PREDICTING CANCELLATIONS AND DELAYS OF UNITED STATES DOMESTIC FLIGHTS



DS-SEA-06

TAYLOR RANDS

MAY 24, 2017

**Problem Statement**

Can we predict the chances of a flight cancellation or delay using Logistic Regression based on the airline company as well as the day and time of the flight?

**Hypothesis**

The chances of a flight cancellation or delay will be higher as the performance of the airline company falls. Also, a cancellation or delay will occur more often during the periods of time where the most number of flights occur or when the factors of a flight cancellation or delay are most prevalent. For instance, the biggest factor of a flight cancellation or delay is most likely the current weather in which the flight departing or arriving must deal with. In this instance, flight conditions are at their worst in Winter months, so more flight cancellations or delays will occur during these months.

**Data Set**

The dataset that I am using is from the United States Department of Transportation’s (DOT) Bureau of Transportation. The statistics that this department provides tracks the on-time performance of domestic flights operated by large air carriers. The department summarizes information on the number of on-time, delayed, cancelled, and diverted flights and publishes it in DOT’s monthly Air Travel Consumer Report.

**Here are highlights of the 2015 dataset that I am using:**

* The data was separated into 3 different .csv files. I was able to combine these files into one .csv file that contained all the data containing the airlines, airports, and corresponding flights.
* The finished dataset contains over 5.8 million flights, as well as 41 columns describing certain elements of each flight.
* The data size is a bit less than 200 MB.

**Identifying Relevant Columns**

Due to the scope of the data, as well as the limited time for research, I was unable to import any weather data into my DataFrame. Instead, I created dummy variables for each month in 2015. I hope that using months as a feature will give me somewhat of an idea of cancellations and delays due to weather across the months. Another feature I used was the time of the scheduled departure of the flight. I used this as one of my features because I believe that more delays will occur during busier hours of the day due to higher traffic. The last feature I used was the airline company themselves. Each company has a different volume of flights as well as a corresponding performance and I believe that using an airline as a feature will help me predict whether a flight is cancelled or delayed.

**Extracting the Data**

The data was available on Kaggle.com in CSV format. I then used the ‘read\_csv’ command in Jupyter Notebook to save it to the variable ‘flights’.

**Data Exploration and Preparation**

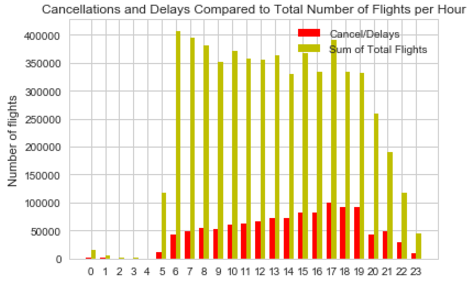
During this phase, I reviewed all the columns in the Dataset and tried to find patterns or similarities between the columns. As I explored the data, I found many of the columns could be used to predict cancellations or delays, but I chose the columns most relevant to predicting my hypothesis. These columns include each period of when a flight takes off to when it lands such as ‘Wheels Off’, ‘Taxi In’, ‘Elapsed Time’, and others. Many of the columns were also highly correlated with each other. Columns were created with equations such as ‘ELAPSED\_TIME = AIR\_TIME + TAXI\_IN +TAXI\_OFF’ and ‘ARRIVAL\_DELAY = ARRIVAL\_TIME – SCHEDULED\_ARRIVAL’, as well as others. Due to the high correlated columns, I decided to not include these further into my research.

One of the challenges I had with preparing the data for a logistic regression model was that there was no column that included both cancellations and delays. A flight is considered late if it arrives at the gate 15 minutes after its scheduled time of arrival. There was no column that included this information, so I had to create one called ‘DELAYED\_15’. Afterwards, I needed to create a column that indicated whether a flight was cancelled or delayed. I defined a function called ‘cancdel’ that ran a simple loop which created this column called ‘CAN\_DEL’ in a new DataFrame called ‘flights\_main’. I created this new DataFrame because I was receiving a Memory Error due to the high amount of memory this original DataFrame was using. I then saved the ‘CAN\_DEL’ column into the original DataFrame and I was able to start my calculations. Before I could start forming and fitting my logistic regression model, I had to create dummy variables for each month ranging from 1-12, as well as dummy variables for each airline which indicate its Airline code.

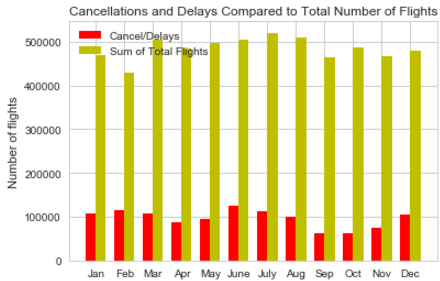
**Observations of Features**

An important part of the research is determining when the occurrence of flight delay or cancellation is highest. I prepared 3 different graphs depicting the ratio between delays and cancellation in relation to the total volume of flights:

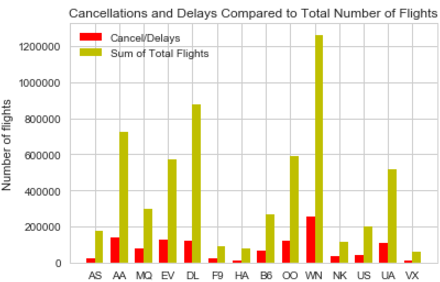
**PER HOUR RATIO OF CANCELLATIONS/DELAYS COMPARED TO TOTAL VOLUME OF FLIGHTS**



**PER MONTH RATIO OF CANCELLATIONS/DELAYS COMPARED TO TOTAL VOLUME OF FLIGHTS**



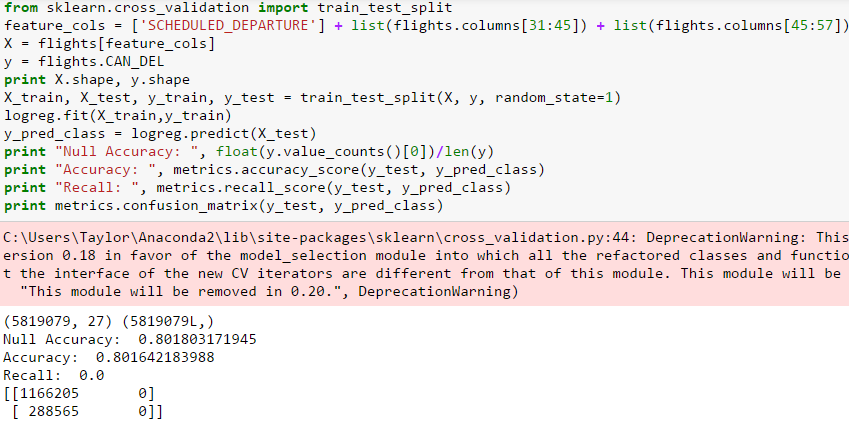
**PER AIRLINE RATIO OF CANCELLATIONS/DELAYS COMPARED TO TOTAL VOLUME OF FLIGHTS**

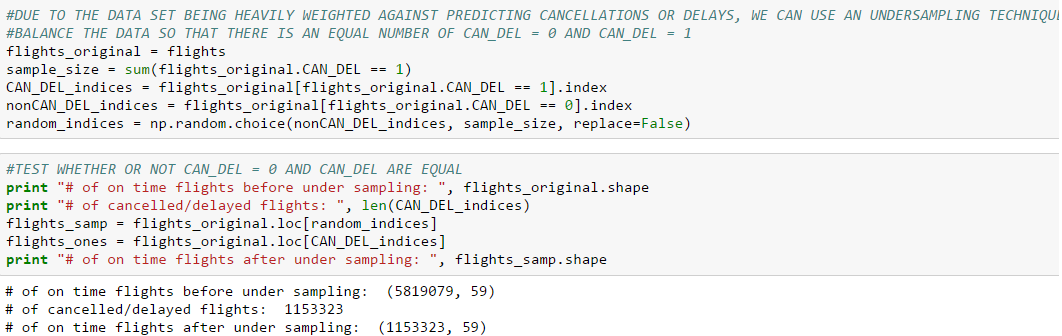


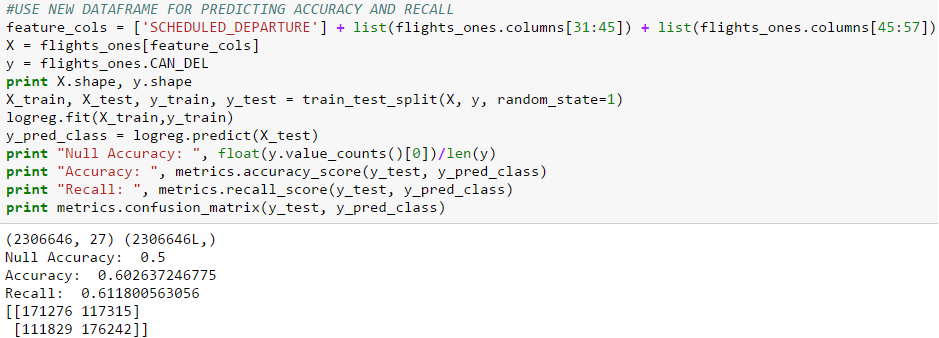
**Using Logistic Regression and Modelling Process**

The model I am using will have one response variable called ‘CAN\_DEL’ which has two possible values. If the value is 1, then the flight is either cancelled or delayed by 15 minutes. If the value is 0, the flight is not cancelled or delayed. Out of the 5.8 million flights, there were about 1.15 million that were cancelled or delayed, which means that close to 20% of all flights in 2015 were cancelled or delayed past 15 minutes.

When using Logistic Regression and using an ‘accuracy’ score to determine the results, I found that the model’s accuracy (80.1%) was about the same as the null accuracy (80.1%). I used a ‘recall’ score metric and found that the model was unable to predict any cancellations or delays.



When a model cannot predict any ‘CAN\_DEL = 1’, then there must be a problem with imbalanced data. To solve the problem of imbalanced data, I found that a method called Undersampling can be used to balance it. The Undersampling problem can be solved by reducing the number of ‘CAN\_DEL =0’ to match the exact number of ‘CAN\_DEL = 1’. To do this, I took a random selection of ‘CAN\_DEL =0’ to match the 1,153,323 instances of ‘CAN\_DEL = 1’. 

After I had an even number of both instances, I ran the model again with the same method. The null accuracy of this model will of course be 50%, and the model performed better than the null with an ‘accuracy’ score of 60%.

**Conclusion and Next Steps**

After running the calculations using Undersampling, I found that the featured columns are good estimators of predicting a cancellation or delay on United States domestic flights. I do believe that with more research and time, other estimators such as weather for each flight will perform better. Also, other models could be used to better predict flight cancellations and delays.