GNR 652 Machine Learning

Real World Assignment

"Flight Delays"

Submitted by

Shubham Bhise   193230015

April 6, 2020

Contents

List of Figures

Chapter 1

Exploratory Data Analysis

The main problem of this assignment is to correctly predict that the flight is going to be delayed or ontime. Its a binary classifier. Therefore the proportion of the delayed (1) and ontime (0) flight is checked from the given dataset.

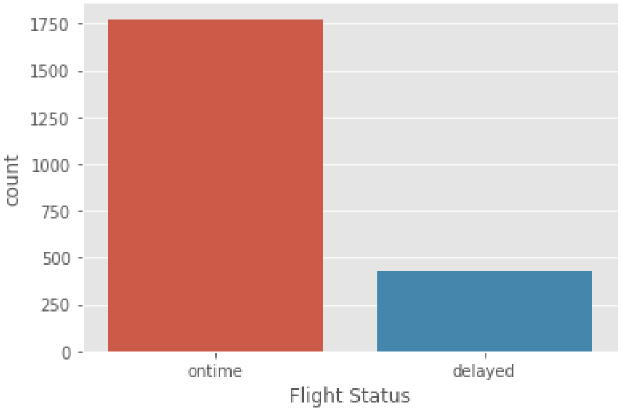


Figure 1.1: Flight Status

It shows that from the total dataset 19% flights are delayed.
Now for checking any null value in dataset a heatmap is used. It showed that there is no vacant or null space in the data.

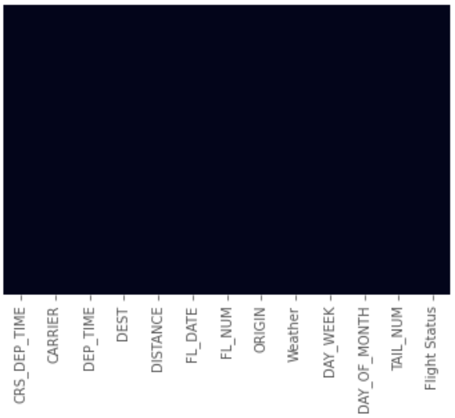


Figure 1.2: Null space checking

Also by plotting the graphs of features vs flight status, it can be seen that which features have more probability for causing the flight delay.

1. Weather

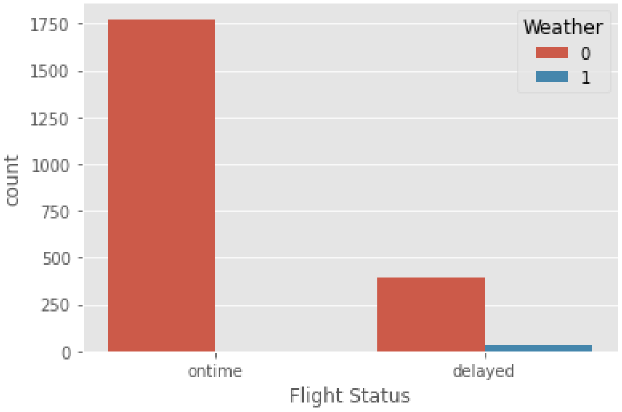


Figure 1.3: Feature - weather

It shows that if weather is bad (1) then there is definitely a delay in flight. So this feature is important in classification.

2. Day of week

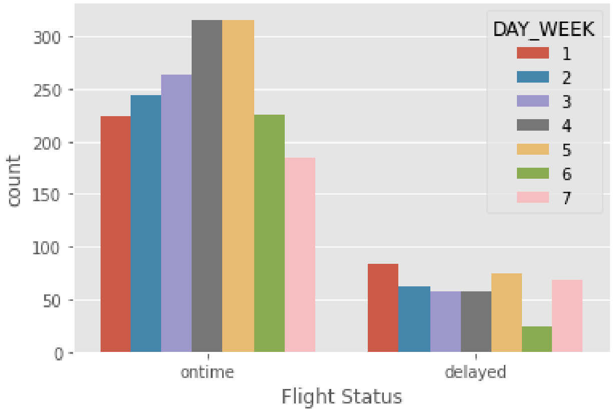


Figure 1.4: Feature - day of week

As compared to other days in week, there is more probability of flight is getting delayed on day 1. Also the trend of graph in delayed part is different than ontime.

