12/15/2018

Predicting Bank Loan Default & Scorecard Development

Using logistic regression in R

**Presented By:**

**Xinyu Wei**

**Sheeba Wali**

**Table of Contents**

Contents

[1. Project Introduction 2](#_Toc532657561)

[1.1 Scope of the project 2](#_Toc532657562)

[2. Objective & Importance of the Project 3](#_Toc532657563)

[3. Data Attributes 4](#_Toc532657564)

[3.1 Data Dictionary 4](#_Toc532657565)

[4. Data Preparation Stage 5](#_Toc532657566)

[4.1 Missing value analysis 5](#_Toc532657567)

[4.2 Outlier Analysis 7](#_Toc532657568)

[4.3 Distribution Analysis 8](#_Toc532657569)

[4.4 Check for correlation 9](#_Toc532657570)

[4.5 Check for Multicollinearity 10](#_Toc532657571)

[5. Data Partition Stage 12](#_Toc532657572)

[6. Logistic Regression Model 12](#_Toc532657573)

[6.1 Anova- Chi Square- Model Selection 14](#_Toc532657574)

[7. Predicting the default rate 15](#_Toc532657575)

[8. ROC and AUC Validation Test 16](#_Toc532657576)

[9. Weight of Evidence (WOE) Conversion 17](#_Toc532657577)

[Calculation of WOE number: 20](#_Toc532657578)

[10. WOE Data Frame Construction 23](#_Toc532657579)

[11. Scorecard Creation 24](#_Toc532657580)

[12. Final Scorecard 28](#_Toc532657581)

[13. A Score calculated 29](#_Toc532657582)

# Project Introduction

Our aim is to develop a logistic regression model to estimate a bank loan default using the application based and behavioral information of the loan. We use the historical dummy bank personal loan data available online to arrive at the target default (0 or 1) value for each of the loan account. We will further go ahead and build a score card for the personal loan segment, which allows the credit manager to arrive at the scoring for each loan application.

This document details the entire process of model building along with statistical inference and outcomes. This document also provides details about each explanatory variable and its relationship with the response/ target variable, further detailing the various data quality and variable selection techniques.

## Scope of the project

In this document we will detail key parameters that impact the default of a bank loan based on the dummy data. We will also study the relationship between different explanatory variables and the response variable (i.e. the default variable), correlation amongst the significant explanatory variables and finally perform some out of sample validation tests like ROC and AUC to assess the goodness of fit of the model.

In this document we will establish the modelling process, the approach and the input data used to develop the model. We will carry out various data quality measures like missing data checks, outlier checks, multicollinearity and correlation check. We will further make model selection based on Anova -Chi Square test to arrive at the final model for prediction.

The final data set is partitioned into training and testing data, while we use the training data to develop the model, testing dataset is used to make predictions about the default probability. The document also details the various data validation techniques like ROC & AUC to ensure that the accuracy of model predictions. Finally, to give a more practical application of our project we have made an attempt to arrive at a scorecard, enabling the credit manager or branch manager to not only decide on the loan applications but also have a credit score w.r.t each applicant.

# Objective & Importance of the Project

In today’s world and the market economy Banks play an important role. They make financing decision based on the credit worthiness of the borrower. Our credit model will help predict the default for any personal loan application and also equip the lending bank to arrive at in individual score for each application. This tool will enable banks to determine whether a loan should be granted.

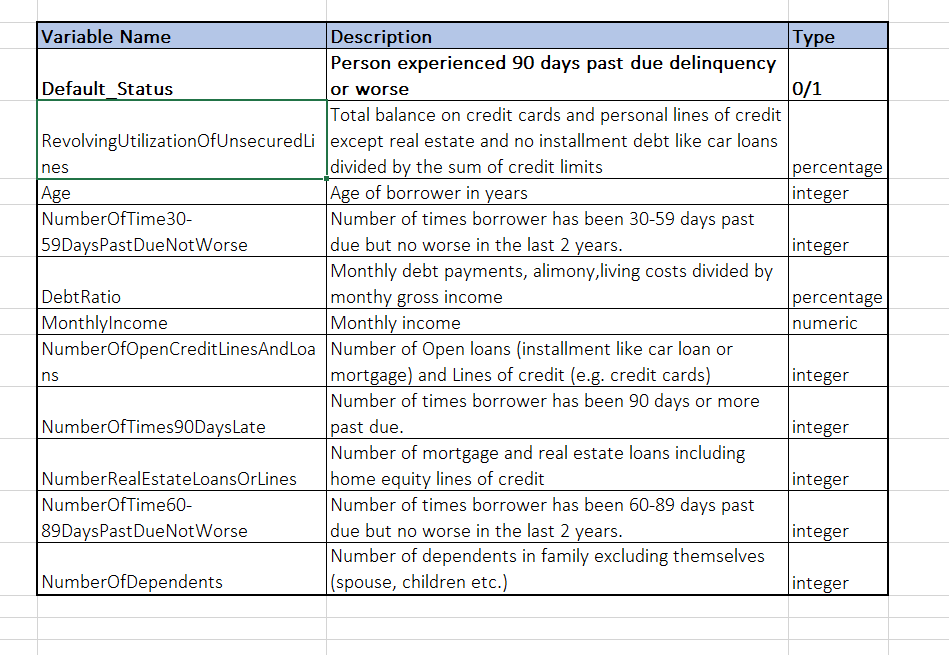
To make it clearer and more understandable we will analyze some of the personal loan data variables and run logistic regression model to arrive at the predictive probabilities of default. We will further go ahead and build a score card for the personal loan segment, which allows the credit manager to arrive at the scoring for each loan application.

# Data Attributes

* Input data is a dummy bank loan data provided by Kaggle Datasets <https://www.kaggle.com/c/GiveMeSomeCredit/data>
* We have used the cs-training.csv file with 150,000 data points.
* The data set has 10 explanatory variables and 1 response variable.
* The data is pertaining to the personal consumer loans for the last 2 years.
* The loan data has a Default\_status column, which is the target variable or the response variable having values 0 = non-default and 1 = default
* This variable captures the default as per the Basel II & III accord.
* Definition of default: overdue period of 90 days or more
* Please refer to the below the data dictionary excel which contains variable names in the loan data set and their descriptions.



