GERMAN CREDIT DATA SCORING USING R

Business Data Analytics Project Report

Submitted by

Laltendu Das [15MCMI22]
Uma Revathi [15MCMI22]
Rosni K V [15MCMI15]

Under the Guidance of

Dr. V.Ravi, Associate Professor, IDRBT

November 2015

Table of Contents

1 Abstract 4

2 Introduction 4

3 Problem definition 4

4 About the data 5

4.1 Preparing data 6

5 Methodology 6

5.1 The model 6

5.2 Data product 6

6 Result and Discussions 6

6.1 Including tables 6

6.1.1 Book quality table 7

6.2 Including figures 7

6.3 How the data product works 7

7 Conclusion 7

8 References 8

List of Figures

1 Awesome figure caption 8

List of Tables

1 This table caption really describes what this table is about and what interesting facts it is revealing. 7

1 Abstract

Credit scoring uses quantitative measures of the performance and characteristics of past loans to predict the future performance of loans with similar characteristics.The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively. As a result, various kinds of credit scoring models are established to evaluate the customers' credit rank.

A credit scoring system should be able to classify customers as good credit those who are expected to repay on time and as bad credit those who are expected to fail. A major problem for banks is how to determine the bad credit, because bad credit may cause serious problems in the future. This leads to loss in bank capital, lower bank revenues and subsequently increases bank losses, which can lead to insolvency or bankruptcy. The categorisation of good and bad credit is of fundamental importance, and is indeed the objective of a credit scoring model. Classification models for credit scoring are used to categorize new applicants as either accepted or rejected with respect to these characteristics. In some cases the final selection of the characteristics was based on the statistical analysis used, i.e. logistic regression, neural network etc.

This study illustrates the use of data mining techniques to construct credit scoring models. Also, it illustrates the comparison of credit scoring models to give a superior final model. The report also highlights each data mining approach using R language.

2 Introduction

Credit scoring is one of the applications for predictive modeling, to predict whether or not credit extended to an applicant will likely result in profit or losses for the lending institution. For instance when a bank provides money to an individual, and expects to be paid back in time with interest commensurate with the risk of default. When a bank grants loan to a new customer, bank uses techniques on the large sample of previous customers with their application details and subsequent credit history available. Applying techniques results in connection between the characteristics of the customers.

Banks use credit risk modeling in order to measure the amount of credit risk which they are exposed to. The most commonly used technique for this purpose is logistic regression. In our study, we applied different techniques like support vector machines, nearest neighbor, decision trees on data to classify the borrowers as good or bad. So that the borrowers which are classified as bad are not granted any credit.

For the experiments, we used the German credit data set which was available in the UCI Repository. The data set consists of 20 attributes (7 numerical and 13 categorical) and there are totally 1000 instances (300 bad and 700 good cases). It was produced by Strathclyde University and is associated with several academic work.

The models were compared based on their accuracy on the German credit data set by using 10-fold cross validation. We divided the data set into ten partitions. Then, we iteratively took one of the ten partitions as the test set and the combination of the other nine were used to form a training set. The accuracy of a hold-out partition was defined as the number of correct classifications over the total number of instances in the partition. Accuracy of the 10-fold cross validation procedure was calculated by dividing the sum of the accuracies of all hold-out partitions by ten.

3 Problem definition

To develop a credit scoring model to predict the credit risk of applicants as bad risk(default) and good risk, which will help credit providers decide whether to grant loan to customers or not. The associated task for this problem is classification, and the German Credit Data set(source::UCI Machine Learning Repository) is using.

4 About the data

To meet with the objective of the analysis, ie, from credit providers perspective, to minimize loss they needs a decision rule regarding who to give approval of the loan and who not to. German Credit Classification dataset obtained from the UCI(University of California,Irvine)Machine Learning Repository, was used in this study. The number of examples in this dataset is sufficient and its values for each attribute are complete or available.

