Predicting Patterns for Flight Delay using Machine Learning Concepts

Diptesh Nath (dnath@uncc.edu)
Vinayak Kolhapure(vkolhapu@uncc.edu)
Yagnesh Vadalia(yvadalia@uncc.edu)

Abstract

Air travel is statistically the safest means of travel and obviously the quickest. In todays world, there are thousands of air planes that take off and land while transporting millions of passengers around the world. As a business, as an operation; it is extremely complex due to the minute margins for error. A minutes delay for a flight in taking off, or landing could have serious financial implications for the airlines and operational difficulties for the Airport authorities. Although, there are several uncontrollable factors that contribute to this, there definitely are many others that can be predicted and avoided. Our goal is to analyze airlines and airport data to determine factors that contribute to flight delays, in an attempt to contribute to the phenomenon of air travel. There is a lot of research conducted to draw models for airports like JFK, ORD and EWR.

1. Introduction

We have studied patterns in flight delay and tried to develop a model to find the cause of their delay and to predict for future delays. We have used flight data from Bureau of Transportation Statistics and weather data from the Weather Underground. Logistic regression and Random forest algorithm were used to detect the pattern of airport delay, aircraft arrival delay and schedule performance. These models are then integrated in the form of a system for aircraft delay analysis and airport delay assessment. We have mainly researched on the data for John F. Kennedy International Airport and HartsfieldJackson Atlanta International Airport. (The results of the research show that the daily average arrival delay at the said airport is highly related to the departure delay at other airports. The daily average arrival delay can also be used to evaluate the delay performance at ATL. The daily average arrival delay at ATL is found to show seasonal and weekly patterns, which is related to the schedule performance. The precipitation, wind speed and visibility also contribute to the arrival delay. This research also investigated the delays at the flight level, including the flights with delay 0 minute and the flights with delay 15min, which provide the delay pattern of single arrival flights. The characteristics of single flight and their effect on flight delay are considered. The precipitation, flight distance, season, weekday, arrival time and the time spacing between two successive arriving flights are found to contribute to the arrival delay. While it was possible to calculate the immediate impact of originating delays, it is not possible to calculate their impact on the cumulative delay. If a late departing aircraft has no buffer time in its upcoming schedule, it will continue to be late. If that aircraft enters a connecting airport, it can pass its lateness on to another aircraft. In the research we also consider purifying only the arrival delay at Origin Airport. The model makes it possible to identify the pattern of the aircraft arrival delay. The weather conditions are found to be a significant factor that influence the arrival delay. But now-a-days due to advances in technology flights are able to land in well-equipped airports even in case of adverse conditions.

1.1. Flight Delay

Flight delay is complex to explain, because a flight can be late due to an array of problems. The most commonly understood ones are attributed to operational errors or difficulties on part of the airlines, infrastructural problems on part of the arrival terminal or airport, bad weather and visibility due to which the aircraft might have to circle around the runway before given to go ahead. The problem that is more difficult than others to analyze is the delay caused by other aircrafts scheduled to arrive on the same runway as the aircraft in question. Also, one airplane if delayed on one of its legs, could cause a delay on its subsequent legs. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

are taken. Even if complex, there exist some pattern of flight delay due to the schedule performance and airline itself.

1.2. Data

As mentioned in the introduction section, we have used the flight data from Bureau of Transportation Statistics and weather data from the Weather Underground. The Bureau of Transportation Statistics API provides data with a wide range of flight attributes which can be downloaded for each month. We used the data for all twelve months of 2015 to train our model and used nine months of 2016 data to test. For the data of each month we extracted different sets of data for diverse experiments,

- We extracted random data from the test and training sets
- For flights departing from ATL and arriving at NYC

The attributes that we considered in the flight data are Month, Day of Month, Day of week, Unique Carrier, Destination airport, Departure time, Days from Holiday The last attribute, Days from Holiday was a derived attribute which we computed based on ten US holidays and calculated this attribute by subtracting each day of flight departure from the nearest holiday. We figured this attribute could act as a significant contributor. The weather data was used from the Weather Underground API. The weather attributes for each day were mapped to that of the flight data by using one to one mapping. The attributes used from the weather data are temperature, humidity, visibility and wind speed.