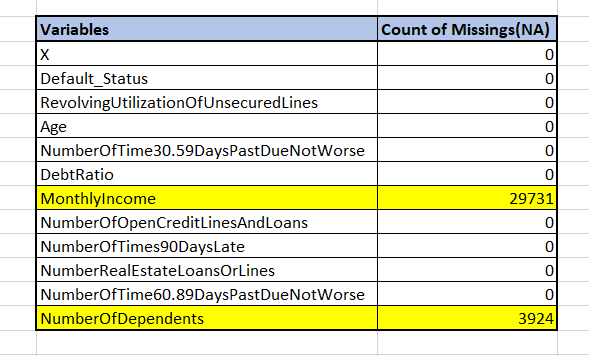
## Data Dictionary



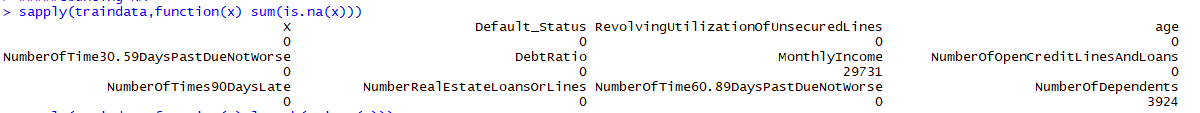
# Data Preparation Stage

## Missing value analysis

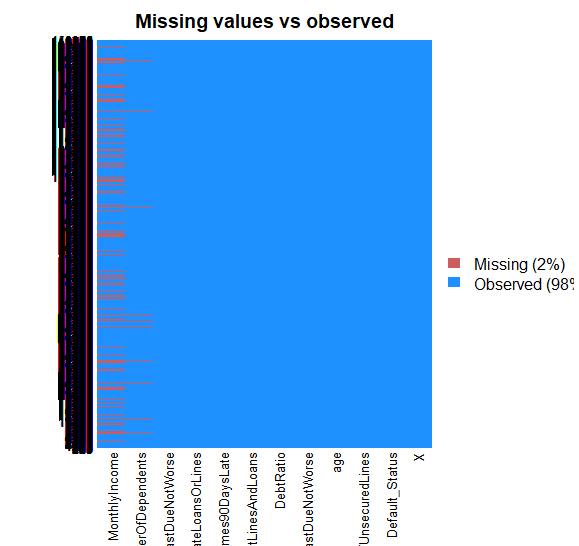
* We perform missing value analysis and find that the only two variables i.e. monthly income and number of dependents has missing values.



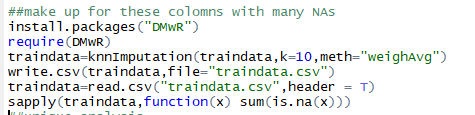
* We identified the variables with NA/ missing values using “sapply” function in R. Count of missing values as given below.



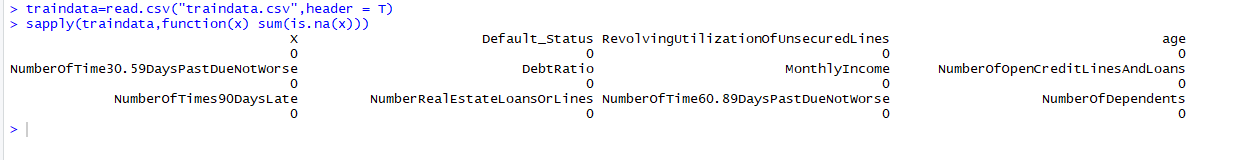
* Used the Amelia package and its “missmap” function to see missing values visually. This helps in cleaning the data.

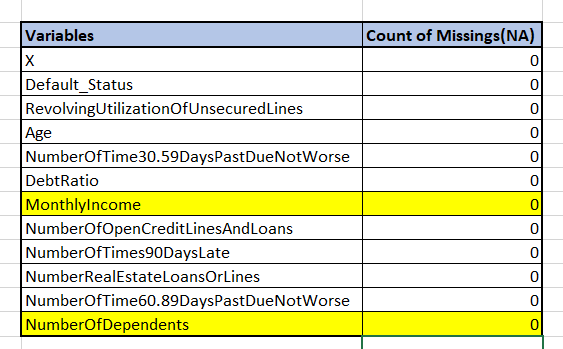


* We did not delete any mossing values from the dataset as we used the KNN function in R to fill the gaps with the weighted average values.



* After using the KNN function above, now when we check and see that all the gaps have been filled.

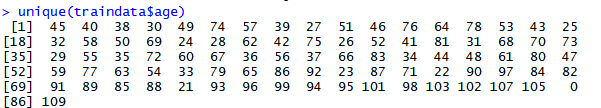




## Outlier Analysis

We conduct the data outlier check on our explanatory variables and remove the outliers. We can see that for age variable has 0 as one of the values and is eliminated as it was an outlier. Similarly, for number of times 30 to 59 Days past due had extreme values of 98 and 96 which again were removed.

