Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPARATIVE STUDY OF DATA MINING MODEL FOR CREDIT CARD APPLICATION SCORING IN BANK

[1] EVARISTUS DIDIK MADYATMADJA, [2] MEDIANA ARYUNI
[1] School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
[2] School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
E-mail: [1] emadyatmadja@binus.edu, [2] mediana.aryuni@binus.ac.id

ABSTRACT

The growth of credit card application needs to be balanced with the anticipation of bad credit risk because it does not use security collateral as warranty. The usage of credit scoring can be used to help the credit risk analysis in determining the applicant's eligibility. Data mining has been proven as a valuable tool for credit scoring. The aim of this research is to design a data mining model for credit scoring in bank in order to support and improve the performance of the credit analyst job. The proposed model applies classification using Naïve Bayes and ID3 algorithm. The accuracy of Naïve Bayes classifier is 82% and ID3 is 76%. So we can conclude that Naïve Bayes classifier has better accuracy than ID3 classifier.

Keywords: *Credit Scoring, Data Mining, Credit Card, Bank, Naïve Bayes, ID3, Classification*

1. INTRODUCTION

Kompas Financial Business Daily News [1] stated that the growth of credit card usage based on data from Bank Indonesia is 14.7 percent per year during 2008 to 2010. At the end of 2010 there were 13.57 million credit cards and their amount became 13.8 million in February 2011. That growth also includes risk increment for the bank. The risk is the possibility of bad debts. In addition, a credit card does not use security collateral as warranty. So, the credit analyst job in determining which credit card application to be approved or rejected is very crucial task.

Credit Scoring is used to support credit risk analysis [2], [3]. With proper Credit Scoring models, the bank will be able to evaluate whether an applicant is feasible to obtain a credit card.

In the literature [4] analyzed the nature and effects of missing data in credit risk modeling and take into account current scarce data set on consumer borrowers, which includes different percent and distributions of missing data.

Data mining has been proven as a valuable tool for the banking and retail industries [5], which identify useful information from a large size data. While previous literatures [6], [7], [2], [3] had applied data mining techniques for credit scoring.

By analyzing the results of previous researches which used data mining for credit scoring, this paper will discuss about how to design a data mining models for credit scoring in the Bank.

The purpose of this study are to determine the proper data mining system for credit scoring credit card application in Bank in order to improve the performance and support the credit analysts job.

The rest of the paper is organized as follows. Section 2 shows few related studies. Section 3 presents the proposed data mining model. Section 4 shows experimental evaluations and results. The paper concludes in Section 5.

2. STUDY LITERATURE

2.1 Data Mining

Data mining is the knowledge extraction from very large size data [8]. It is also called Knowledge Discovery from Data, or KDD. Figure 1 shows the knowledge discovery process in data mining.

The process of knowledge discovery in data mining consists of a sequence of iterative steps [8]: (1) data cleaning (noise and inconsistent data removal), (2) data integration (multiple data sources are combined), (3) data selection (to select the relevant data for the analysis), (4) data transformation (data is transformed into a suitable form for the mining process), (5) data mining (primary process that uses intelligent methods to extract data patterns), (6) pattern evaluation (to identify interesting patterns that represent knowledge), and (7) knowledge presentation (visualization of knowledge representation). In data

Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

preprocessing, we prepare data before mining process. Step one through four are other forms of data preprocessing.

