Detecting and Predicting Behaviors Of Customers Using NJ Flight data

Neel Kumtakar
5/9/2018

1. Abstract

Flight data is used in the modern world by passengers, pilots, flight crew and Airport employees. Using solely flight data, data on various flights, can we use Data mining and Data visualization to make observations and predictions about the behavior of customers? If so, how can these results be useful to Airports and Airlines? To answer these questions, this study will focus on data mining techniques such as Binary Classification, Forecasting, as well as Data Visualization on a dataset collected in April 2017 by the Bureau of Transportation Statistics. The technologies used are MS Excel, SAS EM, and Tableau.

2. Key Words: Data-mining, Data visualization, Flight Data

3. Introduction:

The ability to fly was always a dream of man since early times. In Dec 1903, the Wright Brothers achieved that dream, creating the first “Sustainable, powered” flight. Looking back, the idea of aviation, specifically commercial aviation, has rapidly changed. In the past, they were used solely for war, as seen in WW1 and WW2, but that wasn’t the case years later. Airplanes and Airports were also made for the general public. Today’s Airplanes had the fine amenities most people today take for granted like pressurized cabins, comfortable seats with legroom, and on board beverages. With the rise of Big Data and analytics, the idea of commercial aviation has been revolutionized. Modern Technologies such as sensors, radars, GIS systems have been sending huge volumes of data back and forth from airplane to airport. The evolutions of machines have also allowed this data to be collected and stored in a highly accessible database in real time. In addition, through Big Data and analytics, airports and airlines can better understand the behaviors of their customers, such as our search history, purchase history, number of miles flown, etc. and change marketing tactics based on these behaviors. The hope is that by understanding these behaviors, airlines and airports can receive more revenue than they ever had without huge data analysis. This paper explores a dataset generated by the Bureau of Transportation Statistics, a total of 24 predictors and 20,199 different NJ Flights. In addition, data mining techniques such as binary classification and forecasting, as well as data visualization will be used on the data to answer the question: Can we use data visualization and data mining to make predictions and observations on customer behavior by using just flight describing data?

4. Data Description and Variable Significance:

The dataset was taken from the Bureau of Transportation Statistics and contains 24 predictors. In addition, it contains 20,199 observations. Each observation represents a single flight. The following predictors are shown below:

1. DAY- The Day of the Flight.
   Ex: “Monday”, “Tuesday”, etc.
2. FL_DATE – The Complete Date of the Flight.
   Ex: “4/1/17”
3. CARRIER_NM- Carrier Name
   Ex: “American Airlines”, “United Air”
4. TAIL_NUM- Tail Number which is used to distinguish each plane of each Carrier.
   Ex: “N653AA”
5. FL_NUM- Flight number
   Ex: 1598
6. ORIGIN_AIRPORT_NM- Originating Airport Name
7. ORIGIN_AIRPORT_LAT- Geographic latitude coordinate value of the originating airport
8. ORIGIN_AIRPORT_LONG- Geographic longitude coordinate value of the originating airport
9. ORIGIN_CITY_NM- Origin City name
10. ORIGIN_STATE_NM- Origin State Name
11. DEST_AIRPORT_NM- Destination Airport Name
12. DEST_AIRPORT_LAT- Geographic latitude coordinate of the destination airport
13. DEST_AIRPORT_LONG- Geographic longitude coordinate of destination airport

14. DEST_CITY_NM- Name of the city where the Destination Airport is located
15. CRS_DEP_TIME- Predicted Departure time
16. DEP_TIME- Actual departure time
17. CRS_ARR_TIME- Predicted Arrival time
18. ARR_TIME- Actual Arrival time
19. ARR_DEL15- Binary Variable that shows if the Arrival Delay was more than 15 minutes or not. A value of 1 indicates an arrival delay greater or equal to 15 minutes. A value of 0 indicates a arrival delay of less than 15 minutes, or an arrival delay that doesn't exist (On time flights, and Early arrival flights)
20. ARR_DELAY- The amount of arrival delay in minutes, that is, the difference of time elapsed between the predicted arrival time and the actual arrival time of the Flight
21. DEP_DELAY-The amount of Departure delay, that is, the difference of time elapsed between the predicted departure time and the actual departure time of the Flight
22. DEP_DEL15- Binary Variable that shows if the Departure Delay was more than 15 minutes or not. A value of 1 indicates a Departure delay greater or equal to 15 minutes. A value of 0 indicates a arrival delay of less than 15 minutes, or an departure delay that doesn't exist (On time flights, and Early Departure flights)
23. ARR_DEL_FLAG- Indicates the type of arrival. Ex: "DELAYED ARRIVAL, EARLY DEPARTURE, ON TIME"
24. DEP_DEL_FLAG- Indicates the type of Departure. EX: "EARLY DPEARTURE, ON TIME, DELAYED DEPARTURE."

5. Methodology:

This section will describe the steps of the binary classification, Data visualization, and forecasting techniques that were used to create the tables and graphs seen in the "Results" sections

A Binary Classification:

Overview: *Binary classification in Data Mining is the designation of data into two separate groups using a response variable, or binary variable. A binary variable is a variable that can only take two values, value A and value B. It is important to note that an observation can either have value A or value B, not both.*

In the world of data mining, there are thousands of models, and each model has its own binary classification with regards to the data.

For my binary classification results, I used a technique known as the Confusion Matrix.

A Confusion Matrix is “a table that is often used to describe the performance of a classification model (or ‘classifier’) on a set of test data for which the true values are known” (Data School, 2014)

Start by building a confusion matrix like the one below.