**Age with 0 in the data set**

****

**After the removal of 0**

****

**No of times 30 to 59 DPDs with extreme values like 96 and 98 in the dataset**

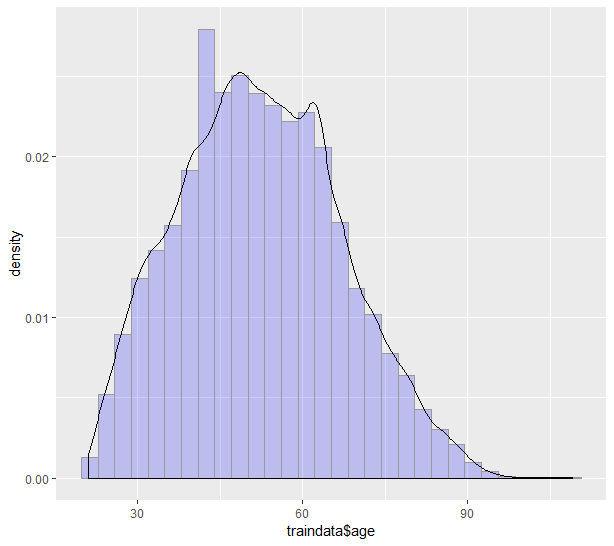
****

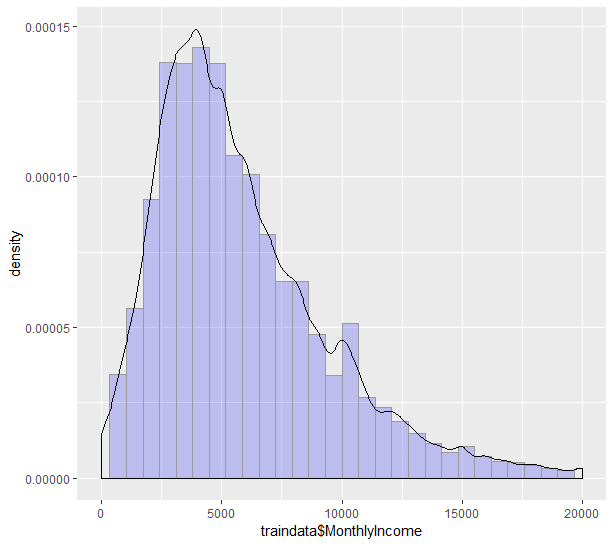
**After the removal of the outliers**

****

## Distribution Analysis

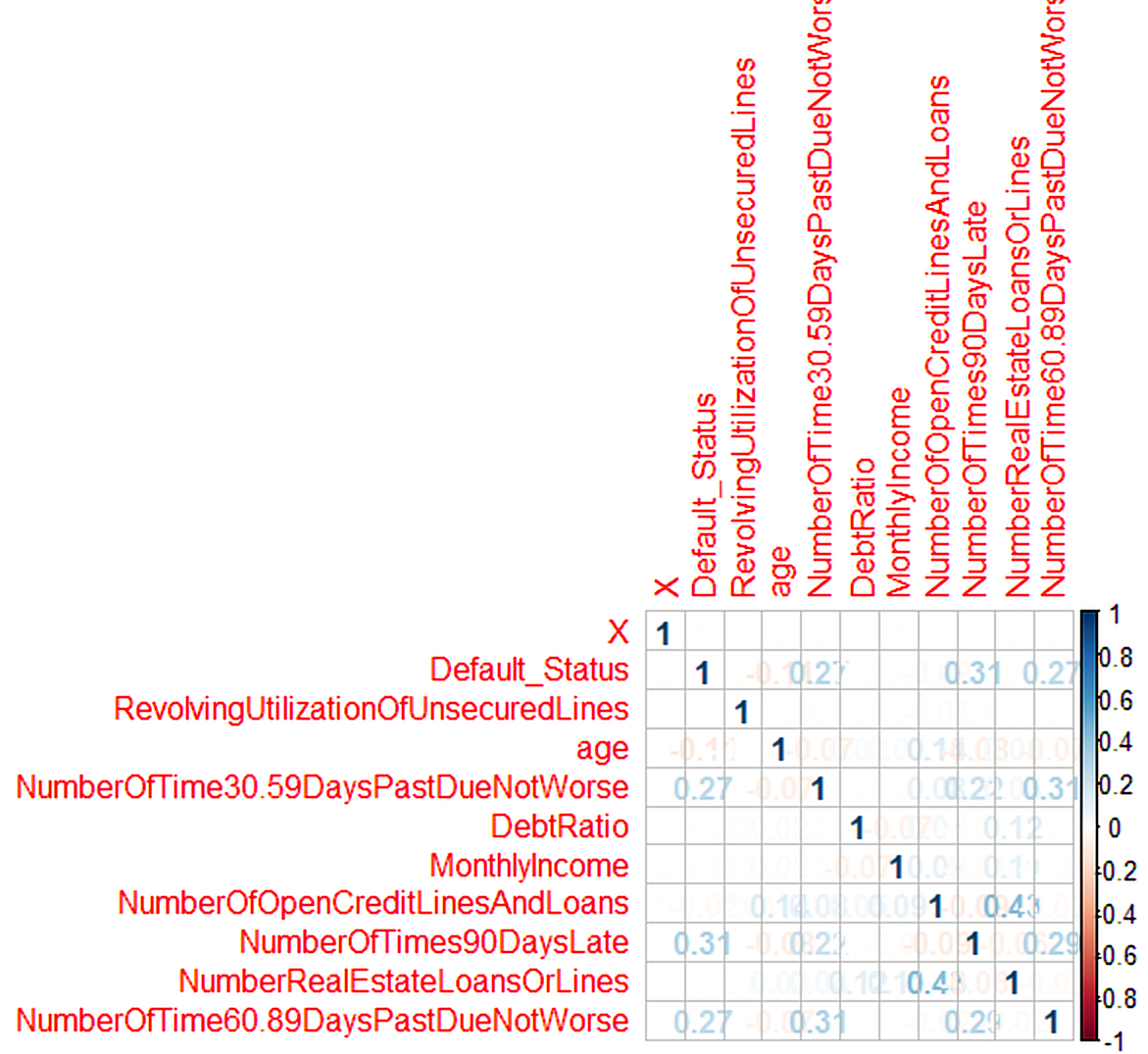
Another way of examining the outliers in the data is to check the distribution of the explanatory variables. We are presenting below distribution for variables age and monthly income.



