DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING

Aboobyda Jafar Hamid[1] and Tarig Mohammed Ahmed[2]

[1]Department of Computer Sciences ,University Khartoum, Sudan
[2]Department of Computer Sciences ,University Khartoum, Sudan

ABSTRACT

Nowadays, There are many risks related to bank loans, for the bank and for those who get the loans. The analysis of risk in bank loans need understanding what is the meaning of risk. In addition, the number of transactions in banking sector is rapidly growing and huge data volumes are available which represent the customers behavior and the risks around loan are increased. Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. In this paper a new model for classifying loan risk in banking sector by using data mining. The model has been built using data form banking sector to predict the status of loans. Three algorithms have been used to build the proposed model: j48, bayesNet and naiveBayes. By using Weka application, the model has been implemented and tested. The results has been discussed and a full comparison between algorithms was conducted. J48 was selected as best algorithm based on accuracy.

KEYWORDS

Loan Risk, Data Mining , Classification

1. INTRODUCTION

There are various areas in which data mining can be used in financial sectors like customer segmentation and profitability, high risk loan applicants, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimizing stock portfolios, cash management and forecasting operations, most profitable Credit Card Customers and Cross Selling. There are many different types of loans you have to take into account when you're looking to borrow money and it's important to know your options. Loan categorization refers to the process of evaluation loan collections and assigning loans to groups or grade based on the perceived danger and other related loans properties. The process of continual review and classification of loans enables monitoring the quality of the loan portfolios and to take action to counter fall in the credit quality of the portfolios. It is required for banks to use more complicated internal classification schemes than the more standardized schemes that bank managers need for reporting reasons and that are intended to make easy observing and interbank evaluation.

There are many types of loans such as: Open-ended loans are loans that you can have a loan of more and more. Credit cards and lines of credit are the famous types of open-ended loans. You have a credit limit that you can buy with both of these two types of loans. In any time you can purchase automatically your available credit will decreases. since you make expenditure, you're on hand increases permitting you to use the credit more and more. Closed-ended loans, this type of loans cannot be on loan once they've been repaid. while you make expenditure on closed-ended loans, the balance of the loan became downward. though, you don't have any existing credit you can employ on closed-ended loans. As an option, if you want to lend more money, you'd have to make application for other loan. widespread types of closed-ended loans involve auto loans, mortgage loans, and student loans. Secured loans

DOI:10.5121/mlaij.2016.3101

Secured loans are loans that rely on an asset. In the state of loan failure to pay, the lender can possess the asset and make use of it to cover up the loan. High wellbeing rates for secured loans may be lower than those for unsecured loans. The asset may need to be evaluated before you can have a loan of a secured loan. Unsecured loans lends may be more complicated to get and have higher concern rates. Unsecured loans rely just on your credit history and your revenue to meet the criteria for the loan. If you failed to return back unsecured loan, the lender has to wear out collection alternatives involving debt collector and claim to recover the loan. Conventional loans or mortgage loans are the loans that aren't insured by a government organization, country Housing Service, or the Veterans management. Conventional loans may be conforming, meaning they tag on the rule set onward by Fannie Mae and Freddie Mac. Non-conforming loans don't meet Fannie and Freddie qualifications [1].

There are many risks related to bank loans, for the bank and for those who get the loans. The analysis of risk in bank loans need understanding what is the meaning of risk. Risk is denotes to the probability of certain outcomes--or the uncertainty of them--especially an existing negative threat for trying to achieve a current monetary operation. Risk in bank loans involve: credit risk, the risk that the loan won't be return back on time or at all; liquidity risk, the risk that too many deposits will be withdrawn too quickly, leaving the bank short on immediate cash; and interest rate risk, the risk that the interest rates priced on bank loans will be too low to earn the bank adequate money [2].

There are two most important goals for data mining prediction and description. Prediction involves using some variables in data set to predict unknown values of other variables and Description concentrates on finding patterns describing the data that can be interpreted by human. Data mining is the process of extracting hidden pattern from large amount of data that used to take a write decisions. The derived knowledge must be new, not obvious, relevant and can be applied in the field where this knowledge has been obtained. It is also the process of extracting useful information from raw data.

Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. In present era, Data Mining is becoming popular in banking field because there is a call for efficient analytical methodology for detecting unknown and useful information in banks data. Skills and knowledge are important requirement for achieving Data Mining task since the success and failure of Data Mining is greatly reliant on the person who managing the process due to unavailability of standard framework. The CRISP-DM (CRoss Industry Standard Process for Data Mining) introduces a framework for doing Data Mining activities. CRISP-DM decomposes the data mining task into 6 phases. The first one is the understanding of the business activities. In the second phase the data used for business activities are collected and analyzed. Data pre-processing and modeling is performed in the third and fourth phase respectively. Fifth phase is the evaluates of the model and last one is the deployment of the designed model [3].

Data mining process consist of three phases:

- Data preparation
- The actual mining
- Interpretation of the results

Data mining techniques aid to distinguish between borrowers who pay back loans at the appointed time from those who don't. It also helps to expect when the borrower is at default, whether providing loan to a particular customer will result in bad loans. All processes related to banking sector could be analyzed using data mining to detect the customers behavior. It also helps to analyze whether the customer will make prompt or delay payment if the credit cards are sold to them.

2. RELATED WORK

Many researches have been conducted based on data mining in the field of financial and banking sector. This section presents briefly some of these techniques which are used in loans risk management and their finding
Sudhakar et al focused on specifying the data mining applications usefulness, these applications are using several data mining techniques such as decision trees and Radial Basis Neural Networks. This study came with in which way to apply these applications in a credit-risk assessment field. McLeod presents Neural networks properties and their fitness for the credit-granting process.

Barney et al made a comparison of the performance of regression analyses and neural networks to identify the farmers who will default on the loans of their Home Administration and those farmers who return back the loans as in the appointment. By using an unstable data, this study proofed that neural networks regarding better logistic regression to classify farmers into two groups, those who pay back on time and those who default to return their loans[8]

Glorfeld and Hardgrave (2001) proposed a complete and useful systematic way to produce an optimal design of a high performance model for neural network estimating of the Credit value related to applications of commercial loan. The neural network constructed using their design was able to classifying 75% of loan applicants correctly.

Tessmer checked credits that offered to small Belgian businesses and he used a decision tree-based learning way. His study focused on the Type I credit errors impact (he mean by type 1 taking good loans as bad loans), and Type II credit errors which mean regarding bad loans as good loans, on the accuracy, conceptual validity and constancy of the learning operation. This study has built on a previous research that compare the efficiency of several data mining tools in several credit risk estimation field.

Efficacy of neural networks and traditional techniques analyzed by Desai et al to build rating models for loans unions. he used consistent sample of data consists of 18 variables belong to three credit unions and his study proved that neural networks were more useful in bad loans detection, whereas logistic regression useful in discovering bad and good loans [9]

Jagielska et al investigated the abilities of neural networks in classifying loans risk, uncertain logic genetic algorithms, rule stimulation software, he concluded that the genetic approach is more favorably than the neuron fuzzy and rough set methods [9].

A study by A.J.Feelders and A.J.F.le Loux about conducted a case study for personal loan evaluation by using data mining techniques. the study carried out in Netherlands in ABN AMRO bank . Data mining capabilities were applied to assess the personal loans. Historical data of clients and their return-back activities and behaviors are used to predict whether a customer will default or not [8].

Emile J. Salame focused on objectives that provide powerful tool to help in decreasing number unauthorized borrowers whose impact on the financial institution is positive. In addition, the paper gave insight in agricultural loan data to help decision-makers and to increase their ability to manage the operation of lending farmers which decreased the time and cost of checking of loan ambiguity and help loan officers to a crucial diction toward the customers [10].

Michael D. Johnson, Anders Gustafsson studied broadly the usages of data mining techniques in banking sectors and its related impacts on several operations. They built a prediction model to predict if the customers will pay back the loans which they borrow using the techniques of neural network and classification.

Jozef Zurada and Martin Zurada focused on the usefulness of the tools of data mining such as decision trees and neural networks, in this study they check and examining how decision trees and neural networks applied in a credit-risk evaluation. Glorfeld and Hardgrave presented a powerful and complete approach. They aim to develop architecture of a neural network model for assessing the operation of creditworthiness related to the applications of commercial loans. The developed neural network model was able to classify 75% of the applicants [12].

