**Will a customer go ‘bad’ and don’t pay back to the company?**

***A Model Creation to predict the probability of risk of the loan defaulting!***

**Mid- Term Project**

**Predictive Analytics (STAT-551)**



**SUBMITTED BY**

**Tahmid Alam**

**Sakib Faisal**

**Md Mominul Islam**

**Pramisha Thapaliya**

**Table of Contents**

Table of Contents

[1. Introduction 2](#_Toc100189649)

[2. Objectives 3](#_Toc100189650)

[3. Methodology 3](#_Toc100189651)

[3.1. Description of data and variables 3](#_Toc100189652)

[3.2. Data reformatting 4](#_Toc100189653)

[3.2.1. Missing values 4](#_Toc100189654)

[3.2.2. Performance variables 5](#_Toc100189655)

[3.2.3. ID Columns variables 5](#_Toc100189656)

[3.2.4. Unneeded variables 5](#_Toc100189657)

[3.3. Summary statistics of categorical variables and continuous variables 5](#_Toc100189658)

[3.5. Training and validation approach 9](#_Toc100189659)

[3.6. Multivariate adaptive regression splines (MARS) 10](#_Toc100189660)

[3.7. Logistic regression model 10](#_Toc100189661)

[3.8. Prediction of dependent variable in test dataset 10](#_Toc100189662)

[3.9. Access fit of the model using ROC and K-S curves 10](#_Toc100189663)

[4. Results 10](#_Toc100189664)

[4.1. Exploratory data analysis 10](#_Toc100189665)

[4.2. Multivariate Adaptive Regression Splines (MARS) 10](#_Toc100189666)

[4.3. Logistic Regression Model 11](#_Toc100189667)

[4.4. Model performance using ROC and KS 12](#_Toc100189668)

[**4.4.1.** **ROC curve (Test set)** 12](#_Toc100189669)

[**4.4.2.** **Kolmogorov-Smirnov (KS) curve and Statistic** 14](#_Toc100189670)

[CONCLUSION 15](#_Toc100189671)

[REFERENCES 16](#_Toc100189672)

[Appendix 1. Summary statistics of continuous variables 16](#_Toc100189673)

[Appendix 2. Summary statistics of categorical variables 17](#_Toc100189674)

[Appendix 3. Histograms showing distribution of data 18](#_Toc100189675)

[Appendix 4. Bar plots showing borrowers’ characteristics 19](#_Toc100189676)

[Appendix 5. Correlation of the variables 22](#_Toc100189677)

# **Introduction**

Companies and big businesses are often risktakers and they will suffer a huge loss if they don’t predict on what kind of customers do, they need to deal with. Thus, it is extremely important for companies to know if their customers tend to pay back in the future or not. This is especially true in case of credit card companies, bank, credit services agencies and loan companies. Hence, it is very important for the companies to predict and analyze if their customer will go “bad” and don’t pay back to the company or not.

The study conducts the exploratory data analysis of the real data called ‘Prosper.csv”. We have used Logistic Regression model and Multivariate Adaptive Regression Splines (MARS) model

to meet our study objective. The results of this study aim to give an essential insight on important predictors that will predict if a customer pays back to the company in the future or not.

# **Objectives**

The major objective of our study is to create a model that will predict the probability that a

customer will go ‘bad’ and don’t pay back to the company. The specific objectives are:

* To understand the information in the dataset and present it as ‘evidence’
* To treat the data and address the issues of data reformatting, categorical data, and missing data
* To create a ‘modeling’ dataset of data with variables of interest
* To identify important predictors that facilitates loan defaulting in the future
* To perform Logistic regression analysis and MARS analysis using training validation approach
* To predict variable of interest using MARS and Logistic Regression models
* To create ROC curve and KS curve, to access the fit and compare the performance of the models

# **Methodology**

## **3.1.** **Description of data and variables**

In our study, we have used ‘Prosper.csv’ data. We have a data of total 18,987 observations with a total number of 30 variables. This is a real dataset from a company called Prosper and we got data from Prosper.com. In this case, our dependent or target variable is an observed outcome derived or observed on a loan after the loan has been issued or in simple words, ‘Did they pay us back?’ The performance metric will be calculated into a ‘target’ variable that is 1/0 binary as to whether the loan is bad or good. ‘Bad’ has two classes in the dataset, that is 0 and 1. 1 means the customer won’t pay back the money in the future or go bad and 0 means the customer would pay back. So, here in our case, we can say that ‘Bad’ is our dependent/ target variable. Our independent variables are debt to income ratio, Is borrower homeowner, current delinquencies, public records last 10 years, employment status and others.

