Airline Flight Delay

**Executive Summary**

In this project, I have tried to use the real flight data from the United States department of Transportation to make a classification model which tries to predict if a flight will be delayed by more than 15 minutes or not.

This significance of such analysis is critical in the airline industry. The airline industry can use this analysis to see if there are any specific reasons that the flights are being delayed and hence, they can try to fix those issues.

In this project I have concluded that there is no systematic delay in the arrival of the flights and any delay that we see is random and cannot be predicted beforehand. I have used the basic approach of decision trees to try and predict the delay. Other more advanced techniques like SVM and logistic regression can also be tried out to see if the results are different.

One thing needs to keep in mind that the data used in the model is only for one month. There might be trends in the data in the long term if we analyse long term (~5 years) of data. The evaluation of the model results can also be improved as well.

**Introduction**

This data is generated from the United States Department of Transportation website for all domestic airlines for the month of February 2019. I will try to predict that whether a flight will be delayed by more than 15 minutes or not based on the independent features that are available before the scheduled take off time of the flight. The technique that I have used is the decision trees. I have tried to predict the delay by using a single decision tree and by the random forest method. The desired outcome is to predict as accurately as possible; the likelihood of a flight being delayed. I have used python for this analysis.

**Methodology**

There are 3 steps in the analysis:

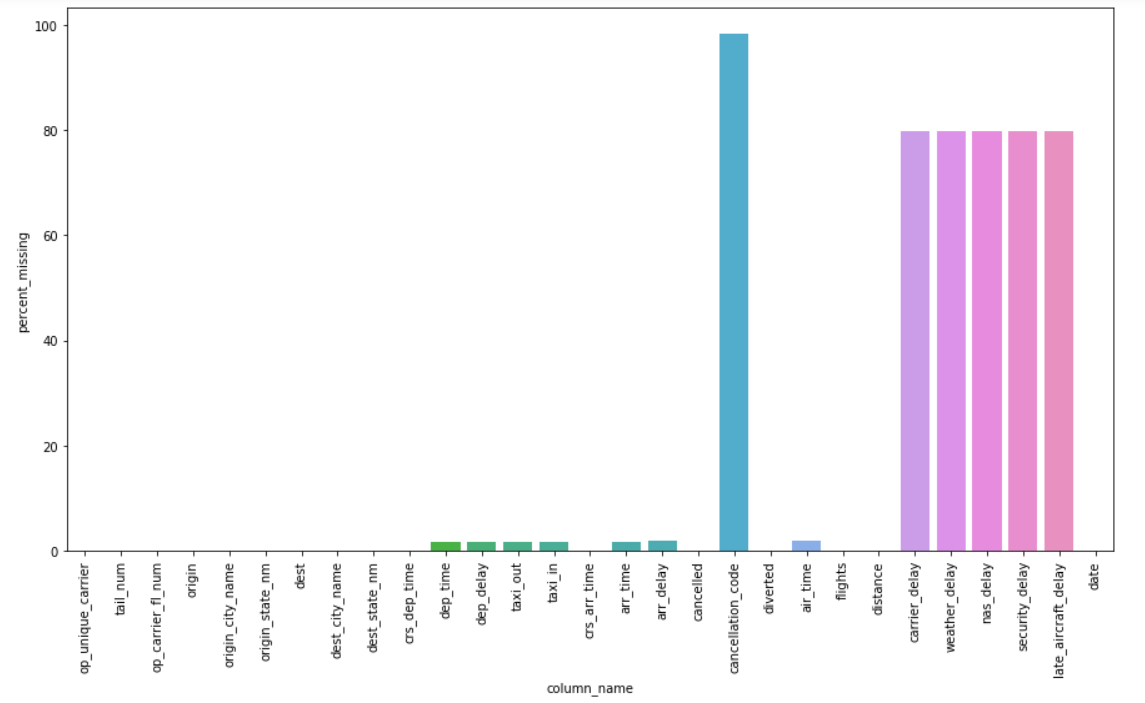
1. Data exploration and cleaning
2. Analyzing the concerned dependent and independent variables
3. Fit the models and interpret the results

I have used the following python libraries in the analysis:

1. Pandas
2. Numpy
3. Scikit-Learn
4. Seaborn
5. Matplotlib
6. Datetime

Our data has 34 columns and ~700k observations. I have only used 100k observations for this analysis as we have enough observations in both categories (delayed and on time).

