**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

[Introduction 3](#_Toc99508887)

[Objectives 3](#_Toc99508888)

[Methodology 3](#_Toc99508889)

[Description of data and variables 3](#_Toc99508890)

[Data reformatting 4](#_Toc99508891)

[Summary statistics 4](#_Toc99508892)

[Exploratory Data Analysis 5](#_Toc99508893)

[Histograms 5](#_Toc99508894)

[Boxplots 6](#_Toc99508895)

[Correlation Plots with coefficients 7](#_Toc99508896)

[Target distribution over home ownership and Employment 8](#_Toc99508897)

[Visual Representation of Income for Bad and Good Customers 13](#_Toc99508898)

[Conclusion 15](#_Toc99508899)

# 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 data. We have used various statistical methods like Logistic Regressions Model, Tree model to meet our 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 examine ‘directionality’ of the input variable
* To create a ‘modeling’ dataset of data with variables of interest
* To identify important predictors that facilitates loan defaulting in the future
* To perform exploratory data analysis
* To predict variable of interest using logistic regression and tree models
* To create ROC curve and KS curve

# Methodology

## Description of data and variables

In our study, 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. 0 means the customer won’t pay back the money in the future and 1 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

We don’t need all the data of our dataset, so we need to reformat the data to keep our variable of interest and filter most important factors to predict the loan default. After removing the unwanted variables in our model prediction, we came up with 15 variables with 14289 number of observations. We also deleted the observations with missing values.

## Summary statistics

The summary statistics of our data set are given below:

**Table 1.** Summary Statistics of the variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Min.** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Max.** |
| **Debt to Income Ratio** | 0.00 | 0.14 | 0.22 | 0.35 | 0.32 | 10.01 |
| **Amount Borrowed** | 1000 | 3000 | 5000 | 6742 | 8850 | 25000 |
| **Current Delinquencies** | 0.00 | 0.00 | 0.00 | 1.14 | 1.00 | 50.00 |
| **Delinquencies Last 7 years** | 0.00 | 0.00 | 0.00 | 5.70 | 6.00 | 99.00 |
| **Public Records Last 10 Years** | 0.00 | 0.00 | 0.00 | 0.42 | 1.00 | 30.00 |
| **Inquiries Last 6 Months** | 0.00 | 0.00 | 2.00 | 2.75 | 4.00 | 46.00 |
| **Revolving Credit Balance** | 0 | 1169 | 5000 | 15478 | 14821 | 1435667 |
| **Bank Card Utilization** | 0.00 | 0.24 | 0.64 | 0.57 | 0.89 | 5.95 |
| **Length Status Months** | 0.00 | 0.00 | 0.00 | 29.82 | 31.00 | 554.00 |
| **Principal Balance** | 0 | 1629 | 2865 | 4104 | 5234 | 25000 |

**Table 2.** Summary Statistics based on employment status

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Employment Status** | **Debt to Income Ratio** | **Amount Borrowed** | **Current Delinquencies** | **Delinquencies Last 7 Years** | **Public Records Last 10 Years** | **Inquiries Last 6 Months** |
| **Full-time** | 0.29 | 6729.0 | 1.17 | 5.81 | 0.42 | 2.79 |
| **Not employed** | 3.17 | 3879.0 | 1.56 | 6.04 | 0.33 | 2.05 |
| **Part-time** | 0.58 | 4679.0 | 0.68 | 3.84 | 0.21 | 1.72 |
| **Retired** | 0.50 | 4914.0 | 1.69 | 7.51 | 0.63 | 2.23 |
| **Self-employed** | 0.87 | 9185.0 | 0.86 | 4.61 | 0.49 | 3.04 |

**Table 3.** Summary Statistics based on home ownership

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Is Borrower Home Owner** | **Debt to Income Ratio** | **Amount Borrowed** | **Current Delinquencies** | **Delinquencies Last 7 Years** | **Public Records Last 10 Years** | **Inquiries Last 6 Months** |
| **FALSE** | 0.33 | 5385.0 | 1.32 | 6.35 | 0.44 | 2.47 |
| **TRUE** | 0.38 | 8309.0 | 0.94 | 4.94 | 0.40 | 3.06 |

# Exploratory Data Analysis

For the exploratory data analysis, histograms and box plots are generated using hist and boxplot functions respectively in R-studio. To examine the relationship between variables, several scatterplots and correlation are created using plot and corPlot function in R-studio.

## Histograms

The figure below shows histograms of numeric variables in the data set. Most of the histograms seems skewed to right.

**Graphical user interface, application

Description automatically generated**

**Figure 1.** Histogram plots to visualize the distribution of the numeric variables in the data set

Among the histograms, we study the distribution for DebtToIncomeRatio. It is interpreted as ratio of total monthly payments towards loans to gross monthly income and is a decisive factor in assessing the credibility of the borrower. For Debt-to-Income ratio, this variable is capped at 10.01. That’s the reason it has a small peak at the point 10. 272 loaners have an extremely high debt level, higher than 10. The whole population has median of 22% (debt) and a mean of 28% (debt). And most people have a debt-to-income ratio less than 50%. It has few peaks, but the overall behavior is right skewed.

**Chart, bar chart, histogram

Description automatically generated**

**Figure 2.** Histogram of Debt-to-Income Ratio

## Boxplots

Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles and averages. The figure below shows boxplot of the numeric variables in the data set. We can see the median for “Good” and “Bad” of the variables are close to each other. But for most of them, we can see outliers which is defined as a data point that is located outside the whiskers of the box plot.