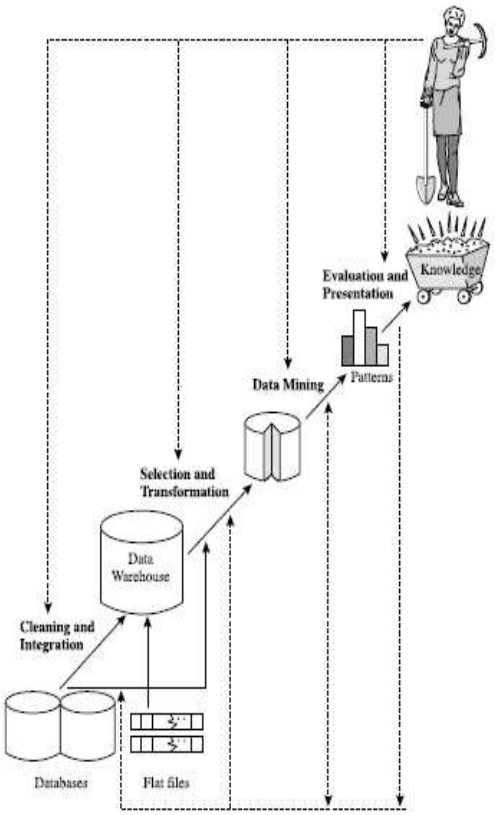

Figure 1: Data Mining as a Step in The Process of Knowledge Discovery [8].

2.2 Data Preprocessing

We need the high-quality data to produce the high-quality mining results [8]. There are many factors which comprise data quality, including accuracy, completeness, and consistency [8]. So, we need to apply data preprocessing first before develop the main data mining process.

The data preprocessing steps is shown in figure 2 [8]. The major steps involved in data preprocessing including data cleaning, data integration, data reduction, and data transformation. Data cleaning task is to "clean" the data by filling in missing values, noisy data smoothing, to identify or remove outliers, and to resolve inconsistencies.

In order to filling missing values, we can select to use the following methods [8]: (1) ignore the tuple (when the class label is missing), (2) fill in the missing value manually, (3) use a global constant to fill in the missing value (replace all missing attribute values by the same constant such as a label like "Unknown"), (4) use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value, (5) use the attribute mean or median for all samples belonging to the same class as the given tuple (for example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice, (6) use the most probable value to fill in the missing value (this may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction).

Data integration is often required in data mining. Its goal is merging data from multiple data stores [8]. In order to help reduce and avoid redundancies and inconsistencies in the resulting data set, we need very careful integration. So, the accuracy and speed of the subsequent data mining process will be improved [8].

Data reduction can be used to get a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Data reduction strategies include dimensionality reduction, numerosity reduction, and data compression [8].

The purpose of dimensionality reduction is to reduce the number of random variables or attributes under consideration [8]. Dimensionality reduction methods are wavelet transforms, principal components analysis, and attribute subset selection.

Numerosity reduction techniques replace the original data volume into smaller forms of data representation. These techniques may be parametric or nonparametric [8]. For parametric methods, a model is used to estimate the data, so we only need to store the data parameters instead of the actual data, which regression and log-linear models are examples. While, nonparametric methods in storing reduced representations of the data include histograms, clustering, sampling, and data cube aggregation [8].

The transformations are applied in data compression to obtain a reduced or "compressed" representation of the original data [8].

The data are transformed or consolidated into forms appropriate for mining in data transformation. Strategies for data transformation include the following [8]: (1) smoothing, to remove

Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

noise from the data (binning, regression, and clustering), (2) attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process, (3) aggregation, where summary or aggregation operations are applied to the data (for example, the daily sales data may be aggregated so as to compute monthly and annual total amounts), (4) normalization, where the attribute data are scaled so as to fall within a smaller range, (5) discretization, where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior), (6) concept hierarchy generation for nominal data, where attributes such as street can be generalized to higher-level concepts, like city or country.

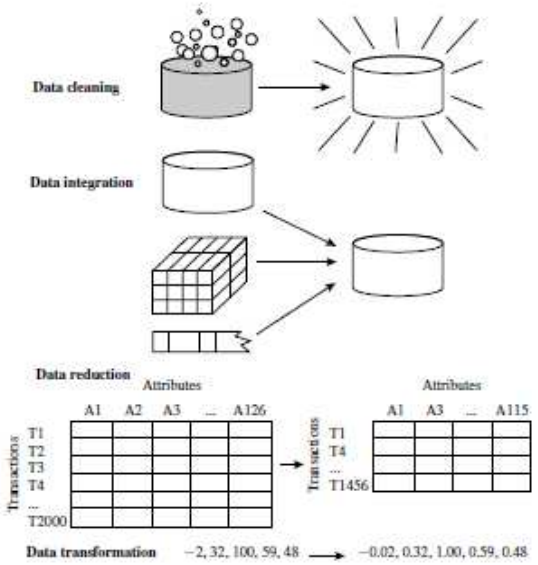

Figure 2: Forms of Data Preprocessing [8].

