GNR : Assignment 2
Flight Status Estimation

Avyakta Wrat, 180070010

April 6, 2020

Contents

1 Exploratory Data Analysis

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.

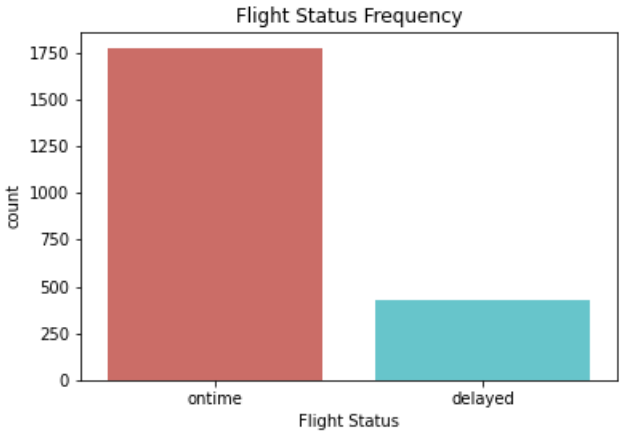

Figure 1: Count of flights

1.1 Days of Week

The table below shows number and percentage of total 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: Weekly Delay

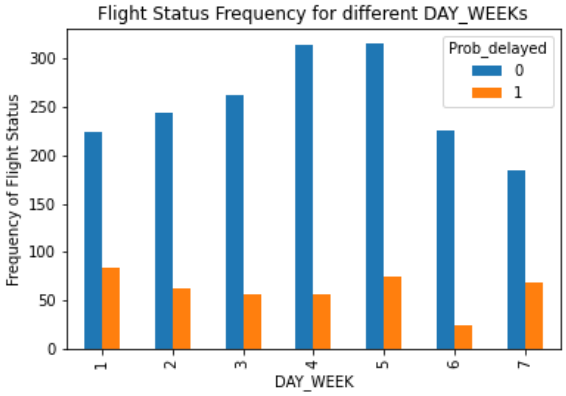

Figure 2: Frequency of flight status

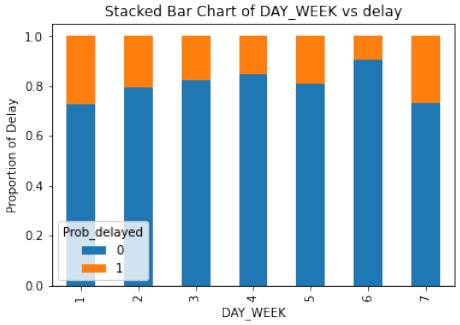

Figure 3: Stacked bar chart for each Carrier

1.2 Carrier Frequency

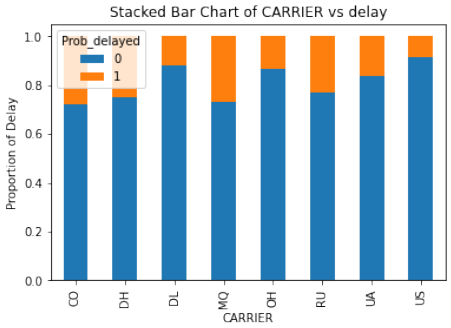

Figure 4: Stacked bar chart for each Carrier

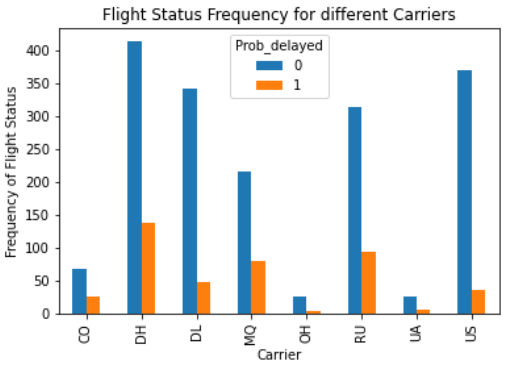

Figure 5: Frequency of flight delay

1.3 Origin & Destination

The table below shows the number and percentage of total 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 2: Delay with Origin