## **Data reformatting**

### **Missing values**

While doing the exploration of the preliminary data, we found missing values in some of the variables in our data set. The missing values are in the form of 999, n/a and NA. For our ease, we converted all the missing value types into a single type i.e. NA. This helps us to visualize the total missing values in our data set.

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 1. Percentage of missing data present in the variables**

The Figure 1 above provides a useful information on missing data of each variable in the data set. From the figure above, we can see that the variables “BorrowerCity”, “AmountDelinquent”, “RevolvingCreditBalance”, “PublicRecordsLast12Months”, “OpenCreditLines”, “CurrentCreditLines”, “BankcardUtilization”, “EmploymentStatus”, “BorrowerOccupation” and “DebtToIncomeRatio” have missing values in the data set. The variables with more than 50% missing values is “BorrowerCity”. The variable with more than 50% of missing values is dropped for further data analysis. However, the variables with less than 50% of missing values i.e., “AmountDelinquent”, “RevolvingCreditBalance”, “PublicRecordsLast12Months”, “OpenCreditLines”, “CurrentCreditLines”, “BankcardUtilization”, “EmploymentStatus”, “BorrowerOccupation” and “DebtToIncomeRatio” are categorized as “Unknown” group.

### **Performance variables**

Performance variables like LPStatus, DPD are dropped for further data analysis. It is because they are not useful information for our model.

### **ID Columns variables**

ID Columns variables like ListingKey, ListingNumber, MemberKey, LoanKey are dropped for further data analysis. It is because they are not useful in prediction of future data.

### **Unneeded variables**

Unneeded variables like BorrowerCity, AmountRemaining and BorrowerOccupation are dropped for further data analysis, to make the modeling easy.

## **Summary statistics of categorical variables and continuous variables**

The summary statistics of categorical variables and continuous variables are given in Appendix 1 and Appendix 2 respectively.

* 1. **Binning of Variables**

Some continuous predictor variables used for building models are binned. Binning is a way to group several more or less continuous values into a smaller number of “bins”. Once the bins are created, the information gets compressed into groups which later affects the final model. These continuous variables now are treated as factor/categorical variables. Below is the visualization of some binned continuous predictors.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Figure 2. Binning of Variables**

## **Training and validation approach**

This study uses training and validation approach to perform analysis. The full dataset is randomly divided into two parts i.e., 60% being training set and 40% being test set. Using similar data for training and testing helps to minimize the effects of data discrepancies and better understand the characteristics of the model. After a model has been processed by using the training set, the model is evaluated by making predictions against the test set. Because the data in the testing set already contains known values for the variable that we want to predict, it is easy to determine whether the model’s guesses are correct or not. After the models have been trained by using the training set, the models are tested by making predictions against the validation set. Because the data in the validation set already contains known values for the response variable, Bad, it is easy to determine whether the models’ guesses are correct or not.

## **Multivariate adaptive regression splines (MARS)**

MARS model provides a convenient approach to capture the nonlinearity aspect of regression by assessing cut points (knots) like step functions. The procedure assesses each data point for each predictor as a knot and creates a linear regression model with the candidate feature(s). MARS model is generated using earth function in R. GCV, RSS, GR squared, R squared values are tabulated to evaluate models’ predictive power.

## **Logistic regression model**

The dependent variable “Bad” is a binary variable. Logistic regression models the probabilities for classification problems with two possible outcomes. Input variables were first transformed into “binned variables” which inherently means the input will be categorical. A backward stepwise logistic regression analysis was conducted for variable selection. Variables selected from backward stepwise logistic regression analysis were used to generate another reduced logistic regression model which consisted binned variables. The details on input, significant variables and analysis are discussed in result section below. These logistic regression models were generated using glm function in R-studio.

* 1. **Prediction of dependent variable in test dataset**

The dependent variable ‘Bad’ is predicted in test dataset using predict function. Both regression models i.e., MARS and logistic regression models are used for predicting variable ‘Bad’ in test dataset. For predicting logistic regression model, type= response is used which generated predicted probabilities of dependent variable.

