PREDICTIONS FOR FLIGHT DATA

SLDM MINI PROJECT

Shruti Deshmukh(21060641047) Yukta Dhankhar(21060641058)

Shubhangi Deshmukh(21060641048) Shivesh Shukla(21060641046)

Executive Summary

Everyday millions of people travel through flights and many of them must have experienced one of issues like delays, last minute cancellations, and diversion due to multiple reasons. Problem is how to help clients to reduce the number of cancelled or delayed flights. Along with this to predict the delay time in departure time.

Approach to solve the problem is to first understand every variable that is included in the dataset and perform basic exploratory data analysis and gain insights. Next step is the pre-processing of data and then moving forward to build models, checking its accuracy and giving relevant suggestions.

Business Problem

For anyone who has travelled on an aeroplane, may have experience with one of the inevitable pains of flying: the delays and cancellations of flights. Sometimes your plane arrives late, other times there may be a queue for takeoff, occasionally the weather forces hour-long delays and even cancellation, regardless of the reason for the delay, they pose a huge inconvenience for travellers.

Hence, given the vast amount of data on flight travels, valuable insights can be drawn from this data to allow us to gain a better understanding of flight delays as well as cancellation. Moreover, with the abundance of data, machine learning models can be trained to possibly predict these delays — something that may prove very valuable for individual travellers and businesses alike. Thus, this topic can help to explore further by extracting potentially meaningful insights from available data and constructing and comparing models to hopefully predict flight delays**.**

To understand the extracted data from the Marketing Carrier On-Time Performance (Beginning January 2018) data table of the "On-Time" database from the TranStats data library. Building three models to predict which all flights will be cancelled or delayed and to predict the delay time. This will help the clients to reduce multiple delays or last minute cancellations.

Now, in order to predict these there are multiple variables present in the dataset and its overview is under data description.

Data Description

The dataset is from the marketing carrier's on-time performance, which has in total 61 columns and 1048575 rows containing both numerical and categorical data. There were a total 27 different airlines for multiple origins and destinations. Out of which, a total of 17047 flights were cancelled, and 2293 flights were diverted. For cancelled flights, there are a total of 493 columns where Arrtime were present in the data, which can be considered as wrong imputed data. The data given includes flight details for 2018 and for two months, i.e. January and October.

Columns in the dataset:

Airline	Origin	Dest	Diverted
ArrTime	ArrDelayMinutes	AirTime	ActualElapsedTime
DayofMonth	DayOfWeek	Marketing_Airline_Network	DOT_ID_Marketing_Airline
IATA_Code_Operating_Airline	Tail_Number	Flight_Number_Operating_Airline	OriginAirportSeqID
OriginStateName	OriginWac	DestAirportID	DestCityMarketID
DestWac	DepDel15	DepartureDelayGroups	TaxiOut
ArrDelay	ArrDel15	ArrivalDelayGroups	ArrTimeBlk
DepTime	DepDelayMinutes	Cancelled	DistanceGroup
Year	Quarter	CRSElapsedTime	CRSDepTime
Flight_Number_Marketing_Airline	Operating_Airline	Operated_or_Branded_Code_Share_Partners	Distance
OriginCityName	OriginState	OriginAirportID	IATA_Code_Marketing_Airline
DestState	DestStateFips	DestAirportSeqID	OriginCityMarketID
WheelsOn	TaxiIn	DepTimeBlk	DestCityName
WheelsOff	DivAirportLandings		

