GNR 652 : Assignment 2
Flight Status Prediction

Devank Rajvanshi, 18D100009

April 7, 2020

Contents

1 Exploratory Data Analysis

We observe that out of 2201 Flights, according to the data given, **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.

To get more insight on the factors responsible for delay, various data representations were made as follows :

1.1 Carrier

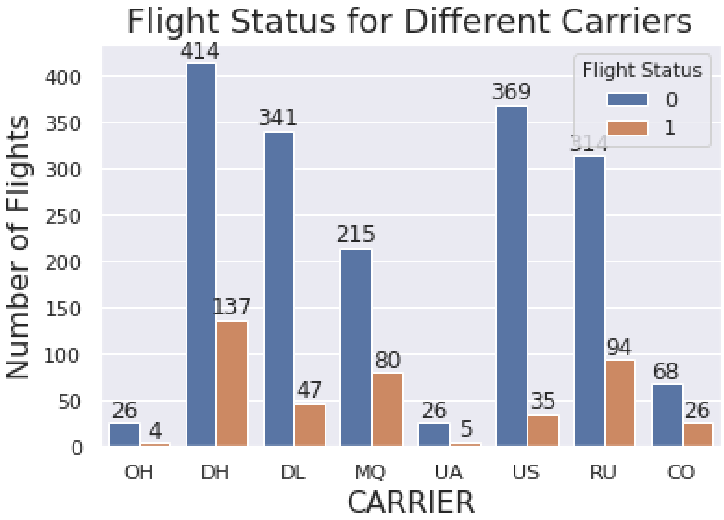


Figure 1: Frequency vs Carrier

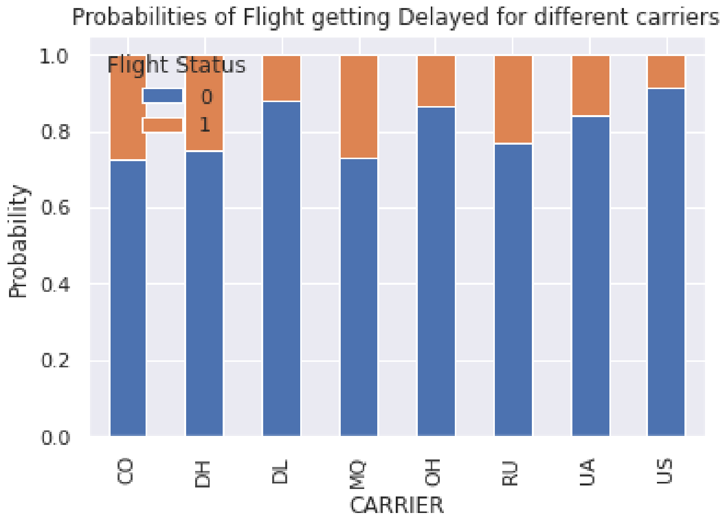


Figure 2: Probabilities vs Carrier

We observe that **US** Carrier service is not only performing really well *(Least probability of getting delayed)* but also one of the most used Airline service. DL is the next airline advised to be considered.

On the other hand **MQ** Flights have most chances to get delayed.

1.2 Origin and Destination

There are 3 Airports in Washington DC which is our Origin City and 3 Airports in New York City which is our Destination.
Thus there can be 9 possible (though 8 practically used) ways to travel the journey on the basis of Route taken. Dependence of Flight Status based on these routes are as follows :

