GNR 652: Assignment 2
Predicting Flights Delay Using Logistic Regression

Jay Tukaram Sawant, Roll number: 18D070050

April 6, 2020

1 Overview

In this Assignment, we are constructing a logistic Regression model to predict the Flight delays for the Airport authorities. We are given a dataset ”Flight-Delays.csv” which consists of 2201 examples and 12 variables. The classification here is a binary classification. Each example has a output/Flightstatus as ’ontime=1’ or ’delayed=0’.
The flow of the Assignment is as follows:

1. Performing Exploratory Data Analysis
2. Pre-processing the dataset
3. Building a Logistic Regression Model based on all the features given and interpreting the results and accuracy of the Model
4. Performing variable selection to reduce the complexity of model
5. Fitting a new model on these selected variables and interpreting the results

2 Question 1 : Exploratory Data Analysis

The variables available in the dataset are as follows :

1. CRS_DEP_TIME (Scheduled Departure time)
2. CARRIER
3. DEP_TIME (Departure Time)
4. DEST (Destination)
5. DISTANCE (in kilometers)
6. FL_DATE (Flight Date)
7. FL_NUM (Flight Number)
8. ORIGIN
9. Weather
10. DAY_WEEK (Day of week)
11. DAY_OF_MONTH
12. TAIL_NUM (Tail number)

The **FL_DATE** column in dataset consists of all the dates of the month January of year 2001.
Hence, the column **DAY_OF_MONTH** and the column **FL_DATE** are equivalent.
Hence, we discard the **FL_DATE** column in the dataset.
In the further Bar Plots,
Brown colour = percent of delayed flights
Green colour = percent of ontime flights

2.1 Feature : Carrier

The following plots show the Distribution of the feature Carrier in the dataset:

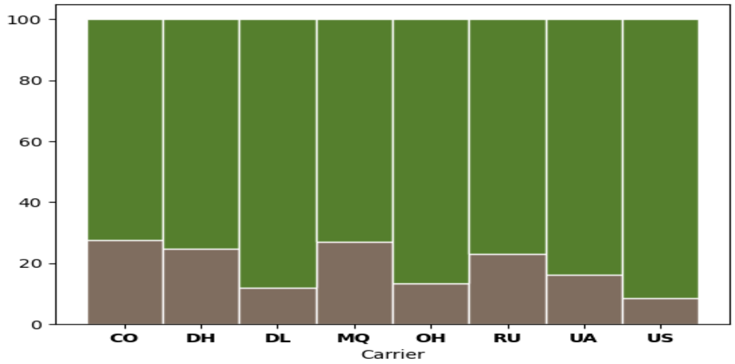


Figure 1: Percent of Flight delayed/ontime by feature Carrier

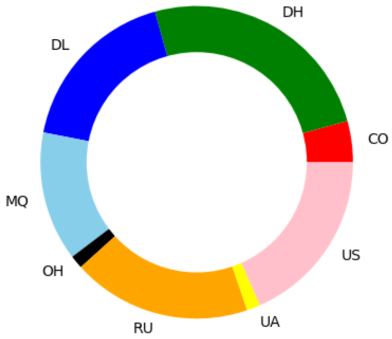


Figure 2: Percentage of count of each Carrier in the whole dataset

The number of examples which are delayed in the dataset are more or less equally distributed among all the carriers. While the count of 'OH', 'UA' and 'CO' carriers in the dataset is very less.

