Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – #7: (LOGISTIC REGRESSION AND LDA/NAÏVE BAYES)

***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 Logistic Regression, Naïve Bayes and Linear Discriminant Analysis.

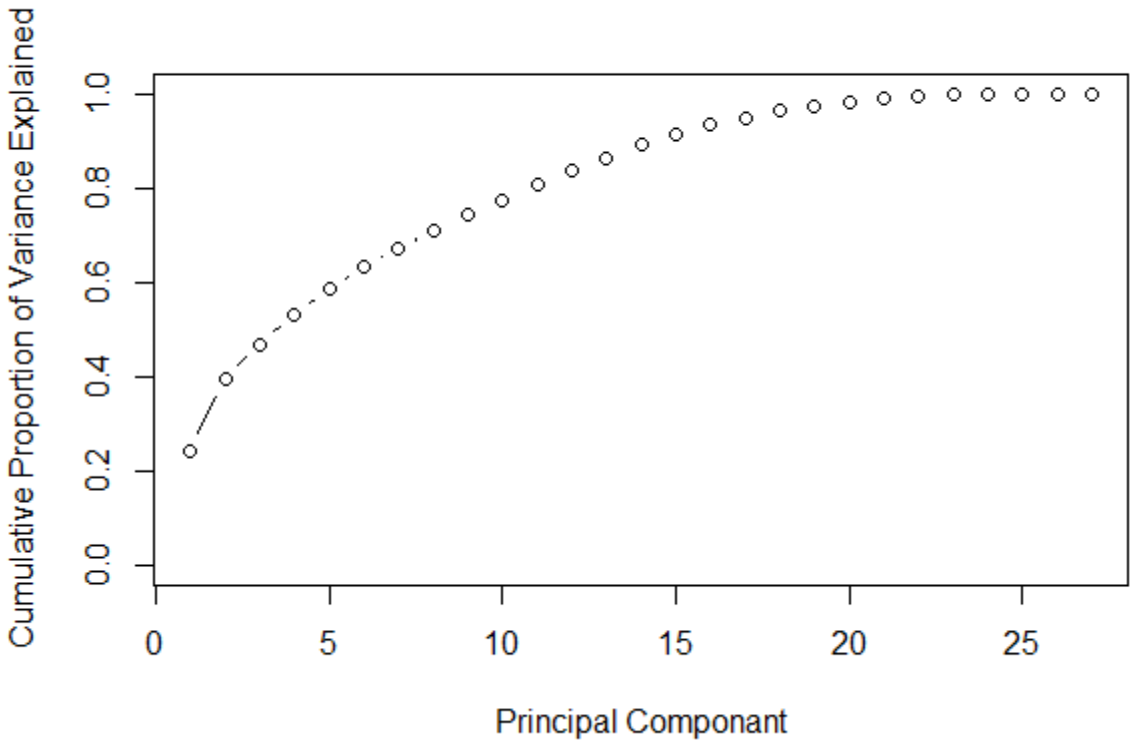
DATASETS: Default of Credit Card Clients Data Set (Link).

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 Logistic Regression method on the data with 'Y' as a response variable and remaining variables as predictors.
- The '*glm*' function is used for the Logistic Regression method. For this method, we indicate family as '*binomial*'.
- The summary of the above model give the details about Null deviance and Residual deviance which are 31705 and 27845 respectively.
- For a good model fit, the Null deviance should be large and Residual should be small. The above results show that the residual value is not very small to conclude it as a good model.
- From the summary of the above model, there are few predictors whose *p-values* are large. So, they are excluded and tried to build another model using same function and parameters.
- The summary of the new model with some of the predictors excluded, gives the null deviance and residual deviance values as 31705 and 27915 respectively. We see that removing some of the variables increased the value of residual rather than reducing.
- Now, we did a simple validation to see how much the model works means up to what percent of the data the model predicts correctly. For this, dataset is divided to training set and testing set with 7:3 ratio.
- First, using the training set we trained the model. Next, the model is used to predict the default payment for the test Data using the probabilities. Then, I tried to classify the probability greater than 0.5 as 1, and 0 otherwise.
- And, I tried to compare the predicted results with the actual test results. The results show that 81.02% of the test data is predicted correctly.
- Now, I applied dimensionality reduction using PCA and tried to find the variation using the principal components. The plot is shown in graph 1. We see that from the graph, variation increases along with the adding of more variables.

- Finally, I applied the Linear Discriminant analysis. For this, I used '*lda*' function from the 'MASS' package.
- To test the model, I did a simple validation. The dataset is divided into training and testing data. The training dataset is used to train a model and the model obtained is used to predict the response variable of test data.
- The predicted values and actual values are compared and found out the model predicted about 81.4% of the test data correctly.
- Finally, Naïve Bayes method is applied to only few of the variables. Similarly, as before, the dataset is divided into training and testing data. The testing data is used to build the model and the testing data is used to predict the data from the model.
- The predicted results from the testing data shows that nearly 80% of the testing data is predicted correctly.

GRAPHS:



(Graph 1)

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 logistic regression technique is applied, I observed that the null deviance is large and Residual deviance is not small. For a good model fit, residual deviance should be small.
- The model is again applied to dataset with the removal of predictors that got large p-value in the above model. This model also has the same variations and the residual deviance is increased in this case.
- When the '*lda*' method is applied, this model predicted about 81% correctly. And also the Naïve Bayes also predicted about 80% of the data correctly.