International
Business Management

SINCE 2000
MEDWELL PUBLICATIONS

Comparison of Data Mining Techniques for the Predictive Accuracy of Credit Risk of Leasing Customers in Sri Lanka

H.A.P.L. Perera
Department of Accountancy, University of Kelaniya, Colombo, Sri Lanka

Key words: Data mining, credit risk, logistic regression, Naive Bayes, decision tree, neural networks

Corresponding Author:
H.A.P.L. Perera
Department of Accountancy, University of Kelaniya, Colombo, Sri Lanka

Page No.: 61-64
Volume: 14, Issue 2, 2020
ISSN: 1993-5250
International Business Management
Copy Right: Medwell Publications

Abstract: This research has been conducted to construct a model which can be used to measure the predictive accuracy of the credit risk of leasing customers in Sri Lanka and to compare different data mining techniques for the purpose of selecting the best and adequate technique to predict the credit risk. The dataset employed in this study was obtained from one of the leading finance/leasing companies in Sri Lanka. Altogether 8235 customers/data instances have been considered for the analysis under 24 variables. The data set was divided in to two different datasets, training (60%) and test (40%). The Waikato environment for knowledge analysis machine learning software was the major software tool used for the entire model construction process. Four main data mining techniques (logistic regression, Naive Bayes, decision tree-J 48 and neural networks) were used to construct models and results from each model were obtained and compared with other techniques. According to the results of the study, we can conclude that, a model constructed using the neural network as the best model to predict the payment accuracy of leasing customers in Sri Lanka.

INTRODUCTION

Every human being today have the opportunity to accumulate enormous amount of data than ever before. With the advancements in the information technology such as from internet of things to integrated management systems, the world is generating vast amount of data at an accelerating pace. Few years back, this massive amount of data was a big issue and a challenge to the corporate world. By now many companies have decided that big data is not an issue anymore but a new opportunity: once they establish strategies in place for managing large volume of both structured and unstructured data. Fan and Bifet[1] the data mining is the area which helps organizations to discover hidden patterns and relationships from vast amount of data.

This study is based on the predictive data mining techniques in the area of finance. The previous studies on data mining in finance domain were mainly focused on financial fraud detection. Phu *et al*.[2] and Sharma and Panigrahi[3], Bankruptcy prediction[4, 5] and customer credit risk analysis[6, 7]. The major limitation found on previous research is that the inadequate attention given to measure the predictive accuracy of the credit risk of the customers of the finance companies and lack of comprehensive comparison of data mining measures and techniques. In this study, an effort is made to develop a model using four main data mining techniques to predict the consumer payment ability of a Sri Lankan leasing company, compare those data mining techniques with each other and finally suggest the best model which can be used to predict the payment ability of leasing customers in Sri Lanka. The results of this study provide the way and model to analyze the customers prior to offer the leasing facility to them and the decision makers of the company can easily identify the best mining technique that can be used in diverse situations without wasting much time.

Literature review: There are core financial tasks, such as currency exchange rate, forecasting stock market, bank bankruptcies, credit rating, loan management, bank customer profiling and money laundering analyses are core financial tasks for data mining[8, 9, 1]. In their study, Bekhet and Eletter studied a credit risk assessment model for Jordanian commercial banks, using neural scoring approach. This study proposes two credit scoring models using data mining techniques to support loan decisions for the Jordanian commercial banks[9]. Similar study has also been conducted by Mandala and coworkers to assess the credit risk in a rural bank and In their study, Mandala and coworkers, proposed a decision tree model and aimed on the reduction of number of non-performing loans[10]. Siami and coworkers conducted a review on credit scoring in banks and financial institutions via. data mining techniques from 2000-2012[11]. According to, the review study conducted by Srivastava and Anchal, credit cards are being widely used for online purchases and also an increase in the risks related to them[12]. In this study, they analyzed the factors causing the operational financial and the security risks and the data mining methods used to minimize them. Lee[13] examined the literature and pointed out the importance of accurately predicting the credit risk. In this study, they reviewed papers which had applied statistic model, neural networks, learning vector, software-computing and hybrid model in credit risk problem. According to the findings of their study, most of the researches prefer to use the neural network method[13].

