Real-world credit scoring: a comparative study of statistical and artificial intelligent methods

Zhou Ying, Tabassum Habib and Guotai Chi*

School of Management and Economics,
Dalian University of Technology,
Dalian 116024, China
Email: zhouying@dlut.edu.cn
Email: tabassum_habib@yahoo.com
Email: chigt@dlut.edu.cn
*Corresponding author

Mohammad Shamsu Uddin

School of Management and Economics,
Dalian University of Technology,
Dalian 116024, China
Email: uddin@mail.dlut.edu.cn
and
Department of Business Administration,
Metropolitan University,
Sylhet, Bangladesh

Abstract: Credit scoring is an integral and crucial part of any lending process that any little development in it can reduce huge potential losses of financial organisations. The assessment of model performance varies because of different performance measures under a variety of circumstances on different nature of datasets. Therefore, this study employed six well-known classification approaches on six real-world credit datasets for comprehensive assessment by combining ten representative performance criterions. The experimental outcomes, statistical significance test and the estimated cost of prediction error confirm the marginal superiority of logistic regression (LR) and TreeNet over CART and MARS, being more robust compared to other two approaches LASSO and RF.

Keywords: credit scoring; performance measures; statistical method; artificial intelligence; AI.

Reference to this paper should be made as follows: Ying, Z., Habib, T., Chi, G. and Uddin, M.S. (2019) 'Real-world credit scoring: a comparative study of statistical and artificial intelligent methods', *Int. J. Knowledge Engineering and Data Mining*, Vol. 6, No. 1, pp.32–55.

Biographical notes: Zhou Ying is an Associate Professor of Accounting Investment and Supervisor of Master's candidates at the Dalian University of Technology, Dalian 116024, China. Her research interest includes asset-liability management, financial risk management, credit rating, etc.

Copyright © 2019 Inderscience Enterprises Ltd.

Tabassum Habib is a Master's candidate of Business Administration at the Dalian University of Technology, Dalian 116024, China. Her research interest includes credit scoring, financial risk management, working capital management, etc.

Guotai Chi is a Professor of Finance and Doctoral Adviser at the Dalian University of Technology, Dalian 116024, China. His research interest includes asset-liability management, financial risk management, credit rating, big data analysis, etc. He is a Visiting Professor at various universities in China and has successfully managed various national sponsored research projects and grants.

Mohammad Shamsu Uddin is an Investment Theory Doctor graduate student at the Dalian University of Technology, Dalian 116024, China. His research interest includes financial risk management, credit analysis, data mining, artificial intelligence and asset liability management, etc.

1 Introduction

Accepting loan application from prospective borrowers is the common foundation of bank business. The massive defaulting loss and intense contest involve financial intermediations to differentiate candidates profitably and adequately. Thus, the bank is supposed to settle on whether to lengthen credit at the time of potential customer selection. Customer data are mostly achieved from the application forms, client demographics, and numerous reports of earlier period borrowing and refunding accomplishment. Generally, the problem of credit scoring is distorted into a binary or multi-class classification problem. In another terminology, to construct decision support systems, classifications are developed with the credit data, thus supporting banks income to a decision regarding granting loans to particular submissions for a loan. For decades as an automatic assessment tool credit scoring has been used by lenders (Li et al., 2017). Credit scoring is playing a significant role in modern credit affairs such as customer selection, risk measurement, post-loan and after loan supervision, inclusive performance assessment and portfolio risk management (He et al., 2018). Especially for any financial organisation credit scoring is considered as an integral part of their lending process. Due to the world over turmoil financial condition, its importance is increasing day by day. Li et al. (2016) mention that, because of the financial crisis and increased capital requirement of the bank, in recent time the importance of credit scoring has increased. As a result of the reducing global economic stability, the demand for credit risk prediction has increased as well as it has become more critical (Khoraskani et al., 2017).

