CUSTOMER CREDIT SCORING: A COMPARISON OF THREE DATA MINING ALGORITHMS

Nafisseh Heiat, College of Business, Montana State University-Billings, Billings, MT 59101, 406-657-1627, nheiat@msubillings.edu

ABSTRACT

In this research Support Vector Machine, Bayes Network, and Decision Tree classification algorithms were used to predict and compare the performance of these three models and to identify the inputs or predictors that differentiate "good credit" from "bad credit". The results of the study indicate that the Support Vector Machine is better at predicting correctly customers with good credits from bad credits.

INRTODUCTION

In the financial industry, customers regularly request credit to make purchases. The risk for financial institutions to extend the requested credit depends on how well they distinguish the good credit applicants from the bad credit applicants. One widely adopted technique for solving this problem is Credit Scoring. Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting credit to customers (5, 6, 8).

In credit business, banks are interested in learning whether a prospective consumer will pay back their credit. The goal of this study is to model or predict a credit applicant can be categorized as a good or bad customer. In this study I have used three different data mining algorithms to identify the inputs or predictors that differentiate risky customers from others on the data set for testing and later deploy those models to predict new risky customers.

Credit scoring has become a critical and challenging business analytics issue as the credit granting businesses have been facing stiffer competition in recent years. Many statistical and data mining methods have been suggested to tackle this problem in the literature [7, 9, 11, 12, 13, 14]. Historically, discriminant analysis and linear regression have been the most widely used techniques for building score-cards. Both have the merits of being conceptually straightforward and widely available in statistical software packages. Other techniques which have been used in the credit scoring field include logistic regression, probit analysis, nonparametric smoothing methods, mathematical programming, Markov chain models, recursive partitioning, expert systems, and genetic algorithms, neural networks and classification models. Hand and Henley [10] examined a wide range of statistical and data mining methods that has been applied to credit scoring and discussed the advantages and disadvantages of these methods. Some researchers believe that the fact that significant portion of credit information is not normally distributed is a critical limitation in applying discriminant analysis and linear regression to credit scoring problems. However, Hand et al on the basis of empirical observation of credit scoring problems, concluded that non-normal distribution of credit information may not be a significant problem. Discriminant analysis also suffers from another weakness that it shares with logistic regression. They merely minimize the number of accepted bad loans given an exogenous acceptance rate, without any rule for picking this rate optimally[1].

On theoretical grounds one may argue that logistic regression is a more appropriate method than linear regression since the goal is to classify good and bad loans. In a comparative study, however, Henley found that logistic regression was no better than linear

regression [10]. Williamson compared logistic regression with discriminant analysis. He concluded that the logistic approach gave superior classification results but that neither method was sufficiently good to be cost effective [2, 3, 4, 15]. There are two other algorithms mentioned in the literature. Artificial neural Networks and
Artificial neural networks are defined as information processing systems inspired by the structure or architecture of the brain. They are constructed from interconnecting processing elements, which are analogous to neurons. The two main techniques employed by neural networks are known as supervised learning and unsupervised learning. In unsupervised learning, the neural network requires no initial information regarding the correct classification of the data it is presented with. The neural network employing unsupervised learning is able to analyze a multi-dimensional data set in order to discover the natural clusters and sub-clusters that exist within that data. Neural networks using this technique are able to identify their own classification schemes based upon the structure of the data provided, thus reducing its dimensionality [10]. Unsupervised pattern recognition is therefore sometimes called cluster analysis [3]. Supervised learning is essentially a two stage process; firstly training the neural network to recognize different classes of data by exposing it to a series of examples, and secondly, testing how well it has learned from these examples by supplying it with a previously unseen set of data. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. It provides projections given new situations of interest and answers "what if" questions. There are disadvantages in using ANN. No explanation of the results is given i.e. difficult for the user to interpret the results. They are slow to train due to their iterative nature. Empirical studies have shown that if the data provided does not contain useful information within the context of the focus of the investigation, then the use of neural networks cannot generate such information any more than traditional analysis techniques can. However, it may well be the case that the use of neural networks for data mining allows this conclusion to be reached more quickly than might ordinarily be the case.

Nonparametric methods, especially nearest neighbor method have been used for credit scoring applications. While the nearest neighbor method has some attractive features for credit scoring applications, they have not been widely used in the credit scoring applications. One reason being the perceived computational demand on the computer resources.