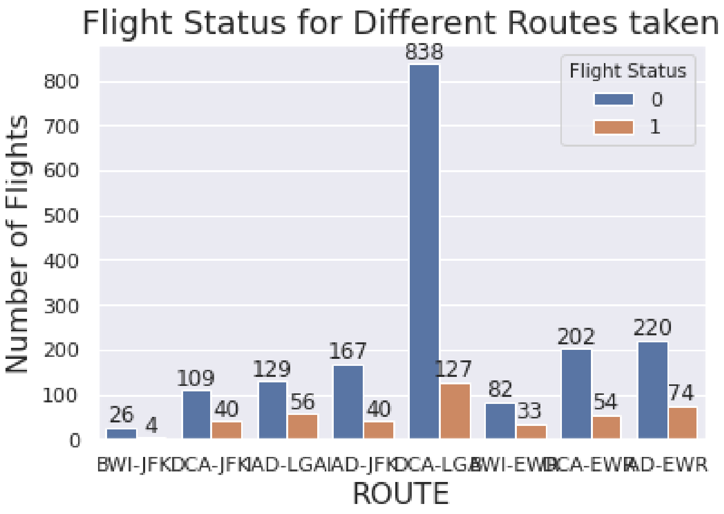


Figure 3: Frequency vs Route Taken

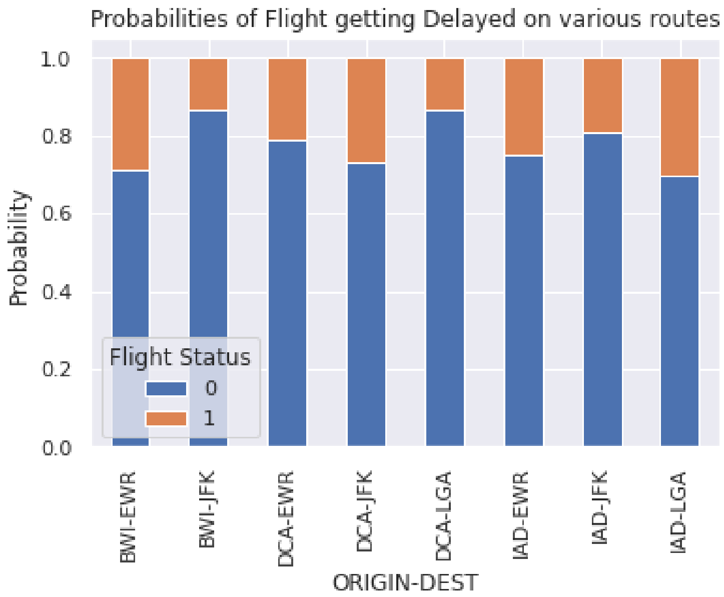


Figure 4: Probabilities vs Route Taken

We observe that **DCA - LGA** is not only people's most popular choice of route but also least probable to get delayed. DCA-EWR and IAD-JFK are next advised Routes.
On the other hand, (BWI - EWR) is very less preferred and flight is most likely to get delayed compared to other options available.

1.3 Distance

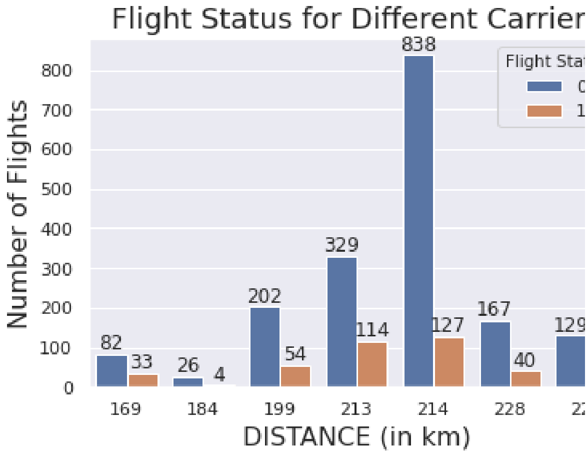


Figure 5: Frequency vs Distance

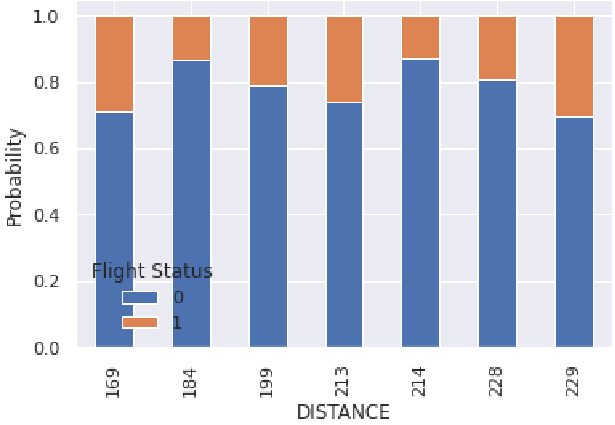


Figure 6: Probabilities of Flight getting Delayed vs Distances

214 km journey is most preferred and least probable to get delayed whereas 169 km is least advised. If we look closely, these graphs are closely related with the previous variable of Route. It is because to Routes have same Distance and thus there are 7 categories instead of 8 in this visualisation. This fact will help us do the feature selection later.

