Estimating Probability of Default & Classifying Credit Quality

Farheen, Nishaat
Kulkarni, Amruta
Zhang, Lanyin

Abstract — **Banks use credit scoring models to assign a credit score to its customers which will represent their overall creditworthiness. To achieve this score, banks use the credit history data along with the demographic details of the customer. A part of the credit scoring process is to estimate the probability of default which is nothing but the probability that the customer will not pay back the loan amount. In this project, we will be estimating the probability of a borrower defaulting in the next two years and classify a borrower into serious delinquent and non-delinquent groups.**

I. INTRODUCTION

The dataset used for this project viz 'Give Me Some Credit' is taken from Kaggle. It has 11 variables altogether which provide the borrower's details such as age, income, debt etc. The variables are described in Figure1. Seriousdlqin2 is the dependent variable and the rest of the 10 variables are the independent variables for the purposes of this project. As the number of independent variables is high, we reduce the dimensions using techniques such as Principal Component Analysis and Factor Analysis. Classification techniques such as KNN, Linear Discriminant Analysis and Logistic Regression are then run on factors. The probability of serious delinquency in the next two years is estimated using the logistic regression technique.

II. DATA PREPARATION

Before analysing the data, we need to make sure that the data is clean and as per our requirements. Upon conducting univariate analysis on all the variables, we found that the dataset has many missing values and outliers. It was also observed that many of the observations had out of range values for ratio variables. All such records were deleted from the dataset. Post data cleaning, the dataset had around 120,000 observations.

After cleaning the data, we standardized the data with mean=0 and std deviation=1 except for the binary dependent variable.

This data was then used to perform Principal Component Analysis and Factor Analysis.

III. DATA ANALYSIS

Principal Component Analysis:

Before doing principal component analysis, we would want to check the correlation between all the variables to get an idea on how the variables will be combined. As shown in Figure2, there is high correlation between the Due3059, Due6089 and Due90 variables. This is expected because borrowers who have been delinquent for 30-59 days are highly likely to be delinquent for 60-89 and 90 days too. There is also significant correlation between OCLL, RELL and Debt.

Now, Princomp procedure is run on all the independent variables. Figure3 shows the eigen values obtained. And the scree plot is shown in Figure 4. There are four principal components whose eigen value is more than 1. And the variance explained by these four components is 74.7% which is satisfactory. Based on this, we have decided to use four factors for further analysis.

The four principal components are plotted with respect to the independent variables to see which variables are getting combined into which component. Figure 5 shows that Due3059, Due6089 and Due 90 load high on PC1 and OCLL, RELL and Debt load high on PC2 which is in accordance with the inference obtained from the correlation matrix. Figure 6 shows the PC3 and PC4 plot. There is no significant pattern in this plot because remaining variance left after PC1 and PC2 are formed, is getting combined in these components.

With this information in hand, we will next run factor analysis.

Factor Analysis:
Proc factor is run on the data such that four factors will be formed. The output is shown in figures 7-9.
Figure 7 shows the communality estimates for the four factors. As expected, the four factors explain 74.7% of the total variance of the independent variables. And at least 50% of every independent variable is getting explained by the four factors with most of the variables getting explained to the extent of 80-90%. This result is satisfactory and it indicates that we can use these factors for further analysis of the data instead of using all the original variables.

Next, we try the various rotation methods to check if the factors can be improved. Figure8 shows the results of the no rotation, varimax rotation and the quartimin rotation. Most of the variance is getting explained by factor1 and the distribution of variance did not improve significantly after using either varimax or quartimin rotation methods. As orthogonality is an important characteristic, we select the varimax output and carry forward the analysis with it.

Figure9 shows the rotated factor pattern for the varimax method. The loadings > 0.5 are highlighted to indicate which variables score high on which factor.

Variables Due3059, Due6089 and Due90 load high on Factor1. This means that this factor causes to the borrower to be more and more delinquent. So, this factor has been named 'Financial Distress'.

Variables Debt, OCLL and RELL load high on Factor2. This means that this factor causes the borrower to take up more debt and more lines of credit. So, this factor has been named as 'Funding Requirements'

Variables Age, Depend and RUL load high on Factor3 with Depend and RUL being negatively related to Factor3. This factor has been named as 'Expendable Income' because as age increases, expendable income increases, as number of dependents increases, expendable income decreases and as utilization of the credit line increases, interest payments increase and thus expendable income decreases.

