Assignment 2

Data Description

Dataset 1: Appliances Energy (https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction)

We have cross-sectional dataset with 19735 observations and 29 variables. The variables are Appliances and lights energy consumption in the house, 9 temperatures and 9 relative humidity in 9 different locations in the house, 6 weather variables such as outside temperature, outside humidity, Pressure, Windspeed, Visibility and Dewpoint. We also have date column, and two random variables that are being ignored in this model. Additional information about the dataset can be found out in the above-mentioned link.

Dependent variable is Appliances which is categorized as binary energy_high variable(1 if energy is greater than median) and Date variable is dropped from the dataset for convenience

None of the variables consists of missing attributes so no missing value imputation is required

Data is partitioned randomly into train and test using 70:30 split percentage(13814 and 5921 observations respectively)

All the features are scaled to the range [0,1] by subtracting each feature from its mean and dividing by standard deviation in order to maintain evenness in the data

Dataset 2: German Credit Data (Attached)

We have cross-sectional dataset with 1000 observations and 21 variables(7 numerical and 14 categorical). Target variable is Default which is renamed as Creditability (binary, 1=Good, 0=bad).

None of the attributes have missing values so no missing value imputation is required

The numerical attributes are rescaled(min-max normalization) and for categorical attributes OneHotEncoding is used to create the dummy variables. The dataset now has 1000 observations and 62 attributes.

Data is partitioned randomly into train and test using 70:30 split percentage(700 train and 300 test observations respectively)

Class imbalance exists in the data. Synthetic Minority Oversampling Technique(SMOTE) is used to balance the data after which there are 972 observations in the train data. All the algorithms performed better after class imbalance is taken care of.

This dataset is chosen given the number of categorical variables it has for analyzing whether an applicant is creditable or not. The dataset provides a credit simulation close to the realistic credit simulation in the modern credit analyses with additional variables like the criminal records of applicants, their health information etc. This is similar to what is done by a bank before a credit is approved.

Modern credit analyses employ many additional variables like the criminal records of applicants, their health information, net balance between monthly income and expenses. A dataset with these variables could be acquired or complementary variables added to the dataset. This will make the credit simulations much realistic, similar to what is done by the banks before a credit is approved.

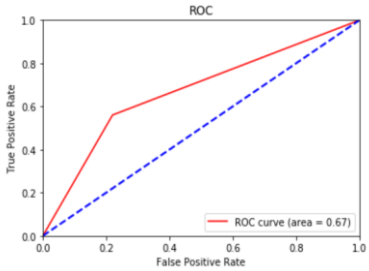
ROC Curves are used to evaluate the performance whereas algorithm learning is observed using Learning Curves

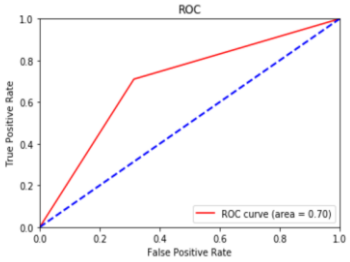
Cross-Validation is implemented to calculate the prediction accuracies of each of the algorithms

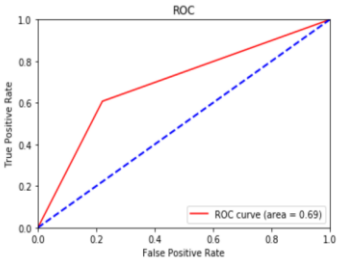
Implementation using various Algorithms

Naïve Bayes'

ROC Curves







Appliances — German Credit – Unbalanced — German Credit – Balanced

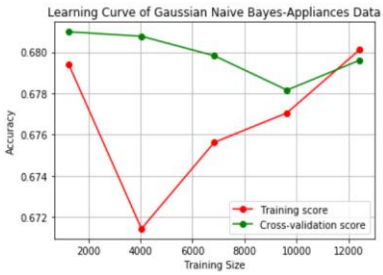
ROC_AUC:0.67 — ROC_AUC:0.70 — ROC_AUC:0.69

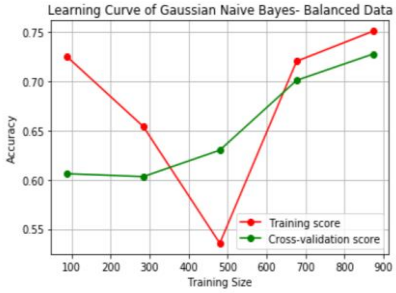
Appliances: Accuracy = 68% , Train Error = 32%, Test Error = 32%

