INST737 – Spring 2016

Digging into Data

Project Report

Predictive Analysis of US Airline Flight Delays

Submitted by:

Karan Kashyap, karan26@umd.edu
Lester Pereira, lester18@umd.edu
Prasad Revalkar, prevalka@umd.edu

Motivation

According to the International Air Transport Association, there would be nearly 3.6 billion annual passengers in 2016. Delayed flights affect the Airline Company in terms of revenue, profit and reputation. According to Joint Economic Committee report the annual cost of domestic flight delays to the US economy was estimated to be $31-40 billion in 2007. Thus, choosing an airline with the least delay is a challenging task. The inefficiency of the airline transportation sector increases the cost of other business sectors making them less productive. Therefore, such high level of associated delay costs motivates us in analyzing and predicting air traffic delays in order to develop a more sophisticated delay management mechanism.

Goal

In our project, we aim to predict and analyze domestic flight delays in the United States from New York to Chicago. To predict whether an individual flight will be delayed or not, we are considering the 15 minute threshold value, i.e. a flight is considered "Delayed" if it is delayed by 15 minutes or more.

Dataset

The data used for this study is the US airline data, which is available at http://www.rita.dot.gov/bts/ which is the website of US Department of Transportation. All the data used is in the public domain and does not require any additional license. The actual dataset can be obtained from the website of the Bureau of Transportation Statistics http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time. The Dataset consist of 14374 domestic flight details of the year 2015, consisting of 25 variables. Since the data was available month wise on the site we had to merge data of all months of the year 2015 using Excel. The response variable is whether or not a flight has been delayed by more than 15 min which is coded as 0 for on-time, and 1 for delay. We have considered New York as the origin city with JFK and LaGuardia (LGA) as the airports and Chicago as the destination city with O'Hare (ORD) and Chicago Midway (MDW) as the airports.

The explanatory variables include:

1. Two different arrival airports in New York City [Kennedy (JFK) and LaGuardia (LGA)]
2. Two different departure airports in Chicago [O'Hare (ORD) and Chicago Midway (MDW)]
3. Seven unique carriers
4. Day of week (1 for Sunday and Monday; and 0 for all other days)
5. Arrival Time

Methodology:

As a part of the predictive modeling, we have used the following techniques to predict if there will be a delay.

1. **Logistic Regression**:

Logistic regression predicts the probability of an outcome that can only have two values. The prediction is based on the use of one or several predictors (numerical and categorical). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Since the variable of interest is a binary dependent variable, it was typical to choose this method to measure the relationship between the categorical dependent variable and one or more independent variables.

2. **Random Forest**:

Now that we had our baseline set using Logistic Regression, we chose to implement Random Forest classification, which is an extension to the decision trees. The main advantage of this method is its high rate of accuracy since it is an ensemble method. Theoretically, it has several advantages over Logistic Regression. It can select important variables automatically no matter how many variables are used initially. Unlike stepwise variable selection in Logistic Regression, Random Forest estimates the output through Bootstrap Aggregation method. It takes into account a number of iterations of the decision trees performed (in our case, 10000 trees) and gives the best aggregated class as the output and it never over fits. Because each tree is constructed using 63% of the dataset selected at random with replacement and each node is split using the best split in a small random sample of available variables (usually a square root of the number of available variables are selected at random as a potential splitter), every tree is constructed at random and is independent from other trees. Therefore, adding trees to the forest does not cause a problem of over fitting.

