**Assignment-2 CS-584 Data Mining**

Team name: Mr. miner

Team members: Murtaza Shareef(G01024452) and Deepak Kanuri(G01070295)

Rank: 29 F1-score:0.64

1)Reading the data:

Both the training data and test was read and stored in panda data frames.





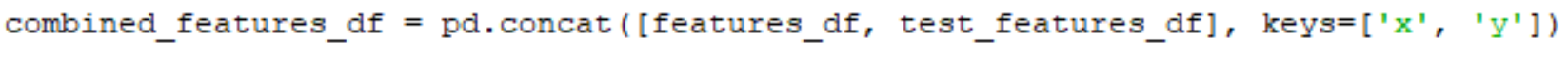
2)Pre-processing the data

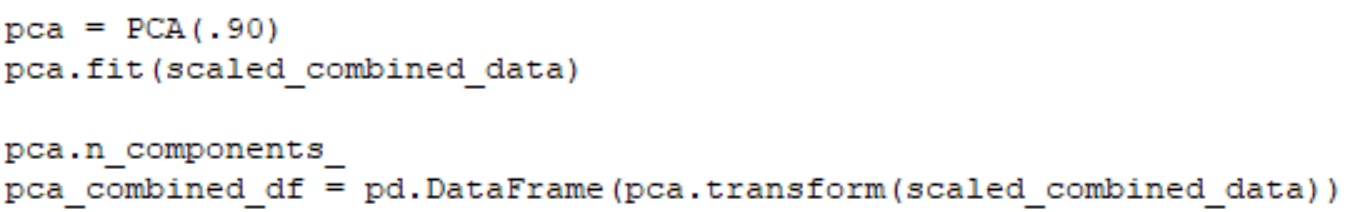
Both the training data and the test data was vectorized using ‘count-vectorizer‘. When vectorizing the training data, the features were vectorized and the acti/non-active data were stored as target data frame.

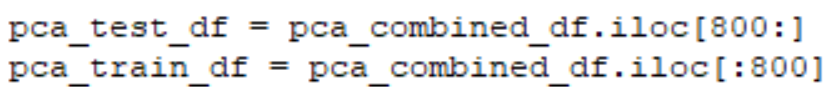


3)Combining the data

The training and test data were combined to form a single dataframe just to perform PCA on it. During PCA 90% of variance was retained. The reason for taking 90% was not to retain much information, and as the data set was big. Later on the training and test data was split and stored separately. The training data is the split for cross-validation purposes. The training data was split into 80% fro training purposes and 20% for testing the classifier.



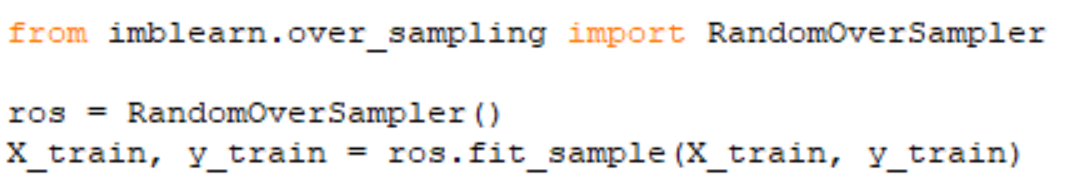






4)Handling imbalanced data

In the beginning we tried SMOTE to balance the data, but did not get great results. So later on we shifted to oversampling. Oversampling got us better results.



5)Classification

We fiddled with so many classification models. We tried SVM grid-search, random forest, decision trees and boosting. When we tried SVM grid search we got less f1-score. We hoped to get better results with random forest classification and we did. When we tried with decision trees we got a really bad result. Then we shifted our attention to boosting. We first tried XGBoosting and got good results on our training data. But later when we applied XGBoosting on the testing data we faced problems and the classification di not work. Then we dddecided to go ahead with gradient-boosting, and when we tried that we got good results while doing cross-validation and when we used the training data we got better results and thus went ahead with gradient-Boosting.

