GNR : Assignment 2
Flight Status Estimation

Meeta, 180070034

April 7, 2020

Contents

1 Visualization of Data-Set

1.1 Division on the basis of Status

From the given data-set we observe that 80.55% of flights are on-time and 19.45% of flights are delayed. Numerically 428 flights were delayed and 1773 flights were on time.

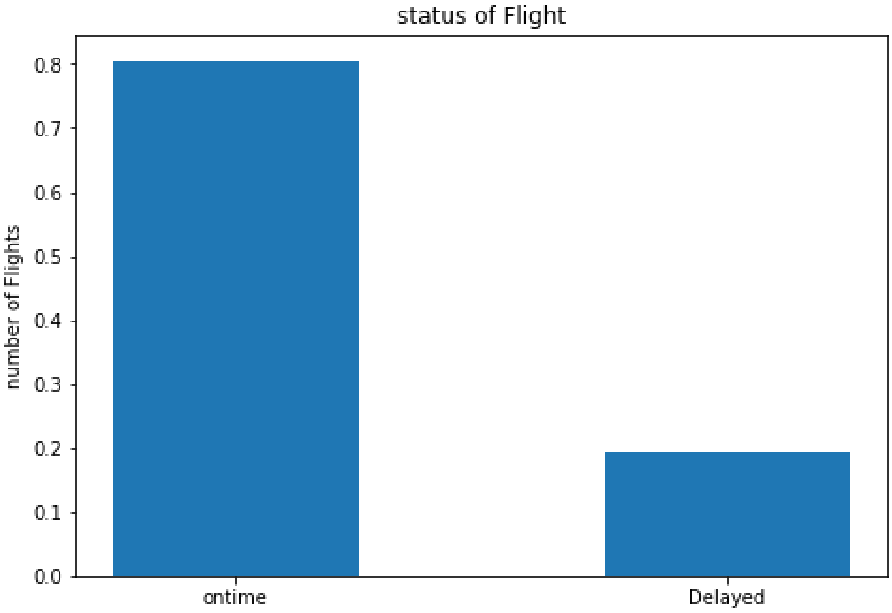

Figure 1: % of flights

1.2 Division on the basis of day of Week

From the given data-set the below table shows number and percentage of flight delay on each day of week.

Day	Number of Fl. delay	% delay of total Fl.
Mon	84	19.63
Tues	63	14.72
Wed	57	13.32
Thurs	57	13.32
Fri	75	17.52
Sat	24	5.61
Sun	68	15.89

Table 1: delay per day of week

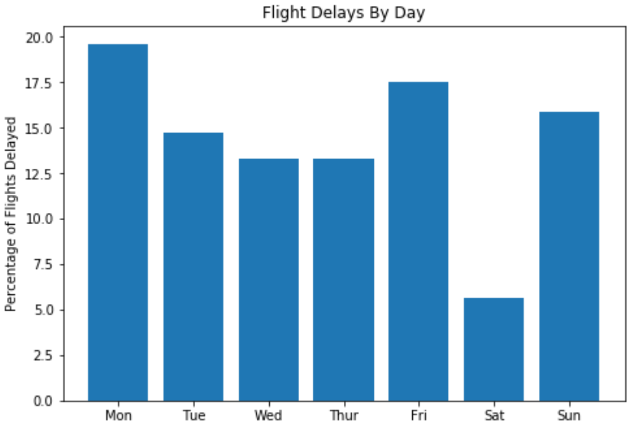

Figure 2: % of flight delay

1.3 Division on the basis of Carrier

From the given data-set the below table shows number and percentage of flight delay with carrier of flight.

Carrier	Number of Fl. delay	% delay of total Fl.
CO	26	6.07
DH	137	32.01
DL	47	10.98
MQ	80	18.69
OH	4	0.94
RU	94	21.96
UA	5	1.17
US	35	8.18