1.3. Methods

We used two different classification algorithms to build our model for predicting flight delays. Logistic Regression and RandomForest classifier. We used the Scikit learn library to train and test our data. The Scikit learn library provides implementations for most of the classification algorithms and they have support for a number of parameters which allow us to tune the model to our liking. Also, as the library is in development, there is a lot of support and discussion surrounding it, which provides with incredible help. The documentation is neat and clear as well. Logistic Regression was used as our base classifier as it was easier to implement and test. To use this and the RandomForest though, we had to categorize all our features, the details of which will be in the subsequent sections. With Logistic Regression though, we need to assume that the features are roughly linear. For this reason, we chose Random Forest as another algorithm to test the prediction accuracy as it doesnt need such an assumption and also the results are easy to understand and the algorithm can be easily tuned. We initially used the data with random samples as mentioned in the previous section and tested both the algorithms and recorded the results. Then we proceeded to use the ATL to NYC data to see if there was a better and more reliable outcome for a set of selected arrival and departure airports. With this, it was also easier to merge the weather data as we could use the attributes for Atlanta and New York. Also, initially we planned to have a binary outcome or dependent variable for Arrival delay, and then scale it to a multi-class attribute for more sophisticated prediction. The binary attribute for Arrival delay was positive is the delay of arrival was more than 15 minutes and negative, if otherwise. This variable was actually given in the data set and hence, we didnt need to compute it. For the multiclass attribute, we planned to use the minutes of arrival delay attribute and put range of delays in different buckets and use them as outcome classes.

2. Data Exploration and Visualization

In order to select which features could impact more on the arrival delay we went on using descriptive statistics and exploratory visualization to get a sense of the impact that various attributes had on the delay. First we verified the correlation between the departure and arrival delays to make sure that we could exclude departure delay from the list of features based on its relation with arrival delay as the inclusion of this feature given a strong correlation would cause a conflict in the analysis.

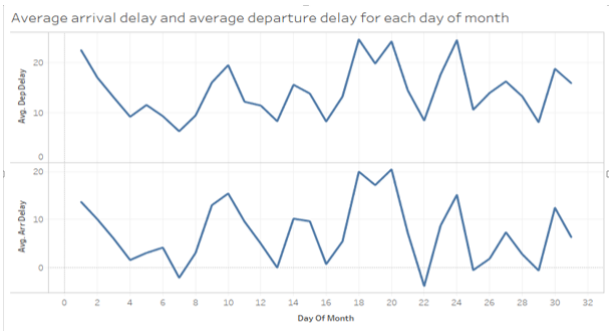

Figure 1. Correlation of Departure and Arrival Delay

Looking at the above figure, it is obvious that the departure delay is strongly related to the arrival delay

and hence we remove it from the list of predicting variables. The above plot also gives a clear visualization of the average arrival delays for each day of the month through the year 2015. Weather attributes might not always be very reliable. Hence before choosing them in building the model, we wanted to visualize if they have consistent effect on the flight arrival delay or not. The below graph shows the average arrival delay across the year of 2015 for the airports JFK and LGA arriving from Atlanta.

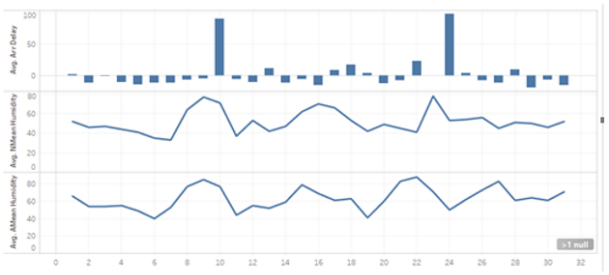

Figure 2. Arrival delay at JFK and LGA v average humidity in NYC at the time

As we see in the above plot that there is no consistent relation between humidity and arrival delay at the airport. We also tried to visualize the relation of the dew point to the arrival delay. This also, as shown in the plot below, has no clear or consistent relation with our dependent variable.