3. **Naive Bayes classification:**

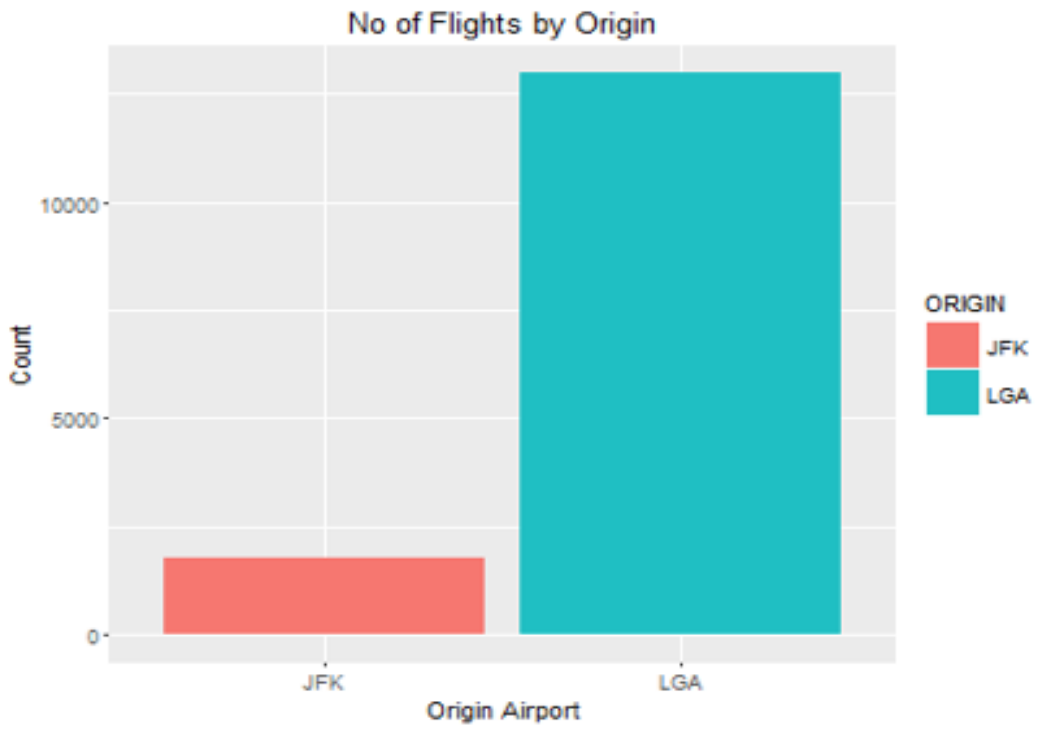
This type of classification assumes conditional independence between the predictors (attributes) and class (target) and uses a maximum likelihood hypothesis and is called a Generative Classification method. The effect of a predictor (attribute) on the class (target) is independent of the effect of the values of other predictors. This is called class conditional independence.

Implementation and Analysis:

Exploratory Data Analysis (EDA):

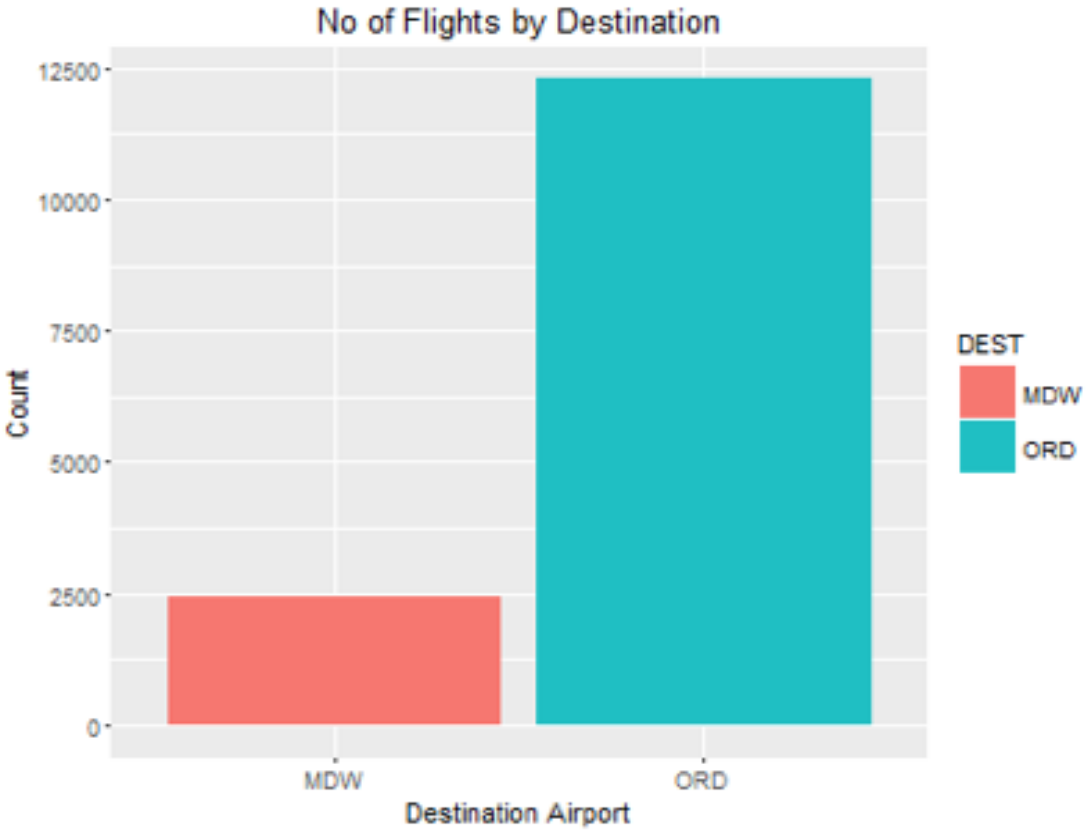
As a part of the exploratory data analysis, our goal was to explore data for variables of interest. We have performed exploratory data analysis by generating plots of explanatory variables to determine how they fare against Delay Flights, which is our response variable of interest.