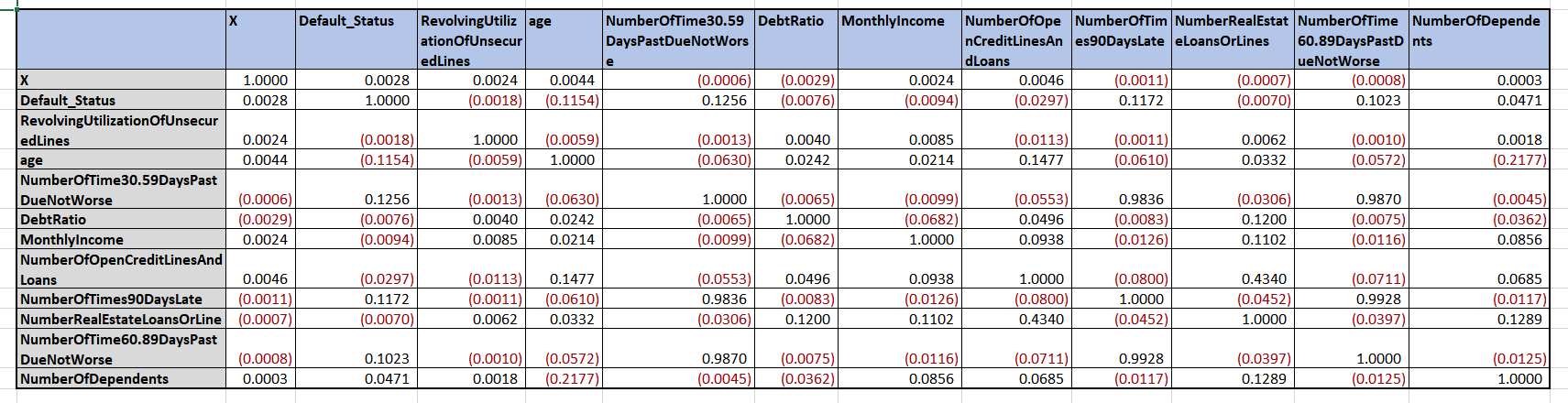


## 4.4 Check for correlation

* We have studied that the correlation of -1 and +1 means very high negative and positive correlation between the variables, which isn’t desirable in our project. Correlation closer to 0 means un correlated variables and is desirable for our analysis.
* We can see below that our data variables have very low correlation between them which means that we can go ahead and use the variables for model building.
* Used “Corrplot” in R to check and plot correlation amongst the explanatory variables.

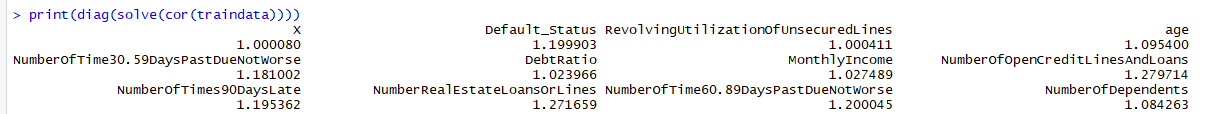






## 4.5 Check for Multicollinearity

* Another aspect to check for is the multicollinearity, its desirable for the variable to showcase low multicollinearity which means that the variables are good to be used in the model.
* When we check multicollinearity, we use VIF, if VIF > 10 we need to drop those variables. VIF below 2 is desirable to consider variables for model building.
* We can see that for our data all the variables have VIF less than 2, hence ensuring that we can use these variables for model building.

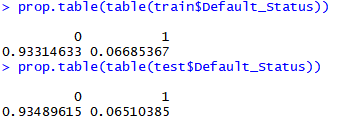


# Data Partition Stage

* We can see the count of number of default and non-default in our data. Our data has only 7% default. This might cause problem with prediction accuracy later.



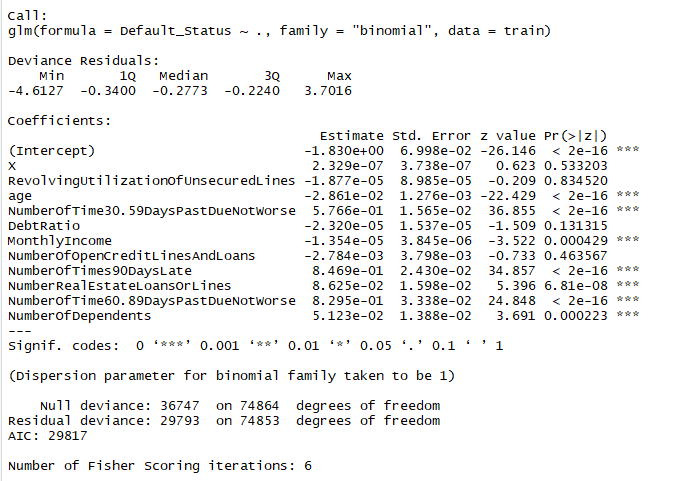
* We use “createDataPartition” function in R to partition our data into training and testing data sets by maintaining the distribution of defaults and non-defaults in both the datasets.



# Logistic Regression Model

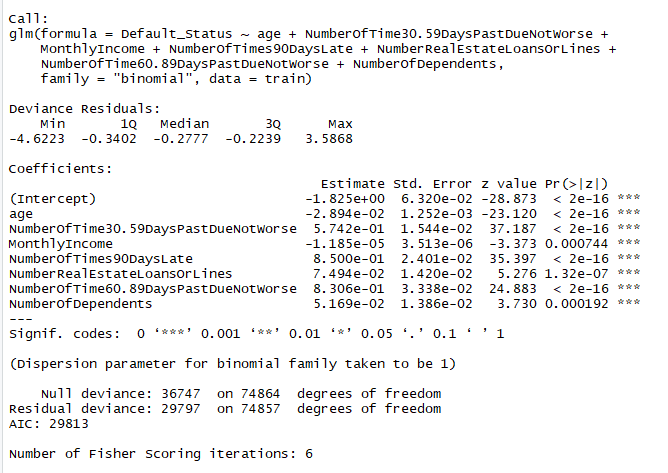
* We are going to construct a logistic regression model for our analysis since the response variable is binary in nature.
* Hence in order to predict the log odds of the defaults we run this logistic regression on full model first and check the p-value of the explanatory variables, if we find that the p-value is less than 0.05 for some explanatory variables we remove them from the model and run logistic regression on reduced model.
* We further carry out Chi-Square Anova to check if the reduced model is better than the full model and that we can go ahead and use the reduced model for model prediction and scorecard development.

