AIRLINE PREDICTION AND COMPARATIVE ANALYSIS OF AIRLINE DELAYS USING MACHINE LEARNING TECHNIQUES

Amit Gupta - axg1629300, Deepan Verma - dxv160430, Komal Mukadam - kjm160030, Saumya Dixit - sxd160630, Somya Singh - sxs161331

Problem Statement

Airline On-Time Schedule Performance is a very crucial in maintaining current customer satisfaction and attracting new ones. However, there are a lot of variables that play an important part in a flight's schedule. Airlines resources are tightly connected, these delays could occur in a domino effect over time and space unless the proper recovery actions are taken.
Solving the issue of flight delay is a complex problem. Even though it's complex, there exist some pattern of flight delay due to the schedule performance and airline itself. The flight delay can show dependency on various factors like, monthly, daily and time of day based delays, and also show some preference according to airborne time, flight distance and origination areas etc. This is the interest of this project.

Description of Input Data

Dataset Extraction: Source – Bureau of Transportation Statistics (United States, Department of Transportation)
Attributes that are mainly used for the prediction of delay in flights includes : 'year','Month','Day_of_month','Day_of_week', 'origin', 'Dest', 'Crs_dep_time', 'Dep_time', 'Dep_delay',crs_arr_time','Arr_time','Arr_delay','Cancelled',distance','CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY', 'SECURITY_DELAY','LATE_AIRCRAFT_DELAY'.

Motivation and challenges

Big data is the driving force behind prediction systems. We need huge amounts of data to learn about any kind of flight delays and to understand the trend behind the delays, then only we can predict future delays for the flights. This prediction system works mainly with the history data and these systems cannot do its job without sufficient data and BIG DATA supplies plenty of data such as what are major factors affecting flight delays and by analysis of trends using these factors for the prediction system to provide efficient and almost relevant predictions, we were able to process the data quickly and it is obvious that traditional technologies will fail processing the huge amounts of data, so it will not suffice to just have big data in order to provide effective predictions.
Few of the challenges that were encountered during the project were data handling refining and preprocessing and identifying the major factors affecting flight delays.

Technical Approach

There are various factors and attributes that are used to analyze trends in delays. The most significant impact factors on the flight delays were chosen for the prediction system.
In pre-processing we have converted the csv data into vector representation, so that the comparative feature analysis and discretion would be easy for the model. Also filters have been applied for appropriate feature selection
Regression techniques are used to determine which factors are contributing to the prediction of a dependent variable. Regression analysis establishes the relationships amongst various independent variables and the target variable. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). We try to analyze the whole model by changing various independent variables and keeping the dependent variable

fixed to see how the dependent variable changes with change in the independent variables. Three major uses for regression analysis are analysis based on cause, effect of forecasting, and forecasting on trends.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Relationship between one dependent binary variable and one or more independent variables is aptly described by the logistic regression.

The most common predictive analysis is linear regression. It models the relationship between a dependent variable y and one or more explanatory variables (or independent variables) denoted X. The linear regression assumes the distribution to be normal and the data to be linearly correlated. Multiple linear regression is used for more than one explanatory variable, the process.

Accuracy estimation refers to the process of determining what all are the correctly predicted values among total number of observation using the current classifier.

Classification involves the precision. The precision for a class is the number of true positives that is, the number of correctly labeled items belonging to the positive class divided by the total number of elements labeled as the positive class that is the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class. Recall in this context is defined as the number of true positives divided by the total number of elements of the positive class that is the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been.

Binary Classification Statistical Analysis, the F1 score measures a test's accuracy. Precision p and the recall r of the test compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The Weighted average of the precision and recall is often referred to as F1, having values 1 and 0.

The balanced F-score is the harmonic mean of precision and recall requires multiplying the constant of both scales the score to 1 when both recall and precision are 1.

The mean squared error calculates the average of the squares of the errors. Error is the difference of actual and the predicted value.

A confusion matrix is a table that is often used to describe the summary of a model and how it performed on the given data. How well the data fit the model. A confusion matrix can also be considered as a visualization of the performance of an algorithm. Each column has attributes of predicted class and each row of the actual class. This forms all the possible combination of actual and predicted and helps in understanding the relationship with each other.