German: Unbalanced: Accuracy = 70%, Train Error = 25%, Test Error: 30%

Balanced: Accuracy = 65%, Train Error = 25%, Test Error: 35%

Learning Curves

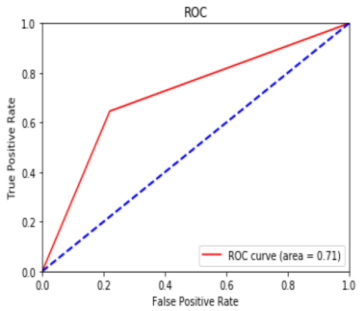


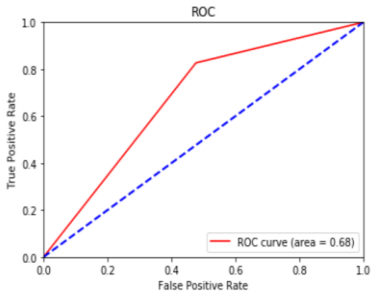


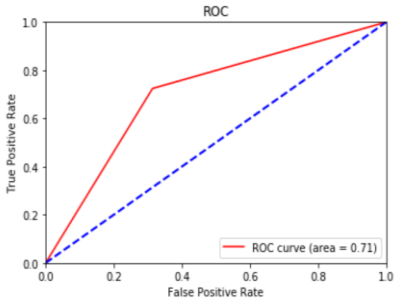
The Train and the validation scores seem to be fluctuating and in case of Appliances Data Cross-Validation score is more than the training score which is not as expected. Since this is Naïve Bayes classification, the algorithm tries to learn whatever the data is provided because of which the train and validation scores keep changing at unexpected rates

Logistic Regression

ROC Curves







Appliances | German Credit – Unbalanced | German Credit – Balanced

ROC_AUC:0.71 | ROC_AUC:0.68 | ROC_AUC:0.71

Appliances: Accuracy=72%, Cross Validation Accuracy=73%, Train Error=27%, Test Error=28%

Confusion Matrix:
```
[[2482  701]
 [ 970 1768]]
```
(2482+1768=4250) observations are classified correctly whereas (970+701 = 1671) observations are misclassified

German Credit:

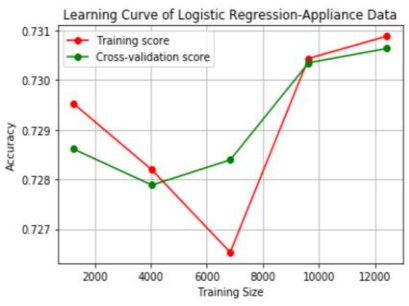
Unbalanced: Accuracy=74%, Cross Validation Accuracy=75%, Train Error=20%, Test Error=26%

Confusion Matrix:
```
[[ 45  41]
 [ 37 177]]
```
222 observations are classified correctly where as 78 are misclassified

Balanced: Accuracy=75%, Cross Validation Accuracy=75%, Train Error=21%, Test Error=29%

Confusion Matrix:
```
[[ 59  27]
 [ 59 155]]
```
214 observations are classified correctly where as 86 are misclassified

Learning Curves



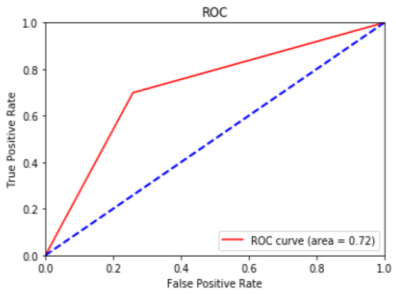
Appliances: After certain number of observations(4000) train score falls down suddenly whereas validation score keeps improving and the accuracy is almost similar when there are enough number of observations in the data(>8500)

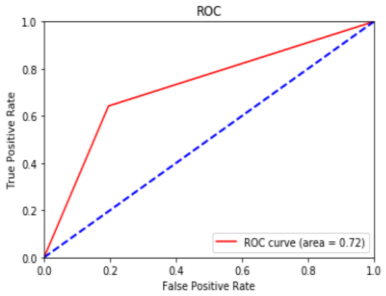
German: Unbalanced data has high bias and high variance where as the balanced dataset has validation accuracy improving as the number of observations crosses 500

