Effective Analysis of Financial Data using Knowledge Discovery Process

G. Arutjothi
Research Scholar
Department of Computer Science
Govt. Arts College(Autonomous)
Salem-7, TamilNadu, India

C. Senthamarai
Assistant Professor in Computer Applications
Department of Computer Science
Govt. Arts College(Autonomous)
Salem-7, TamilNadu, India

***Abstract*- Finance is the biggest factor of the Banking Industry. In Banking Industry success and failure is based on the credit. Banking Industries are competitive today with increase in volume, velocity and variety of new and existing data. Managing and analyzing the massive data is more difficult. One of the critical problem in financial organization is to properly evaluate the credit risk. Credit risk is the biggest challenge for the Banking Industry. Credit risk encompasses the borrower ability and willingness to pay and it is one of the main factor for defining a lenders credit policy. This research paper focuses on reducing the credit risk using the credit evalution model. This model uses data mining techniques such as decision tree, Support vector machine and logistic regression and it provides the information to make decision on loan proposals using weka tool.**

Keywords- Big data, Credit Evaluation, Decision Process, Decision Tree (C4.5), Logistic Regression, Support Vector Machine.

I. INTRODUCTION

Financial data analysis is used in many financial organizations for accurate analysis and assessment of the loan proposal**.** Credit risk is the biggest challenge for the Banking Industry. Credit evaluation is the process an individual customer must go through to become eligible for a loan. It also refers to the process, credit lenders undertake when evaluating a request for credit. Granting credit approval depends on borrowers credit risk. Credit risk which encompasses the borrowers ability and willingness to pay and it is one of the main factor for defining a lenders credit policy.

Creditors and lenders utilize a number of financial tools to evaluate the credit risk of a potential borrower. A lender analyze the borrowers balance sheet, cash flow statements and character before the credit is approved. Credit evaluation system was aimed to reduce the rate of credit risk on decision making to a minimum via analysis of existing all types of loan customers and estimate potential customer’s payment performances. The decision to grant credit to a certain customer must be evaluated for credit risk.

The credit decision process is a very difficult task on credit lenders. The credit managers provide a response approving or rejecting a credit request. A credit manager evaluate the risk associated with extending credit and declining an applicant based on numerous factors. The reliable information obtained by the credit manager is challenging task [1]. The manual decision making processes takes a long time. The proposed credit evaluation model is used to evaluate the credit risk in existing customers and provide a good decision on new applicant.

Many credit scoring models has been developed by bankers and researchers for the credit decision making. Initially personal credit scoring was evaluated subjectively according to personal experiences, and later it was based on 3Cs -Character, Capacity and Collateral [2].

A. Statement of the problem

Finance is the biggest factor of the Banking Industry. In Banking Industry success and failure is ased on the credit . If the credit amount could not be recoverd from the customer the bank will be in loss. To avoid this problem, the credit risk is checked before the credit is approved. The study of design and development of a system is used to increase the efficiency of the loan application evaluation process. The aim of the study is to design and develop an efficient system for analyzing the big data prevailing in Banking Industry. It also discovers the information for evaluating the loan.

B. Big data

Big data concerns huge-volume, complex, growing data sets with multiple data sources. Data comes from everywhere. Big data analytics is the process of examining big data to dicover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. The dimensions of big data is volume, variety and velocity.

Volume: Volume refers to the massive quantities of data that organizations are trying to improve decision making across the enterprise.

Variety: Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data.

Velocity: Data in motion. The speed at which data is created, processed and analyzed continues to accelerate.

Nowadays there are two more V's

Variability: The increase in the range of values typical of a large data set.

Value: which addresses the need for valuation of enterprise data.

Big data must be processed with advanced tools and provide the meaningful information. Big data accuracy may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

II. Literature Review

For obtaining the credit score, one has to undergo a process of credit evaluation before the credit score is sanctioned [1]. Financial industries have increasing volume of data. These massive data can be analyzed and managed using data mining techniques [3]. Statistical Classification methods have become important with dramatic growth in consumer credit in recent years. Statistical methods have been applied for the effective credit scoring system [4]. In commercial banks , neural network systems are efficient inorder to overcome the problems based on the competition in banking market and increasing number of consumer demands[5]. Even though there are anumber of credit evaluation systems developed the effective evaluation process is critical for all the banks.