A receiver operating characteristic curve, i.e. ROC curve, is a graphical plot that tells how accurate our prediction is. The closer the graph is to one the better the prediction. ROC curve is plotted between the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is ratio of true positives and the sum of true positives and false negatives. The false-positive rate is a false alarm indicator and can be calculated as ratio of false positives and sum of false positives and true negatives. The ROC curve is thus the sensitivity as a function of fall-out. ROC analysis helps us remove the suboptimal models and only keep the ones having the area under the curve close to 1.

The area under the curve is a major tool in deciding the accuracy of the model. This area equals the probability of selecting a positive example randomly than selecting a negative random example. An area of 1 is a perfect test for accuracy; an area of .5 shows a highly ambiguous model.

We have implemented Logistic Regression to predict Departure Delay, Arrival Delay with and without given origin and destination. Accuracy, Precision, Recall, F1 and Area under ROC. Curve have been calculated for the prediction model in this project. We have also implemented Linear Regression to predict the number of minutes for both Departure and Arrival Delay

Experiments and Results:

Arrival Delay for DFW:
1. Precision = 0.25
2. Recall = 0.59
3. F1 = 0.35
4. Accuracy = 0.54
5. Area under ROC = 0.55877
6. Mean Square Error = 0.2351
7. Confusion Matrix

		Predicted Class	
		True	False
Actual Class	True	80241.0	72134.0
	False	16739.0	24183.0

Departure Delay for DFW:
1. Precision = 0.24
2. Recall = 0.61
3. F1 = 0.35
4. Accuracy = 0.55
5. Area under ROC = 0.5736
6. Mean Square Error = 0.2278
7. Confusion Matrix

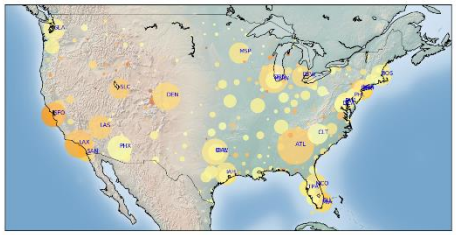
		Predicted Class	
		True	False
Actual Class	True	83682.0	71913.0
	False	14901.0	23264.0

Departure Delay for Flights from DFW-JFK
1. Precision = 0.27
2. Recall = 0.40
3. F1 = 0.33
4. Accuracy = 0.59
5. Area under ROC = 0.5250
6. Confusion Matrix

		Predicted Class	
		True	False
Actual Class	True	236.0	127.0
	False	72.0	48.0

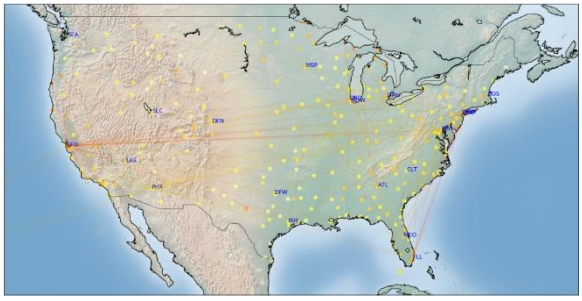
Analysis:
Airports with in comparative delay proportion

Airports with average delay is represented in this plot below. The airports with bigger circle represent more flights from the airport and the dark colored airports show more delays.



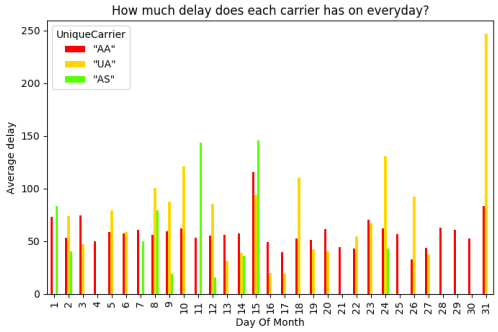
The airports with maximum delays include
ATL – Atlanta
ORD – Chicago
LAX – Los Angeles
SFO – San Francisco
DEN – Denver

Routes with delay : The plot below presents the routes with delay in flights being running on these routes. The darker routes represent more delay in the flights operating on that route.