## **Access fit of the model using ROC and K-S curves**

An **ROC curve (receiver operating characteristic curve)** is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False positive Rate. Along with AUROC, another measure is the KS statistic. It is the maximum difference between TPR and FPR. Higher KS stat value is indicative of better model.

# **Results**

## **Exploratory data analysis**

In this section, the histograms, bar plots and correlation plot generated in Final Project Part I are given in Appendix 3, Appendix 4, and Appendix 5 respectively. For full interpretation of exploratory data analysis, please refer to the Final Project Part I.

## **Multivariate Adaptive Regression Splines (MARS)**

The results from MARS model shows us that 9 terms were used from 3 predictors. A glm argument with family binomial is passed in MARS model since the dependent variable “bad” was binary. Looking into the terms in our model, we can see that the variables, “Principal Balance”, “Amount Borrowed”, and “Income” are included.

Residual sum of square (RSS) is found to be 3139.506 whereas R2 is found to be 0.1423213. GCV value of the MARS model is calculated using the penalty argument and is found to be 0.1566817. Generalized R2 which is a predictive power of the model is found to be 0.1406927 (Table 1).

**Table 1. Results of multivariate adaptive regression splines (MARS) model**

|  |  |
| --- | --- |
| **Variable** | **Bad** |
| (Intercept) | -1.2059975 |
| PrincipalBalance\_Bins(6.25e+03,1.25e+04) | 5.0899534 |
| PrincipalBalance\_Bins(1.25e+04,1.88e+04) | 10.6707240 |
| PrincipalBalance\_Bins(1.88e+03,2.5e+04) | 27.3337580 |
| AmountBorrowed\_Bins(5.8e+03,1.06e+04) | -0.7233100 |
| AmountBorrowed\_Bins(1.06e+04,1.54e+04) | -4.9618021 |
| AmountBorrowed\_Bins(1.54e+04,2.02e+04) | -5.9901291 |
| AmountBorrowed\_Bins(2.02e+04,2.5e+04) | -10.6746360 |
| Income\_Bins [1.17, 2.33) | 0.2896362 |
| Income\_Bins [2.33, 3.5) | 0.2709326 |
| GCV | 0.1656817 |
| RSS | 3139.506 |
| Generalized R2 | 0.1406927 |
| R2 | 0.1423213 |

## **Logistic Regression Model**

Logistic regression, also called a logit model, is used to model the dichotomous outcome of credit delinquency. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables.

Input variables were first transformed into “binned variables” which inherently means the input will be categorical. A backward stepwise logistic regression analysis was conducted for variable selection. This resulted in significance of only four variables i.e., PrincipalBalance, AmountBorrowed, CurrentDelinquencies, IncomeBins. And again, an analysis of reduced logistic regression model with significant variables was performed. The result of reduced logistic regression analysis with binned variables is shown in the Table 2 below.

The table below provides the summary of logistic regression model which contains coefficient estimate. The AIC value of the logistic regression model is found to be 11281.

**Table 2. Results of logistic regression model**

|  |  |
| --- | --- |
|  | **Coefficient Estimate** |
| (Intercept)\*\*\* | -1.29473 |
| PrincipalBalance\_Bins(6.25e+03,1.25e+04)\*\*\* | 5.25077 |
| PrincipalBalance\_Bins(1.25e+04,1.88e+04)\*\*\* | 10.71944 |
| PrincipalBalance\_Bins(1.88e+03,2.5e+04) | 27.41586 |
| AmountBorrowed\_Bins(5.8e+03,1.06e+04)\*\*\* | -0.74545 |
| AmountBorrowed\_Bins(1.06e+04,1.54e+04)\*\*\* | -5.09771 |
| AmountBorrowed\_Bins(1.54e+04,2.02e+04)\*\*\* | -5.95064 |
| AmountBorrowed\_Bins(2.02e+04,2.5e+04)\*\*\* | -10.71883 |
| CurrentDelinquencies\_Bins[21.3,42.7)\*\*\* | 1.18761 |
| CurrentDelinqencies\_Bins[42.7,64.1) | -14.64085 |
| Income\_Bins [1.17, 2.33)\*\*\* | 0.45629 |
| Income\_Bins [2.33, 3.5)\*\*\* | 0.36951 |
| Income\_Bins [3.5, 4.67)\* | 0.16484 |
| Income\_Bins [4.67, 5.83) | 0.14268 |
| Income\_Bins [5.83, 7.01) | -0.02826 |

Note: \*, \*\*, \*\*\* represents significance at 5%, 1% and 0.1% level

The estimates of logistic regression model represent log-odds value. The log odds value of intercept term is found to be -1.29473 and significant. The significant variables were also shown in the table.