Table 2: delay per carrier

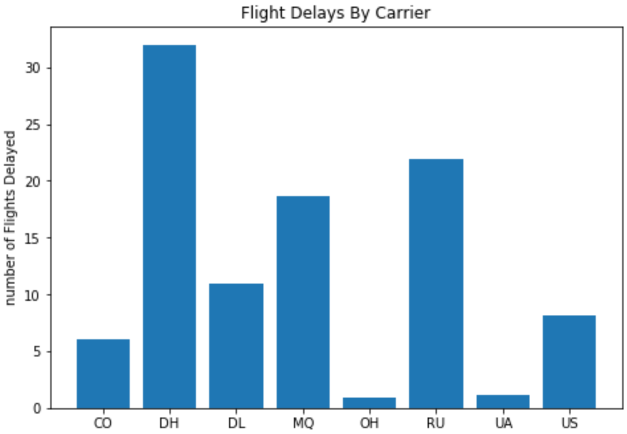

Figure 3: % of flight delay

1.4 Division on the basis of Origin

From the given data-set the below table shows number and percentage of flight delay with origin of flight.

Origin	Number of Fl. delay	% delay of total Fl.
BWI	37	8.65
DCA	221.0	51.64
IAD	170.0	39.72

Table 3: delay per origin

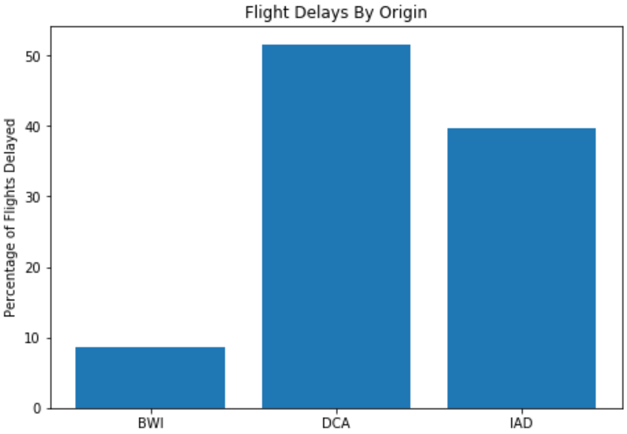

Figure 4: % of flight delay

1.5 Division on the basis of Destination

From the given data-set the below table shows number and percentage of flight delay with Destination of flight.

Origin	Number of Fl. delay	% delay of total Fl.
EWR	161	37.62
JFK	84	19.63
LGA	183	42.76

Table 4: delay per destination

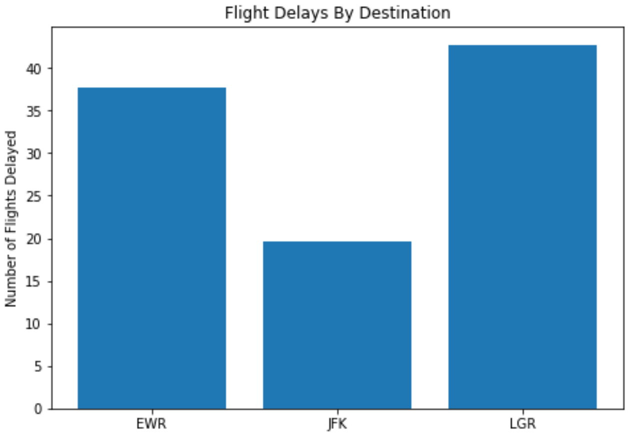

Figure 5: % of flight delay

1.6 Division on the basis of Weather

From the given data-set the below table shows number and percentage of flight delay with weather.

Weather	Number of Fl. delay	% delay of total Fl.
Good	32	7.48
Bad	396	92.52

Table 5: delay per weather

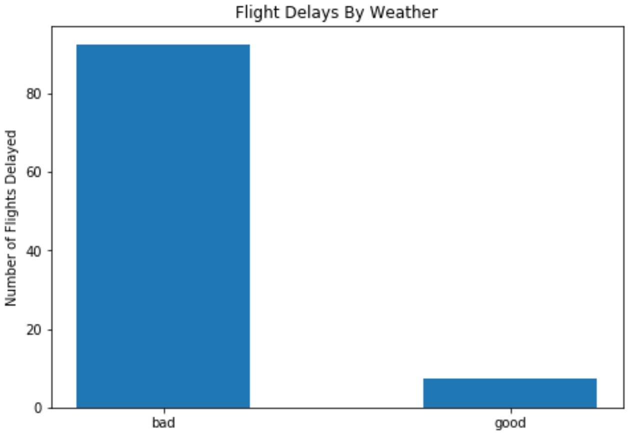

Figure 6: % of flight delay

1.7 Division on the basis of Day of Month

From the given data-set the below table shows number and percentage of flight delay with Day of Month.