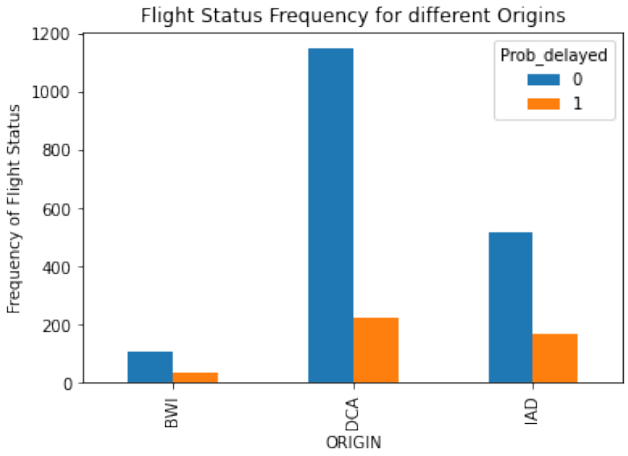

Figure 6: Frequency of flight delay

The table below shows number and percentage of total flight delay with Destination of flight.

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

Table 3: Delay with Destination

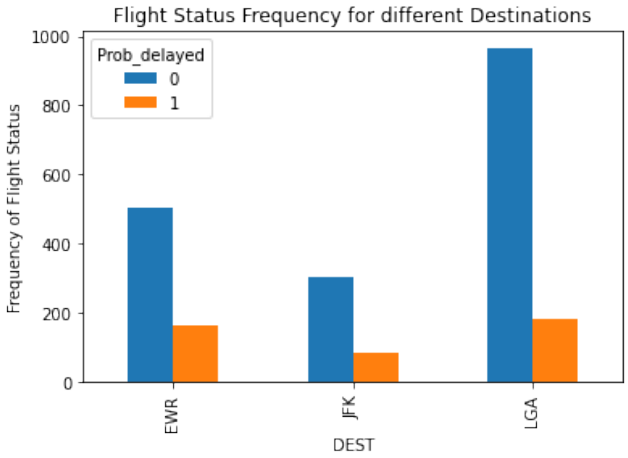

Figure 7: Frequency of flight delay

The ratio is better observed in the stacked bar chart where total number of flights for each day is scaled to 1.0 and delayed and ontime flights are shaded accordingly.

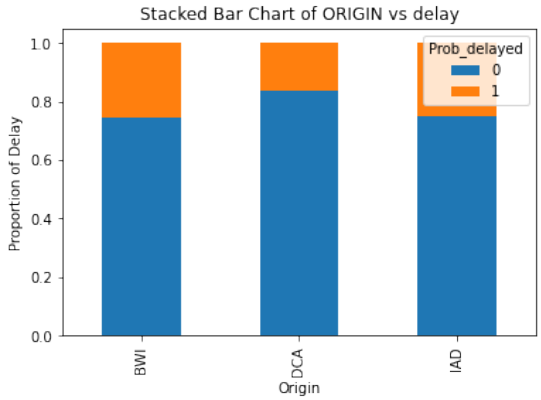

Figure 8: Ratio of flight delayed

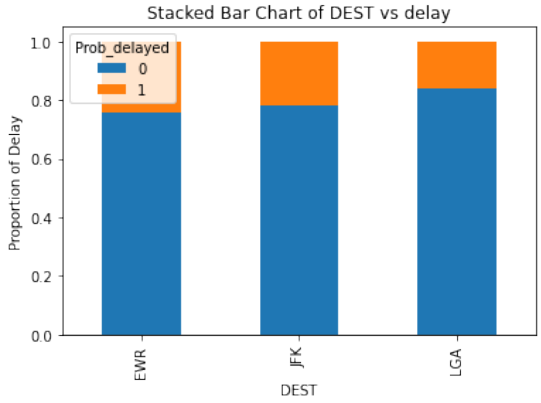

Figure 9: Ratio of flight delayed

1.4 Weather

The table below shows the 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 4: Delay due to Weather