2.2 Feature : Destination

The following plots show the Distribution of the feature Destination in the dataset:

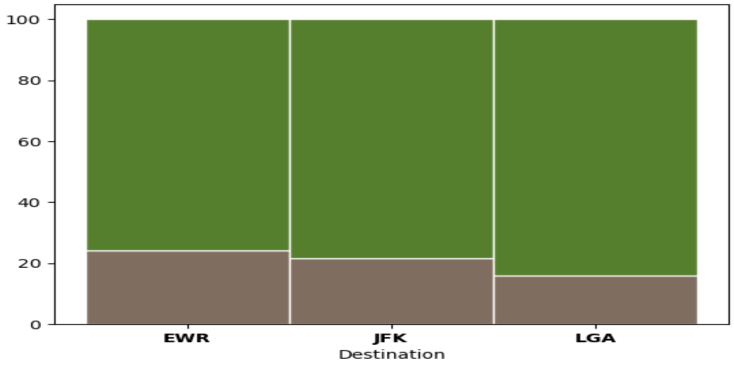


Figure 3: Percent of Flight delayed/ontime by feature Destination

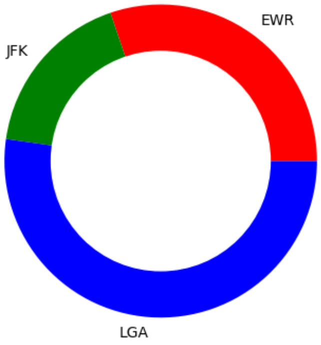


Figure 4: Percentage of count of each Destination in the whole dataset

The number of examples which are delayed in the dataset are more or less equally distributed among all the Destinations. While the count of Destination 'LGA' in the dataset is very large.

2.3 Feature : ORIGIN

The following plots show the Distribution of the feature ORIGIN in the dataset:

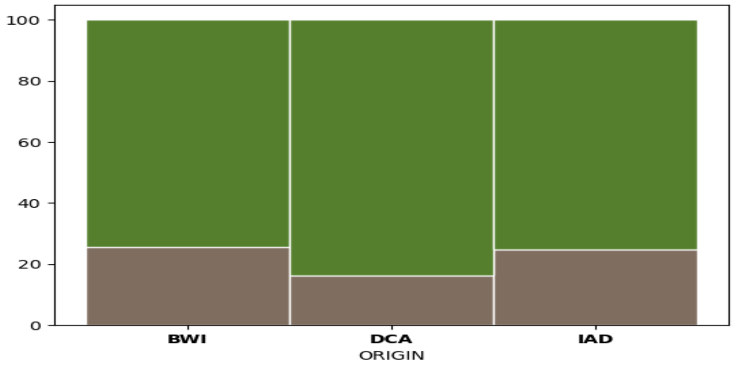


Figure 5: Percent of Flight delayed/ontime by feature Origin

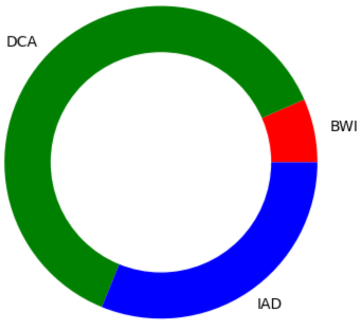


Figure 6: Percentage of count of each Origin in the whole dataset

The number of examples which are delayed in the dataset are more or less equally distributed among all the Origin. While the count of Origin 'DCA' in the dataset is very large and the count of Origin 'BWI' is very small.

2.4 Feature : Days of Week

The following plots show the Distribution of the feature DAY_WEEK in the dataset:

