INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 5)
Available online at: www.ijariit.com

Calculation of client credit risk prediction in banking sector using data mining

Manpreet Kaur
mkaur272382@gmail.com
Institute of Engineering and Technology, Bhaddal, Punjab

Gurinderpal Singh
gurinderpal.singh@ietbhaddal.edu.in
Institute of Engineering and Technology, Bhaddal, Punjab

ABSTRACT

The use of credit scoring can be used to assist the analysis of credit risk in assessing the eligibility of the applicant. As a valuable method for credit rating, data mining has been proven. Many credit scoring models for determining the creditworthiness of loan applicants have been established over the last few years. With an obligation to repay over a period of time, credit provides access to capital today, There may be financial sources, or they may consist of goods or services. Today, credit has become a very necessary component of daily life. Although credit cards are currently the most common type of loan, other credit plans include, among others, residential mortgages, car loans , student loans, small business loans, commercial lending and bonds.

Keywords: *Credit Card, Financial, DecisionTree, DataMining, Prediction*

1. INTRODUCTION

The Business of providing loan is increasingly flattering a foremost goal for many banks; as a consequence there is lofty rivalry in the middle of the monetary establishments in India foremost to non-payment of the majority loans. In sort to lift the excellence of charitable loans and decrease the peril engross in charitable advance, credit score sculpt have been urbanized by banks and researchers to get enhanced the procedure of analysis credit value throughout the credit assessment procedure. This study uses chronological data on expenditure, demographic individuality and numerical system to create logistic deterioration model and to make out the imperative demographic distinctiveness correlated to credit threat. The logistic failure model was used to devise a logistic deterioration model calculator, which was used to estimate the likelihood of evasion. Customer's age, sex, employment, quantity of reliant, marital class and sum of credits are used. Married patrons fail to pay additional the patrons who are not marital and the privileged number of reliant, the upper the non-payment tempo. The self-employment customers evade extra than wage earners. It was found out that the elevated the sum of loan composed the superior the likelihood of default.

2. REVIEW SURVEY

K. Kavitha, 2016: The corresponding credit risk cost for three real-life credit scoring data sets. Besides the well-known classification algorithms (e.g. linear discriminate analysis, logistic regression, neural networks and k-nearest neighbour), we also investigate the suitability and performance of some recently proposed, advanced data mining techniques such as support vector machines (SVMs), classification and regression tree (CART), and multivariate adaptive regression splines (MARS). Using the classification accuracy and cost of credit scoring errors assesses the performance.

Adnan Dzelihodzic, Dzenana Donko: Credit risk assessment is very important research field with wide application in the practice. Even if there is a hundreds of research, models and methods, it is still hard to say which model is the best or which classifier or which data mining technique is the best. Each model depends on particular data set or attributes set, so it is very important to develop flexible model, which is adaptable to every dataset or attribute set. In order to have better accuracy of model every model should be tested by credit staff because their knowledge can help to improve our models.

Ling Kock Sheng1 and Teh Ying Wah: This observation and relationship could be validated if more records are obtained for testing the models. This could further improve predictive accuracy. It will be good if an incremental decision tree technique is made available to see if it will exceed what the current tool is able to improve on. The successful completion of this research has provided new insights into the use of data mining tools and techniques in a bank. Factors such as the sample size, the equal-length partition sizes and the training and testing partition size have had an effect on predictive accuracy. The objectives put forward was successfully met with the identification of a batch classifier and model that attain a higher level of predictive accuracy beyond that is initially expected.

3. PROBLEM DEFINITION

Valuable information from historical data can yield knowledge and predict future attributes with the help of data mining. A decision support system based on data mining techniques can be employed to improve the quality and to enhance the performance of credit risk processing in banking sector. The objectives for this synopsis are as follows:

- To study and collect information from the data-set of customers.
- To identify the best classifier algorithm into good and bad classes.

4. METHODOLOGY

Step 1 Data selection.
Step 2 Data pre-processing.

Splitting training and test datasets.
Step 3 Feature selection.
Step 4 Identify best classification model.
Step 5 Predicting class levels of test datasets and evaluate prediction.