SVM

ROC_AUC

Appliances





Linear — ROC_AUC:0.72

Polynomial — ROC_AUC:0.72

Appliances:

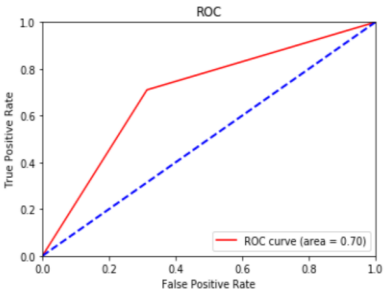
Linear SVC: Cross Validation Accuracy=73%, Train Error=27%, Test Error=28%

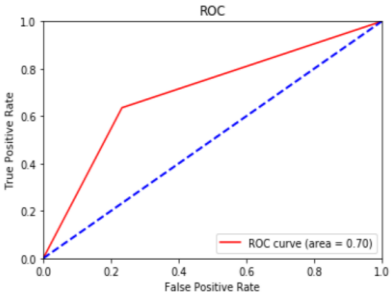
Confusion Matrix: `[[2361 822] [ 825 1913]]` 4274 obs are classified correctly whereas 1647 are misclassified

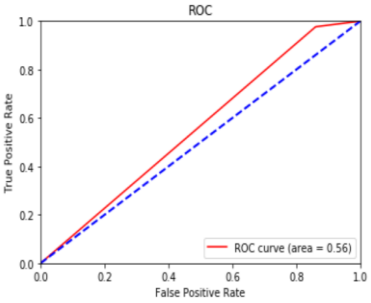
Poly: Cross Validation Accuracy=72%, Train Error=28%, Test Error=27%

Confusion Matrix: `[[2561 622] [ 980 1758]]` 4319 observations are classified correctly where as 1602 are misclassified

German Credit







Linear — ROC_AUC:0.70

Polynomial Kernel — ROC_AUC:0.70

rbf Kernel — ROC_AUC:0.51

German Credit:

Linear SVC: Cross Validation Accuracy=76%, Train Error=21%, Test Error=30%

Confusion Matrix: `[[ 59 27] [ 62 152]]` 211 observations are classified correctly where as 89 are misclassified

Poly: Cross Validation Accuracy=73%, Train Error=25%, Test Error=33%

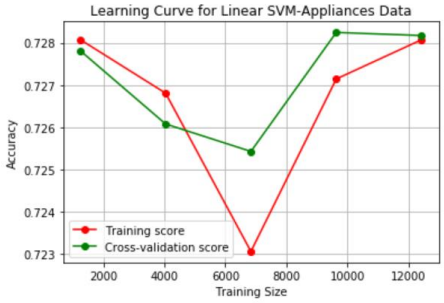
Confusion Matrix: `[[ 66 20] [ 78 136]]` 202 observations are classified correctly where as 98 are misclassified

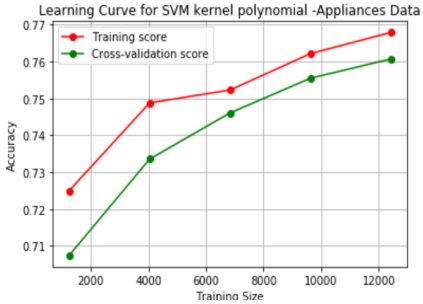
rbf: Cross Validation Accuracy=92%, Train Error=0.1%, Test Error=26%, which might be due to overfitting

Confusion Matrix: `[[ 12  74]`
`[  5 209]]` 221 observations are classified correctly where as 79 are misclassified