Origin Airport	%
JFK	12%
LGA	88%



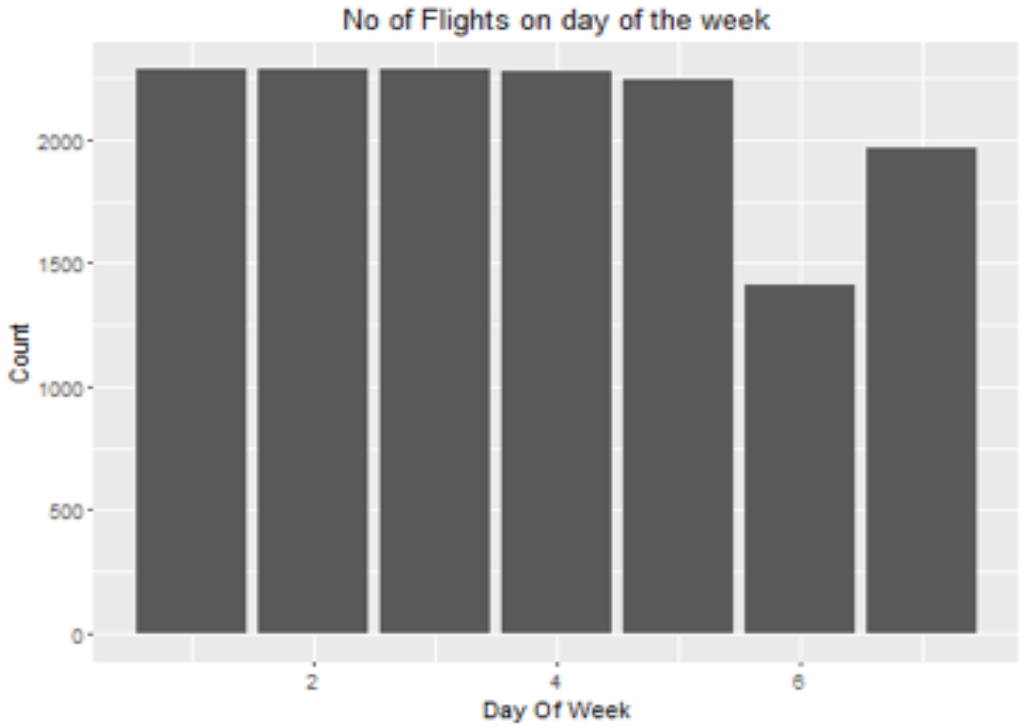
We performed exploratory analysis in the form of a bar plot for the number of flights from the origin airport. As we can clearly see that the number of flights from LaGuardia airport (88%) is very much higher than the number of flights from Kennedy airport (12%)

Destination Airport	%
MDW	16%
ORD	84%



Similarly we obtained a bar plot for the number of flights by destination. As we can see the number of flights from Chicago Midway (84%) is much higher than the number of flights from O'Hare airport (16%)