		Model		
		Binary Value=0	Binary Value=1	Total
Actual Data	Binary Value=0	X	Y	X+Y
	Binary Value=1	Z	W	Z+W
	Total	X+Z	Y+W	n

W represents the number of observations that the model got correct when the actual binary value was 1. In other words, when the model predicted the binary value to be 1 for an observation, the actual binary value for the observation it was trying to predict on was 1. In statistics, Y is known as number of **True Positive (TP) tests.**

X represents the number of observations that the model got correct when the actual binary value was 0. When the model predicted the binary value to be 0 for an observation, that actual binary value for that observation was 0. In statistics, X is known as number of **True Negative (TN) tests.**

Y represents the number of observations that the model got incorrect when the actual binary value was 0. When the model predicted the binary value to be 1 for an observation, that actual binary value for that observation was 0. In statistics, Y is known as number of **False Positive (FP) tests.**

Z represents the number of observations that the model got incorrect when the actual binary value was 1. When the model predicted the binary value to be 0 for an observation, that actual binary value for that observation was 1. In statistics, Z is known as number of **False Negative (FN) tests.**

From the confusion matrix, accuracy, specificity, and sensitivity can be derived. These Sensitivity and specificity are used to generate the ROC curve for the model. On the next page, the steps to perform binary classification using SAS-EM are shown.

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+FP+TN+FN)}$$

$$\text{Sensitivity}=\frac{(TP)}{(TP+FN)}$$

$$\text{Specificity} = \frac{(TN)}{(TN+FP)}$$

Binary Classification using SAS EM:

Step 1: Designating a target variable:

To perform the binary classification, a response variable must be selected. This response variable must be binary, meaning it must have only two categories.

In the data, two binary variables, DEP-DEL15, and ARR-DEL15 exist. The question becomes on what binary variable to select as the response/target variable. There is no set answer, and one has a variety of options to choose from. The first option would be to set DEP-DEL15 as a target and keeping ARR-DEL15 as an input variable. The second option would to keep ARR-DEL15 as the target, and setting DEP-DEL15 as the input. The last option would be keep only one binary variable while rejecting the other. It is important to note that two choice will most likely yield different results in the model.

For the sake of argument, DEP-DEL15 will be chosen as the target variable. Since DEP-DEL15 as our target variable, the other binary variable will be set as an input variable. The other binary variable can come in handy for models like the Decision Tree which can use that binary variable to create more rules about the data.

Step 2: Model comparison

. The challenge of performing binary classification is finding out what models the best to use. To solve this, perform a model comparison by connecting the model nodes to the model comparison node and running the diagram. The top 3 models are the ones with the lowest valid misclassification rate. Having a lower misclassification translates into higher accuracy in the model, since $Accuracy = 1 - misclassification\ rate$

Step 3: View and Interpret the Confusion Matrix of each Model

Using SAS EM click on the model nodes, and right click to review the results. Under the output, you can find the confusion matrix of each model under the "Output" windows. You can also find out the number of TP, TN, FP, FN tests. From this, derive specificity, sensitivity, and accuracy of the model. The next section deals with the Data Visualization methodology.

B. Data visualization:

Overview: *Data visualization is the ability to visualize data. It is helpful when the data is large and complex to interpret. By providing graphical images based on the data, viewers can better understand the data much clearer. Tableau is a technology that is well known for Data Visualization. Tableau was used to generate the ascetic but meaningful images in this project.*

Unfortunately, there is no set methodology to Tableau. It is primarily a drag and drop technology, meaning that by dragging the appropriate variables into the right place in a correct combination and by setting certain features, you can generate the graph you are looking for. In other words, this technology requires a bit of trial and error to get the right graph. The data fields to drag and the features to set to generate the graphs are show below.

The last methodology to be discussed in this paper is methodology regarding a technique known as forecasting.

Instructions to generate Map 6.1:

1. Drag “Origin Airport Long” to the Column plane and “Origin Airport Lat” to Rows plane. Make sure both variables are set to “Dimension” and “Continuous”. To do this, simply hover over the variable, and click the tiny upside arrow.
2. Drag “Fl Number” from “Dimensions” to the “Color” Pane. Make sure this variable is set to “Continuous” and “Measure (Count Distinct)”
3. Drag “Fl Number” again from “Dimensions” but this time drag it to the “Size” pane. Make sure this variable is set to “Discrete”, and “Measure (Count Distinct)”
4. Go to the “Color” Panel and Click “Edit Colors”. The Pallete that was used to make both graphs were “Orange- Gold”. Also, Start was set to 500 and End was set to 644.
5. Finally, make sure that your map is a “Symbol Map”. You can do this by clicking the “Show Me” and hovering the cursor to the 4th map.
6. If the graph is not the same as the one in this paper, then it is most likely “Origin Airport Long” and “Origin Airport Lat” aren’t set to Geographic. To do this, you can hover your mouse over any variable, click the upside triangle, and under “Geographic role”, Latitude and Longitude are available. Select the appropriate one for each of the two variables.

Instructions to generate Map 6.2:

The instructions to generate Graph 6.2 are the same as that of Graph 6.1 but in the first step, replace “Origin Airport Long” with “Dest Airport Long” and “Origin Airport Lat” with “Dest Airport Lat”

Instructions to generate Graph 6.3

1. Drag “Arr Time” and “Dept Time” to Columns, and make sure they are set to continuous. Also since they are time variables, select “HOUR”
2. Drag Fl Number to Rows, set it to “Count(Distinct)” and “Continuous”

3. Drag Origin Airport Name to filters, click “Edit filter” and select “Newark Liberty International”.
4. Make sure your graph is a “lines (discrete)” map. These steps if followed correctly produce this kind of map.

Instructions to generate Map 6.4

1. Drag Origin Airport Long to Columns and Origin Airport Lat to rows.
2. Drag Dep Time to filters. Make sure the “Hours” feature is selected. For this map, click the checkboxes for 12-23.
3. Make sure that your “Marks Pane” has a “Path” Panel, if not make sure “Line” is selected. You should be able to see it there in the bottom right.
4. Drag Fl num to Path and make sure it is set to attribute and continuous.
5. Also to see the carriers of each flight, drag “Carrier name” to “Colors.”
6. Lastly, to make the map dark go to “Map” and then go to “Map Layers”. Make sure “Dark” is selected under “Styles”

Instructions to generate Graph 6.5

1. Drag Carrier Name to Columns and Fl Num to Rows (set Fl num to Count(Distinct))
2. Drag “Origin Airport to filters” and filter it to “Newark Liberty Int.”
3. Make sure this map is selected as “Horizontal bars”

Instructions to generate Graph 6.6:

1. Drag Carrier Name to columns, and drag CNTD (Fl Num) to rows.
2. In the list of maps to the right, select “Packed Bubbles”

C. Forecasting:

Overview: Forecasting is the ability to predict future values given a current set of values.

