**Introduction to Data Mining in R – Credit Scoring Example**

Banks play a crucial role in market economies. They decide who can get finance and on what terms and they can make or break investment decisions. For markets and society to function, individuals and companies need access to credit. Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted.The data set, **CreditScoring.csv**, to be investigated to better understand the characteristics of somebody who is known to have experienced financial distress in the past 2 years.

* An initial exploratory analysis should therefore be carried out on the data to investigate this.
* This information should then be used to understand which variables are significant in predicting if a borrower will experience distress in the next two years.
* Finally, based on the results obtained a decision should be made as to which customers should be accepted for a loan.

The data set contains the following variables.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| ID | Unique borrower ID | Integer |
| Serious\_2yrs | Person experienced financial distress in the last 2 years (90 days past due delinquency or worse) | Y/N |
| Balanace\_Unsecured\_Lines | Total balance on credit cards and personal lines of credit except real estate and no instalment debt like car loans divided by the sum of credit limits. | Percentage |
| Age | Age of borrower in years. | Integer |
| No\_30\_59\_DaysPast | Number of times borrower has been 30-59 days past due delinquency but no worse in the last 2 years. | Integer |
| Debt\_Ratio | Monthly debt payments, alimony, living costs divided by monthly gross income. | Percentage |
| Monthly\_Income | Monthly income. | Integer |
| No\_OpenLines | Number of Open loans (instalment like car loan or mortgage) and Lines of credit (e.g. credit cards) | Integer |
| No\_90\_DaysLate | Number of times borrower has been 90 days or more past due. | Integer |
| No\_Real\_Estate\_Loans | Number of mortgage and real estate loans including home equity lines of credit. | Integer |
| No\_60\_89\_DaysPast | Number of times borrower has been 60-89 days past due but no worse in the last 2 years. | Integer |
| No\_Dependents | Number of dependents in family excluding themselves (spouse, children etc.). | Integer |

There are a number of things you should consider as you complete this task. When undertaking a project it will not always be clear what functions in R to use, and you may not have any prior knowledge of it. Having a general idea of what you want to accomplish, and turning that into code is a difficult task.

**Task 1 – Data Pre-Processing**

* You should produce some summary statistics of your data.
* How is your data distributed? Histograms may help here.
* Clean your data, are there outliers and missing values? How will you address these?
* Have the summary statistics and distributions of the data improved. Is the data normally distributed across the variables?

**Task 2 – Exploratory Data Analysis**

* Now that the data has been cleaned and saved as a new dataset, investigate the relationships between variables. Perhaps try box or scatter plots.
* Identify the most important variables within the data in relation to the research question.
* Carry out initial statistical tests to determine which variables show significant differences by Serious\_2yrs.

**Task 3 – Model & Model Prediction**

* Use your knowledge of Logistic Regression in order to predict whether someone has experienced financial distress in the last 2 years. Use only a small sample of the variables available (the most significant 2).
* Research the predict.glm, prediction, and performance functions in R Studio help or online. Determine how you would use them to calculate the accuracy of your logistic regression model. Plotting the ROC Curve.
* Attempt a second model, which contains further variables.