Santander Customer Satisfaction – Classification problem

Author: Anirban Roy, April 2016

Problem Statement:

Which customers are happy customers?
More details at Santander Customer Satisfaction Challenge

Environment Setup:

The project is built using Python. To setup the environment please install the below dependencies:

1. Anaconda (Scikit learn, pandas, scipy, numpy) - Download and install
2. PyBrain - Download and install
3. Python 2.7 (not python 3) - Download and install This comes with Anaconda bundle and maybe skipped. In case there difficulties, install separately and make sure the environment paths are configured as per the installation guidelines.

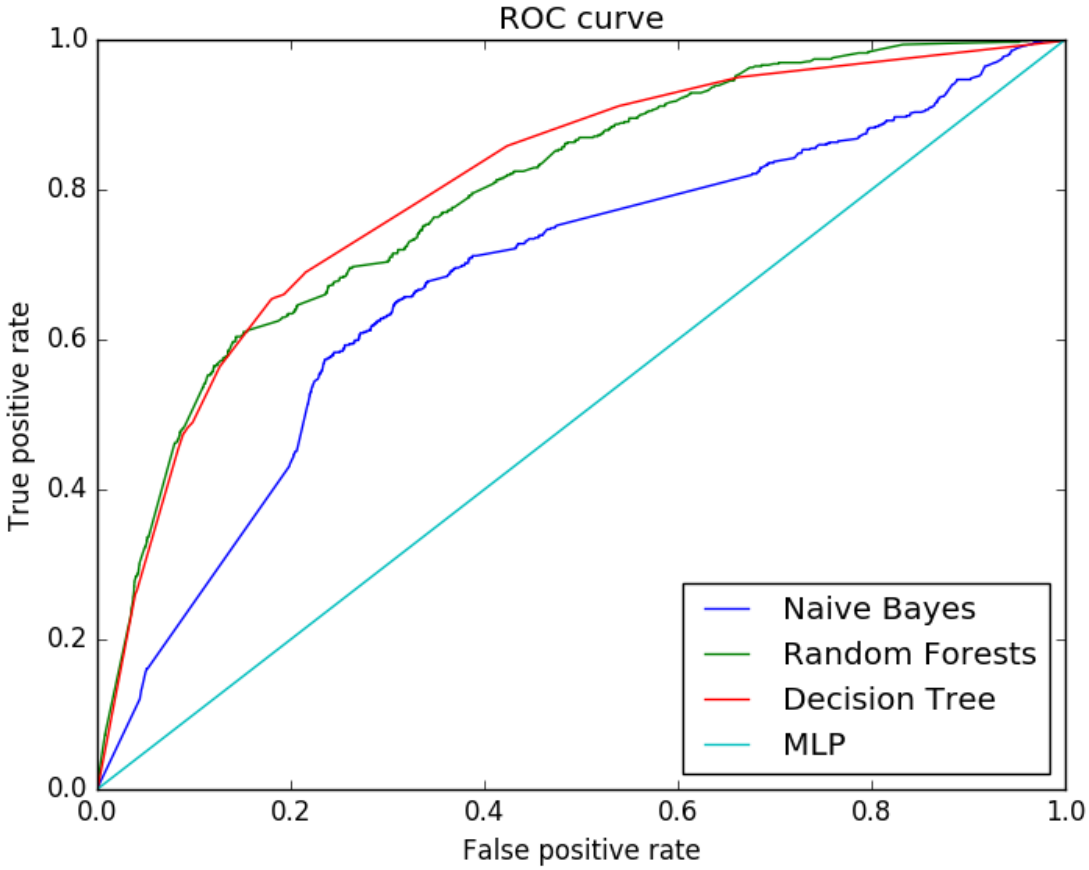
Once the above dependencies are resolved, the program may be executed from command line as below:
../SantanderCustomerSatisfaction> python ClassifyData.py
The data must be present in the /input directory of the working directory of the program.

Note: Execution will take some time to train the MLP classifier. Pycharm (download Pycharm) may installed and then the program may be executed from its runtime as well.

Once completed, the statistics will be printed as below (may vary) and the ROC curve will be produced as an image:

```
Training MLP...
Root Mean Square Error, epoch 1: 0.138362466815  and Error: 0.019144172223
Root Mean Square Error, epoch 2: 0.137821474894  and Error: 0.0189947589418
Training MLP complete for  2  epochs

================================================
Using Classifier:  Naive Bayes
Accuracy= 0.73802946593
Roc AUC for Naive Bayes :  0.680032824491

================================================
Using Classifier:  Random Forests
Accuracy= 0.959155485399
Roc AUC for Random Forests :  0.796799551416

================================================
Using Classifier:  Decision Tree
Accuracy= 0.958958168903
Roc AUC for Decision Tree :  0.805817673348

================================================
Using Classifier:  MLP
Accuracy= 0.959155485399
Roc AUC for MLP :  0.5
MLP RMSE of prediction:  0.202100258786
```

ROC Curve


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes
Random Forests
Decision Tree
MLP