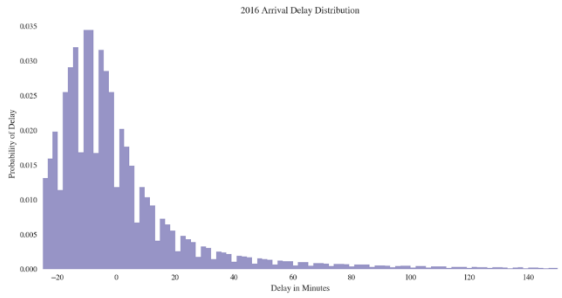


The identified routes with maximum delays include
SFO-BWI (San Francisco – Washington)
JFK-FLL (NewYork – Florida)
SFO-JFK (San Francisco – NewYork)

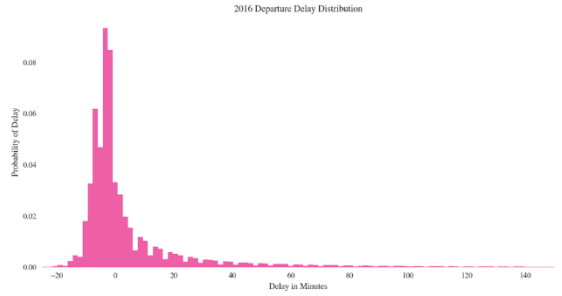
Every day Delay with carriers : Everyday delay contribution of each carrier across the month.



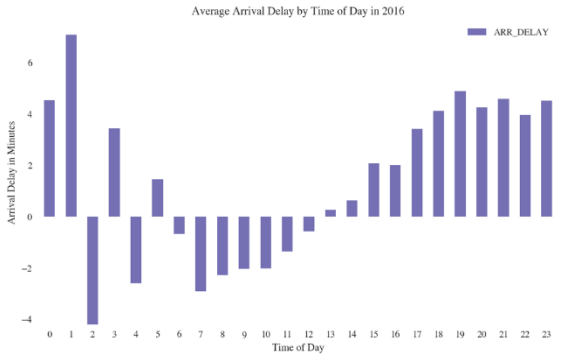
Arrival Delay Distribution: Plot below shows probabilistic distribution of flights with Departure delay in minutes.



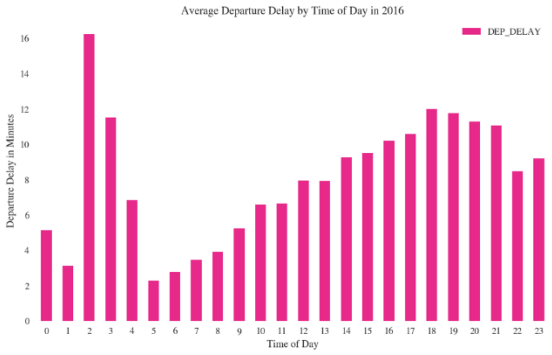
Departure Delay Distribution: Plot below shows probabilistic distribution of flights with Arrival delay in minutes.



Average Arrival Delay By Time Of Day : Plot shows average arrival delay at different times of day for flights distribution with delay in minutes.

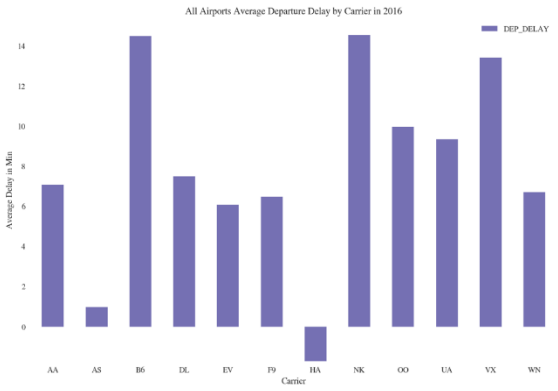


Average Departure Delay By Time Of Day : Plot shows average departure delay at different times of day for flights distribution with delay in minutes.



Airports Arrival Delay By Carrier : Average arrival delay in minutes distribution for all airlines carriers.

Airports Departure Delay By Carrier : Average departure delay in minutes distribution for all airlines carriers.



Conclusion

[1] With the accuracy and MSE we have got as the result of our Regression, we could conclude that the delay is being satisfactorily predicted.

[2] Various analysis plots included gives a good idea of delay vs contributing factors.

[3] Analysis also represents the majorly affected airports, routes, cities, carriers etc by flight delays.

Future Work

[1] Real-time interface wherein user while searching for flights and during reservation can get delay estimates.

[2] The model could be refined to include separate predictions for holidays, time and days during which more people prefer to travel etc.

[3] To include more data as the training dataset to improve the predictions further.

References

[1]https://scipy.github.io/oldwiki/pages/PyLab

[2] https://spark.apache.org/docs/1.5.0/mllib-statistics.html

[3]https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time