3. PROPOSED MODEL

We mean by loan evaluation process, the sequence of steps that taken to take diced about granting a loan to the customer or not. When the customer apply for a loan granting application, the bank officer must investigate about what called 5 C's which are Character (or Credit History), Cash Flow (or Capacity), Collateral, Capitalization and Conditions. It is helpful for evaluation loan application and it regarded as a helpful framework for estimate the credit risk related to a probable creditor [16]

3.1 Dataset

A collection of data from banking sector has been selected. The Data set format that acceptable by Weka is (ARFF) which stand for Attribute-Relation File Format ARFF which composed of tags that include the name of attributes, types of attributes, the values and the data itself. After doing preprocessing tasks and applying feature selection to choose the most important attribute we used 8 attributes Credit_history, Purpose, Gender, Credit_amount, Age, Housing, Job and the Class. The following table describes them:

No	The attribute	Description	Data type
1	Credit_history	Previous history of customer credit	Nominal
2	Purpose	The loan purpose	Nominal
3	Gender	Male or female	Nominal
4	Credit_amount	The amount of credit	Numeric
5	Age	Customer Age	Numeric
6	Housing	Rent, own or for free	Nominal
7	Job	Is the customer has a job	Nominal
8	Class	The class of loan good/bad	Nominal

Table 1 Data set Description

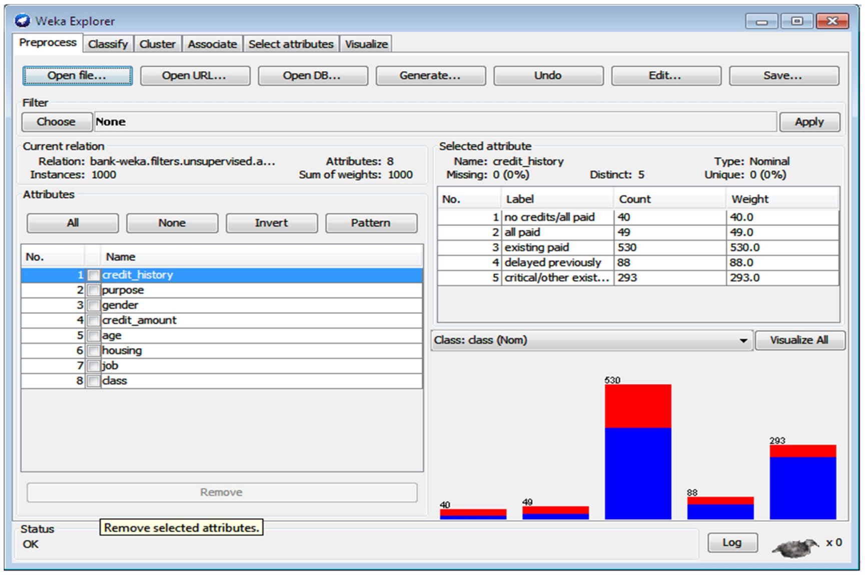
4. MODEL IMPLEMENTATION

The process of classification crowd the data set into groups of classes according to their dissimilarity. There are several classification algorithm or classifiers like Naïve Bayes Classifier, Neural Network Classifier, decision Tree Classifier. There are several algorithms in each of this technique which used to produce a model to predict the class of unknown class tables. The major goal of this algorithm is the provision by a model for predicting the class of unknown records.

The common steps for each of these classifiers consist of the following:

- Prepare the training set, a records that are already have known class label.
- Build the model by applying one of learning algorithm using training set.
- Applied the model upon unknown data test set class.
- Evaluate the accuracy of the model.