There are many forecast methods, but I used the Exponential Smoothing forecast model for my methodology. The Exponential Smoothing forecast offered by both SAS as well as Excel takes care of seasonality, which are constant fluctuations in the data.

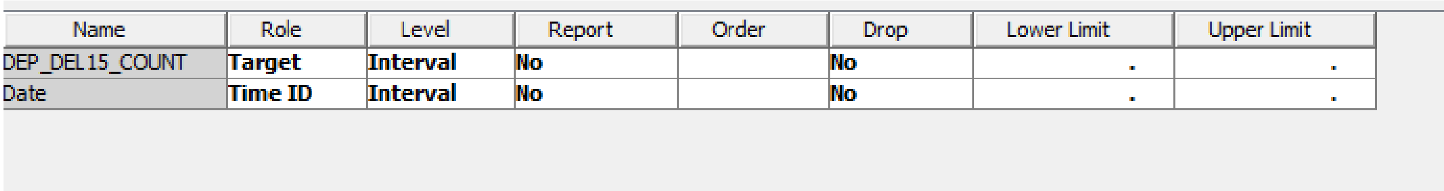
How to perform Exponential Smoothing using Excel

To perform the forecast, you will need time series data. Time series data has two columns. One column is for Time related values like hours, days, etc., the other column relates to the values of the variable that is to be observed. For this dataset, the variable "Fl_Date" is the time variable, and the variable to be observed is a variable that will contain the total count of "DEP_DEL15 = 1" per day. To extract the time series data quickly, you use what is known as a Pivot Table on the original data. Pivot Tables are useful in that they help to quickly analyses the data, or rather manipulate the data by drag and drop. In addition, you can get the SUM or the COUNT of a particular variable from the Pivot Table's field list. Select the Pivot Table, go to Data, and click "Forecast Sheet"

Reference to Pivot Table: https://bit.ly/2FrQJxr
Reference to Excel Forecasting:
https://support.office.com/en-us/article/create-a-forecast-in-excel-2016-for-windows-22c500da-6da7-45e5-bfdc-60a7062329fd

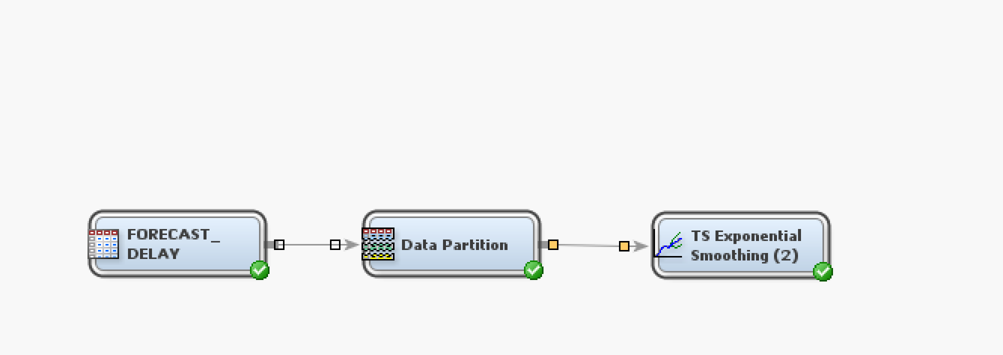
How to perform Exponential Smoothing Forecast using SAS-EM

1. Set your observed variable to "Target", and your time variable to "Time ID"



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DEP_DEL15_COUNT	Target	Interval	No		No	.	.
Date	Time ID	Interval	No		No	.	.

2. Build a flow like this one and run the Exponential Smoothing node.



6. Main results:

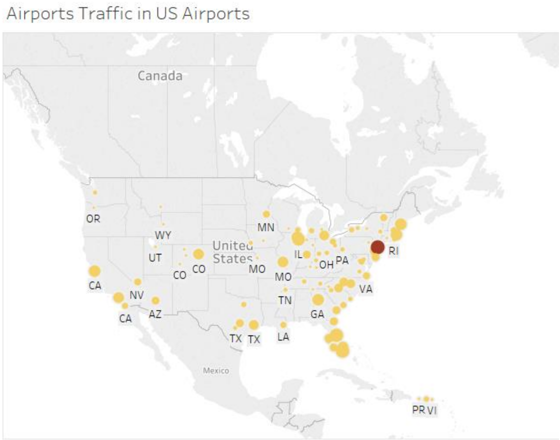
A. Data Visualization:

In this section, we will analyze several graphs generated by the software Tableau. Each graph or each set of graphs will reveal insights about the data that would never been derived from looking at the raw dataset.

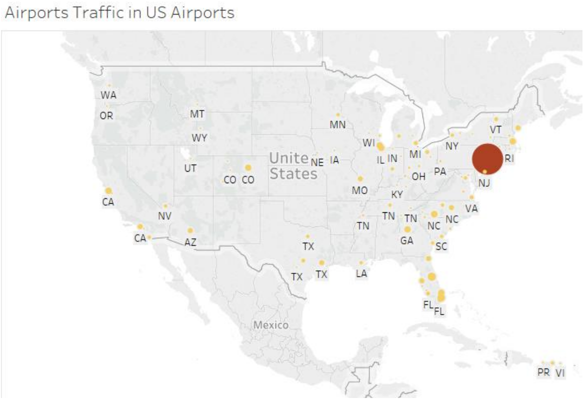
“Which Airport do most customers go to for NJ Flights?”