5. OPEN SOURCE TOOLS FOR DATA MINING

Data mining is an array of a mixture of performance like blueprint enjoyment, statistics, machine learning, etc. Whereas there is a good amount of association flank by mechanism learning and data mining, as in cooperation go hand in hand and machine learning algorithms are used for mining data and will imprison in this leader to only those tools exacting for data mining.

5.1 Weka

Weka is a Java based free and open source software comprises a collection of machine learning algorithms for data mining. It posts tools for data pre-processing, classification, regression, clustering, association rules and visualization. It lets you import the raw data from a variety of file formats, and ropes well known algorithms for different mining actions like filtering, clustering, classification and attribute selection.

5.2 Weka Results Output: Write Parameters used

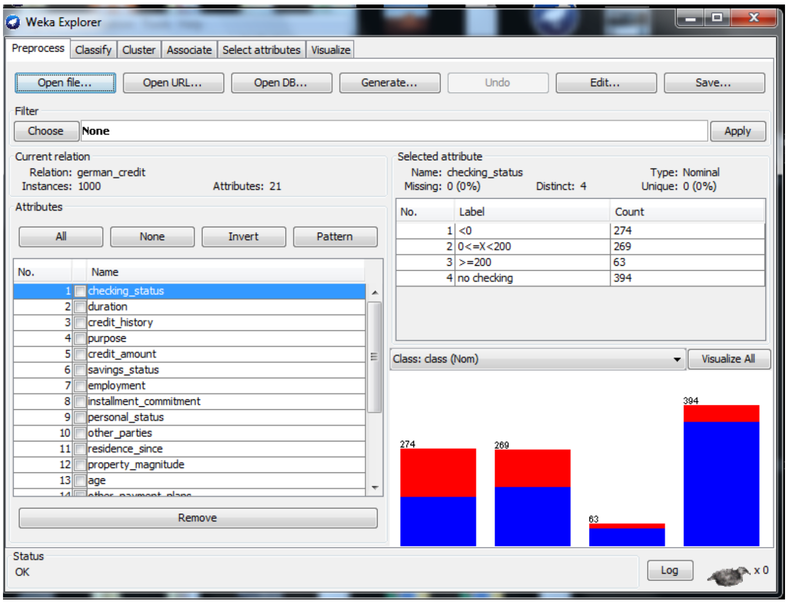
TP = true positives: number of examples predicted positive that are actually positive
FP = false positives: number of examples predicted positive that are actually negative
TN = true negatives: number of examples predicted negative that are actually negative
FN = false negatives: number of examples predicted negative that are actually positive

5.3 Accuracy

Accuracy is how lock a calculated rate is to the actual rate.
Accuracy is distinct in the stipulations of correctly confidential occurrence divided by the total number of instances at hand in the dataset.
Accuracy= (TP+TN)/total
Where TP- True Positive, TN- True Negative

6. RESULT IMPLEMENTATION

All the Class label outputs are copied to an excel file for each test instance or row, and the actual Class labels are also pasted into the same file in the next column. When comparing the two rows, this file is identified in reading by a java code and the correctly and incorrectly labeled instances.



Filter: None

Current relation
Relation: german_credit
Instances: 1000 Attributes: 21

Selected attribute
Name: checking_status Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count
1	<0	274
2	0<=X<200	269
3	>=200	63
4	no checking	394

Attributes

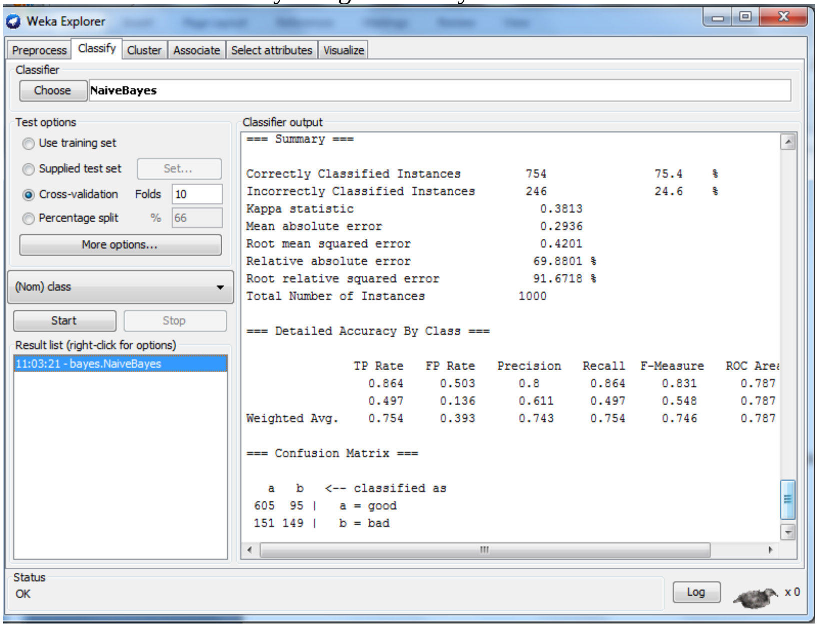
No.	Name
1	checking_status
2	duration
3	credit_history
4	purpose
5	credit_amount
6	savings_status
7	employment
8	installment_commitment
9	personal_status
10	other_parties
11	residence_since
12	property_magnitude
13	age
14	other_payment_plans