In our research we use three different classification algorithm to build three different models. These algorithms are j48, bayesNet and naiveBayes. We will give more details about them later. The data will appear in weka as shown in figure 1:



Filter: None

Current relation
Relation: bank-weka.filters.unsupervised.a...
Attributes: 8
Instances: 1000
Sum of weights: 1000

Attributes

No.		Name
1	□	credit_history
2	□	purpose
3	□	gender
4	□	credit_amount
5	□	age
6	□	housing
7	□	job
8	□	class

Selected attribute
Name: credit_history
Type: Nominal
Missing: 0 (0%)
Distinct: 5
Unique: 0 (0%)

No.	Label	Count	Weight
1	no credits/all paid	40	40.0
2	all paid	49	49.0
3	existing paid	530	530.0
4	delayed previously	88	88.0
5	critical/other exist...	293	293.0

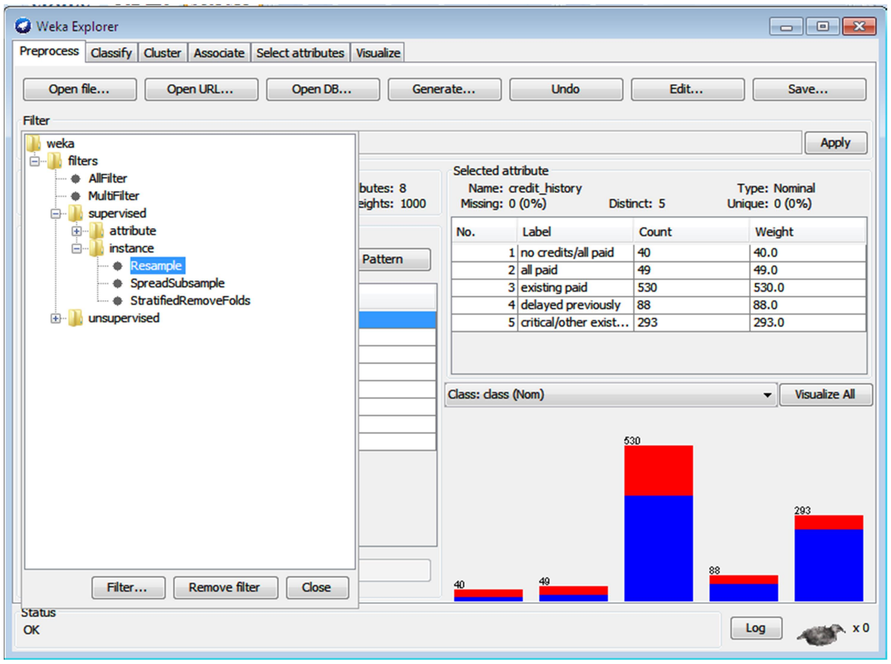
Class: class (Nom)

Status
OK
Remove selected attributes.

Figure 1 Data exploration

We divide the original data set into two groups, training set which represent 80% from all data and testing set which represent 20% of the data set. the operation of spiriting happened as follow using weka.

We open the original file as shown in figure 2 press choose bottom behind filter menu



butes: 8
eights: 1000

Selected attribute

Name: credit_history
Type: Nominal
Missing: 0 (0%)
Distinct: 5
Unique: 0 (0%)

No.	Label	Count	Weight
1	no credits/all paid	40	40.0
2	all paid	49	49.0
3	existing paid	530	530.0
4	delayed previously	88	88.0
5	critical/other exist...	293	293.0

Class: class (Nom)
Visualize All

OK

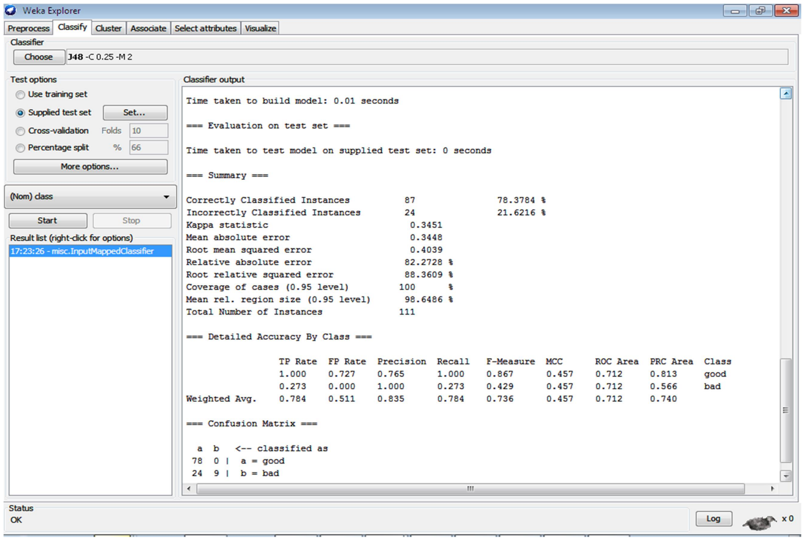
Figure 2 step one to spilt data set to training and test sets

