**Data Mining Project Report**

**German Credit**

**2202 MSA 6440 5001 Data Mining**

**Project Documentation**

**Venkata Sai Suneeth Ravilla**

**Academic Supervisor:**

**Shuchismita Sarkar**

**Spring 2020**

**Bowling Green State University, Ohio**

**Dataset Background:**

Dataset represents person who takes credit from a bank. The response variable is classified as good or bad credit risks. The dataset contains 1000 records with 20 variables provided by Dr. Hans Hofmann. These 20 variables contain 7 numerical and 13 categorical variables.

**Variables:**

**Numerical:** Duration of month, credit amount, Instalment rate in percentage, residence, Age, existing credits in the bank, Liable to provide maintenance.

**Categorical:** Credit history, Purpose, Account, Employment, Gender, guarantors, Property, Installment plan, Housing, Job, Telephone, Foreign worker or not

**Goal:**

The main aim of this report is to find the method which can estimate whether the customer is good or bad to approve credit. The performance of the models is the key as better the accuracy of the prediction the banks gets profitable.

**Approach:**

In this report, focusing on designing a logistic regression, regression tree, random forest for predicting if the customer is good or bad. The dataset is partitioned into training and validation, the models are built on training data and validation is performed on validation data.

**Major Findings:**

From the analysis in the report, based on the models the customer can be estimated either by using a logistic model or by a random forest with better accuracy. But by comparing some key parameters we identified that Random forest gives better results. Hence the customer can be estimated using Random Forest might give good results.

Below report gives detailed analysis and approach followed for estimating customer.

**Data Preprocessing:**

Below steps are performed to ensure the dataset is cleaned and ready for building the models. That is because while gathering the data, it is not always perfect, and it will help in advance to find any missing values and irrelevant data.

**Missing data:**

Aim in this task is to analysis the data whether the data has any missing values to make sure the data has the values completely for the analysis.



No missing values are found in the given dataset. This ensures that there is no need to pass any mean or median values for the empty cells.

**Noisy data:**

To predict the model more accurately without any noise we need to remove the variables that are just meaning less to the data. From the data we can see Gender and Phone does not seem appropriate to fit the model so removing while predicting the model.

**Data Transformation:**

Given above that preprocessing helps in identifying irrelevant data and missing values, data transformation helps to transform the data into appropriate justification for the analysis.

**Normalization:**

This technique is performed in preprocessing which provides linear transformation on original range of the data

**Categorical data:**

The categorical data present in the dataset has various levels which belongs to specific sets and these need to be scaled for better fit of the model.

**Numerical data:**

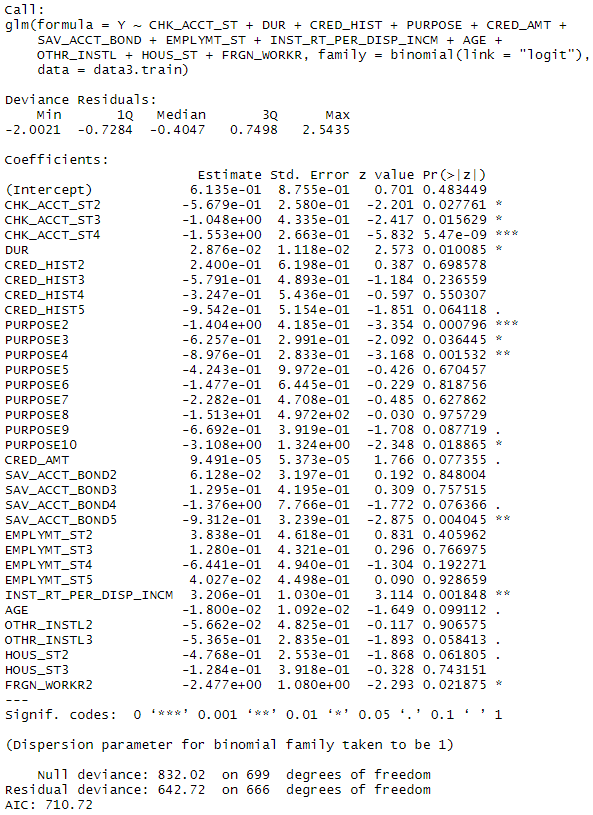
Feature scaling is required as the values in numerical data has different range, like credit history ranges from 3 digits to 5 digits but the age column ranging in 2 digits which some models might get effected.

**Data Configuration:**

The target predictor Y given as 1 representing good and 2 representing bad is changed as 0 representing good and 1 representing bad.