## 6.1 Logistic regression Full Model:



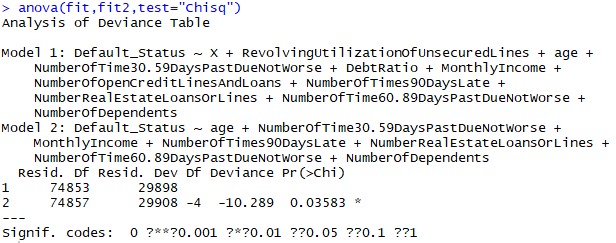
## 6.2 Logistic Regression Reduced Model:

The full model logistic regression shows that 3 variables have “**p-value”** greater than 0.05 thus failing the hypothesis testing. Now we run reduced model eliminating these 3 variables



## Anova- Chi Square- Model Selection

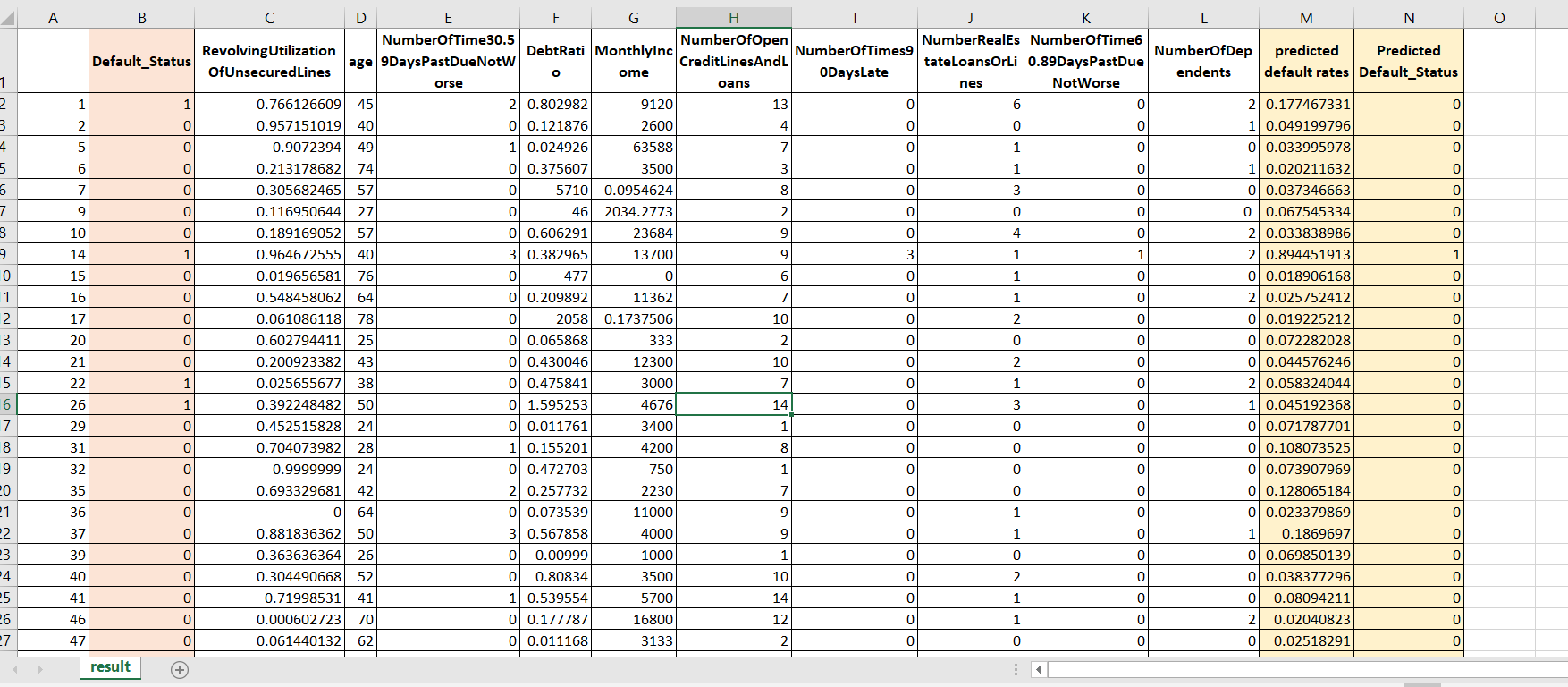
* The Anova table below shows that the reduced model is better to work with as the Pr (>Chi) is 0.035 which is less than 0.05 hence we reject the null hypothesis that the full model and reduced models show no difference.



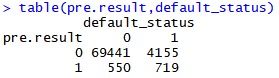
# Predicting the default rate

* We ran the “predict” function in R to arrive at the predicted values of default
* We have considered 0.5 as the critical point to assign 0 or 1 to each loan application.
* Sample predictions are given below for reference



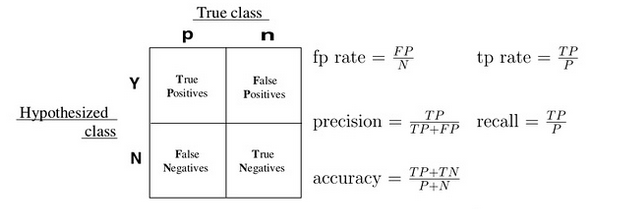


* The accuracy of the prediction can be seen from below table as well (93.72% Accuracy)

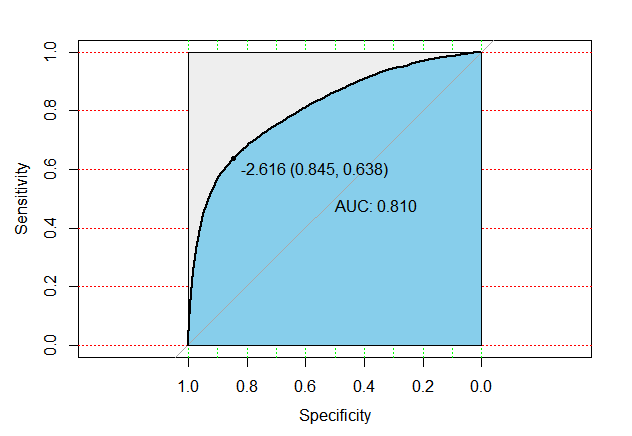


# ROC and AUC Validation Test

* To test the prediction made on the test data run “pROC” package in R
* Also give the AUC= Area under the ROC curve number. AUC ranges between 0.5 to 1.
* The closer the AUC to 1 the better prediction capabilities of the model.



