DATA MINING TOOLS FOR CREDIT

Abbas Heiat, Montana State University-Billings, 1500 University Drive, Billings, MT 59101, 406-657-1627, aheiat@msubillings.edu

ABSTRACT

In this research a binary classification set of models that included SVM, C5, Bayesian Network, C&R Tree, CHAID, Logistic Regression, and Neural Net data mining models were used to predict and compare the performance of these methods and to identify the inputs or predictors that differentiate customers with "good credit" from customers with "bad credit" in a German bank database. The results indicated that while all models yielded acceptable results, the SVM model was superior to other models in terms of correctly classifying good credits from the bad ones.

INTRODUCTION

In the financial industry, consumers 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 of consumer credit.

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 Bayes network to identify the inputs or predictors that differentiate risky customers from others on the training data, and later deploy those models to predict new risky customers.

REVIEW OF THE LITERATURE

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. 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, recursive partitioning, expert systems, and genetic algorithms, neural networks and classification models (Hand and Henley, 1997). Hand and Henley 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 (Hand, Oliver and Lunn, 1998) 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.

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 (Henley, 1995) found that logistic regression was no better than linear regression. Wiginton (Wiginton, 1980) 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.

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.

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 (Shmueli, Patel, and Bruce, 2010).

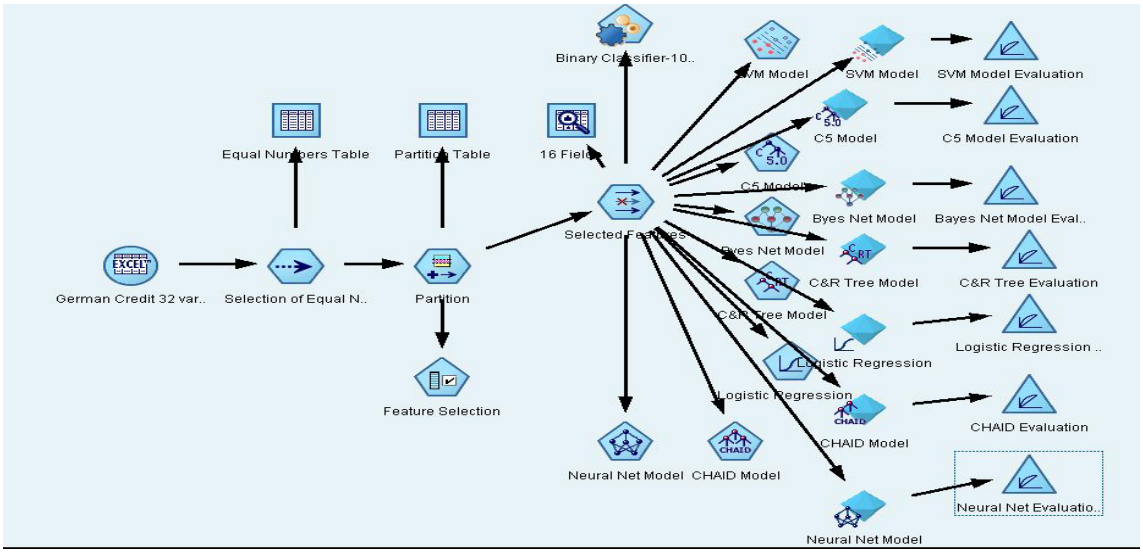
METHODOLOGY

I was interested to use a several data mining models and compare the performance of these models. I used Clementine data mining software for this research. Clementine has a binary classification feature that uses 10 different methods and reports the results of the ones with best performances. The followings methods were used: SVM, logistic regression, decision tree, C5, C&R tree, CHAID, and ANN.

FINDINGS

Figure 1 shows the data mining model developed and used in this research.

FIGURE 1. THE DATA MINING MODEL FOR CREDIT SCORING



The diagram starts with selecting the data set for the analysis. Next, data was randomly sampled to include equal numbers of good and bad credits. Data partitioned through Partition node into training and testing (validation) sets.

