3-22-2019

The credit risk evaluation models: an application of data mining techniques

Cuong Nguyen
West Texas A&M University, ccnguyen2@buffs.wtamu.edu

THE CREDIT RISK EVALUATION MODELS: AN APPLICATION OF DATA MINING TECHNIQUES

Cuong Chi Nguyen
West Texas A&M University
ccnguyen2@buffs.wtamu.edu

ABSTRACT

In the banking sector, credit risk assessment is an important operation in ensuring that loans could be paid on time, and banks could maintain their credit performance effectively; despite restless business efforts allocated to credit scoring yearly, high percentage of loan defaulting remains a major issue. With the availability of tremendous banking data and advanced analytics tools, classification data mining algorithms can be applied to develop a platform of credit scoring, and to resolve the loan defaulting problem. With the dataset of 5,960 observations representing information about characteristics of underlying-collateral loans, the paper sets out a data mining process to compare four classification algorithms, including logistic regression, decision tree, neural network, and XGboost in performance. Via the confusion matrix and Monte Carlo simulation benchmarks, the XGboost outperforms as the most accurate and profitable model, displaying a high consistency about the major factors which could be attributable for default possibilities of the credit scoring.

Keywords

Credit risk evaluation, data mining, classification, simulation.

INTRODUCTION

From the beginning of the financial market, the demand of credit analysis emerged with the need in borrowing and lending of money and the purchasing authorization to payback loan in future (Louzada, 2016). However, in today's business world where the banks and financial institutions all have to face highly competitive threats, the credit risk evaluation has become an important operational task determining the bank's overall performance and initiating a lot of challenges to the banks and the decision makers. In reality, the risk of credit in many financial intermediates can account for 60% of their business activities (Morales et al., 2013). On the other hand, big challenges come with greater business opportunities; credit scoring also offers big competition advantages for whom those can effectively develop and utilize the prediction models in assessing and preventing potential default risk. Undoubtedly, when successfully evaluate the credit risk, a bank can minimize their loan default losses, and maximize lending profits, which helps to maintain the sustainable existence of the firm. Theoretically, the purpose of credit scoring (or credit risk assessment) is to identify different groups in a population (Morales et al., 2013). According to Thomas et al. (2002), credit risk evaluation is "a set of decision models and their underlying techniques that aid credit lenders in granting of credit". Under the current competition context, financial intermediates need to effectively select and implement credit scoring efforts for optimal credit granting. In this paper, the two major classification groups include good payers (who keep up with their frequent payments), and bad payers (who are likely to default).

Nowadays, with the availability of tremendous data collected from customer relationship management operations, and advanced data mining algorithms, financial companies can discover business insights, causality and correlation from business information, contributing to resolve the credit risk issues (Brown and Mues, 2012). Throughout empirical studies, a wide range of classification techniques have been developed for the credit risk evaluation, from traditional statistical analysis to machine learning/data mining tools (Xia et al., 2017). In spite of the large amount of financial and business resources allocated to credit scoring every year, the inconsistency in empirical results stands still, and high rate in loan defaulting (loan bankruptcy) remains a major issue in bank management. Therefore, banks have always tried to investigate the customers' background, as well as monitor their interactions with the banks to detect potential signs of credit defaulting, including negative credit records, high debt/income ratio, or ambiguous credit behaviors, and propose suitable manner to avoid infeasible loan offers (Lee et al., 2006). The paper is aimed to design and test the appropriateness of some classification models, including logistic regression, decision tree, artificial neural network, and extreme gradient boosting, by investigating the loan applications' characteristics and the borrowers' background information. Following this approach, the banks would be able to propose better predictions for the feasibility of loan applications, as well as to develop efficient lending policies.