MATERIALS AND METHODS

According to the literature, there are significant number of researches have been conducted on this area and various researcher's have compared the performance of different data mining techniques and conclude with the prediction accuracy of the each technique. However not much research has been done using the data from Sri Lankan companies, especially for the leasing industry. So, the aim of the current study is to construct a data mining model to measure the predictive accuracy of payments of leasing customers in Sri Lanka. Two others main objectives of the current study are to compare data mining techniques that can be used for financial data analysis and to make recommendations on best data mining techniques for the analysis of financial data.

The dataset employed in this study is obtained from one of the leading finance/leasing companies in Sri Lanka. All the agreements which were matured on December 2015 were considered for the study under 24 variables. Altogether 8235 customers/data instances have been considered for the analysis. Those who were active on the maturity date were considered as "Active" customers and those who were not in the active mode (either ceased or legal transfer) on the maturity date were considered as "Sink". All the variables used in the sample were extracted from the original data set which was obtained from the ERP system of the company. Eight variables were removed from the original variable list due to lack of relatedness and sixteen variables were taken for the variable refining process. There are two types of independent variables used in this study. Some variables are related with the loan/leasing agreement and the others are related with the demographic and socio-economic conditions of the customer/lessee. The backward elimination process is used to refine variables using SPSS. There were nine independent variables and the only dependent variable of FacSTS (Facility Status)-The Status of the agreement, refined from the process.

The selected variables then be examined for the multicollinearity. Before perform the collinearity diagnostics the data set converted in to standardize values. For the standardization or the normalization process, the binary encoding has been used in the current study. Since, categorical variables contain more than two possible values, 1-of-C dummy encoding method has been employed. To encode numerical variables min-max normalization was applied. The standardized data set was used to calculate the multicollinearity using SPSS. The tolerance level and the VIF factor were the two measures used to identify the problem of multicollinearity. According to the values obtained from the collinearity diagnostic test, values of both tolerance level and the VIF factor were recorded inside the threshold values of both the measurements. So, it was indicated that an unavailability of high correlation between independent variables. The data set was divided in to two different datasets, training (60%) and test (40%) data sets. WEKA machine learning software was the major software tool used in the entire model construction process.

RESULTS AND DISCUSSION

Four main predictive data mining techniques; Logistic regression, Naive Bayes, decision tree-j 48 and

Table 1: Summary of the performance of four classifiers

Data mining technique	Classification accuracy (%)	Kappa statistic (%)	AUC values (%)	F-measure (%)	MAE (%)	RMAE (%)
Logistic regression	**91.92**	0.57	62.40	**95.80**	**14.48**	**27.00**
Naïve Bayes	91.89	0.26	61.15	95.75	14.70	27.10
Decision tree- J48	91.90	6.30	63.35	95.75	14.09	27.08
Neural networks	91.58	**13.00**	**66.70**	95.55	16.02	27.53

Bold values are significant

Table 2: The best model to predict the credit risk of leasing customers in Sri Lanka

Data mining technique	Classification accuracy (%)	Kappa statistic (%)	AUC values (%)	F measure (%)	MAE (%)	RMAE (%)
Neural networks	91.58	13	66.70	95.55	16.02	27.53

neural networks are used to construct models. From Table 1 compares the main measures used to analyze the performance in accurate prediction and classification of all four data mining models used in the current study.

The performance of four classifiers was evaluated in terms of their classification accuracy, Kappa statistic, AUC value, F-measure, MAE and RMAE values. As seen in Table 1, almost all the algorithms generated high classification accuracy with the highest amount of 91.92%. Which was recorded in logistic regression. However, the Kappa statistic values of all four models were very low. The highest Kappa statistic in our study was recorded in neural networks (13%). The overall classification ability of individual models, (especially when the data is highly skewed in nature) is represented by their AUC values. The highest average AUC value was recorded in neural networks (66.7%). F-measure is represented the highest classification accuracy by minimizing the misclassification costs (recall and precision). As the results revealed in Table 1, the logistic regression model has the highest F-measure (95.8%) in comparison with other three models. But all the other models also achieved similar F-measures with more than 95% level of accuracy. Therefore, the logistic regression model can successfully reduce the possible risk of extra losses due to high misclassification.

