HIGHER INSTITUTE OF MANAGEMENT OF TUNIS

Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years.

ECONOMETRICS

Research master degree in modeling of information system and decision making

By
Omri Soufiane
Oussema Ben Brahem

Under the supervision of:
Prof. Mohammed Kriaa

Table of Contents

List of the Figures

1. Introduction

Banks present one the main building blocks of today's economy. In fact Banks, manages money flow for the world population thus can either give or deny access to such resources. Loans, which are a major part of a banks daily operation are very delicate thus requires a high level treatment to determine who will get a loan and who's going to be denied one. Here rises the term credit scoring which is a set of algorithm based on mathematical and statistical method to predict or guess the probability of default, as to determine loan worthiness.

This report describes set of tools and procedure used to improve the Credit scoring algorithms through building a model that financial institution can use to make the best financial decisions. This developed model is based on prior knowledge of financial distress that some of early borrowers encountered.

2. Description of Database

This database is provided by the kaggle team as a part of competition that included 925 teams. In fact, the data was provided by actual banks who are facing issue of accuracy in regard to scoring. Historical data are provided on 150,000 borrowers this data is divide into a training set of 120000, around 80%, and a test (validation) set of 30000 observation around 20%. The goal is to use the training set to develop a model and the validation set to test the prediction quality of the model. Through our analysis we will train the model on the training set then validate it on the test set in order to diagnose its accuracy. In addition, to make the utmost of this study we will be experimenting with different model ranging from cluster analysis, logistic regression , discriminant analysis, Probit. Below the data dictionary of the different variables

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 monthly 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

Table 1: Data Dictionary

3. Methodology

There are many techniques that have been used in the credit scoring industry including logistic regression, mathematical programming, and Markov chain models. In our experiment we will use discriminant analysis and the logistic regression given that tow discrete classes (YES/NO) have been identified. The experiment will be conducted using STATA-programming.

We have data provided on 150,000 borrowers divided into a training set (120000) and validation set (30000) using the 80-20% rule, the goal is to use the training set to develop a model and the validation set to test the prediction quality of the model.

During our analysis we have proceeded through four main steps; data cleaning and checking, verification of assumptions, model diagnostics and potential follow-up analyses. We have developed different model including a logit and probit model which are similar thus probit will be reported in the appendix. Furthermore, we used discriminant analysis model and cluster analysis as exploratory tools to determine whether the given data variable can classify or distinguish correctly the different classes.

Finally we have used a set of test and graphs to test each model including tests for global fit for the regression model and Roc analysis on both training and validation set. In addition we used only the roc to determine the accuracy of the discriminant analysis.

4. Exploratory Analysis

4.1. Data cleaning and checking

4.1.1. Missing data

Upon importing the data to STATA we have noted red column this due to the absence of data which is labeled by NA, thus STATA have labeled the variables **monthlyincome** and **numberofdependents** as string. Converting the late variables to numeric STATA have detected the presences of 29731 missing values for the **monthlyincome** variable and 3924 missing values for the **numberofdependents** variable. This further showed through the command INSPECT

```
monthlyincome:  MonthlyIncome                    Number of Observations
_______________________________             ________________________________
                                             Total   Integers   Nonintegers
 |  #                        Negative            -          -             -
 |  #                        Zero             1634       1634             -
 |  #                        Positive       118635     118635             -
 |  #                                       ______     ______        ______
 |  #                        Total          120269     120269             -
 |  #   .   .   .   .        Missing         29731
 +-----------------------                   ______
0                3008750                    150000
(More than 99 unique values)
```

Figure 1: Monthlyincome Description

```
numberofdependents:  NumberOfDependents         Number of Observations
_______________________________________    ______________________________
                                            Total   Integers   Nonintegers
 |  #                       Negative            -          -             -
 |  #                       Zero            86902      86902             -
 |  #                       Positive        59174      59174             -
 |  #                                      ______     ______        ______
 |  #                       Total          146076     146076             -
 |  #   .   .   .   .       Missing          3924
 +----------------------                   ______
0                     20                   150000
  (13 unique values)
```

Figure 2: Numberofdependents

4.1.2. Dealing with outliers

Using the extremes command in stat to detect the maximum and lower values in a given variable. Here we mainly focus on late payment in a certain period.

```
. extremes numberoftimes90dayslate numberoftime6089dayspastduenotwo numberoftime3059dayspastduenotwo
```

obs:	number~e	n~6089~o	n~3059~o
1.	0	0	2
2.	0	0	0
4.	0	0	0
5.	0	0	1
6.	0	0	0

147775.	98	98	98
149154.	98	98	98
149240.	98	98	98
149440.	98	98	98
149770.	98	98	98

```
note: 141662 values of 0
note: 264 values of 98
```

Figure 3: Extremes

A clear analysis of the above outcome show the existence of 264 values of 98 for the three variables; "NumberOfTime30-59DaysPastDueNotWorse"," NumberOfTime60-89DaysPastDueNotWorse" and" NumberOfTimes90DaysLate". In addition we have found a few values of 96.