In general there is no overall best method for classification application. The choice of the method or methods will depend on the nature of the problem, on the data structure, the variables selected and the objective of the classification and the measures like misclassification rate used to evaluate the performance of the method.

METHODOLOGY

In this study we are going to use the following three classification algorithms that was recommended by Auto Classifier of SPSS Modeler and has proved to be very efficient in many areas of business analysis.

Decision Tree. Decision tree is the most commonly used approach to discovering logical patterns within data sets. Decision trees may be viewed as a simplistic approach to rule discovery because of the process used to discover patterns within data sets. Decision tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. Initially, you start with a training set in which the classification label (say, "bad

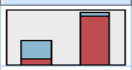
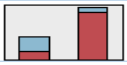
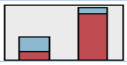
credit" or "good credit") is known (pre-classified) for each record. All of the records in the training set are together in one big box. The algorithm then systematically tries breaking up the records into two parts, examining one variable at a time and splitting the records on the basis of a dividing line in that variable (say, income > 50,000 or income <=50,000). The object is to attain as homogeneous set of labels (say, "good credit" or "bad credit") as possible in each partition. This splitting or partitioning is then applied to each of the new partitions. The process continues until no more useful splits can be found. The heart of the algorithm is the rule that determines the initial split rule [14]. The process starts with a training set consisting of pre-classified records. Pre-classified means that the target field, or dependent variable, has a known class or label: "diabetic" or "non-diabetic. The goal is to build a tree that distinguishes among the classes. For simplicity, assume that there are only two target classes and that each split is binary partitioning. The splitting criterion easily generalizes to multiple classes, and any multi-way partitioning can be achieved through repeated binary splits. To choose the best splitter at a node, the algorithm considers each input field in turn. In essence, each field is sorted. Then, every possible split is tried and considered, and the best split is the one which produces the largest decrease in diversity of the classification label within each partition. This is repeated for all fields, and the winner is chosen as the best splitter for that node. The process is continued at the next node and, in this manner, a full tree is generated.

Bayesian Network. A Bayesian network is a graphical model that displays variables (often referred to as nodes) in a dataset and the probabilistic, or conditional, independencies between them. Causal relationships between nodes may be represented by a Bayesian network; however, the links in the network (also known as arcs) do not necessarily represent direct cause and effect. For example, a Bayesian network can be used to calculate the probability of a customer being of good credit, given the presence or absence of certain attributes and other relevant data, If the probabilistic independencies between attributes and status of the customer as displayed on the graph hold true. Networks are very robust where information is missing and make the best possible prediction using whatever information is present. A Bayesian network is a graphical model that displays variables (often referred to as nodes) in a dataset and the probabilistic, or conditional, independencies between them. Causal relationships between nodes may be represented by a Bayesian network; however, the links in the network (also known as arcs) do not necessarily represent direct cause and effect.

Support Vector Machine. Support Vector Machine (SVM) is a classification and regression algorithm that uses machine learning theory to maximize predictive accuracy without over-fitting the data. SVM uses an optional nonlinear transformation of the training data, followed by the search for regression equations in the transformed data to separate the classes (for categorical targets) or fit the target (for continuous targets). SVM is particularly suited to analyzing data with very large numbers (for example, thousands) of predictor fields. SVM has been used in applications in many disciplines, including customer relationship management (CRM), facial and other image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition.

DATA

In this research I have used the data set with information pertaining to past and current customers who borrowed from a German bank for various reasons in this research. The data set contains information related to the customers' financial standing, reason to loan, employment, demographic information, etc. The German Credit data set (available at ftp.ics.uci.edu/pub/machine-learning-databases/statlog/) contains observations on 30 variables for 1000 past applicants for credit. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases). New applicants for credit can also be evaluated on these 31 "predictor" variables. We want to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. The original data set had a number of categorical variables, some of which have been transformed into a series of binary variables so that they can be appropriately handled by the data mining software.
The unimportant variables were removed by Filter node before starting to use algorithms. The Auto Classifier tool of Modeler was used to pick the best algorithms for our dataset.

