

# Decision Support for Banking Industry through Data Mining Techniques

1. Pankaj Pathak \* 2. Hemant Pal

\*Research Scholar, Pacific University Udaipur (Raj), Asst. Professor, SVIM, Indore

## ABSTRACT

In present scenarios every financial industry is facing a critical competition. Recently, with the financial crisis becoming serious, the trend of financial globalization and financial market volatility has attracted people's attention, especially banks and investors who suffered from the unprecedented challenges of credit risks. So each industry wants to increase their valuable customers by knowing the credit capability of the customers which are based upon number of factors. To derive this knowledge about the customers with the help of existing data sets and information is possible by Data Mining Techniques. We can make important strategic decisions on the basis of results of these data mining techniques. One of the important data mining techniques is decision tree a hierarchical structure which contains nodes and directed edges. The Predictive models can build with the help of Decision Trees.

## Keywords

Decision, Data Mining, Decision Tree, credit risk, Algorithm

## 1. INTRODUCTION

Data mining techniques are used to discover hidden knowledge, unknown patterns and new rules from large data sets, which maybe useful for a variety of decision making activity. With the increasing economic globalization and improvements in information technology, large amounts of financial data are being generated and stored. These can be subjected to data mining techniques to discover hidden patterns and obtain predictions for trends in the future and the behavior of the financial markets. This in turn would result in an improved market place responsiveness and awareness leading to reduced costs and increased Revenue.

Data mining is different from conventional statistical analysis. Data mining requires building a BI decision-support application, specifically a data mining

application, using a data-mining tool. The data mining application can then use a sophisticated blend of classical and advanced components like artificial intelligence, pattern recognition, databases, traditional statistics, and graphics to present hidden relationships and patterns found in the organization's data pool (Koyuncugil, 2004).

Recently, with the financial crisis becoming serious, the trend of financial globalization and financial market volatility has attracted people's attention, especially banks and investors who suffered from the unprecedented challenges of credit risk [1]. The credit crisis caused by American showed that international banking has been great challenged because of their lack of effective methods for assessment in controlling credit risk. Statistical model and artificial intelligence model are widely used in the study of credit risk [1].

## 2. PAGE THE NEED FOR PPLYING DATA MINING TO BANKING

As banking competition becomes more and more global and intense, banks have to fight more creatively and proactively to gain or even maintain market shares. Banks which still rely on reactive customer service techniques and conventional mass marketing are far below from those which not do so. The banks of the future will use one asset, knowledge and not financial resources, as their leverage for survival and excellence [2]. Surprisingly, most of this knowledge are currently in the banking system and generated by daily transactions and operations. This valuable information need not be gathered by intrusive customer surveys or expensive market research programs. The only problem is that this storehouse of data has to be mined for useful information.

Normally unmined and unappreciated, these terabytes of transaction data are collected, generated, printed, stored, only to be filed and discarded after they have served their short-lived purposes as audit trails and paper trails. Most data generated by the bank's information systems, manual or automated

like ATM's and credit card processing, were designed to support or track transactions, satisfy internal and external audit requirements, and meet government or central bank regulations. Few are gathered intentionally and originally to generate useful management reports.

Current information systems are not designed as decision support systems (DSS) that would help management make effective decisions to manage resources, compete successfully, and enhance customer satisfaction and service. One of the fundamental tasks in credit risk management is to assign a credit grade to a borrower. The results are management reports that are late, inaccurate, and incomplete. Executive decisions [5] based on these misleading reports can lead to millions of dollars in short and long term losses and lost opportunities and markets.