Learning Curves

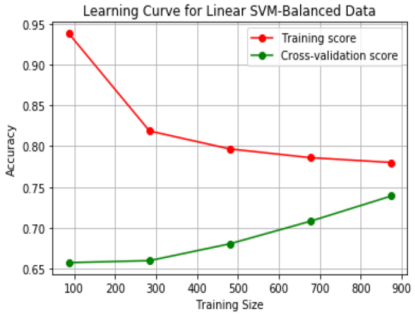
Appliances

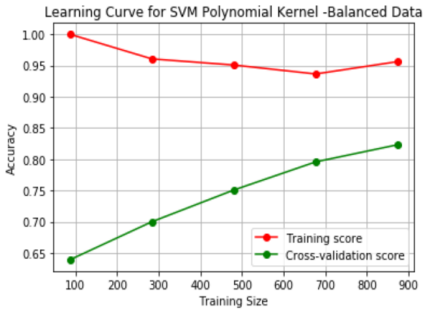


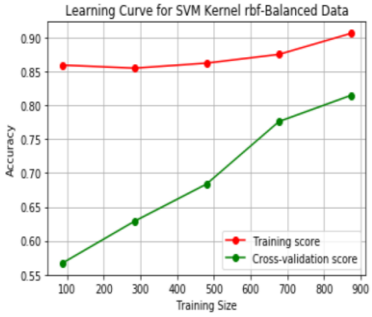


Linear SVM shows fluctuations in the learning where as polynomial has both training and validation score increasing at a gradual rate where as the area under ROC is same for both the linear and polynomial

German Credit

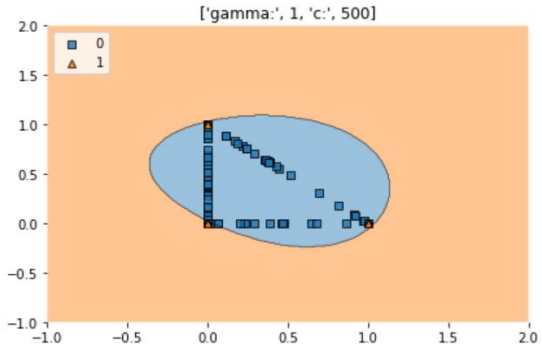


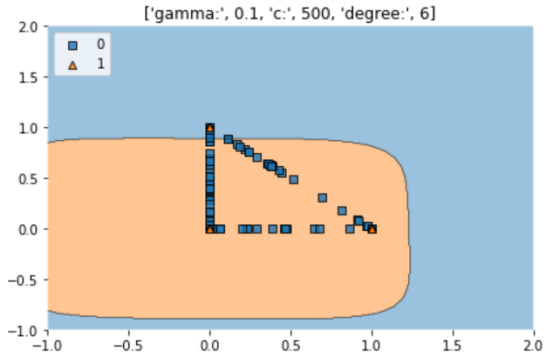




Rbf shows constant improvement in learning but the accuracy is very low when compared to linear and polynomial. Since polynomial kernel has improved validation accuracy this has the high learning rate

After an experimentation, for a certain German data, it is found that rbf kernel with gamma=1 and C=500 and polynomial kernel with gamma=0.1, degree=6 and C=500 are the best values



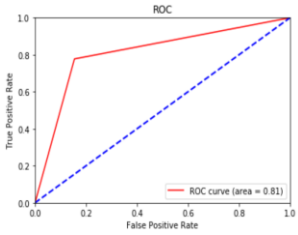


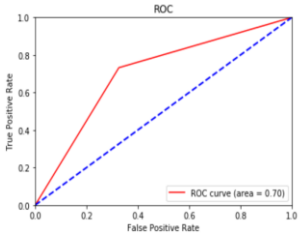
Since the best estimates always tend to overfit the data, we decide on the parameters based on the ROC and Learning Curves which suggests small values for all three parameters.

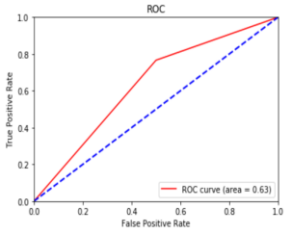
Both Appliances and German Credit Data are linearly separable because of which Linear SVM gives better results compared to Polynomial and rbf Kernels

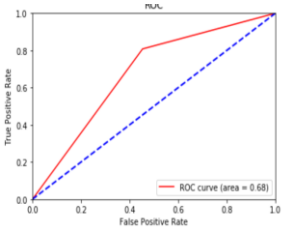
Decision Trees

ROC Curves









Appliances-GINI	Appliances-Entropy	GermanCredit-GINI	GermanCredit-Entropy
AUC_ROC:0.81	AUC_ROC:0.70	AUC_ROC:0.63	AUC_ROC:0.68