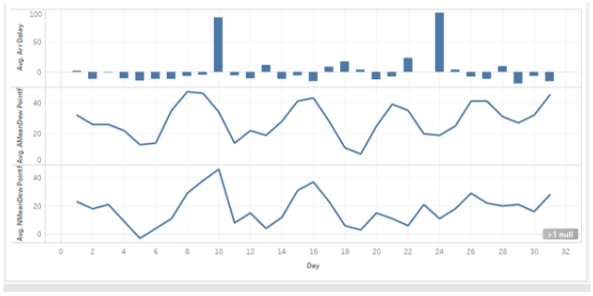

Figure 3. Arrival delay at JFK and LGA v dew point

In the plot below, we try to find relation between visibility and delay. We expect visibility to have some consistent effect on the arrival delay as experienced by many of us travelling. We can clearly see that during the end of the month, there was correlation between visibility and arrival delay. We choose to use this variable as one of the predictor attributes in the model building. In the following 4 plots, we plot the average arrival delay for the year of 2015 for each of the 4 seasons of the year winter, summer, fall and spring.

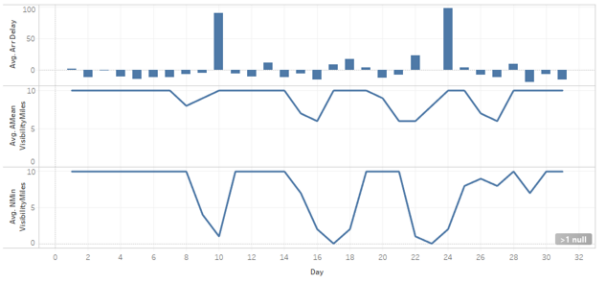

Figure 4. Average arrival delay against visibility

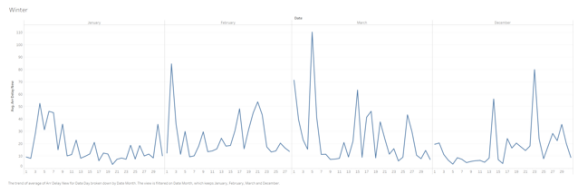

Figure 5. Average arrival delay in winter

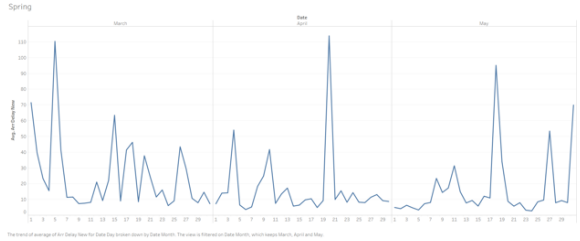

Figure 6. Average arrival delay in Spring

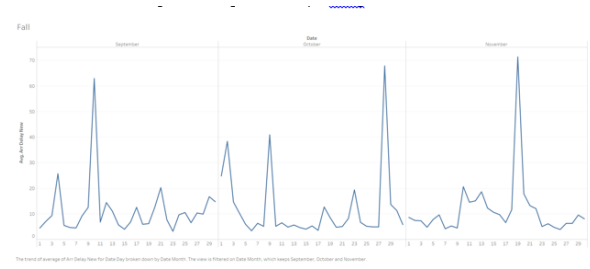

Figure 7. Average arrival delay in Fall

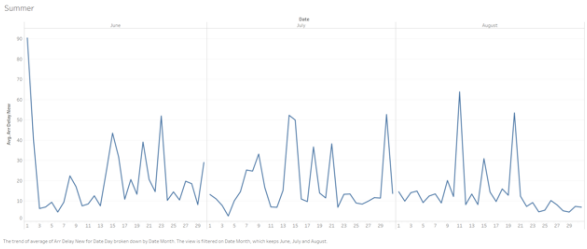

Figure 8. Average arrival delay in Summer

As we see in the above plots, there is some sequence in the delays in each of the seasons of the year. We think this observation would be valuable to do across many years than just one, but as we have used only 2015 data to train our model, we visualized this plot only for this year. According to this, we believe that time, day and month of departure or arrival are key indicators for predicting delays on a consistent basis. Delay based on behavior might not be characteristic always.