The analytical models functional to credit scoring can generally be alienated into two groups: statistical methods and artificial intelligence (AI) approaches. In credit scoring, the statistical technique is the first and most commonly used method for credit risk prediction. Chen et al. (2016) clarified that the rationale behind the statistical model is to find the optimal linear grouping of explanatory input variables, which can model, analyse, and predict enterprise default risk. To develop credit scoring model many researchers employed statistical model (Altman, 1968; Altman and Sabato, 2007; Banasik et al., 2001; Boyes et al., 1989; Durand, 1941; Ewert, 1968; Makowski, 1985; Myers and

Forgy, 1963; Orgler, 1970; Wiginton, 1980). With the recent progress of information and computational technologies, credit scoring model has been developed based on sophisticated intelligence approaches. In accounting and finance, credit risk modelling is one of the key areas in which artificial intelligent technologies have been applied effectively (Angelini et al., 2008; Arminger et al., 1997; Chatterjee and Barcun, 1970; Desai et al., 1996; Piramuthu, 1999; Tsai and Wu, 2008; West, 2000; Pisharody et al., 2015; Huang et al., 2015; Mishra and Srivastava, 2012; Siddiky et al., 2012). So many studies have highlighted modelling techniques that recommend a new algorithm to improve the accuracy of credit scoring.

According to related literature, for constructing credit scoring models the frequently used statistical technique is a logistic regression (LR). For the first time, Ohlson (1980) was utilised in the default prediction study. The core advantages of LR above other models are, it highlights the perspective of few restraining modelling hypotheses. The other benefits of LR are, it is not assumed the linearity, normality conditions, as well as, independence between independent variables in approach which leaves additional flexibility in functioning real-life application. The numerous study stated LR is a sound and robust statistical approach for credit risk prediction. Additional studies of models for predicting business credit risk using LR are evaluates and executed by Johnsen and Melicher (1994), Dimitras et al. (1996), Laitinen and Laitinen (2000), Altman and Sabato (2007), Kumar and Ravi (2007), Chen (2011b) and Nikolic et al. (2013). In the last decade, the widespread development of credit scoring models has been done by LR. Even with the existence of more sophisticated several classification models for credit scoring such as neural network (NN) (Derelioglu and Gurgen, 2011; Lee et al., 1996; Leshno and Spector, 1996), support vector machines (SVM) (Kim and Ahn, 2012), and case-based reasoning (Vukovic et al., 2012). The attractiveness and acceptance of LR have sustained typically because of its flexibility, functionality and theoretical dependability. In this study, we employ classification and regression trees (CART), least absolute shrinking and selection operator (LASSO), multivariate adaptive regression splines (MARS), Random forest (RF), TreeNet (TN) for credit risk prediction modelling. We utilised these models because of well acceptance, efficiency and frequently used in the previous literature of the credit scoring. Finally, the results of all approaches are compared with industry standard LR.

In addition, the existing literature on credit scoring stated that there is a variety of performance standards for credit prediction modelling, and these standards have been developed to asses unique things on different features. A few evidence have clarified that the prediction algorithm achieved the most excellent performance according to specified criterions on the specific dataset, and may not be the best technique for using different criterions on a different dataset. For example, Chen (2011a) evaluated SVM with a few traditional statistical techniques, using Taiwan Stock Exchange listed companies' data, and he established that the rankings of the approaches different on overall accuracy, precision, true positive rate, and true negative rate. The similar study was done by Tinoco and Wilson (2013) on some logit models with unequal groups of descriptive attributes using some performance measures like; ROC, GINI index and Kolmogorov-Smirnov statistics (KS) as a measure of discriminatory power and drawn the same ending. In addition, Zhong et al. (2014) got the same result that SVM on rating distribution and NN approaches outperform SVM on reliability, they used SVM and MLP with two other algorithms for credit rating analysis. Recently, Moula et al. (2017) mentioned that no particular performance measure might achieve the best-ranked index for the every study

area. However, there have been some pieces of indication evaluating the performance of assessment techniques for binary and multi-class scoring problems. For instance, Sokolova and Lapalme (2009) and Cuadros-Rodríguez et al. (2016) used 24 and 21 different performance metrics to review the performances of the classifiers.