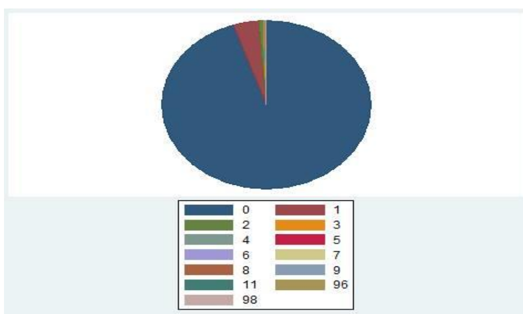

Figure 4: outlier pie chart

After a thorough search we have come to the conclusion that this value have a qualitative explanation rather than quantitative. In fact, the 98 means the interviewee did not provide an answer while 96 for other reason data was not available. During our analysis such extreme value may have a huge impact on our models thus we will eliminate and consider this data as missing and leave its treatment for stata.

Proceeding to the "RevolvingUtilizationOfUnsecuredLines" variable. In fact, revolving utilization, also known as your "debt-to-limit ratio" or "credit utilization," measures the amount of your revolving credit limits that you are currently using. Based on the proceeding definition we have identified extreme value to be true for the analysis as shown below by stat

```
+--------------------+

149280.      20514

149161.      22000

16957.      22198

31415.      29110

85490.      50708
```

Here we have the first column representing the observation number and the second its corresponding value which are considered very high even if allowed to pass limit for this reason and given that the variables is a proportion "percentage" the rule will be as follow if the data is superior to 2.5 then would be considered as missing.

```
. replace revolvingutilizationofunsecuredl=. if ( revolvingutilizationofunsecuredl>=2.5)
(311 real changes made, 311 to missing)
```

Exploring the other variables we came on 0 ages which is non logical as shown below

```
. extremes age
```

obs:	age
65696.	0
1732.	21
2792.	21
3369.	21
3717.	21

Figure 5: Extremes age

Of course this extreme value will be dropped in the analysis or rather replaced to missing for better analysis

```
. replace age=. if ( age<10)
(1 real change made, 1 to missing)
```

Furthermore, exploring the debt ratio and the monthly income we found extreme value for the debt ratio, tracking those people down we found that they hold an income of 0, missing or 1. This is either due to wrong estimate of income used for the computation of the debt ratio or the value were imputed by 1 this for missing and 0 to avoid division by such values thus resulted in the extreme values rises in debt ratio. Of course for the debt ratio value that are due to missing data will be excluded stata without a need to label them. Note we will drop debt ratio value that are proportional to a missing income.

4.2. Univariate Profiling: examining the variables distribution

4.2.1. Descriptive statistic

The starting point for understanding the nature of any variable is to characterize the shape of its distribution. Using the summarize command we have got the bellow table.

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
seriousdlq~s	150000	.06684	.2497455	0	1
revolvingu~l	149689	.3210776	.3573331	0	2.494658
age	149999	52.29556	14.7713	21	109
n~3059days~o	149731	.2457941	.6977798	0	13
debtratio	120269	26.59878	424.4465	0	61106.5
monthlyinc~e	120269	6670.221	14384.67	0	3008750
numberofop~s	150000	8.45276	5.145951	0	58
numberofti~e	149731	.0904556	.4855273	0	17
numberreal~s	150000	1.01824	1.129771	0	54
n~6089days~o	149731	.0648229	.3300732	0	11
numberofde~s	146076	.7572223	1.115086	0	20

Figure 6: Descriptive statistic table

The above table contain main feature including the mean, Standard deviation and the min and max value. We can note tow high Standard dev which are the monthly income and the debt ratio. Also, the mean of serious very low due to the highest number of refused loans 93.23 % as shown by the figure 6.

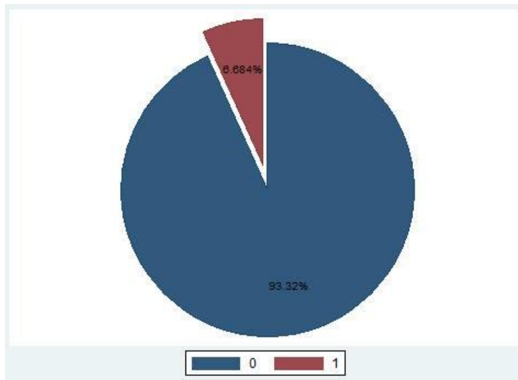

Figure 7: seriousdlinq pie chart

4.2.2. **Distribution description**

The main tool for a researcher to understand the shape of the distribution is the histogram. In this section a set of histogram of the variables is presented below