3. Data Preparation

For this project, as mentioned earlier, we get the data from the US Department of Transportations Bureau of Transportation Statistics here, and as you will see in the link, you can choose which features to be downloaded (we have listed the features we used in a previous section). Unfortunately, the way the Department of Transportation processes the downloading, we have to select which features we want and download them only a month at a time. Hence, this was an initial tedious task to download data for each month, take random samples from each set, modify the features and then merge to compile the yearlong data set. After downloading the dataset by selecting our features, we used R to pre-process the data and to setup our new days from holidays variable. At the end of the preparation, we categorized variables like unique carrier into numerical categories to be used in the classification models. Once we import each data set using R, we first see how many NA values are present with respect to the total number of samples. In each case, it was less than 2 percentage of the samples, and hence we went ahead and removed the NA valued attributes from the data set. Then we see the summary of the data to make sure everything is alright and most importantly that we have no NA values anymore. After this, we proceed to create our holidays variable. We first take the exact dates of the 10 US holidays in consideration and convert them into date type in R. We go ahead and create a function to make things easier. We have gotten rid of the YEAR attribute earlier as we know what year we are dealing with here. The function takes inputs of month and day from the dataset and finds the minimum distance between the date and our holidays. Instead of running this function on all the samples of the data set, we run it on the 365 different days of the year and apply it to the corresponding records in our dataset. Once we have this feature ready, we categorize the other variables like the unique carrier ID into numerical categories. After this step, we take random samples of the data and merge data for all of the months together first for the training data set and then the testing dataset.

4. Classification Model

As mentioned earlier we used Logistic Regression as our baseline classifier and we moved on to use RandomForest later in search for better accuracy and score. Logistic Regression, as far as we know is more suited to be a binary classifier but sci-kit learn does support the one vs rest classifier. We will have a look at it later. The scikit learn library also provides support for different solvers in the Logistic Regression Class like newton-cg, sag and lbfgs which support only L2 regularization. The liblinear solver supports both L1 and L2 regularization. There are many resources out there to understand the difference between L1 and L2 regularization and we will reference some of them in the section. For Random Forests, we have seen in studies that it has lower classification error, hence, more accuracy and better f-scores. RandomForest classifier is also easy to understand as it is an amalgamation of many decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and to control over-fitting. It is almost like an implicit cross-validation. We used RandomForest classifier to try get better accuracy over the baseline classifier and we planned to use the confusion matrix for both these classifiers to see which one predicts more delays. We arent much interested in the model predicting more flights that didnt suffer delay but interested in predicting more instances of delays true positive.

5. Interpretation and Comparison of Results

As mentioned earlier, we started using the Logistic Regression classifier first on the data and the attributes that are described in the previous sections. With this and the RandomForest, we used the 2015 data for training and 2016 data for testing before going on to calculate the precision, recall, f-score and accuracy. For Logistic Regression classifier, with the random data, we found that the classifier was able to predict the binary arrival delay with an accuracy of 60 percentage, and for the data which had flights only from Atlanta to New York with an accuracy of 55 percentage, with the confusion matrix below-

In the first matrix we can see that the number of True-positive samples are very encouraging in the sense that the classifier predicts a good number of flights that got delayed. Moving on, we used the same data for training the Random Forest classifier and found a largely improved overall classification accuracy. For random