This study, consequently, argues that the performance evaluations in use presently do not entirely reflect the demand of the respective sectors. We try to make a composite mixture of performance measures from different fields to accomplish the complete evaluation process. As per example, discriminant power (Blakeley and Oddone, 1995) and likelihood ratios (Biggerstaff, 2000) are usually used in the field of clinical diagnosis. In our study the core objective behind to used multiple traditional and new measures are to give a comprehensive assessment of prediction algorithms on several credit datasets. On the other hand, most of the existing study used a few performance measures, which are not sufficient to describe model efficiency entirely. This study also tries to evaluate the capacity of LR with other sophisticated new algorithms like CART, MARS, and RF, such as a few previous studies (Lessmann et al., 2015; Ala'raj and Abbod, 2016) clarified that, although there have been developed new technology, LR is still considered as an industry standard.

Therefore, by utilising SME and farmer credit with four real-world public datasets, this study contributes to the growing literature by assembling composite performance metrics for comprehensive assessment of prediction classifiers. This process can be evaluated all the model's performance entirely on a different perspective. However, the more specific objective of this study is to construct a reliable credit scoring model by using statistical and AI methods. Six well-known classifiers are used to develop models, namely CART, LASSO, LR, RF, MARS and TN. The predictive performance is evaluated against ten performance measures: accuracy, AUC, type I error, type II error, F-score, kappa, positive likelihood ratio, negative likelihood ratio, G-mean, DP. Finally, a statistical significance test and the cost of prediction error have been done for each single models to compare and find out the most reliable model.

The structure of the rest of the paper is as follows: Section 2 provides an overview of the introduction of statistical and AI techniques used in this study. Section 3 explains the experimental set-up carried out for this study. Section 4 presents the empirical results from comparing the models. Section 5 summarises the paper and future research trends.

2 Models

2.1 Classification and regression trees

The CART has proven to be robust and effective decision tree-based method for classification problem, introduced by Breiman et al. (1984). The CART is based on landmark mathematical theory and is considered the essential tools in modern data mining. It has shown superiority regarding accuracy, performance, feature set, build in automation and ease of use. The CART is competent to distinguish nonlinear connections among input variables, which radically enhances the kinds of associations that can be captured and some independent variables that can be used. In addition, CART model produces easily interpretable decision rules.

2.2 LASSO

LASSO, a regression analysis technique, which is related to statistics and machine learning that can be used for variable selection and regularisation. In 1996 Tibshirani introduced it based on Leo Breiman's non-negative garrote to improve prediction accuracy and interpretability of the regression model by altering the model fitting process. Before LASSO stepwise selection technique was broadly used for covariates selection, it can perform better in specific cases, and it can make prediction error inferior. On the other hand, by shrinking the large regression coefficient to reduce over-fitting ride regression improves prediction error but it does not have the interpretable model capacity and does not perform covariance selection. By focusing the summation of the absolute value of the regression coefficient on being less than a fixed value, LASSO can achieve both of these objectives.

2.3 Logistic regression

LR is a regression model for the categorical dependent variable. Cox developed LR in 1958. LR is handy for credit classification as it is used to model a binary outcome variable, usually represented by 0 or 1 (non-default and default loan). LR is a predictive analysis which is suitable to conduct when the dependent variable is dichotomous and to solve classification and regression problem. Until now, this has been regarded as the industry standard for credit scoring model development (Lessmann et al., 2015). The principal reason behind the continuous usage of LR over the other methods of estimation is that it provides an appropriate balance of accuracy, efficiency, and interpretability of results (Crone and Finlay, 2012). LR model identifies the probability of a default event to occur is expressed as (Atiya and Parlos, 2000):

$$\mathrm{Log}\left[p(1-p/p)\right] = \beta_{0+}\beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{1}$$

where p is the probability of default, β_0 is intercepted, β_i stands for coefficient of independent variables X_i (i = 1 …. 2 ……. n) and log [p(1 – p / p)] is the dependent variable.