LITTERATURE REVIEW

In general, an efficient data mining framework is an essential component of the customer relationship management business foundation (Xu and Qiu, 2008), via suitably applying, validating, and evaluating predictive models to solve business challenges. Particularly, the need for efficient risk management requires banks to seek a continuous enhancement in data mining techniques

applied for credit risk analysis (Louzada, 2016), and in the frame of this paper, the credit scoring prediction and determinants of loan default. In the topic of credit scoring, a large number of well-known studies, such as Arminger et al. (1997), West (2000), Baesens et al. (2003), or more recently, Xia et al. (2017) have been researched, and suggested valuable solutions for the risk management operation of the banking sector. Associated with this is also a wide range of distinct classification techniques for loan risk assessment, which have been applied in recent years; some typical techniques are listed as following:

Researchers	Data Mining Techniques
Arminger et al. (1997)	Logistic Regression, Decision Tree, Neural Network
Altman (1994)	Neural Network, Linear Discriminant Analysis, Quadratic Discriminant Analysis
Baesens et al. (2003)	Logistic Regression, Decision Tree, Neural Network, Linear Discriminant Analysis, Quadratic Discriminant Analysis, k-Nearest Neighbors, Support Vector Machine
Desai et al. (1996)	Logistic Regression, Neural Network, Linear Discriminant Analysis
West (2000)	Logistic Regression, Decision Tree, Neural Network, Linear Discriminant Analysis
Yobas et al. (2009)	Decision Tree, Neural Network, Linear Discriminant Analysis
Brown and Mues (2012)	Logistic Regression, Decision Tree, Neural Network, Gradient Boosting, Linear Discriminant Analysis, Quadratic Discriminant Analysis, k-Nearest Neighbors, Support Vector Machine, Random Forest
Xia et al. (2017)	Logistic Regression, Neural Network, Gradient Boosting, Support Vector Machine, Random Forest

Table 1. Data mining techniques used in previous empirical researches

Moreover, many benchmarking researches have been conducted to empirically compare the credit risk assessment techniques (Louzada et al., 2016). Throughout these papers, the empirical results indicate that machine learning and data mining techniques are suitable in tackling the good/bad distribution classification topics as credit risk evaluation (Brown and Mues, 2012).

METHODOLOGY

From the table 1, some popular data mining techniques preferred for loan default classification and forecasting, are logistic regression (in the researches of Arminger et al. (1997), Baesens et al (2003), Desai et al. (1996)), decision tree (Brown (2012), Xia et al. (2017), and West (2000)), neural network (Arminger et al. (1997), Altman (1994), Yobas et al. (2009)), gradient boosting (Brown and Mues, (2012), Xia et al. (2017)). To comparing the performance of classification techniques in credit scoring, listed four classifiers are utilized, ranked from traditional algorithms, such as logistic regression, decision tree, to more newly established ones, including neural network and extreme gradient boosting. The following sentences provide some brief explanations of data mining algorithms used in this paper.

Logistic regression: With the application of binary classification purpose of credit risk assessment, the logistic regression is suitable in categorizing whether a credit borrower is a good payer (non-defaulter) or is less likely to payback the loans (defaulter) (Brown and Mues, 2012). In a logistic regression model, a binary dependent variable, y, is defined as y=0 if the customer is a bad creditor, or y=1 if he would keep the regular payment normally. Here, we assume x is a column vector of independent variables and $\pi = \Pr(y = 1|\, x)$ is the response probability of the dependent variable. The logistic regression model then takes the form: $logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^T * x$, with α is the intercept parameter and β^T is row vector of the independent variable coefficients (Hosmer and Stanley, 2000).

Decision tree: a system of internal nodes, edges, and leaves that specify tests on individual input variables. The algorithm splits the entire training dataset into smaller subsets and produce individual trees (Xia et al., 2017), in which each of the observations is assigned to leaf node classes in the resulting segments. Although there are some algorithms specified for the decision tress, such as ID3 (interactive Dichotomiser), CART (classification and regression trees) (Gupta et al., 2017), the paper implements the popular decision tree algorithm C4.5, which constructs decision trees by applying Entropy (information gain) (Brown and Mues, 2012). The entropy of a sample D of classified observations is given by Entropy (D) =$\sum_{i=1}^{c} -p_i\, log_2(p_1)$, where p_i is the probability that an arbitrary tuple in D is assigned to the class values ranking from 0 to1 in the sample D. C4.5 algorithm determines the quality of normalized information gain (entropy difference) after choosing attributes for data splitting task. The attribute representing the highest normalized information gain is then used in the decision-making implementation, then a smaller subset is selected for calculation to be recurred.

