Evaluating the dataset

I used carrier_ct as the prediction factor. I created a new variable to evaluate if there was a carrier delay (1) or not (0) based on the values of carriert_ct. If carrier_cnt was not null and a valid value greater than 0, the carrier delay variable was set to 1 else 0. The numeric variables of interest were arr_del15 (delay over 15 minutes as reported to RITA), weather_ct (delays caused by extreme weather conditions), nas_ct (delays caused by NAS directives), security_ct (delays caused by security issues), late_aircraft_ct (delays caused by the airline carrier)
A simple decision tree classifier is used to train and test the data. Based on the output below we can say the most important feature to determine flight delays is the variable arr_del15. Another variable tht affects flight delay are nas_ct and late_aircraft_ct. Weather and Security do not play a significant part in the delay times. Thus we can conclude that our analysis is on the right track to determine trends with airports and airlines to determine the pain points for a particular carrier. At some airports our selected carrier experienced considerable delays. We will focus our visulaizations on this aspect. If we can pinpoint the delay times and location, we might be able to do further research to eliminate the delays for the airline.

```
training time for all data: 0.425 s
Decision Tree Accuracy on All the data:  1.0
training time: 0.184 s
prediction time: 0.008 s
no. postive predictions:  82060
F1 Score:  0.996
Precision score:  0.996
Recall score:  0.996
Decision Tree Classifier Accuracy:  0.992
output: [0, 1, 2, 3, 4]
importance: of  arr_del15  is  0.716
importance: of   weather_ct  is  0.053
importance: of  nas_ct  is  0.12
importance: of  security_ct  is  0.004
importance: of  late_aircraft_ct  is  0.107
```

PCA Fit

Using the PCA Fit we come to the similar conclusion. Weather and Security are not the leading causes for flght delays. The features we have used are the same as above. The

numeric variables of interest were arr_del15 (delay over 15 minutes as reported to RITA), weather_ct (delays caused by extreme weather conditions), nas_ct (delays caused by NAS directives), security_ct (delays caused by security issues), late_aircraft_ct (delays caused by the airline carrier)
PCA Fit confirms that the Decision Tree Classifier is the best fit for this data set.

```
Prediction accuracy for the normal test dataset with PCA using the Decision Tree
Classifier
99.01%


Prediction accuracy for the standardized test dataset with PCA using the
Gaussian Naive Bayes Classifier
66.76%


PC 1 without scaling:
 [0.86776321 0.20001608 0.02510475 0.30166796 0.00134976 0.33962471]

PC 1 with scaling:
 [0.45851309 0.44154183 0.39242163 0.42110894 0.26603731 0.43897976]
Carrier Delay:  [0.93628354 0.91977291 0.92658247 ... 0.92730209 1.02254084
0.92235982]
slope: [-0.04989449  0.05111411  0.0497768   0.04979733  0.04961322  0.04960349]
intercept:  0.9199668519849167

##### stats on test dataset #######

r-squared score: 0.015727406648427977

##### stats on training dataset #######

r-squared score: 0.015813212459995896
```