3. Carrier company

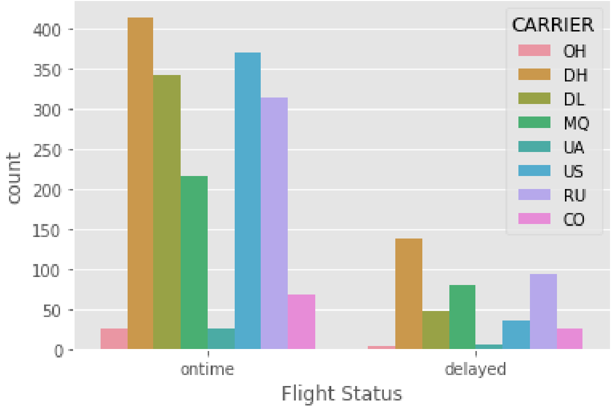


Figure 1.5: Feature - carrier company

Some carrier companies or services are showing different trends in ontime and delayed or it can be said that there is high variance.

4. Distance between origin and destination

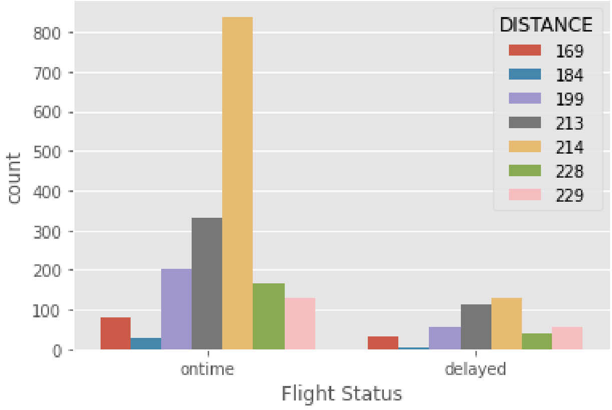


Figure 1.6: Feature - Distance

If distance is 213 and 229 then compared to others there are more chances of flight is getting delayed.

5. Origin airport

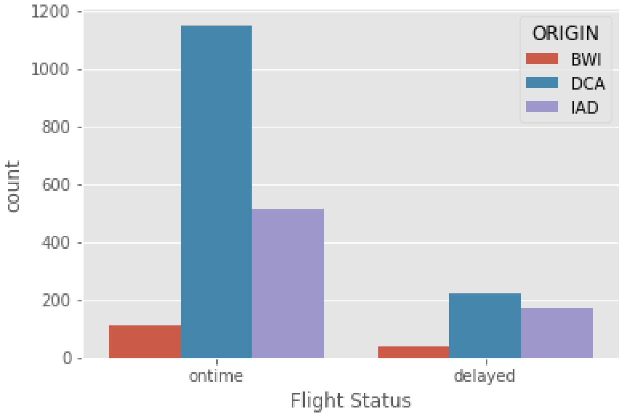


Figure 1.7: Feature - Origin airport

6. Destination airport

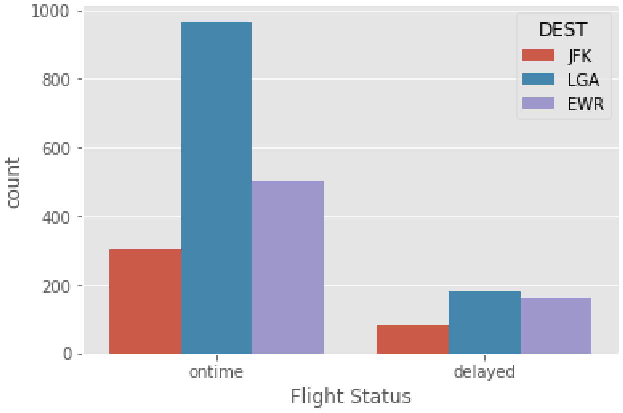


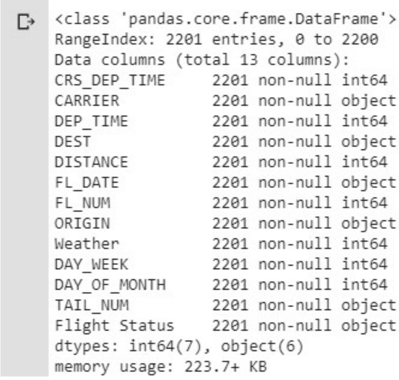
Figure 1.8: Feature - Destination airport

From the graphs it can be seen that there not much variance in feature 6 and also same for feature 5. So these features are of less important for the classification problem.