1.4 Weather

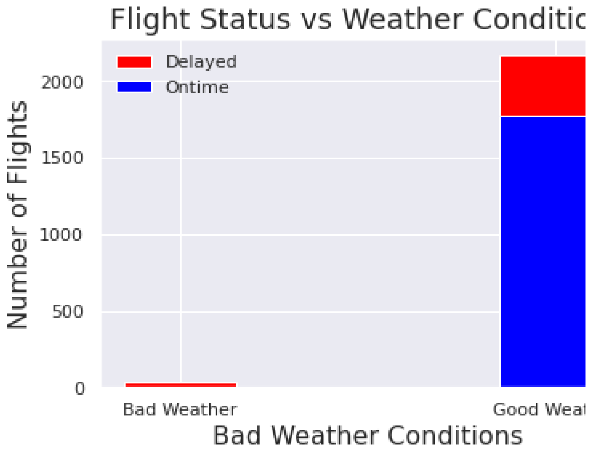


Figure 7: Frequency vs Weather

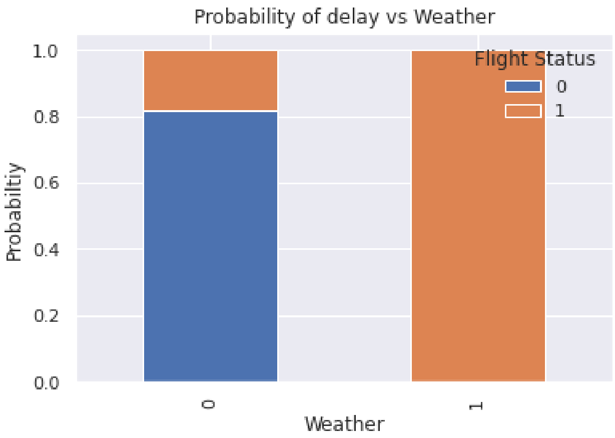


Figure 8: Probabilities of Flight Status

We observe that Weather is a great Classifier because if Weather is bad, flights are always delayed. Though data for bad weather flights is less but we can safely assume the above mentioned fact. This is because, intuitively, bad weather makes, working at airports, operation and maneuvering of Flights difficult. Thus it is advised to travel under Good Weather Conditions. :)