Figure 2- Statistics on Selected Features

Field	Graph	Type	Min	Max	Mean	Std. Dev
CHK-ACCT		Set	0.000	3.000	--	--
Duration		Range	4.000	72.000	21.843	12.191
HISTORY		Set	0.000	4.000	--	--
NEW-CAR		Flag	0.000	1.000	--	--
USED- CAR		Flag	0.000	1.000	--	--
RADIO/TV		Flag	0.000	1.000	--	--
AMOUNT		Range	339.000	18424.000	3379.352	2961.232
SAV-ACCT		Set	0.000	4.000	--	--
EMPLOYMENT		Set	0.000	4.000	--	--
INSTALL-RA...		Range	1.000	4.000	3.000	1.116
REAL-STATE		Flag	0.000	1.000	--	--
PROP-UNKN...		Flag	0.000	1.000	--	--
AGE		Range	19.000	75.000	35.263	11.570
RENT		Flag	0.000	1.000	--	--
OWN-RES		Flag	0.000	1.000	--	--
RESPONSE		Flag	0.000	1.000	--	--

Once the dataset is partitioned into training and validation (test) datasets, the statistically important predictor variables were selected using feature selection algorithm in the Clementine software. Figure 2 shows the 16 selected variables, their distribution graphs, their data types and corresponding basic statistics. Finally, the Binary Classification node which applies a variety of data mining models to the dataset with selected features is used. Table 1 reports the results for the seven most accurate models. Overall accuracy of predicting the customers with good credit ranges from 88.77 for SVM to 75.23 for the Neural Net. While the performances of all seven models are acceptable, SVM model is definitely superior to other models.

TABLE 1. PERFORMANCE RESULTS OF THE MODELS

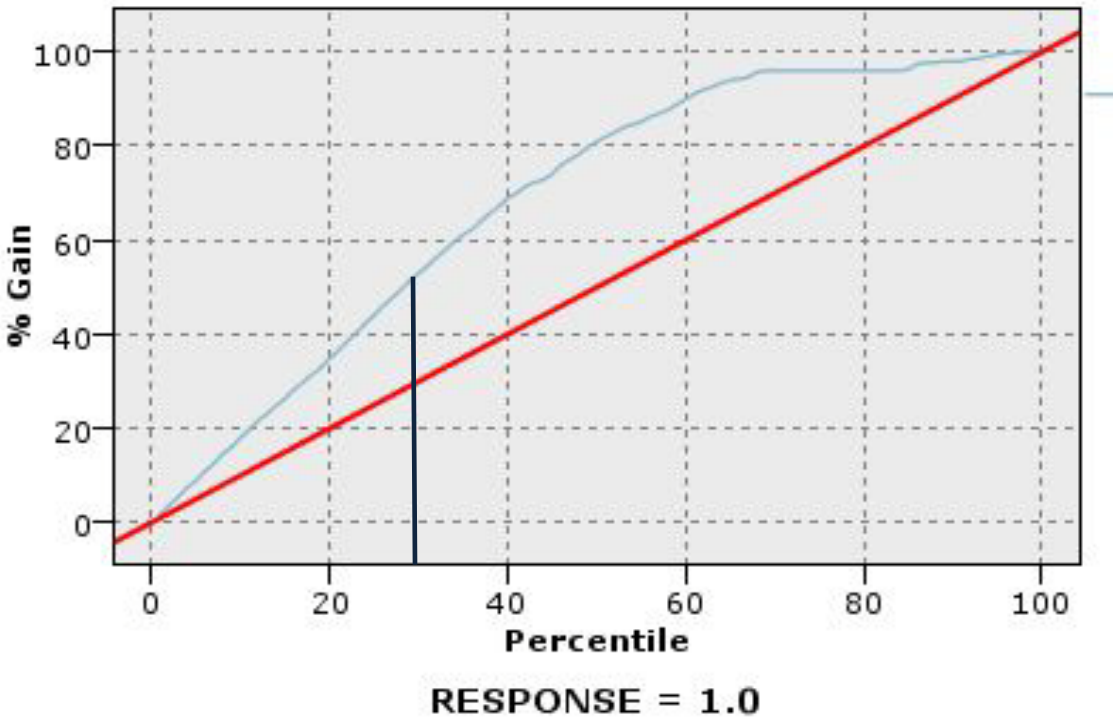
Model	**Lift (Top 30%)**	**Overall Accuracy (%)**
SVM	1.771	88.769
C5	1.655	83.231
Bayesian Network	1.667	78.462
C&R Tree	1.59	77.692
CHAID	1.631	76
Logistic Regression	1.581	75.846
Neural Net	1.657	75.231