Class: class (Nom)

Status
OK

Fig. 1: Detail of Dataset

Prediction accuracy and confusion matrix obtained by using Naïve Bayes



Classifier: NaiveBayes

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

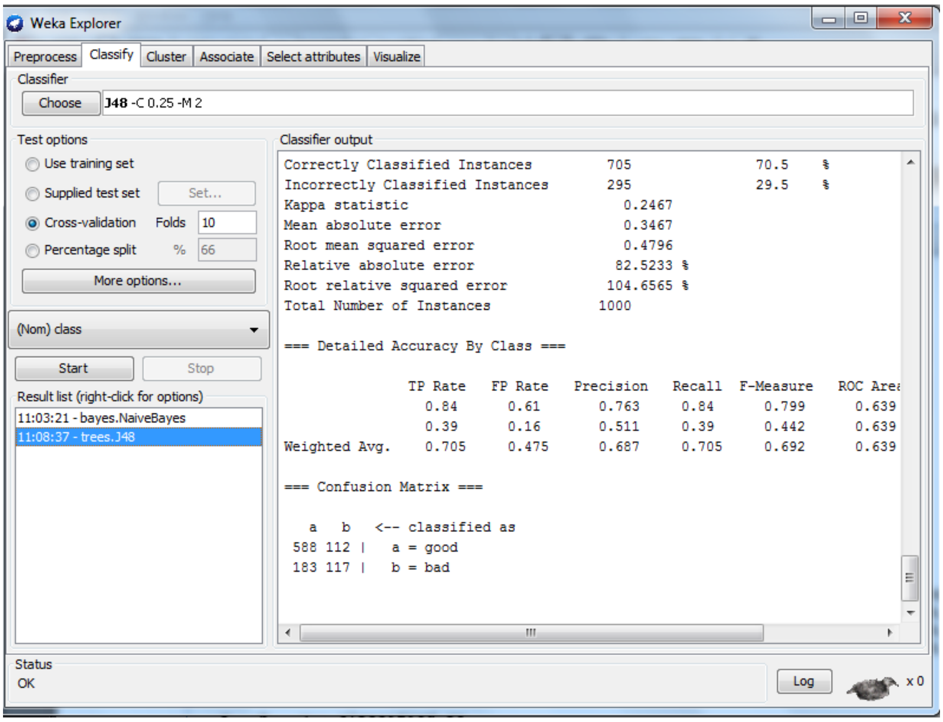
(Nom) class

Result list (right-click for options): 11:03:21 - bayes.NaiveBayes

Classifier output

```
=== Summary ===

Correctly Classified Instances         754               75.4    %
Incorrectly Classified Instances       246               24.6    %
Kappa statistic                          0.3813
Mean absolute error                      0.2936
Root mean squared error                  0.4201
Relative absolute error                 69.8801 %
Root relative squared error             91.6718 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.864     0.503      0.8       0.864     0.831      0.787
                 0.497     0.136      0.611     0.497     0.548      0.787
Weighted Avg.    0.754     0.393      0.743     0.754     0.746      0.787

=== Confusion Matrix ===

   a   b   <-- classified as
 605  95 |   a = good
 151 149 |   b = bad
```

Naïve – Bayes Accuracy – 75.4% accuracy and confusion matrix obtained by using J48 (Decision Tree)



