

Data Mining Techniques and its Applications in Banking Sector

Dr. K. Chitra¹, B. Subashini²

¹Assistant Professor, Department of Computer Science, Government Arts College, Melur, Madurai.

²Assistant Professor, Department of Computer Science, V.V. Vanniaperumal College for Women, Virudhunagar.

Abstract— Data mining is becoming strategically important area for many business organizations including banking sector. It is a process of analyzing the data from various perspectives and summarizing it into valuable information. Data mining assists the banks to look for hidden pattern in a group and discover unknown relationship in the data. Today, customers have so many opinions with regard to where they can choose to do their business. Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations for the banking sector to avoid customer attrition. Customer retention is the most important factor to be analyzed in today's competitive business environment. And also fraud is a significant problem in banking sector. Detecting and preventing fraud is difficult, because fraudsters develop new schemes all the time, and the schemes grow more and more sophisticated to elude easy detection. This paper analyzes the data mining techniques and its applications in banking sector like fraud prevention and detection, customer retention, marketing and risk management.

Keywords— Banking Sector, Customer Retention, Credit Approval, Data mining, Fraud Detection,

I. INTRODUCTION

Technological innovations have enabled the banking industry to open up efficient delivery channels.

IT has helped the banking industry to deal with the challenges the new economy poses. Nowadays, Banks have realized that customer relationships are a very important factor for their success. Customer relationship management (CRM) is a strategy that can help them to build long-lasting relationships with their customers and increase their revenues and profits. CRM in the banking sector is of greater importance. The CRM focus is shifting from customer acquisition to customer retention and ensuring the appropriate amounts of time, money and managerial resources are directed at both of these key tasks. The challenge the bank face is how to retain the most profitable customers and how to do that at the lowest cost. At the same time, they need to find and implement this solution quickly and the solution to be flexible. Traditional methods of data analysis have long been used to detect fraud. They require complex and time-consuming investigations that deal with different domains of knowledge like financial, economics, business practices and law. Fraud instances can be similar in content and appearance but usually are not identical. In developing countries like India, Bankers face more problems with the fraudsters. Using data mining technique, it is simple to build a successful predictive model and visualize the report into meaningful information to the user. The following figure 1 illustrates the flow of data mining technique in our system model.

Problem Understanding	<ul style="list-style-type: none"> • Knowledge about the bankink sector • Its challenges and the bank's most important concerns
Data Understanding	<ul style="list-style-type: none"> • Data gathering about the attributes needed • The customer data
Data Filtering	<ul style="list-style-type: none"> • Data filtering, preparation of the attributes • Selecting the data set, checking missed data, making it ready to be mined
System Modelling	<ul style="list-style-type: none"> • Modelling system for customer retention, fraud prevention and detection • Finding their most applicable attribute
System Evaluation	<ul style="list-style-type: none"> • Checking data of new customers with the designed model and • Verification of the results
Analyzing Results	<ul style="list-style-type: none"> • Comparing with the other models and • Analyze the performance

Data mining tools, using large databases, can facilitate

1. Automatic prediction of future trends and behaviors and
2. Automated discovery of previously unknown patterns

II. DATA MINING TECHNIQUES AND ALGORITHMS

Data mining algorithms specify a variety of problems that can be modeled and solved. Data mining functions fall generally into two categories:

1. Supervised Learning
2. Unsupervised Learning

Concepts of supervised and unsupervised learning are derived from the science of machine learning, which has been called a sub-area of artificial intelligence. Artificial intelligence means the implementation and study of systems that exhibit autonomous intelligence or behavior of their own. Machine learning deals with techniques that enable devices to learn from their own performance and modify their own functioning. Data mining applies machine learning concepts to data.

A. Supervised Learning

Supervised learning is also known as directed learning. The learning process is directed by a previously known dependent attribute or target. Directed data mining attempts to explain the behavior of the target as a function of a set of independent attributes or predictors. Supervised learning generally results in predictive models. This is in contrast to unsupervised learning where the goal is pattern detection.