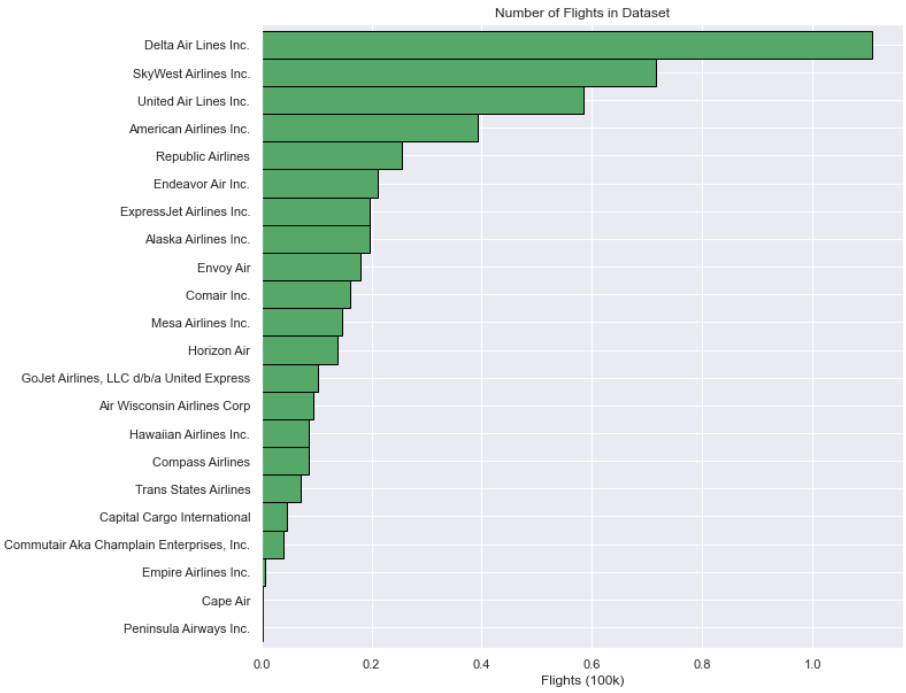
Exploratory Data Analysis (EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and to check assumptions with the help of summary statistics and graphical representations.

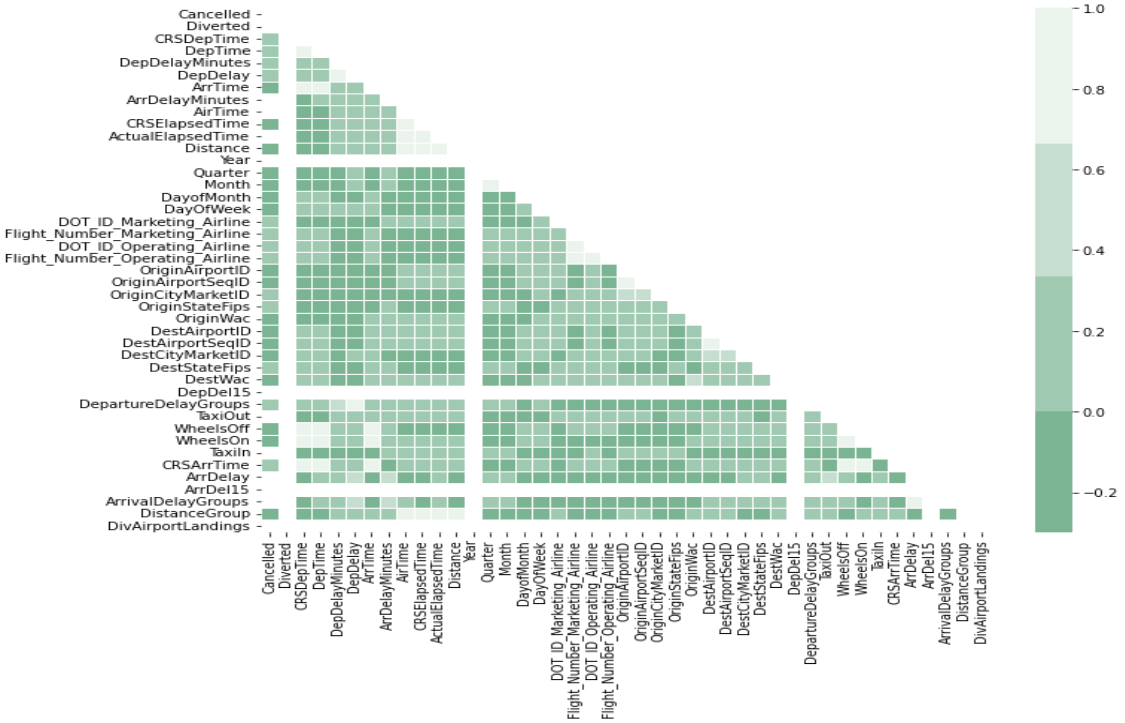
- First basic details were gathered on the total number of flights, the number of cancelled flights, delayed flights, and the percentage of both cancelled and delayed flights which can be seen in the below table. The Number of cancelled flights are very low.

```
Total number of flights: 1048575
Number of cancelled flights: 17047
Number of delayed flights: 342300
Number of diverted flights: 2293
Number of not cancelled flights: 1031528
Number of not delayed flights: 706275
Percentage of cancelled flights: 1.625730157594831%
Percentage of delayed flights: 32.644302982619266%
```