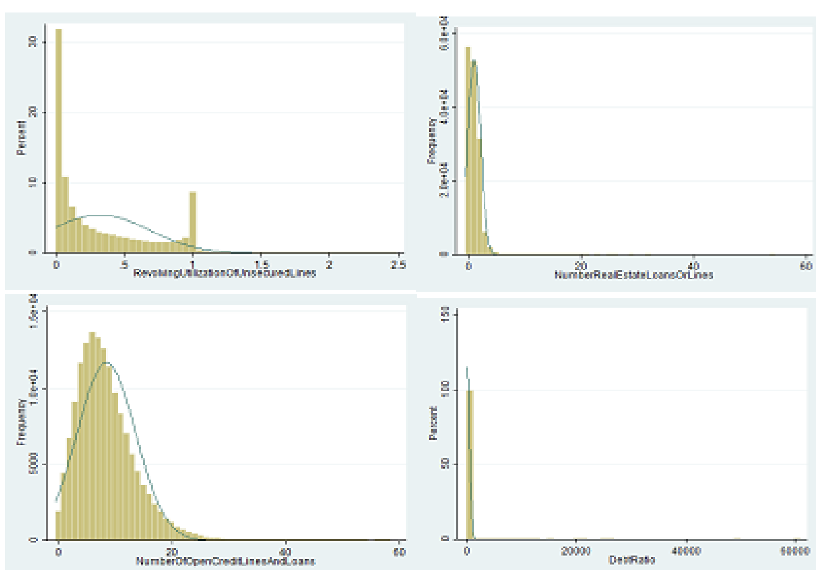

Figure 8: Histogram

The height of the bars represent frequencies of data values within each category. The normal curve is also superimposed on the distribution. Most of the variable above are within the bell shaped line, but we can note some skewness for the numberofopencredit and shortage of values for the revolving utilization in the middle of the distribution. In fact, we can say that the distrubtions are normally distributed this will be further confirmed by the shapiro wilk test.

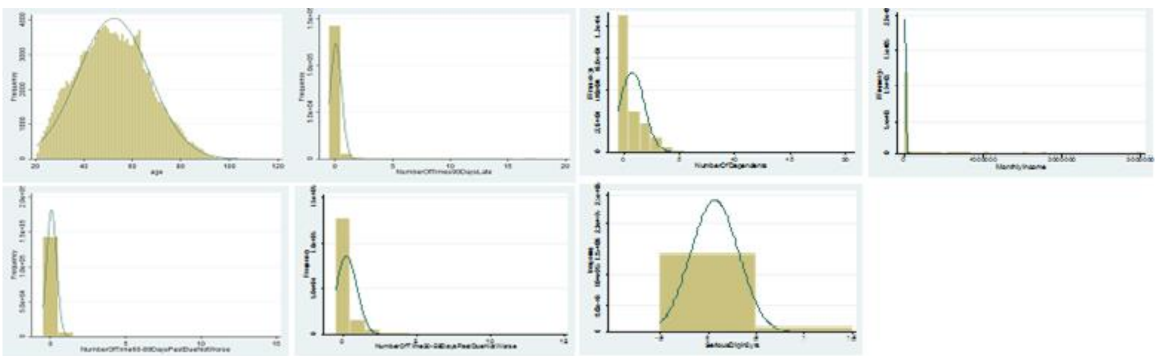

Figure 9: Histogram continued

The variable age seems perfectly normal with all the data within the normal curve. In addition we can note some skewness but within the normal distribution curve except for the binary variable in which also 93.5% of the data are within the curve. For further clearance in regard to the shape of the distribution we have below the shapiro wilk test table which proves that all variable are normally distributed this can be further explained by the huge number of observation in our sample about 150000 observations.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
seriousdlq~s	150000	0.99981	7.675	5.739	0.00000
revolvingu~l	149689	0.83799	6486.762	24.715	0.00000
age	149999	0.99155	338.797	16.403	0.00000
n~3059days~o	149731	0.87025	5196.086	24.090	0.00000
debtratio	120269	0.03375	3.3e+04	29.264	0.00000
monthlyinc~e	120269	0.13041	3.0e+04	28.968	0.00000
numberofop~s	150000	0.93840	2469.941	21.997	0.00000
numberofti~e	149731	0.76343	9473.930	25.781	0.00000
numberreal~s	150000	0.89200	4330.205	23.577	0.00000
n~6089days~o	149731	0.87008	5202.911	24.094	0.00000
numberofde~s	146076	0.97084	1149.226	19.837	0.00000

Figure 10: Shapiro wilk normality test

4.3. Bivariate Profiling: examining the relationship between variables

The most popular method for examining bivariate relationship is the scatterplot, a graph of data points based on tow metric variables one on the X axis the other on the Y. in this part we will use the scatterplot matrix and the correlation matrix to examine bivariate relationships.