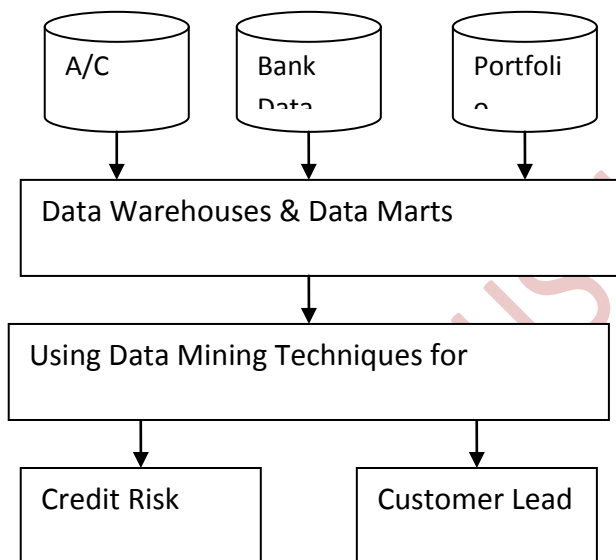


Fig1: Use of Data Mining For Business Activity

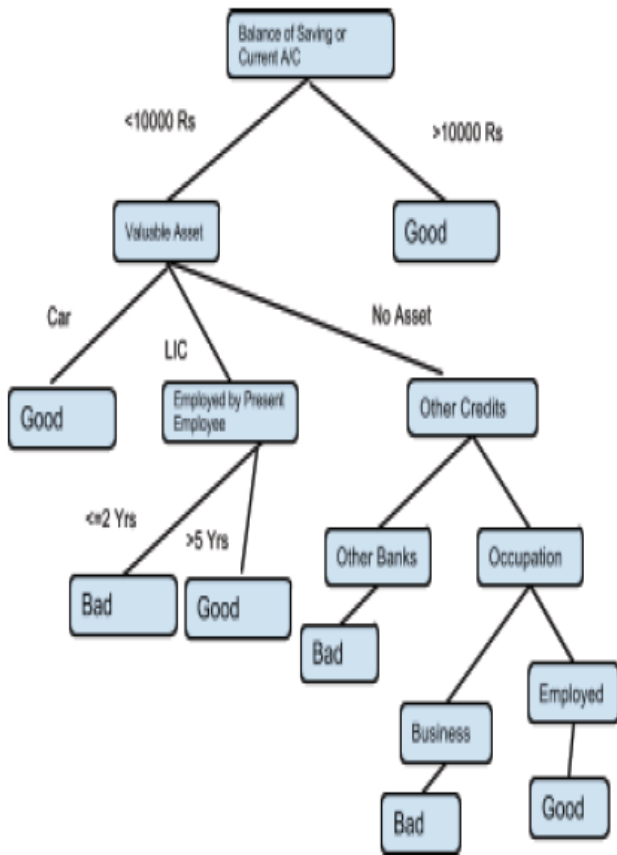
### 3. METHODOLOGY

#### 3.1 Classification Methods

In this approach, risk levels are organized into two categories based on past default history. For example, customers with past default history can be classified into "risky" group, whereas the rest are placed as "safe" group. [4] Using this categorization information as target of prediction, Decision Tree and Rule Induction techniques can be used to build models that can predict default risk levels of new loan applications.

Decision trees [3] are built up in a simple five step process by increasing information contained in the reduced data set following each split. Data by its nature contains uncertainties. We may be able to systematically reduce uncertainties and thus increase information by activities like sorting or classifying. When we have sorted or classified to achieve the greatest reduction in uncertainty, we have basically achieved greatest increase in information.

1. The root node - Balance of Current Account - manages to classify nearly 94% of the data set. Specifically, if someone has a Balance of Current Account  $\geq 10000$  Rs, then the chances of them having a "good" score. However, the tree is unable to clearly pick out good or bad scores, if there is "no running account" (Less chance). A similar conclusion results if someone has "no balance".
2. If the Balance of Current Account is less than 10000 Rs. then the other parameters come into effect and play an increasingly important role in deciding if someone is likely to have "good" or "bad" credit.
3. However, the fact that there are numerous terminals leaves with frequencies of occurrence as low as 2 (for example, "Other credit"), it implies that the tree suffers from "over fitting". One way we could have avoided this situation is by changing the Decision Tree criterion "Minimal leaf size" to something like 10 (instead of default, 2). But doing so, we would also lose the classification influence of all the other parameters, except the root node.



**Fig2: Decision tree for assessing credit risk**

**4. IMPLEMENTATION METHODS OF DECISION TREE FOR CREDIT RISK ANALYSIS**

**4.1 ID3 Algorithm:** ID3 algorithm [6] assumes that a good decision tree is the simplest decision tree. Preferring simplicity and avoiding unnecessary assumptions which is Known as Occam’s Razor. It believes that we should always accept the simplest answer that correctly fits our data. ID3 selects a property to test at the current node of the tree and uses this test to partition the set of examples. The algorithm then recursively constructs a sub tree for each partition. This continuous until all members of the partition are in the same class. That class becomes a leaf node of the tree.

Steps of ID3 Decision Tree Induction:

1. A ← the “best” decision attribute for next node
2. Assign A as decision attribute (=property) for node.
3. For each value of A create new descendant.

4. Sort training examples to leaf node according to the attribute value of the branch.

5. If all training examples are perfectly classified (same value of target attribute) stop, else iterate over new leaf nodes.

**4.2 C4.5 Algorithm:** C4.5 Decision Tree Algorithm was proposed by Quinlan by extending and improving the ID3 algorithm. Besides the functions provided by ID3, C4.5 [8, 9] is also able to deal with continuous attributes and default attributes. In addition, unbalanced trees are avoided via pruning technology while cross certification is enabled. C4.5’s simplicity, efficiency and reliability have made C4.5 become the most important algorithm in machine learning and classification. However, C4.5 is not perfect. The divide-and-conquer approach [10] makes it achieve not global optimization but local optimization through only local search strategy. Moreover, it is difficult to restructure or make further improvement on a constructed tree because C4.5 evaluates a decision tree while building it. C4.5 consists of three groups of algorithm: C4.5, C4.5-no-pruning and C4.5-rules. In this paper, we will focus on the basic C4.5 algorithm.