The building of a supervised model involves training, a process whereby the software analyzes many cases where the target value is already known. In the training process, the model "learns" the logic for making the prediction. For example, a model that seeks to identify the customers who are likely to respond to a promotion must be trained by analyzing the characteristics of many customers who are known to have responded or not responded to a promotion in the past.

1) *Supervised Data Mining Algorithms*: Table I describes the data mining algorithms for supervised functions.

Table I
Data Mining Algorithms for Supervised Functions

Algorithm	Function	Description
Decision Tree	Classification	Decision trees extract predictive information in the form of human-understandable rules. The rules are if-then-else expressions; they explain the decisions that lead to the prediction.
Generalized Linear Models (GLM)	Classification and Regression	GLM implements logistic regression for classification of binary targets and linear regression for continuous targets. GLM classification supports confidence bounds for prediction probabilities. GLM regression supports confidence bounds for predictions.
Minimum Description Length (MDL)	Attribute Importance	MDL is an information theoretic model selection principle. MDL assumes that the simplest, most compact representation of data is the best and most probable explanation of the data.
Naive Bayes (NB)	Classification	Naive Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence, as observed in the data.
Support Vector Machine (SVM)	Classification and Regression	Distinct versions of SVM use different kernel functions to handle different types of data sets. Linear and Gaussian (nonlinear) kernels are supported. SVM classification attempts to separate the target classes with the widest possible margin. SVM regression tries to find a continuous function such that the maximum number of data points lie within an epsilon-wide tube around it.

B. Unsupervised Learning

Unsupervised learning is non-directed. There is no distinction between dependent and independent attributes. There is no previously-known result to guide the algorithm in building the model.

Unsupervised learning can be used for descriptive purposes. It can also be used to make predictions.

1) *Unsupervised Data Mining Algorithms*: Table II describes the unsupervised data mining algorithms.

Table II
Data Mining Algorithms for Unsupervised Functions

Algorithm	Function	Description
Apriori	Association	Apriori performs market basket analysis by discovering co-occurring items (frequent itemsets) within a set. Apriori finds rules with support greater than a specified minimum support and confidence greater than a specified minimum confidence. For example Find the items that tend to be purchased together and specify their relationship
k-Means	Clustering	k-Means is a distance-based clustering algorithm that partitions the data into a predetermined number of clusters. Each cluster has a centroid (center of gravity). Cases (individuals within the population) that are in a cluster are close to the centroid. For example, segment demographic data into clusters and rank the probability that an individual will belong to a given cluster
Non-Negative Matrix Factorization (NMF)	Feature Extraction	NMF generates new attributes using linear combinations of the original attributes. The coefficients of the linear combinations are non-negative. During model apply, an NMF model maps the original data into the new set of attributes discovered by the model. For example, given demographic data about a set of customers, group the attributes into general characteristics of the customers
One Class Support Vector Machine	Anomaly Detection	One-class SVM builds a profile of one class and when applied, flags cases that are somehow different from that profile. This allows for the detection of rare cases that are not necessarily related to each other. For example, given demographic data about a set of customers, identify customer transaction behavior that is significantly different from the normal.

III. TOP 10 FRAUDS IN INDIAN BANKING SECTOR

The Reserve Bank of India – RBI maintains data on frauds on the basis of area of operation under which the frauds have been perpetrated. According to such data pertaining, top 10 categories under which frauds have been reported by banks are as follows

1. Credit Cards
2. Deposits – Savings A/C
3. Internet Banking
4. Housing Loans
5. Term Loans
6. Cheque / Demand Drafts
7. Cash Transactions
8. Cash Credit A/c(Types of Overdraft A/C]
9. Advances
10. ATM / Debit Cards

IV. DATA MINING APPLICATIONS IN BANKING SECTOR

Figure 2 depicts the data mining techniques and algorithm that are applicable to the banking sector. Customer retention plays vital role in the banking sector. The supervised learning method Decision Tree implemented using CART algorithm is used for customer retention. Preventing fraud is better than detecting the fraudulent transaction after its occurrence. Hence for credit card approval process the data mining techniques Decision Tree, Support Vector Machine (SVM) and Logistic Regression are used. Clustering model implemented using EM algorithm can be used to detect fraud in banking sector.