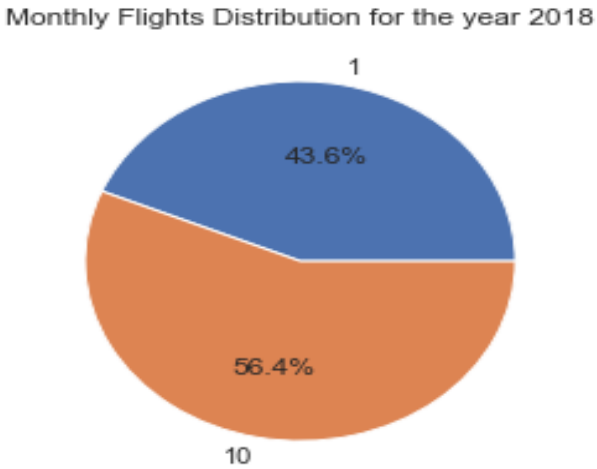
- Airline distribution:

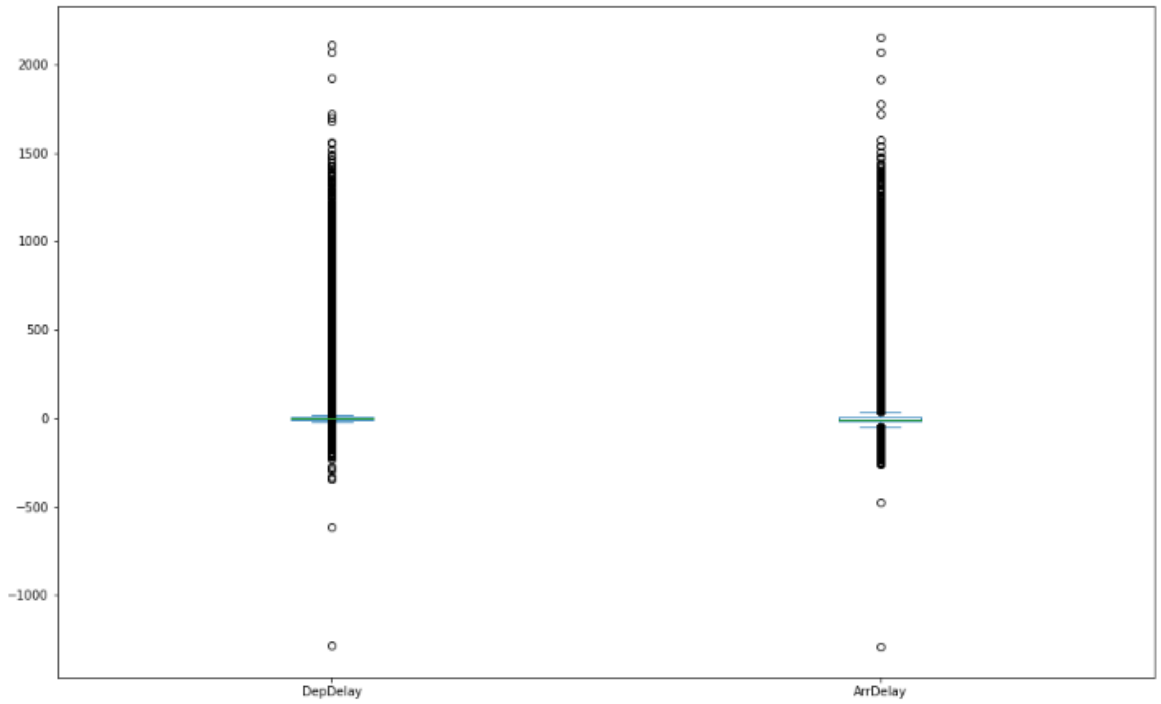


- Next, along with the datatypes of each variable, missing values percentages were calculated, from which it was noticed that 10 variables had missing values with a filling factor percentage of around 98% and the rest all the variables had zero null values. This helped to reach a conclusion that maximum flights which were cancelled had no arrival or delay time. 493 rows had data on cancelled flights along with departure time which can be considered as an input error.
- From the heat map, it can be seen that variables are correlated which means multicollinearity can be present. Variable values between 1 to 0.8 are highly correlated.