The research work carried out earlier on data mining techniques have also been successfully applied to credit risk assessment problems. Bank data can be analyzed and provides specialized information which to improve the credit risk. Risk management is primarily concerned with reducing earning volatility and avoiding large losses [6]. Accurate credit scoring system has developed using hybrid data mining techniques such as clustering and classification (Decision Tree) [2]. This system uses a two stage process and provides the effective information. The importance of the credit risk assessment problem is, especially after the global economic crisis in 2008. So it is very important to have a proper way to deal with the credit risk and provide the accurate model for credit risk assessment [7]. Decision Tree algorithm C4.5 is worked on loan proposals. It is based on three parameters accuracy, precision and recall. The loan is approved based on high accuracy. High accuracy data efficiently works on loan decisions [8].

III. Proposed Model

The research work focus on reducing the rate of credit risk using the credit evaluation model. The proposed work is to identify the highly recommended classification techniques such as decision tree, support vector machine and logistic regression to evaluate the credit risk. Managing and analyzing the massive data is more difficult. Financial industries compraised on huge amount of data. In order to make decisions on loan proposals an efficient model is designed using data mining techniques. Tthese three techniques ara compared to find the best model.

A. The Proposed Architecture

Figure 1 deplitcs the proposed architecture, the whole credit data are taken from Banking Industry(big data) [9], it is analyzed to provide the effective information. This is really a complex or critical work on the banks. The proposed work is make decision whether the loan can be approved or rejected for the new potential customer.

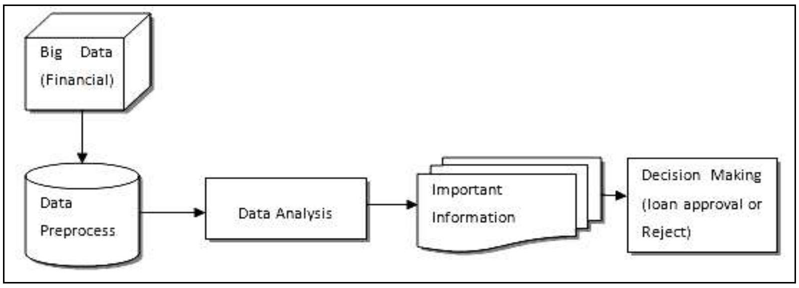


Figure 1. The proposed architecture for a credit evaluation system using data mining.

B. Classification

Classification is a process of generalizing the data according to different instances. Several classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, neural networks and regression models.

1. Decision tree

A decision tree is a tree structure that includes a root node, branches, and leaf nodes. Each internal node represents a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. It does not require any domain knowledge. It is easy to comprehend. The learning and classification steps of a decision tree are simple and fast.

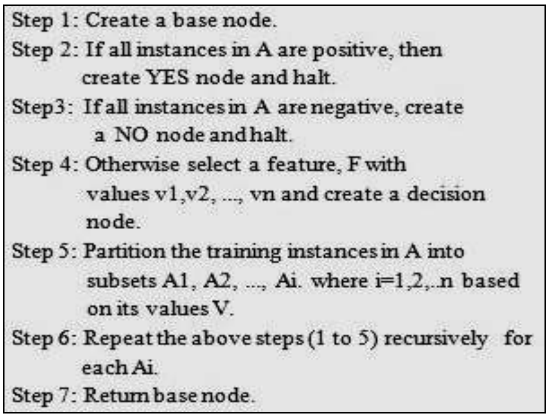
- Decision Tree Induction Algorithm

In Figure 2 A machine learning researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm, there is no backtracking. The trees are constructed in a top-down recursive divide-and-conquer manner. The proposed algorithm over a set of training instances A:

- Tree Pruning

Tree pruning is performed to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex. Here are the Tree Pruning Approaches listed below

  - Pre-pruning − The tree is pruned by halting its early construction.
  - Post-pruning - This approach eliminates a sub-tree from a fully grown tree.



```
Step 1: Create a base node.
Step 2: If all instances in A are positive, then
        create YES node and halt.
Step3:  If all instances in A are negative, create
        a NO node and halt.
Step 4: Otherwise select a feature, F with
        values v1,v2, ..., vn and create a decision
        node.
Step 5: Partition the training instances in A into
        subsets A1, A2, ..., Ai. where i=1,2,..n based
        on its values V.
Step 6: Repeat the above steps (1 to 5) recursively for
        each Ai.
Step 7: Return base node.
```

Figure 2. Decision Tree Induction Algorithm

2. Logistic Regression

The logistic regression model predicts the probability of an outcome that can only have two values. The prediction is based on the use of one or several predictors such as numerical and categorical. A logistic regression provides a logistic equation, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the equation is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability.