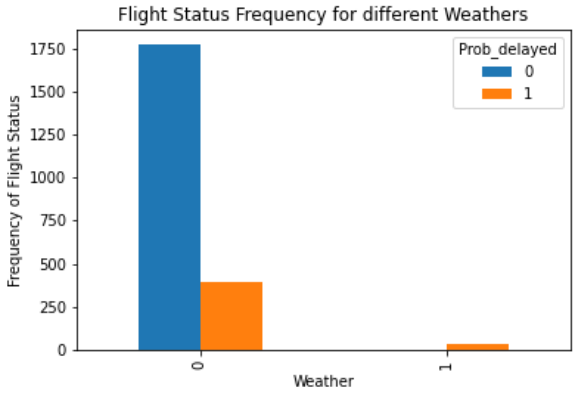

Figure 10: Frequency of flight delay

1.5 Time-difference in Departure

This was divided in 6 parts: Flights that departed before time, in the first five minutes of scheduled departure time, next 10 minutes and so on. The percentages are calculated out of total flights departed in individual time interval.

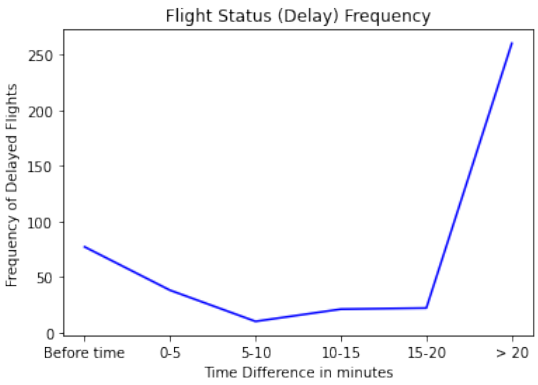

Figure 11: Frequency of flight delayed

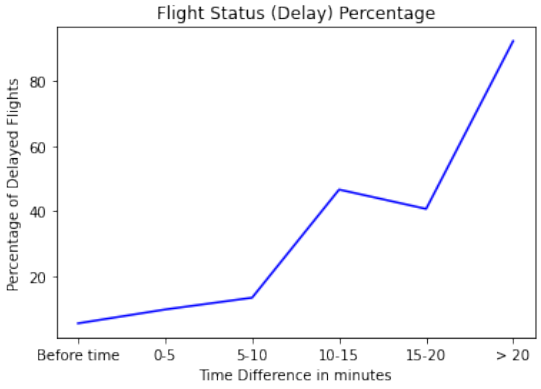

Figure 12: Percentage of flight delayed

1.6 Day of Month

The table below 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

Date of month	Number of Fl. delay	% delay of total Fl.
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 5: Delay with date

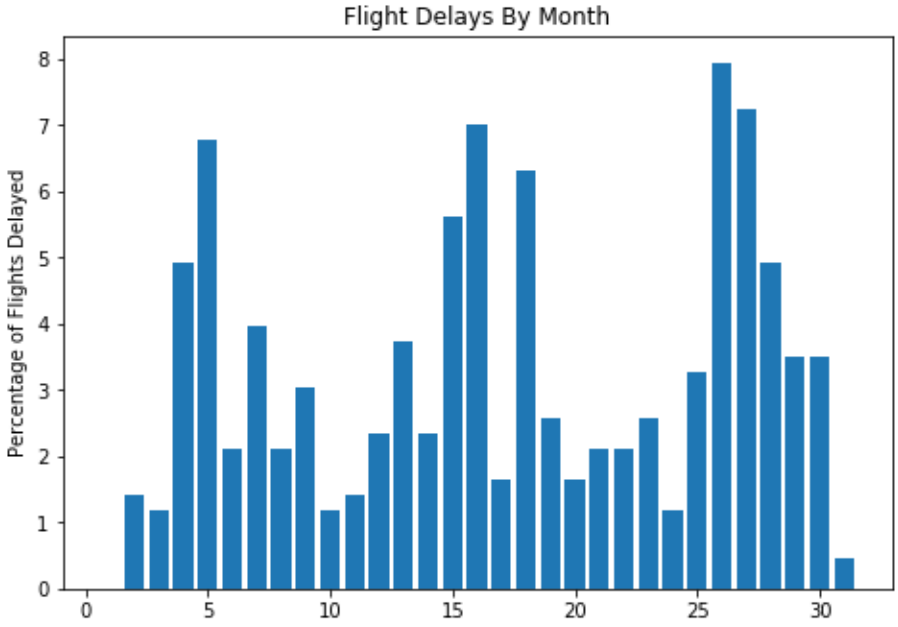

Figure 13: % of flights

1.7 Distance

Maximum count of flight delay occurred at a distance of 214 km.
Minimum count of flight delay occurred at a distance of 184 km.