Random data for random source and destination flights

CONFUSION MATRIX		
	0	1
0	3708	2588
1	533	900

60% accuracy

Confusion matrix for New York and Atlanta

CONFUSION MATRIX		
	0	1
0	2602	2646
1	386	1079

55% accuracy

Figure 9. Confusion Matrix for Logistic Regression

data, we got an accuracy of 82 percentage and an accuracy of 80 percentage for the flight data from Atlanta to New York city. Although, the confusion matrix clears the picture about accuracy. Despite the accuracy being really high for this classification(Figure 10), we notice a startling statistic in terms of the number of samples for true-positive. It is very low. Hence, for the objective of this project, this is something we cannot use. Hence, we decide that logistic regression is the better fit for our data which is justified by the fact that it is a good binary classifier. We took these results forward and tried to work with a multi-class classifier with each class having a range of arrival time delays-

- For no delay or early arrivals, the class is 0
- For delays of less than 15 minutes, the class is 1
- For delays of more than 15 minutes and less than 30 is class 2
- For delays of more than 30 minutes and less than 45 is class 3
- And finally for delays of more than 45 minutes is class 4

We tried this approach based on the feedback during the class presentation. Also, this was done for data with destination as JFK to control the number of samples for quicker analysis. This analysis led us

Confusion matrix for random data,

CONFUSION MATRIX		
	0	1
0	6213	83
1	1303	130

Accuracy increased from 60% to 82%

Atlanta to NewYork data,

CONFUSION MATRIX		
	0	1
0	4707	541
1	819	646

Accuracy increased from 55% to 80%

Figure 10. Confusion matrix for Random Forest classifier

to a rather strange set of results, we only got predictions for classes 0 and 4 and accuracy of 61 percentage for the Logistic Regression classifier, indicating that we stretched it too far with this. We then tried the Random Forest classifier with the same parameters as before (50 sub-trees), and got the following results(Figure 11) given below.

Again, the accuracy doesnt spell everything, the number of samples that are predicted correctly for actual delay are less compared to the samples classified for 0. Here, we discover a shortcoming in our data. The number of samples for class 0 significantly out-weigh the number of samples for other classes. This is one point of improvement that we will take for work beyond this report.

6. Conclusion

Overall, based on the classification results and the exploratory analysis done on the features, we discover that the features most contributing for arrival delay are time of arrival during the day and time of the year of the flight. The days from holidays feature was also interesting, but could be optimized and made more sophisticated to the approach that we used and described earlier. In the end of the report, we think that to an-

Confusion Matrix					
	0	1	2	3	4
0	35444	820	78	6	747
1	9516	251	26	1	295
2	4414	140	14	1	162
3	2418	80	3	0	134
4	6137	180	19	0	630

Figure 11. Random Forest for multi-class

alyze and to report more interesting results, we need to explore more knowledge of the domain and derive more features that can correlate better to the arrival delay. We hope that this report can work as a baseline for other analysts to take forward, tune the classifiers or use different classifiers in search of more interesting results. After this work, we have a lot of new items that we would like to add to this to take this analysis forward. We will list those items in the future work section.

7. Future Work

In machine learning and data analysis, like in many things in life, there is always scope for improvement. We have discovered many mistakes through the course of this work and would continue to optimize and improve our solution to get more interesting results in this space. Some of the important points or tasks are-

- Use larger data samples for more reliable prediction and use appropriate Big-Data framework to manage and test these samples, to avoid losing interesting samples to randomization.
- Explore more domain knowledge to use better features in designing an improved and more comprehensive predictive model.
- An improved model can be extended to be used as an API, which can in turn be used as an App to plan travel.

8. References

- sklearn.linear model.LogisticRegression.html
- sklearn.ensemble.RandomForestClassifier.html
- data-science-apacheh-hadoop-predicting-airline-delays
- jessesw.com/Air-Delays/
- www.usna.edu/Users/math/dphillip/jsm.pdf
- CastilloLawson-PredictingFlightDelays.pdf