The number of examples in the dataset is 1000.The dataset is classified into two classes:good and bad class. The good class has 700 examples whereas the bad one has 300. The dataset has 20 attributes, Seven of the attributes are of continuous(numerical) types, while the other 13 are of categorical types. The summary of data is given below:

```
data<-read.csv("/root/BDA_project/Data/german.csv")
summary(data)

##  status       duration     history      purpose        credit
##  A11:274   Min.   : 4.0   A30: 40   A43    :280   Min.   :  250
##  A12:269   1st Qu.:12.0   A31: 49   A40    :234   1st Qu.: 1366
##  A13: 63   Median :18.0   A32:530   A42    :181   Median : 2320
##  A14:394   Mean   :20.9   A33: 88   A41    :103   Mean   : 3271
##            3rd Qu.:24.0   A34:293   A49    : 97   3rd Qu.: 3972
##            Max.   :72.0             A46    : 50   Max.   :18424
##                                     (Other): 55
##  bonds     jobex          rate         sex      guarantor    residence
##  A61:603   A71: 62   Min.   :1.000   A91: 50   A101:907   Min.   :1.000
##  A62:103   A72:172   1st Qu.:2.000   A92:310   A102: 41   1st Qu.:2.000
##  A63: 63   A73:339   Median :3.000   A93:548   A103: 52   Median :3.000
##  A64: 48   A74:174   Mean   :2.973   A94: 92              Mean   :2.845
##  A65:183   A75:253   3rd Qu.:4.000                        3rd Qu.:4.000
##                      Max.   :4.000                        Max.   :4.000
##
##  property        age          install     house        nocredit
##  A121:282   Min.   :19.00   A141:139   A151:179   Min.   :1.000
##  A122:232   1st Qu.:27.00   A142: 47   A152:713   1st Qu.:1.000
##  A123:332   Median :33.00   A143:814   A153:108   Median :1.000
##  A124:154   Mean   :35.55                         Mean   :1.407
##             3rd Qu.:42.00                         3rd Qu.:2.000
##             Max.   :75.00                         Max.   :4.000
##
##    job            no           ph        nri       credibility
##  A171: 22   Min.   :1.000   A191:596   A201:963   Min.   :1.0
##  A172:200   1st Qu.:1.000   A192:404   A202: 37   1st Qu.:1.0
##  A173:630   Median :1.000                         Median :1.0
##  A174:148   Mean   :1.155                         Mean   :1.3
##             3rd Qu.:1.000                         3rd Qu.:2.0
##             Max.   :2.000                         Max.   :2.0
##
```

The data may not be tidy and we may have to preprocess the data before our analysis can be done. We will discuss how we prepared the data in the following section.

4.1 Preparing data

After examine the whole data,it is found that there is no missing values for all attributes. The next step in this study is the statistical analysis of the data.

5 Methodology

This section will include the methods you are planning to use for your analysis. You should include some theoretical justification here. For example, why you think the method is applicable, what are the assumptions about the methods, whether your data satisfies those assumption or not etc.

5.1 The model

These theories may require you to type mathematical equations and we need to refer them in the text like equation 1.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

where $\epsilon \sim N(0, 1)$.

You should discuss the exploratory steps and the logical conclusion of adopting equation 1 for fitting to your data. Clearly mention the conditions and the assumptions of the model. Do not write any result of the model in this section. This section is only for theoretical discussion and any results of these models should be discussed in results section.

5.2 Data product

You may end up building a data product in your project. You may discuss about the plan here.

6 Result and Discussions

In result section you can start with an overview of what you have found during the exploration of data.

6.1 Including tables

Include some summary tables of the data as as shown in table 1. Make sure you discuss about the table you have included and explain the facts it is revealing. You have to sell your table in a way that the reader will understand that this table was awesome and it reveals a fact the reader would otherwise not recognize.

Notice that we used the function `xtable()` form the **R** package `xtable` [1] to generate a pretty table. `knitr` does this using LaTeX codes generated by `xtable` and automatically put it in a nicer we and we don't have to worry about its position. Also notice how we write the caption of the table as well as refer the table 1 from the text.

```r
# Creating and printing summary data table
library(xtable)
summary_data <- apply(trees, 2, function(x) {
    return(c(Average = mean(x), Median = median(x), SD = sd(x), Range = range(x)))
})
print(xtable(summary_data, digits = 2, caption = paste("This table caption really",
    "describes what this table is about and what interesting facts it is revealing."),
    label = "summary-data"), caption.placement = getOption("xtable.caption.placement",
    "top"))
```

Table 1: This table caption really describes what this table is about and what interesting facts it is revealing.

