistic Regression	0.869530	0.850665	0.859640
4	Random Forest Classifier	0.880095	0.838747	0.857290

Table 12 - Precision, Recall and F1 scores of all models

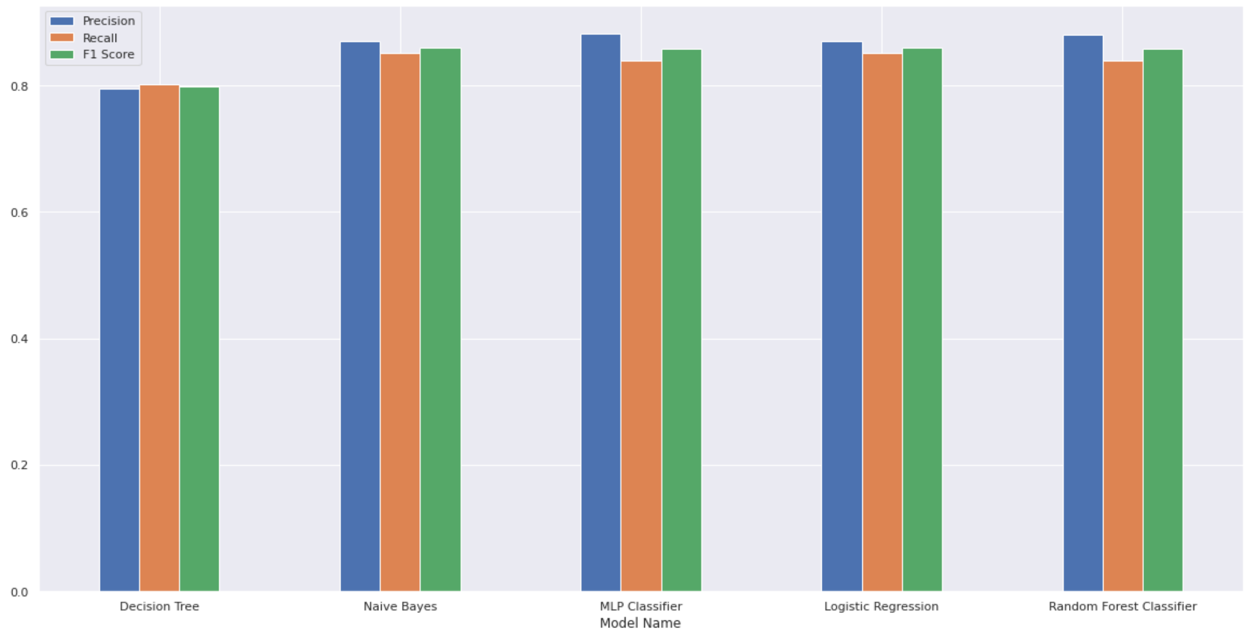


Figure 20 - Grouped bar plot comparing the precision, recall and F1 scores

Impact of the Project Outcomes

In this Project, the goal was to be able to identify if a given flight gets delayed or not accurately. The class of interest represented by label 1 was that of delayed flights. This implies that the end result of the analysis would help us identify which of the flights were likely to get delayed, based upon which the appropriate actions could be taken by consumers such that they reach their destinations on time. Out of all the classification models, the Multi-layer Perceptron outperformed all the remaining classifiers due to its high overall accuracy, high precision, high recall and a high F-1 score. From this analysis, the MLP Classifier was the best classification model which could be used in predicting flight delays.

References

[1]Kaggle Dataset-
https://www.kaggle.com/datasets/divyansh22/flight-delay-prediction?select=Jan_2020_ontime.csv

[2] Jiage Huo, K. L. Keung, C. K. M. Lee, Kam K.H. Ng - *The Prediction of Flight Delay: Big Data-driven Machine Learning Approach* - 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)