2.3 Classification

Classification is one of data mining functionalities. It finds a model or function that separates classes or data concepts in order to predict the class of an unknown object [8]. For example, a loan officer requires data analysis to determine which loan applicants are "safe" or "risky". The data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as "safe" or "risky" for the loan application data. These categories can be represented by discrete values, where the ordering among values has no meaning. Because the class labels of training data is already known, it is also called supervised learning [8].

Classification consist two processes: (1) training and (2) testing. The first process, training, builds a classification model by analyzing training data containing class labels. While the second process, testing, examines a classifier (using testing data) for accuracy (in which case the test data contains the class labels) or its ability to classify unknown objects (records) for prediction [9].

2.4 Naïve Bayes

A naïve (or simple) Bayesian classifier based on Bayes' theorem is a probabilistic statistical classifier [9], which the term "naïve" indicates conditional independence among features or attributes. Its major advantage is its rapidity of use because it is the simplest algorithm among classification algorithms. Hence, it can readily handle a data set with many attributes.

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows [8]:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if
   $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m; j \neq i$.
   Thus we maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem,
   $$P(C_i|X) = P(X|C_i)P(C_i)/ P(X) \quad (1)$$
3. As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = … = P(Cm), and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i)=|C_i,D|/|D|$,where $|C_i,D_j|$ is the number of training tuples of class C_i in D.
4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes).

2.5 Decision Tree

Ross Quinlan introduced a decision tree algorithm (known as Iterative Dichotomiser (ID3) in 1979. Decision tree classifiers construct a flowchart-like tree structure in a top down, recursive, divide-and-conquer, manner [9]. Using The Attribute Selection Method (ASM), it selects a splitting criterion (attribute) that best splits the given records into each of the class labels, then selected attributes become nodes in a decision tree.

Figure 3 shows basic algorithm for decision tree induction [8].

Algorithm: Generate decision tree. Generate a decision tree from the training tuples of data partition D.
Input:
Data partition, D, which is a set of training tuples and their associated class labels;
attribute list, the set of candidate attributes;
Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.
Output: A decision tree.
Method:
(1) create a node N;
(2) **if** tuples in D are all of the same class, C **then**
(3) return N as a leaf node labeled with the class C;
(4) **if** attribute_list is empty **then**
(5) return N as a leaf node labeled with the majority class in D;
// majority voting
(6) apply Attribute_selection_method(D, attribute list) to find the "best" splitting_criterion;
(7) label node N with splitting_criterion;
(8) **if** splitting_attribute is discrete-valued and
multiway splits allowed **then** // not restricted to binary trees
(9) attribute_list ← attribute_list – splitting_attribute;
// remove splitting attribute
(10) **for each** outcome j of splitting_criterion
// partition the tuples and grow subtrees for each partition
(11) let Dj be the set of data tuples in D satisfying outcome j;
// a partition
(12) **if** Dj is empty **then**
(13) attach a leaf labeled with the majority class in D to node N;
(14) **else** attach the node returned by Generate_decision_tree(Dj, attribute_ list) to node N;
endfor
(15) return N;

Figure 3: Basic Algorithm for a Decision Tree Induction [8].

2.6 Credit Scoring

Credit scoring is formally defined as a statistical method (or quantitative) which is used to predict the probability of the applicant's credit worthiness [6]. Credit scoring goal is to measure the financial risk of the loan so that the loan provider can make credit lending decisions quickly and objectively.

