GNR638: Assignment 2

Hitesh Kandala, 180070023

April 7, 2020

Contents

1 Exploratory Data Analysis 2
1.1 Pie Charts 2
1.2 Categorical Plots 3
1.3 Box Plots 4

2 Logistic Model 5
2.1 With One-hot encoding 5
2.2 With Label encoding 5

3 Model Interpretation 6

4 Variable Selection 7

5 New Logistic Model 8
5.1 With One-hot encoding 8
5.1.1 Without feature ”DELAY” 8
5.1.2 With feature ”DELAY” 8
5.2 With Label encoding 8
5.2.1 Without feature ”DELAY” 8
5.2.2 With feature ”DELAY” 8

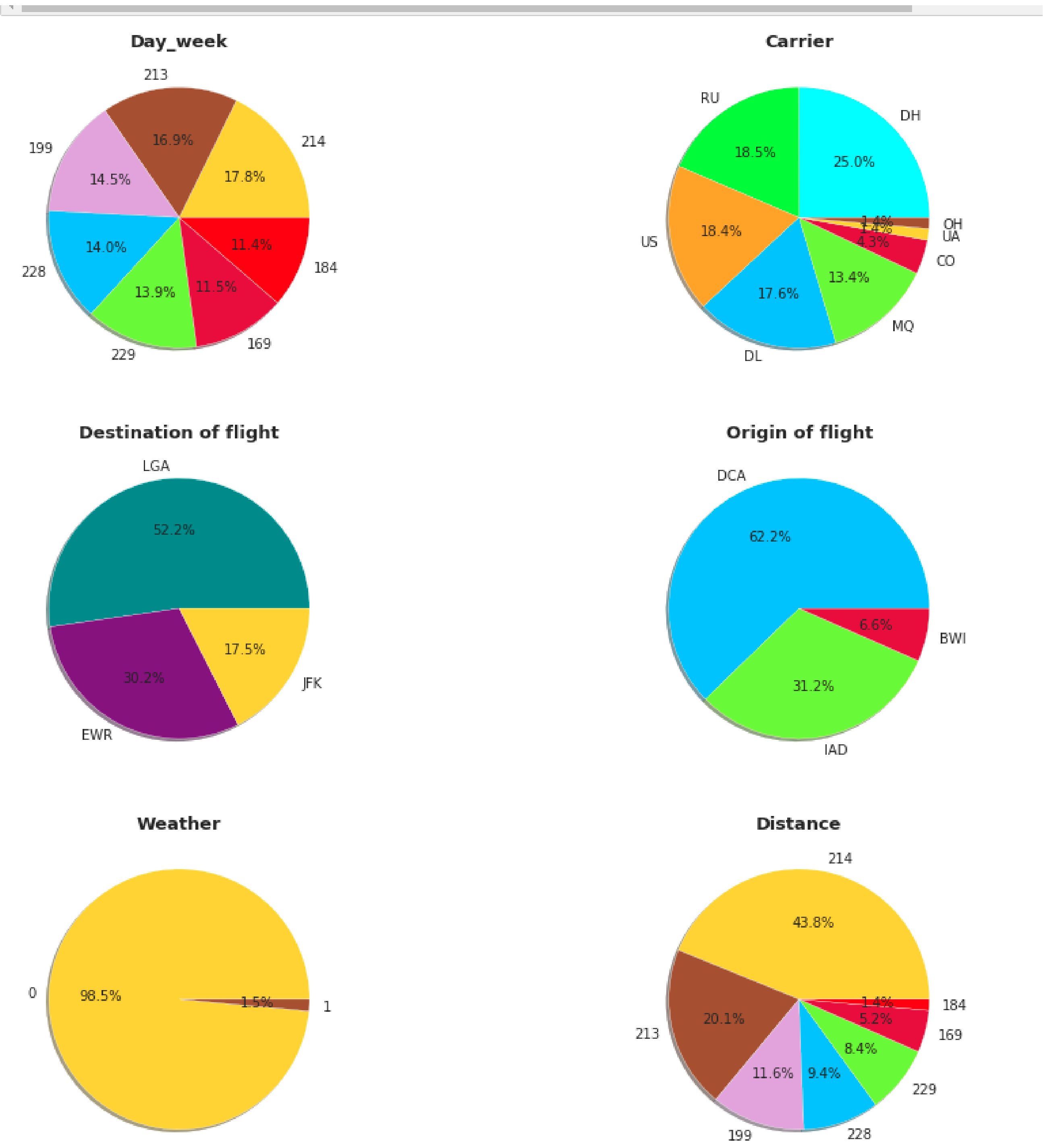
6 Ideal Weather Conditions 9

7 Bonus Questions 9

1 Exploratory Data Analysis

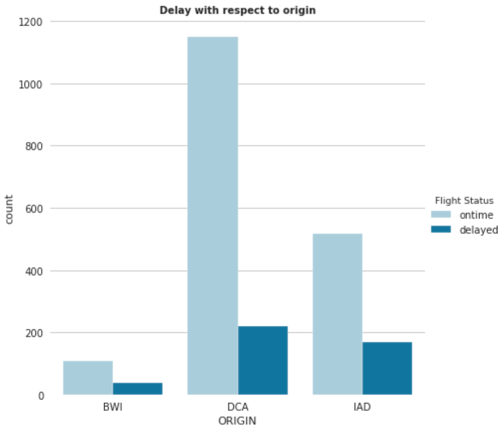
1.1 Pie Charts

Here we see some underlying distributions and their percentages in the data.

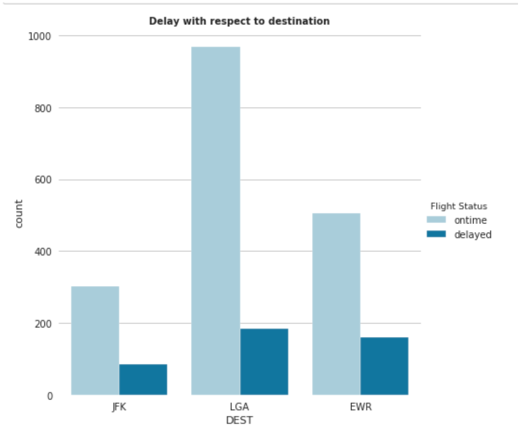


1.2 Categorical Plots

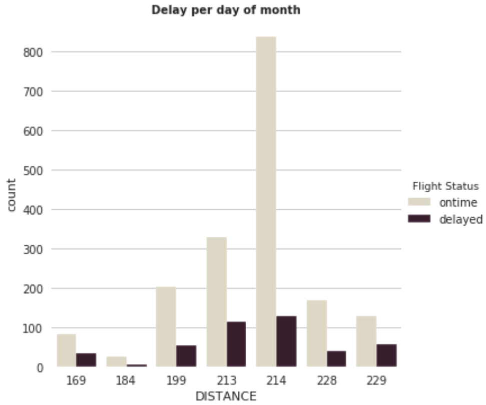
Now we see some categorical plots:



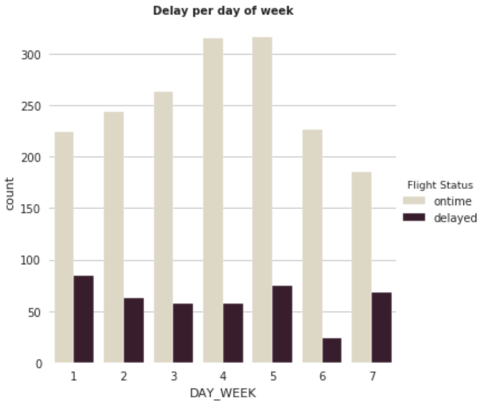
(a) Delay wrt origin



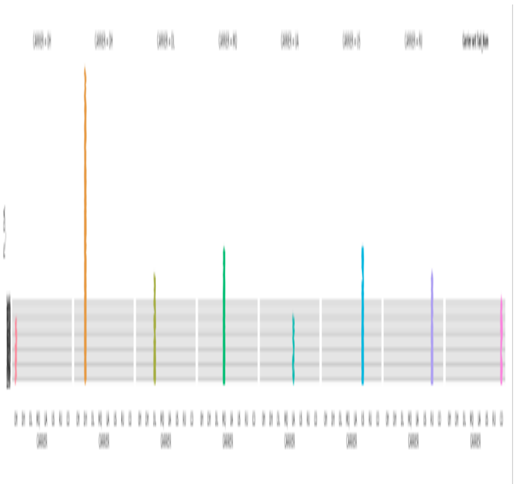
(b) Delay wrt destination



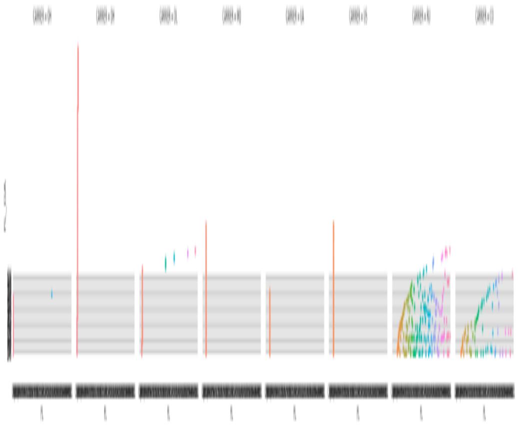
(a) Delay wrt day of month



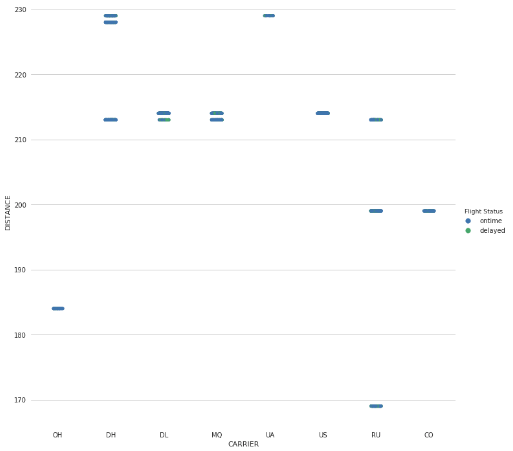
(b) Delay wrt day of week



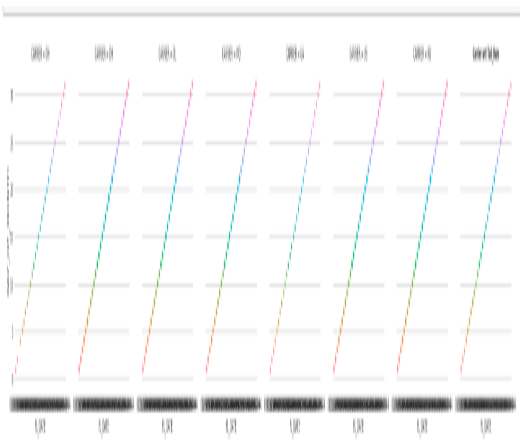
(a) Tail_Num with FL_Num



(b) Carrier, Tail_Num, FL_Num relation



(a) Distance wrt Carrier



(b) Day of month wrt FL_DATE for different carriers

1.3 Box Plots

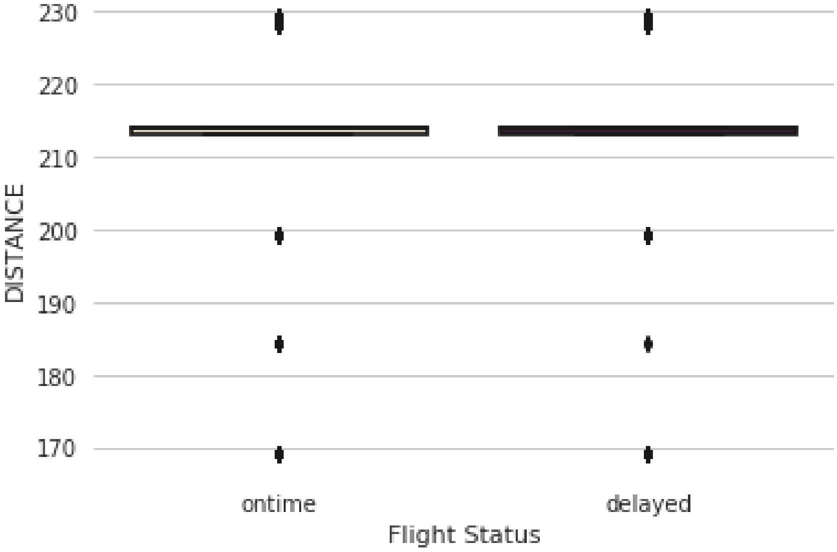


Figure 5: Boxplot of Distance

2 Logistic Model

We use the logistic regression to predict the flight status of 40% test dataset.