The first two graphs we will analyze are geographic maps. Both graphs describe the Airline Traffic on each airport. Airline Traffic in this paper is the number of planes arriving and departing from each airport. Map 6.1 describe the number of planes departing from each airport while Map 6.2 describes the number of planes arriving at each airport. Both quantities are represented by dots on the maps, for each airport. The color and the Size of each dot both serve to indicate how large of airline traffic each airport has. A large red circle, or a red circle in general indicates high airline traffic. In both graphs, there is a red dot, and that red dot belongs to Newark Airport (EWR). **From these two graphs, Newark Airport**

is inferred to be the most popular airport, possessing the greatest Flight traffic out of all the other airports.



Map 6.1



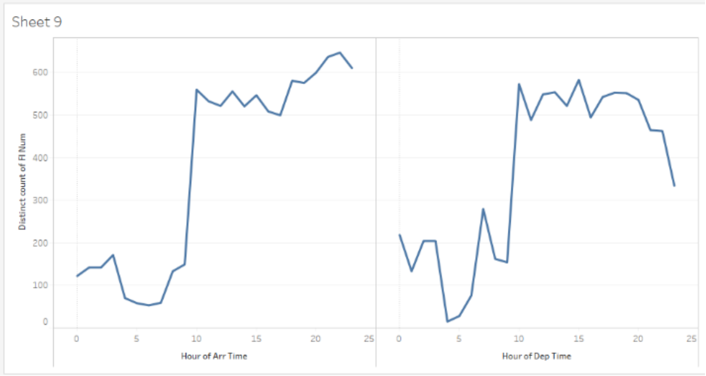
Map 6.2

It should not come off as a surprise that Newark Int. was ranked number #1 in terms of NJ Flight Traffic. Since the dataset was made up of NJ Flights, the flights had to arrive and depart from a NJ airport at some point. But why Newark? Why wasn't it some other NJ airport? According to an article written by BuisnessWire in May 2017, Newark Int. was one of the busiest airports stating: "Airports that leverage Everbridge for critical event management include the world's busiest....**Newark Liberty International Airport** (EWR) and LaGuardia Airport (LGA)" (BuisnessWire, 2017). In addition, an airport traffic report published by Port Authority of NJ and NY, Newark is ranked 14th out of the 114 US airports and 43rd out of the world's airports for hosting the most passengers. Moreover, Newark appears to have experienced some major renovations in 2017, hosting new restaurants such as "Wanderlust, a burger bar with a long list of gourmet burgers and beer offerings, and Saison, a French bistro" (Spellman, 2017). These new eateries could have been an incentive for more fliers to come to Newark, thereby increasing the number of flights at that airport.

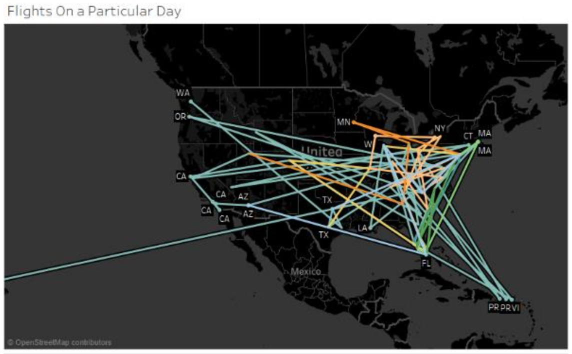
"Do NJ Customers prefer Morning or Evening Flights?

From the previous graphs, Newark International Airport (EWR) was shown to be the most popular airport in terms of departing and arriving flights. Graph 6.3 reveals what time is most common for the NJ Flights at Newark Airport. Specifically, the graph shows the number of flights departing and arriving over a

24-hour interval at EWR. After 12 PM, there is a gigantic spike in the number of flight that depart and arrive from and at EWR. The peaks of both graphs occur in the time interval from 12 PM-12 AM. On the other hand, Map 6.4 shows the geographic mapping of various flights on the US. This map represents a night map, which shows the flights during evening departure hours (12 PM-12 AM). From this graph you can see a lot of flights appearing, especially in the Northeast region of the country. **These two graphs imply that evening flights are more frequent than morning flights.**