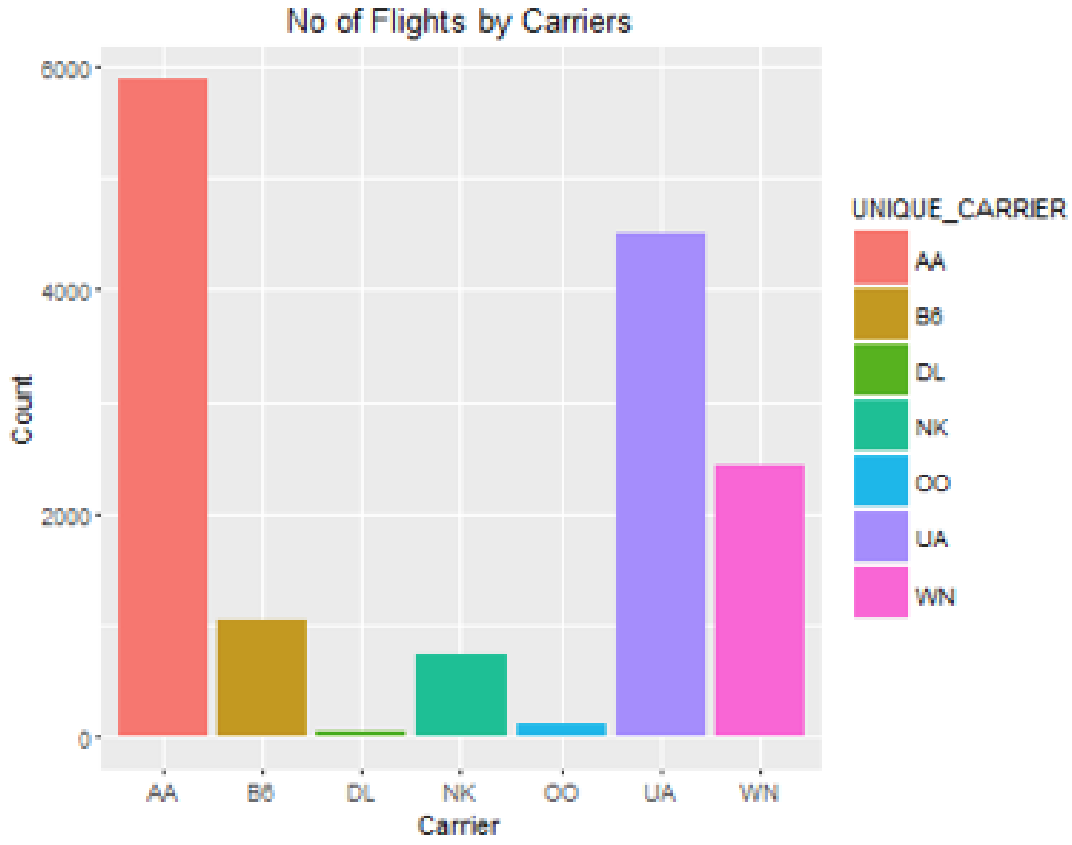
Day	%
Mon	4.06%
Tue	4.06%
Wed	4.06%
Thu	4.04%
Fri	3.99%
Sat	2.50%
Sun	3.50%



We performed exploratory analysis in the form of bar plot for the number of flight each on day of the week. As we can see the number of flights on weekends (2.50% on Saturday and 3.5% on Sunday) is less compared to the rest of the days in the week (4.06% from Monday – Thursday and 3.99% on Friday)

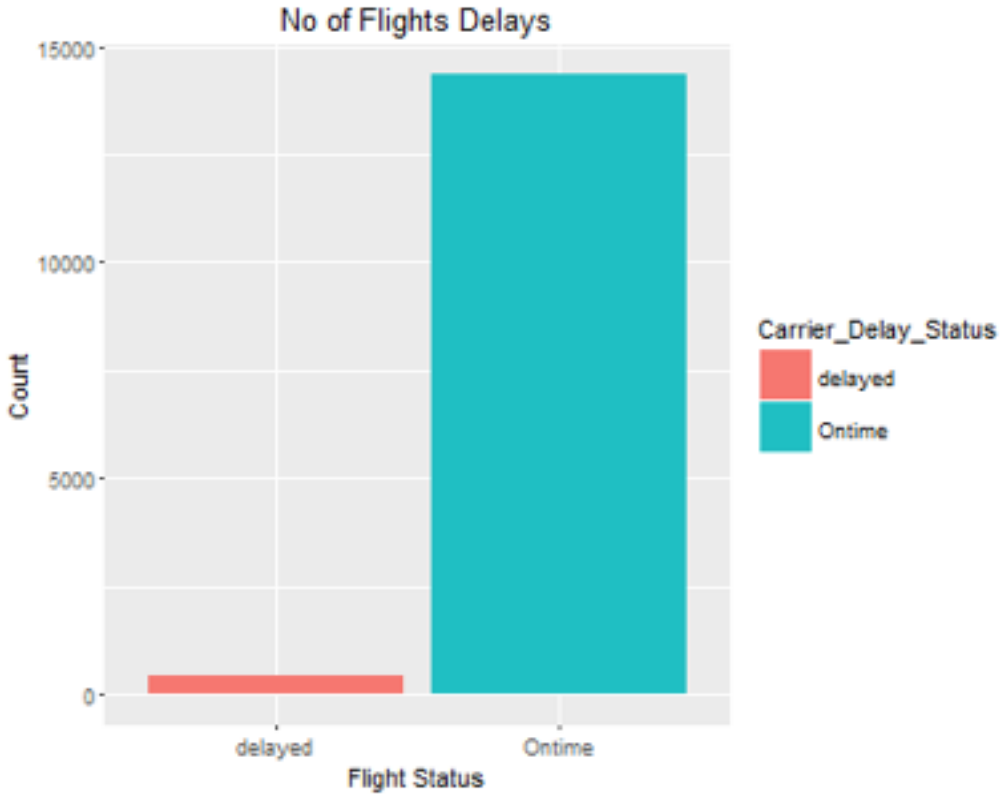
Airline	%
AA (American Airlines)	39.98%
B6(Jet Blue)	7.02%
DL (Delta)	0.22%
NK(Spirit)	4.95%
OO(SkyWest)	0.71%
United Airlines	30.62%
WN(southwest)	16.49%