Date of month	Number of Fl. delay	% delay of total Fl.
1	0	0
2	6	1.4
3	5	1.17
4	21	4.91
5	29	6.77
6	9	2.1
7	17	3.97
8	9	2.1
9	13	3.04
10	5	1.17
11	6	1.4
12	10	2.34
13	16	3.74
14	10	2.34
15	24	5.61
16	30	7.01
17	7	1.64
18	27	6.31
19	11	2.57
20	7	1.64
21	9	2.1
22	9	2.1
23	11	2.57
24	5	1.17
25	14	3.27
26	34	7.94
27	31	7.24
28	21	4.91
29	15	3.51
30	15	3.51
31	2	0.47

Table 6: delay per date

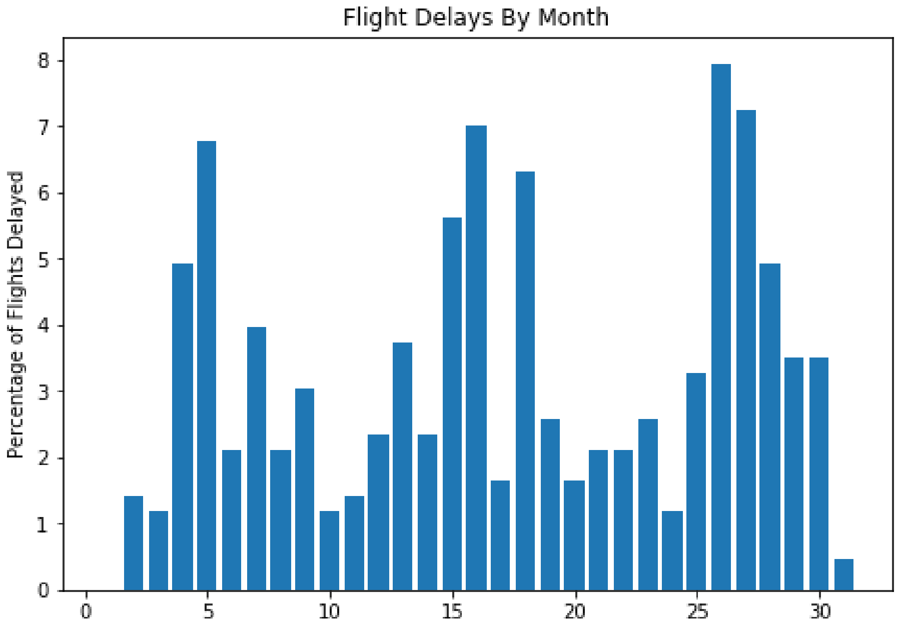

Figure 7: % of flights

1.8 Division on the basis of Distance

From the given data-set the below table shows number and percentage of flight delay with distance.
Maximum occurs at distance of 214Km.
Minimum occurs at distance of 184Km.

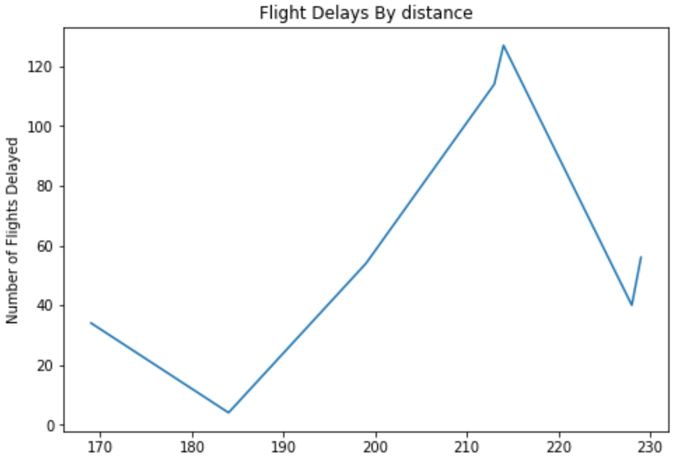

Figure 8: % of flights

1.9 Division on the basis of Scheduled Departure time

From the given data-set, graphs is made between number of flight delay and Scheduled Departure time.
Maximum number of delay occurs at scheduled departure time of 1455.