Graph 6.3

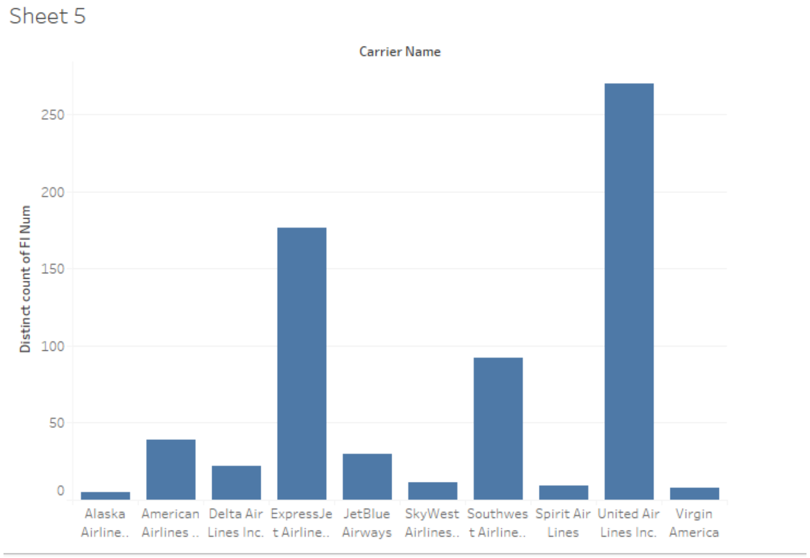


Map 6.4

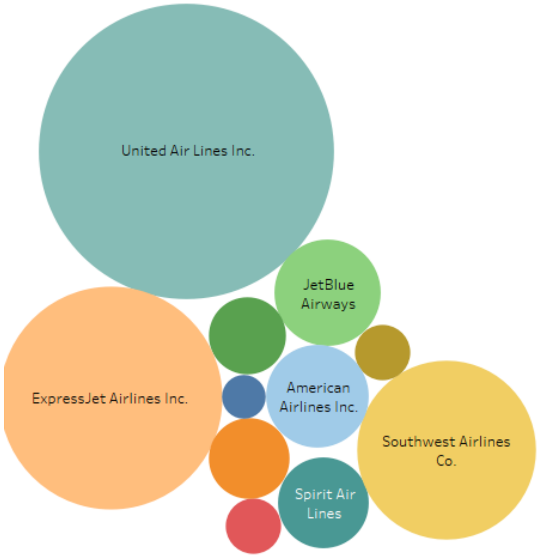
Red eye flights are flights that depart late at night and arrive early morning. According to *Red-Eye Flight Survival Guide*, Jessica Padykula states, “Often, red-eye flights are cheaper than flights at other times. For people who have no problem sleeping during a flight, this can be a good, cheap option. Red-eye flights also mean you aren’t paying for accommodation for the night you’re in the air, which can be a money-saving strategy. Red-eye flights also appeal to business travelers who want to avoid losing a work day in transit.” (Padykula, 2017). Perhaps the reason evening flights are more popular than day flights because they are cheaper, and because passengers reap the benefits such as sleep, and save work productivity. Padykula also states, “As far as the experience, red-eye flights are generally less crowded than regular morning or midday flights tend to be, so you might luck out with an empty seat next to you and have more space to spread out (making it easier to sleep).” (Padykula, 2017)

“Which carriers are most favored by Customers when they take evening flights?”

While it was discovered in the previous maps that evening flights are more prevalent than morning flights, what carrier is responsible for most of these evening flights? Graph 6.5 shows the all the airline carriers categorized in terms of the number of night flights that the carriers hosted. The originating airport of these flights was EWR. Analyzing these graph, two carriers are shown to be the highest number of night flights. The first being United Airlines and the second being Express Jet Airlines. This observation is indeed conspicuous in Graph 6.6. This graph shows the number of evening flights for each carrier in the form of bubbles. A bigger bubble for a carrier shows that the carrier hosts a large number of evening flights. United Airlines and Express Jet both have the largest bubbles out of all the bubbles in the graph. **Hence, it can be inferred that United Airlines and ExpressJet are the most popular airlines for Evening NJ flights.**



Graph 6.5



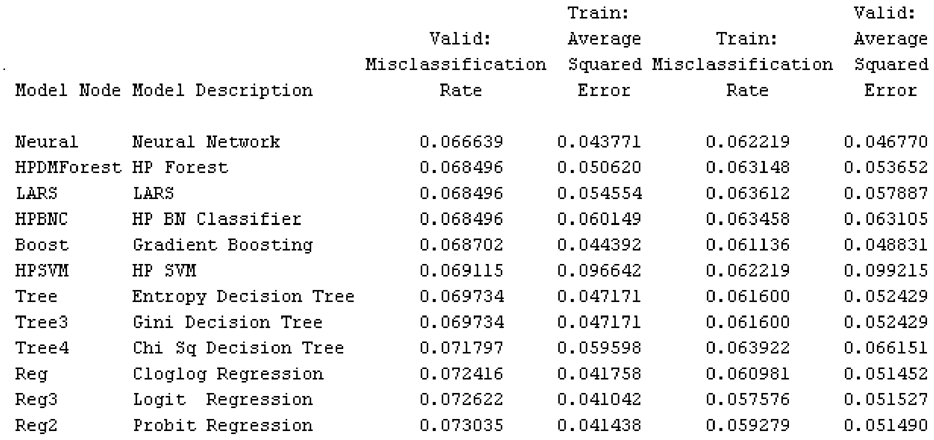
Graph 6.6

Indeed, United and Express are both desired flights sought out by customers. In an article titled, *The Best and Worst Airlines in 2017*. Analyst Julian Kheel ranks United at #2 out of the 10 most popular domestic flights he evaluated. Kheel states the following about United, “But what about some of these other results? United coming in second may surprise some people, especially since the carrier didn’t win outright in any particular area. However, it performed solidly in a number of our criteria, including route network, cabin comfort, baggage and change fees and lounges...” (Kheel, 2017) In addition, according to Benjamin Zhang of Business Insider, he ranked ExpressJet #10 in his Article, *Here are the 12 best airlines in America*.

B. Data Mining:

In this section, we will discuss the results of three models with regards to binary classification, a data mining technique, on the data. Our Binary Variable is **DEP_DEL15**. Recall that this variable indicates if a Flight’s Departure Delay is 15 minutes or more. If the value of this variable is 1, then the Delay is at least 15 minutes. If the value is 0, the Delay is less than 15 minutes or nonexistent.

When the model comparison was run, a total of 12 models tested: Gradient Boosting, Decision Trees (Entropy, Gini, and ChiSq), HP Forest, LARS, HP SVM, Neural Network and Regression (Clog, Probit, and Logit), HP BN Classifier**.** Comparing all the models, **Neural Network, HP Forest, and LARS** are the top 3 models have the lowest misclassification rates. While it is true that the Random Forest, LARS, and HP BN Classifier have the same misclassification rate, HP Forest and LARS have lower Valid Average Square Errors than the BN model.



```
                                                Train:                    Valid:
                                   Valid:       Average      Train:       Average
                              Misclassification Squared Misclassification Squared
Model Node Model Description        Rate         Error         Rate        Error

Neural     Neural Network         0.066639     0.043771      0.062219     0.046770
HPDMForest HP Forest              0.068496     0.050620      0.063148     0.053652
LARS       LARS                   0.068496     0.054554      0.063612     0.057887
HPBNC      HP BN Classifier       0.068496     0.060149      0.063458     0.063105
Boost      Gradient Boosting      0.068702     0.044392      0.061136     0.048831
HPSVM      HP SVM                 0.069115     0.096642      0.062219     0.099215
Tree       Entropy Decision Tree  0.069734     0.047171      0.061600     0.052429
Tree3      Gini Decision Tree     0.069734     0.047171      0.061600     0.052429
Tree4      Chi Sq Decision Tree   0.071797     0.059598      0.063922     0.066151
Reg        Cloglog Regression     0.072416     0.041758      0.060981     0.051452
Reg3       Logit  Regression      0.072622     0.041042      0.057576     0.051527
Reg2       Probit Regression      0.073035     0.041438      0.059279     0.051490
```