Chapter 2

Preprocessing of dataset and implementation of logistic model

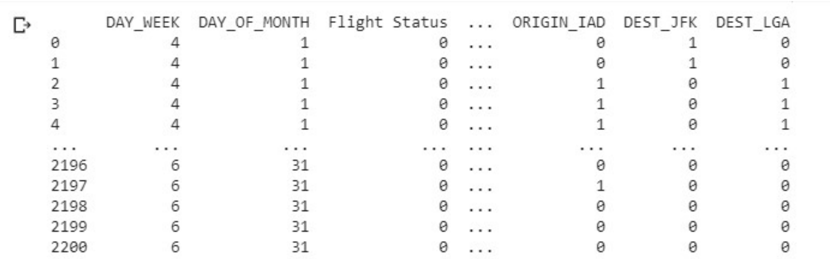
1. Null values: The fig.1.2: Null space checking in data analysis showed that there is no null value in the dataset. Therefore no need to perform the removal of null values. Also following result shows the availability of null values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2201 entries, 0 to 2200
Data columns (total 13 columns):
CRS_DEP_TIME     2201 non-null int64
CARRIER          2201 non-null object
DEP_TIME         2201 non-null int64
DEST             2201 non-null object
DISTANCE         2201 non-null int64
FL_DATE          2201 non-null object
FL_NUM           2201 non-null int64
ORIGIN           2201 non-null object
Weather          2201 non-null int64
DAY_WEEK         2201 non-null int64
DAY_OF_MONTH     2201 non-null int64
TAIL_NUM         2201 non-null object
Flight Status    2201 non-null object
dtypes: int64(7), object(6)
memory usage: 223.7+ KB
```

Figure 2.1: Checking null values

2. Dummy variables: These are generated for the all categorical variables like ORIGIN, DEST. It gives values 0 or 1 and coverts them into integers.



```
      DAY_WEEK  DAY_OF_MONTH  Flight Status  ...  ORIGIN_IAD  DEST_JFK  DEST_LGA
0            4             1              0  ...           0         1         0
1            4             1              0  ...           0         1         0
2            4             1              0  ...           1         0         1
3            4             1              0  ...           1         0         1
4            4             1              0  ...           1         0         1
...        ...           ...            ...  ...         ...       ...       ...
2196         6            31              0  ...           0         0         0
2197         6            31              0  ...           1         0         0
2198         6            31              0  ...           0         0         0
2199         6            31              0  ...           0         0         0
2200         6            31              0  ...           0         0         0
```

Figure 2.2: Dummy variables

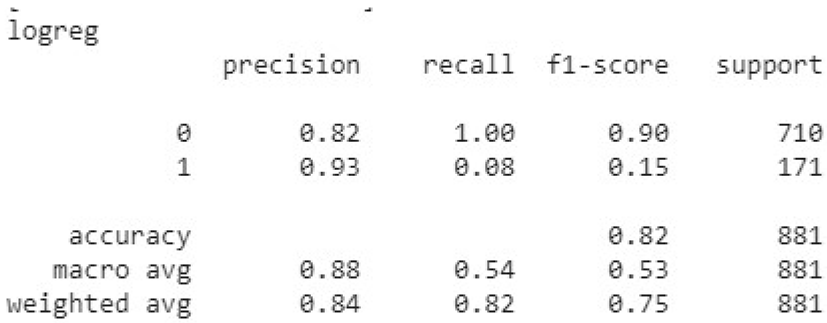
3. Train-Test split: Divided the data into 60-40 using sklearn library.

4. Logistic Regression: Implemented this model with the help of sklearn library. In this features which are redundant or not affecting the classifications are neglected such as DEP-TIME, TAIL-NUM, FL-DATE, FL-NUM. And for all other features, one hot encoding is applied for converting it from categorical to integers.

Chapter 3

Results

Overall Accuracy for flight ontime and delayed: 0.82.
But the recall and f1 score for delayed (1) is very low.



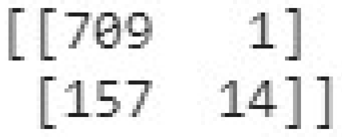
```
logreg
              precision    recall  f1-score   support

           0       0.82      1.00      0.90       710
           1       0.93      0.08      0.15       171

    accuracy                           0.82       881
   macro avg       0.88      0.54      0.53       881
weighted avg       0.84      0.82      0.75       881
```

Figure 3.1: Accuracy

It can be seen from the confusion matrix that true negatives are very less.