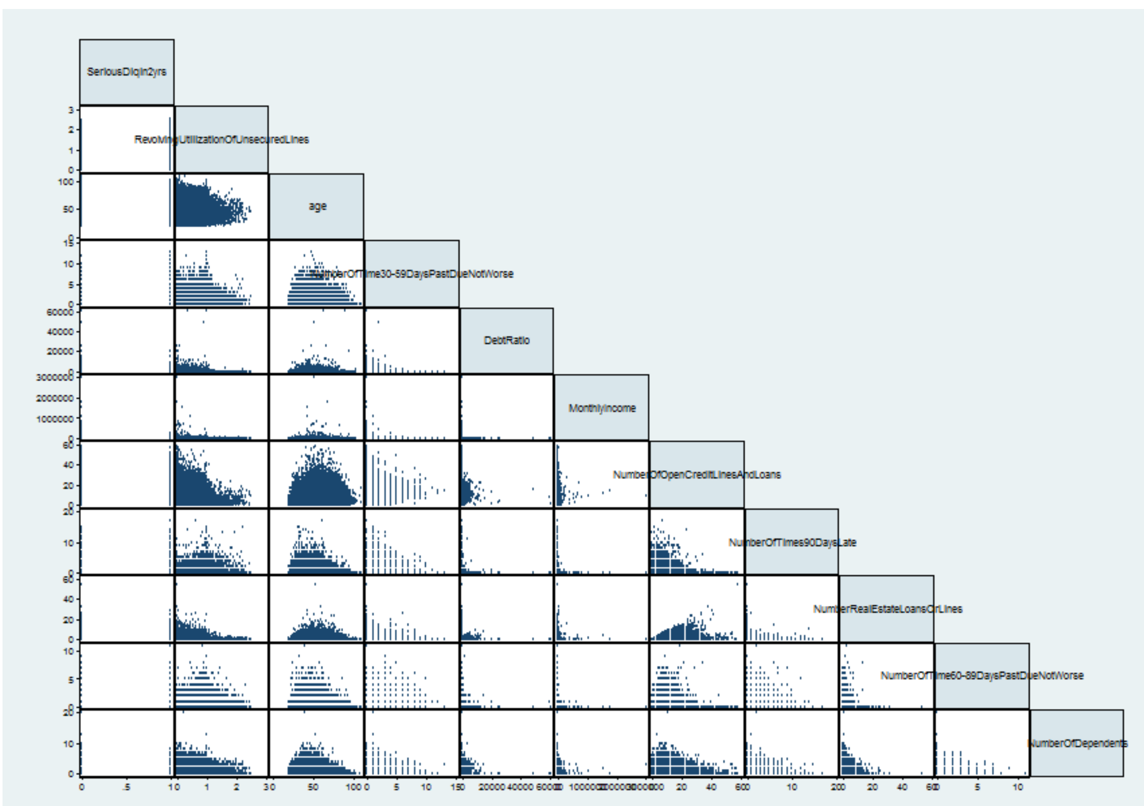

Figure 11: scatter plot matrix

By a mere look at the scatter plot matrix we can note that some variables exhibit the same behavior, this is clear for the "debtratio" and "monthlyincome" variables. In fact the debt ratio equals debt divided by monthly income which explain the similarity. In addition, delay variables "90 days delay", "30-59 delay" and "60-89 delay" also have similar distribution and relation with the other variables. In fact, we can only note very weak relationship between the variables which is further confirmed by the correlation matrix (figure 11) in fact the highest correlation can be noted between "numberofrealestate" and "numberofopencredit" about 0.4 followed by 0.3 for the "90 days delay" and "60-89", "30-59" and "60-89 days delay" and finally between a "90 days delay" and decision of assigning a loan "seriousdlinq2yrs"

```
. spearman, stats(rho)
(obs=95929)
```

	serious~s	revolv~l	age	n~3059~o	debtra~o	monthl~e	numbe~ns	number~e	numbe~es	n~6089~o	numbe~ts
seriousdlq~s	1.0000										
revolvingu~l	0.2318	1.0000									
age	-0.1012	-0.2573	1.0000								
n~3059days~o	0.2456	0.2288	-0.0745	1.0000							
debtratio	0.0562	0.1997	-0.0652	0.0996	1.0000						
monthlyinc~e	-0.0639	-0.0793	0.1336	-0.0085	-0.1336	1.0000					
numberofop~s	-0.0323	-0.1004	0.2025	0.0634	0.3882	0.3111	1.0000				
numberofti~e	0.3264	0.2244	-0.0884	0.2389	-0.0213	-0.0810	-0.1230	1.0000			
numberreal~s	-0.0308	-0.0451	0.0955	0.0187	0.5850	0.3892	0.4618	-0.0950	1.0000		
n~6089days~o	0.2588	0.1751	-0.0714	0.2608	0.0311	-0.0472	-0.0385	0.2935	-0.0384	1.0000	
numberofde~s	0.0461	0.1075	-0.2171	0.0676	0.1132	0.2021	0.0658	0.0368	0.1592	0.0386	1.0000

Figure 12: Correlation matrix

Based on the above analysis of the correlation and scatter plot matrices we may conclude that the data does not exhibit a co-linearity issue and thus we may proceed to building our models.

4.4. Discriminant Analysis

In this section we are using the DA as an exploratory approach. DFA takes a similar approach to the PCA but seeks a function that will maximize the differences among the groups. The function will show how well the borrowers can be distinguished, as well as where the classification is more robust and where it is more likely to fail. Running the Linear discriminant analysis in stata produces the below table.