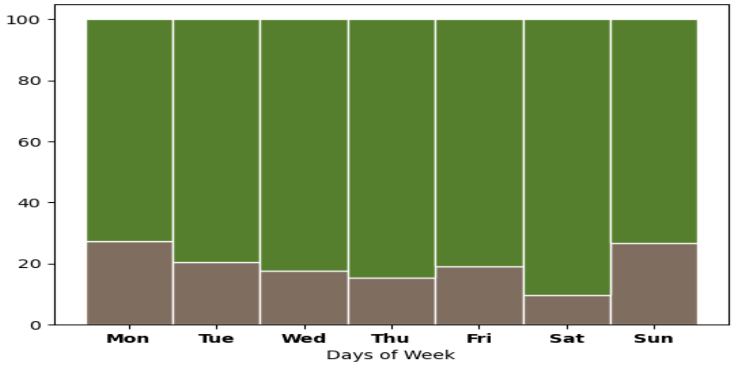


Figure 7: Percent of Flight delayed/ontime by feature DAY_WEEK

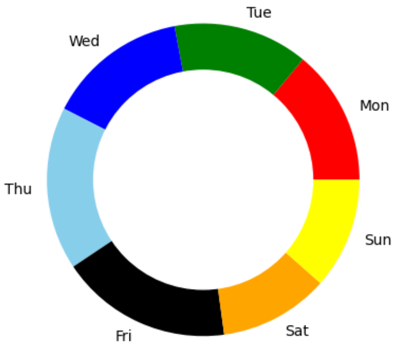


Figure 8: Percentage of count of each day in the whole dataset

The number of examples which are delayed in the dataset are more or less equally distributed among all the days of the week. While the count of all the days in the dataset is more or less similar.

2.5 Feature : Weather

The following plots show the Distribution of the feature Weather in the dataset:

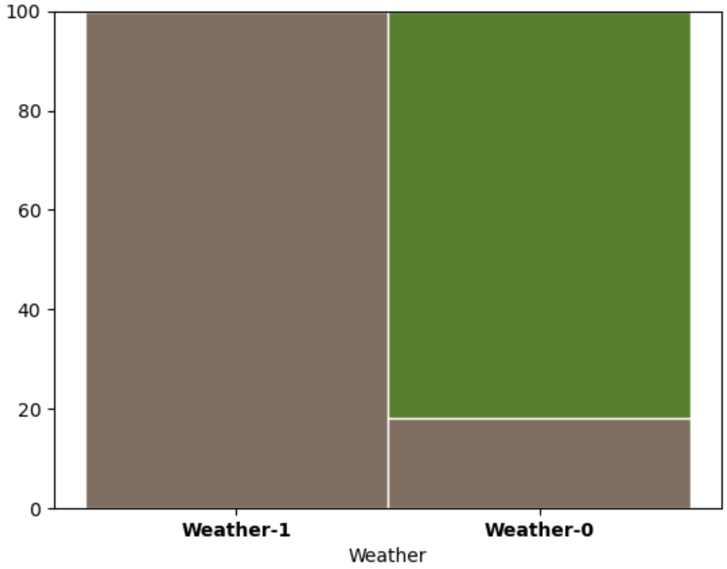


Figure 9: Percent of Flight delayed/ontime by feature Weather

In the Weather distribution, all the flights are delayed when the feature 'Weather' takes up the value 1. When the feature 'Weather' has value 0, a small portion of the total flights is delayed.

3 Question 2

1. We label encode the column **Flight Status**, in which 'ontime'= 1 and 'delayed' = 0 using the LabelEncoder() function from scikit library

2. We perform One Hot Encoding using the function get_dummies() on the features Carrier, Destination, Origin, Days of Week, Flight Number, Days of Month, Tail number

3. The features **CRS_DEP_TIME** and **DEP_TIME** are first converted from 24 hour format to minute format and then these both the features are normalized before passing as input.

4. The feature **DISTANCE** is also normalized before passing it as input.

5. The Data is randomly splitted as 60% training set and 40% test set using the train_test_split() function from scikit library

6. Te cost function used is as folows :

$$L(\mathbf{b}) = -\frac{1}{m}\sum_{i=1}^{m} y_i \log \sigma(\mathbf{b}^T\mathbf{x}_i) + (1-y_i)\log(1-\sigma(\mathbf{b}^T\mathbf{x}_i))$$

$$\frac{\delta L}{\delta b_j} = \frac{1}{m}\sum_{i=1}^{m}(\sigma(\mathbf{b}^T\mathbf{x}^{(i)}) - y^{(i)})x_j^{(i)}$$

7. Parameter $\mathbf{b}$ is initialized to $\mathbf{0}$ and then is updated using gradient descent as follows :

$$b_j := b_j - \alpha * \frac{\delta L}{\delta b}$$

8. Here, α is the learning rate and $\sigma()$ is the sigmoid function

9. We train the above model using the training set from dataset and obtain the coefficient vector $\mathbf{b}_{new}$

10. We calcuate the value of $\sigma(\mathbf{X}_{test}\mathbf{b}_{new})$. If the values in this vector are greater than 0.5, the prediction is labelled as 1 or else it is labelled as 0.

