A COMPARATIVE STUDY OF DATA MINING TECHNIQUES IN PREDICTING CONSUMERS CREDIT CARD RISK IN BANKS

Ling Kock Sheng¹, Teh Ying Wah²
¹ Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: ksling99@siswa.um.edu.my
²Department of Information Science, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: tehyw@um.edu.my

ABSTRACT

It is increasingly important for banks to analyze and understand their risk’s portfolio, in particularly those related to credit card as it is one of those instruments that provide the highest return as well as potentially, most risky. The competitive level for the credit card business have increased significantly in the last few years due to the push by the banks to brought forth many new differentiated services and products to win over market share. Other strategies employed by banks are also to improve their current portfolio by extending the application of their services through, for instance, extending credit limit where it is deemed credible and appropriate, improve performance, for example, by predicting future payment behavior of consumers and lastly, proper management of bad debts. All such emphasis are usually managed using a typical score known mostly as credit scoring by the banks. Credit score is usually established or derived from three major sources of information and data, they are historical credit information from Credit Bureau or Central bank, current and historical transactions extracted from the bank’s database and collaterals and guarantees provided. A measure of consumer who will default in their payment or bad credit and those would not, can therefore be used to provide a good yardstick in substantiating the level of score for credit worthiness. The focus of the banks should therefore be, finding the best classifier from the model generated that provides the best predictive accuracy.

There were substantial literature advocating the use of data mining techniques such as Logistic Regression, C5 and Neural Network for providing the best predictive accuracy for such cause. Prior to generating all the models for comparison, the initial set of data is also loaded to an ETL system customized to perform feature selection or attribute relevancy analysis using ID3 algorithm, compiling a subset of data with the highest information gain and gain ratio. An extended test is performed to use equal length binning on some attributes to find if it affect the relevancy of each attribute. The selected subset of data is used to generate various data mining models using different training and testing sizes. C5 emerged consistently as the technique that has generated the best models with an average of predictive accuracy at 91.64%.

Keywords: Data Mining techniques, Credit Scoring, Predictive Accuracy, Credit Risk, C5, Logistic Regression and Neural Networks.

1.0 INTRODUCTION

One of the primary services of a bank is to provide credit facilities or lending to consumers. This risk for the bank to extend the requested credit depends on how well they distinguish between good and bad credit borrowers. The management of risks is vital to the financial soundness of the bank. It is usually a very important aspect of the bank’s business strategy. The management philosophy for banks would therefore, emphasize primarily on identifying, measuring, monitoring and managing their portfolio within a robust risk management framework. Returns has to commensurate based on risks taken.

With increasing competition, the recent credit crunch and world wide recession, banks are tightening their lending policies. Banks are now not only facing the highly competitive markets within their industry, they are also hard pressed to balance between liquidity, market share, application in terms of higher credit provided to borrowers or investment, performance and also bad debt’s recovery. One of the most critical credit risks has been an explosive growth of credit card usage. According to Malaysia Rating Corp. Bhd. Chief economist, Nor Zahidi Alias, the number of credit cards in circulation for primary holders, have tripled to 9.67 million as of March 2009 from a mere 2.92 million seven years ago[1]. In addition, the credit line for card holders more than quadrupled to a whooping RM103.8 billion versus RM 24.3 billion in
January 2002[1]. Such phenomena growth suggests Malaysia and other regions, ie Korea [11] is relying more on credit cards in their consumption, thereby creating a burden not only from the consumer’s point of view, but also the level of risks that the bank has to take in. It is therefore important for banks to understand their current portfolio and the risks level that they are in. Such understanding provides insights into balancing their portfolio for[9]:

1. Market - Targeting the right candidate as credit card holders
2. Application – The level of credit extended to such card holders
3. Performance – Understanding and predicting the future payment behavior
4. Bad Debt Management – Understanding and deriving the right collection policy for maximizing amount recovered.

All mentioned thrust above are usually managed using a typical score known mostly as credit scoring. Credit scoring is usually established or derived from three major sources of information and data, they are historical credit information from Credit Bureau or Central bank, current and historical transactions extracted from the bank’s database and collaterals and guarantees provided. A measure of those who will default in their payment and who would not does provide the needed yardstick in substantiating the level of score for credit worthiness. The focus of the banks should therefore be, finding the best classifier from the model generated that provides the best predictive accuracy

2.0 DATA MINING TECHNIQUES AND CREDIT RISKS

Various data mining techniques are employed in data mining on various domains of the banking and credit card’s management of risks. Sinha and Zhao[8], uses several techniques to examine the performance on business problems, mostly involving binary classification of two categories, ie. bankrupt or non-bankrupt, bad credit or good credit and others. In classification, a set of training data is use as input to build a model describing the predetermined set of data classes. Once the predictive accuracy of the model is acceptable, the model can be used to predict future data tuples[3][4].