As shown in figure 2 , the number of instance is 1000(see the red area),we want to spilt them into training set which represent 80% of the hole data, and test set which represent the rest(20%).
In this research we applied three different classification algorithms and make a comparison between them according to their accuracy in classifying the data correctly and these algorithms are j48, bayesNet and naiveBayes.

4.1 J48 classification algorithm

J48 is the enhanced edition of C4.5 algorithms or can be viewed as C4.5 implementation.J48 takes as an input the set of tables and generate a decision tree as an output. The generated decision tree is alike to the structure of tree. It consists of root, intermediate and leaf node. The nodes in the generated tree contain a decision which guide to the result. It split the input data set into mutually exclusive sets, each set with a label. Splitting measure is applied to determine which attribute lead to the optimal splitting like information gain criterion which is already found in weka.

By using j48 algorithm, the model has been implemented and figure 3 present the result.



J48 -C 0.25 -M 2

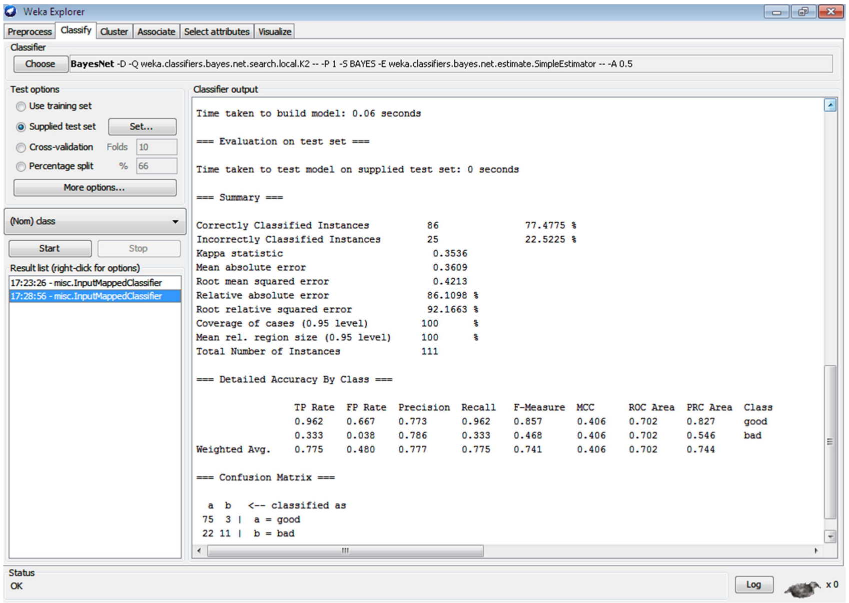
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          87               78.3784 %
Incorrectly Classified Instances        24               21.6216 %
Kappa statistic                          0.3451
Mean absolute error                      0.3448
Root mean squared error                  0.4039
Relative absolute error                 82.2728 %
Root relative squared error             88.3609 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      98.6486 %
Total Number of Instances              111

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.727    0.765      1.000    0.867      0.457    0.712     0.813     good
                 0.273    0.000    1.000      0.273    0.429      0.457    0.712     0.566     bad
Weighted Avg.    0.784    0.511    0.835      0.784    0.736      0.457    0.712     0.740