1.5 Day of the Week

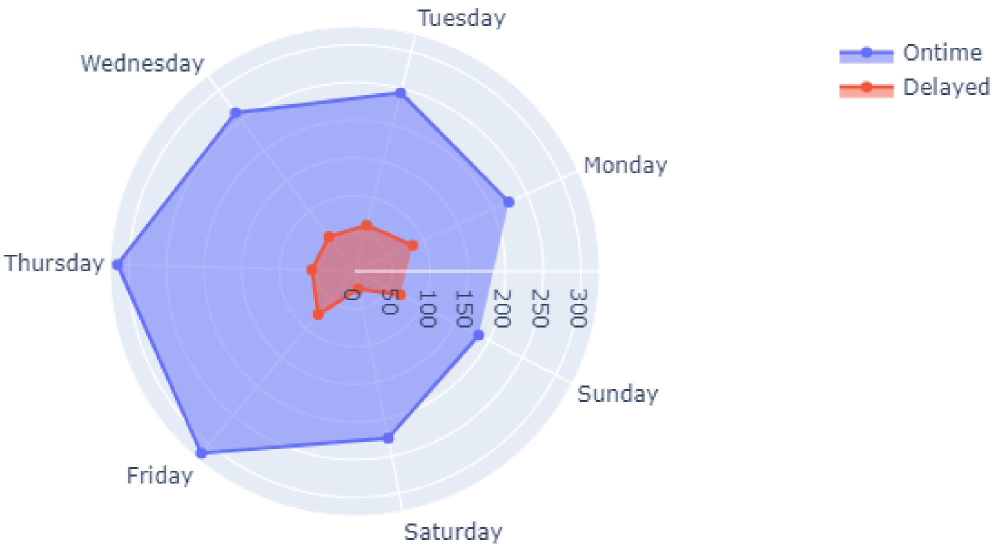


Figure 9: Frequency vs Day of travel

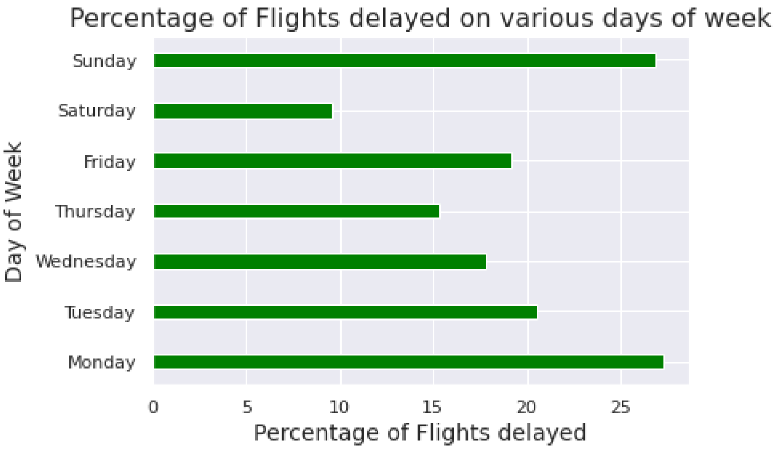


Figure 10: Percentage of Flights Delayed each Day

We observe that **Saturday** is the best day to travel because more or less same number of flights fly on that day but probability of them getting delayed is least. Contrary to this, Sunday and Monday are worst day to travel because even though with similar number of Flights being flown, more than one-fourth of them are delayed.

1.6 Time Interval of the Day

Departure Time of each Flight was allotted a Time division of 1 hour in which it left. Flight Status observed is as follows:

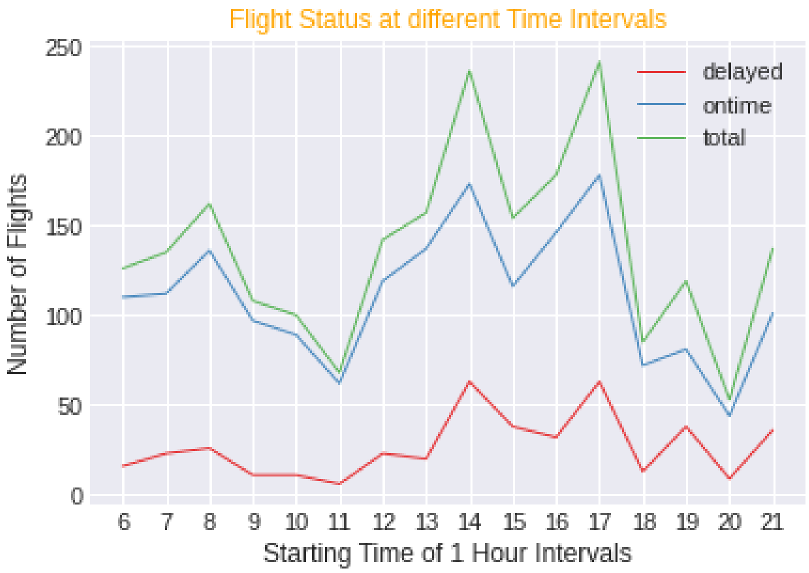


Figure 11: Frequency vs Time Interval

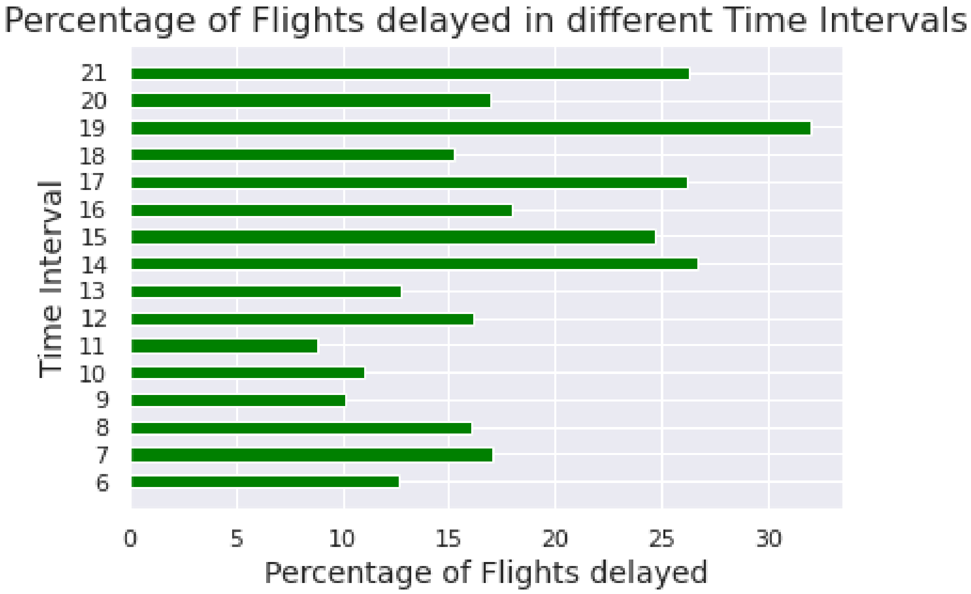


Figure 12: Probability of Flight getting delayed in particular Time Interval

We observe that (19:00 - 19:59) at night is the worst time slot to travel even though many flights are available this time. Maybe that can be the reason of delays. Thus evenings should be avoided.