Appliances:

GINI: Cross Validation Accuracy=83%, Train Error=7%, Test Error=18.6%

Confusion Matrix: `[[2691  492]`
`[ 609 2129]]`4820 observations are classified correctly where as 1101 are misclassified

Entropy: Cross Validation Accuracy=70%, Train Error=29%, Test Error=31%

Confusion Matrix: `[[2145 1038]`
`[ 733 2005]]`4150 observations are classified correctly where as 1771 are misclassified

German Credit:

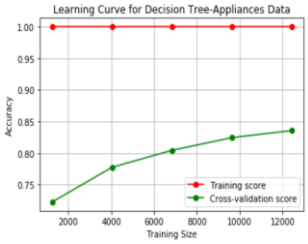
GINI: Cross Validation Accuracy=75%, Train Error=9%, Test Error=26.6%

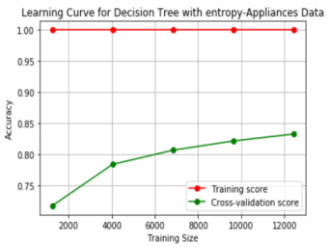
Confusion Matrix: `[[ 47  39]`
`[ 41 173]]`220 observations are classified correctly where as 80 are misclassified

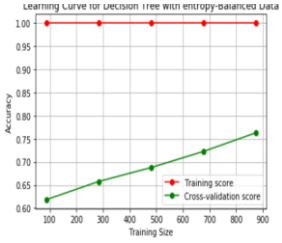
Entropy: Cross Validation Accuracy=69%, Train Error=28%, Test Error=27.6%

Confusion Matrix: `[[ 11  75]`
`[  8 206]]`217 observations are classified correctly where as 83 are misclassified

Learning Curves







Appliances: For validation set, GINI's score is better than entropy's but there is a very slight variation

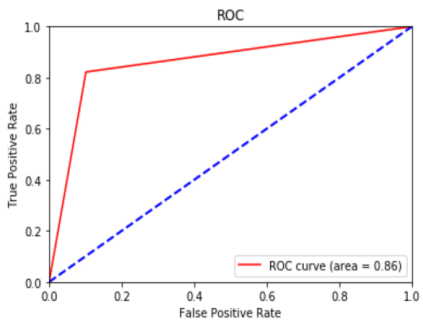
German Credit: For validation set, Entropy's score is better than GINI's but again only a slight variation

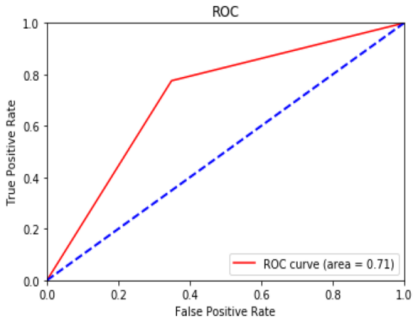
There is a high variance between the train and validation scores so there is a possible overfitting in the data because of which the prediction accuracies might not be appropriate

According to the area under ROC curves for Appliances, GINI index gives good prediciton accuracy and for German Data, Entropy seems to give good prediction accuracy whereas according to the validation accuracy both the datasets gives better prediciton accuracy while using GINI index

Random Forest

ROC Curves





Appliances

AUC_ROC:0.86

German Credit

AUC_ROC:0.71

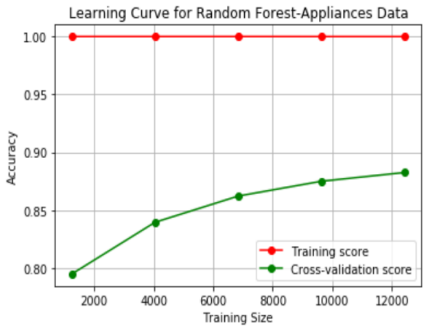
Appliances: Cross Validation Accuracy=87%, Train Error=0.8%, Test Error=13.7%

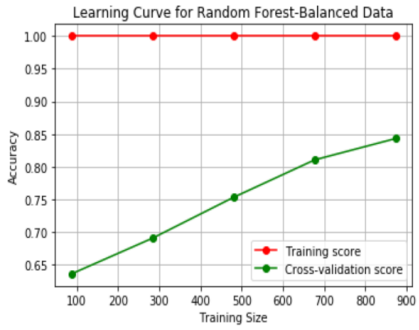
Confusion Matrix: `[[2860  323]`
`[ 487 2251]]` 5111 observations are classified correctly where as 810 are misclassified