## **Model performance using ROC and KS**

### **ROC curve (Test set)**

The ROC curve is created by evaluating the class probabilities for the model across a continuum of thresh-holds. For each candidate threshold, the resulting true-positive rate(sensitivity) and the false-positive rate (specificity) are plotted against each other. The figures below show the results of this process for the credit card data for two models: MARS and logistic. The ROC plots is a helpful tool for choosing the threshold that appropriately maximizes the trade-off between sensitivity and specificity. In comparing the two models with ROC curves, a perfect model would have a sensitivity and specificity of 100% - Graphically, the curve would be a single steep between (0,0) and (1,1) and remain constant from (0,1) to (1,1). The area under the curve (AUC) of such a perfect model would be equal to 1. An ineffective model will have its ROC curve that follows the 45 degrees diagonal line and would have an AUC of approximately 0.5.

ROC curves with corresponding Area Under Curve (AUC) values are made from the training and validation datasets for each model. In comparing the logistic and the MARS model, ROC plots and AUC was generated from the validation dataset. The logistic model has a little higher AUC value than the MARS model so, we can say that Logistic model is better in this case.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

**Figure 3. ROC curves**

### **Kolmogorov-Smirnov (KS) curve and Statistic**

The Kolmogorov-Smirnov (KS) statistic is a performance statistic which measures the discriminatory power of a model. It is the largest difference between the True Positive Rate (TPR) and False Positive Rate (FPR) at a given percentile. It looks at the maximum difference between the distribution of cumulative events and cumulative non-events. It is a very popular metric used in credit risk and response modeling. The Kolmogorov–Smirnov test a very efficient way to determine if two samples are significantly different from each other. In predictive analytics, the test is used to determine if predictions from different models differ significantly from each other. **The higher the value, the better the model.**

Chart, line chart

Description automatically generated

**Figure 4. KS Charts**

In this case, the Logistic Model and MARS has the same KS statistic. This means that in this case, both models can be said to have the same predictive power in this case.

## **Conclusion**

From the above study, we can see that the MARS and Logistic models are two good models in predicting if the customer goes bad or not. Even though the KS Statistics of both models on the validation dataset are equal, based on the AUC value, the logistic model outperforms the MARS model since it has the highest value.

**REFERENCES**

“The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)”, by Hastie, Tibshirani, and Friedman.

“An Introduction to Statistical Learning with Applications in R’ James, Witten, Hastie, Tibshirani.

prosper.com

**APPENDIX**

## **Appendix 1. Summary statistics of continuous variables**

|  |  |  |
| --- | --- | --- |
| Variables | Target (Frequency) | |
| **Good (0)** | **Bad (1)** |
| Borrower State | | |
| AA | 5 | 0 |
| AE | 8 | 0 |
| West | 4380 | 1641 |
| South | 4361 | 1562 |
| AP | 7 | 1 |
| Northeast | 1273 | 322 |
| IA | 128 | 37 |
| ID | 108 | 46 |
| Midwest | 3216 | 1214 |
| IN | 278 | 71 |
| ME | 69 | 11 |
| ND | 30 | 6 |
| NE | 76 | 11 |
| TN | 96 | 30 |
| Is Borrower Homeowner | | |
| FALSE | 7860 | 2700 |
| TRUE | 6175 | 2252 |
| Employment Status | | |
| Full-time | 9304 | 3446 |
| Not available | 2670 | 770 |
| Not employed | 91 | 26 |
| Part-time | 452 | 134 |
| Retired | 209 | 103 |
| Self-employed | 749 | 368 |
| Income | | |
| Level 0 | 3308 | 894 |
| Level 1 | 296 | 145 |
| Level 2 | 1335 | 536 |
| Level 3 | 4064 | 1598 |
| Level 4 | 2744 | 988 |
| Level 5 | 1203 | 398 |
| Level 6 | 1012 | 373 |
| Level 7 | 73 | 20 |