3. Support vector machine

A support vector machine (SVM) is a supervised learning method that analyze data and recognize patterns. It used for classification and regression analysis in statistics and computer science.

Support vector machine was first presented by Vladimir vapanik (1992). It uses nonlinear mapping to transform the original training instances into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by hyperplane. SVM has an effective method for classification in terms of the credit approval process.

IV. Result

A. Data analysis

The classification techniques uses categorical and discreate data types. Other data types can be changed into categorical and discreate data types. The following are the attributes can be analyzed using and specification used in the data analysis of debetors eligibility. Attributes can be classified as dependent and independent attributes.

Independent attributes

1) Account balance
2) Duration of credit
3) Payment status
4) Purpose
5) Credit amount
6) Value savings/ stocks
7) Length of current employement
8) Sex/marital status
9) Gurantor
10) Most valuable available asset
11) Age
12) Number of credit at this bank
13) Occupation
14) type of apartment

Dependent attribute

15) Good credit or bad credit

B. System Requirements

This Credit evaluation system requires the following components to work with massive data: computer system, customer, data and WEKA software. Sample data taken from UCI machine learning repository [9]. It consists of fifteen attributes and 1000 instances. Efficient Credit evaluation system used for building on three classifier models such as Decision tree, Logistic Regression, Support Vector Machine model.

C. Techniques used

A credit evaluation system uses a three techniques. Data preprocess methods are used to preprocess the data. The preprocessed data can be analyzed using the three analyzing techniques. This system uses ID3, Logistic Regression and Sequential minimal optimization techniques. Data is divided into four partitions, that is percentage split in 90% training data and 10% test data. Percentage split in 80% training and 20% test data. Percentage split in 70% training and 30% test data. Percentage split in 60% training and 40% test data. Each partition is tested using different classification techniques.

D. Result

Information gain method is used on this system. The table I describes the decision tree classification percentage.

Table I. Result for Decision tree

Data partition	Classification percentage	Accuracy	Precision	Recall
90% :10%	72.0	72.0	73.4	72.0
80%: 20%	74.5	74.5	74.7	74.5
70% :30%	68.6	68.6	67.1	68.7
60%: 40%	71.7	71.7	70.4	71.8

.

The table II describes the logistic regression classification percentage.

Table II. Result for Logistic Regression

Data partition	Classification percentage	Accuracy	Precision	Recall
90%:10%	70.0	70.0	70.5	70.0
80%: 20%	72.0	72.0	71.2	72.0
70% :30%	71.0	71.0	69.8	71.0
60%: 40%	72.0	72.0	70.7	72.0

The table III describes the support vector machine classification percentage

Table III. Result for Support vector Machine.

Data partition	Classification percentage	Accuracy	Precision	Recall
90%:10%	69.0	69.0	69.4	69.0
80%: 20%	72.5	72.5	72.2	72.5
70% :30%	70.3	70.3	69.1	70.3
60%: 40%	71.0	71.0	69.4	71.0

E. Result analysis

Figure 3 shows the classification accuracy. The highest accuracy is 80%:20% with decision tree technique.

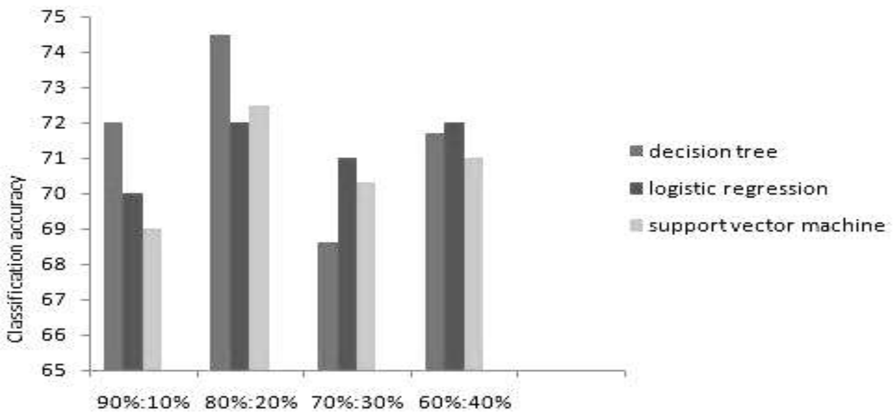


Figure 3. Classification accuracy analyzed.