Variables Income and Depend load high on Factor4. This factor has been named as Lifestyle because as income increases lifestyle improves and as number of dependents increase, it implies that the borrower has a family.

Splitting the Data:
Once the factors were obtained, the SAS output dataset was split into 70-30% ratio to get training and testing data respectively. We used random sampling to do the split.

Clustering and Classification:
Firstly, cluster analysis was done to get an idea of how well the data was clustered. Post which, KNN, LDA and logistic regression models were created using the training data.

Lastly, testing data was used to test the accuracy of the classification models created.

Cluster Analysis
K means cluster analysis is performed on the data set using the four factors extracted from the ten original predictor variables.

The data set includes two types of customers:
1. Those who will face serious delinquency in next two years (Sdl2 = 1)
2. Those who will not face any serious delinquency in next two years (Sdl2 = 0)

Therefore, we decided to see if the factors can divide the data set into two clusters. The output is shown in figures 10-11.

Our data set contains 93% customers with no serious delinquency and 7% customers with serious delinquency, however the four factors divide 99.87% customers in one cluster and 0.13% customers in another cluster.

K Nearest Neighbour:
In KNN, a customer is classified by a majority vote of its neighbours, with the customer being assigned to the class most common among its K nearest neighbours.

While choosing the value of number of neighbours, we tried to minimize the misclassification of Sdl2 = 1 as Sdl2 = 0. Because for a bank or credit card company, classifying a customer who will default as not default will have more impact than classifying a customer who will not default as default. We obtained the best result for K = 20.

The output for the KNN model is shown in Figure12. In training data, the model classified 79.2% of the customers with no serious delinquency correctly and 81.35% of the customers with serious delinquency correctly. Overall accuracy on training data is 80.27%.

In the test data, the model classified 78.31% of the customers with no serious delinquency correctly and 66.37% of the customers with serious delinquency correctly. Overall accuracy on test data is 71.84%.

As the difference between the accuracy of training data and testing data is high, we can say that the model overtrained.

Linear Discriminant Analysis:
From the cluster analysis, we confirmed that the data is clustered and the two clusters are well separated. As LDA works on the principle of increasing the distance between the means of the two groups, we expect this technique to perform well.

Proc discrim is run to get the LDA output which is shown in figures 13-14. The probability of Mahalanobis Distance being zero is less than 0.5, therefore null hypothesis that "The squared distance between means of two groups is zero" is rejected. Further discriminant analysis shows that all the factors are significant in classifying the customers.

Linear Discriminant Functions:
Equation for group 0 (no serious delinquency) = -0.00237 - 0.03037*Factor1 – 0.00421*Factor2 + 0.06108*Factor3 + 0.01939*Factor4

Equation for group 1 (serious delinquency) =
- 0.43442 + 0.41572*Factor1 + 0.08205*Factor2 - 0.83849*Factor3 - 0.20242*Factor4

Interpretation:
The coefficients of the factors suggest that, higher values of Factor 1 and Factor 2 and lower values of Factor 3 and Factor 4 lead to classification into group 1.

High values of financial distress (number of times person has been late for credit payments) and Funding requirements (debt ratio and number of open loans or credit lines and low values of expendable income (young age and high revolving utilization of unsecured lines) and lifestyle (low income and less number of dependents) lead to classifying a customer into 'Serious Delinquency' group.

Accuracy of LDA on training data is 70.94% while that on test data is 69.5%

Logistic Regression
Since the dependent variable is binary, we expect logistic regression to perform well. Logistic regression will be used to:

- Classify the Sdl2 into two group
- Test which factors are significant
- Find the probability of serious delinquency in the next two years

Logistic regression model is created using the training data. As shown in figure16, the p-value indicates that all the coefficients are significant. And as shown in figure17, the chi-square of the model is significant as well. All the tests show that the model is good, now we dig into the meaning of the model.

Equations & Interpretation

$$\log\left(\frac{p}{1-p}\right) = -2.92 + 0.155 * Factor1 + 0.051 * Factor2 - 0.858 * Factor3 - 0.157 * Factor4$$

As we mentioned before, the Factor1 means Financial Distress, which consists of three due time, the Factor2 means Funding Requirement, Factor3 means the Expendable Income, and Factor4 refers to the Lifestyle.