Whereas **11:00-11:59** in the late morning is the best time slot to travel if we want to reach on time as maximum flights are reached ontime though availability of flights in this slot is less.

1.7 Delay in Departure

Delay in Departure was calculated by finding the Difference in CRS-DEPT-TIME and DEPT-TIME.

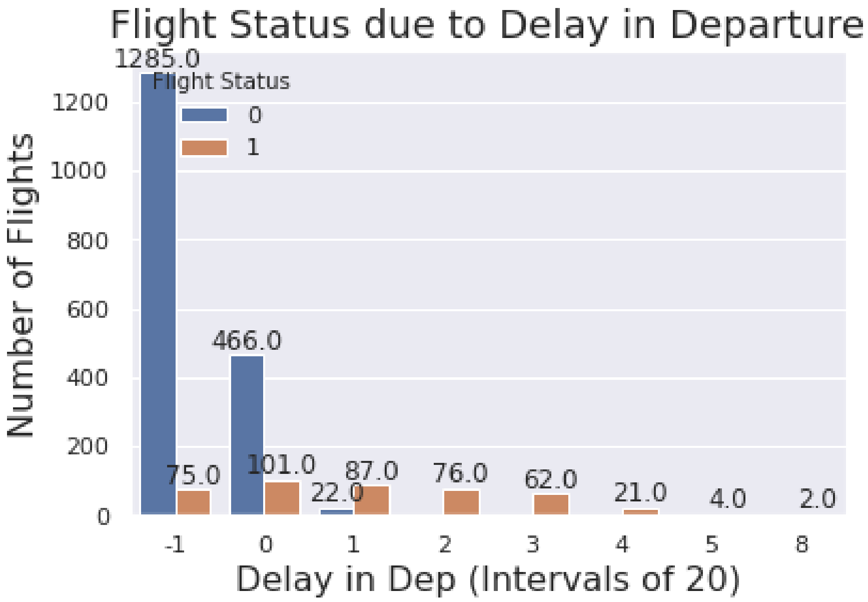


Figure 13: Frequency vs Time Interval

Using the calculations of Gini Index for this standalone classifier, it turns out if Flight is delayed by more than 30 minutes from the scheduled departure time, Flight is delayed, as evident from the graph too.

1.8 Day of Month

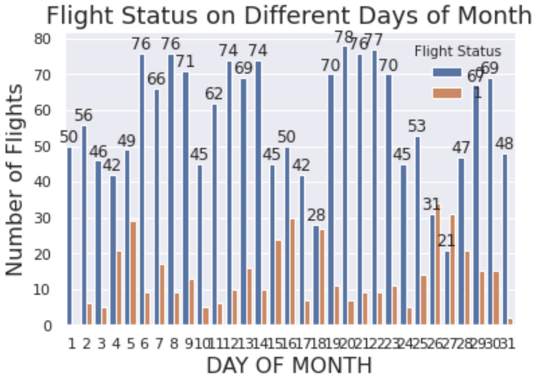


Figure 14: Frequency vs Day of Month

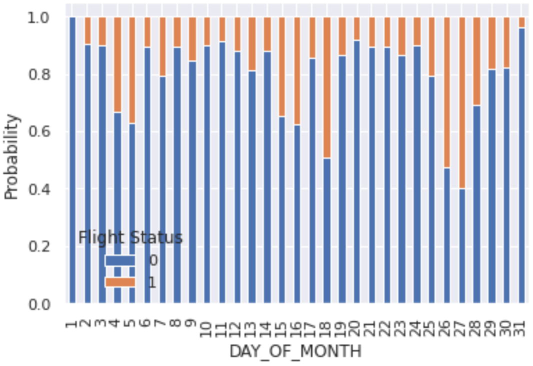


Figure 15: Probabilities vs Day of Month

It turns out Day of Month has very varying Data and nothing significant can be extracted from it. As a feature for our model, intuitively, its impact can be accounted by Days of Week. Also, number of Flights and they getting delayed depends a lot on external factors like public holidays, festivals, whether its weekend or weekday, political conditions etc. Thus its not a great variable to rely on.