Neural network (NN): a system based on input variables, which are interacted under linear or non-linear correlations. Besides, the network also includes one or more hidden computing layers, leading to output variables. According to West (2000), neural networks were credited as efforts to mimic the automatic learning ability of human being brain, which simulates sending electronic signals between enormous number of neurons and linkages. As a human brain network basically includes major

elements, including stimuli, neurons, and responses, a artificial neural network receives information from input variables (sources of stimuli), then constructs synapses in neurons (activation function implementation of neurons in the hidden layers), and finally produces output variables (responses) (Brown and Mues (2012)). During the transmission processes, each linkage between neurons is assigned a weight for training, and then output values of hidden nodes are calculated based on these weights and specified activation functions. Theoretically, neural networks are different regarding their basis structures, according to the layer number and the activation functions implemented.

Extreme gradient boosting: With the advancement in computing power and new algebraic algorithms, more complicated benchmark models, such as Gradient Boosting and its extreme variations, have been introduced to enhance the quality of classification tasks, and generate more economy business modeling. Particularly, extreme gradient boosting is a machine learning classification technique, in which an ensemble of weaker prediction models, typically decision trees, is used in a stage-wise boosting fashion. The technique generalizes the component classification decision trees by utilizing arbitrary differentiable loss function for model optimization. Under the pseudo-residual minimization at each iteration of the boosting, the loss function is minimized and help to improve the approximation accuracy and execution speed of model implementation (Friedman, 2001 and 2002).

In this paper, the dataset is divided into training and testing datasets by a proportion of 70/30 to implement the classification modelling. Next, precision and predictive accuracy testing (confusion matrix) will be implemented to determine the best performance model in predicting the default possibility of loan application. Finally, a Monte Carlo simulation model will be developed under several specified assumptions to assess the appropriateness of the model under the profitability criteria of banks. According to Morales et al. (2002), the Monte Carlo is a suitable approach in evaluating the matching between theoretical modeling fit and the real-life business context requirements.

DATA AND DATA PREPROCESSING

The dataset used in the project was retrieved from website www.creditriskanalytics.net with 5,960 observations represented in 13 attributes. The dataset indicates information about characteristics of home equity loans and can be summarized into 2 groups: customer information and loan-related information.

	Variable Name	Variable Type	Description
Dependent variable	BAD	Nominal	Loan default 1- Applicant defaulted on loan or seriously delinquent 0 -Applicant paid the loan
Customer information variables	JOB	Nominal	Customer's occupational categories
	YOJ	Discrete	Number of years at present job
	DEROG	Discrete	Number of major derogatory reports
	DELINQ	Discrete	Number of delinquent credit lines
	CLAGE	Discrete	Age of oldest credit line in months
	NINQ	Discrete	Number of recent the credit inquires
	CLNO	Discrete	Number of credit lines
	DEBTINC	Continuous	Debt-to-income ratio
Loan-related information variables	LOAN	Continuous	Amount of the loan requested
	MORTDUE	Continuous	Amount due on existing mortgage
	VALUE	Continuous	Value of current property
	REASON	Nominal	Motives for applying the loan