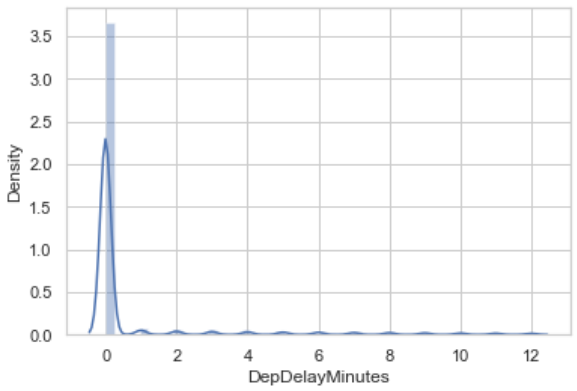


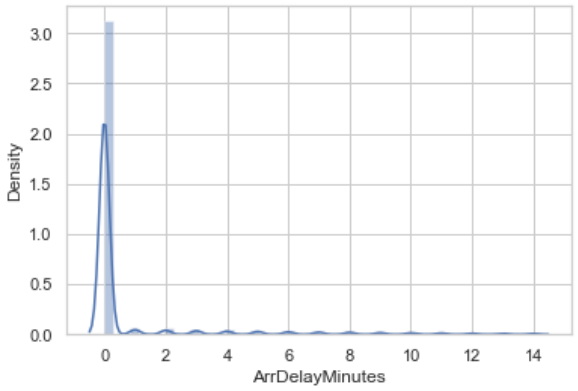
- From the box plot, we can see that there are some negative values which mean there are some flights that took off before the expected departure time.
- Monthly Flights Delay distribution for the year 2018: it can be seen that in the month of october maximum flights are delayed.



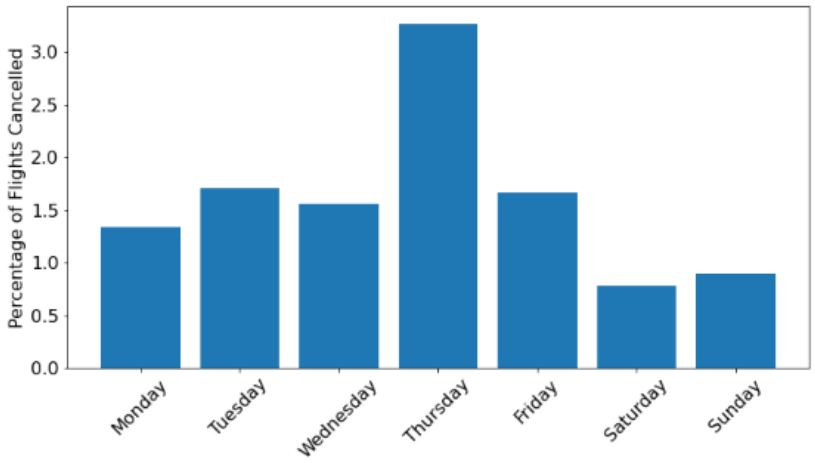


- Next, on-time flight percentage and cancelled percentage were calculated for each airline which helped to understand that Cape Air airline had the lowest flight counts with no flight cancellations and the highest on-time percentage of 78%. Southwest Airlines co. has maximum flight counts with 1.44% cancellations and 64.56% on-time departures.
- It can be seen on the histogram and by the skewness and kurtosis indexes, that delays are mostly located on the left side of the graph, with a long tail to the right. The majority of delays are short, and the long delays, while unusual, are more heavily loaded in time.