Earlier papers on the development of application on the use of data mining techniques are closely related to credit scoring[6] and bank’s credit analysis, has been found in areas such as bankruptcy prediction[5][7], credit risk assessment[10], and credit evaluation[8] are found.

There is ample evidence of on data mining techniques that were employed extensively in business applications in areas such as bankruptcy, credit scoring and risk assessment. The more common techniques are, Neural Network[12][13], Logistic regression[14][15] and decision tree[16].

3.0 DATA MINING TECHNIQUES, CLASSIFIER AND PREDICTIVE ACCURACY

The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data. This can be estimated using one or more data set or test sets[2].

Looking at the three data mining techniques under this research study, ie Logistics Regression, Decision Tree and Neural Networks, there are varying predictive accuracy of each of the technique under differing circumstances. Decision Tree has a better predictive accuracy in a study by Goh[16] on credit scoring whereas Neural Networks has a predictive accuracy in Zurada and Lional[17] on bad debt recovery.

Due to the complexity involved with each technique, the above ranking should not be deemed as conclusive as their results may differ from dataset sizes or sample sizes, generalization and discretization level and other pre-mining processes such as smoothing, aggregation and aggregation construction, each according to the strength where each model will make an impact with.

4.0 RESEARCH DESIGN

The performance measure use to evaluate Logistic Regression, C5 and Neural Networks data mining techniques is based on the highest level of predictive accuracy amongst the data mining models
generated. The model which is consistently scoring the highest predictive accuracy using different percentage of training and testing data is deemed having the best classifier[2]. A method that is quite similar to Holdout Method and random sampling[2] is used to partition the bank’s data comprising 5000 customer payment records into 50%, 80% and 90% respectively as training set and 50%, 20% and 10% respectively as the test set. The training set is used to derive the model whereas the test set is to check on the accuracy of it.

Prior to obtaining the right subset of data that is relevant for generating data models, the full set of data is loaded to a customized Extraction, Transformation and Loading (ETL) as well as attribute relevancy system to compute information gain and gain ratio level by using ID3 (Iterative Dichotomiser) algorithm.

5.0 DATA ANALYSIS

Information computed includes info gain and gain ratio based on different binning sizes of the attributes. The best information gain obtained for each attribute, age, location, sex, accrued amount and credit limit is recorded as 0.00600, 0.00669, 0.00342, 0.17848 and 0.00363 for bin of sizes 50, 50, 2, 50 and 50 respectively. The best gain ratio obtained for each attribute is recorded as 0.00127, 0.00122, 0.00342, 0.12053 and 0.00109 for bin of sizes 50, 50, 2, 10 and 50. This is as shown in figure 5.1.

The information gain data recorded for different attributes also indicate an increasing gain when bin size is increased from 10 to 50. This similar pattern is also found in gain ratio except for attribute Acr – accrued amount, which dwindle as the bin size increases. This is as shown in Figure 5.2

On a whole, except for Acr – accrued amount, all other attributes are having more or less the same level of information gain and gain ratios. Since there are not many attributes that are available, all five attributes should be used for generating the models. The results also denote the need to use the right binning sizes so that the level of maximum gains can be accurately ascertained, especially if there are more attributes to consider.
The testing results indicated a trend amongst the models in which C5 achieve a better predictive accuracy on a whole for most of the model with different equal-length binning sizes. The best predictive accuracy is at 91.64% at three different binning sizes. Attributes of different binning sizes does have an influence on the test results.

6.0 CONCLUSION

C5 emerged consistently as the technique that has generated the best models with an average of predictive accuracy at 91.64%.

REFERENCES


**BIOGRAPHY**

Ling Kock Sheng has 18 years of working experience with several multi-nationals in various industries and in various capacities. He is currently pursuing the Master of Computer Science degree from University of Malaya and is researching topics in data mining.

Teh Ying Wah is a senior lecturer at the Faculty of Computer Science and Information Technology in University of Malaya. He holds a PhD in computer science from University of Malaya, Malaysia. He is currently carrying out research in data mining, data warehouses and E-Commerce.