11. You can find the code for this question in all_features.py

4 Question 3

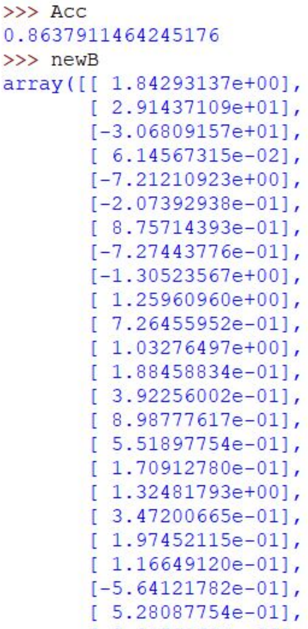
1. As the data is splitted into train and test set randomly, the accuracy obtain at each run is different. Hence, the code has been ran for 10 times and average of accuracies is taken.

2. Average Accuracy obtained : 80.85 %

3. Highest Accuracy obtained in the 10 runs : 86.37 %

4. As we used one hot encoding, the number of features has been blown up to 708 features. Blown up of features is mainly caused due to the variable 'TAIL_NUM' in the dataset

5. Some of the coefficients (considering the best case out of 10 runs) of initial features are as follows :



```
>>> Acc
0.8637911464245176
>>> newB
array([[ 1.84293137e+00],
       [ 2.91437109e+01],
       [-3.06809157e+01],
       [ 6.14567315e-02],
       [-7.21210923e+00],
       [-2.07392938e-01],
       [ 8.75714393e-01],
       [-7.27443776e-01],
       [-1.30523567e+00],
       [ 1.25960960e+00],
       [ 7.26455952e-01],
       [ 1.03276497e+00],
       [ 1.88458834e-01],
       [ 3.92256002e-01],
       [ 8.98777617e-01],
       [ 5.51897754e-01],
       [ 1.70912780e-01],
       [ 1.32481793e+00],
       [ 3.47200665e-01],
       [ 1.97452115e-01],
       [ 1.16649120e-01],
       [-5.64121782e-01],
       [ 5.28087754e-01],
```

Figure 10: Coefficients

6. From the above image, we can see that the weather coefficient is highly negative (-7.21). Hence, when weather takes value 1, the model is likely to predict a delay.

7. We can see the distance coefficient here is 0.0614. Hence the output will not be affected by the variations in Distance.

8. The coefficient of CRS_DEP_TIME is highly positive (29.14) and the coefficient of DEP_TIME is highly negative (-30.68)

9. The rest of the dummy coefficients of TAIL_NUM, FL_NUM,etc have small values and thus have a small impact on the output individually.

5 Question 4

1. We saw that creating the dummy variables of the feature **TAIL_NUM**, the total number of features was blown up. Hence, we discard the **TAIL_NUM** feature.

2. We perform One Hot Encoding on the features Carrier, Destination, Origin, Days of Week, Flight Number, Days of Month

3. The features **CRS_DEP_TIME** and **DEP_TIME** are also converted from 24 hour format to minute format and then these both the features are normalized before passing as input.