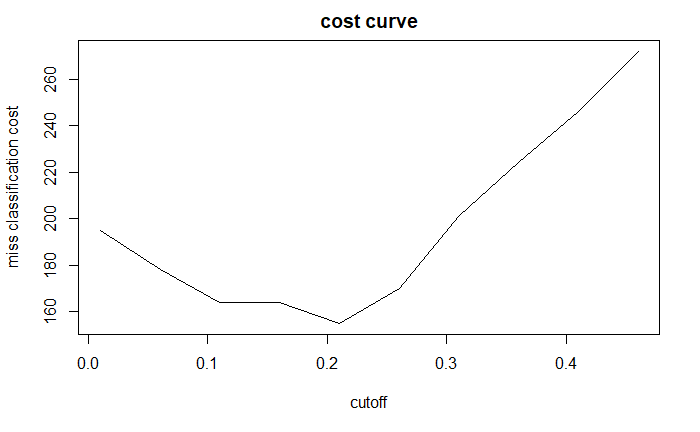
**LOGISTIC MODEL**

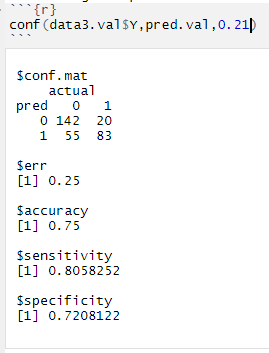
Building logistic regression model that can be used to predict Y in the future. For this logistic model performed preprocessing steps – Categorical data which is non numerical which is specified as categories changed to factor, Data configuration of changing Y and splitting the data to 70% as train set and 30% as validation set.

Logistic model is built with response variable Y to train the training set. From the logistic model, there are many non-significant variables fitted to the model, so performing stepwise regression to eliminate the non-significant variables.

Beside is the stepwise fitted logistic model on training dataset. As the stepwise is performed many non-significant variables are removed from the model. From the logistic model, the null deviance is 832.02, residual deviance is 642.72 and AIC of the model is 710.72.

Based on the fitted model, the validation set is predicted to check the performance of the model.

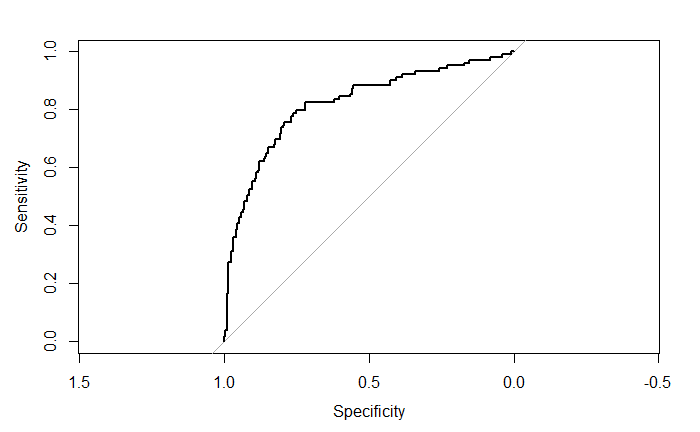
Cost curve is plotted using predicted values and validation set values. The curve decreasing shows that the misclassification cost has decreased when the cutoff value has increased then it turned into another direction as the cutoff value increased misclassification cost increased.



From the developed random forest model which is generated on training set we are now checking the performance of the model with the left over 30% validation data. From the confusion matrix, out of 300 observations, the model has predicted 83 as bad customers.

* 142 observations are predicted as good customers
* 55 records which has actual outcome as bad are wrongly predicted as good.

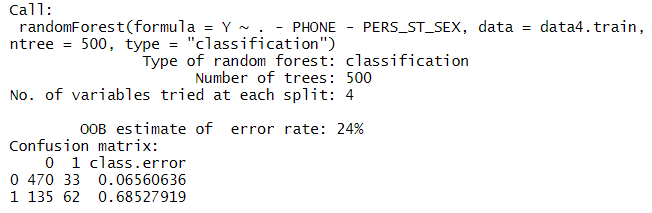
The sensitivity of the model is 0.8 and accuracy of the model is 75% which is significant.