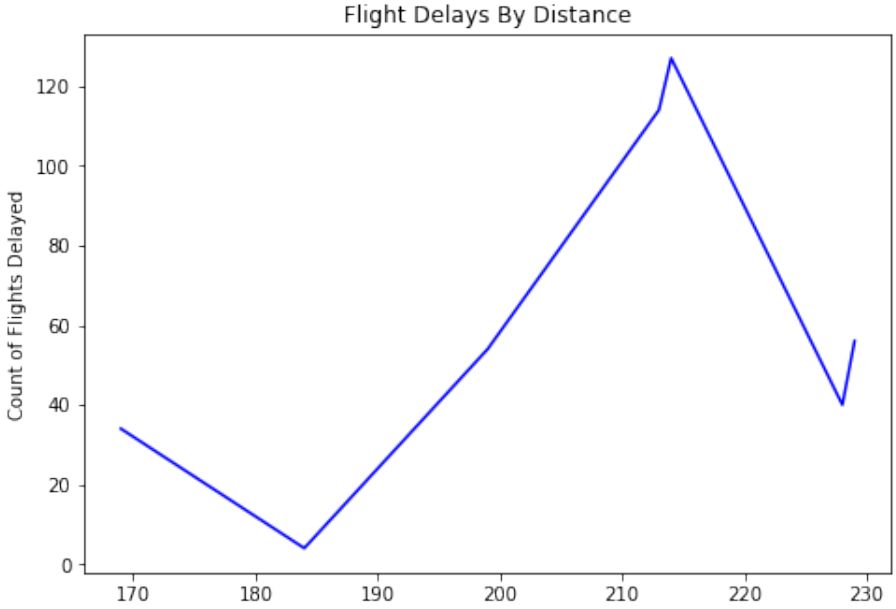

Figure 14: Count of flights delayed

1.8 Scheduled Departure Time

Maximum number of delay occurred at scheduled departure time of 1455.
The graph below captures the frequency and percentage of flight delay with x-axis

as minutes taken from 0000 hrs. This done to ensure continuity in the x-axis data. 0 - 24 hours now map to 0 to 1440 minutes (24*60).

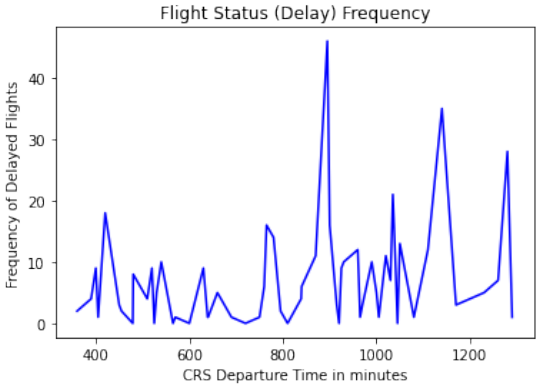

Figure 15: Frequency of flight delay

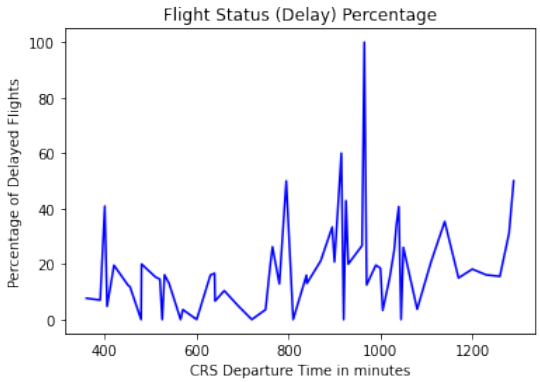

Figure 16: Percentage of flight delayed

1.9 Departure time

Similar analysis as the scheduled departure time was followed in this. However the graph were too random. Hence I have converted them into hour blocks where all the flights in a one hour range are counted and analysed as shown is 18 and 19.