Neural Network (NN):

The following table gives a distribution of the True Positives, True Negative, False Positives and False Negative tests when the Neural Network model was classifying the data based on the binary variable. Most of the tests (approximately 70%) are True Negative, where the Outcome and Target are both 0, and True Positive (approximately 23%) where the Outcome and Target are both 1. This table indicates that the Neural Network is a sufficient model for binary classification as the prediction of the model matches the actual observation value about 93% of the time. In other words, the accuracy of the model is about 93%. There is about 7% error in the NN's ability to classify the binary variable. Also, below the table, indicates the specificity and sensitivity. Readers should be familiar with these two terms.

Test	Target	Outcome	Total Percentage *	Count
True Negative	0	0	70.15	3400
False Negative	1	0	2.97	144
False Positive	0	1	3.69	179
True Positive	1	1	23.19	1124
Total			100	4847

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} = \frac{(1124+3400)}{(3400+144+179+1124)} = 0.9333$$

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)} = \frac{1124}{1124+144} = 0.8864$$

$$\text{Specificity} = \frac{(TN)}{(TN+FP)} = \frac{3400}{3400+179} = 0.949986$$

*Total Percentage is the % of tests out of the total number of classification tests in the data. *

Because the accuracy rate is about 93%, it makes one wonder about how the NN is making the binary classifications. Running a decision tree on the Neural Network Mode reveals 4 decision rules.

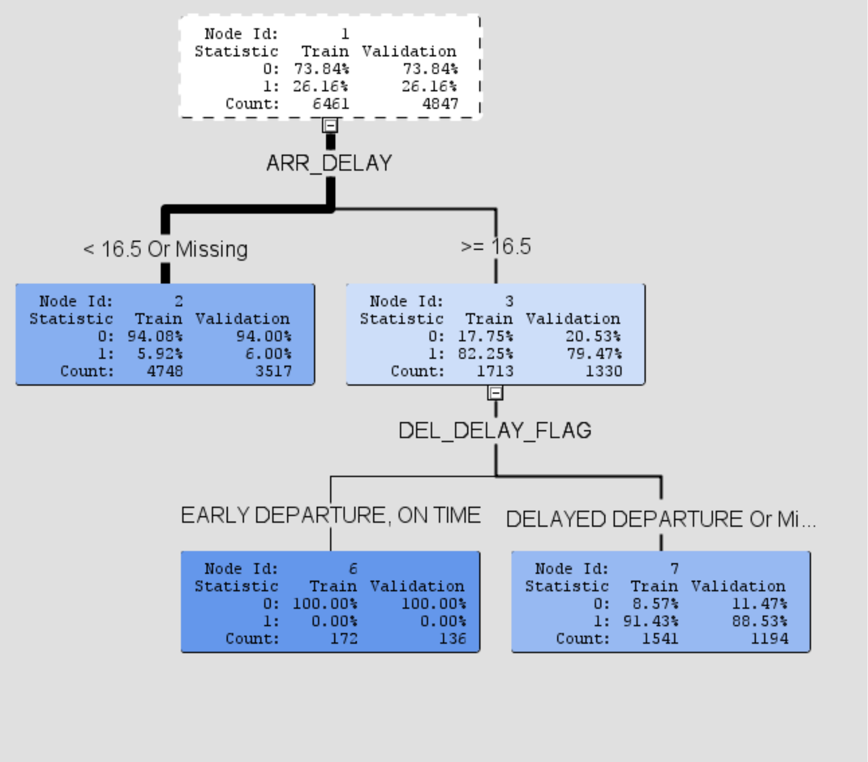
1. ARR_Delay < 16.5 or Missing. If a flight had an arrival delay of less than 16.5 minutes, it was classified into Node number #2. 3517 flights with arrival delays less than 16.5 minutes were put into this node, with 6% having a departure delay of 15 minutes or greater, while the remaining 94% flights did not have this type of delay. They were either flights early departure flights, on time flights, or flights that had a delay less than 15 minutes. **According to the NN model, if a flight has an arrival delay less than 16.5 min, it will most likely had a departure delay of less than 15 minutes or not even a delay at all.**

2. ARR_DELAY ≥ 16.5. Conversely, if a flight had an arrival delay that was 16.5 minutes or greater, it was classified into node #3. 1330 flights were put into this node, with 20.53% having a delay less than 15 minutes, while 79.47% of the 1330 flights had a delay greater than 15 minutes. **According to NN, if a flight has arrival delay of greater than 15 minutes, it most likely had a departure delay of at least 15 minutes.**

3. DEL_DEL_FLAG= EARLY DEPARTURE, ON TIME- Node #3 split into two other nodes. Node # 6 has 136 flights, and all of these flights have departure Delays less than 15 minutes**. However, these flights were classified as Early Departure and On Time flights as the label implies**. It's most probable that these flights most likely had negative departure delay values or departure delays of 0. *

4. DEL_DEL_FLAG= DELAYED DEPARTURE- Node # 7 is the child node of Node 3, and has 1194 flights. Out of the 1194, 11.47% had departure delays less than 15 min, and 88.53% had departure delays greater than 15 min. **As the label implies, these flights are delayed departure flights.**

* Recall that it is possible to have negative departure delay values. Departure delay is defined as $Departure\ Delay\ =\ Actual\ Departure\ Time -$ $Predicted\ Departure$ time in terms of minutes. When Actual Time < Predicted Time, then Departure Delay < 0.



HP Forest:

Like the Neural Network, most of the tests when using the binary variable for classification are True Positive and True Negative. However, when compared to the NN, the Accuracy of the HP Forest model is slightly lower, and Sensitivity (True Positive Rate) is lower. However, Specificity (True Negative Rate) is higher in the HP Forest Model than the NN Model.