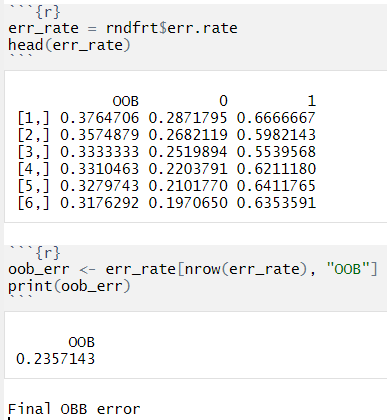
From above, we identified confusion table only for one probability which is cut off probability. Beside is the ROC curve for the logistic model using the validation data. Through ROC curve we can examine the same for full range of cut off values from 0 to 1. The AUC value for the ROC curve is 0.8132 which indicates that 81.32% of the time a randomly selected pair of subjects will be correctly predicted by the model.

**RANDOM FOREST MODEL**

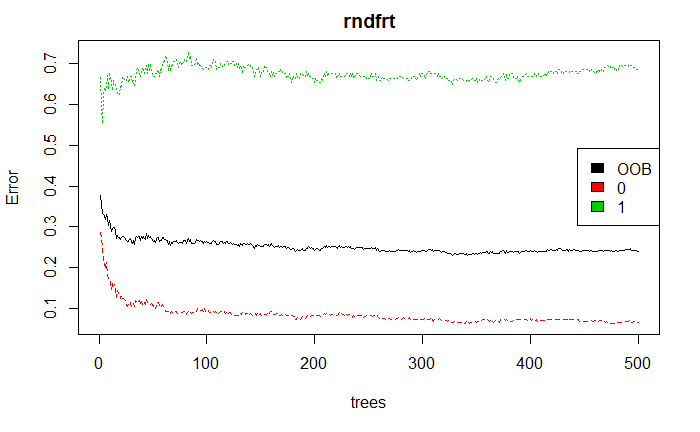
In the random forest model, we check the performance of the model in predicting the outcome of the variables. In the above logistic regression model, scaling is not performed for the numerical variables. But for this random forest model, scaling is performed to evaluate the better efficient model. Even though scaling is not required for RF model, I think normalizing or scaling the attributes will help the classification task. Because, normalizing the attributes will result in more linear relationship. Then followed with the same steps done for Logistic model.

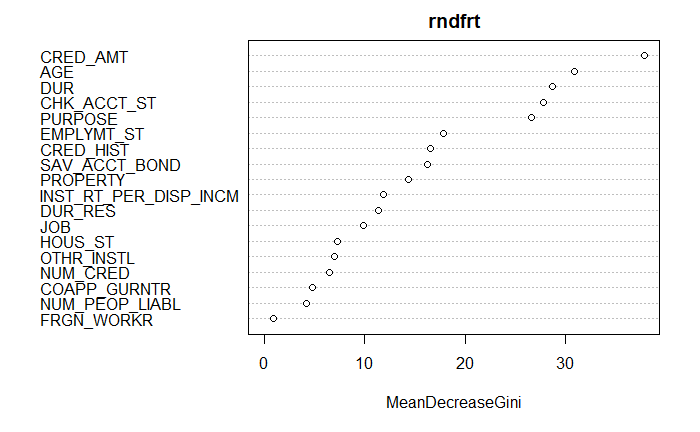
Data is partitioned randomly into training set 70% and validation set 30% from the original data.

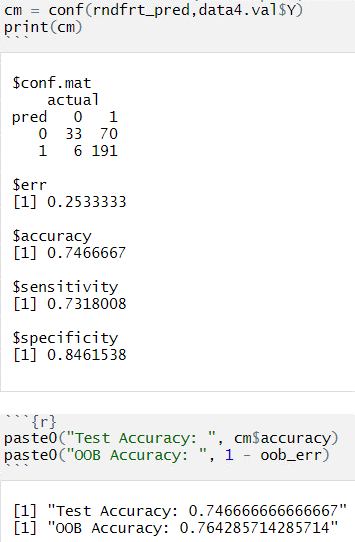


Random forest each tree is trained on roughly 2/3rd of the data. Number of trees grown was fixed at 500 to achieve as much gain as possible. and the number of features used as potential candidates for each split is 4. The model will automatically attempt to classify each of the samples in the Out-Of-Bag dataset and display a confusion matrix with the results. The output of the random forest, each tree is the left-over is used for calculating out of bag error. The OBB error rate is 24%

Each tree gives the classification over the left-over data which is OBB. The OBB error starts with 0.3764 and the last value of OBB in the model is 0.3257.