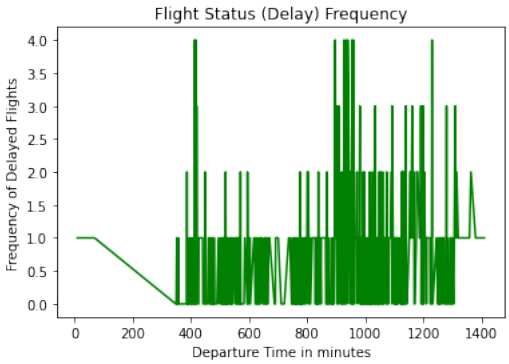

Figure 17: Frequency of flight delay

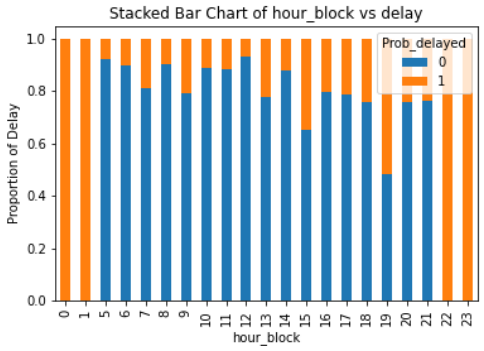

Figure 18: Ratio of flight delay in hour blocks

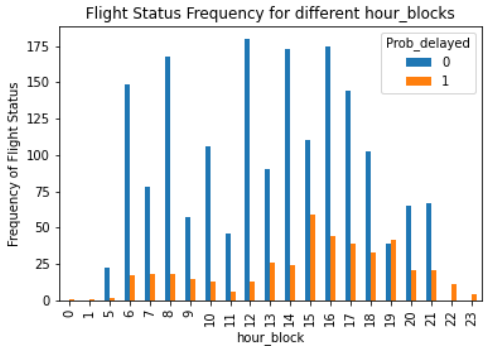

Figure 19: Frequency of flight delay in hour blocks

1.10 Inferences from Exploratory Data Analysis

1. Classification on the basis of days give low to moderate variation.
2. Carriers show moderate variation but as they are 8 in number, it can increase the dimensions of the feature when one hot encoded.
3. As observed in the stacked figures 8 and 9, destination and origin show low variations in the ratio of flight delayed to total flights arriving/leaving a destination/ origin. Hence they may not serve as a good estimator.
4. Weather is a good classification criterion due to high variations. However there are very few datasets available for bad weather.
5. Time difference, Distances and Departure time (hour blocks) show reasonable variations.

2 Logistic Regression

2.1 Data Preprocessing

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

A new feature (Time difference or delay time) was added to the model, which is just the difference between the departure time and correct departure time in minutes. This is continuous valued.

The scheduled departure time and departure time are converted to 1440 minutes in a day so as to make their Exploratory Data Analysis easier. Departure time was however changed into hour blocks (24 hour blocks in a day) and then one hot encoded. The reason for one hot encoding and not passing it as numerical value is that I feel all the hour blocks independently decide if the flight will be delayed or not. There shouldn't be any correlation with the increasing departure time to the outcome. However time diff is passed as a continuous variable as the flights are likely to be delayed with increasing difference.

Other categorical data like weather, day of week, day of month, carrier, destination, origin, tail number, flight number are one hot encoded.

Distance is passed as a continuous variable (numerical data) to the model.

Scheduled Departure time was dropped as departure time and time difference together also account for the scheduled time, making it a redundant feature.

Lastly, flight date was also dropped from the features.

2.2 Training the Model

The data-set was applied to logistic regression model in Python.

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

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.
Train-test split ratio of 60:40 was used in the logistic regression model and all features apart from date of flight was considered.