1.9 Flight Number

There were variety of Flight Numbers available. I Filtered out the Flight Numbers that were most and least probable to get delayed.

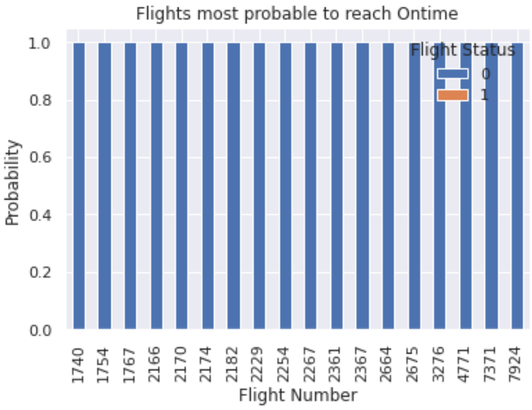


Figure 16: Frequency vs Day of Month

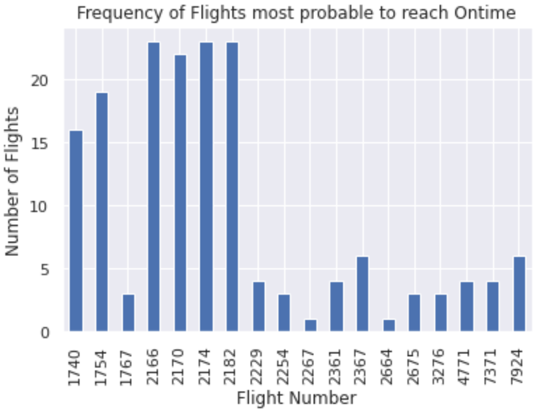


Figure 17: Probabilities vs Day of Month

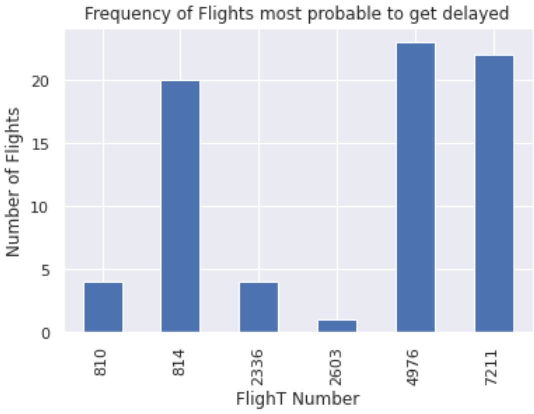


Figure 18: Frequency vs Day of Month

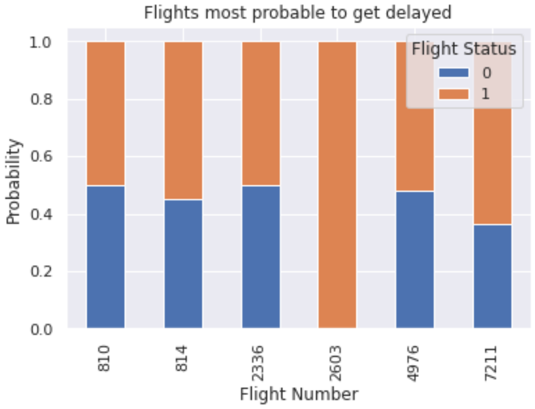


Figure 19: Probabilities vs Day of Month

As evident from first two graphs, Flights numbered **2166, 2170, 2174, 2182** are most reliable and preferred flights that never get late and frequently available too. Whereas Flight numbered 2603 is the worst flight which is very less frequent and always delayed.

2 Logistic Regression

2.1 Data Preprocessing

The target variable (Flight Status) is categorical type. The status labeled 'delayed' is assigned '1' and 'ontime' is assigned '0'. Then all the rows with null values were removed using df.dropna() command but interestingly data provided was good and no Null value was found.

2.1.1 Adding New Classes

New features were added to the model based on the given data. In accordance with Data Processing Inequality. no extra information was generated but variables closely related (Eg. Distance, Origin and Destination) were combined to study there effect later.