Credit scoring is not only useful for credit providers, but also for the credit borrowers. For example, credit scoring help reduce discrimination because the credit scoring model provides an objective analysis of the feasibility of the applicant. In addition, the credit providers focus only on information related to credit risk and avoid subjectivity of credit analysts [6]. In the United States, the variables related to discrimination such as race, sex, and age are not included in the credit scoring model. Only information that is not related to discrimination and have proven predictive for the performance of credit payments, can be included in the model. Credit scoring also supports to increase the speed and consistency of the credit application process and enables the automation of the credit application process. So, human intervention in the credit evaluation and cost can be reduced. The usage of credit scoring will support the financial institutions to measure the risk associated to lending to the applicant in a short time. In addition, the financial institutions can make better decisions [6].

2.7 Data Mining Applications for Credit Scoring

In literature [6] discussed the advantages and usage of credit scoring as well as the development its model using data mining. Data mining techniques which used for credit scoring models such as logistic regression, neural networks, and decision tree.

The selection of important features or attributes that influence the performance of credit scoring model was done in [7]. The process of selecting the best features or attributes used four data mining methods for feature selection such as ReliefF, Correlation-based, and Consistency-based Wrapper algorithms for improving three aspects of credit scoring model performance like simplicity, speed and accuracy.

Other previous researches about feature selection for credit scoring model were conducted in literatures [3], [11], [12], [13].

According to [10], customer identification by a behavioral scoring model is helpful characteristics

Journal of Theoretical and Applied Information Technology
20th January 2014. Vol. 59 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of customer and facilitates marketing strategy development.

Credit risk analysis became the main focus on the financial and banking industries [2]. To improve accuracy, the research developed a hybrid method that combined several representative algorithms and then used selective voting methodology.

Using a retail credit data of banks in Czech Republic, the two credit risk models are built based on logistic regression and Classification and Regression Trees (CART) [14].

The research [3] was used logistic regression, neural networks, C5, naïve bayes updateable, IBK (instance-based learner, k nearest neighbor) and raced incremental logit boost in order to select the best classifier which is used to improve the predictive accuracy of credit risk of credit card users in Malaysian Bank. In addition, feature selection using ID3 algorithm performed to select subsets of data that has the highest information gain and gain ratio values.

3. THE PROPOSED MODEL

The data were obtained from credit card application dataset of Bank XYZ in Indonesia. Available records in the dataset are classified into two class labels, 'approve' and 'reject'. The class label is determined by credit experts' knowledge.

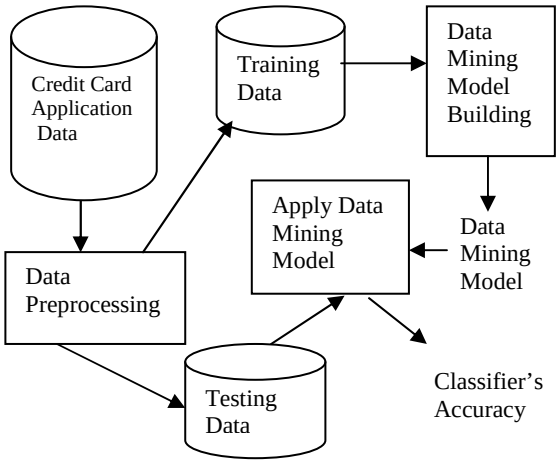

Figure 4: The Proposed Model.

Figure 4 shows the proposed data mining model that applies classification using Naïve Bayes and ID3 algorithm. We use dataset consists of 1000 records with two class labels ("Approved" or "Rejected"). Before building the model, data preprocessing is applied to the dataset. Then, the dataset is divided into 600 records for training data (to build the model) and 400 records for testing data (to apply the model and know the accuracy of the model).

Data preprocessing consists of data cleaning (fill in missing values), data reduction (feature selection) and data transformation.

Data mining model using Naïve Bayes and ID3 classifier to know which method gives the best performance.

4. RESULTS

The confusion matrix of the model using Naïve Bayes Classifier is shown in table 1. Table 2 shows ID3's confusion matrix.