* The more the curve is leaning towards the left side top corner the better the prediction capabilities of the model. Our model gives a good AUC of 0.81 .

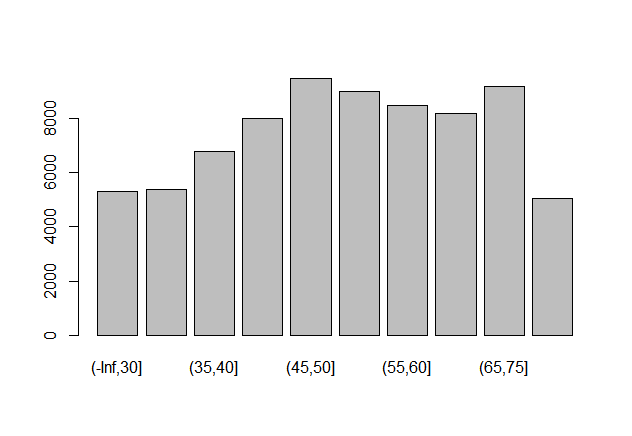


# Weight of Evidence (WOE) Conversion

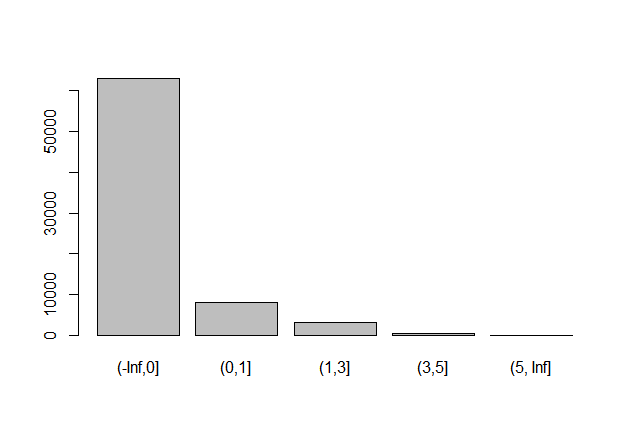
* The WOE transformation helps in transforming the logistic regression model into a scorecard model.
* We replace each of the X variables in the logistic regression model into WOE(x) by using the formula: WOE = LN [(Bad/Total Bad)/ (Good/Total Good)]
* Before we do the WOE transformation, we begin with binning each variable into groups or bins which is suitable for arriving at a score later on.
* Given below are the bins creation for each variable:

## 9.1 Binning of each variable

1. **Age**



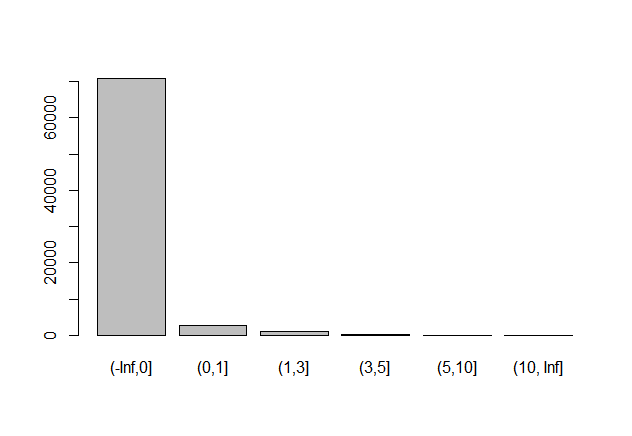
1. **Number of Time 30-59 Days Past Due Not Worse**



1. **Monthly Income**



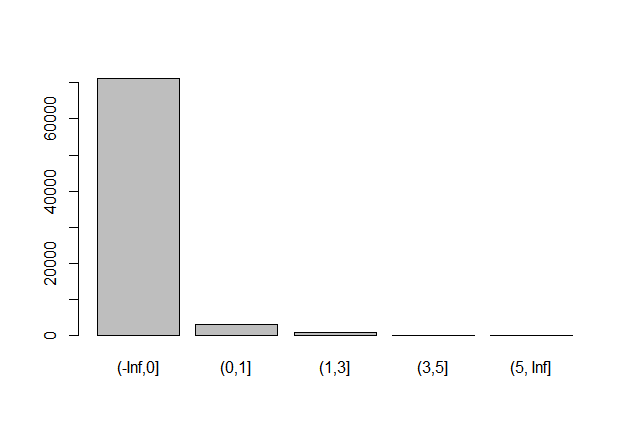
1. **Number of Times 90 Days Late**



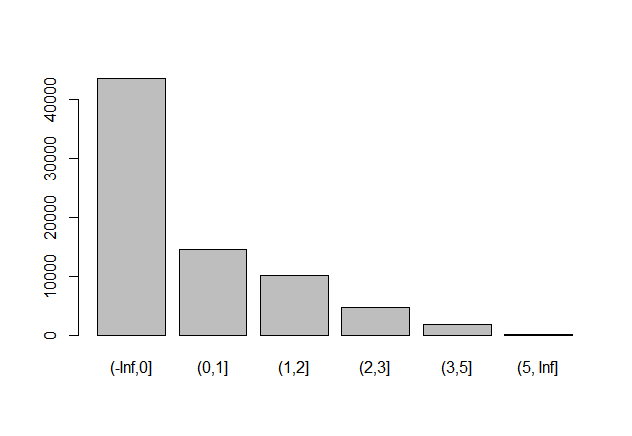
1. **Number Real Estate Loans or Lines**