The log of odds ratio is positively related to factor1 and factor2. This makes sense because it is reasonable to expect that when a borrower undergoes financial distress and/or his funding requirement increases, the probability of him being seriously delinquent would also increase.

The log of odds ratio is negatively related to factor 3 and factor 4. This also complies with our factors' definition. As expendable income and lifestyle improve, the probability of serious delinquency decreases.

Odds Ratios

However, log of probability is not a good way to interpret the coefficients. In order to interpret the output better, we calculate the odds ratio of the model under different confident interval.

The 95% confidence interval (CI) is used to estimate the precision of the OR. OR>1 means exposure associated with higher odds of outcome. OR<1 means exposure associated with lower odds of outcome. The output is shown in Figure18. The odds ratio of factor 1 is 1.17, which means that when the factor 1 increase 1 unit, the odd of someone default vs someone not default will increase 1.17.

Meanwhile, there are different odds ratios calculated according to different confident interval. A large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR.

ROC Curves

ROC curves show the accuracy of the model, we obtain the ROC curves for training data and testing data as shown in figure19-20. The AUC values of both are pretty high, which means our model works well.

IV. CONCLUSION

All the three classification techniques are compared in the following table. KNN model seems to have got overtrained. However, both LDA and Logistic regression have performed similarly for both training and testing data. Between LDA and logistic regression, logistic regression has performed better. This could be because the dependent variable is binary and logistic regression would have modelled it better.

Classification Model	Accuracy on training data	Accuracy on test data
KNN	80.27%	71.84%
Linear Discriminant Analysis	70.94%	69.5%
Logistic Regression	75.4%	74.2%

Hence, we can recommend the bank to use the logistic regression model both for estimating the probability of serious delinquency and for classifying borrowers into non delinquents and delinquents.

V. APPENDIX

Figure 1: Variables Description

Variable Name	Description	Type
SeriousDlqin2yrs	**Person experienced 90 days past due delinquency or worse**	**Y/N**
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Figure2: Correlation Matrix

Pearson Correlation Coefficients, N = 116594
Prob > |r| under H0: Rho=0

	RUL	AgeY	Due3059	Debt	Income	OCLL	Due90	RELL	Due6089	Depend
RUL	1.00000	-0.26273 <.0001	0.11385 <.0001	0.09995 <.0001	-0.08761 <.0001	-0.17250 <.0001	0.09983 <.0001	-0.08186 <.0001	0.08481 <.0001	0.08288 <.0001
AgeY	-0.26273 <.0001	1.00000	-0.04996 <.0001	-0.04242 <.0001	0.09110 <.0001	0.18474 <.0001	-0.04844 <.0001	0.06386 <.0001	-0.04405 <.0001	-0.20581 <.0001
Due3059	0.11385 <.0001	-0.04996 <.0001	1.00000	-0.01447 <.0001	-0.02769 <.0001	-0.04893 <.0001	0.97598 <.0001	-0.02679 <.0001	0.98074 <.0001	0.00339 0.2466
Debt	0.09995 <.0001	-0.04242 <.0001	-0.01447 <.0001	1.00000	-0.15454 <.0001	0.23555 <.0001	-0.03032 <.0001	0.36818 <.0001	-0.02751 <.0001	0.04195 <.0001
Income	-0.08761 <.0001	0.09110 <.0001	-0.02769 <.0001	-0.15454 <.0001	1.00000	0.22942 <.0001	-0.03391 <.0001	0.31151 <.0001	-0.03000 <.0001	0.16176 <.0001
OCLL	-0.17250 <.0001	0.18474 <.0001	-0.04893 <.0001	0.23555 <.0001	0.22942 <.0001	1.00000	-0.07799 <.0001	0.43090 <.0001	-0.06683 <.0001	0.04657 <.0001
Due90	0.09983 <.0001	-0.04844 <.0001	0.97598 <.0001	-0.03032 <.0001	-0.03391 <.0001	-0.07799 <.0001	1.00000	-0.04306 <.0001	0.98958 <.0001	-0.00398 0.1740
RELL	-0.08186 <.0001	0.06386 <.0001	-0.02679 <.0001	0.36818 <.0001	0.31151 <.0001	0.43090 <.0001	-0.04306 <.0001	1.00000	-0.03635 <.0001	0.12601 <.0001
Due6089	0.08481 <.0001	-0.04405 <.0001	0.98074 <.0001	-0.02751 <.0001	-0.03000 <.0001	-0.06683 <.0001	0.98958 <.0001	-0.03635 <.0001	1.00000	-0.00524 0.0736
Depend	0.08288 <.0001	-0.20581 <.0001	0.00339 0.2466	0.04195 <.0001	0.16176 <.0001	0.04657 <.0001	-0.00398 0.1740	0.12601 <.0001	-0.00524 0.0736	1.00000

