Loan Defaulter's Application in R Programming

Mahesh Mardolkar
Principal
Bharatesh College of Computer Applications
Belagavi, Karnataka, India- 590 016

Abstract - The bad loan in India's banking industry is widening and even the capital burden of these banks would increase multifold due to bad loans. CIBIL(Credit Information Bureau(India) Limited) is assisting the banks to evaluate and approve loan applications. The paper "Loan Defaulter's application in R programming" which help the bank to predict defaulters using K Nearest Neighbor classification for clear loan repayment process

Keywords: CIBIL, KNN Classification, R programming

I. INTRODUCTION

The bad loan situation in India's banking industry is around Rs. 4.4 Lakh Crore declared as bad loan till December 2015, solid action against the loan recovery for defaulters is unfortunately missing. The bad loan issue of banks is still pending and the RBI(Reserve Bank of India) finally has put a deadline of March 2017 for banks to have clear balance sheet. The banks are now disclosing their NPA(Non Performing Assets) to comply with RBI.
With total bad loan stock of Rs. 4,43,691 Crore in banking industry, the capital burden of these bank would increase multifold. The government has infused Rs. 90,000 Crore in India's 27 public sector banks in the last eight years

II. CIBIL(CREDIT INFORMATION BUREAU(INDIA) LIMITED)

Founded in August 2000 CIBIL is India's first Credit Information Company(CIC). Records of loan and credit card payment of an individual is collected and maintained by CIBIL. The member banks and credit institution submit the records to CIBIL on a monthly basis, which is used to create Credit Information Report(CIR) in order to help evaluate and approve loan applications credit scores are provided to credit institutions.

III. CIBIL SCORE

The very crucial and important part of everyone's financial journey is CIBIL score or CIBIL records, when availing loan or credit card, person's eligibility is determined on the CIBIL score or CIBIL records. The person's credit history is represented in numeric form as CIBIL score, to determine loan or credit card eligibility the banks and financial institution depends on CIBIL score. In order to be eligible for credit an individual should maintain average or above average CIBIL score. In India 750 and above is considered as average CIBIL score and anything below that is considered as poor CIBIL score, when it comes to lenders personal loans is at high risk for CIBIL defaulters.

IV. R PROGRAMMING LANGUAGE

R is a programming language used for developing statistical software and data analysis, widely used among statisticians and data miners. R programming is basically used for statistical computing. Ross Ihaka and Robert Gentleman created R at University of Auckland, New Zealand, which is currently developed by R development core team. R is an implementation of S programming which was created by John Chambers at Bell Labs.

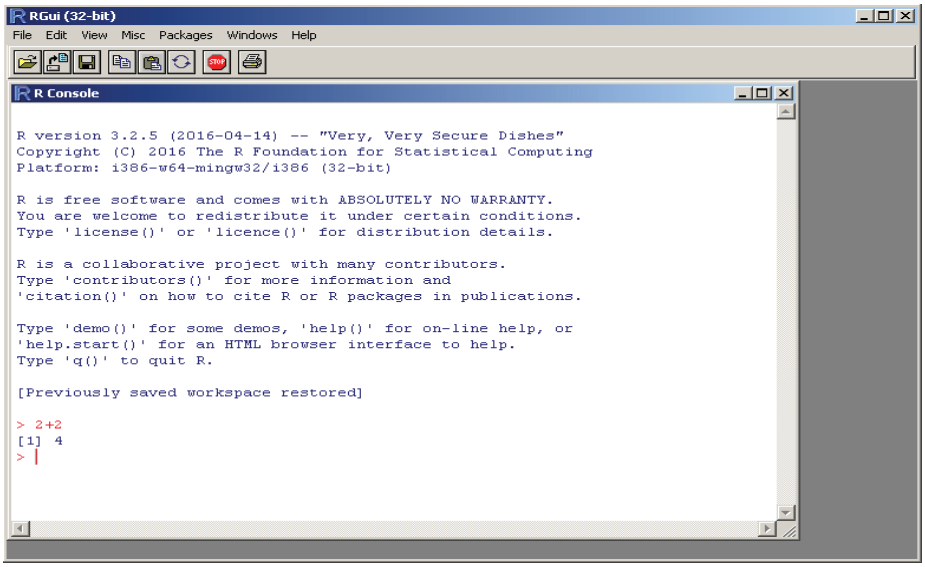

Fig. 1- R programming console

R is freely available under GNU General Public License, R is a GNU project whose software environment is written in C, Fortran and R.

V. Statistical Features

A wide variety of statistical and graphical techniques is implemented using libraries in R which includes linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering. R can also produce static graphs of high quality including mathematical symbols.

VI. Programming Features

Through command line interpreter, since R is a interpreted language, if user types 2+2 on command prompt and presses enter, the computer replies with 4 as given in Fig. 1.
R also supports matrix arithmetic, data structures which include vectors, matrices, arrays, data frames which makes it similar to APL and MATLAB.

VII. Packages

Through user created packages the different capabilities of R are extended. These packages are developed in R and other languages like JAVA, C, C++ and Fortran. Packages allow specialized statistical techniques to be incorporated easily like graphical devices (ggplot2), Import/ Export capabilities, reporting tools (knitr, Sweave), etc. R installation includes core set of packages and additional packages of more than 7,801 available at Comprehensive R Archive Network(CRAN), Bioconductor, Omegahat, GitHub and other repositories.