Linear discriminant analysis
Resubstitution classification summary

Key
Number Percent

True seriousdlqin 2yrs	Classified 0	1	Total
0	87,365 97.75	2,009 2.25	89,374 100.00
1	4,593 70.07	1,962 29.93	6,555 100.00
Total	91,958 95.86	3,971 4.14	95,929 100.00
Priors	0.9317	0.0683	

Figure 13: Classification summary

The output includes the means of the groups and a classification table. Values in the diagonal of the classification table reflect the correct classification of individuals into groups based on their scores on the discriminant dimensions. Prior probability is based non frequency computation on the classes, a 0.9308 Priors if rejected and 0.0692 if accepted.

We can note from the above that the classification is more robust when classifying borrowers as rejected and it highly fail to distinguish those who were accepted to get a loan. This can be further demonstrated by the error rate classifier as shown in figure.

	seriousdlqin2yrs 0	1	Total
Error rate	.0224786	.7006865	.0688217
Priors	.9316682	.0683318	

Figure 14: Classification error rate

Based on the above table we can further accentuate the miss classification of the accepted individual as rejected, around 70%, exactly about 5745 of the 8200 "accepted" individuals. Whereas only 2.25 %"rejected" were classified otherwise. In total around 6.68% of the observations were misclassified around 8256 individual.

As a final comment we have reached the conclusion that the borrowers are not well distinguished and thus we need to use other method to explore the difference as we believe a third class should be identified. For such purpose we will use a cluster analysis to identify three groups as to better distinguish them.

4.5. Cluster Analysis

We have a very large simple size of historical data to classify borrowers into tow category "accepted" or "rejected". However we believe that we may detect instead of the given tow class, a third class in which customer are favorable but still be rejected due to minor issue. This, in fact, is due to the multiple barriers presented by financial institution to ensure safety. In this part we will use K-median cluster as our variable are mixed between percentage, integer, continues.

```
. tab _clus_1
```

_clus_1	Freq.	Percent	Cum.
1	39,373	41.04	41.04
2	21,221	22.12	63.17
3	35,335	36.83	100.00
Total	95,929	100.00	

Figure 15: cluster groups

Clearly and based on historical data and literature review we know that the minority group is the one with accepted profiles thus 22.12 percent based on the cluster, the majority are rejected thus 41.04% and finally those with favorable but not expected profiles around 36.83%.

Based on this results we will first conduct a binary logit, then we will use the cluster results to construct a multinomial model based on logistic regression and determine the impact of the different variables on the loan demand classes given the new class.

5. Model building and diagnosis

In this analysis, we will start by building a model based on the given class which are tow, then a model based on cluster, and one based on prior. The dependent variable “seriousdelinq2yrs” is dichotomous and coded 1 for loan accepted and 0 otherwise. Thus we may use either logistic regression, probit model or discriminant analysis. Furthermore, the sample size is very large enough to conduct either of this analysis. In addition, probit will provide similar output to that of logistic regression thus will be reported in the appendix.

5.1. Dichotomist Logistic regression

5.1.1. Model generation and interpretation

Below we use the logit command to estimate a logistic regression model.

```
. logit seriousdlqin2yrs revolvingutilizationofunsecuredl age numberoftime3059dayspastduenotwo debtratio monthlyincome numberofopencreditline
> sandloans numberoftimes90dayslate numberrealestateloansorlines numberoftime6089dayspastduenotwo numberofdependents

Iteration 0:   log likelihood = -23915.317
Iteration 1:   log likelihood = -22662.052
Iteration 2:   log likelihood = -21734.309
Iteration 3:   log likelihood = -19536.829
Iteration 4:   log likelihood = -18765.016
Iteration 5:   log likelihood = -18591.883
Iteration 6:   log likelihood = -18591.669
Iteration 7:   log likelihood = -18591.669

Logistic regression                             Number of obs   =      95929
                                                LR chi2(10)     =   10647.30
                                                Prob > chi2     =     0.0000
Log likelihood = -18591.669                     Pseudo R2       =     0.2226
```

seriousdlqin2yrs	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
revolvingutilizationofunsecuredl	1.881206	.040738	46.18	0.000	1.801361	1.961051
age	-.0155871	.0011787	-13.22	0.000	-.0178974	-.0132769
numberoftime3059dayspastduenotwo	.4312098	.0136762	31.53	0.000	.404405	.4580146
debtratio	-.0000996	.0000522	-1.91	0.057	-.0002019	2.79e-06
monthlyincome	-.0000293	3.70e-06	-7.93	0.000	-.0000366	-.0000221
numberofopencreditlinesandloans	.0294354	.0031674	9.29	0.000	.0232274	.0356434
numberoftimes90dayslate	.653041	.0207986	31.40	0.000	.6122765	.6938055
numberrealestateloansorlines	.1047223	.0129195	8.11	0.000	.0794005	.1300441
numberoftime6089dayspastduenotwo	.5831361	.0285553	20.42	0.000	.5271687	.6391035
numberofdependents	.0466625	.0119992	3.89	0.000	.0231444	.0701806
_cons	-3.414753	.0682329	-50.05	0.000	-3.548487	-3.281019

Note: 8 failures and 0 successes completely determined.