Figure 3: PCA Output - Eigenvalues

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.00774080	1.15898348	0.3008	0.3008
2	1.84875732	0.42869024	0.1849	0.4856
3	1.42006708	0.22624065	0.1420	0.6277
4	1.19382643	0.39459420	0.1194	0.7470
5	0.79923223	0.11578808	0.0799	0.8270
6	0.68344414	0.09649613	0.0683	0.8953
7	0.58694802	0.16171555	0.0587	0.9540
8	0.42523247	0.40021190	0.0425	0.9965
9	0.02502057	0.01528963	0.0025	0.9990
10	0.00973095		0.0010	1.0000

Figure 4: PCA Output – Scree Plot

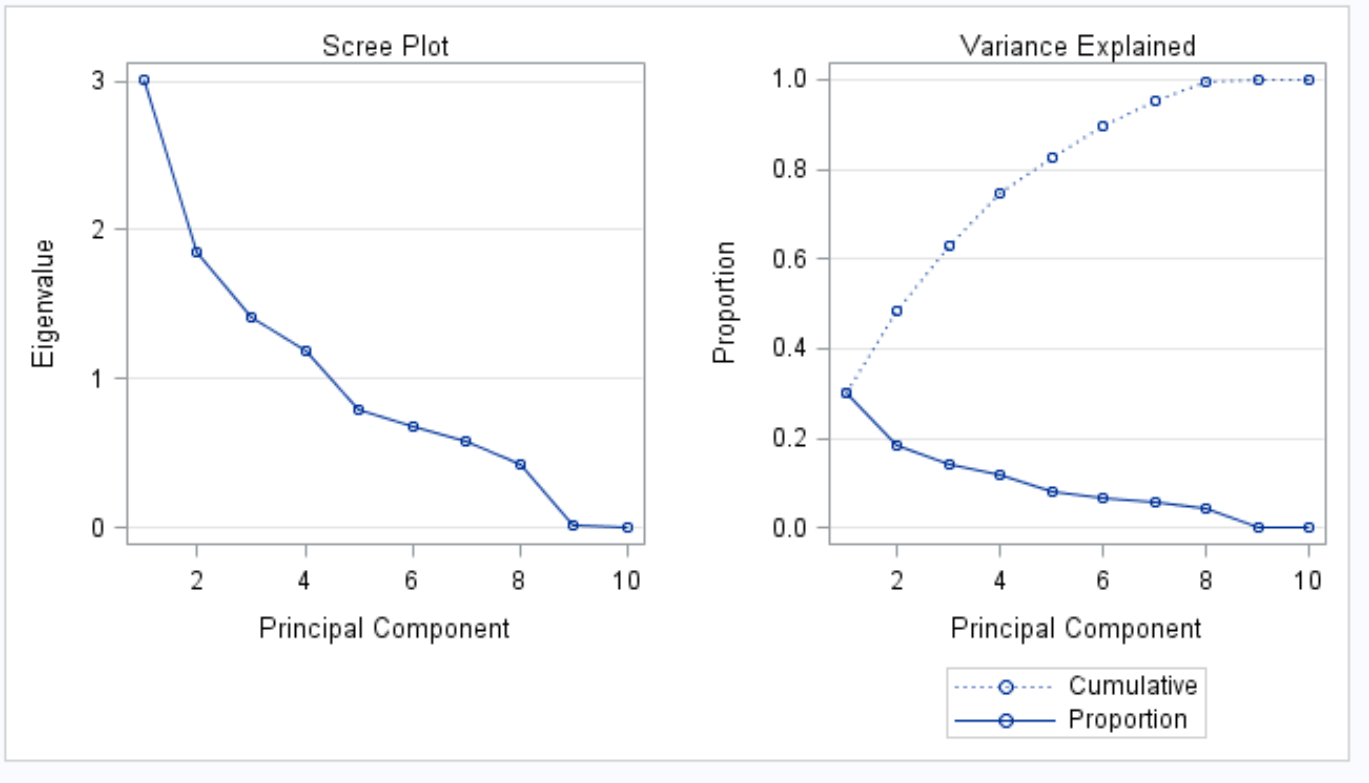

Figure5: PC1 vs PC2

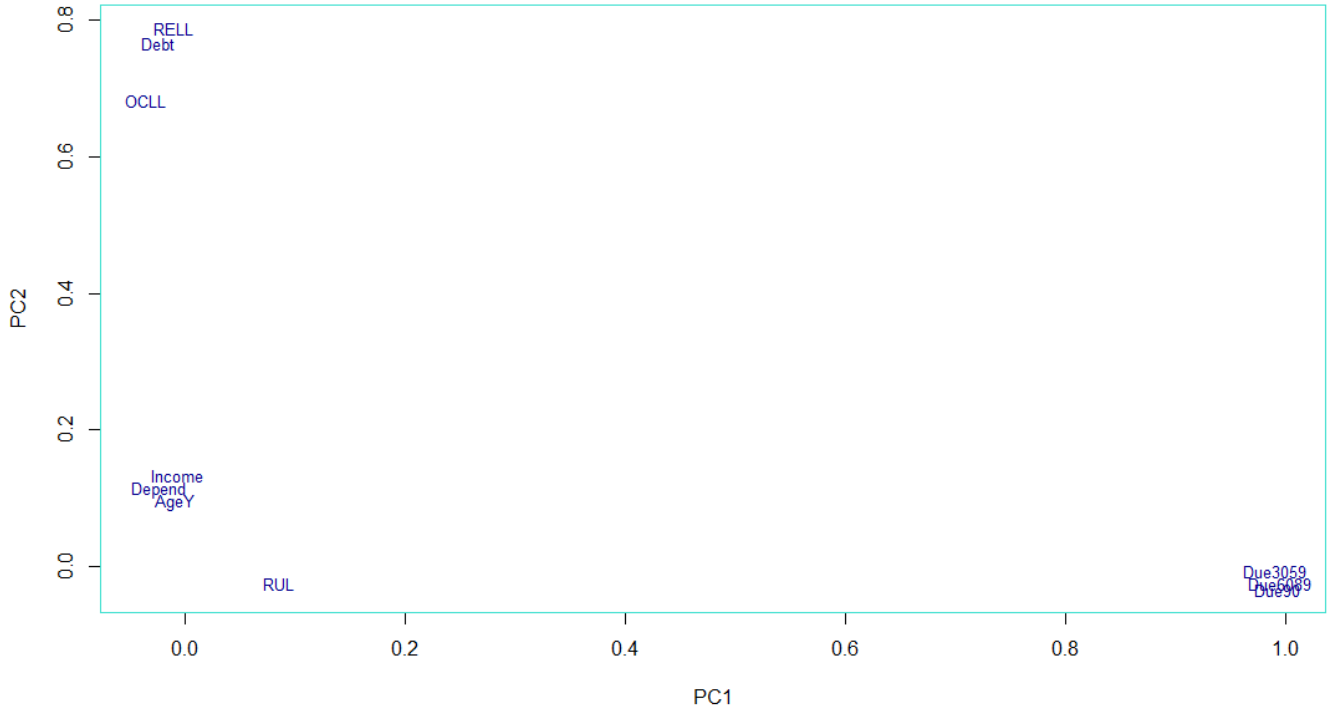

Figure6: PC3 vs PC4