German Credit: Cross Validation Accuracy=81%, Train Error=0.3%, Test Error=26%

Confusion Matrix: `[[ 56  30]`
`[ 48 166]]` 222 observations are classified correctly where as 78 are misclassified

Learning Curves





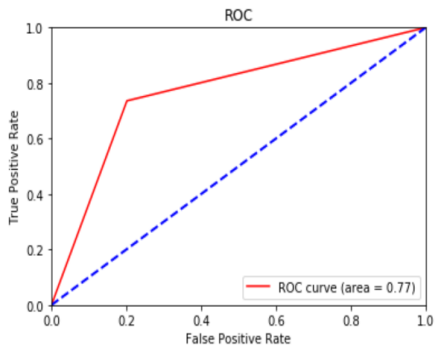
So far, Random Forest gave the highest prediction accuracies for both Appliances and German Credit Data, number of misclassifications are very less when compared to the other algorithms, train and test error rates very small

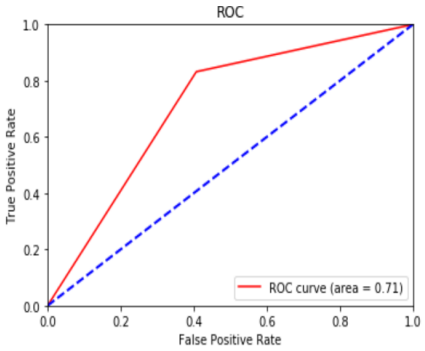
There is not much variation between the train and validation scores so we can say that the algorithm has a high learning rate, algorithm learns the test data as good as the training data and the prediction accuracies are appropriate

Area under the ROC curve is also high when compared to the algorithms for both the datasets

Gradient Boosting

ROC Curves





Appliances

AUC_ROC:0.77

German Credit

AUC_ROC:0.71

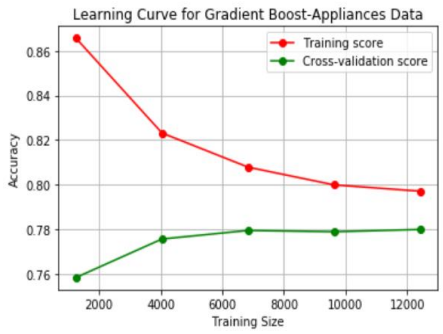
Appliances: Cross Validation Accuracy=78%, Train Error=20%, Test Error=23%

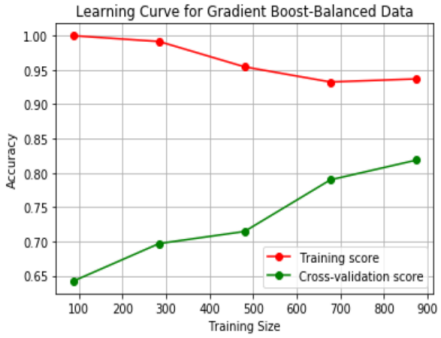
Confusion Matrix: `[[2541  642]`
`[ 723 2015]]`4556 observations are classified correctly where as 1365 are misclassified

German Credit: Cross Validation Accuracy=82%, Train Error=7%, Test Error=23.6%

Confusion Matrix: `[[ 51  35]`
`[ 36 178]]`229 observations are classified correctly where as 71 are misclassified

Learning Curves





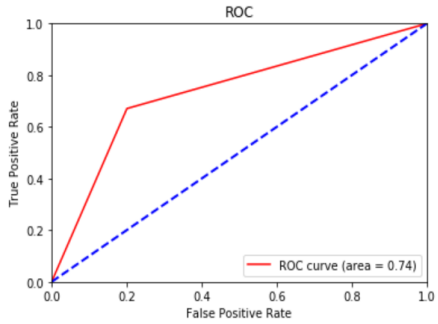
Learning curves suggest that as the number of observations increases training accuracy is decreasing.