1. 'Delay in departure time' was added to the model, which is just the difference between the departure time and correct departure time in minutes.
2. After analysing the values it take, the delay obtained was further classified into categories of 20 minutes interval to capture it as a categorial class rather than continuous.
3. Scheduled Departure Time was allotted to Time Intervals of 1 Hour.
4. Since Origin, Destination and Distance are, intuitively, closely related as a representation of route taken (discussed in EDA too), these were combined together to obtain new class of ROUTE+DIST

2.1.2 Creating Dummy Variables

Since all the new features analysed and added were converted to categorical classes, these and all other seemingly relevant classes were One-Hot Encoded to create dummy variables to be used in Classification Model later.
Variables One-Hot Encoded were :

- Carrier
- Weather
- Day of Month
- Day of Week
- Intervals of Delay in Departure
- Intervals of Scheduled Departure Time
- Flight Number
- ROUTE(DIST) taken

Lastly Scheduled Departure Time, Departure Time, Tail-Number, Origin, Distance, Destination, Fl-Date, Delay in Departure were dropped from the features DataFrame. These were dropped because either they had already been accounted in some other variable, or they were creating hell lot of unnecessary dummy variables or they were converted into their categorical representation.

2.2 Training the Model

The data-set was applied to logistic regression model imported from Sklearn library 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. Using a threshold between 0 and 1, we can classify Flights with certain parameters as Delayed or Ontime depending on the output value of above function being above or below the threshold value respectively.
Train-test split ratio of 60:40 was used in the logistic regression model. I played with this ratio by taking it as 80:20 also. Much difference in accuracy wasn't observed. All the dummy variables listed above were used as features to train the model.

2.3 Results

1. Accuracy of the Model: 90.24% with 183 features.

2. I tried passing all the features to the Decision Tree Classifier. Though an accuracy of 88.54% was reported, the depth of the tree was 32 which suggested over-fitting probably due to one hot encoding of categorical features like flight number.

3 Interpreting the Model

One hot encoding of Flight Number have increased the dimensions by a great number leading to 188 features in the model.
Few of the coefficients have been listed below :

Feature	Coefficient
Carrier CO	0.69
Carrier DH	-0.10
Carrier DL	-0.33
Carrier MQ	0.39
Carrier OH	-0.41
Carrier RU	0.15
Carrier UA	-0.30
Carrier US	-0.08
Weather 0	-1.27
Weather 1	1.28
Day 1	0.14
Day 2	0.034
Day 3	0.150
Day 4	-0.133
Day 5	-0.0719
Day 6	-0.344
Day 7	0.215
FlNo 7211	1.21
FlNo 2174	-0.53

Feature	Coefficient
Route BWI-EWR(169)	0.24
Route BWI-JFK(184)	-0.41
Route DCA-EWR(199)	-0.19
Route DCA-JFK(213)	0.53
Route DCA-LGA(214)	-0.11
Route IAD-EWR(213)	0.18
Route IAD-JFK(228)	-0.46
Route IAD-LGA(229)	0.22
Delay Interval -1	-3.58
Delay Interval 0	-2.38
Delay Interval 1	0.38
Delay Interval 2	2.32
Delay Interval 3	1.78
Delay Interval 4	1.13
Delay Interval 5	0.13
Delay Interval 8	0.22
DEP Interval 6	-0.16
DEP Interval 7	0.04
DEP Interval 8	-0.24

Consider sigmoid function shown in Training the Model section. It suggests for same values of X, if magnitude of coefficient of one particular X is high, it would be more significant feature as its weightage in function is more. Using this fact and EDA done before, we can have following interpretations.

1. Major contribution comes from weather as evident from the coefficients listed above. Thus, Weather is a very crucial variable.

2. Delay in departure seems the most promising features after weather due to high magnitude of coefficients of dummy variable.

3. Dummy variables of Departure time hour block have low magnitude varying coefficients suggesting less dependence of output on them. Not only this but they also add extra dimensions in our model.

4. Flight number have increased the feature dimension to a great extent, making it difficult for us to analyze every coefficient individually. Each of its dummy variable show a large variation in coefficient. Though as predicted in EDA, few flight numbers have very high magnitude coefficients as they are most likely to get delayed.