```
[[709   1]
 [157  14]]
```

Figure 3.2: Confusion Matrix

Chapter 4

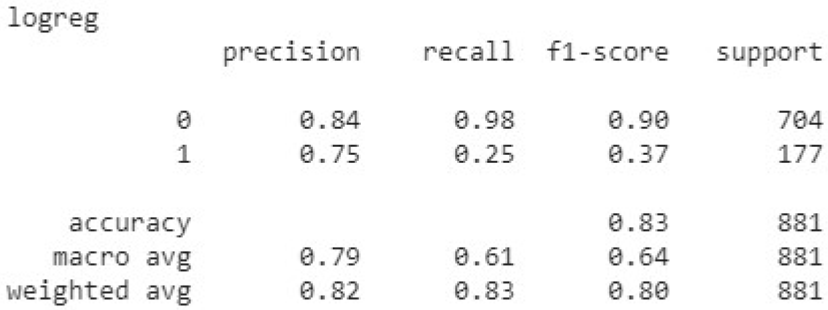
Feature selection and New model

It is performed by checking variance based on graphs in chapter 1.
Therefore following features are selected for increasing the true negatives or improving the recall and f1-score.
Significant features: Weather, CRS-DEP-TIME, DAY-OF-MONTH, CARRIER, DAY-WEEK, DISTANCE
Non significant features: ORIGIN, DEST (Because a feature DISTANCE can represent these two)

Chapter 5

New results

By using selected features, results were improved little bit.

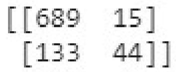


```
logreg
              precision    recall  f1-score   support

           0       0.84      0.98      0.90       704
           1       0.75      0.25      0.37       177

    accuracy                           0.83       881
   macro avg       0.79      0.61      0.64       881
weighted avg       0.82      0.83      0.80       881
```

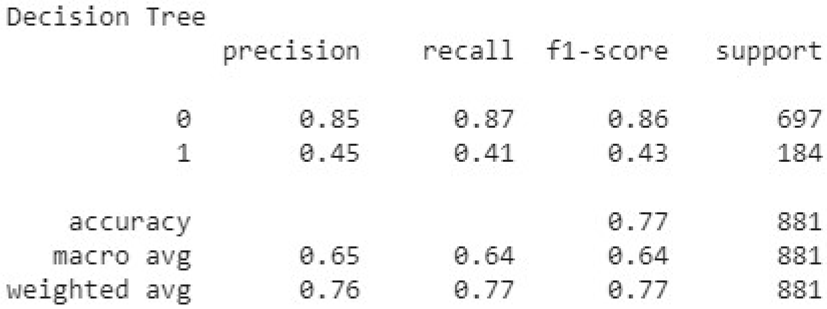
Figure 5.1: Accuracy of new model



```
[[689  15]
 [133  44]]
```

Figure 5.2: Confusion Matrix of new model

But true negatives increased after implementing decision tree model.

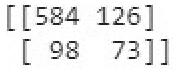


```
Decision Tree
              precision    recall  f1-score   support

           0       0.85      0.87      0.86       697
           1       0.45      0.41      0.43       184

    accuracy                           0.77       881
   macro avg       0.65      0.64      0.64       881
weighted avg       0.76      0.77      0.77       881
```

Figure 5.3: Accuracy of decision tree model



```
[[584 126]
 [ 98  73]]
```

Figure 5.4: Confusion Matrix of decision tree model

True negatives are not increasing significantly because of imbalanced data (In flight delays only 19% flights are delayed). Therefore it is difficult for model to get more accurate results for delayed flight.

This can be improved by further analysis such as Randomized-Synthetic Minority Over-sampling Technique (R-SMOTE). In this technique, the minority label i.e., label with lesser number of instances (here label 1) in the dataset is over-sampled. In other words, new artificial examples of the minority label are created to reduce the imbalance and match up with the majority label.

Chapter 6

Ideal conditions

From the variance we can understand that
Weather- 1, day of week- 4, carrier- US, origin- DCA, destination- LGA, time interval between 8 to 9

Chapter 7

Bonus

1. Ultron
2. For data analysis or machine learning purpose we process or gain some information from raw data. And information can only be lost and never increases as we process it. For example in the traditional method, we extract feature Y from an image X with a deterministic function. Given the feature, we estimate the outcome Z. So, if we lost some information from the first feature extraction stage, we cannot regain the lost information from the second stage. According to the data processing inequality, the mutual information between X and Z, I(X;Z) cannot be greater than that between X and Y, I(X;Y).
3. The rule of two
4. C-3PO and R2-D2
5. For Black Friday, Cards Against Humanity's writers are battling an A.I. to see who can write a better pack of cards. For A.I. they trained a powerful neural network to write the cards. And the interesting thing is that writers sold only 2 percent more than A.I.