Results

1. Accuracy of the Model: 91.60% with 729 features.

2. I tried passing all the features to the Decision Tree Classifier. Though an accuracy of 89.44% was reported the depth of the tree was 56 which suggested over-fitting probably due to one hot encoding of categorical features like flight number and tail number.

3 Interpreting the Model

One hot encoding of Flight Number and Tail number have increased the dimensionality by a great extent leading to 729 features in the model. Few of the coefficients have been listed below.

Feature	Coefficient
Time diff / delay	0.15457
DISTANCE	-0.00886
Carrier is CO	0.52360
Carrier is DH	-0.37663
Carrier is DL	0.139187
Carrier is MQ	0.331800
Carrier is OH	-0.80128
Carrier is UA	0.04499
Carrier is US	0.06198
hour block is 1	0.0
hour block is 6	0.61347
hour block is 13	-0.88524
hour block is 16	-0.63628
hour block is 17	-0.03319
Day month is 16	0.89368
Day month is 17	0.63918
Day month is 18	0.81608
Day month is 19	-1.0387
Day month is 20	-0.4934
Day month is 21	-0.0082

Feature	Coefficient
Dest is EWR	0.27444
Dest is JFK	-0.13554
Dest is LGA	-0.15579
Origin is BWI	-0.21119
Origin is DCA	0.011103
Origin is IAD	0.18319
Weath is 0	-0.77946
Weath is 1	0.76257
Day is 1	-0.08232
Day is 2	-0.08430
Day is 3	0.30879
Day is 4	-0.07255
Day is 5	-0.01485
Day is 6	-0.34001
Day is 7	0.26836
Day month is 1	-0.78474
Day month is 2	-0.9843
Day month is 6	-0.8670
Day month is 7	-0.4936
Day month is 8	-0.1654

1. Major contribution comes from weather, if the weather is bad then the flight is most likely to get cancelled. This is very evident as the coefficient is high for the same (around 0.7 in magnitude in both).

2. Destination and Origin had little variance, as concluded from EDA. Their low coefficients suggest that their weight is low and the outcome has low dependence on these parameters.

3. Distance has a low coefficient. However, this does not indicate low dependence on this feature as distance wasn't normalized before passing it as a feature. Distance are in the order of 10^2. Scaling the coefficient by multiplying with this value increases it to 0.8 in magnitude which is quite significant and matches conclusion drawn from the EDA analysis.

4. Dummy variables of Departure time hour block have varying coefficients.

5. Tail number and flight number have increased the feature dimension to a great extent, making it difficult for us to analyze every coefficient individually. Each of their dummy variable show a large variation in coefficient.

4 Feature Selection

From the inferences drawn using EDA and Logistic Regression Model, the following can be concluded:

1. Origin and destination together capture the information encoded in distance or vice versa. The conclusions made during the coefficient analysis suggest distance to be a better factor over origin a destination.

2. There are very few datasets when they are spread over the day of month. Also, day of month and day of week are overlapping features

3. The features flight number and tail number increase dimensionality with one hot encoding and due to low number of dataset, this can lead to over-fitting. Hence these features can be dropped.

4. There is high dependence on weather time difference between departure and scheduled departure time. Hence these features are important.

5. Features selected are:

   (a) Day of month

   (b) Distance

   (c) Departure time in hour blocks

   (d) Weather

   (e) Delay time

Results

1. The total number of features reduced to 57 with one hot encoding of Day of Month, Hour blocks and Weather. Distance and delay time as continuous variable.

2. Accuracy of the Logistic Regression: 91.94%

5 Fitting new model - Decision Tree

The decision tree was fit using the selected features and its accuracy and depth have been reported below. However, none of the features were one-hot encoded to avoid over-fitting and growth of tree in one direction. When continuous variables are used, tree generally splits in both directions.
For the reasons mentioned above, I converted the categorical information into labels (nominal value).

Results

1. Accuracy for the selected features: 87.29%

2. Tree Depth: 23

6 Predicting ideal weather conditions for a flight

- Weather : Good
- Day : Saturday
- Time : Between 10 and 11 in the morning
- Carrier : US

7 Bonus Questions

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

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.
Ans: X is 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.