Figure 16: Logit model

At the top of the output we see that only 95929 observations in our data set were used in the analysis instead of 120000 this mainly due to missing data discussed in the first part of this report.

The likelihood ratio chi-square of 10647.30 with a p-value of 0.0000 tells us that our model as a whole fits significantly better than an empty model.

The next part of the output shows the coefficients, their standard errors, the Wald z-statistic, and the associated p-values. We can notice that most of the variables are significant with a significance level of 0.000 except debt ratio is not a significant predictor with a p-value=0.057>0.05.

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. In fact, we can note that only some variable contribute highly to the odds of acceptance including "revolvingutilizatyion", "numberoftimes3059", "numberoftimes90", numberoftimes6089"

- For every one unit change in revolving utilization of unsecured lines, the log odds of loan acceptance (versus refusal) increases by 1.88.
- For every one unit increase in Number of times borrower has been 90 days or more past due, the log odds of loan acceptance (versus refusal) increases by 1.88.

Furthermore we can compute the odds using the logistic command in other words exponentiate the coefficients and interpret them as odds-ratios.

```
Logistic regression                             Number of obs   =      95929
                                                LR chi2(10)     =   10647.30
                                                Prob > chi2     =     0.0000
Log likelihood = -18591.669                     Pseudo R2       =     0.2226
```

seriousdlqin2yrs	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
revolvingutilizationofunsecuredl	6.561413	.2672989	46.18	0.000	6.057886	7.106792
age	.9845337	.0011605	-13.22	0.000	.9822618	.9868109
numberoftime3059dayspastduenotwo	1.539118	.0210492	31.53	0.000	1.498411	1.580932
debtratio	.9999004	.0000522	-1.91	0.057	.9997981	1.000003
monthlyincome	.9999707	3.70e-06	-7.93	0.000	.9999634	.9999779
numberofopencreditlinesandloans	1.029873	.003262	9.29	0.000	1.023499	1.036286
numberoftimes90dayslate	1.921375	.0399619	31.40	0.000	1.844626	2.001317
numberrealestateloansorlines	1.110402	.0143459	8.11	0.000	1.082638	1.138879
numberoftime6089dayspastduenotwo	1.791648	.0511611	20.42	0.000	1.694129	1.894781
numberofdependents	1.047768	.0125724	3.89	0.000	1.023414	1.072702
_cons	.0328845	.0022438	-50.05	0.000	.0287681	.0375899

Note: 8 failures and 0 successes completely determined.

Figure 17: Logistic model

Now we can say that for a one unit increase in revolving utilization of secure lines, the odds of being getting a loan proposal accepted (rejected) increase by a factor of 6.561.

5.1.2. Diagnosis of accuracy

In this section we will diagnose the model starting with global fit, classification table and roc curve. We will start by analyzing global fit using the hosmer & lemeshow test and pearson test.

```
Logistic model for seriousdlqin2yrs, goodness-of-fit test

       number of observations =     95929
 number of covariate patterns =     95864
          Pearson chi2(95853) =    140786.58
                  Prob > chi2 =        0.0000
```

Figure 18: Pearson Goodness of fit

Based on the pearson test (figure) the model is globally significant for a chi-square value of 140786 and 10 degrees of freedom resulting in a p-value <0.0005. furthermore, we have the lemeshow test below which confirms the above result eventhough the test won't work well for huge amount of data.

```
Logistic model for seriousdlqin2yrs, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =     95929
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =       144.66
                  Prob > chi2 =         0.0000
```

Figure 19: Hosmer & lemeshow Goodness of fit

Proceeding to comparing the null model to the full using the fitstat comand in stata. Below we have the output

```
Measures of Fit for logistic of seriousdlqin2yrs

Log-Lik Intercept Only:   -23915.317     Log-Lik Full Model:       -18591.669
D(95918):                  37183.337     LR(10):                    10647.297
                                         Prob > LR:                     0.000
McFadden's R2:                 0.223     McFadden's Adj R2:             0.222
Maximum Likelihood R2:         0.105     Cragg & Uhler's R2:            0.268
McKelvey and Zavoina's R2:     0.296     Efron's R2:                    0.174
Variance of y*:                4.676     Variance of error:             3.290
Count R2:                      0.935     Adj Count R2:                  0.049
AIC:                           0.388     AIC*n:                     37205.337
BIC:                      -1.063e+06     BIC':                     -10532.583
```

Figure 20: Null vs full model

The log-likelihood multiplied by -2 and is commonly used to explore how well a logistic regression model fits the data. The lower this value is the better the model is at predicting the binary outcome variable this value will be later compared to the one of the model 1. From the above output we can see that the model improved much compared to the null with just intercept. Other measure are also provided including different R-square values.

Now proceeding to the classification table which list the number of classified observation and under which.