5



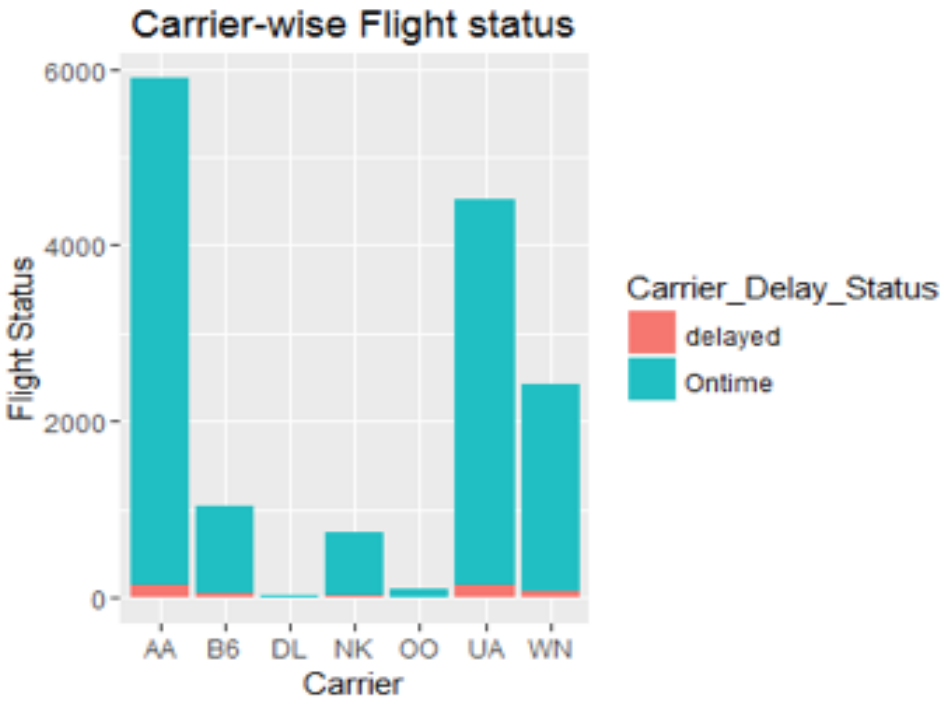
We created a bar plot for the number of flights by carriers. American Airlines flights took off the highest number of times (39.98%), followed by United Airlines (30.62%), followed by Southwest (16.49%), followed by Jet Blue (7.02%) , followed by Spirit Airlines (4.95%) and lastly SkyWest (0.71%)

Delay Status	%
No Delay	97%
Delay	3%



We created a bar plot of the flight status i.e. the number of flight delayed versus the number of flights not delayed. As we can see the numbers of flights delayed were much lower than the number of flights not delayed

Airline	%
American Airlines	34.31%
Jet Blue	8.19%s
Delta	0.00%
Spirit	8.88%
Sky West	0.22%
United Airlines	30.43%
Southwest	17.97%