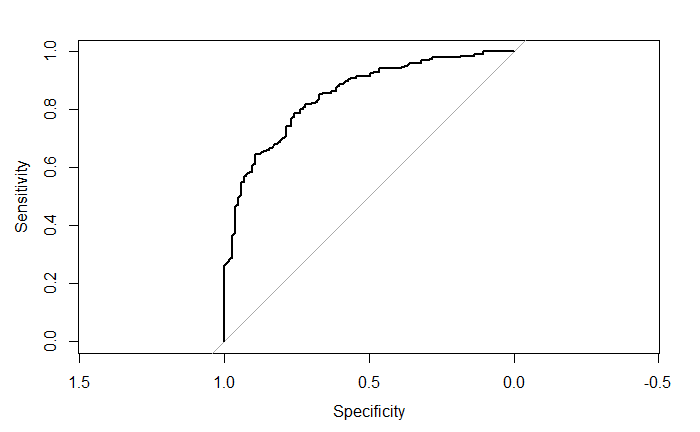
From the random forest plot, we can see the error for the Bad credit which shows in green color as “1” and the OOB error is shown in the black color.

Gini is the measure of node impurity. Highest purity shows that how pure each node contains only elements of a single class. From the Mean decrease Gini, we can identify how important the variables after overall splits are performed. From the plot we can see Credit amount, Age, Duration and checking account are top 5 important variables in the model.

From the developed random forest model which is generated on training set we are now checking the performance of the model with the left over 30% validation data. From the confusion matrix, out of 300 observations, the model has predicted 191 as bad customers.

* Only 33 observations are predicted as good customers
* 6 records which has actual outcome as bad are wongly predicted as good which is very less when comapred to logistic model.

The Test Accuracy is 74.6% and the OBB accuracy is 0.76%. The sensitivity of the model is very good with 0.73 which is 73%.

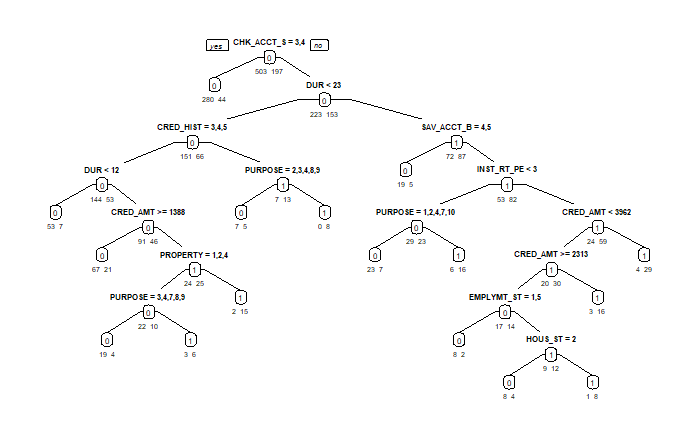
Plotting the ROC curve for the random forest model using the output of the validation data. The AUC for the random forest model is 0.8508 which is 85.08% of the time a randomly selectd pair of subjects will be correctely predicted by the model.

**DECISION TREE MODEL**

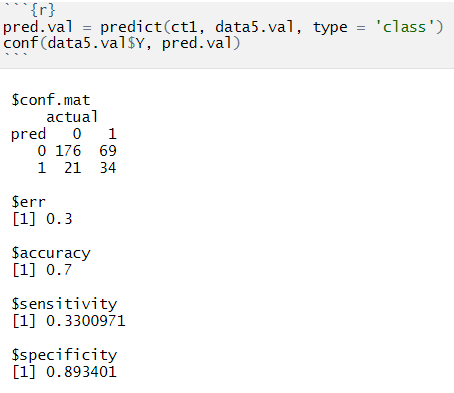
In the Decision tree final model, building the model to predict the outcome of the target variable Y using the validation data. In the previous model we have scaled for both categorical and numerical but here I’m scaling only the categorical varibles even though decision tree does not require scaling in general.

The dataset is partitioned into training set with 70% of the data and validation set with the remaining 30% of the data.

Decision tree model is built on training dataset and below is the classification tree plot for the model.