Table 2. Attribute Names and Descriptions

Data preprocessing

Firstly, the descriptive analysis is set to gain the first data understandings. Appendix 1 provides information about min, max, mean, and median of 10 numerical variables, and the numbers of observations in the same groups of categorical attributes of the credit application profiles. Additionally, for some numerical variables, which displays serious problem in interval-scaled inconsistency, a rescaling function (x-min(x))/(max(x)-min(x)) is utilized to convert these variables to unitless measures. Then, the problems of missing values and observation outliners were identified in the data cleaning stage after first data understanding tasks. In fact, some numerical attributes such as DEBTINC and DEROG display a remarkable percentage of missing data with 21.26% and 11.88%, respectively (appendix 1); hence, the paper uses the mean values of variable to replace the missing records. Moreover, to measure how independent variables interact, calculation for the correlation coefficients is also implemented; this task is significantly important before implement some traditional econometrics models such as logistic regression, which requires very strict assumptions in multi-collinearity. Based on the correlation matrix, the correlation degree between

independent variables are relatively small (the absolute values are less than 0.8), implying that there is not the potential of multi-collinearity and none of these attributed should be eliminated from the data mining models (refer the appendix 2).

RESULTS

Modeling benchmark

From the perspective of a bank manager, accurately identifying the BAD applications is likely more important for the bank's performance than the GOOD ones; when a BAD application is mistakenly classified as a good case, potentials of loss are significant (Morales et al., 2013). Hence, the precision and recall values are based on the BAD cases of the dataset. Among the 4 classification models, the XGboost outperforms over others in accuracy, precision, and recall values (regrading we are interested in the BAD credit applications). Hence, the extreme gradient boosting model could be considered as the most suitable classification model for the given dataset, and it could provide the highest prediction reliability for the credit risk.

In general, all the four model well perform in terms of classification application; their accuracy rates all higher than the preference level 75% if we intuitionally base on the random guess (4471/5960). In particular, the results and validating performance degrees between logistic regression, neural network and are relatively consistent; the accuracy, precision, and recall of logistics regression are 84.17%, 69.72%, and 29.2%, compared to those of the artificial neural network, 85.46%, 41.35%, and 38.94%, respectively. While the prediction performance could be quite similar between the first two models, decision tree can be considered as a more advanced approach, with significantly higher rate of accuracy (89.32%), and also the precision and recall rates, with 73.72% and 67.85%, in turn. However, by applying the XGBoost, remarkable improvement in accuracy rate could be experienced (9.4% higher than the worst model – logistic regression, and 4.25% higher compared to the 2nd-best model – decision tree). The recall rate of XGBoost (88.36%) is also a remarkable improvement compared to that of other classification models. This result is consistent with the studies of Brown and Mues (2012) and Xia et al. (2017), which also indicated that the gradient boosting technique was ranked top in terms of prediction performance when the authors compared the data mining models for credit scoring task in the banking sector.

	Logistics Regression		Neural Network		Decision Tree		XGBoost	
	BAD	GOOD	BAD	GOOD	BAD	GOOD	BAD	GOOD
BAD	99	43	132	53	230	82	258	81
GOOD	240	1406	207	1396	109	1367	34	1415
Accuracy	0.8417		0.8546		0.8932		0.9357	
Precision	0.6972		0.7135		0.7372		0.7611	
Recall	0.2920		0.3894		0.6785		0.8836	

Table 3. Confusion Matrix representing the accuracy of data mining methods based on the testing subset data in R

The simulation results:

Although the accuracy, recall and precision could serve as good criteria for classification model comparison in general, the specific context of the banking sector requires not only the highly accurate prediction, but also profitable models. Therefore, for profit-oriented organizations like banks, a Monte Carlo simulation could be utilized to make the comparison of certain properties between models (Morales et al., 2013), and to interpret how a model could be best fit for the profit goals of business.

The Monte Carlo simulation is based on some specified assumptions as following:

- Number of iterations: 50,000.
- Return rate: 20%.
- If a BAD credit record is mistakenly classified as GOOD by the model, the loss rate will be 100%.
- The bank finances the credit records based on the historical data of credit scoring.

The simulation models apply the prediction results from data mining models and measure the profit earned if the banks implement each model in their loan granting decisions. The profits are calculated based on the listed assumptions following the formula: Profit = $\sum(20\% * loan\ amount\ of\ good\ loan_i) - \sum loan\ amount\ of\ bad\ loan_k$.

