ASSIGNMENT – #8: (DECISION TREES AND SVM)

PURPOSE: In this assignment, we analyse the Credit Card Clients Data Set and try to predict the default payment of all clients. For the prediction, we use the classification techniques like bagging, Random Forests and Support Vector Machines.

DATASETS: Default of Credit Card Clients Data Set (Link).

APPROACH:

- There are 24 variables and 30000 attributes in the Dataset. The default payment (0 or 1), is the response variable.
- The predictors that are factors are levelled. The dataset is clean, that is there are no NA's or null values and each variable is properly formatted.
- First, I have applied the Random Forest method on the data with 'Y' as a response variable and including all other remaining variables as predictors.
- For this method, '*randomForest*' function is used from the package with same name.
- First, I have tried with all the predictors included in the model. Now, we will divide the dataset into training and testing data and check how well the model works.
- Now, using the training set we train the randomForest model with all the predictors included in the model. The model is used to predict the testing dataset.
- The confusion matrix of the actual values and the predicted values shows that about 81.7% of the testing data is predicted correctly.
- Next, I have included 10 predictors from the training set to train the randomforest model. This model is used to predict the testing data and confusion matrix shows about the 81.3% of the data is predicted correctly.
- The importance of the random Forest model gives mean decrease accuracy values for all the predictors showing that how much the model would affect removing the variable from the model.
- Next, I have tried Support Vector Machine to do the classification problem.
- I have used '*svm*' function from the '*e1071*' package. First, I have applied svm technique with the linear kernel at cost of 1 on the training set of data.
- Now the trained modal is used to predict the testing set data. The confusion matrix of the actual values and the predicted values of the testing data shows that 80.9% of the data is predicted correctly.
- Finally, I have applied the radial kernel of the svm technique. The result of the confusion matrix of the predicted values of the testing data with the actual values show that about 80% of the data is predicted correct.

SUMMARY:

- In this assignment, we want to classify data based on default payment type (0 or 1) whether the client is credible or not.
- When Random Forest technique is applied, with including all the predictors and with only few of them. The prediction results of the testing data show that about 81% of the data is predicted correctly.
- The SVM technique is applied to the dataset. *Linear* and *Radial* kernels are used to classify the data. In both the cases, the model predictions of the data is about 80% of the actual data.