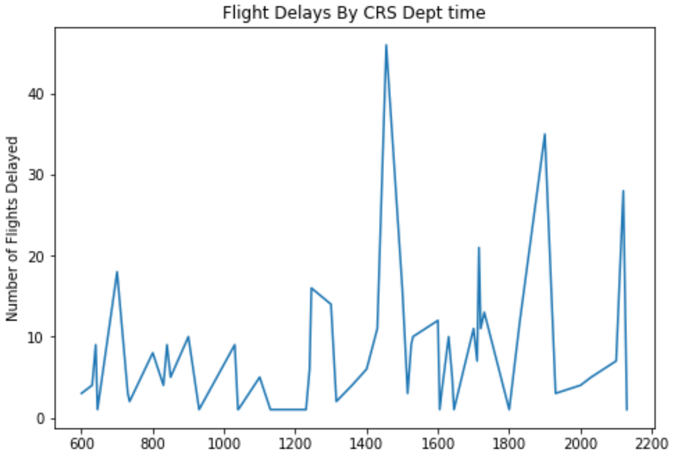

Figure 9: % of flights

1.10 Division on the basis of Departure time

From the given data-set, graphs is made between number of flight delay and Departure time.

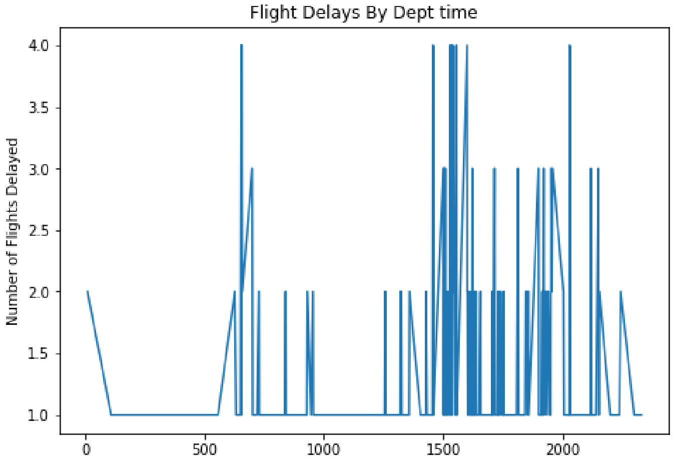

Figure 10: % of flights

2 Logistic Regression

2.1 Data Processing

The target variable (Flight delay status) is categorical type. The status labeled ‘delayed’ is assigned ‘1’ and ‘ontime’ is assigned ‘0’.

Weather was changed to numerical form, good weather was denoted by ’0’ and bad weather was denoted by ’1’.

Distance crs-dept-time and dept-time are passed as a continuous variable (numerical data) to the model.

Day of week, day of month, carrier, destination, origin, tail number, flight status was converted into numerical data using one-hot key where each column corresponded to one one attributes of the feature.

The data-set was applied to logistic regression model. After changing the to categorical data of to numerical we find that there are 740 features.

2.2 Training the model

Data-set with 740 features was trained via logistic regression and implemented via sci-kit library of Python. The model applied gives outcome using the following function:-

$$\sigma(x) = \frac{1}{1 + e^{-Wx}} \tag{1}$$

When Wx is large and positive then value of function will be close to 1 and when Wx is large and negative then the function value is close to 0.
Data was split into 60:40 = Train: test ratio and fed into the model.

Results

1. The model works fair enough and got the accuracy of approximately 89.33%.
2. The coefficients of scheduled and actual departure time were equal and opposite in sign, which indicated that their difference is important.
3. A column for difference in scheduled and actual departure time was added and departure time was removed and model was trained again, this resulted in accuracy of 90.6%

3 Interpreting the Model

1. Major contribution comes from weather, if the weather is bad then the flight is most likely to get cancelled. This also can be observed from the graph between delays and weather.
2. These 740 features tends to over-fit the data. The model barely depended upon features such as tail-num or flight-num as they showed larger rfe ranking.
3. When tail-num or flight-num were removed the accuracy comes out to be 91.4%. Which means that the model barely depended on them.
4. Destination and Origin had little variance, as concluded from EDA. Their low coefficients suggest that their weight is moderate and the outcome has moderate dependence on these parameters.
5. Some days of week and months shows higher coefficient than other days suggesting more delay has occurred on those days.
6. The coefficients of scheduled and actual departure time were equal and opposite in sign, which indicated that their difference is important. When model was trained using difference of timings and scheduled time, the coefficient of scheduled time was very low.
7. The coefficient of distance was low but when multiplied with distance which are in order of 100's gives a significant contribution in predicting the delay.