## **Appendix 2. Summary statistics of categorical variables**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Target | Variables | Min. | 1st Quantile | Median | Mean | 3rd Quantile | Max. | N |
| 0  (Good) | Debt to Income Ratio | 0.00 | 0.13 | 0.20 | 0.32 | 0.31 | 10.01 | 14035 |
| Amount Borrowed | 1000 | 2550 | 5000 | 6285 | 8000 | 25000 | 14035 |
| Current Delinquencies | 0.00 | 0.00 | 0.00 | 1.20 | 1.00 | 50.00 | 14035 |
| Delinquencies Last 7 years | 0.00 | 0.00 | 0.00 | 5.80 | 6.00 | 99.00 | 14035 |
| Public Records Last 10 years | 0.00 | 0.00 | 0.00 | 0.39 | 1.00 | 21.00 | 14035 |
| Total Credit Lines | 2.00 | 13.00 | 22.00 | 23.74 | 32.00 | 108.00 | 14035 |
| Inquiries Last 6 Months | 0.00 | 0.00 | 1.00 | 2.44 | 3.00 | 46.00 | 14035 |
| Amount Delinquent | 0.00 | 0.00 | 0.00 | 1068 | 20 | 190585 | 14035 |
| Public Records Last 12 Months | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 7.00 | 14035 |
| Current Credit Lines | 0.00 | 5.00 | 9.00 | 9.49 | 13.00 | 46.00 | 14035 |
| Open Credit Lines | 0.00 | 4.00 | 7.00 | 8.14 | 11.00 | 43.00 | 14035 |
| Revolving Credit Balance | 0.00 | 1338.00 | 5411.0 | 15570 | 15213 | 1435667 | 14035 |
| Employment Status | 9304 | 2670 | 91 | 452 | 209 | 749 | 14035 |
| Income | 4064 | 3308 | 2744 | 1335 | 1203 | 1012 | 14035 |
| Principal Balance | 0 | 1357 | 2529 | 2254 | 4312 | 16755 | 14035 |
| 1  (Bad) | Debt to Income Ratio | 0.00 | 0.14 | 0.22 | 0.40 | 0.34 | 10.01 | 4952 |
| Amount Borrowed | 1000 | 2600 | 5000 | 7019 | 9500 | 25000 | 4952 |
| Current Delinquencies | 0.00 | 0.00 | 0.00 | 2.07 | 2.00 | 64.00 | 4952 |
| Delinquencies Last 7 years | 0.00 | 0.00 | 1.00 | 7.32 | 9.00 | 99.00 | 4952 |
| Public Records Last 10 years | 0.00 | 0.00 | 0.00 | 0.55 | 1.00 | 30.00 | 4952 |
| Total Credit Lines | 2.00 | 14.00 | 23.00 | 25.45 | 34.00 | 129.00 | 4952 |
| Inquiries Last 6 Months | 0.00 | 1.00 | 3.00 | 4.17 | 6.00 | 105.00 | 4952 |
| Amount Delinquent | 0.00 | 0.00 | 0.00 | 1847.2 | 590.5 | 444745.0 | 4952 |
| Public Records Last 12 Months | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 7.00 | 4952 |
| Current Credit Lines | 0.00 | 5.00 | 8.00 | 9.34 | 13.00 | 52.00 | 4952 |
| Open Credit Lines | 0.00 | 4.00 | 7.00 | 8.00 | 11.00 | 48.00 | 4952 |
| Revolving Credit Balance | 0.00 | 769 | 3992 | 16827 | 14851 | 493300 | 4952 |
| Employment Status | 3446 | 770 | 26 | 134 | 103 | 368 | 4952 |
| Income | 1598 | 988 | 894 | 536 | 398 | 373 | 4952 |
| Principal Balance | 0.00 | 2108 | 3753 | 5514 | 7119 | 25000 | 4952 |

## 

## **Appendix 3. Histograms showing distribution of data**

**Chart, histogram

Description automatically generated**

## 

## **Appendix 4. Bar plots showing borrowers’ characteristics**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Appendix 5. Correlation of the variables**

A picture containing text, white, light, keyboard

Description automatically generated