Classified	True D	True ~D	Total
+	1033	713	1746
-	5522	88661	94183
Total	6555	89374	95929

Classified + if predicted Pr(D) >= .5
True D defined as seriousdlqin2yrs != 0

Sensitivity	Pr(+| D)	15.76%
Specificity	Pr(-|~D)	99.20%
Positive predictive value	Pr(D| +)	59.16%
Negative predictive value	Pr(~D| -)	94.14%
False + rate for true ~D	Pr(+|~D)	0.80%
False - rate for true D	Pr(-| D)	84.24%
False + rate for classified +	Pr(~D| +)	40.84%
False - rate for classified -	Pr(D| -)	5.86%
Correctly classified		93.50%

Figure 21: Classification table

We can see that 93.5% of the data were correctly classified by our model which have a 99.20% specificity in addition to a low sensitivity of 15.76 % further confirmed by the graph below

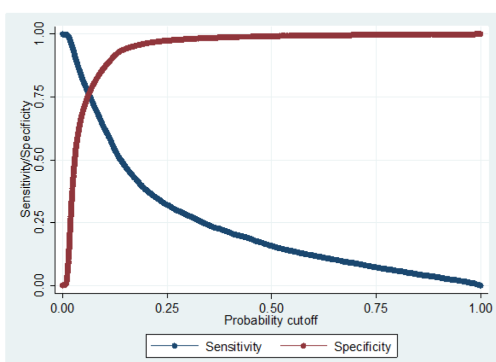

Figure 22: sensitivity curve

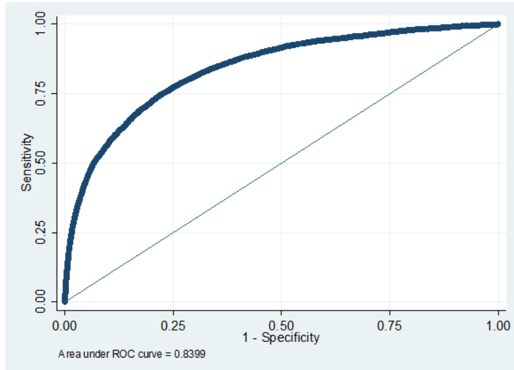

Figure 23: Roc curve

Here, the area under the curve measures discrimination, that is, the ability of the test to correctly classify those with and without financial distress. Based on the area under the curve (0.8399~0.84) we found that the rule performed predict correctly 84% of the time presence of financial distress, thus our model is quite accurate.

Proceeding to validation on the test set we see that our model accuracy decreased t 76.88% this normal and may be due to over fitting on the training set

```
. roctab real2 seriousdlqin2yrs
```

Obs	ROC Area	Std. Err.	Asymptotic Normal [95% Conf.	Interval]
23966	0.7688	0.0115	0.74633	0.79129

Figure 24: Roc test set

5.1.3. Conclusion

The developed model developed is a good model to predict financial distress even though its accuracy decreased when tested. This model can be further improved by examining over fitting problems and improving the data cleaning process.

5.2. multinomial Logistic regression based on cluster

Below we use the Mlogit command to estimate a logistic regression model based on the cluster groups

Multinomial logistic regression — Number of obs = 95929
LR chi2(20) = 58491.67
Prob > chi2 = 0.0000
Log likelihood = -73121.568 — Pseudo R2 = 0.2857

_clus_1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1	(base outcome)					
2						
revolvingutilizationofunsecuredl	-.4510794	.0252038	-17.90	0.000	-.5004779	-.401681
age	.0093449	.0005874	15.91	0.000	.0081937	.0104962
numberoftime3059dayspastduenotwo	.0104358	.0122157	0.85	0.393	-.0135065	.0343781
debtratio	-.000814	.0000588	-13.84	0.000	-.0009292	-.0006987
monthlyincome	.0001122	2.74e-06	40.97	0.000	.0001068	.0001175
numberofopencreditlinesandloans	.0625265	.0019253	32.48	0.000	.0587529	.0663001
numberoftimes90dayslate	-.0791331	.0188316	-4.20	0.000	-.1160423	-.0422238
numberrealestateloansorlines	.3735226	.0102653	36.39	0.000	.353403	.3936422
numberoftime6089dayspastduenotwo	-.064913	.0259995	-2.50	0.013	-.1158711	-.0139548
numberofdependents	.3060658	.008047	38.03	0.000	.290294	.3218376
_cons	-2.230451	.0387617	-57.54	0.000	-2.306423	-2.154479
3						
revolvingutilizationofunsecuredl	-.5721191	.0326431	-17.53	0.000	-.6360985	-.5081397
age	.0212254	.000804	26.40	0.000	.0196496	.0228013
numberoftime3059dayspastduenotwo	-.0501736	.0155391	-3.23	0.001	-.0806297	-.0197176
debtratio	-.0005007	.0000562	-8.90	0.000	-.0006109	-.0003905
monthlyincome	.0002359	2.98e-06	79.15	0.000	.0002301	.0002417
numberofopencreditlinesandloans	.0837299	.0022197	37.72	0.000	.0793794	.0880805
numberoftimes90dayslate	-.1252639	.0292605	-4.28	0.000	-.1826133	-.0679145
numberrealestateloansorlines	.7922326	.0116122	68.22	0.000	.7694731	.8149921
numberoftime6089dayspastduenotwo	-.1966854	.0376281	-5.23	0.000	-.270435	-.1229357
numberofdependents	.5119531	.0093442	54.79	0.000	.4936388	.5302675
_cons	-5.028026	.0552407	-91.02	0.000	-5.136296	-4.919756