2.4 Multivariate adaptive regression splines

MARS is a type of non-parametric and nonlinear regression analysis initiated by Friedman in 1991, which models the multifaceted association among independent input variables and dependent target variables. By piecing together a succession of straight lines, MARS constructs its model with each of its allowed slope. This allows tracing out any pattern detected in the data. Automatically MARS performs the variable selection, variable transformation, interaction detection and self-testing with at high speed. MARS is constructed in the structure of:

$$y = c_0 + \sum_{i=1}^{k} c_i Bi(x) \tag{2}$$

where c_o is a constant coefficient, Bi(X) is the basis function and c_i is a coefficient of the basis function. In basis function, it gets various forms of independent variables' connections; the familiar functions used are the hinge functions that are used to discover variables, which are chosen as knots, therefore the function obtains the following form (Friedman, 1991).

$$\max(0, X - c) \tag{3}$$

Or

$$\max(0, c - X) \tag{4}$$

where c is a constant, threshold or knot location, X is the independent variable. The objective at the back the basis function is to convert the independent variables X into new variables (e.g., X′). According to the equation (3) moreover, (4), X will take the value of X if X is larger than c and it will take the value of zero if the value of X is less than c (Briand et al., 2004). MARS re-evaluate the model following all terms concerning the variable are detached to be appraised and computes the lessening in the model's error, and then every variable are classified according to their supremacy on the outcome of the model; the best MARS model is standing on lowest the generalised cross-validation (GCV) measure (Briand et al., 2004). For more details about the MARS model, please refer to (Friedman, 1991; Hastie et al., 2005).

2.5 Random forest

For classification and regression, RF is considered as an ensemble learning method and advanced techniques of DTs, as proposed by Breiman (2001). RF produces numerous classification trees. To categorise a novel object from an input vector, put the input vector down on each one of the trees in the forest. Each tree gives a classification; it considers votes for that class. According to most votes, forest chooses the classification. Among the current algorithms, it is superior in accuracy and efficient for large databases, missing data and unbalances datasets.

2.6 TreeNet

Friedman (1999a, 1999b) introduced the TN, which is the most flexible and influential data-mining tool. 'Ultra-slow learning' in which layer of information is systematically peeled off to expose structure in data use by TN. For regression and classification is reveal outstanding performance. The algorithm classically produces thousands of small decision construct in a chronological error-correcting method to congregate to an authentic model. TN is not required to time-overwhelming data training; the accusation of missing values as well as it is not susceptible to data errors. It has an interaction detection system, which facilitates to improve model performance and helps in the detection of critical new divisions and earlier unrecognised patterns.

3 Experimental design

3.1 Real-world credit datasets

We highlight six real-world credit databases to authenticate the effectiveness and feasibility of planned credit prediction classifiers. The Australian, German, Japanese and Taiwan are four well known, broadly accepted, accessed merely and publicly available at the UCI marching-learning repository; but the processed database is collected from Chi et al. (2017). Additionally, two project datasets, grounded on a historical loan dataset related with SME and farmer credit, was collected from one leading public commercial bank of China. These datasets contain customers data from 28 major cities of China. The datasets encompass an instance of non-default and default customers with a double objective variable, exemplified by a set of risk drivers, which imprison the information from the customer's application, form; personal, financial, non-financial and macroeconomics. A synopsis of the six datasets is offered in Table 1.

Table 1 Description of databases used in the experiment

Databases	*Total cases*	*Non-default/default cases*	*No. of attributes*
Australian credit	690	307/383	14
Chinese SME credit	3,111	3,040/71	81
Chinese farmer credit	2,036	2,012/24	44
Japanese credit	690	307/383	15
German credit	1,000	700/300	20
Taiwan credit	30,000	23,364/6,636	23

In standard, any training set with jagged allocation between the two classes can be considered as imbalanced. Though, sample ratios 1:5 (minority samples: majority samples) or upper have generally been judged in the trail as insignificant datasets (He and Garcia, 2009). In our study, Australian and Japanese datasets are not imbalanced, German credit datasets are slightly imbalanced, and Chines SME credit, Chinese farmer credit, and Taiwan credit are mostly imbalanced. The experimental databases, consequently, are the excellent combination of balance and imbalance example even though data sampling approach is away from the range of the existing study.