The simulation results are evaluated based on the mean, median, min, max, and standard deviation values of profit earned from each model. Moreover, based on the distribution probability of the models resulted from 50,000 iterations, probability distribution graphs for each model are also created.

Model	Mean	Median	Min	Max	Standard Deviation
Logistics Regression	$444.70	$444.84	$367.13	$508.03	$16.36
Artificial Neural Network	$786.93	$787.02	$724.92	$858.24	$15.38
Decision tree	$1800.02	$1800.03	$1748.73	$1848.06	$12.08
XGBoost	$2233.46	$2233.47	$2186.63	$2279.36	$11.11

Table 4. Descriptive analysis of profit based on the simulation results

In short, the all the simulation models generated by four techniques are relatively reliable and display low rates of variation in distribution (the standard variable is relatively small compared to the mean values, ranging from 11.11 to 16.36). Regarding the profitability of models, the ranks of four models are: XGBoost (average profit = 2233.46) > Decision Tree (mean profit = 1800.02) > Neural Network (average profit = 786.93) > Logistic Regression (mean = 444.7), meaning the extreme gradient boosting is also the most profitable model for the credit risk classification operation. Besides, regarding the variation issue, the XGBoost model also represents a relatively low standard deviation in the profit probability distribution compared to those of other algorithms. After all, this simulation result is highly compatible with the performance criteria ranking above, implying that the XGboost is the most appropriate model in both terms accuracy performance and profitability consistency.

CONCLUSION

In conclusion, under the risk exposure, loss-cutting, and intensively competitive pressures in the banking sector, risk management has become the new focus of financial institutions; and discovering the unknown future behaviors, such as the loan cheating/ default, is placed a high priority. Through the analysis on credit scoring data mining models, the extreme gradient boosting performs as the most accurate, and most profitable technique in predicting those who express higher potential of not paying back the bank. By identifying the key factors which could be attributable for the loan default, the decision makers would develop proper credit policies applied to classify different borrower groups:

Firstly, the bank should be aware that some features in profile of home equity application really matter and could serve as good predictors for high potential of loan loss, implying the bank needs to be more careful when granting these credit applications. Moreover, in terms of new customer attracting and credit stimulation, when the decision makers can efficiently distinguish the customer groups who have low probability of loan default, bank managers can allocate more resources to attract them to increase their credit without the threat of non-payment, and exploit proper strategies to expose to these target customers. Nowadays, with the development in the inter-bank information systems, as well as the enhance in network analysis, the bank would not find it hard to collect data about the prospects. By filtering qualified cases by applying the data mining classification, the bank can encourage these customers to enlarge their credit build-up through bank's services with suitable business policies.

This study contributes by introducing the concepts of risk scoring and the applications of data mining classification models to address the related problems of credit granting decisions. Additionally, the extreme gradient boosting (XGBoost) was implemented besides traditional classification methods. The study also applies Monte Carlo simulation benchmarks as a more appropriate approach in model comparison.

Regarding the limitations of the study, because the dataset obtained from the website was basically a secondary data source, while some important attributes are under missing, such as demographic attributes of the customer, financial indicators of customer, some other irrelevant attributes of the dataset were not really useful and suitable for the data mining process (such as the REASON variable). Additionally, missing values (such as in the DEBTINC and DEROG variables) and outliers are also other serious problems of the data, which could adversely decrease the predictability of the classification models, and detrimentally influence the decision-making effectiveness. Lastly, with the limit in timing and computational power, only four classification models have been implemented with a small number of model benchmark criteria, like accuracy, precision, and recall; this, in some ways, can limit the benchmarking quality between models.

For the future improvements, more attempts could be allocated into the data collection stage; for example, collecting a larger dataset, basing on a more reliable database, and seeking more relevant attributes included. Moreover, with the development of more advanced data mining algorithms with stricter benchmark platforms, more appropriate techniques could be proposed to explain the loan default likelihood of the credit application for better results and better business strategies in the future.