The lift statistics in Table 2 confirms also the overall accuracy results. The Gain chart plots the values in the Gains (%) column from the table. Gains are defined as the proportion of hits in each increment relative to the total number of hits in the tree, using the equation: (hits in increment / total number of hits) x 100%. The chart effectively illustrates how widely you need to cast the net to capture a given percentage of all the hits in the database. The diagonal line plots the expected response for the entire sample, if the model were not used. In this case, the response rate would be constant, since one person is just as likely to respond as another. To double your yield, you would need to ask twice as many people. The curved line indicates how much you can improve your response by including only those who rank in the higher percentiles based on gain. For example, including the top 30% might net you more than 70% of the positive responses. Steeper curves mean more gain over random selection of records. Cumulative gain charts, which show the lift for the model at certain percentage level, always start at 0% and end at 100% as you go from left to right. For a good model, the gain charts will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal (base line) from lower left to upper right. The interpretation of an evaluation chart depends to a certain extent on the type of chart, but there are some characteristics common to all evaluation charts. For cumulative charts higher lines indicate better models, especially on the left side of the chart.

Figure 3 shows the Gain chart for SVM Model. The vertical dark blue line intersects the horizontal axis and the base line (red diagonal line, not using a model) at 30% Gain. This line intersects the curve (using SVM model) at 53.13 Gain. The Gain (0.771 %) shown in Table 2 is calculated by:

$$30X=53.13-30 \text{ or } X=23.23/30=0.771 \text{ and lift}=1+0.771=1.771$$

FIGURE 3. SVM GAIN CHART



CONCLUSION

The classification rate and the Lift statistics show that the SVM model is superior at predicting correctly good credits from bad credit. To avoid bias, this paper used equal number of good and bad credits which reduced the total number of records to six hundred. We need to apply model to a much larger data set to be able to generalize the results.

In future, to evaluate appropriately the effectiveness of the models used in this research, we need to include the cost of misclassifications and the benefits of correct classifications in our analysis. Including costs and benefits might change our results about the most effective and/or most profitable) model(s).

REFRENCES

Arminger, G., D., and Bonne, T., 1997, "Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feed- forward networks", *Computational Statistics*, Vol. 12, pp. 293-310.

Caudill, M., and Butler, A., 1990, *Naturally Intelligent Systems*, the MIT Press Cambridge, Massachusetts.

Darwiche, Adnan, 2003, "A differential approach to inference in Bayesian networks", *Journal of the ACM*, Vol. 50, Issue 3, pp. 280-305.

Hand, D. J. and Henley, W. E., 1997, "Statistical Classification Methods in Consumer Credit Scoring: a Review", *Journal of Royal Statistics Society*, Part 3, pp. 523-541.

Hand, D. J., Oliver, J. J., and Lunn, A. D., 1998, "Discriminant analysis when the classes arise from a continuum", *Pattern Recognition*, Vol. 31, pp. 641-650.

Henley, W. E., 1995, *Statistical Aspects of Credit Scoring*, PhD theses, The Open University, Milton, Keynes, UK.

Sabzevari, H., Soleymani, M., Noorbakhsh, E., "A comparison between statistical and Data Mining methods for credit scoring in case of limited available data", *CRC Conference, 2007.*

Shmueli, G., Patel, N., and Bruce, P., 2010, Data *Mining for Business Intelligence*, Wiley, New Jersey.

Wiginton, J. C., 1980, "A note on the comparison of logit and discriminant models of consumer credit behavior", *Journal of Financial Quantitative Analysis*, Vol. 15, pp. 757-770.

Williamson, S., 1987, "Costly monitoring, loan contracts and equilibrium credit rationing", *Quarterly Journal of Economics, VOl.*102 (1), pp. 135-145.