3.2 Performance evaluation

Ten accuracy measures are employed to assess the experimentation of six classifiers over six different datasets, which are derived from a 2 × 2 confusion matrix as that given in Table 2, where every unit comprises the figure of non-default/default prediction. The assessment measures are illustrated correspondingly as the following:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{TN} + \text{FP}) \tag{5}$$

$$\text{AUC} = (1/2)(\text{Sensitivity} + \text{Specificity}) \tag{6}$$

$$\text{Type I error} = \text{FN}/(\text{TP} + \text{FN}) \tag{7}$$

$$\text{Type II error} = \text{FP}/(\text{FP} + \text{TN}) \tag{8}$$

$$\text{F-score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \tag{9}$$

$$\text{Kappa} = (\text{Total accuracy} - \text{Random accuracy}) / (1 - \text{Random accuracy}) \tag{10}$$

$$\text{Positive likelihood ratio (P+)} = \text{Sensitivity} / (1 - \text{Specificity}) \tag{11}$$

$$\text{Negative likelihood ratio (P−)} = (1 - \text{Sensitivity}) / \text{Specificity} \tag{12}$$

$$\text{G-mean} = \sqrt{\text{Sensitivity} * \text{Specificity}} \tag{13}$$

$$\text{DP} = \sqrt{3} / \pi (\log X + \log Y) \tag{14}$$

TP, TN, FP, and FN correspond to the figure of true positive (non-default), true negative (default), false positive, false negative correspondingly. The accuracy (ACC) in equation (5) is one of the most frequent assessment procedures, which evaluates the overall efficiency of the classifier with all possible outcomes. Through ACC cannot be considered a single standard since it cannot differentiate good and bad applicants.

Table 2 Confusion matrix

Actual	*Predicted*		
	Good	*Bad*	
Good	TP	FP	TP + FP
Bad	FN	TN	FN + TN
	TP + FN	FP + TN	TP + FP + FN + TN

Therefore, the area under the curve (AUC) in equation (6) is an additional discrimination capacity measure depending on the receiver operating characteristics (ROC) curve, without any prior information about the error costs, AUC is used to estimate model performance (Hand, 2009). Type I and type II error rates in equations (7) and (8) are also employed as performance measures. Type I error appears when good applicants are misclassified as bad ones, while type II occurs when bad applicants are misclassified as good. The misclassification cost related to type I and type II are not the same. The F-score in equation (9) which is accounting for two extents of the predictive accuracy is the composite evaluation of recall and precision. Cohen's kappa (kappa) in equation (10), consider the accuracy that would be produced merely by chance. Kappa considers values from –1 to +1, with a value of 0 meaning there is no conformity among the real and classified classes. A value of 1 specifies perfect concordance of the model prediction and actual class and a value of –1 indicates a total discrepancy between forecast and actual. Positive (p+) and negative (p–) likelihood ratios, equations (11) and (12), correspondingly, are well known with clinical sector those are used to choose proper diagnostic investigation and thus are helpful to contrast two classifiers. The employment of positive and negative likelihood ratios, rather than sensitivity and specificity, a measure of credit prediction capacity, has some recompense (Schaefer and Strebulaev, 2008). Superior values of P+ and inferior value of P– indicate improved performance on non-default and default classes, respectively, and consequently favoured. The geometric mean (G-mean) in equation (13) it is a metric of true positive and true negative rate. G-mean can get a high value when sensitivity and specificity are large enough. Low G-mean is an indicator of poor performance. Similarly, the discriminate

power (DP) in equation (14) is another global performance measure that summarises sensitivity and specificity. It has been considered as a superior measure of diagnostic discrimination and widely used in epidemiological studies, very rarely used in credit scoring. Where X = sensitivity / (1-sensitivity) and Y = specificity / (1-specificity). The DP evaluates how well classifiers differentiate between the positive and negative cases. The classifier is considered as a poor classifier if DP < 1, limited if DP < 2, fair if DP < 3 and good in other cases.