4 Feature Selection

1. The Tail-num and flight-num didn't any any value to the model and increased the number of features by about 600, hence resulting in over-fitting the data set. The coefficients suggests that they are insignificant and hence they were removed from the data-set.
2. I have introduced a new feature which signifies origin and destination pairs and then converted these 8 pairs via one hot labeling.
3. As origin and destination is fixed thus is the distance between them, hence distance acts as a redundant feature.
4. Major contribution comes from weather, if the weather is bad then the flight is most likely to get cancelled.
5. The model gives a significant drop in accuracy if the difference in time is dropped as a feature hence is a major contribute.
6. Days of month and week shows moderate dependence on the model as interpreted from the model.
7. Features selected from observations and trends are:-
   (a) Delay time.
   (b) Origin- Destination pair
   (c) Weather
   (d) Carrier
   (e) Days of Week
   (f) Days of Month

Selecting these features reduced number of features from 740 to 56 features, giving accuracy of 91.83%.

5 Fitting new model - Decision Tree

For every tree-based algorithm, there is a sub-algorithm that is used to split the data-set into two bins based on a feature and a value. The splitting algorithm considers all possible splits (based on all features and all possibles values for each feature) and finds the most optimum split based on a criterion.

If a continuous variable is chosen for a split, then there would be a number of choices of values on which a tree can split and in most cases, the tree can grow in both directions.

Categorical variables have only a few options for splitting which results in very sparse decision trees. One-hot encoding has just two levels. The trees generally tend to grow in one direction because at every split of a categorical variable there are only two values (0 or 1).

For the reasons mentioned above, I converted the categorical information into labels (nominal value).

Conclusion

1. Accuracy for the selected features: 85.81%
2. Tree Depth: 19

6 Ideal Conditions for an ON-TIME Flight fro DC to NYC

From equation (1), we know that for the flight to be on time the $\sigma(x)$ must be 0, hence Wx ought to be large and negative. I trained the model with weather, origin, destination, day of week, difference in time, scheduled departure time and distance, i.e. total of 17 features.
I printed the coefficient of the model, which are as follows-

Parameter	coff
Time diff	1.43223803e-01
Arrival	2.44951763e-04
Distance	-1.31851603e-02
Weather	8.95293095e-01
LGA	-2.74959636e-01
DCA	2.18445879e-01
BWI	-5.97363686e-01
Saturday	-2.02737567e-01
US	1.33640245e-01

Table 7: coff of model

For flight to be on-time we select the feature such that Wx calculated results in maximum negative value. Or can be observed fro the graphs and trends.
Hence the result obtained is;-

1. Weather = good
2. Carrier - US
3. Day of Week = Saturday
4. Origin = DCA
5. Destination = LGA
6. Scheduled time = 1100 hrs

From the graph between scheduled time and delay, we see that the minimum number of delay occur when the scheduled time is 1040 hrs amongst other possible fights for the given origin and destination.
As origin and destination is fixed, and for Thursday, carrier available are US, MQ, DL, and among these three US has the least number of delays as observed from the data set.

7 Bonus Questions

Q1 Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY and EDITH.
Ans: Few other examples include DUMMY, VERONICA, KAREN, TADASHI .

Q2 Data Processing Inequality
Ans: The contents of the signal/data cannot be increased by a local physical operation (in some sense information can be only lost after processing the data.)

Q3 In Star Wars Universe, X was a Sith philosophy mandating that only two Sith Lords could exist at any given time: a master to represent the power of the dark side of the Force, and an apprentice to train under the master and one day fulfill their role.Who is X?
Ans: X is Darth Bane: Rule of Two

Q4 In Star Wars Universe, name this robotic duo :-
Ans: Left one is R2-D2, Right one is C-3PO

Q5 What is special about Cards against Humanity: Black Friday 2019?
Ans: They teach a computer to write cards, using Artificial Intelligence (trained on their brain storming session) and compete with the writers for the most popular collection of the cards.