Below we see this model applied in two ways:

2.1 With One-hot encoding

Here we use all the features except "FL_DATE" as it is similar to "DAY_OF_MONTH" to perform regression by encoding them in one-hot format.
By this encoding, we have total 570 encoded features.

$$Accuracy = 0.887627695800227 \tag{1}$$

∗ There are a total of 570 encoded features. Therefore for weights, please refer this link.

2.2 With Label encoding

Here again we use all the features except "FL_DATE" but this time encoded with numeric labels, thus we have 11 features.

$$Accuracy = 0.8796821793416572 \tag{2}$$

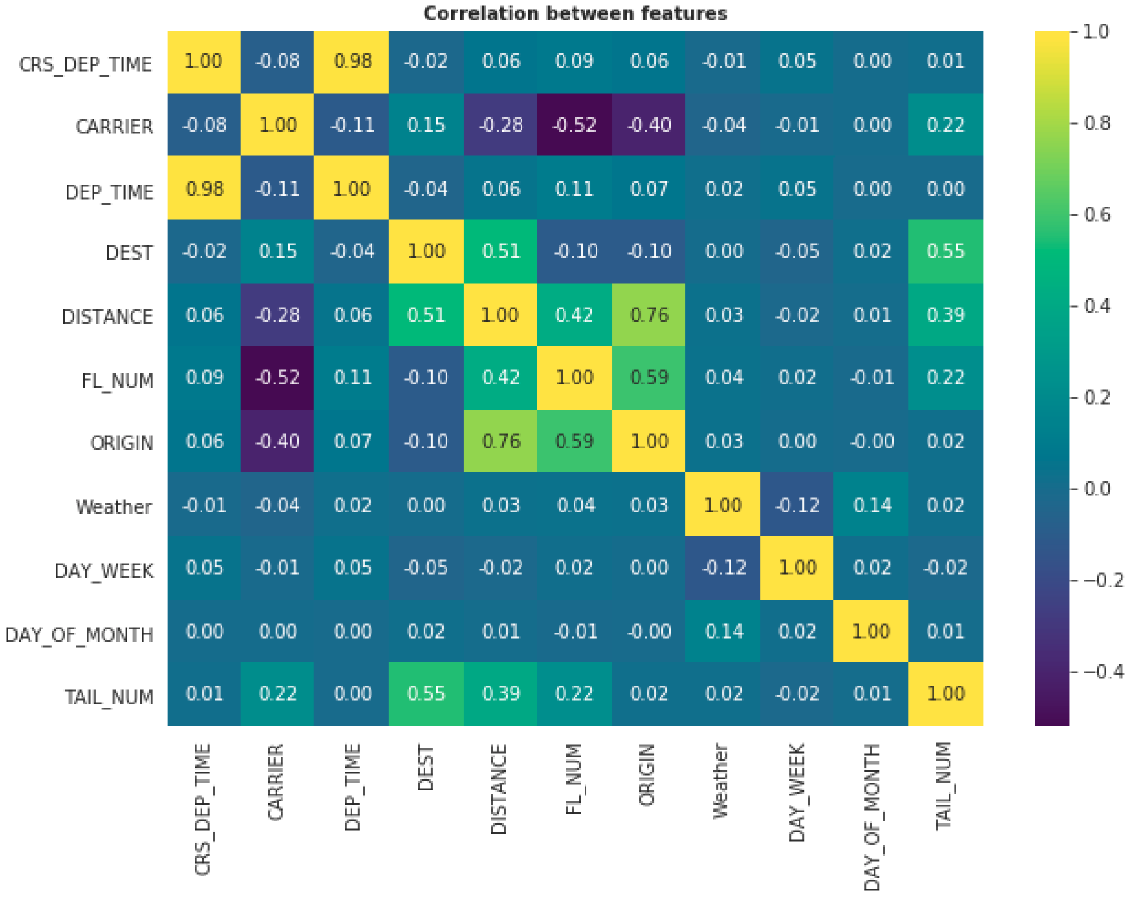
Weights obtained are listed below:

$$\begin{aligned}
\text{CRS Dep Time} &= -2.26297565e-02\\
\text{Carrier} &= -7.38888053e-02\\
\text{Dep Time} &= 2.31130722e-02\\
\text{Destination} &= -2.61639645e-01\\
\text{Distance} &= -1.28222471e-02\\
\text{Flight Number} &= -5.58873009e-05\\
\text{Origin} &= 2.67995890e-01\\
\text{Weather} &= 2.39055202e-01\\
\text{Week Day} &= -6.95762886e-02\\
\text{Day Of Month} &= 3.28264788e-02\\
\text{Tail Number} &= 1.91006038e-03
\end{aligned}$$

3 Model Interpretation

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model.

Some more insight into the correlation of features is shown below:



	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	TAIL_NUM
CRS_DEP_TIME	1.00	-0.08	0.98	-0.02	0.06	0.09	0.06	-0.01	0.05	0.00	0.01
CARRIER	-0.08	1.00	-0.11	0.15	-0.28	-0.52	-0.40	-0.04	-0.01	0.00	0.22
DEP_TIME	0.98	-0.11	1.00	-0.04	0.06	0.11	0.07	0.02	0.05	0.00	0.00
DEST	-0.02	0.15	-0.04	1.00	0.51	-0.10	-0.10	0.00	-0.05	0.02	0.55
DISTANCE	0.06	-0.28	0.06	0.51	1.00	0.42	0.76	0.03	-0.02	0.01	0.39
FL_NUM	0.09	-0.52	0.11	-0.10	0.42	1.00	0.59	0.04	0.02	-0.01	0.22
ORIGIN	0.06	-0.40	0.07	-0.10	0.76	0.59	1.00	0.03	0.00	-0.00	0.02
Weather	-0.01	-0.04	0.02	0.00	0.03	0.04	0.03	1.00	-0.12	0.14	0.02
DAY_WEEK	0.05	-0.01	0.05	-0.05	-0.02	0.02	0.00	-0.12	1.00	0.02	-0.02
DAY_OF_MONTH	0.00	0.00	0.00	0.02	0.01	-0.01	-0.00	0.14	0.02	1.00	0.01
TAIL_NUM	0.01	0.22	0.00	0.55	0.39	0.22	0.02	0.02	-0.02	0.01	1.00

4 Variable Selection

To perform variable selection, we find features which are more correlated to the output i.e. ”Flight Status” by using Pearson Correlation method.
This is listed below:

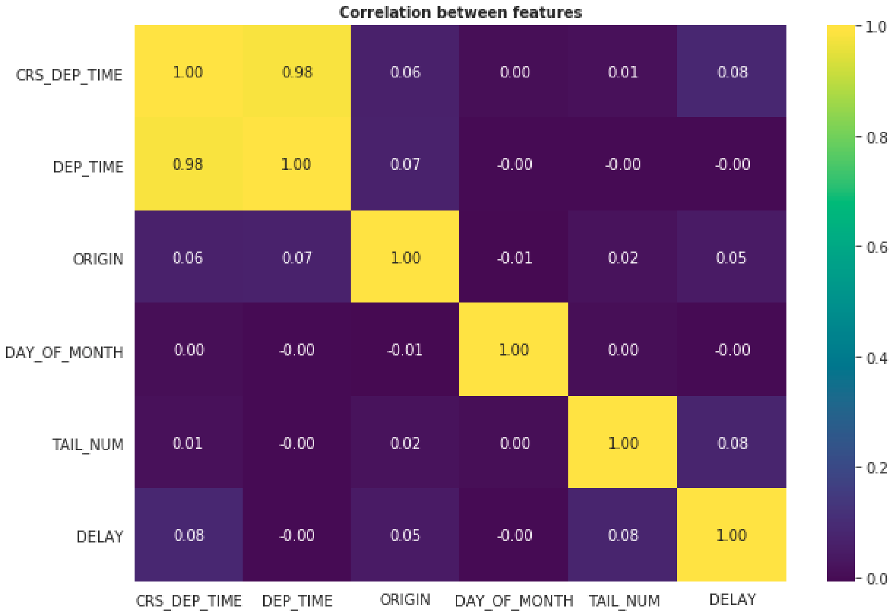
$$\begin{aligned}
\text{CRS Dep Time} &= 1.1377064772637745e - 01 \\
\text{Carrier} &= 2.1743435714333705e - 02 \\
\text{Dep Time} &= 1.661547742824352e - 01 \\
\text{Destination} &= 3.843964546936009e - 02 \\
\text{Distance} &= 1.8793865337225355e - 02 \\
\text{Flight Number} &= 2.279215208774348e - 02 \\
\text{Origin} &= 5.666613992603034e - 02 \\
\text{Weather} &= 2.472166382647505e - 01 \\
\text{Day Of Week} &= 4.0756223907865506e - 02 \\
\text{Day Of Month} &= 6.659849597821402e - 02 \\
\text{Tail Number} &= 8.644776214350594e - 02
\end{aligned}$$

We select variables by keeping a threshold value of 5e-02 on the correlation coefficient i.e. select variables whose correlation coefficient exceeds 5e-02.
Thus the variables we retain are

CRS_DEP_TIME’, ’DEP_TIME’, ’ORIGIN’, ’DAY_OF_MONTH’, ’TAIL_NUM’

Apart from these variables, we create an additional variable and see its effect on the accuracy of the new model. This new variable is:

$$DELAY = abs((CRS\_DEP\_TIME) - (DEP\_TIME)) \tag{3}$$



5 New Logistic Model

After selecting variables in the above section, we train our logistic model again on these variables.

5.1 With One-hot encoding

5.1.1 Without feature ”DELAY”

$$\boxed{Accuracy = 0.869815668202765} \tag{4}$$

∗ There are a total of 549 encoded features. Therefore for weights, please refer this link.

5.1.2 With feature ”DELAY”

$$\boxed{\textbf{Accuracy} = \mathbf{0.9089861751152074}} \tag{5}$$

∗ There are a total of 550 encoded features. Therefore for weights, please refer this link.

5.2 With Label encoding

5.2.1 Without feature ”DELAY”

$$\boxed{Accuracy = 0.8778801843317973} \tag{6}$$

Weights obtained are listed below:

$$\begin{aligned}
\text{CRS Dep Time} &= -0.0203133\\
\text{Dep Time} &= 0.02077522\\
\text{Origin} &= 0.11241854\\
\text{Day Of Month} &= 0.02219817\\
\text{Tail Number} &= 0.00060492
\end{aligned}$$

5.2.2 With feature ”DELAY”

$$\boxed{Accuracy = 0.9032258064516129} \tag{7}$$

Weights obtained are listed below:

$$\begin{aligned}
\text{CRS Dep Time} &= -0.03543157\\
\text{Dep Time} &= 0.03578089\\
\text{Origin} &= -0.0420204\\
\text{Day Of Month} &= 0.02950807\\
\text{Tail Number} &= -0.00063077\\
\text{Delay} &= 0.03825846
\end{aligned}$$

∗ **Therefore, the best accuracy comes with including ”Delay” feature and encoding them in one-hot format**

6 Ideal Weather Conditions

The ideal weather conditions (weather, time, day, carrier) for the highest chance of an ontime flight from DC to New York are:

◦ Weather = 0 (i.e. no flight delay due to bad weather)
◦ Time = 21:30 (CRS Departure Time)
◦ Day = 1st day of the month
◦ Carrier = Any airline (but specially ”USAirways” due to its low percentage of delayed flights)

