Name: Taha Shahid

Capstone Project 1: In-depth Analysis (Machine Learning)

For the in-depth analysis of the default of credit card client’s dataset retrieved from uci repository following machine learning models and techniques were applied.

1. Decision Tree
2. Random Forest
3. XGBoost
4. ADABoost
5. Logistic Regression
6. Ensemble
7. Bagging
8. Feature Engineering
9. ADABoost after Feature engineering
10. XGboost after Feature engineering
11. Ensemble after Feature engineering

In order to fit the machine learning models, first, the dataset, which was imported as a dataframe using pandas library, was used to make X(predictors) and y(target) using the following commands:

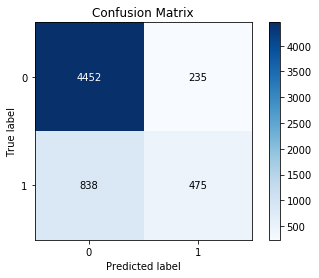
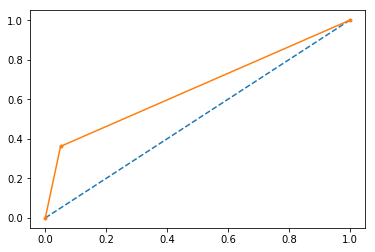
X=dfs.iloc[:-1].values

y=dfs.iloc[:,[24]].values

Second, the data was split into test and training sets using the sklearn library.

#### Decision Tree:

Decision tree algorithm did quite well for being a simple algorithm as compared to the other algorithms performed in the analysis and it scored roc-auc score of 0.6558.



A 5-fold cross-validation score mean was 81.41%. And on the test set, decision tree accuracy achieved was 82.11%

#### Random Forest:

For random forest ‘gini’ criterion was used with 4 jobs and 400 estimators. An accuracy of 82.08% was achieved on the test set. Also a roc-auc score of 0.6523.A 5-fold cv score mean was 81.41 %.

#### XGBoost:

For XGboost, three parameters were optimized using RandomizedSearchCV, which is also a function of sklearn library. From optimization n\_estimators were set to 57, min\_child\_weight was set to 9, and max\_depth was set to 3. A 5-fold cv score mean was 82.19%. And test set score of 82.02%.

#### ADABoost:

For ADABoost similar procedure of optimization was used as XGBoost and n\_estimators was set to 162, and learning\_rate of 0.001. A 5-fold cv score mean was 81.96%. Test set score was 81.93%.

#### Logistic Regression:

After optimization, using the above-mentioned optimization in sklearn following parameters were set:

penalty='l2',dual=False,max\_iter=145,C=2,fit\_intercept=False

A 5-fold cv score mean was 77.81%. A test set score was 78.11 %.

#### Ensemble:

For ensemble, a voting classifier was used with logistic regression, random forest, decision tree, xgboost,adaboost and parameter voting was set to hard.

Accuracy score of 81.8% was used. The best classifier was xgboost if used as is without changing any parameters.

#### Bagging:

bag\_clf = BaggingClassifier(RandomForestClassifier(), n\_estimators=500,

bootstrap=True, n\_jobs=-1, oob\_score=True)

81.95% accuracy score was achieved using above classifier.

bag\_clf = BaggingClassifier(DecisionTreeClassifier(), n\_estimators=500,bootstrap=True, n\_jobs=-1, oob\_score=True)

81.53% accuracy scored was achieved using above classifier.

#### Feature Engineering:

Feature 1 was addition of first three payments names PAY\_0, PAY\_2 and PAY\_3.

Feature 2 was using above three payments and making a binary column if any of the three above payments had -2 or +2.

Feature 3 was sixth month payment minus the sixth month bill amount.

Feature 4 was first month payment minus the first month bill amount.

After feature engineering xgboost performed the best and achieved an accuracy of 82.22 % on the 5-fold cv score mean and the best confusion matrix. Also, a solid 82% accuracy on the test set.