VIII. K Nearest Neighbor Classification

In the beginning of 1970's KNN has been used in statistical estimation and patterns recognition as a non-parametric technique.

Algorithm

The distance functions stated below are only valid for continues variables. Hamming distance is used in case of categorical variables, it also brings the issue of standardization of variables between 0 and 1 when there is a mixture of numerical in the dataset and categorical variables

Fig. 2- Console Script of Loan defaulters in R programming

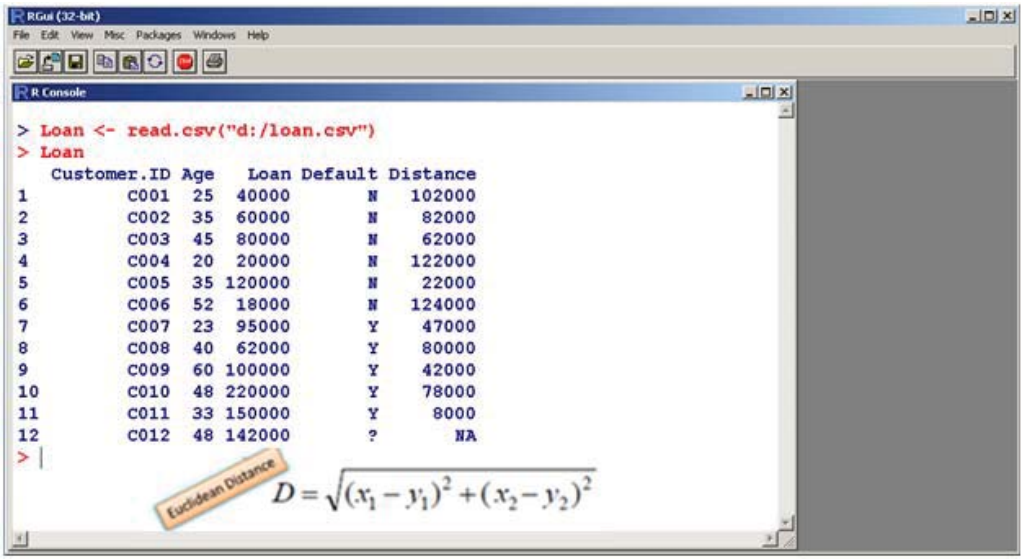

Fig. 3- Reading Bank loan database created in MS Excel – Loan.csv

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y

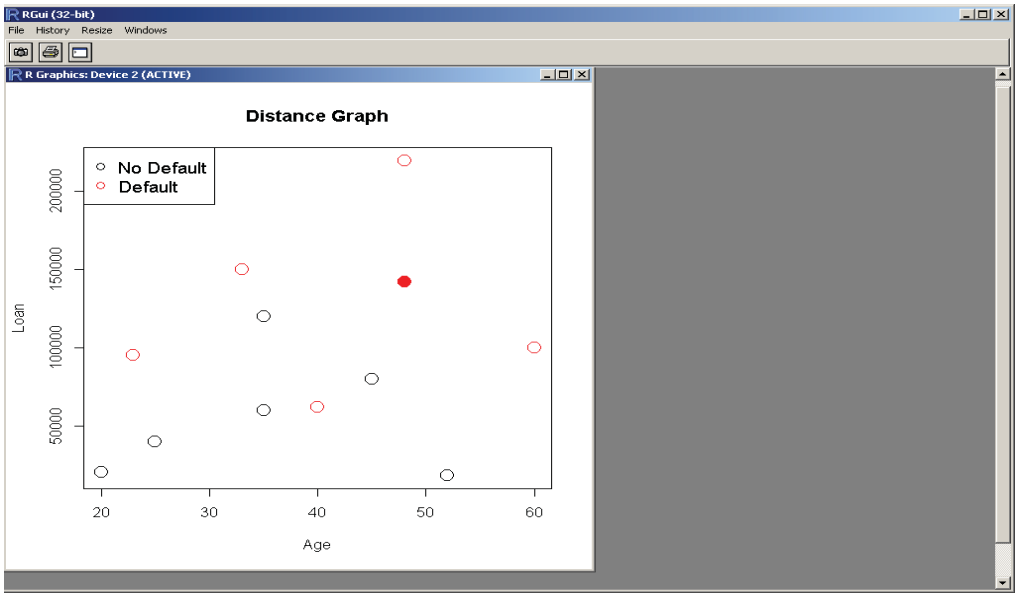

Fig. 4- Graph representing loan defaulters

IX. STANDARDIZED DISTANCE

Considering two variables one with annual income in rupees and other based on age in years, then income will have much higher influence on the distance calculated. The solution is to standardize the training set as give below.

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

Table 1- Standard Distance Table

X. CONCLUSION

Bad loan is a major worry for the Banks in India, CIBIL can track for prospective customers for the Banks, using KNN algorithm the bank can predict the behavior of its customer towards loan repayment. R programming supports high quality graphs and has a unique feature of graphical representation to identify loan defaulters.

REFERENCES

[1] Discovering knowledge in Data – Daniel T. Larose, Chantal D. Larose – Second Edition John Wiley
[2] The Elements of Statistical Learning - Trever Hstie, Robert Tibshirani, Jerome Friedman – First Edition 2001 Springer
[3] Learning R – Richard Cotton – O' REILLY
[4] Introductory Statistics with R – Peter Dalgaard – Second Edition Springer
[5] http://www.cibil.com
[6] http://en.wikipedia/org/wiki/CIBIL.