Classifier: J48 -C 0.25 -M 2

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

Result list (right-click for options): 11:03:21 - bayes.NaiveBayes; 11:08:37 - trees.J48

Classifier output

```
Correctly Classified Instances         705               70.5    %
Incorrectly Classified Instances       295               29.5    %
Kappa statistic                          0.2467
Mean absolute error                      0.3467
Root mean squared error                  0.4796
Relative absolute error                 82.5233 %
Root relative squared error            104.6565 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.84      0.61       0.763     0.84      0.799      0.639
                 0.39      0.16       0.511     0.39      0.442      0.639
Weighted Avg.    0.705     0.475      0.687     0.705     0.692      0.639

=== Confusion Matrix ===

   a   b   <-- classified as
 588 112 |   a = good
 183 117 |   b = bad
```

J48 Algorithm Result

Decision Tree Accuracy – 70.5%

Table 1: Correct and Incorrect Classified Result:

Classifier Algorithm	Correct classified	Incorrect Classified	Accuracy
Naïve Bayes Algorithm	745/1000	246/1000	75.4%
Decision Tree Algorithm	705/1000	295/1000	70.5%
Zero R Algorithm	700/1000	300/1000	70%

7. RESULTS AND DISCUSSION

The prediction accuracy of defaulter instances obtained by using Cost Sensitive Learning is considerably good as compared to results obtained by not using it. The overall classification results also seem to be relatively balanced. ‘Duration’, ‘employment’ and ‘age’ are the most important factors for predicting the class of the loan applicant (whether the applicant would ‘default’ or ‘not’) in case of the credit dataset. ’Job Personal_ status’ and own telephone’ are found to be the least significant attributes for prediction. Credit history’ and ‘interest rate’ are the most important factors for predicting the class of the loan applicant (whether the applicant would ‘default’ or ‘not’).

8. REFERENCES

[1] M. Sudhakar, and C.V.K. Reddy, “Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5(3), pp. 705-718, 2016.

[2] J. H. Aboobyda, and M.A. Tarig, “Developing Prediction Model Of Loan Risk In Banks Using Data Mining”, Machine Learning and Applications: An International Journal (MLAIJ), vol. 3(1), pp. 1–9, 2016.

[3] K. Kavitha, “Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016.

[4] Z. Somayyeh, and M. Abdolkarim, “Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran”, Journal UMP Social Sciences and Technology Management, vol. 3(2), pp. 307–316, 2015.

[5] Abell`an, J.n, J., Mantas, C., 2014. Improving experimental studies about ensembles of classier for bankruptcy prediction and credit scoring. Expert Systems with Applications 41 (8), 3825–3830.

[6] Sudhakar M and Dr. C. V. Krishna Reddy, "CREDIT EVALUATION MODEL OF LOAN PROPOSALS FOR BANKS USING DATA MINING," International Journal of Latest Research in Science and Technology, pp. 126-131, July 2014.

[7] Abhijit A. Savant and P. M. Chowan, "Comparison of Data Mining Techniques used for Financial Data Analysis," International Journal of Emerging Technology and Advanced Engineering, June 2013.

[8] .R.Mahammad Shafi, A Tool for Enhancing Business Process in Banking Sector, 3rd ed.: International Journal of Scientific & Engineering Research, 2012

[9] Abdoun, O., Abouchabaka, J., 2011. A comparative study of adaptive crossover operators for genetic algorithms to resolve the traveling salesman problem. International Journal of Computer Applications 31 (11), 49–57

[10] Classification methods applied to credit scoring: A systematic review and overall comparison: Francisco Louzadaa Anderson Arab Guilherme B. Fernandez, a Department of Applied Mathematics & Statistics, University of S˜ao Paulo, S˜ao Carlos, Brazil b Department of Statistics, Federal University of S˜ao Carlos, S˜ao Carlos, Brazil c P&D e Innovation in Analytics, Serasa-Experian, S˜ao Paulo, Brazil

[11] Data Mining Techniques for Credit Risk Assessment Task.: ADNAN DZELIHODZIC, DZENANA DONKO International Burch University Francuske revolucije bb, Ilidza, Sarajevo BOSNIA AND HERZEGOVINA