1. **Number of Time 60-89 Days Past Due Not Worse**

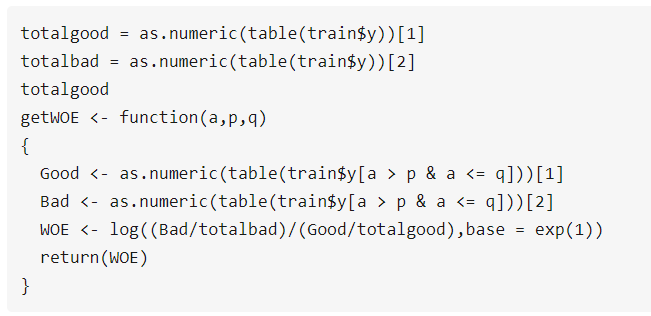


1. **Number of Dependents**

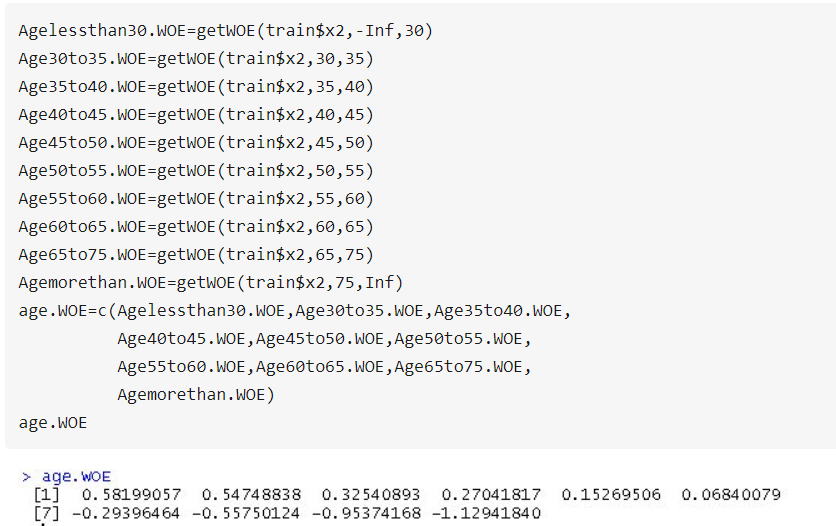


## 9.2 Calculation of WOE number:

* We have calculated the WOE for each of the bins created earlier using the below mentioned rule.



* Given below is the R code for arriving at WOE for each bin of AGE variable. Similarly, WOE is calculated for rest of the variables. We have only shown the output WOE for the remaining variables.



1. **Age**



1. **NumberOfTime30-59DaysPastDueNotWorse**



1. **Monthly Income**



1. **Number of Times 90 Days Late**



1. **Number Real Estate Loans or Lines**



1. **Number of Time60-89 Days Past Due Not Worse**



1. **Number of Dependents**



## 9.3 WOE for the entire dataset

We further show the WOE values for the entire data for each variable.



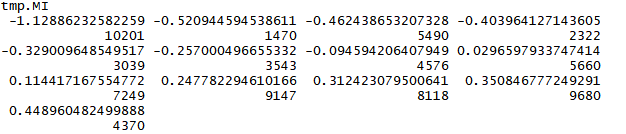
1. **Age**



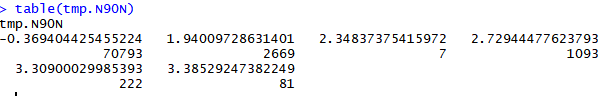
1. **Number of Time 30-59 Days Past Due Not Worse**



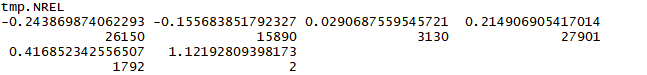
1. **Monthly Income**



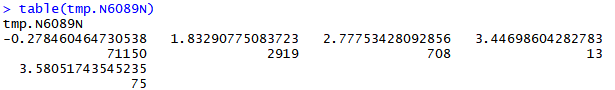
1. **Number of Times 90 Days Late**



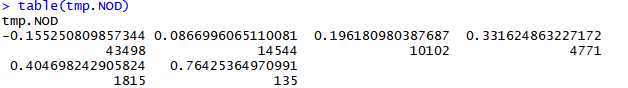
1. **Number Real Estate Loans or Lines**



1. **Number of Time 60-89 Days Past Due Not Worse**



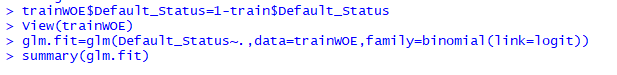
1. **Number of Dependents**

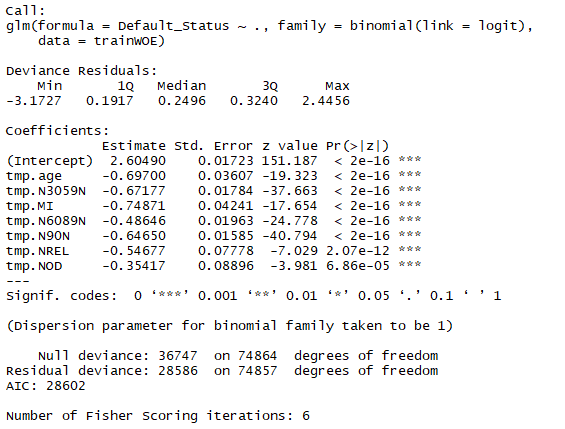


# WOE Data Frame Construction

Once all the variable WOEs are constructed, we combine them into a data frame and run logistic regression model to arrive at the beta for each of the newly transformed WOE variables. Further these beta values are multiplied by the WOEs to arrive at the score for each variable.





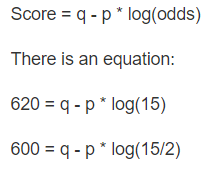
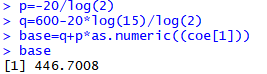


# Scorecard Creation

## 11.1 Base score creation:

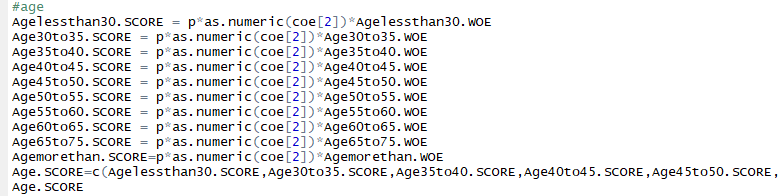
Each variable in the scorecard follows IF-THEN rule, value of each variable determine the score assigned to them, and the total score is the sum of score of each variable. Also remember for scorecard development we need to have a threshold or base score value such that each applicant score will be compared to this threshold score and the applicants above this threshold will only be permitted into the loan system.