First, I will only be using those columns which will be used in this model, so we drop the remaining columns. Then I move to identify any NULL values. The NULL values in the desired dependent variable only occur where the flight is either diverted or cancelled. So, I have filtered the data to only keep observations where the flight has reached the destination uninterrupted. Here is a snapshot of NULL values in all the columns in the data set.

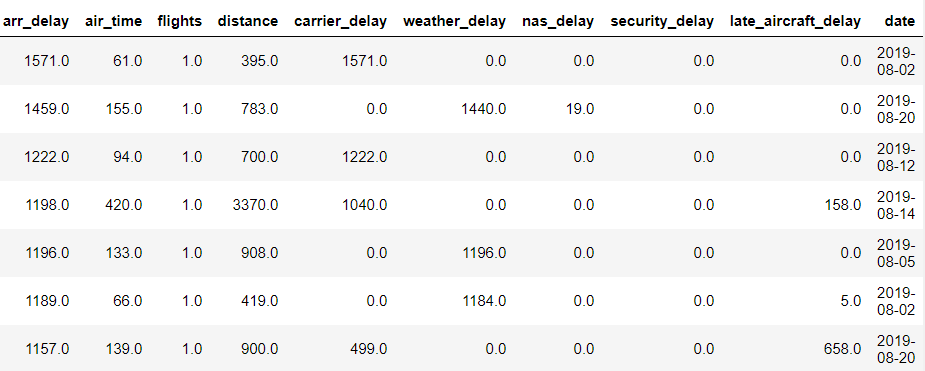


We can clearly see that all the columns that have NULL values are the flights which have been cancelled. The last few columns which have high NULL percentage, are only available after flight has landed so we will not use them in the analysis.

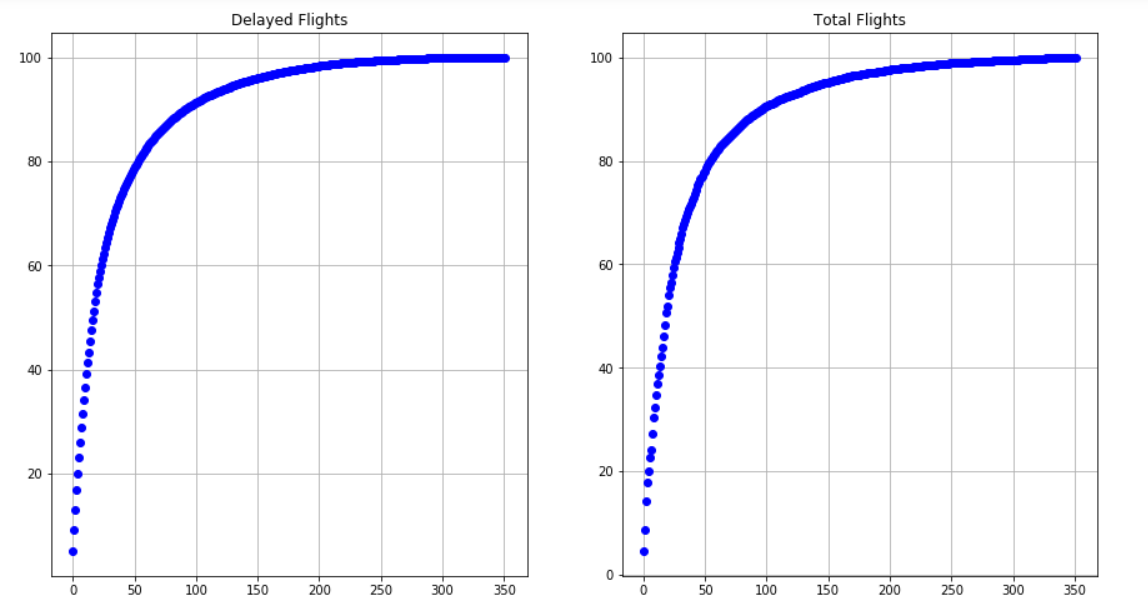
The final list of features that I have used in the predictions is:

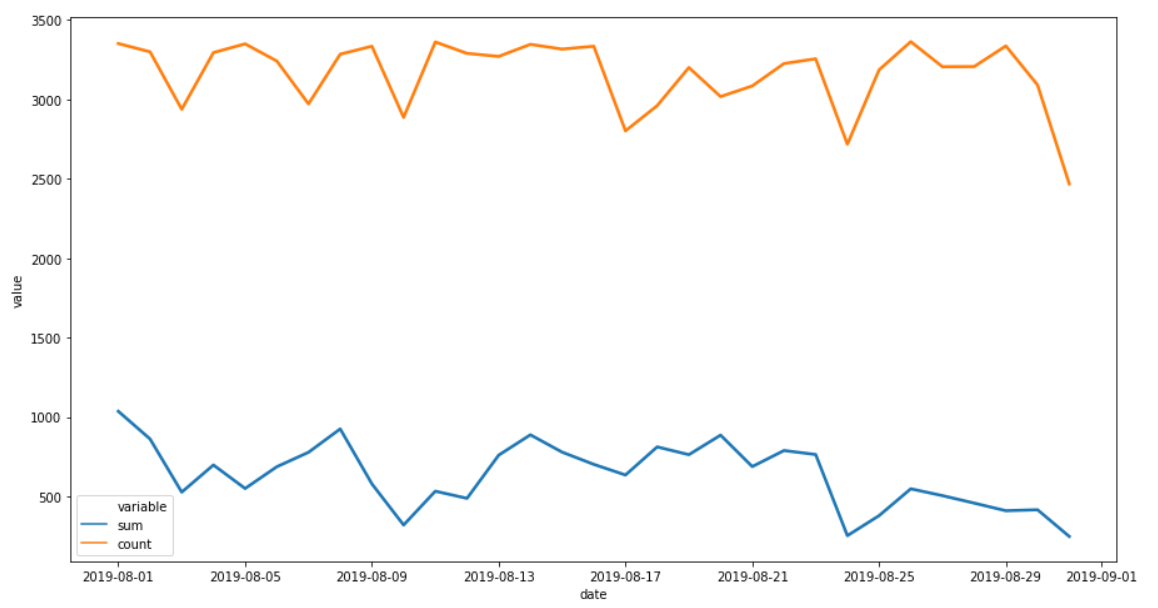
1. Unique Carrier
2. Tail Number
3. Flight Number
4. Airport, city and State of flight origin
5. Airport, city and State of flight destination
6. Scheduled Arrival and departure times
7. Total scheduled airtime
8. Total distance between airports
9. The date of scheduled flight departure

After the first stage, I analyzed the dependent variable to find out that arrival delay time has some negative values. After some exploration, I found out that these are legitimate values and represent the flights that landed earlier than scheduled time. I also checked to verify some extreme delay values, which also seem fine as the delay is legitimate, explained in the delay variables below.

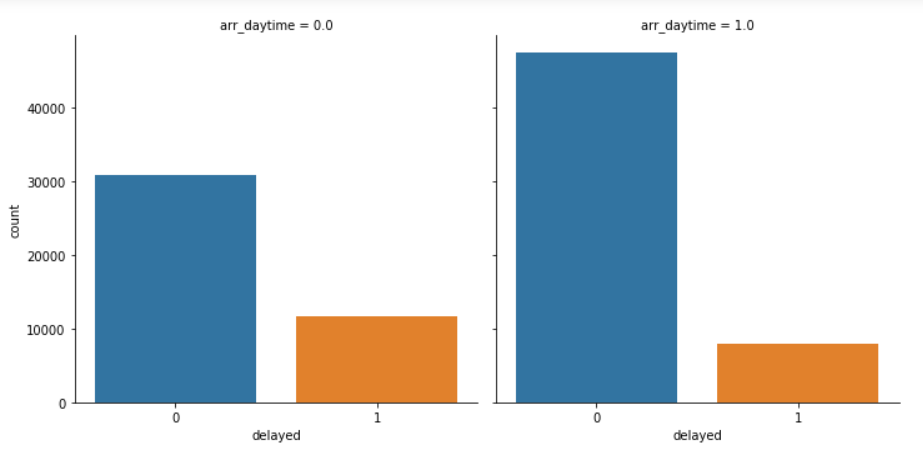


I explored the departure and destination airports to find out that around top 100 airports have ~90% of the flights out of a total of 360 airports.



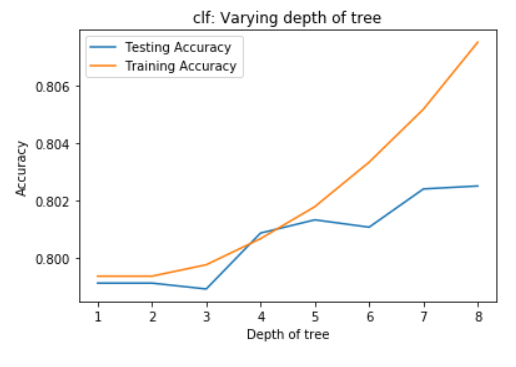
I don’t see any specific day within the month where we see any anomalous trend as well. 