3.3 Statistical significance test

According to García et al. (2015), as the different models employed various splitting methods, it is not sufficient to authenticate one model gets results better than other. To assess performance thoroughly, it would appear appropriate to use some hypothesis testing to emphasise that the experimental variations in returns are statistically significant. A statistical test can be parametric and non-parametric (Demšar, 2006). As parametric tests theoretically unsuitable and statistically hazardous, Demšar (2006) recommended that using non-parametric tests is preferable to parametric tests since they do not suppose normality of the data or homogeneity of the variance. We use Friedman's test to rank all the techniques, which is a non-parametric test. Friedman's (1940) analysis ranks the classifiers for each dataset independently. The rank 1 for best ranking classifier rank 2 for second best and so on. Under the null hypothesis of Friedman, the test is that every classifier from this group performs identically and all distinction is merely random fluctuations. The Friedman statistics x_F^2 is distributed according to x_F^2 with K–1 degrees of freedom when N (number of datasets) and K (number of classifiers). We used a post-hoc test Holm to compare all the methods when the null hypothesis is rejected. This test is various evaluation measures to contrast all techniques that can be used. Holm's test is a step-up measure that serially tests the hypothesis prearranged by their significance. Each p-value p_i ($p_0 \leq p1 \leq \ldots. \leq Pnc - 1$) is compared with α (NC – i), with the number of pairwise comparisons. All the linger hypothesis are retained as well when a particular null hypothesis cannot be rejected. In our experiments, on all the test carried out, the level of significance has been $\alpha = 0.1$.

4 Results and discussion

The vital pace before constructing the model is to organise the data for training, initially; a test is finished for any missing variables, which are substituted through an assertion approach by replacing the missing variables by the average or mean value of the respective variables.

In this paper, the six algorithms compared, these are a CART, MARS, LASSO, LR, RF, and TN. The k-fold cross-validation (CV) was applied concerning the experimental setting and data splitting method, i.e., the real dataset being divided into k-folds or subgroup of the same dimensions where each split has to be trained and tested. Therefore, the outcome is calculated by taking the average of all folds/subgroup that has been tested. We repeated 50 times five-fold CV was applied for each dataset, as (He and Garcia, 2009; Jones et al., 2015; Ala'raj and Abbod, 2016) recommended that 5–10 folds can be a reasonable alternative with datasets of various sizes.

However, the model choice and parameter determination is a very significant part of the performance of the classifiers. In case of the CART, the initial phase requires growing the tree by a recursive division technique and partition points through GINI. It also utilises the lowest cost complexity CV as a pruning algorithm, and 'one standard error rule' was adjusted to choose the optimal tree. For the LR model, 0.50 cut off point has been applied. For all others, classifiers provided by Salford Predictive Modeler 8.2 were used with their default configurations.

Our purpose in this empirical assessment is to examine the performance of six algorithms for comparison purposes and to find out which one is the best. To validate our approach and to reach a reliable conclusion, Tables 3–8 review the performance indicator measures of the six classifiers. The top algorithm for every dataset is highlighted using bold fonts; the second best is market with italic and bold fonts.

Table 3 Classifier results for all classifiers for the Chinese SME dataset

Name	*Accuracy*	*AUC*	*Type I error*	*Type II error*	*F-score*	*Kappa*	*P+*	*P–*	*G-mean*	*DP*
CART	0.9402	0.6245	0.0042	0.7468	0.9686	0.3660	1.3335	0.0165	0.5022	1.0522
LASSO	0.7891	0.5426	0.0033	0.9114	0.8794	0.1248	1.0936	0.0376	0.2972	2.2703
LR	***0.9431***	**0.6421**	**0.0003**	**0.7154**	***0.9700***	**0.4211**	**1.3972**	**0.0012**	**0.5333**	**7.7983**
MARS	0.8351	0.5539	0.0031	0.8891	0.9081	0.1632	1.1212	0.0284	0.3325	***5.6712***
RF	0.0948	0.5101	0.0044	0.9754	0.1373	0.0030	1.0207	0.1807	0.1565	0.4146
TreeNet	**0.9460**	***0.6383***	***0.0028***	***0.7207***	**0.9717**	***0.4043***	***1.3837***	***0.0099***	***0.5277***	1.1824

Starting with the Chinese SME dataset in Table 3 from experimental results, we find that, TN and LR is a close competitor in all of the performance measures. Regarding accuracy, TN has the highest prediction accuracy, the score is 94.60%, and the close opponent is LR with 94.31% score. In the case of F-score TN also shows the highest prediction accuracy with 97.17% score. In the other performance measures, like AUC (63.83%), type I error (0.28%), type II error (72.07%) kappa (40.43%), P+ (1.38), P– (0.0099), G-mean (52.77%) and DP (1.1824) results confirm that TN has second highest prediction accuracy. On the other hand, LR shows supremacy in all of the performance measures with following scores: AUC (64.21%), type I error (0.03%), type II error (71.54%) kappa (42.11%), P+ (1.39), P– (0.0012), G-mean (53.33%) and DP (7.7983). For the other classifiers it shows moderate performances except for RF, throw RF is one of the top classifiers, but it shows very miserable performances for this dataset.