4. The feature **DISTANCE** is also normalized before passing it as input.

5. We use the statistic of Pearson Correlation to correlate between the data of each feature vector and the Flight Status

6. We select only those features whose correlation with the output Flight Status is greater than 0.1.

7. Hence, from a total of 159 features, we select the 9 best features which are as follows :

   (a) CRS_DEP_TIME

   (b) DEP_TIME

   (c) Weather

   (d) CARRIER_US

   (e) ORIGIN_DCA

   (f) FL_NUM_7211

   (g) DAY_OF_MONTH_18

   (h) DAY_OF_MONT_26

   (i) DAY_OF_MONTH_27

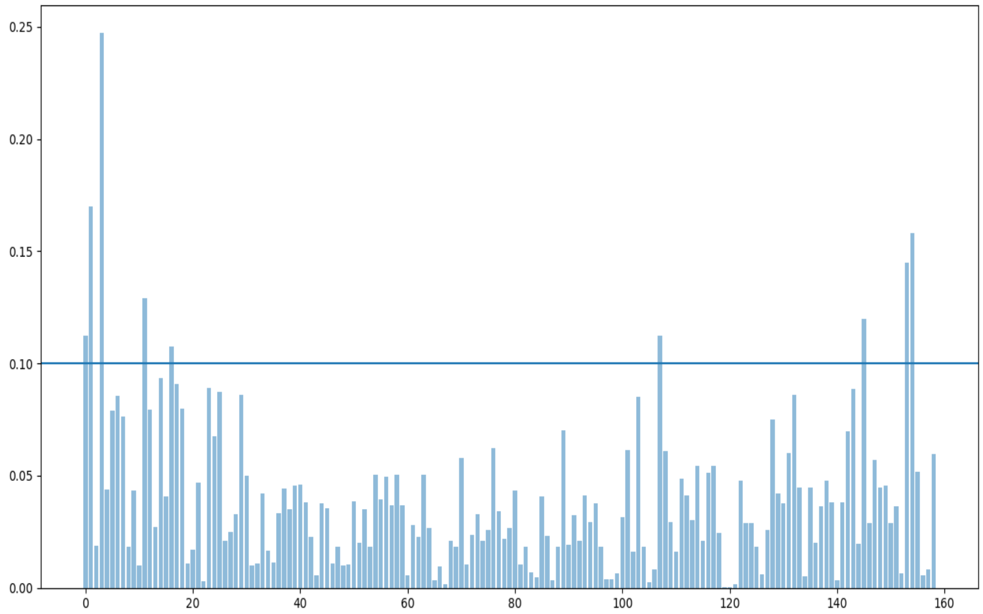


Figure 11: Correlation Bar Graph

The above plot shows the Correlation of data of each feature with the output column of Flight Status. We selected the only features whose correlation value lies above the horizontal line.

6 Question 5

1. As the data is splitted into train and test set randomly, the accuracy obtain at each run is different. Hence, the code has been ran for 10 times and average of accuracies is taken.

2. Average Accuracy obtained : 90.58 %

3. Highest Accuracy obtained in the 10 runs : 91.71 %

Feature	**Coefficient**
CRS_DEP_TIME	31.625
DEP_TIME	-32.34
Weather	-5.768
CARRIER_US	0.224
FL_NUM_7211	-0.364
ORIGIN_DCA	-0.315
DAY_OF_MONTH_18	-1.343
DAY_OF_MONTH_26	-0.350
DAY_OF_MONTH_27	-1.488

The above coefficients are taken from the run which had highest accuracy out of 10 runs

4. Hence, when the number of features are reduced, we reduce the complexity of the model and prevent overfitting.

5. Thus we can see a huge jump in the average accuracy of our new model which have selected features from the model which used all the features.

6. You can find the code for this question in selected_features_model.py

7 Question 6

1. We are interested to find the best conditions such that the flight from DC to New York will be ontime

2. For the ideal conditions, we need $\mathbf{b}^T\mathbf{x}$ must have a highly positive value so that its sigmoid will have high chances of being greater than 0.5 (ontime)

3. Hence, the ideal conditions are :

   (a) Weather = 0

   (b) Carrier = US

   (c) DAY_OF_MONTH $\neq$ 18,26,27

   (d) FL_NUM $\neq$ 7211

   (e) CRS_DEP_TIME = 20:30

   (f) DAY_WEEK = 4 (Thursday)

8 Bonus Question

Q1 : Ans : Veronica, Ultron
Q3 : Ans : The rule of Two
Q4 : Ans : R2D2 and C3PO