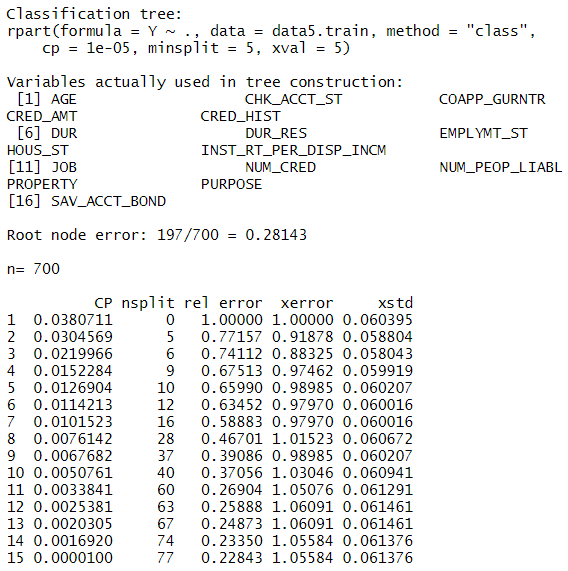


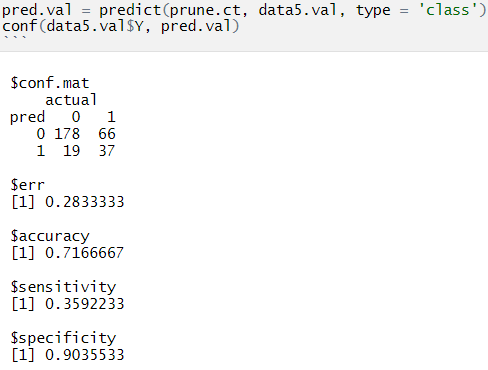
From the above tree, the new observations will be classified whether the customer is good or bad. The plot shows the generated tree has 14 leaves. These leave will help in predicting the outcomes from the validation data.

Based on the decision tree model built on the training set, we are now checking the performance of the model with the left over 30% validation data. From the confusion matrix, out of 300 observations, the model has predicted 34 as bad customers which is less compared to previous model.

* 176 observations are predicted as good customers out of 300
* 21 records which has actual outcome as bad are wongly predicted as good which is very less when comapred to logistic model.

The accuracy of the model is 70% and the sensitivity of the model is very good with 0.33 which is 33% very less compared to the random forest model.

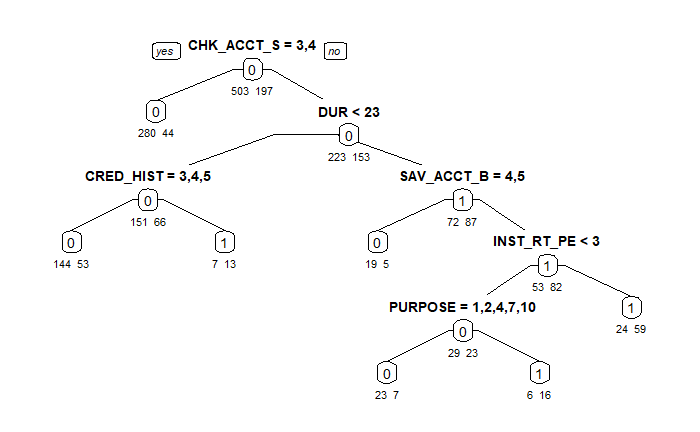
Cost complexity purning is performed in the training set in order to check the optimal size of the tree which has the lowest cost complexity. The ouput shows the prune result, based on the xerror value choosing the cost complexity. The values in the table gives information about the cross complexity and the xerror value for all the trees in the training set model. The root node error in this model is 0.28 and the minimun value of xerror is 0.91878 and the corresponding nsplit is 5

From the pruned decision tree model which is generated on training set we are now checking the performance of the model with the left over 30% validation data. From the confusion matrix, out of 300 observations, the model has predicted 37 as bad customers.

• 178 observations are predicted as good customers

• 19 records which has actual outcome as bad are wongly predicted as good which is very less when comapred to logistic model.

The accuracy is 71.6% and the sensitivity of the model is very good with 0.35 which has improved from previous model.

From the pruned tree we can see the number of leaves has reduced to 5. The above confusion matrix has also proved that the model has improved after the pruning.

The AUC for the decision tree model is 0.7364 which is 73.64% of the time a randomly selectd pair of subjects will be correctely predicted by the model.

**Justification:**

* Random Forest has higher sensitivity with equal accuracy as logistic model then followed by decision tree model.
* Therefore I would like to conclude that for the german credit data set, random forest classifier gives better prediction over other two methods for credit risk.

**Improvement:**

* Ensemble model of Random forest can be tuned for the better results in order to predict better bad credit risk.
* Scaling is performed for both numerical and categorical in Random Forest even though it is not needed, performing analysis without gives better idea to predict the credit risk.
* Therefore, we would have to come up with tuning the Random Forest to increase the performance of the model.