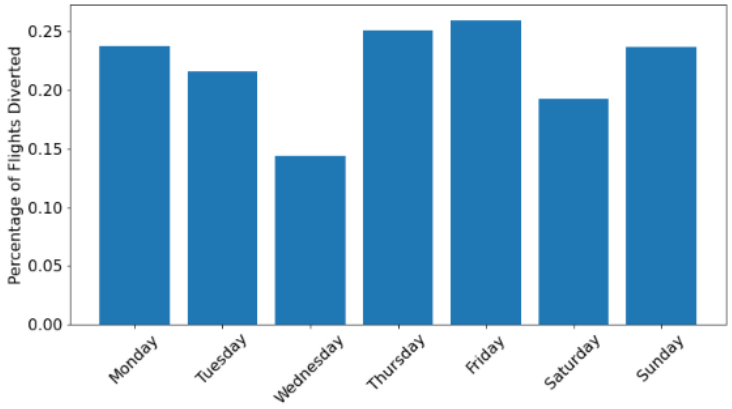




- Here Saturday has the lowest percentage of cancelled flights, and Thursday has the highest.



- Here, Wednesday has the lowest percentage of diverted flights.



- From the state CA, maximum number (122800) of flights are cancelled and minimum number(315) from state VI

Airline	Cancelled Flights	Diverted Flights
Air Wisconsin Airlines Corp	191	34
Alaska Airlines Inc.	322	104
Allegiant Air	84	38
American Airlines Inc.	572	150
Capital Cargo International	224	24
Comair Inc.	441	69
Commutair Aka Champlain Enterprises, Inc.	289	12
Compass Airlines	41	29
Delta Air Lines Inc.	1613	225
Empire Airlines Inc.	13	6
Endeavor Air Inc.	1158	53
Envoy Air	523	76
ExpressJet Airlines Inc.	1061	80
Frontier Airlines Inc.	371	25
GoJet Airlines, LLC d/b/a United Express	145	26
Hawaiian Airlines Inc.	24	19
Horizon Air	346	61
JetBlue Airways	1613	135
Mesa Airlines Inc.	712	50
Peninsula Airways Inc.	41	11
Republic Airlines	968	68
SkyWest Airlines Inc.	1573	389
Southwest Airlines Co.	3023	346
Spirit Air Lines	401	35
Trans States Airlines	238	32
United Air Lines Inc.	957	189
Virgin America	103	6
Grand Total	**17047**	**2293**

	variable	missing values	filling factor (%)
0	AirTime	20220	98.071669
1	ArrivalDelayGroups	19561	98.134516
2	ArrDel15	19561	98.134516
3	ArrDelay	19561	98.134516
4	ArrDelayMinutes	19561	98.134516
5	ActualElapsedTime	19359	98.153780
6	ArrTime	17515	98.329638
7	DepDelay	17061	98.372935
8	DepartureDelayGroups	17061	98.372935
9	DepDel15	17061	98.372935
10	DepDelayMinutes	17061	98.372935
11	DepTime	16573	98.419474
12	DOT_ID_Marketing_Airline	0	100.000000
13	Operating_Airline	0	100.000000
14	FlightDate	0	100.000000
15	DayOfWeek	0	100.000000

Data Pre-Processing

- Before building any model, it is crucial to perform data pre-processing to feed the correct data to the model to learn and predict. Model performance depends on the quality of data feeded to the model to train.
- This Process includes

  a) Handling Null/Missing Values

  b) Handling Skewed Data

  c) Outliers Detection and Removal

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

a) Remove duplicate or irrelevant observations

b) Filter unwanted outliers

c) Renaming required attributes

Methodology

Theory:

Logistic regression models a relationship between predictor variables and a categorical response variable. For example, we could use logistic regression to model the relationship between various measurements of a manufactured specimen (such as dimensions and chemical composition) to predict if a crack greater than 10 mils will occur (a binary variable: either yes or no). Logistic regression helps us estimate the probability of falling into a certain level of the categorical response given a set of predictors.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction. The random forest classifier can be used to solve regression or classification problems.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.The fundamental Naive Bayes assumption is that each feature makes an: Independent and equal contribution to the outcome.

1. **For solving the problem, prediction of cancellation of the flights following approach was followed.**

   Data Pre-processing:

   First, by observing the whole data, the identification of columns which appeared irrelevant to the cancellation of the flights were deleted. The presence of the null values was then observed. The contains a total 1048576 rows out of which 2751 i.e. 0.26% were null for the column "Tail Number", thus those rows were removed.