A. Customer Retention in Banking Sector

Today, customers have so many opinions with regard to where they can choose to do their business. Executives in the banking industry, therefore, must be aware that if they are not giving each customer their full attention, the customer can simply find another bank that will. Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics.

These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts. This is a natural source of ideas, since the machine learning task can be described as turning background knowledge and examples (input) into knowledge (output).

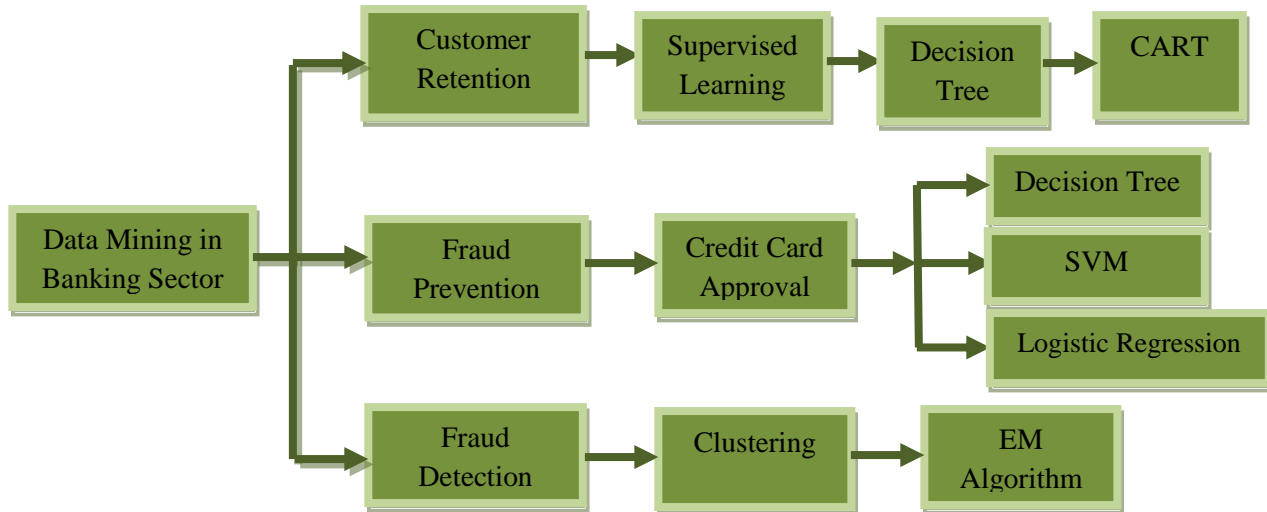
Data mining can help in targeting ‘new’ customers for products and services and in discovering a customer’s previous purchasing patterns so that the bank will be able to retain existing customers by offering incentives that are individually tailored to each customer’s needs. Churn in the banking sector is a major problem today. Losing the customers can be very expensive as it costs to acquire a new customer. Predictive data mining techniques are useful to convert the meaningful data into knowledge. In this section, we discuss the predictive data mining techniques for the churn problem in banking sector.

To improve customer retention, three steps are needed: 1) measurement of customer retention; 2) identification of root causes of defection and related key service issues; and the 3) development of corrective action to improve retention. Measurement of existing customer retention rates is the first significant step in the task of improving loyalty. This involves measuring retention rates and profitability analysis by segment.

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. 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 Tree

Decision trees are the most popular predictive models (Burez and Van den Poel, 2007). A decision tree is a tree-like graph representing the relationships between a set of variables.