Steps of C4.5 Decision Tree Induction:

1. Check if algorithm satisfies termination criteria.
2. Computer information-theoretic criteria for all attributes.
3. Choose best attribute according to the information-theoretic criteria.
4. Create a decision node based on the best attribute in step
5. Induce (i.e. split) the dataset based on newly created decision node in step 4.
6. For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call).
7. Attach the tree obtained in step 6 to the decision node in step 4
8. Return tree.

**4.3 CART Algorithm:** Classification and Regression Trees (CART) methodology was developed in 80s by Breiman, Freidman, Olshen, Stone in their paper “Classification and Regression Trees” (1984). For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations. For example, learning sample for credit scoring system would be fundamental information about previous borrows (variables) matched with actual payoff results (classes).

Decision trees [9] are represented by a set of questions which splits the learning sample into

smaller and smaller parts. CART asks only yes/no questions. A possible question could be: "Is age greater than 50?" or "Is sex male?" CART algorithm will search for all possible variables and all possible values in order to find the best split – the question that splits the data into two parts with maximum homogeneity. The process is then repeated for each of the resulting data fragments.

CART can easily handle both numerical and categorical variables. Among other advantages of CART method is its robustness to outliers. Usually the splitting algorithm will isolate outliers in individual node or nodes.

Steps of CART Decision Tree Induction

1. Let  $A$  be a feature with domain  $A$ . Ensure a finite number of binary splitting

For  $X$  by applying the following domain partitioning rules:

– If  $A$  is nominal, choose  $A_0 \dots A_n$  such that  $0 < j < n$

– If  $A$  is ordinal, choose  $a \in A$  such that  $x_{\min} < a < x_{\max}$ , where  $x_{\min}$ ,  $x_{\max}$  are the minimum and maximum values of feature  $A$  in  $D$ .

– If  $A$  is numeric, choose  $a \in A$  such that  $a = (x_k + x_{k+1})/2$ , where  $x_k, x_{k+1}$  are consecutive elements in the ordered value list of feature  $A$  in  $D$ .

2. For node  $t$  of a decision tree generate all splitting of the above type.

3. Choose a splitting from the set of splitting that maximizes the impurity reduction.

**4.4 CHAID Algorithm:** CHAID [7] is a type of decision tree technique, based on adjusted significance testing (Bonferroni testing). The technique was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic. CHAID can be used for prediction (in a similar fashion to regression analysis, this version of CHAID being originally known as XAID) as well as classification, and for detection of interaction between variables. CHAID stands for Chi-squared Automatic Interaction Detection, based upon a formal extension of the SAID (Automatic Interaction Detection) and THAID (Theta Automatic Interaction Detection) procedures of the 1960s and 70s, which in turn were extensions of earlier research, including that performed in the UK in the 1950s.

The best input attribute to be used for splitting the current node is then selected, such that each child node is made of a group of homogeneous values of the selected attribute. Note that no split is performed if the adjusted  $p$  value of the best input attribute is not less than a certain split threshold. This procedure also stops when one of the following conditions is fulfilled:

1. Maximum tree depth is reached.
2. Minimum number of cases in node for being a parent is reached, so it can not be split any further.
3. Minimum number of cases in node for being a child node is reached.

CHAID handles missing values by treating them all as a single valid category [8].

CHAID does not perform pruning. Like other decision trees, CHAID's advantages are that its output is highly visual and easy to interpret. Because it uses multiway splits by default, it needs rather large sample sizes to work effectively, since with small sample sizes the respondent groups can quickly become too small for reliable analysis.

## 5. CONCLUSION

In this paper, we discuss about Data Mining techniques, and the problem of determining the credit rating of customers by the banks and other financial companies. This problem can be sort out with the help of technology. As the banks having massive information about their customers and this information can be used to determine the credit risk of particular customer. We can also develop a decision support system which can help managers to take credit related decisions. Further we have described several decision tree models which can be implemented for designing the above mentioned decision support system.

## 5. REFERENCES

- [1] Yi Jiang et al "A New Approach based on a Rough Set and a Decision Tree to Bank Customer Credit Evaluation, Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education.
- [2] Seema Purohit, Anjali Kulkarni "Credit Evaluation Model of Loan Proposals for Indian Banks", 2011 World Congress on Information and Communication Technologies
- [3] Introduction to Data Mining by Pang-Ning Tan, Tan
- [4] Vivek Bhambri: IMPLEMENTATION OF DATA MINING IN BANKING SECTOR- A FEASIBILITY STUDY, IJRIM Volume 2, Issue 9 (September 2012) (ISSN 2231-4334).
- [5] Rajanish Das: DATA MINING IN BANKING AND FINANCE: A NOTE FOR BANKERS, IIM Ahmadabad
- [6] Data Mining Practical Machine Learning Tools and Techniques by Ian H. Witten, Eibe Frank, Mark A. Hall
- [7] <http://tigger.uic.edu/~georgek/HomePage/Nonparametrics/timofeev.pdf>
- [8] [www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf](http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf)

[9] Data Mining with Decision trees: Theory and applications by Lior Rokach and Oded Maimon .

[10] Xingdong Wu et. Al: Top 10 algorithms in data mining, Springer-Verlag London Limited 2007.

IJSHRE