We created a bar plot for the carrier wise flight status i.e. the number of fights delayed versus not delayed of each carrier. The number of flights were delayed the most for American Airlines (34.31%), followed closely by United Airlines (30.43%), followed by Southwest (17.97%), followed by Spirit Airlines (8.88%), followed by Jet Blue (8.19%) which is followed by SkyWest (0.22%). Lastly delta had no delays in the year 2015

Results:

Logistic Regression

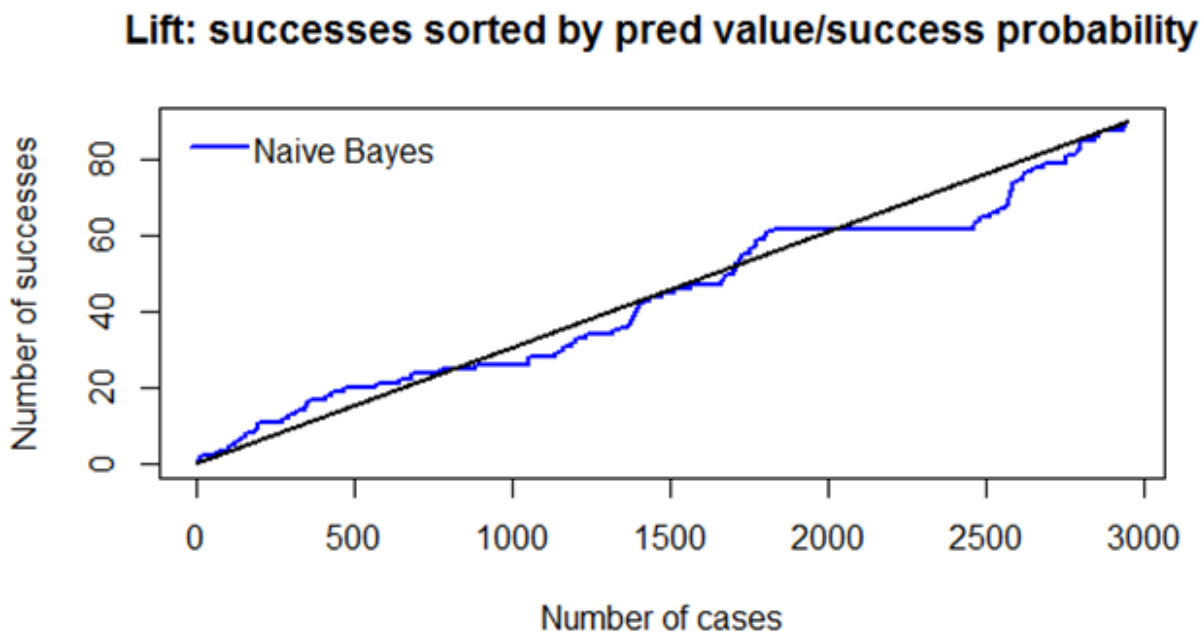
Logistic regression predicts the probability of an outcome that can only have two values. The prediction is based on the use of one or several predictors (numerical and categorical). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability.

Combination	Accuracy %
Carrier_Delay_Status~Day_of_week	97.3
Carrier_Delay_Status~Day_of_week+Origin	97.7
Carrier_Delay_Status~Day_of_week+Origin+Arr_time	98.1
Carrier_Delay_Status~Day_of_week+Origin+Arr_time+Dest+Unique_carrier	99.2

The feature combination of DAY_OF_WEEK + ORIGIN + ARR_TIME + DEST + Unique_CARRIER has the best accuracy.

Naïve Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. They can handle both continuous and discrete predictors; it assumes a distribution (can be chosen) for the continuous predictor



The accuracy for Naïve Bayes is 98.7 %

Random Forest

Random Forests classify instances by creating many decision trees and then combining the results across these trees to make a prediction. Random Forest has many advantages over single decision trees. For example, it can handle many input variables, both categorical and continuous. It also "runs efficiently on large databases. The accuracy for the model created using Random forest is 99.47%

Limitations

As the number of flights that were delayed was much less compared to the number of flights that were not delayed, we get a very high accuracy rate Delay can be due to other factors which are not dependent on the carrier. For our analysis we have not taken into consideration the weather delay, but if we had considered it our results could have been different.

Conclusion

Out of all the three classification algorithms used random forest gives us the best result with an accuracy of **99.47 %**