5. Carrier is also not a promising feature as apart from few coefficients, rest are really low in magnitude.

6. It seems from this model that Day of Week is not a great parameter because we have Day of Month variable in our model which overlaps with this thus over fitting our model and reducing the coefficients of these. (since they are less in number)

4 Feature Selection

The less varying percentage of sub categories for a particular feature suggests about variance and we can safely remove those features as there is not much dependence on them.
Using the Interpretation of previous Model and the percentage graphs plotted (probability stacked bar graphs) while doing EDA helps us select the necessay features for our final optimised model.

1. Origin and destination together capture the information encoded in distance or vice versa. Combined ROUTE+DIST variable was created and used. This is important because it is a very important real life parameter to consider while travelling and EDA also supports this as there is quite a variation in percentage garph.

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. Once other irrelevant features were removed, model was made for both and accuracy was checked. It was found Day of Week is much suitable variable. Moreover it greatly reduced the no. of variables used.

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 and time difference between departure and scheduled departure time. Hence these features are important. Model was tested with and without considering Carrier, and it turned out to be weaker estimator and thus was eliminated.

5. Features selected are:

   (a) Day of Week

   (b) ROUTE+DIST

   (c) Weather

   (d) Delay in Departure time

5 Final Optimised Model

Weather-0	-1.39
Weather-1	1.39
Route–BWI-EWR(169)	0.18
Route–BWI-JFK(184)	-0.19
Route–DCA-EWR(199)	0.25
Route–DCA-JFK(213)	0.59
Route–DCA-LGA(214)	-0.30
Route–IAD-EWR(213)	0.33
Route–IAD-JFK(228)	-0.85
Route–IAD-LGA(229)	-0.01
Delay-Interval-no—1	-3.58
Delay-Interval-no–0	-2.33
Delay-Interval-no–1	0.359
Delay-Interval-no–2	2.195
Delay-Interval-no–3	1.82
Delay-Interval-no–4	0.96
Delay-Interval-no–5	0.19
Delay-Interval-no–8	0.37
Week-Day-1	0.23
Week-Day-2	-0.025
Week-Day-3	-0.15
Week-Day-4	-0.26
Week-Day-5	0.218
Week-Day-6	-0.369
Week-Day-7	0.361

Results

1. Accuracy for the new Linear Logistic model with selected features: 92.05%
2. Accuracy for the new Decision Tree model with selected features: 91.26%
3. Depth of the tree : 16

6 Predicting ideal weather conditions for a flight

- Weather : Good
- Day : Saturday
- Time : Between 11 to 12 in the morning
- Carrier : US
- Route : DCA - LGA

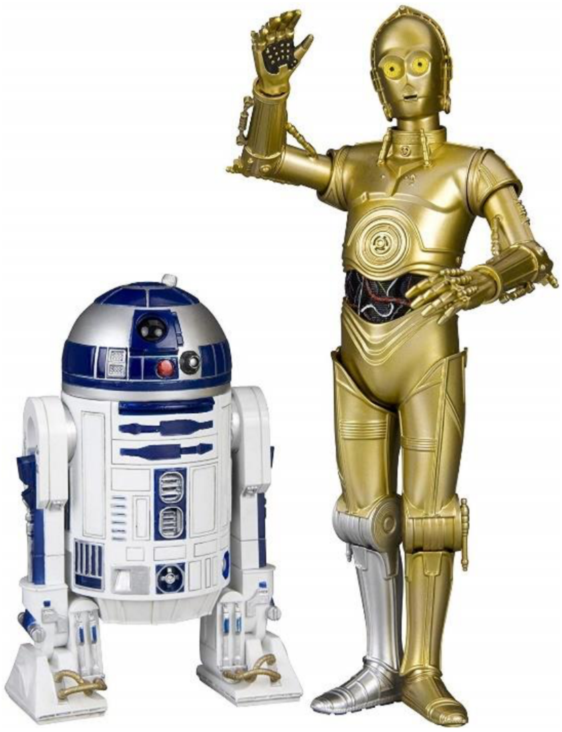
7 Extra Questions

Q – Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY and EDITH.
A – DUMMY, VERONICA, KAREN, TADASHI are some other examples.

Q – Data Processing Inequality
A – The content of the signal/data cannot be increased by a local physical operation (in some sense information can only be lost after processing the data.)

Q – 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.
A – X is Rule of Two (Darth Bane)

Q – In Star Wars Universe, name this robotic duo :-



A – Left one is R2-D2, Right one is C-3PO

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

8 Link to Colab Notebook