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
☑		SVM 1	< 1	2,990	70	1.395	89.1	15	0.944
☑		C5 1	< 1	2,724.4	77	1.252	84.5	14	0.819
☑		Bay...	< 1	2,560	73	1.381	81.1	15	0.858

Bayes Network Results. Two Bayes models were used are TAN network and Markov. The confusion matrix shows that on average TAN model 81.1% of time correctly determine the status of credit and 18.9% of the time gets it wrong.

Markov Network. The second model used is Markov Bayes Network. Amount, CHK-ACCT, History, Duration, and Real State variables are the most important variables in descending order. It indicates that the accuracy rate of Markov model (78.3%) is lower than the TAN (81.1%). The Markov network assigns the same importance to the CHK-ACCT, History variables as the TAN network does. However, Amount of loan in this model is not as important as The TAN model.

Decision Tree Results. C5 model is used in this study. This model works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned. Accuracy rate is 84.5% which shows improvement over the first two Bayes networks. C5 also creates a decision tree graph and a decision rule set that can be used to design a guideline or policy for granting loans.

Support Vector Machine Results. Confusion matrix in Figure 11 states that SVM has the best accuracy rate among the three classification algorithms. It is 89.1% which is superior not only to other algorithms we used in this study but much improved compared to the model used in the literature. SVM accuracy rate is 3.5% higher than Decision Tree mode. However C5 creates the

decision rule set that informs more details and could be used for making policies for loan granting.

CONCLUSION

The confusion matrix accuracy rate shows that the SVM model is superior at predicting correctly good credits from bad credits. However, C5 model is only 3.5 of percentage points behind the SVM model. We may conclude that it would be better to use the C5 model since it yields more detailed information that could be used in designing policies for granting loans.

The number records with the good and bad credits are not equal in the German Credit dataset. To avoid bias, we used balancing based on the distribution of the target variable. In addition, in future studies, a larger dataset, should be used in the analysis. This would lead to more reliable results and a generalization of the results.

To evaluate the effectiveness of these models a cost-benefit analysis should be conducted. The gains from correct classification should be compared to costs of misclassification.

REFRENCES

[1] Altman, E.I., R.B. Avery, R.A. Eisenbeis and J.F. Sinkey, (1981), Application of classification techniques in business, banking and finance, JAI Press, Greenwich, CT.

[2] Amemiya, Y., (1985), *Advanced Econometrics*, Harvard University Press, Cambridge MA.

[3] Arminger, G., D. Enache and T. Bonne, (1997), Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feed-forward networks, *Computational Statistics* 12, 293-310.

[4] Bermann, G., (1993), Estimation and inference in bivariate and multivariate ordinal probit models, dissertation, Department of statistics, Uppsala University.

[5] Bernanke, B., and M. Gertler, (1995), Inside the Black Box: The Credit Channel of Monetary Policy Transmission, *Journal of Economic Perspectives* 9(4), 27-48.

[6] Boyes,W.J., D.L. Hoffman and S.A. Low, (1989), An Econometric Analysis of the Bank Credit Scoring Problem, *Journal of Econometrics* 40, 3-14.

[7] Dionne, G., M. Artis and M. Guillen, (1996), Count data models for a credit scoring system, *Journal of Empirical Finance* 3, 303-325.

[8] Gale, D., and M. Hellwig, (1985), Incentive compatible debt contracts: The one period problem, *Review of Economic Studies* LII, 647-663.

[9] Greene, W.E., (1993), *Econometric Analysis*, 2nd edition, Macmillan, New York.

[10] Henley, W.E., and D.J. Hand, (1996), k-nearest-neighbor classifier for assessing consumer credit risk, *The Statistician* 45 (1), 77-95.

[11] Tor Jacobson and Kasper Roszbach, Bank lending policy, credit scoring and Value at Risk, Research Department, Sveriges, Riksbank, Sweden, July 1998.

[12] Roszbach, K.F., (1998), Bank lending policy, A credit scoring and the survival of loans, Manuscript, Stockholm School of Economics.

[13] Steenacker, A., and M.J. Goovaerts, (1989), A credit scoring model for personal loans, *Insurance: Mathematics and Economics* 8, 31-34.

[14] Stiglitz, J.E., and A. Weiss, (1981), Credit rationing in markets with imperfect information, *American Economic Review* 71, 393-410.
[15] Williamson, S., (1987), Costly monitoring, loan contracts and equilibrium credit rationing, *Quarterly Journal of Economics* 102 (1), 135-145.