=== Confusion Matrix ===

  a  b   <-- classified as
 78  0 |  a = good
 24  9 |  b = bad
```

Figure 3   j48 results

4.2 BayseNet algorithm:

This algorithm depends on the theorem of Bayse. We build Bayesian Network after calculating condition probability to all nodes. It represent a directed acyclic graph. Diverse forms of algorithms are applied to approximation conditional probability for e.g. Genetic Algorithm and K2, Hill Climbing, Simulated Annealing, Tabu Search. The output of this algorithm can be seen as graph.

By using BayseNet algorithm, the model has been implemented and figure 3 present the result.



BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

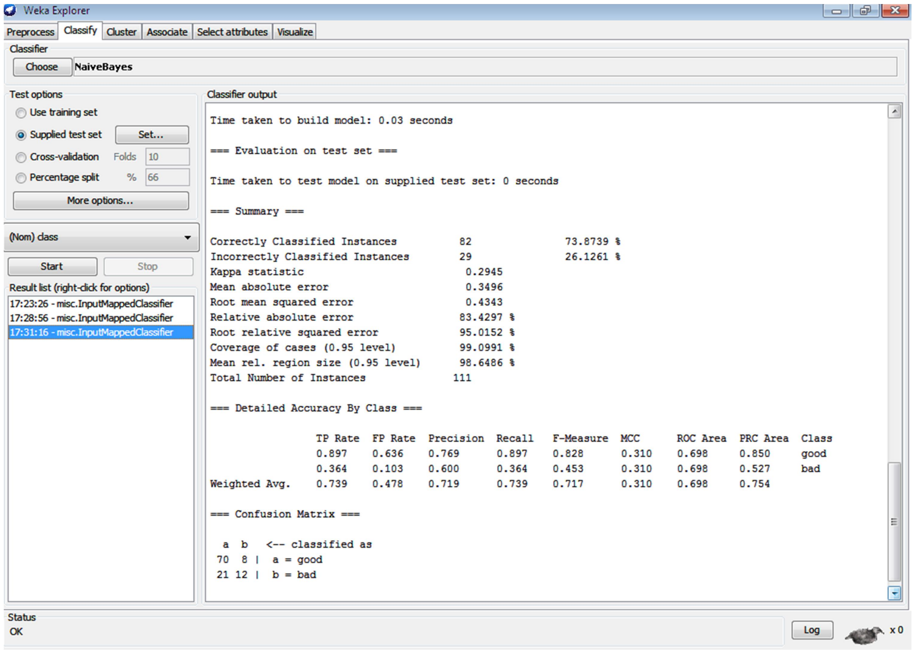
```
Time taken to build model: 0.06 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          86               77.4775 %
Incorrectly Classified Instances        25               22.5225 %
Kappa statistic                          0.3536
Mean absolute error                      0.3609
Root mean squared error                  0.4213
Relative absolute error                 86.1098 %
Root relative squared error             92.1663 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances              111

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.667    0.773      0.962    0.857      0.406    0.702     0.827     good
                 0.333    0.038    0.786      0.333    0.468      0.406    0.702     0.546     bad
Weighted Avg.    0.775    0.480    0.777      0.775    0.741      0.406    0.702     0.744

=== Confusion Matrix ===

  a  b   <-- classified as
 75  3 |  a = good
 22 11 |  b = bad
```

Figure 4 bayesNet results

4.3 NaiveBayes

The Bayesian is a supervised learning method.It characterized with it is elegance, simplicity, and robustness. For this reason it is became widely used in classification purposes. it supposes that attributes of a class are self-determining in real life.

By using BayseNet algorithm, the model has been implemented and figure 5 present the result.



Classifier: NaiveBayes

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

(Nom) class

Result list:
17:23:26 - misc.InputMappedClassifier
17:28:56 - misc.InputMappedClassifier
17:31:16 - misc.InputMappedClassifier

```
Time taken to build model: 0.03 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          82               73.8739 %
Incorrectly Classified Instances        29               26.1261 %
Kappa statistic                          0.2945
Mean absolute error                      0.3496
Root mean squared error                  0.4343
Relative absolute error                 83.4297 %
Root relative squared error             95.0152 %
Coverage of cases (0.95 level)          99.0991 %
Mean rel. region size (0.95 level)      98.6486 %
Total Number of Instances              111

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.897    0.636    0.769      0.897    0.828      0.310    0.698     0.850     good
                 0.364    0.103    0.600      0.364    0.453      0.310    0.698     0.527     bad
Weighted Avg.    0.739    0.478    0.719      0.739    0.717      0.310    0.698     0.754