Test	Target	Outcome	Total Percentage	Count
True Negative	0	0	70.7448	3429
False Negative	1	0	3.7549	182
False Positive	0	1	3.0947	150
True Positive	1	1	22.4056	1086
Total			100%	4847

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} = \frac{(1086+3429)}{(4847)} = 0.9315$$

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)} = \frac{1086}{1086+182} = 0.8565$$

$$\text{Specificity} = \frac{(TN)}{(TN+FP)} = \frac{3429}{3429+150} = 0.9581$$

LARS (Least Angular Regression):

Below is a binary classification table for the LARS model. Like the previous two models, a respectable number of the classification tests came out True Negative or True Positive. Comparing LARS with Random Forest, Accuracy of LARS is the same. The sensitivity of LARS surpasses that of the HP Forest but performs poorly in terms of specificity.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} = \frac{(1088+3427)}{(1088+3427+180+152)} = 0.9315$$

$$\text{Sensitivity}= \frac{(TP)}{(TP+FN)} = \frac{1088}{1088+180} = 0.8580$$

$$\text{Specificity} = \frac{(TN)}{(TN+FP)} = \frac{3427}{3427+152} = 0.9575$$

Test	Target	Outcome	Total Percentage	Count
True Negative	0	0	70.70	3427
False Negative	1	0	3.71	180
False Positive	0	1	3.14	152
True Positive	1	1	22.45	1088
Total			100	4847

Binary Classification Summary:

The table below is a summary of the binary classification of the three models explored and the performance of each one, with regards to Accuracy, Sensitivity, and Specificity. All the models have gotten most of the binary classifications correct, but the NN model received the most correct classifications, and has the highest accuracy than the other two models. Sensitivity is known as the true Positive Rate, while Specificity is known as the true Negative Rate. Recall that the true Positive rate is the proportion of true positive tests out of the total number of positive tests. The true Negative rate is the proportion of true negative tests out of the total number of negative tests. In the context of the binary variable DEP_DEL15, sensitivity refers to the model's ability to correctly identify flight that have 15 minute departure delays or more .On the other hand, specificity refers to the model's ability to correctly detect flights that don't have the 15 minute delay, or a delay greater than 15 minutes. All the three models have high sensitivities and high specificities (between 80-100%). However, NN while highest accuracy and sensitivity, fares poorly in specificity when compared to the Random Forest, and LARS models.

	Count			Rate		
Model	number of Correct Classifications (TP+TN)	number of incorrect Classifications (FP+FN)	Classification Count	Accuracy	Sensitivity	Specificity
NN	4524	323	4847	0.9333	0.8864	0.9500
HP Forest	4515	332	4847	0.9315	0.8565	0.9581
LARS	4515	332	4847	0.9315	0.8580	0.9575

Variable Importance:

While the models are impressive with regards to classifying the binary variable, (DEP_DEL15) it is important to view the top 5 predictors that were highly responsible in causing the models to make above average classifications.

Note that for these models, relative importance was used instead of variable importance.

	Variable Importance				
	Rank				
Model	1	2	3	4	5
NN	ARR_DELAY	DEL_DELAY_FLAG	DEST_AIRPORT	ORIGIN_AIRPORT_NM	DAY
HP Forest	ARR_ DELAY	DEL_DELAY_FLAG	DEST_AIRPORT	ORIG_AIRPORT_NM	DAY
LARS	ARR_DEL15	DEL_DELAY_FLAG	ARR_DELAY	DAY	ARR_DELAY_FLAG

Scoring

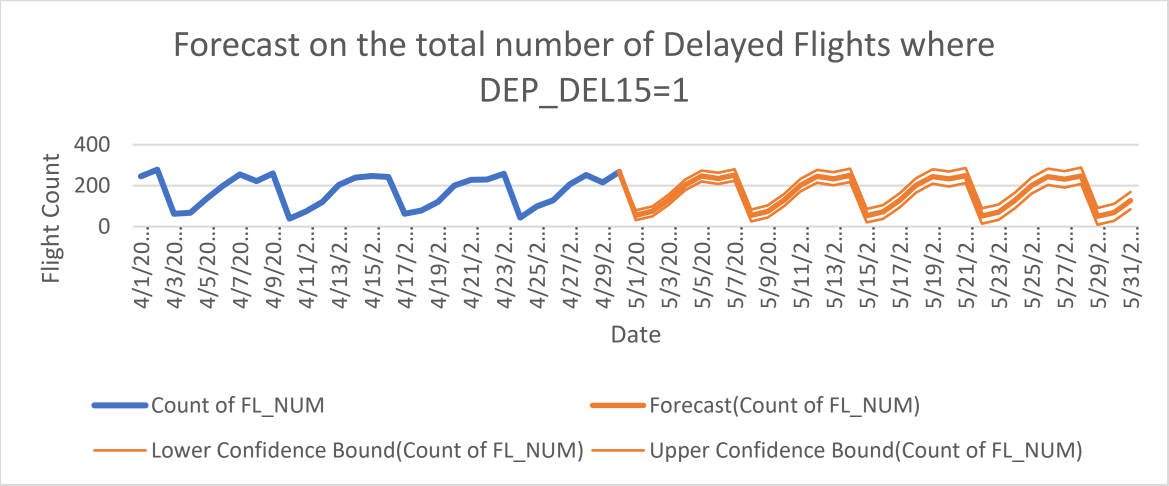
Below is a graph that shows the scoring of the observations (flights) according to each of the 3 of the models. The scoring table indicates how likely each flight was to experience a flight departure delay of 15 minutes or more. Interestingly enough each model gives a different observation (flight) for each score. For instance, for the most likely observation to experience the 15 departure delay, NN ranked Obs 17,868, Random Forest ranked Obs 16077, and LARS ranked Obs 17868. The likelihood percentage (parenthesis) below each observation is displayed in this table,

	Score according to likelihood of experiencing 15 minute Delay Departure				
	1	2	3	4	5
Neural Network	Obs 17868 (0.99776)	Obs 17925 (0.99774)	Obs 10843 (0.99774)	Obs 19576 (0.99774)	Obs 8691 (0.99774
HP Forest	Obs 16077 (0.87265)	Obs 2636 (0.87257)	Obs 8389 (0.87257)	Obs 8412 (0.87257)	Obs 14230 (0.87257)
LARS	Obs 17868 (1)	Obs 8691 (1)	Obs 10843 (1)	Obs 5032 (1)	Obs 17925 (0.99)

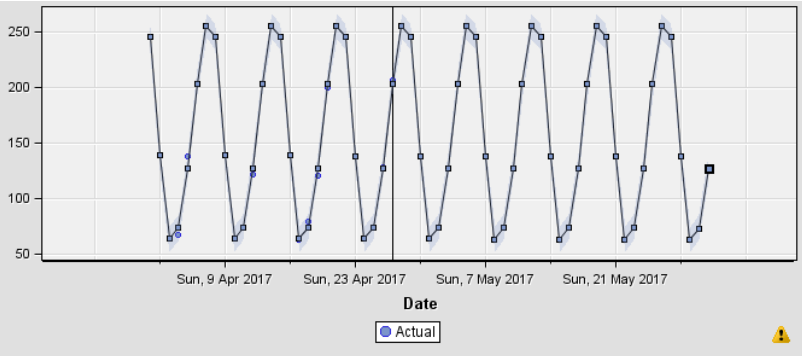
C. Forecasting: SAS vs Excel

This section is discusses the concept of forecasting.

The exponential smoothing forecast has been performed on the data using MS Excel and SAS EM. The first forecast is a forecast on that shows the number of Delayed Flights with 15 minute Departure delays or more. According to the Excel Forecast, at the end of May, the graph ends just above the 100 mark. In fact, the predicted number of such flights is 126 flights. The 95% CI is (84,168).



The forecast generated by SAS EM produces a slightly different graph than the one generated by Excel. The seasonality exists in both graphs as evidenced by the repeating peaks in the graph. According to this graph, approximately 127 flights will have Departure Delays of 15 or more. (115,138) is the Confident Interval for this value.



7. Concluding Remarks:

From this paper, it is clear to see that data visualization and data mining are powerful tools for predicting and observing customer behavior. With data visualization, airports and airlines can see trends in their data clearly. They can use these trends to change their marketing efforts to entice their customers. With regards to binary classification, instead of a person counting the number of delayed and non-delayed flights an accurate predictive model like the Random Forest, Neural Network or LARS can classify the flights (delayed or not), and inform airlines and airports on the performance of the aircrafts. Such performance is crucial in securing and gaining customers. Lastly, forecasting can be used for predicting values that can give connection to customer behavior. In the previous section, a forecast model was generated with regards to the number of delayed flights with 15 minute departure delays or more. Despite what one may think, 127 delayed flights is a significant number and airports might have to expend their resources on improving and optimizing the performance of the aircrafts if they don't want to lose their main source of revenue. The world of predictive and prescriptive analytics is the world of today, and to be frank, I don't see airports and airlines giving up the pursuit of understanding their customers any time soon.

8. Work Cited

Buisness Wire. "100% Of the 25 Busiest Airports in North America Now Utilize Everbridge for Critical Event Management." *William Lyon Homes Announces Agreement to Acquire RSI Communities, a Southern California and Texas Based Homebuilder | Business Wire*, 27 July 2017, www.businesswire.com/news/home/20170727005289/en/100-25-Busiest-Airports-North-America-Utilize.

Zhang, Benjamin. "Here Are the 12 Best Airlines in America." Business Insider, Business Insider, 12 Apr. 2017, www.businessinsider.com/best-airlines-america-2017-4

Markham, Kevin. "Simple Guide to Confusion Matrix Terminology." Data School, Data School, 8 June 2016, www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

Kheel, Julian Mark. "The Best and Worst US Airlines in 2017." The Points Guy, The Points Guy, 27 Feb. 2017, thepointsguy.com/guide/best-and-worst-airlines-2017/.

Padykula, Jessica. “Red Eye Flights - What They Are, How to Book & Tips to Survive a Red Eye Flight.” Cheapflights, 25 Jan. 2017, www.cheapflights.com/news/red-eye-flights.

Spellman, Spencer. “Newark Airport: A Complete Guide to EWR.” TripIt Blog, 28 Mar. 2017, www.tripit.com/blog/2017/03/newark-airport-a-complete-guide-to-ewr.html.

Bureau of Transportation Statistics,” On-Time : On-Time Performance”, Feb 2018, https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

Port Authority of NY & NJ, “Airport Traffic Report”, 2017, https://www.panynj.gov/airports/pdf-traffic/ATR2017.pdf

9. Appendix:

A. How to do Data Modification (MS Excel) – VLOOKUP Function, Conditional Formatting, If statements.

Sometimes the dataset you download isn’t adequate or you want to understand the raw dataset more. Below are references that you could use on your dataset to understand it more, granted your dataset is in Excel Format.

Vlookup: https://support.office.com/en-us/article/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1
Conditional Formatting: https://support.office.com/en-us/article/use-formulas-with-conditional-formatting-fed60dfa-1d3f-4e13-9ecb-f1951ff89d7f
IF Statements: https://www.pcworld.com/article/2971613/software-productivity/excel-logical-formulas-5-simple-if-statements-to-get-started.html

Code for Scoring:

```
Data goodapps;
  Set &EM_Import_Score curobs = n; obsnum=_n_; /*curobs: current
observation */
  If P_DEP_DEL151<= 0.75 Then delete;
Run;
Proc Sort data= goodapps;
```

```
  By descending P_DEP_DEL151;
Run;
Proc Print data= goodapps noobs split='*';
   Var obsnum P_DEP_DEL151;
   Label P_DEP_DEL151='Likelihood of 15-min Delay';
   Title "'Likelihood of 15-min Delay";
Run;
```