* Any Borrower’s score is calculated as:
  + **Individual Score = Base Score + each variable score**
* For Base Score computation:
  + Assume that the ratio is 15 or better for 600 points,
  + The ratio of good or bad for every 20 points is doubled to calculate P, Q.
  + For our data the base score was 446.70



## 11.2 Score creation for each variable:

Now that we have a base score, we continue to score each binned WOE variable, given below is the sample R code for only one variable Age and the score output for each of the variables.



**1. Age:**



**2. Number of Time 30-59 Days Past Due Not Worse- N3059N**



**3. Monthly Income - MI:**



**4. Number of Times 90 Days Late- N90N:**



**5. Number Real Estate Loans or Lines- NREL:**



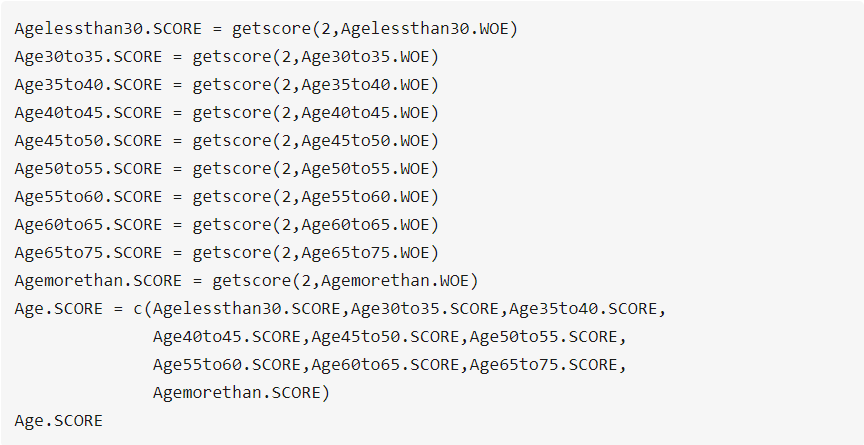
**6. Number of Time 60-89 Days Past Due Not Worse- N6089N:**

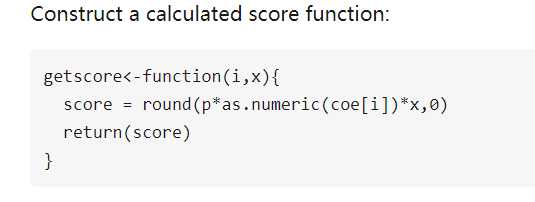


**7. Number of Dependents- NOD:**



## 11.3 Roundup the scores for each variable:

In the end we round up each of the scores for every variable using the following function.



**1. Age:**



**2. Number of Time 30-59 Days Past Due Not Worse- N3059N**



**3. Monthly Income - MI:**



**4. Number of Times 90 Days Late- N90N:**



**6. Number of Time 60-89 Days Past Due Not Worse- N6089N:**

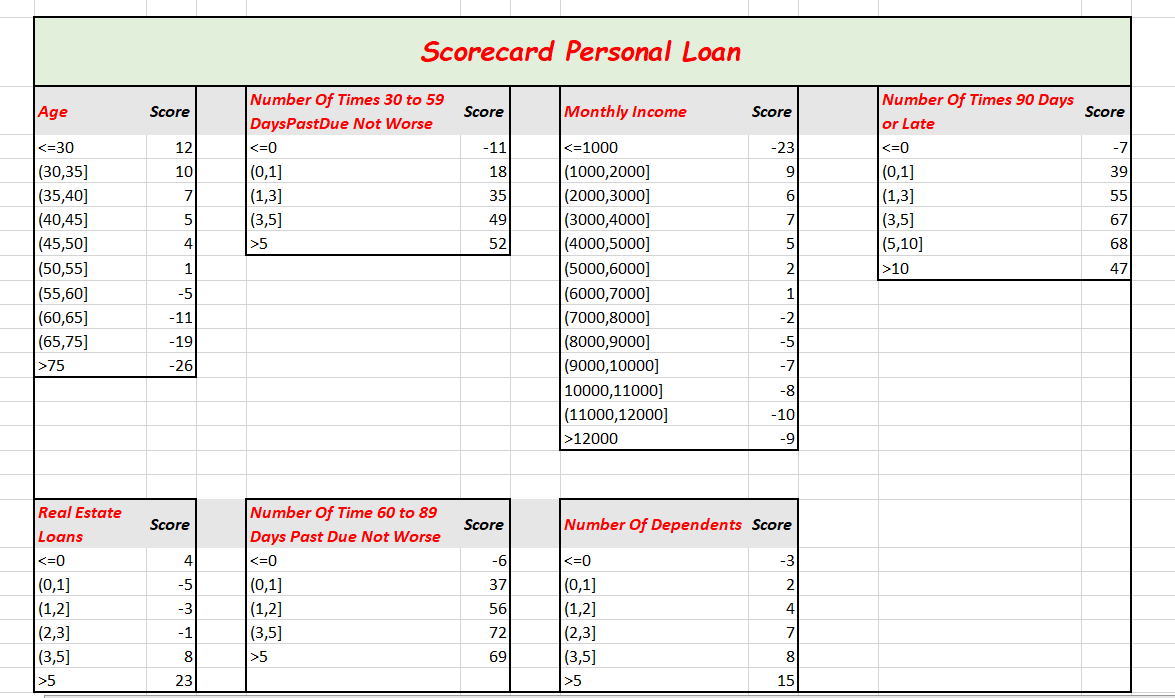


**7. Number of Dependents- NOD:**



# Final Scorecard

Lastly, based on the scores rounded up in the above step we are able to form the score card for our personal loan data. The scorecard is given below:



# A Score calculated

We have also calculated a hypothetical applicant score with the following attributes and the result is as follows:

