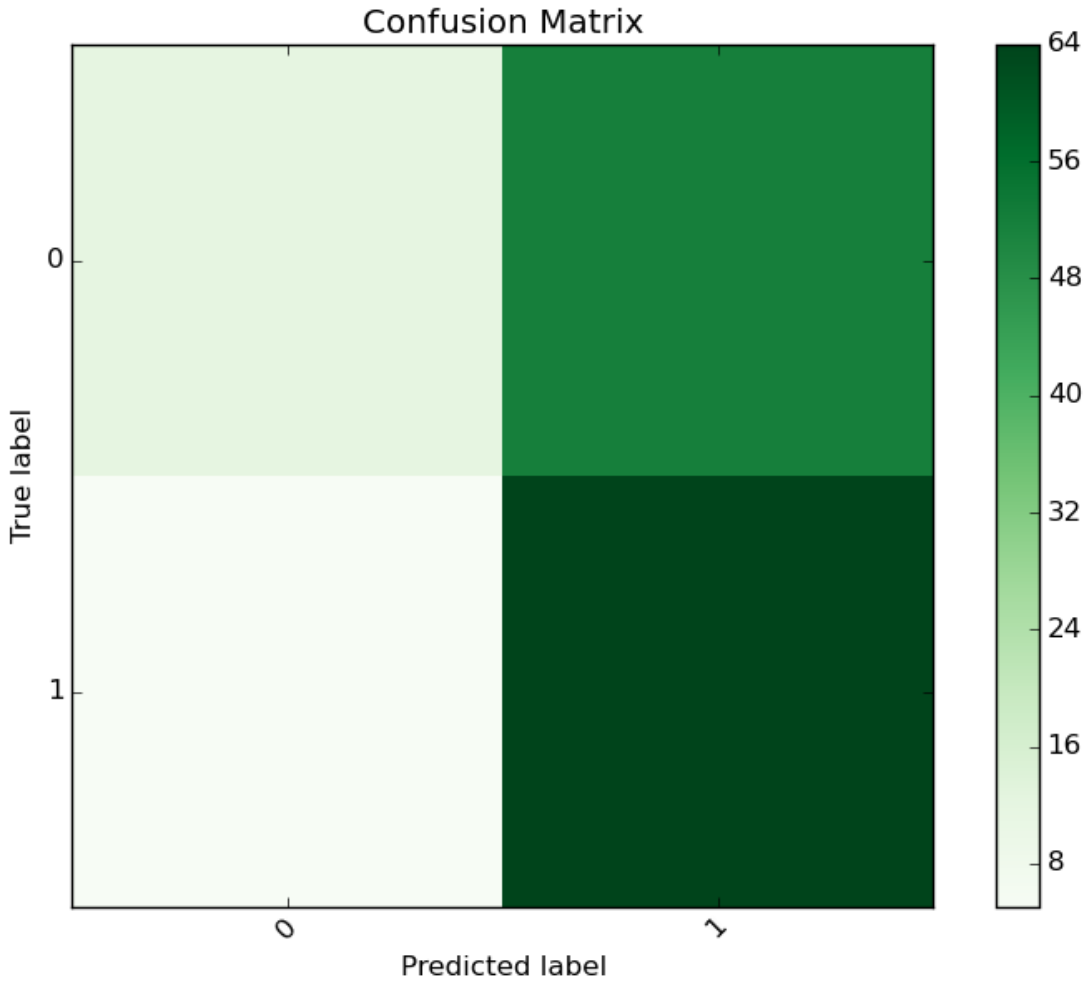
PROJECT RECAP

This report summarizes the results and findings on the project from June 20th to August 20th of 2016. Logistic regression and Decision Tree was performed on the dataset provided. The dataset was split into 85% training and 15% cross validation set. Following were the results for the company named SPY.

Logistic Regression:

Class	Precision	Recall	F1-Score	Support
0	0.71	0.19	0.30	64
1	0.55	0.93	0.69	69
Avg/Total	0.63	0.57	0.50	133

The confusion matrix generated is as shown below



The numerical results of the above figure:

	Class 0	**Class 1**
Class 0	12	52
Class 1	5	64

The above table tells that, when logistic regression was tested on our data set we found that the algorithm was able to predict 64 correctly of 69 examples for class ‘1’ and 12 examples of 64 examples of class ‘0’.

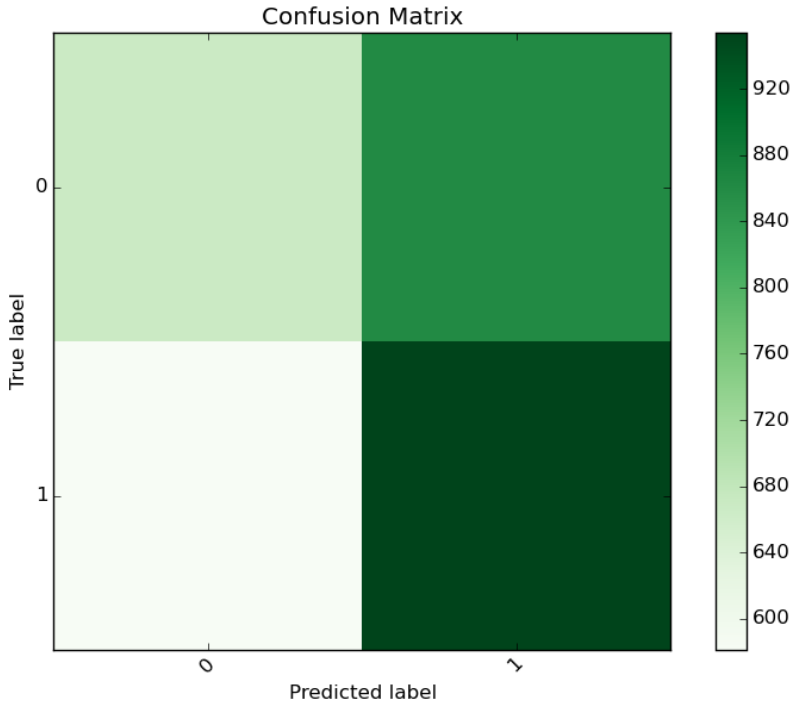
The overall accuracy of algorithm was calculated to be around 58%.

The script for the logistic regression is ‘psych_sig_log_reg.py’ and the cleaned data used for machine learning algorithm is found in SASAndPrice.csv

Following were the results for all the companies when logistic regression was done on the data set that was split as 75% training set and 25 % testing data set.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.53	0.44	0.48	1528
1	0.53	0.62	0.57	1535
Avg/Total	0.53	0.53	0.53	3063

The confusion matrix generated is as shown below:



The numerical results of the above figure:

	Class 0	**Class 1**
Class 0	667	861
Class 1	581	954

The above table tells that, when logistic regression was tested on our data set we found that the algorithm was able to predict 954 correctly of 1535 examples for class ‘1’ and 667 examples of 1528 examples of class ‘0’.

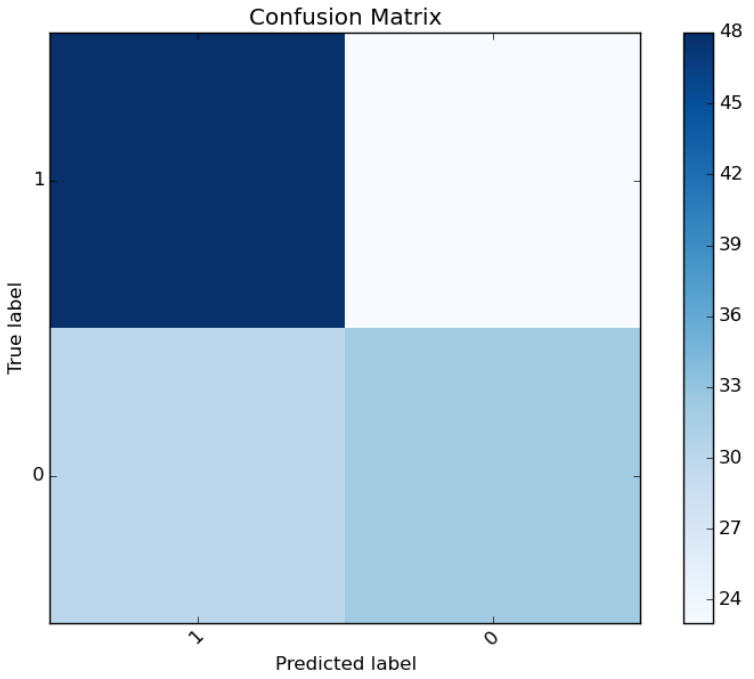
The overall accuracy of algorithm was calculated to be around 53%.

The script for the logistic regression is ‘psych_sig_log_reg_3.py’ and the cleaned data used for machine learning algorithm is found in AllSASAndPrice.csv

Decision Tree:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.58	0.52	0.55	62
1	0.62	0.68	0.64	71
Avg/Total	0.60	0.60	0.60	133

The confusion matrix generated is as shown below



The numerical results of the above figure:

	Class 1	**Class 0**
Class 1	48	23
Class 0	30	32

The above table tells that, when decision tree was tested on our data set we found that the algorithm was able to predict 48 correctly of 71 examples for class ‘1’ and 30 examples of 62 examples of class ‘0’.

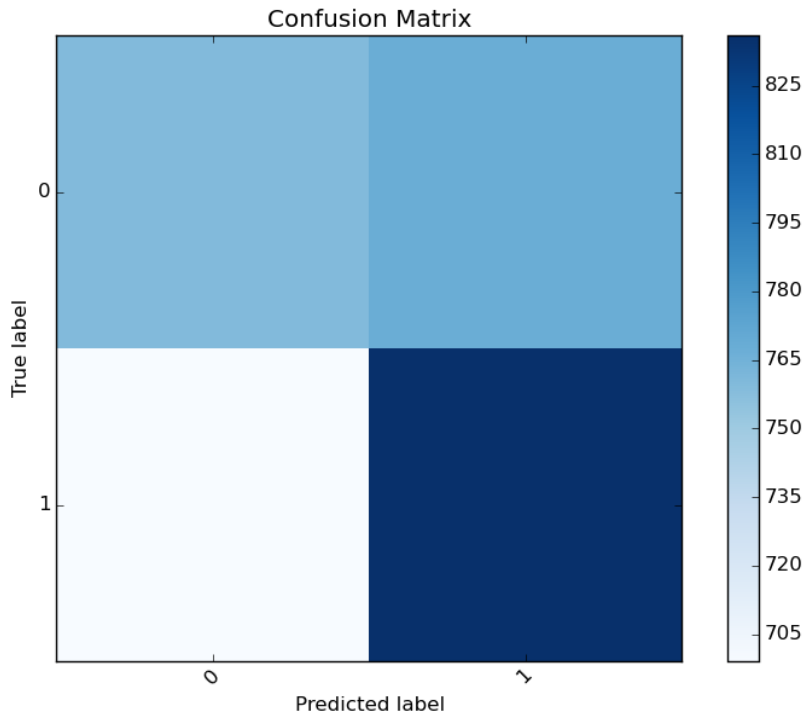
The overall accuracy of algorithm was calculated to be around 60%.

The script for the decision tree is ‘psych_sig_dec_tree.py’ and the cleaned data used for machine learning algorithm is found in SASAndPrice.csv. The corresponding constructed decision tree has been drawn on the image dt.png and the rules have been written in dt.dot files available in the folder.

Following were the results for all the companies when decision tree was done on the data set that was split as 75% training set and 25 % testing data set.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.52	0.50	0.51	1528
1	0.52	0.54	0.53	1535
Avg/Total	0.52	0.52	0.52	3063

The confusion matrix generated is as shown below:



The numerical results of the above figure:

	Class 0	**Class 1**
Class 0	760	768
Class 1	699	836

The above table tells that, when decision tree was tested on our data set we found that the algorithm was able to predict 836 correctly of 1535 examples for class ‘1’ and 760 examples of 1528 examples of class ‘0’.

The overall accuracy of algorithm was calculated to be around 52%.

The script for the logistic regression is ‘psych_sig_dec_tree_2.py’ and the cleaned data used for machine learning algorithm is found in AllSASAndPrice.csv. The corresponding constructed decision tree has been drawn on the image dt_complete.png and the rules have been written in dt_complete.dot files available in the folder.

[PS: The png image needs to be zoomed in for a better view]