The MAE and the RMAE can be used together to diagnose the variation in the errors in a set of forecasts. Since, both measures represent error values, lowest would be the better. According to the Table 1, lowest MAE and RMAE values were recorded in logistic regression (14.48 and 27%, respectively). However, these values are not particularly useful in a classification task because they are used to assess performance when the task is numeric prediction.

Based on the threshold values of each measure, logistic regression achieved the best level of results among all four constructed models. Although, the logistic regression model achieved the best level of results among used measures, the values of Kappa statistic and AUC has recorded very low. Since, the data set is highly skewed in nature these two values are the most important and other values could be misleading. And also, if the actual need of the model is to classify no of "Active" instances correctly then a classifier with a lower Kappa but better rate of classifying "Active" might be good. But in the current study our focus is to measure both "Active" as well as "Sink" instances with the same weightage. It has been observed that the highest value for both Kappa statistic and the AUC values were obtained by the neural network model. Therefore, we can conclude that, as per Table 2 with the classification accuracy of over 91%, Kappa statistic of 13%, AUC value of 67% and F-measure of over 95%, a model constructed using the neural network as the best model to predict the credit risk of leasing customers in Sri Lanka.

CONCLUSION

As per the objectives of the study, we developed a data mining model to measure the predictive accuracy of the credit risk of leasing customers in Sri Lanka which enables decision makers in the leasing industry to identify customers with healthy future financial behaviors with the company. And also we compared four data mining techniques for the above mentioned objective and made recommendations on the best technique for the predictions of credit risk in the finance domain. The scope of the current study was limited only for the comparison of selected measurements of four data mining techniques.

RECOMMENDATIONS

The future research should be more focused on conducting a rigorous analysis of these techniques in terms of their underline concepts. Due to the time and resource limitations, the problem identification has been done only using the literature survey. If future researches can contact company representatives, industry experts, monitoring agencies and policy makers of the industry, the problem could be more specific. This would help to construct models to predict customer behaviors and to understand which technique is genuinely better than others. Finally, the scope of the study should be enhanced to predict the credit risk of the customers who are dealing with any kind of a finance transaction. It can be in the leasing industry, banking industry or insurance industry.

REFERENCES

01. Fan, W. and A. Bifet, 2013. Mining big data: Current status and forecast to the future. ACM. SIGKDD. Explor. Newsl., 14: 1-5.
02. Phua, C., V. Lee, K. Smith and R. Gayler, 2014. A comprehensive survey of data mining-based fraud detection research. Comput. Eng. Finance Sci., 1: 1-14.
03. Sharma, A. and P.K. Panigrahi, 2013. A review of financial accounting fraud detection based on data mining techniques. Intl. J. Comput. Appl., 39: 39-47.
04. Atiya, A.F., 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE. Trans. Neural Networks, 12: 929-935.
05. Bapat, V. and A. Nagale, 2014. Comparison of bankruptcy prediction models: evidence from India. Accounting Finance Res., 3: 91-98.
06. Alam, M., C. Hao and K. Carling, 2010. Review of the literature on credit risk modeling: Development of the past 10 years. Banks Bank Syst., 5: 43-60.
07. Sheng, L.K. and T.Y. Wah, 2013. A comparative study of data mining techniques in predicting consumers credit card risk in banks. Afr. J. Bus. Manage., 5: 8307-8312.
08. Yeh, I.C. and C.H. Lien, 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Exp. Syst. Appli., 36: 2473-2480.
09. Bekhet, H.A. and S.F.K. Eletter, 2014. Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. Rev. Dev. Finance, 4: 20-28.
10. Mandala, I.G.N.N., C.B. Nawangpalupi and F.R. Praktikto, 2012. Assessing credit risk: An application of data mining in a rural bank. Procedia Econ. Finance, 4: 406-412.
11. Sadatrasoul, S.M., M. Gholamian1, M. Siami1 and Z. Hajimohammadi, 2013. Credit scoring in banks and financial institutions via. data mining techniques: A literature review. J. Al Data Min., 1: 119-129.
12. Srivastava, S. and A. Garg, 2013. Data mining for credit card risk analysis: A review. Intl. J. Comput. Sci., 3: 193-200.
13. Lee, M.C., 2012. Enterprise credit risk evaluation models: A review of current research trends. Intl. J. Comput. Appl., 44: 37-44.