=== Confusion Matrix ===

  a  b   <-- classified as
 70  8 |  a = good
 21 12 |  b = bad
```

Status
OK

Figure 5 naiveBayes results

The following table represents the accuracy measure for the above three techniques for test:

Technique	Correctly classified instance percent
j48	78.3784 %
bayesNet	77.4775 %
naiveBayes	73.8739 %

Table 2 the result from the three algorithm

After applying classification's data mining techniques algorithms which are j48, bayesNet and naiveBayes, we obtained the results from the three experiments in table 2 and after comparing the correctly classified instance percent we find that the best algorithm for loan classification is j48 algorithm. J48 algorithm is best because it has high accuracy and low mean absolute error as shown in the result in figure 3. Also it is capable to classify the instances correctly than the other techniques. Confusion matrix of the three algorithm in figures 3 ,4 and 5 showed that the j48 algorithm is the best one.   The experiments have been done several times and in each time the training and test sets size have been changed (80% training 20% test set,60% training 40% test and 70% training 30% test) and we obtain the same result which is J48 algorithm is best in classifying loans to good and bad loan. this model help bank manager to accept or reject loan applications by predicting that if the transaction will lead bank to risk or not and support decision maker to make a write decisions.

4. CONCLUSION

In this paper, three algorithms - j48, bayesNet and naiveBayes algorithms was used to build a predictive models that can be used to predict and classify the applications of loans that introduced by the customers to good or bad loan by investigate customer behaviors and previous pay back credit. The model has been implemented by using Weka application. After applying classification's data mining techniques algorithms which are j48, bayesNet and naiveBayes, we find that the best algorithm for loan classification is j48 algorithm. J48 algorithm is best because it has high accuracy and low mean absolute error as shown in the result in figure 5.

REFERENCES

[1] Ogawa, Ms Sumiko, et al. Financial Interconnectedness and Financial Sector Reforms in the Caribbean. No. 13-175. International Monetary Fund, 2013.

[2] Strahan, Philip E. "Borrower risk and the price and nonprice terms of bank loans." FRB of New York Staff Report 90 (1999).

[3] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[4] Pulakkazhy, Sreekumar, and R. V. S. Balan. "Data mining in banking and its applications-a review." Journal of computer science 9.10 (2013): 1252.

[5] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[6] Sharma, Poonam, and Gudla Balakrishna. "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Pattern." International Journal of Computer Science and Engineering Survey 2.4 (2011): 111.

[7] Chitra, K., and B. Subashini. "Data Mining Techniques and its Applications in Banking Sector." International Journal of Emerging Technology and Advanced Engineering 3.8 (2013): 219-226.

[8] Zurada, Jozef, and Martin Zurada. "How Secure Are “Good Loans”: Validating Loan-Granting Decisions And Predicting Default Rates On Consumer Loans."Review of Business Information Systems (RBIS) 6.3 (2011): 65-84.

[9] İkizler, Nazlı, and H. Altay Guvenir. "Mining interesting rules in bank loans data." Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks. 2001.

[10] Abhijit A. Sawant and P. M. Chawan, "Comparison of Data Mining Techniques used for Financial Data Analysis," International Journal of Emerging Technology and Advanced Engineering, june 2013.

[11] 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.

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

[13] Islam, Md Samsul, Lin Zhou, and Fei Li. "Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring." MSc, School of Management Blekinge Institute of Technology(2009).

[14] Vikas Jayasree and Rethnamoney Vijayalakshmi Siva Balan, A REVIEW ON DATA MINING IN BANKING SECTOR.: American Journal of Applied Sciences, 2013.

[15] Abhijit A. Sawant and P. M. Chawan, Comparison of Data Mining Techniques used for Financial Data Analysis.: International Journal of Emerging Technology and Advanced Engineering, june 2013.

[17] Moin, Kazi Imran, and Dr Qazi Baseer Ahmed. "Use of data mining in banking."International Journal of Engineering Research and Applications Vol 2 (2012): 738-742.

[18] Chu, Wesley, and Tsau Young Lin. Foundations and advances in data mining. Vol. 180. Springer Science & Business Media, 2005.