Decision tree models are used to solve classification and prediction problems where instances are classified into one of two classes, typically positive and negative, or churner and non-churner in the churn classification case. These models are represented and evaluated in a top-down manner. Developing decision trees involves two phases: Tree building and tree pruning.

Tree building starts from the root node that represents a feature of the cases that need to be classified. Feature selection is based on evaluation of the information gain ratio of every feature. Following the same process of information gain evaluation, the lower level nodes are constructed by mimicking the divide and conquer strategy. Building a decision tree incorporates three key elements:

- 1- Identifying roles at the node for splitting data according to its value on one variable or feature.
- 2- Identifying a stopping rule for deciding when a sub-tree is created.
- 3- Identifying a class outcome for each terminal leaf node, for example, ‘_Churn’ or ‘_Non-churn’.

Decision trees usually become very large if not pruned to find the best tree. The pruning process is utilized not only to produce a smaller tree but also to guarantee a better generalisation. This process involves identifying and removing the branches that contain the largest estimated error rate and can be regarded as an experimentation process. The purpose of this process is to improve predictive accuracy and to reduce the decision tree complexity (Au, Chan and Yao, 2003).

Once the model is built, the decision about a given case regarding to which of the two classes it belongs is established by moving from the root node down to all the leaves and interior nodes. The movement path is determined by the similarity calculation until a leaf node is reached, at which point a classification decision is made.

2) *Value Prediction Methods*: In this method, for example, instead of classifying new loan applications, it attempts to predict expected default amounts for new loan applications. The predicted values are numeric and thus it requires modeling techniques that can take numerical data as target (or predicted) variables. Neural Network and regression are used for this purpose. The most common data mining methods used for customer profiling are:

- Clustering (descriptive)
- Classification (predictive) and regression (predictive)
- Association rule discovery (descriptive) and sequential pattern discovery (predictive)

B. Automatic Credit Approval using Classification Method

Fraud is a significant problem in banking sector. Detecting and preventing fraud is difficult, because fraudsters develop new schemes all the time, and the schemes grow more and more sophisticated to elude easy detection. Bank Fraud is a federal crime in many countries, defined as planning to obtain property or money from any federally insured financial institution. It is sometimes considered a white collar crime.

All the major operational areas in banking represent a good opportunity for fraudsters with growing incidence being reported under deposit, loan and inter-branch accounting transactions, including remittances. In developing countries like India, Bankers face more problems with the fraudsters. There is lack of technique to detect the banking fraud.

Automatic credit approval is the most significant process in the banking sector and financial institutions. Fraud can be prevented by making a good decision for the credit approval using the classification models based on decision trees (C5.0 & CART), Support Vector Machine (SVM) and Logistic Regression Techniques. It prevents the fraud which is going to happen.

1) *Classification Methods*: Classification is perhaps the most familiar and most popular data mining technique. Estimation and prediction may be viewed as types of classification. There are more classification methods such as statistical based, distance based, decision tree based, neural network based, rule based[4].

C5.0

C5.0 builds decision trees from a set of training data in the same way as ID3, using the concept of Information entropy. The training data is a set $S=S_1, S_2, \dots$ of already classified samples. Each sample S_i consists of a p -dimensional vector $(x_{1,i}, x_{2,i}, \dots, x_{p,i})$, where the x_j represent attributes or features of the sample, as well as the class in which s_i falls. At each node of the tree, C5.0 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C5.0 algorithm then recurses on the smaller sublists. Gain is computed to estimate the gain produced by a split over an attribute. The gain of information is used to create small decision trees that can identify the answers with a few questions[5].

CART

A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.

Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

Support Vector Machine (SVM)

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables. For degree- d polynomials, the polynomial kernel is defined as

$$K(x,y) = (x^T y + c)^d$$

Where x and y are vectors in the input space, i.e. vectors of features computed from training or test samples, $c > 0$ is a constant trading off the influence of higher-order versus lower-order terms in the polynomial.