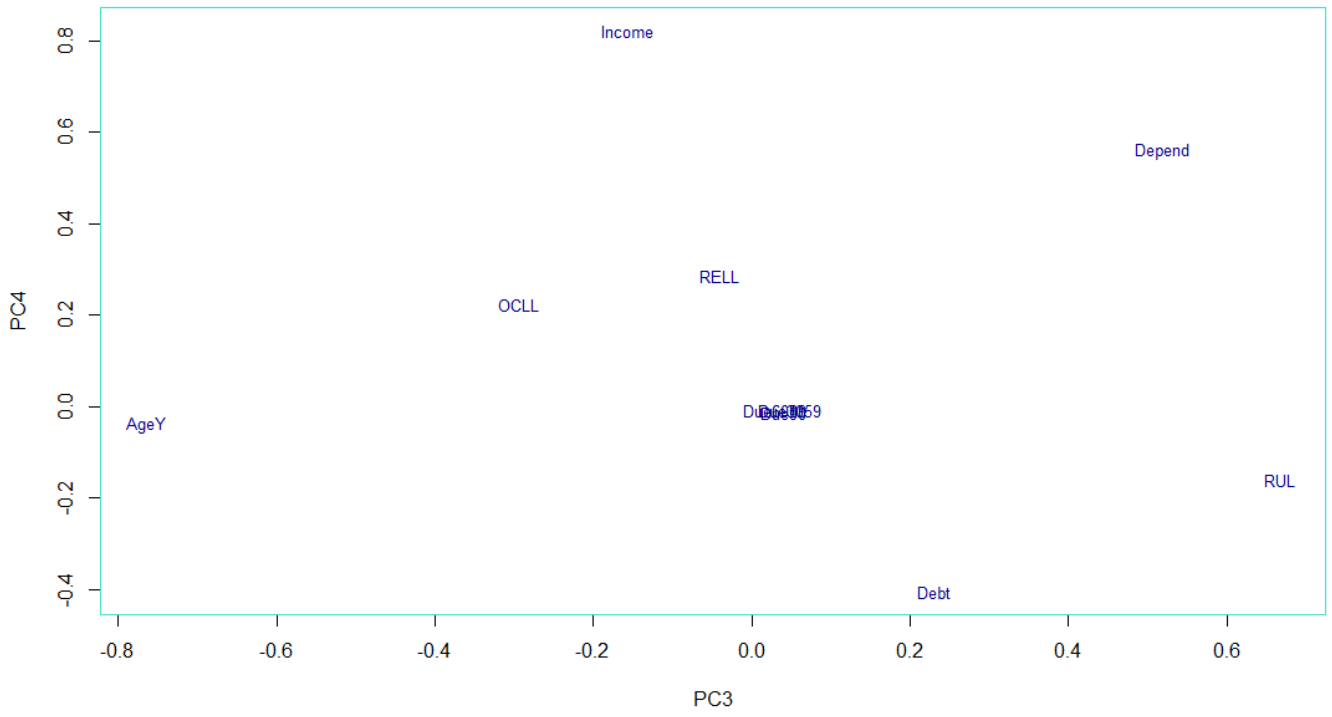

Figure7: Factor Analysis Output – Communality Estimates

Final Communality Estimates: Total = 7.470392									
RUL	AgeY	Due3059	Debt	Income	OCLL	Due90	RELL	Due6089	Depend
0.47945425	0.59519885	0.98331980	0.80572604	0.72036031	0.60506118	0.98863788	0.70645860	0.99137534	0.59479937

Figure8: Factor Analysis – Rotation Methods

No rotation

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
3.0077408	1.8487573	1.4200671	1.1938264

Varimax Rotation

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
2.9659062	1.7191921	1.4718141	1.3134792

Quartimin Rotation

Variance Explained by Each Factor Eliminating Other Factors			
Factor1	Factor2	Factor3	Factor4
2.9475478	1.6734089	1.4509066	1.2866476

Figure9: Factor Analysis – Rotated Factor Pattern

Rotated Factor Pattern				
	Factor1	Factor2	Factor3	Factor4
RUL	0.08555	-0.02406	-0.66798	-0.15927
AgeY	-0.00937	0.09429	0.76467	-0.03873
Due3059	0.99030	-0.00694	-0.05008	-0.00809
Debt	-0.02422	0.76704	-0.23242	-0.40346
Income	-0.00709	0.13338	0.15806	0.82313
OCLL	-0.03567	0.68335	0.29371	0.22484
Due90	0.99275	-0.03388	-0.04224	-0.01269
RELL	-0.01004	0.78880	0.04119	0.28716
Due6089	0.99483	-0.02553	-0.03099	-0.00879
Depend	-0.02370	0.11212	-0.51835	0.55945

