Assignment #07:

Prediction of Credit Card Payment Default

Team Members

Harika Addagada (hxa150930)

Ravali Nallamasu(rxn152730)

Aman Miryala (axm151830)

Himanshu Parashar(hxp151330)

Purpose: To create a model that will predict the default of credit card payment for a given dataset using the techniques Logistic Regression, Naïve Bayes/LDA/QDA.

Dataset: Credit Card Payment dataset available in below link

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Data Preprocessing:

- First, we removed the ID column from the dataset as it is the client identifier and doesn't add value to the analysis
- Renamed the "default payment next month" attribute to "default_payment" (response variable) for easy access of the column.
- Then we fixed the datapoint of education and marriage - combined few categories of Education together to get only 4 types i.e., 1 = graduate school; 2 = university; 3 = high school; 4 = others
- And, aggregated the marriage attribute values into 3 categories which are 1 = married; 2 = single; 3 = others
- Then, converted the character attributes Education, Marriage and Sex into factors for analysis.

Some Observations:

- From summary of data, we saw that there are more female than male in the dataset.
- There are more single people than married, i.e (53.2%:45.5%)
- From the correlation plots with demographic data (sex, education and marriage) against default_payment we can observe that married woman are more likely to default payment.
- The default_payment is approx 22% of the total recordings in the dataset.

Approach/Model Building:

Logistic Regression:

- Constructed a model using glm command for the entire dataset to predict the response variable "default_payment" based on the predictors available in the dataset.
- We observed there is not really much reduction in the deviance. The values we obtained were as follows:
  Null deviance: 31705, Residual deviance: 27845
- We had 30000 entries in the dataset, which we divided 24000 as training set and rest 6000 entries as test data and built the model on training set to predict the test data.
- The model did a moderate job and resulted in below predictions when compared against the original values available in the dataset:

	0	1
0	4544	139
1	993	324

- The best predictors for the model based on the summary of the model generated are LIMIT_BAL, SEX , Education, MARRIAGE AGE PAY_0 PAY_2 PAY_3 BILL_AMT1 PAY_AMT1 PAY_AMT2 PAY_AMT4

- Then, we applied Principal Component Analysis on predictors(see Graph 1.1) obtained and seemed it did not help much in this situation.

LDA Technique:

- Built a model using the LDA technique by dividing the dataset into training and test sets.
- The class values of the predictions were obtained and compared the results with the original values:

	0	1
0	4516	999
1	152	333

- The model did a good fit thus with a mean value of 0.808 on test data set.

QDA Technique:

- Built a model using the LDA technique by dividing the dataset into training and test sets.
- The class values of the predictions were obtained and compared the results with the original values:

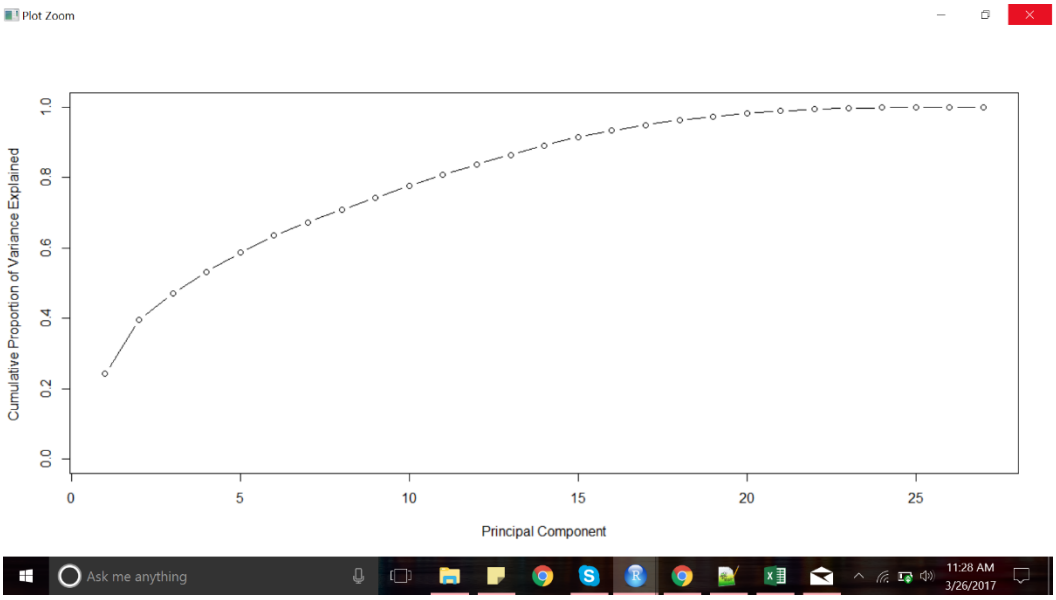
	0	1
0	1640	202
1	3028	1130

- As we can see from the values above, the model did bad in predicting the response variables. It wrongly interpreted approximately 3000 entries as payable while they were not when we compared with the actual values.
- In addition, when tried to calculate the mean value for the predictions, the result was 0.461 which is not really a good fit.

Naïve Bayes:

- Naïve Bayes is one of the very efficient techniques in predicting the binomial result for a given dataset.
- The command naivebayes() in E1071 package does built a model automatically on giving the dataset and required information.
- We implemented this technique on both training and test dataset (by dividing them in ratio of 80:20) with 24000 and 6000 entries respectively.
- The model predicted did a very good job by predicting 4050 as 0 (non-payable) and 1950 as 1 (payable).
- On further calculating the mean for the results obtained, we got it as 0.72.

Graphs:



1.1 Variance vs Principal Components

Summary:

On comparing the values obtained after building the models using various techniques, we found Naïve Bayes and LDA did a really good job in predicting the values with a mean of 0.72 and 0.808 respectively. We also observed there is not really much significant difference on removing the predictors that did not add value to the analysis. This was also proved by obtaining the PCA values.