   The next step was to convert the columns into suitable data types, categorical for the modelling purpose. The selection of Dependent and Independent variables was done. Here, the Dependent variable is the column "Cancelled". The feature selection was done for the whole data, as it still contains 26 features even after removing irrelevant columns. The feature selection gave features which influenced the cancellation the most.

   In case of a cancelled column, the false values were much higher than the True values. Thus, the data is unbalanced. If the modelling on such data, the overfitting of the data will occur. To deal with this unbalanced data, resampling was done. This resampled data was thus then split into training and testing data.

Modelling:

After the preprocessing of the data, three models were fit on the data.

1.Logistic Regression Model

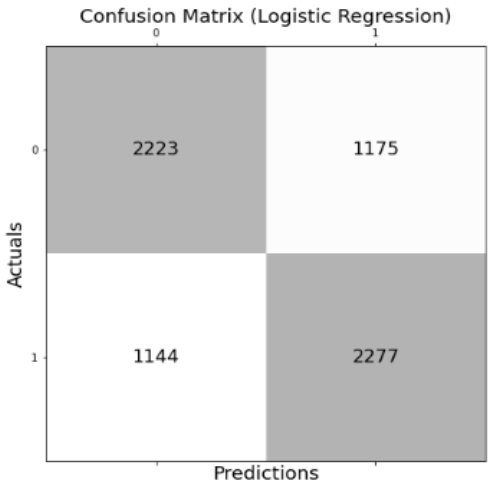
2.Random Forest Classifier

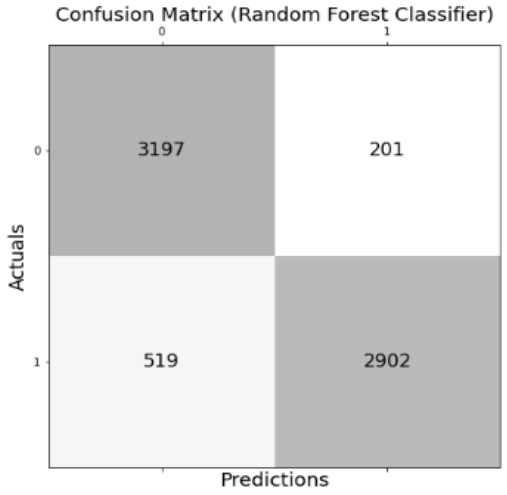
3.Naive Bayes Classifier

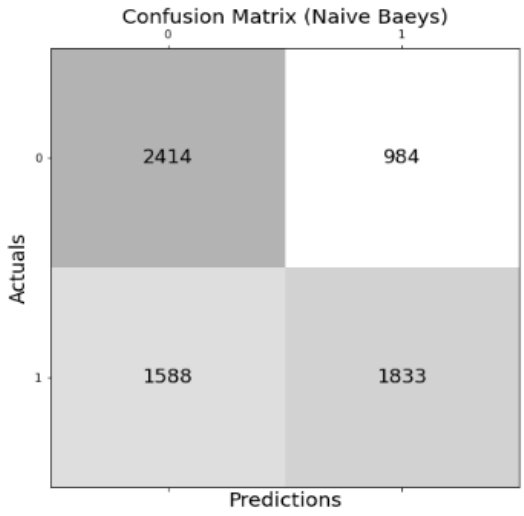
The data was splitted into independent variables and dependent variable which is 'Cancelled'. Then data was splitted into train and test and a model was fitted on it to classify flights into Cancelled or not. Then classification was predicted using test data. Confusion matrix of each model is given below.

Outputs

Models	Accuracy
Logistic Regression Model	66%
Random Forest	89%
Naïve Bayes Classifier	62%







After observing all the outputs, we can conclude that Random Forest Classifier gave best results.

Thus, Cross validation was done for Random Forest Classifier, with 10 folds.

```
Cross Validation Scores are [0.88892962 0.88416422 0.88526393 0.88416422 0.88
746334 0.88668867
 0.89365603 0.88302164 0.88815548 0.89145581]
Average Cross Validation score :0.8872962968339844
```

Thus, the Random Forest Classifier gives best predictions with a good cross validation score.