Three model results are analyzed. Figure 4 shows the classification Receiver Operating Curve with 80%: 20% and figure 5 shows the classification cost benefit curve with 80%: 20%. It is found that the biggest precision, recall and accuracy value is reached with data partition of 80% : 20% in decision tree technique. The value is 74.5% is accuracy and 74.7% is precision and 74.5% is recall.

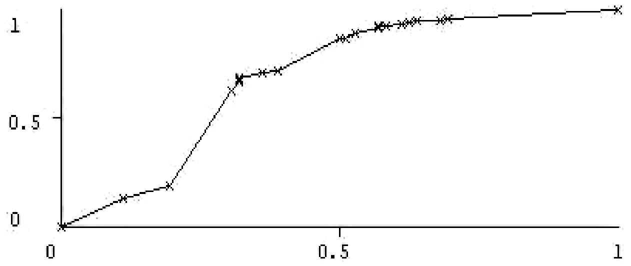


Figure 4. Classification ROC curve with 80% :20%

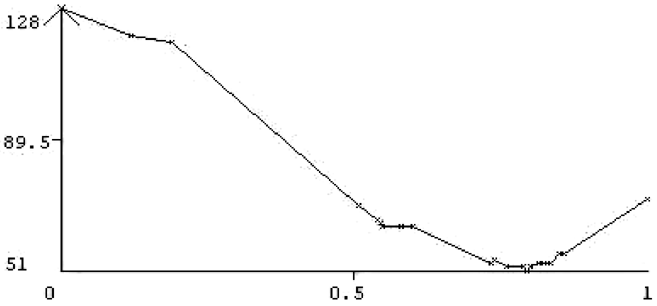


Figure 5. Classification cost benefit curve with 80% :20%.

V. CONCLUSION

Credit evaluation process is a biggest challenge for all banks. This study is to compare the decision tree, logistic regression and support vector machine technique and used to find the best credit evaluation technique. We have used three metrics to evaluate these techniques. Decision tree technique is best for credit evaluation system. The result of this system is effective on credit evaluation process.

REFERENCES

[1] Seema Purohit, Anjali Kulkarani, Credit Evaluation Model of Loan Proposals for Indian Banks, Proceeding of IEEE World Congress on Information and Communication Technology, 2011.

[2] Weimin Chen, Guocheng Xiang, Youjin Liu, Kexi wang, Credit Risk Evaluation by Hybrid Data Mining Technique, Systems Engineering Procedia, Vol.3, 2012.

[3] Bharti Thakur, Manish Mann, Data Mining for Big Data: A Review, International Journal of Advanced Research in Computer Science and Engineering, Vol.4, No.5, 2014.

[4] D.J.Hand and W.E. Henley, Statistical Classification Methods in Consumer Credit Scoring: a Review, J.R. Statistics Society, pp.523-541, 1997.

[5] Shorouq Fathi Eletter, Saad Ghaleb Yaseen, Neuro Based Artificial Intelligence Model for Loan Decision, American Journal of Economics and Business Administration, 2010.

[6] Tony Van, Credit risk management: Basic Concepts, Rating analysis, models, economic and regulatory capital, Oxford University press, 2008.

[7] Adnan, Dzenana, Data mining Techniques for Credit Risk Assessment Task, Recent Advances in Computer Science and applications, 2013.

[8] Rafik Khairul Amin, Indwiarti, Yuliant Sibaroni, Implementation of Decision Tree Using C4.5 Algorithm in Decision Making of Loan Application by the debtor (Case Study: Bank parser of Yogyakarta Special Region), 3rd International Conference on information and communication Technology, 2015.

[9] http://mlr.cs.umass.edu/ml/datasets.html