Figure10: Confusion Matrix for clusters

Frequency Percent Row Pct Col Pct	Table of CLUSTER by Sdl2		
	Sdl2		
CLUSTER(Cluster)	0	1	Total
1	108373 92.95 93.07 99.94	8073 6.92 6.93 98.97	116446 99.87
2	64 0.05 43.24 0.06	84 0.07 56.76 1.03	148 0.13
Total	108437 93.00	8157 7.00	116594 100.00

Figure11: Plot of clusters

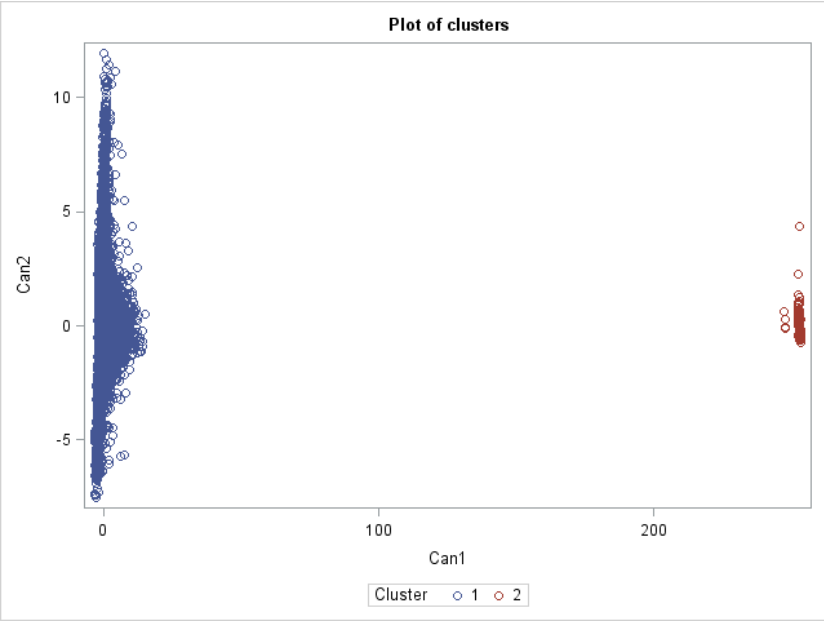

Figure12: Result of KNN for training and test data

Number of Observations and Percent Classified into SdI2			
From SdI2	0	1	Total
0	60136 79.20	15792 20.80	75928 100.00
1	1061 18.65	4627 81.35	5688 100.00
Total	61197 74.98	20419 25.02	81616 100.00
Priors	0.5	0.5	

Error Count Estimates for SdI2			
	0	1	Total
Rate	0.2080	0.1865	0.1973
Priors	0.5000	0.5000	

Number of Observations and Percent Classified into SdI2			
From SdI2	0	1	Total
0	25457 78.31	7052 21.69	32509 100.00
1	855 34.63	1614 65.37	2469 100.00
Total	26312 75.22	8666 24.78	34978 100.00
Priors	0.5	0.5	

Error Count Estimates for SdI2			
	0	1	Total
Rate	0.2169	0.3463	0.2816
Priors	0.5000	0.5000	

Figure13: Mahalanobis Distance

Squared Distance to SdI2		
From SdI2	0	1
0	0	1.00172
1	1.00172	0

Prob > Mahalanobis Distance for Squared Distance to SdI2		
From SdI2	0	1
0	1.0000	<.0001
1	<.0001	1.0000

Figure14: Standardized coefficients from LDA

Linear Discriminant Function for SdI2		
Variable	0	1
Constant	-0.00237	-0.43442
Factor1	-0.03037	0.41572
Factor2	-0.00421	0.08205
Factor3	0.06108	-0.83849
Factor4	0.01939	-0.20242

Figure15: Result of LDA for training and test data

Number of Observations and Percent Classified into Sdl2			
From Sdl2	0	1	Total
0	54883 72.28	21045 27.72	75928 100.00
1	1729 30.40	3959 69.60	5688 100.00
Total	56612 69.36	25004 30.64	81616 100.00
Priors	0.5	0.5	

Error Count Estimates for Sdl2			
	0	1	Total
Rate	0.2772	0.3040	0.2906
Priors	0.5000	0.5000	