2. **For predicting which flights will be delayed, following approach was followed**

   Data Pre-processing:

   First, by observing the whole data, the identification of columns which appeared irrelevant for predicting which flights will be delayed were deleted. The presence of the null values was then observed. Out of 1048576 rows 19564 rows have missing values. As the number of missing values are too less compared to whole data, those rows were

deleted. The column ‘DepDelayMinutes’ have numerical values of by how much time the flight was delayed than actual departure time. For converting that column into category ‘Delayed’ or ‘On-time’, all values were converted into 1 or 0. The DepDelayMinutes which are greater than 0 are the delayed flights hence they are categorised as 1 and DepDelayMinutes with value 0 are the flights which are not delayed(On time) hence kept as it is which is 0.

The next step was to convert the columns into suitable data types, categorical for the modelling purpose. By using Decision tree as feature selection method for selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy variables. After feature selection , the best set of 9 features were selected to build different models.

List of features :

1.**ArrTime**

2.AirTime

3.DepartureDelayGroups

4.TaxiOut

5.WheelsOff

6.WheelsOn

7.ArrDel15

8.ArrivalDelayGroups

9.DistanceGroup

The data was unbalanced having 323664 delayed flights and 704010 On time flights. For dealing with unbalancedness ,randomly duplicated examples in the minority class.Then model was fitted on this balanced data.

Modelling :

After the preprocessing of the data, three models were fit on the data.

1.Logistic Regression Model

2.Random Forest Classifier

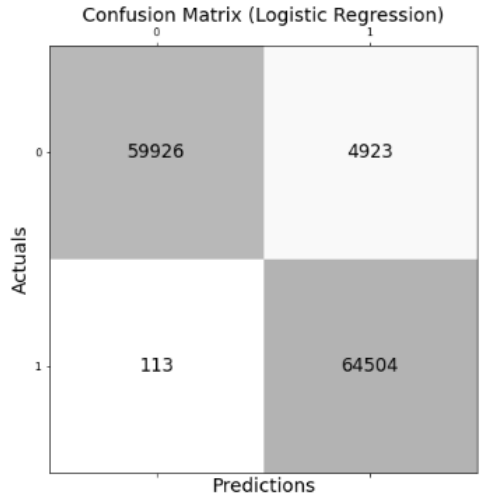
3.Naive Bayes Classifier

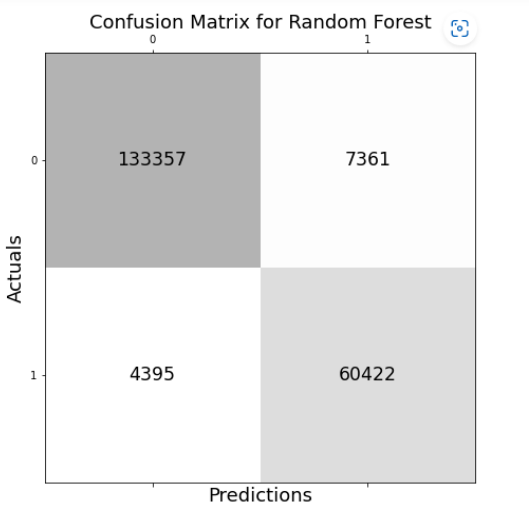
The data was splitted into independent variables and dependent variable which is 'DepDelayMinutes'. Then data was splitted into train and test and a model was fitted on it to classify flights into delayed or On time. Then classification was predicted using test data. Confusion matrix of each model is given below.

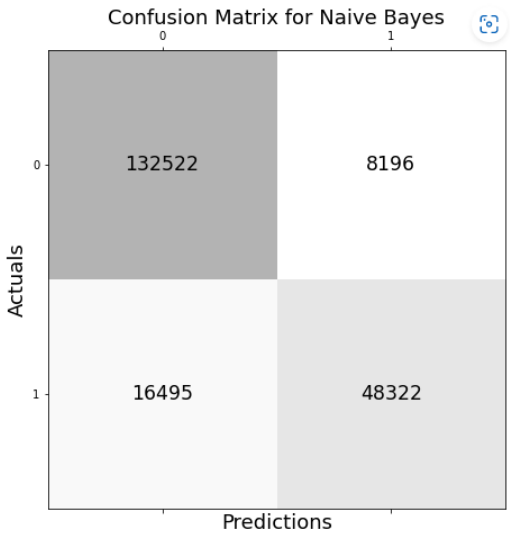
Output:

Models	Accuracy
Logistic Regression Model	96%
Random Forest	95%
Naïve Bayes Classifier	85%

Confusion matrix :







After observing all the outputs, we can conclude that Logistic Regression gave best results to predict if flight is delayed or not.