Logistic Regression

Logistic regression or logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. Instead of fitting the data to a straight line, logistic regression uses a logistic curve. The formula for a univariate logistic curve is

$$p = \frac{e^{c_0 + c_1 x_1}}{1 + e^{c_0 + c_1 x_1}}$$

To perform the logarithmic function can be applied to obtain the logistic function

$$\log_e = \frac{p}{1-p} = c_0 + c_1 x_{10}$$

Logistic regression is simple, easy to implement, and provide good performance on a wide variety of problems [6].

C. Fraud Detection in Banking Sector

Sometimes the given demographics and transaction history of the customers are likely to defraud the bank. Data mining technique helps to analyze such patterns and transactions that lead to fraud. Banking sector gives more effort for Fraud Detection. Fraud management is a knowledge-intensive activity. It is so important in fraud detection is that finding which ones of the transactions are not ones that the user would be doing.

So the mining system checks which ones of the transactions do not fit into a specific category or are not standard repeat transactions.

It is required to decide which of the user's actions correspond to his natural behavior and which are exceptional, without any assistance. With the help of unique algorithms, it is able to detect suspicious activity within the data in a non prescriptive way. While the system observes the user's transactions, it discovers common behavior patterns by grouping similar transactions together.

In order to discover anomalous transactions, new transactions are compared with the user's common behavior patterns. A transaction that does not correspond with one of these patterns will be treated as a suspicious activity and trigger precautionary steps accordingly.

An important early step in fraud detection is to identify factors that can lead to fraud. What specific phenomena typically occur before, during, or after a fraudulent incident? What other characteristics are generally seen with fraud? When these phenomena and characteristics are pinpointed, predicting and detecting fraud becomes a much more manageable task. Because of the nature of the data, traditional machine-learning techniques are not suitable. Traditional techniques may detect fraudulent actions similar to ones already recognized as fraud, but they will rarely detect fraudulent activities that were not learned beforehand. The clustering model and the probability density estimation methods, described in the following sections, can be well utilized for detecting fraud in the banking sector.

1) The Clustering model: Clustering helps in grouping the data into similar clusters that helps in uncomplicated retrieval of data. Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible. This model is based on the use of the parameters' data clusterization regions.

In order to determine these regions of clusterization first its need to find the maximum difference ($DIFF_{max}$) between values of an attribute in the training data. This difference ($DIFF_{max}$) is split into $N_{interval}$ segments. $N_{interval}$ is the binary logarithm of the attribute values account N_{points} . In general, $N_{interval}$ can be found using another way of looking. Such calculation of $N_{interval}$ is based on the assumption that a twofold increase of N_{points} will be equal to $N_{interval}$ plus one.

For each found segment the calculation of the average value and the corresponding deviation for hit attribute values is made.

Thus $N_{interval}$ centers and corresponding deviations that describe all values of the certain attribute from the training data appears.

The final result of classification of the whole transaction is the linear combination of classification results for each parameter:

$$\text{Result} = w_1 \times \text{Class}_1 + w_2 \times \text{Class}_2 + \dots + w_n \times \text{Class}_n$$

2) Probability density estimation method: To model the probability density function, Gaussian mixture model is used, which is a sum of weighted component densities of Gaussian form.

$$p(x) = \sum_{j=1}^M p(x | j) P(j)$$

The $p(x | j)$ is the j^{th} component density of Gaussian form and the $P(j)$ is its mixing proportion. The parameters of the Gaussian mixture model can be estimated using the EM algorithm (Computes maximum-likelihood estimates of parameters). This method specialize the general model by re-estimating the mixing proportions for each user dynamically after each sampling period as new data becomes available. Whereas the means and the variances of the user specific models are common, only the mixing proportions are different between the users' models. In order to estimate the density of past behavior, it is necessary to retrieve the data from the last k days and adapt the mixing proportions to maximize the likelihood of past behavior. But this approach requires too much interaction with the billing system to be used in practice. To avoid this burdensome processing of data, this method formulates the partial estimation procedure using on-line estimation. The on-line version of the EM algorithm was first introduced by Nowlan.