7 Bonus Questions

1) Name any AIs made by Tony Stark in the Marvel Cinematic Universe besides JARVIS, FRIDAY and EDITH.
Ans. **Veronica**, **Ultron**

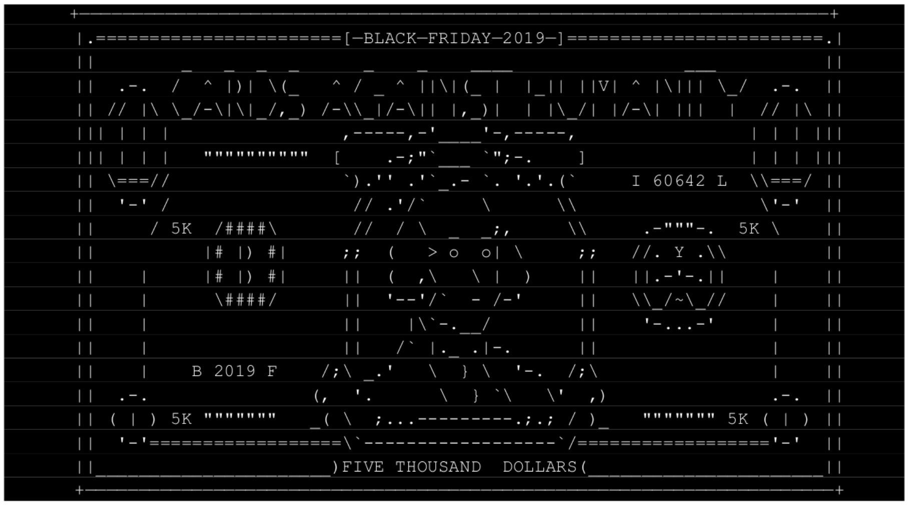
2) Explain the Data processing inequality.
Ans. The Data processing inequality states that the information content of a signal cannot be increased by any clever manipulation of the data. This can be expressed concisely as ’post-processing cannot increase information’.

3) 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. What is X?
Ans. **Rule of Two**

4) In Star Wars Universe, name the robotic duo.

Ans. **C-3PO** and **R2-D2**.

5) What is special about Cards against Humanity: Black Friday 2019?



```
+--------------------------------------------------------------------+
|.=======================[-BLACK-FRIDAY-2019-]=========================.|
||         _   _  _  _      _    _    ____                ___          ||
||  .-.  /  ^ |)| \(_   ^ / _ ^ ||\|(_ |  |_|| ||V| ^ |\||| \_/  .-.  ||
|| // |\ \_/-\|\|_/,_) /-\\_|/-\|| |,_)|  | |\_/| |/-\| |||  |  // |\ ||
||| | | |               ,-----,-'____'-,-----,              | | | |||
||| | | |   """"""""""  [    .-;"`___ `";-.    ]              | | | |||
|| \===//               `).'' .'`_.- `. '.'.(`      I 60642 L  \\===/ ||
||  '-' /                // .'/`      \      \\                 \'-'  ||
||     / 5K  /####\      //  / \  _  _;,     \\     .-""""-.  5K \    ||
||          |# |) #|    ;;  (   > o  o| \     ;;   //. Y .\\        ||
||     |    |# |) #|    ||  (  ,\   \ |  )    ||   ||.-'-.||    |   ||
||     |     \####/     ||  '--'/`  - /-'     ||   \\_/~\_//    |   ||
||     |                ||    |\`-.__/        ||    '-...-'     |   ||
||     |                ||   /` |._ .|-.      ||                |   ||
||     |   B 2019 F   /;\ _.'   \  } \  '-.  /;\                |   ||
||  .-.              (,  '.      \  } `\   \'  ,)               .-.  ||
|| ( | ) 5K """""""  _( \  ;...---------.;.; / )_   """"""" 5K ( | ) ||
||  '-'==================\`------------------`/==================='-'  ||
||______________________)FIVE THOUSAND  DOLLARS(_____________________||
+--------------------------------------------------------------------+
```

Ans. For Black Friday, a computer was taught on how to write Cards Against Humanity cards. Over the next 16 hours, the writers battled this powerful cardwriting algorithm to see who can write the most popular new pack of cards.