Number of Observations and Percent Classified into Sdl2			
From Sdl2	0	1	Total
0	23398 71.97	9111 28.03	32509 100.00
1	814 32.97	1655 67.03	2469 100.00
Total	24212 69.22	10766 30.78	34978 100.00
Priors	0.5	0.5	

Error Count Estimates for Sdl2			
	0	1	Total
Rate	0.2803	0.3297	0.3050
Priors	0.5000	0.5000	

Figure16: Significance of factors in Logistic Regression

```
glm(formula = cs_training$Sdl2 ~ cs_training$Factor1 + cs_training$Factor2 +
    cs_training$Factor3 + cs_training$Factor4, family = binomial(link = "logit"),
    data = cs_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3908  -0.4034  -0.3000  -0.2133   3.2420

Coefficients:
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.921608   0.017610 -165.910  < 2e-16 ***
cs_training$Factor1  0.155101   0.009832   15.776  < 2e-16 ***
cs_training$Factor2  0.050952   0.012431    4.099 4.16e-05 ***
cs_training$Factor3 -0.857788   0.014979  -57.265  < 2e-16 ***
cs_training$Factor4 -0.156829   0.013929  -11.259  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41272  on 81615  degrees of freedom
Residual deviance: 36970  on 81611  degrees of freedom
AIC: 36980
```

Figure17: Chi-square of the model

```
                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                81615      41272
cs_training$Factor1  1    379.9     81614      40892 < 2.2e-16 ***
cs_training$Factor2  1     36.4     81613      40856 1.582e-09 ***
cs_training$Factor3  1   3753.4     81612      37102 < 2.2e-16 ***
cs_training$Factor4  1    132.2     81611      36970 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure18: Odds ratios and 95% CI

```
## odds ratios and 95% CI
exp(cbind(OR = coef(glmreg), confint(glmreg)))
```

```
## Waiting for profiling to be done...
```

```
##                                OR      2.5 %     97.5 %
## (Intercept)          0.05384703 0.05200943 0.05572651
## cs_training$Factor1  1.16777595 1.14664035 1.19198559
## cs_training$Factor2  1.05227206 1.02689887 1.07818220
## cs_training$Factor3  0.42409895 0.41179590 0.43670053
## cs_training$Factor4  0.85485013 0.83174650 0.87842461
```

Figure19: ROC for the training data

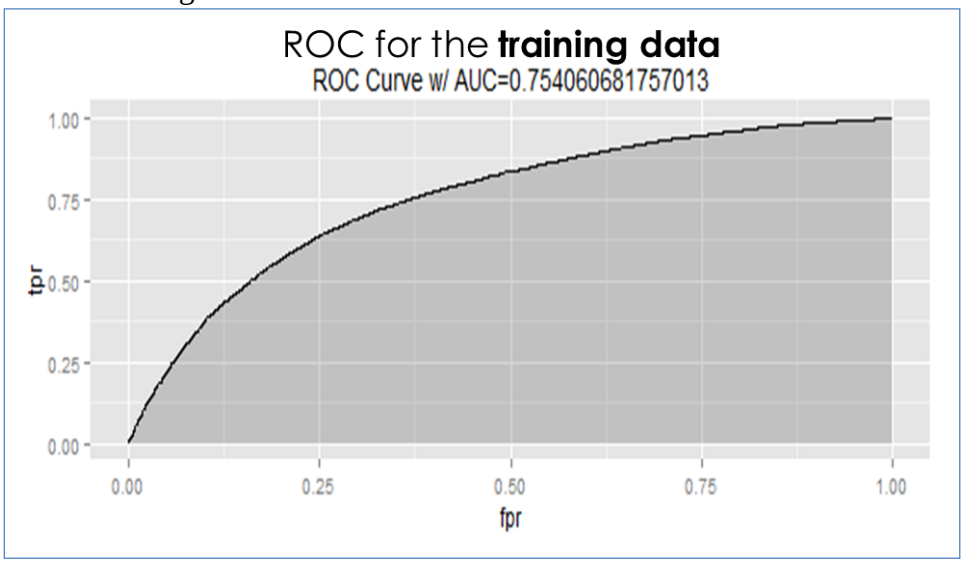

Figure20: ROC for the testing data