$$P(j)^{new} = \alpha P(j)^{old} + P(j | x)$$

Remembering that the new maximum likelihood estimate for $P(j)$ is computed as the expected value of $P(j | x)$ over the whole data set with the current parameter fit, this model can easily formulate a recursive estimator for this expected value as can be seen in Equation 3. The decay term α determines the efficient length of the exponentially decaying window in the past. The approach performs statistical modeling of past behavior and produces a novelty measure of current usage as a negative log likelihood of current usage. The detection decision is then based on the output of this novelty filter.

D. Marketing

Bank analysts can also analyze the past trends, determine the present demand and forecast the customer behavior of various products and services in order to grab more business opportunities and anticipate behavior patterns. Data mining technique also helps to identify profitable customers from non-profitable ones. Another major area of development in banking is Cross selling i.e banks make an attractive offer to its customer by asking them to buy additional product or service.

E. Risk Management

Data mining technique helps to distinguish borrowers who repay loans promptly from those who don't. It also helps to predict when the borrower is at default, whether providing loan to a particular customer will result in bad loans etc. Bank executives by using Data mining technique can also analyze the behavior and reliability of the customers while selling credit cards too. It also helps to analyze whether the customer will make prompt or delay payment if the credit cards are sold to them.

V. CONCLUSION

Data mining is a technique used to extract vital information from existing huge amount of data and enable better decision-making for the banking and retail industries. They use data warehousing to combine various data from databases into an acceptable format so that the data can be mined. The data is then analyzed and the information that is captured is used throughout the organization to support decision-making. Data Mining techniques are very useful to the banking sector for better targeting and acquiring new customers, most valuable customer retention, automatic credit approval which is used for fraud prevention, fraud detection in real time, providing segment based products, analysis of the customers, transaction patterns over time for better retention and relationship, risk management and marketing.

Those banks that have realized the usefulness of data mining and are in the process of building a data mining environment for their decision-making process will obtain huge benefit and derive considerable competitive advantage in future.

REFERENCES

- [1] K. Chitra, B.Subashini, Customer Retention in Banking Sector using Predictive Data Mining Technique, International Conference on Information Technology, Alzaytoonah University, Amman, Jordan, www.zuj.edu.jo/conferences/icit11/paperlist/Papers/
- [2] K. Chitra, B.Subashini, Automatic Credit Approval using Classification Method, International Journal of Scientific & Engineering Research (IJSER), Volume 4, Issue 7, July-2013 2027 ISSN 2229-5518.
- [3] K. Chitra, B.Subashini, Fraud Detection in the Banking Sector, Proceedings of National Level Seminar on Globalization and its Emerging Trends, December 2012.
- [4] K. Chitra, B.Subashini, An Efficient Algorithm for Detecting Credit Card Frauds, Proceedings of State Level Seminar on Emerging Trends in Banking Industry, March 2013.
- [5] Petra Hunziker, Andreas Maier, Alex Nippe, Markus Tresch, Douglas Weers, and Peter Zemp, Data Mining at a major bank: Lessons from a large marketing application <http://homepage.sunrise.ch/homepage/pzemp/info/pkdd98.pdf>
- [6] Michal Meltzer, Using Data Mining on the road to be successful part III, http://www.dmreview.com/editorial/newsletter_article.cfm?nl=bireport&articleId=1011392&issue=20082, October 2004.
- [7] Fuchs, Gabriel and Zwahlen, Martin, What's so special about insurance anyway?, published in DM Review Magazine, http://www.dmreview.com/article_sub.cfm?articleId=7157, August 2003.