I have created 2 new features which seem to have a very good correlation with flight delay. Assuming that in the month of February, the day starts at 7 AM and it is dark again by 5 PM, I have marked each flight’s scheduled arrival and departure time to be during the day or night.

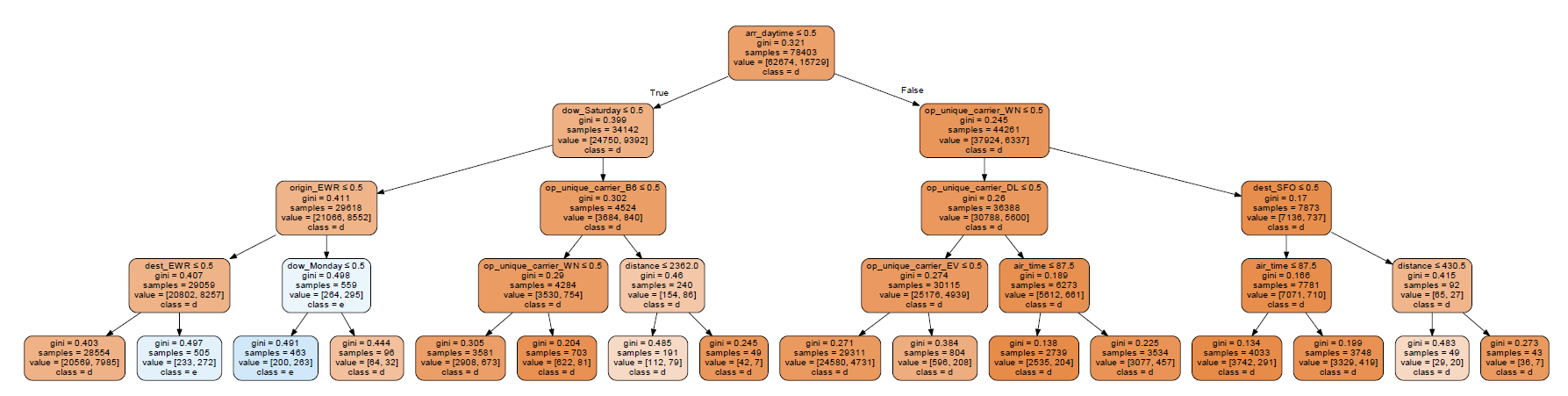


**Results**

For a single decision tree, the results are not encouraging. I tried o find out the optimum depth of the tree by iterating the accuracy.

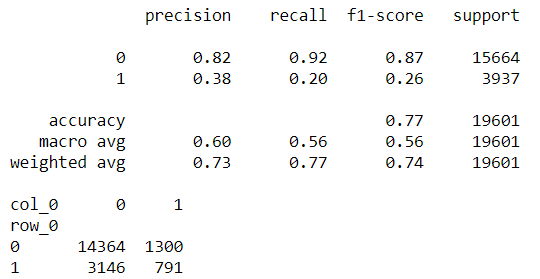


I used k=4 as the accuracy is reasonable stable. The model has a recall of 3% when predicting a flight delay. Recall is the ratio of true positives to total positives in the test data set. If we were to reduce the depth, the recall would drop to 1%. Increasing the depth did not really improve performance. From this model, I concluded that there is not clear pattern in the delay of flights.

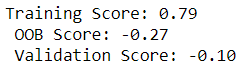


Single Decision tree with max depth of 4

Then I tried the random forest decision tree to validate my findings. The random forest gave some better results as the recall for delayed flights increased to 20% with this model. If the probability of the random forest was more than 50%, I considered the prediction as delayed.



Finally, I used the scikit-learn accuracy measure to check the model accuracy for predictions on the training and test data separately. While the score on the training data is close to 80%, the validation set had a negative score. This clearly showed that the model has failed to find any systematic patterns in flight delays. The OOB score is also negative. This means that the Out Of Bag samples in the training data set also do not predict the outcomes correctly. This also verifies that the model is highly overfitting and cannot be used to make any predictions.



**Conclusion**

After analysing the results of both the models, I concluded that there is no pattern in the delayed flights and the delay is completely random. The accuracy of the model that we saw on the training set shows that the model is overfitting and if we limit the depth of the random forest tree, it fails to predict any delayed flight!

This analysis can be extended to predict when a flight is more likely to be cancelled or what is the expected delay in minutes for each flight. This remains a future objective.