APPENDICES

Variable	Min	Max	Mean	Median	Percent of missing values
LOAN	1100	89,900	16,300	18,608	0
MORTDUE	2,063	399,550	73,761	65,019	8.69
VALUE	800	855,909	101,776	89,236	1.88
YOJ	0	41	8.922	7	8.64
DEROG	0	10	0.2546	0	11.88
DELINQ	0	15	0.4494	0	9.73
CLAGE	0	1,168.2	179.8	173.5	5.17
NINQ	0	17	1.186	1.000	8.56
CLNO	0	71	21.3	20	3.72
DEBTINC	0.5245	203.3121	33.7799	34.8183	21.26

Variable	Number of records in each group	
BAD	Yes: 1189, No: 4771	0
REASON	DebtCon: 3928, HomeImp: 1780, Other: 252	0
JOB	Mgr: 767, Office: 948, Other: 2388, ProfExe: 1276, Sales: 109, Sel: 193, NA: 279	0

Appendix 1. Description statistics and missing value analysis

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
BAD	**1.00**										
LOAN	-0.08	**1.00**									
MORTDUE	-0.05	0.22	**1.00**								
VALUE	-0.03	0.33	0.58	**1.00**							
YOJ	-0.06	0.10	-0.08	0.01	**1.00**						
DEROG	0.26	0.00	-0.04	-0.04	-0.06	**1.00**					
DELINQ	0.34	-0.03	0.00	-0.01	0.04	0.18	**1.00**				
CLAGE	-0.17	0.09	0.13	0.17	0.19	-0.08	0.02	**1.00**			
NINQ	0.17	0.04	0.03	0.00	-0.07	0.16	0.06	-0.11	**1.00**		
CLNO	0.00	0.07	0.31	0.26	0.02	0.06	0.16	0.23	0.09	**1.00**	
DEBTINC	0.12	0.07	0.13	0.12	-0.05	0.01	0.03	-0.04	0.11	0.16	**1.00**

Appendix 2. Correlation matrix between attributes

REFERENCES

1. Altman, E. (1994). Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, 18 (3), 505-529.
2. Arminger, G, Enanche, D. & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discriminant, classification tree analysis, and feed forward networks. *Computational Statistics*, 12, 293-310.
3. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54 (6), 627-635.
4. Brown, I. & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39, 3446-3453.
5. Desai, V.S., Crook, J.N. & Overstreet, G.A. Jr. (1996). A comparison of neural networks ad linear scoring models in the credit union environment. *European Journal of Operational Research*, 95 (1), 24-37.
6. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29 (5), 1189-1232.
7. Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38 (4), 367-378.
8. Gupta, B., Rawat, A., Jain, A., Arora, A. & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163 (8), 15-19.
9. Hosmer, D.W., & Stanley, L. (2000). Applied logistics regression (2nd edition). *Chichester, New York: Wiley.*
10. Lee, T.S., Chiu, C.C, Chou, Y.C. & Lu, C.J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, 50, 1113-1130.
11. Louzada, F., Ara, A. & Fernandes, G.B. (2016). Classification methods applied to credit scoring: systematic review and overall comparison. *Surveys in operations research and management science*, 21, 111-134.
12. Morales, D., Pérez-Martin, A. & Vaca, M. (2013). Monte Carlo simulation stud of regression models used to estimate the credit banking risk in home equity loans. *WIT Transaction on Information and Communication Technologies*, 45, 141-153.
13. Thomas, L.C., Edelman, D., & Crook, J. (2002). Credit scoring and its applications in monographs on mathematical modeling and computation. *SIAM*, 2002.
14. West, D. (2000). Neural network credit scoring models. *Computer & Operations Research*, 27 (11-12), 1131-1152.
15. Xia, Y., Liu, C., Li, Y. & Liu, Nana. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78, 225-241.
16. Xu, S. & Qiu, M. (2008). A privacy preserved data mining framework for customer relationship management. *Journal of Relationship Marketing*, 7(3), 309-321.