Figure 25: Multinomial logit

At the top of the output we see that only 95929 observations in our data set were used in the analysis instead of 120000 this mainly due to missing data discussed in the first part of this report.

The likelihood ratio chi-square of 58497.64 with a p-value of 0.0000 tells us that our model as a whole is globally significant

The next part of the output shows the coefficients, their standard errors, the Wald z-statistic, and the associated p-values. We can notice that most of the variables are significant with a significance level of 0.000 except “numberoftimes30-59” ratio is not a significant predictor with a p-value=0.393>0.05. Here the modality of references is the one with the highest frequency which is as identified by the cluster to be the rejected individual class.

Starting with the second class which represent the one offered loans. We can note that some variables have a negative impact on the odds of acceptance compared to those rejected. In fact, RevolvingUtilizationOfUnsecuredLines,DebtRatio,NumberOfTimes90DaysLate, NumberOfTime60-89DaysPastDueNotWorse have a negative coefficient thus unit increase would decrease the odds of acceptance compared to those rejected. While Age, NumberOfTime30-59DaysPastDueNotWorse, MonthlyIncome, NumberOfOpenCreditLines, NumberRealEstateLoansOrLines, NumberOfDependents contribute positively to the odds of acceptance compared to those rejected.

In fact, For every one unit change in revolving utilization of unsecured lines, the log odds of loan acceptance (versus refusal) decreases by 0.45. Whereas, for every one unit increase in NumberRealEstateLoansOrLines, NumberOfDependent, the log odds of loan acceptance (versus refusal) increases by 0.37 and 0.30 respectively.

Going to the third class which is the partially accepted loans. Here unlike the second class all the variables are significant with p-value <0.05. In addition, we have variable with negative impact including; RevolvingUtilizationOfUnsecuredLines, NumberOfTime30-59DaysPastDueNotWorse, DebtRatio, NumberOfTimes90DaysLate, NumberOfTime60-89DaysPastDueNotWorse with Revolvingutilization holding the highest contribution around 0.57 for each unit increase succeeded by NumberOfTimes90DaysLate, NumberOfTime60-89DaysPastDueNotWorse as they contribute around 0.12 increase in odds. Furthermore, we have the rest of the variables contribute positively increase the odds of acceptance (versus rejection), Age, MonthlyIncome, NumberOfOpenCreditLines, NumberRealEstateLoansOrLines, NumberOfDependents. Unlike the second class, both NumberRealEstateLoansOrLines, NumberOfDependent contribute highly around 0.79 and 0.51 for every unit increase compared to the rejected class.

6. Conclusion

Through this report we went through different steps to build accurate models we have built two main model one based on two classes and another based on three classes. We came to the conclusion that the data does not accurately distinguish between groups thus we used a cluster analysis to distinguish between groups.

7. References

Multivariate Data Analysis (7th Edition) 7th Edition by Joseph F. Hair Jr, William C. Black, Barry J. Babin

Econometrics by Example, by Damodar Gujarati

An Introduction to Applied Multivariate Analysis By Tenko Raykov, George A. Marcoulides

8. Appendix: probit output

Probit regression	Number of obs = 95929
	LR chi2(10) = 10861.82
	Prob > chi2 = 0.0000
Log likelihood = -18484.408	Pseudo R2 = 0.2271

seriousdlqin2yrs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
revolvingutilizationofunsecuredl	.9331985	.0199198	46.85	0.000	.8941563	.9722407
age	-.0073541	.0005656	-13.00	0.000	-.0084625	-.0062456
numberoftime3059dayspastduenotwo	.2401978	.00737	32.59	0.000	.2257528	.2546428
debtratio	-.0000325	.0000222	-1.47	0.143	-.000076	.000011
monthlyincome	-.0000107	1.43e-06	-7.53	0.000	-.0000135	-7.94e-06
numberofopencreditlinesandloans	.0135684	.0015425	8.80	0.000	.0105452	.0165915
numberoftimes90dayslate	.3439942	.0102271	33.64	0.000	.3239494	.3640391
numberrealestateloansorlines	.0459639	.0061601	7.46	0.000	.0338903	.0580375
numberoftime6089dayspastduenotwo	.3316098	.0150408	22.05	0.000	.3021304	.3610892
numberofdependents	.0194442	.0060159	3.23	0.001	.0076532	.0312352
_cons	-1.897882	.0329728	-57.56	0.000	-1.962508	-1.833257

Figure 26: Probit output