	Girth	Height	Volume
Average	13.25	76.00	30.17
Median	12.90	76.00	24.20
SD	3.14	6.37	16.44
Range1	8.30	63.00	10.20
Range2	20.60	87.00	77.00

6.1.1 Book quality table

We can add tables that look like the tables in the book. For this we need to add package `booktabs` in the preamble of this .Rnw file. This will include a package called `booktabs` onto LaTeX. Once we add that we can now put option `booktabs = TRUE` in the **R** code as below.

```
library(knitr)
x <- head(mtcars)
kable(x,format = ’latex’, booktabs = TRUE)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

6.2 Including figures

Please don’t forget to add nice data plots in your documents. Plots are nice to conveying message and much better than tables. Discuss what facts the figure is revealing and refer the figure from the text as figure 1.

```
plot(trees)
```

6.3 How the data product works

If you build a data product you may discuss here how it works and what it provides. For data product being your main purpose, your main section may be different from just saying `Results`. You may think how you rename your sections to naturally fit in your work and the purpose.

7 Conclusion

The conclusion is an elaboration of your abstract. Here you will discuss what you have done and how. The gist of the results need to be mentioned here. It needs to be convincing and the reader will never regret forgetting the date. Please keep it in mind that there may be readers who only read your conclusion. So, make your conclusion complete so that no reader misses anything even if they don’t want to read the whole document.

Each paragraph of the conclusion may discuss one result you have found or one concept you are proposing. Discuss your findings and why it is better and how it is compared to any existing methods may exist.

Please don’t forget to cite the works of others if you used it in your analysis. The citation is important for two reasons. Fist of all it acknowledges the good works other people have done which encourages them

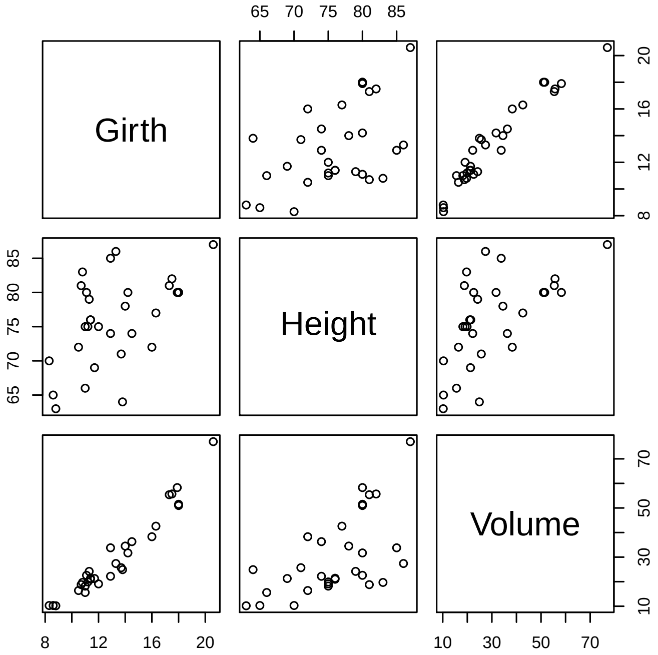

Figure 1: Awesome figure caption

keep continue doing their good work. Second, it protects you from plagiarism which is a very nasty task everyone should avoid.

There should be one paragraph about the future direction of the work you have done. You would like to make it so fascinating that the reader would wish to be involved in this work in future.

Finally this is just a template. Your exact document may have a very different outlook. It demonstrates how you can start to write a document. Our biggest problem is to figure out where to start from. And this documents provides a guide for that. I hope it turns out to be helpful for some of the readers. If you have any comments or concern about this document please let me know so that I can improve this document.

8 References

References

[1] David B. Dahl, *xtable: Export tables to LaTeX or HTML*, R package version 1.7-3, http://CRAN.R-project.org/package=xtable, 2014

[2] Leslie Lamport, *LaTeX: A Document Preparation System*. Addison Wesley, Massachusetts, 2nd Edition, 1994.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ , 2014

[4] Yihui Xie *knitr: A general-purpose package for dynamic report generation in R*, http://yihui.name/knitr/, 2014