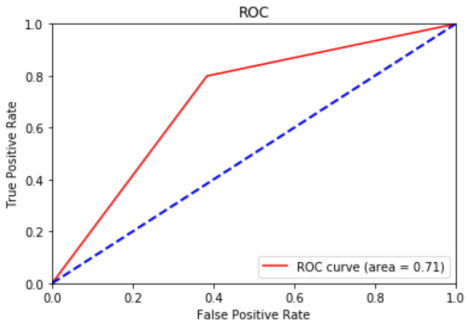
For Appliances data, after 12000 observations both training and validation accuracies are becoming constant and the gap suggests that there is a chance of overfitting

For German Credit data, as the number of observations reaches a certain number, train and validation scores are decreasing and increasing respectively at a certain rate. Once the number of observations is close to 700, both train and validation accuracies seem to increase. The overall accuracy is high for both training and validation sets.

AdaBoost

ROC Curves





Appliances

AUC_ROC:0.74

German Credit

AUC_ROC:0.71

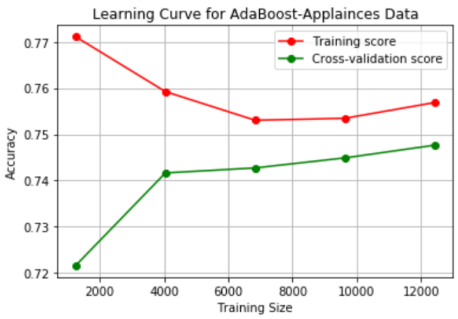
Appliances: Cross Validation Accuracy=75%, Train Error=24%, Test Error=26%

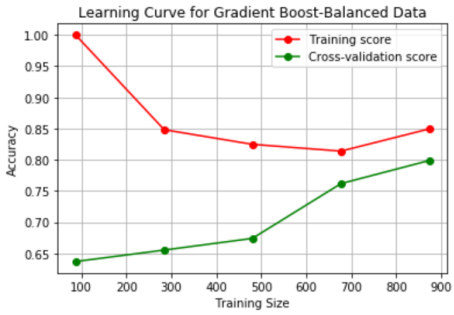
Confusion Matrix: `[[2544 639] [ 901 1837]]` 4381 observations are classified correctly where as 1540 are misclassified

German Credit: Cross Validation Accuracy=79%, Train Error=15.6%, Test Error=25.3%

Confusion Matrix: `[[ 53 33] [ 43 171]]` 224 observations are classified correctly where as 77 are misclassified

Learning Curves





The learning curves are similar to the ones obtained in Gradient Boosting. As the number of observations increases, the accuracy for both train and validation sets are increasing

Area under the ROC curves is also high which indicates that using AdaBoost increased the prediction accuracies in both Appliances and German Credit Data

Conclusion:

Appliances

The performance of the seven algorithms achieved an average accuracy of 76.57%. Three algorithms (Decision Trees, Gradient Boosting and Random Forest) achieved the top three accuracy scores after modeling. Of them, Random Forest turned in the top result using the training data. It achieved an average accuracy of 87%, followed by Decision Tree with GINI index with 83%. Random Forest algorithm has the area under ROC curve 0.86 which is the highest of all the other algorithms.

From the model-building activities, the Random Forest ensemble algorithm yielded the top-notch training and validation results. It is the recommended algorithm to use from the accuracy perspective.

German Credit

The performance of the seven algorithms achieved an average accuracy of 76%. Three algorithms (Random Forest, Gradient Boosting and AdaBoost) achieved the top three accuracy scores after modeling. Of them, Random Forest and Gradient Boosting turned in the top result using the training data. It achieved an average accuracy of 81%, followed by AdaBoost with 79%. Logistic Regression, Random Forest, Gradient Boosting and AdaBoost algorithms have the area under ROC curve 0.71 which is the highest of all the other algorithms. The low score on area is result of the size of the data.

From the model-building activities, the Random Forest and Gradient Boosting ensemble algorithms yielded the top-notch training and validation results along with small training and test error rates which implies very less misclassification. These two are the recommended algorithms to use from the accuracy perspective.

Appliances data showed better results compared to German Credit in terms of prediction accuracies because for the Appliances data 10 best features are selected based on the correlations with the target and the German Credit Data is modelled using all the 62 features. Class imbalance has been taken care of, but the number of features caused a problem in the prediction accuracies. Dimension Reduction algorithms such as PCA can be used to select the best features.