2. Develop a model to predict the delay time

Data Pre-processing:

- For predicting departure delay in minutes
- Intuitively independent variables were selected
- Rows having missing values were deleted
- Numerical variables were converted into categorical using encoding

Modelling :

For checking assumption of multicollinearity VIF values of dependent variables were calculated. Variables having VIF values less than 5 are selected to remove multicollinearity.

Variables included in model :

1.Airline

2 Month

3 DayOfMonth

4 Flight_Number_Marketing_Airline

5 Operating Airline

6. OriginAirportID

7 OriginAirportFips

8 DestAirportID

9 DestAirportFips

10 DestWac

11 TaxiIn

12 ArrDelay

Data splitted into Train and test then Multiple Linear Regression model was fitted on data. Then departure delay in minutes predicted using test data and RMSE value calculated which is 12.2

Output:

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.917
Dependent Variable: y                AIC:                8073929.1755
Date:               2022-11-24 18:08 BIC:                8074083.1320
No. Observations:   1027674          Log-Likelihood:     -4.0370e+06
Df Model:           12               F-statistic:        9.469e+05
Df Residuals:       1027661          Prob (F-statistic): 0.00
R-squared:          0.917            Scale:              151.20
------------------------------------------------------------------
             Coef.    Std.Err.       t       P>|t|     [0.025    0.975]
------------------------------------------------------------------
const       17.1589     0.1560    109.9738   0.0000   16.8531   17.4647
x1          -0.0147     0.0024     -6.2478   0.0000   -0.0193   -0.0101
x2          -0.3297     0.0028   -118.7968   0.0000   -0.3352   -0.3243
x3           0.0481     0.0014     35.0416   0.0000    0.0454    0.0508
x4           0.0001     0.0000     14.7496   0.0000    0.0001    0.0001
x5          -0.1022     0.0023    -44.8329   0.0000   -0.1067   -0.0978
x6           0.0001     0.0000     13.8787   0.0000    0.0001    0.0001
x7           0.0126     0.0008     16.7893   0.0000    0.0111    0.0141
x8          -0.0001     0.0000    -16.5154   0.0000   -0.0002   -0.0001
x9          -0.0177     0.0008    -23.5476   0.0000   -0.0192   -0.0162
x10         -0.0064     0.0005    -13.1893   0.0000   -0.0074   -0.0055
x11         -0.6021     0.0022   -270.6185   0.0000   -0.6064   -0.5977
x12          0.8904     0.0003   3355.2106   0.0000    0.8899    0.8910
------------------------------------------------------------------
Omnibus:            428741.106     Durbin-Watson:        1.747
Prob(Omnibus):      0.000          Jarque-Bera (JB):     300360179.074
Skew:               0.537          Prob(JB):             0.000
Kurtosis:           86.746         Condition No.:        234126
==================================================================
* The condition number is large (2e+05). This might indicate
strong multicollinearity or other numerical problems.
```

```
r2 socre is  0.9193862775736159
mean_sqrd_error is== 149.05433925173472
root_mean_squared error of is== 12.20878123531316
```

An r-squared of 91% reveals that 91% of the variability observed in the target variable is explained by the regression model.

Limitations:

1. In the dataset multiple variables were highly correlated which lead to multicollinearity.
2. Data given is just of 2 months, so finding out trends related to flight delay are very inadequate.
3. Data entry issue for a few rows in a cancelled column.

Recommendations:

1. Inorder to find delays/cancellations, weather data is the most important factor missing.
2. Along with this columns like NAS delay , carrier delay, security delay are missing which also contributes to delay/ cancellation factor will help to understand the reasons for delay/ cancellation.