Table 1: The Confusion Matrix of Naïve Bayes Classifier

	true Approved	true Rejected	Class Precision
Predicted Approved	163	35	82.32%
Predicted Rejected	37	165	81.68%
Class Recall	81.50%	82.50%	

Table 2:The Confusion Matrix of ID3 Classifier

	true Approved	true Rejected	Class Precision
Predicted Approved	160	56	74.07%
Predicted Rejected	40	144	78.26%
Class Recall	80.00%	72.00%	

The accuracy of Naïve bayes classifier is (81.50%+82.50%+82.32%+81.68%)/4 =82%, and ID3 has (80.00%+72.00%+74.07%+78.26%)/4 =76% of accuracy.

5. CONCLUSIONS

We have presented a data mining model which applies classification methods using Naïve Bayes and ID3 algorithm for credit scoring in credit card application. The best accuracy is achieved by Naïve bayes classifier (82%), while ID3 has 76% of accuracy. So we can conclude that Naïve Bayes classifier has better accuracy than ID3 classifier. In addition, the proposed data mining model able to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

improve the performance and support the credit analyst's job.

The selection of the important features is a challenge. In our further work, we plan to conduct some feature selection methods to know which method can give best classification performance.

REFERENCES:

[1] Kompas Financial Business Daily News, "Mudahnya Mendapat Kartu Kredit", http://bisniskeuangan.kompas.com/read/2011/04/15/09562379/Mudahnya.Mendapat.Kartu.Kredit, 2011.

[2] S. Kotsiantis, "Credit Risk Analysis using Hybrid Data Mining Model", *Int. Journal Intelligent Systems Technologies and Applications*, Vol. 2, No. 4, 2007.

[3] L.K. Sheng and T.Y. Wah, "A comparative study of data mining techniques in predicting consumers' credit card risk in banks", *African Journal of Business Management*, Vol. 5, No. 20, Available online at http://www.academicjournals.org/AJBM, ISSN 1993-8233, 2011, pp. 8307-8312.

[4] R.F. Lopez, "Effects of missing data in credit risk scoring, A comparative analysis of methods to achieve robustness in the absence of sufficient data", *Journal of the Operational Research Society*, Vol. 61, No. 3, 2010, pp. 486 -501.

[5] A.M. Hormozi and S. Giles, "Data Mining: A Competitive Weapon for Banking and Retail Industries", *Information Systems Management, Spring*, Vol. 21, No. 2, ProQuest Research Library, 2004.

[6] G.C. Peng,"Credit scoring using data mining techniques", *Journal of Singapore Management Review*, ISSN: 0129-5977, 2004.

[7] Y. Liu and M. Schumann, "Data mining feature selection for credit scoring models", *Journal of the Operational Research Society*, Vol. 56, 1099–1108 r 2005 Operational Research Society Ltd, 2005.

[8] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", San Francisco, Morgan Kaufmann Publishers, 2012.

[9] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. Chang, and L. Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", *J Med Syst*, Vol. 36, 2012, pp. 2431–2448.

[10] N.C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers", *Expert Systems with Applications*, Vol. 27, 2004, pp. 623–63.

[11] J. Li, H. Wei, and W. Hao, "Weight-selected attribute bagging for credit scoring", Hindawi Publishing Corporation, *Mathematical Problems in Engineering*, Volume 2013, doi: 10.1155/2013/379690, 2013.

[12] P. Somol, B. Baesens, P. Pudil, and J. Vanthienen, "Filter- versus wrapper-based feature selection for credit scoring", *International Journal of Intelligent Systems*, Vol. 20, Issue 10, October 2005, doi: 10.1002/int.20103, pp. 985–999.

[13] B. Waad, B.M. Ghazi, and L. Mohamed, "Rank aggregation for filter feature selection in credit scoring", *International Conference on Control, Engineering, and Information Technology (CEIT'13), Economics, Strategic Management of Business Process*, Vol. 1, 2013, pp. 64-68.

[14] E. Kocendaand M. Vojtek, "Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data", *William Davidson Institute Working Paper Number 1015*, Electronic copy available at http://ssrn.com/abstract=1912049, 2011.