**Graphical user interface

Description automatically generated**

**Figure 3.** Boxplot of Numeric variables

## Correlation Plots with coefficients

The figure below shows correlation plot with correlation coefficients of the numeric variables. The correlation between the Amount Borrowed and Revolving Credit Balance is the strongest between other variables, which is 0.243, a moderate correlation.

A picture containing text, white, light, keyboard

Description automatically generated

**Figure 4.** Correlation plot

## Target distribution over home ownership and Employment

Here we are going to explore the people for two categories. First those who are opting for loan to renovation of home when they have a house. Second those who opt for home loans even though they don't have house.

**Table 4.** Home Ownership Category

|  |  |  |
| --- | --- | --- |
| **Category** | **ACTUAL NUMBER** | **PERCENTAGE** |
| **FALSE (Not homeowner)** | 7658 | 53.59% |
| **TRUE (Homeowner)** | 6631 | 46.40% |

Next, we can divide the dataset into 5 categories based on their employment status.

**Table 5. Employment Status Category**

|  |  |  |
| --- | --- | --- |
| **Employment Status** | **ACTUAL NUMBER** | **PERCENTAGE** |
| **Full-time** | 12508 | 87.54% |
| **Not employed** | 75 | 0.52% |
| **Part-time** | 551 | 3.86% |
| **Retired** | 307 | 2.15% |
| **Self-employed** | 8 | 5.93% |

The pie chart below depicts the percentage of employment status.

**Figure 5.** Employment Status based on Prosper Data

We can further categorize based on income range (Low Range, Mid-Range, Upper range) of employment status.

**Table 5. Employment Status based on Range**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Low Range** | | | **Mid-Range** | **Upper Range** | | |
| **Employment Status** | **0** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Full-time** | 65 | 1141 | 5097 | 3499 | 1499 | 1207 | 0 |
| **Not employed** | 3 | 8 | 4 | 0 | 0 | 0 | 60 |
| **Part-time** | 10 | 379 | 133 | 21 | 5 | 3 | 0 |
| **Retired** | 1 | 136 | 128 | 30 | 8 | 4 | 0 |
| **Self-employed** | 18 | 181 | 259 | 155 | 85 | 150 | 0 |

In our data, the income variable was termed as ‘Categorical Data’. So, we have further classified our Income variable into three levels.

* Low Range Income: 0-3
* Mid- Range Income: 4
* Upper Range Income: 5-7

Out of 12508 customers, 6303 people are within low range income and 2706 people are in the upper range income level. Number of low range income person are almost three times greater than the upper range class which is normal.

**Figure 6.** Relationship between Employment Status vs. Income

Now, we plotted boxplots for DebtToIncome Ratio and AmountBorrowed vs Homeowener, we can see that the median for of the variables are close to each other indicating no difference.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

**Figure 7.** Boxplots for Debt To Income Ratio and Amount Borrowed vs Homeowener,

Next, we plotted parallel boxplots for region vs Amount Borrowed. In the plot, all the boxplots appear to have similar centers. All the boxplots are reasonably symmetric, but boxplot for West Region is skewed to the right. We can see obvious outliers in all the samples.

Chart, box and whisker chart

Description automatically generated

**Figure 8.** Boxplots for Region vs Amount Borrowed

Chart, box and whisker chart

Description automatically generated

**Figure 9.** Boxplots for Employment Status vs Debt To Income Ratio

Amount Borrowed is almost equal for every employment status category except Self-Employed. The variance within the self-employed is high.

|  |  |  |
| --- | --- | --- |
| Category | Number | Percentage |
| Bad | 3897 | 27.27% |
| God | 10392 | 72.72% |

**Table 7. Target variables’ statistics**

About 72.72% of customers using credit card turned out to be good and the bad percentage is low with a rate of 27.27% which is about one-third of good customers.

**Figure 10.** Statistics of the target variable

**Table 8.** Summary Statistics of the targeted variable ‘Good’

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Min.** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Max.** |
| **Amount Borrowed** | 1000 | 3000 | 5000 | 6543 | 8200 | 25000 |
| **Revolving Credit Balance** | 0 | 1349 | 5430 | 15394 | 15101 | 1435667 |
| **Principal Balance** | 0 | 1454 | 2705 | 3483 | 4446 | 16755 |

# Visual Representation of Income for Bad and Good Customers

We have defined ‘Good’ and ‘Bad’ based on our target variables and considered few parameters to visualize either there is any pattern or not.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated | Chart  Description automatically generated |

**Figure 11:** Bar plot for Income and Employment on Bad Loan Customers

If we look at the left figure, we can see that most of the the people with a mid-range income are not paying back on time. But with high income rate, we can say that they tend to be good on paying loans on time. Right figure tells us something which is not expected. Although people are employed, they are not paying back. So, variable ‘Employment’ can’t be enough to say either the customer is bad or good.

Chart, histogram

Description automatically generated

**Figure 12:** Histogram for Amount Borrowed By ‘Good’ Customer

We have also analyzed customer behavior who are called ‘Good’. Left one depicts that the data is right skewed. Most good customers borrowed money within a range of $ (0-10,000). We can say that it might be correlated as being good or bad. More than 2000 people borrowed money less than $5000.

# Conclusion

The study conducts exploratory data analysis using the company data to predict the probability that a customer will go ‘bad’ and don’t pay back to the company. Data reformatting and data processing is done before performing analysis in the dataset. In our model, the dependent variable is ‘Bad’ whereas all other variables of the given data could be considered as independent variables.