Table 4 Classifier results for all classifiers for the Chinese farmer dataset

Name	*Accuracy*	*AUC*	*Type I error*	*Type II error*	*F-score*	*Kappa*	*P+*	*P–*	*G-mean*	*DP*
CART	0.8497	***0.5214***	***0.0052***	0.9519	0.9181	0.0689	***1.0450***	***0.1086***	0.2187	***0.5422***
LASSO	0.0157	0.4504	0.1111	0.9882	0.0079	–0.0009	0.8995	9.3843	0.1026	–0.5615
LR	***0.8669***	**0.5319**	**0.0029**	**0.0000**	***0.9280***	**0.1035**	**1.0684**	**0.0428**	**0.2578**	**0.7702**
MARS	0.8438	0.5188	0.0058	**0.0000**	0.9147	0.0603	1.0394	0.1342	0.2079	0.4902
RF	0.0123	0.2559	0.5000	0.9882	0.0010	–0.0010	0.5060	42.3958	0.0768	–1.0603
TreeNet	**0.8703**	0.5213	0.0062	0.9511	**0.9302**	***0.0695***	1.0448	0.1272	***0.2204***	0.5043

Looking at the Chinese farmer dataset in Table 4, LR shows the highest prediction accuracy in seven performance measures out of ten. These are: AUC (53.19%) type I error (0.29%), type II error (0.00%), P+ (1.0684), P– (0.0428), G-mean (25.78%) and DP (0.7702). In the accuracy (87.03%), F-score (93.02%) and kappa (6.95%) TN shows top performance. Beside this CART has been emerged as a second classifier in some of the performance measures like AUC (52.14%), type I error (0.52%), kappa (6.89%), P+ (0.1086), P– (0.1086) and DP (0.5422). LASSO and RF have shown miserable performance.

Table 5 Classifier results for all classifiers for the Australian dataset

Name	*Accuracy*	*AUC*	*Type I error*	*Type II error*	*F-score*	*Kappa*	*P+*	*P–*	*G-mean*	*DP*
CART	0.8551	0.8584	0.2133	0.0699	0.8503	0.7116	11.2532	0.2293	0.8554	0.9322
LASSO	0.8551	0.8579	0.2117	0.0725	0.8499	0.7114	10.8720	0.2282	0.8551	0.9251
LR	**0.9319**	**0.9301**	**0.0988**	**0.0410**	**0.9255**	**0.8628**	**21.9901**	**0.1030**	**0.9297**	**1.2843**
MARS	***0.9087***	***0.9077***	***0.1391***	***0.0455***	***0.9023***	***0.8170***	***18.9409***	***0.1457***	***0.9065***	***1.1655***
RF	0.7478	0.7923	0.3525	0.0628	0.7705	0.5121	10.3160	0.3761	0.7790	0.7929
TreeNet	0.8464	0.8464	0.2087	0.0986	0.8374	0.6928	8.0294	0.2315	0.8446	0.8491

In the Australian dataset, as the experimental results demonstrated in Table 5, LR prediction model has shown significantly better in all criterions. This can clearly be seen from accuracy (93.19%), AUC (93.01%), type I error (09.88%), type II error (04.10%), F-score (92.55%), kappa (86.28%), P+ (21.9901), P– (0.1030), G-mean (92.97%), DP (1.2843). MARS also has reveal better performance in all of the cases as a second-best classifier as accuracy (90.87%), AUC (90.77%), type I error (13.91%), type II error (04.55%), F-score (90.23%), kappa (81.70%), P+ (18.9409), P– (0.1457), G-mean (90.65%), DP (1.1655). All other classifiers have revealed better results but have a meaningful distinction with the finest model LR.

Table 6 Classifier results for all classifiers for the German dataset

Name	*Accuracy*	*AUC*	*Type I error*	*Type II error*	*F-score*	*Kappa*	*P+*	*P–*	*G-mean*	*DP*
CART	0.6970	0.6650	0.1650	0.5049	0.7660	0.3445	1.6537	0.3333	0.6430	0.3836
LASSO	***0.7190***	***0.6912***	0.1392	***0.4785***	***0.7810***	***0.3983***	***1.7991***	0.2669	***0.6700***	0.4569
LR	**0.7430**	**0.7113**	0.1300	**0.4475**	**0.8025**	**0.4411**	**1.9441**	***0.2353***	**0.6933**	0.5056
MARS	0.7030	0.6909	***0.1203***	0.4979	0.7591	0.3903	1.7670	0.2396	0.6646	0.4784
RF	0.5650	0.6564	**0.0870**	0.6003	0.5748	0.2389	1.5210	**0.2176**	0.6041	***1.3102***
TreeNet	0.7140	0.6911	0.1327	0.4851	0.7741	0.3965	1.7879	0.2578	0.6683	**2.1459**

For the German dataset is shown in Table 6, LASSO exceptionally has demonstrated better performance in accuracy (71.90%), AUC (69.12%), type II error (47.85%), F-score (78.10%), kappa (39.83%), P+ (1.7991) and G mean (67.005) as second-best classifiers. The more consistent classifiers LR also express superiority in case of accuracy (74.30%), AUC (71.13%), type II error (44.75%), F-score (80.25%), kappa (44.11%), P+ (1.9441) and G-mean (69.33%). RF exposes supremacy in the type I error (8.70%) and P– (0.2176). Finally, TN confirms the best result in the perspective of DP (2.1459).

The outcomes for the Japanese dataset in Table 7 reveal the dominance of the MARS approach over other classifiers in maximum performance criterions. The performances of MARS reaches in accuracy (86.67%), AUC (86.95%), type II error (6.04%), F-score (86.19%), kappa (73.45%), P+ (13.2308), G-mean (86.67%) and DP (0.9881). In the other two cases; type I error and P– TN express supremacy with the score of 17.96% and 0.1980. RF and all other classifiers also express modest performances in the entire performance criterion.

Table 7 Classifier results for all classifiers for the Japanese dataset

Name	*Accuracy*	*AUC*	*Type I error*	*Type II error*	*F-score*	*Kappa*	*P+*	*P–*	*G-mean*	*DP*
CART	0.8551	0.8584	0.2133	0.0699	0.8503	0.7116	11.2533	0.2293	0.8554	0.9322
LASSO	0.8580	0.8608	0.2089	***0.0695***	0.8529	0.7172	***11.3848***	0.2245	0.8580	0.9401
LR	0.8623	0.8632	0.1971	0.0765	0.8554	0.7251	10.4989	0.2135	0.8611	0.9327
MARS	**0.8667**	**0.8695**	0.2006	**0.0604**	**0.8619**	**0.7345**	**13.2308**	0.2135	**0.8667**	**0.9881**
RF	***0.8652***	***0.8657***	***0.1925***	0.0760	***0.8580***	***0.7307***	10.6214	***0.2084***	***0.8638***	***0.9413***
TreeNet	***0.8652***	0.8638	**0.1796**	0.0927	0.8549	0.7295	8.8499	**0.1980**	0.8627	0.9099

Table 8 Classifier results for all classifiers for the Taiwan dataset

Name	*Accuracy*	*AUC*	*Type I error*	*Type II error*	*F-score*	*Kappa*	*P+*	*P–*	*G-mean*	*DP*
CART	0.7516	0.6681	0.1191	0.5447	0.8316	**0.3641**	1.6172	0.2615	0.6333	***0.4362***
LASSO	0.6848	0.6249	0.1264	0.6239	0.7747	0.2717	1.4003	0.3361	0.5732	0.3417
LR	0.6828	0.6243	0.1260	0.6254	0.7727	0.2703	1.3976	0.3363	0.5722	0.3411
MARS	***0.7575***	***0.6692***	0.1245	***0.5372***	**0.8376**	0.3629	***1.6297***	0.2689	***0.6365***	0.4314
RF	0.4442	0.6013	**0.0713**	0.7261	0.4650	0.1230	1.2791	***0.2602***	0.5044	0.3813
TreeNet	**0.7585**	**0.6764**	***0.1138***	**0.5333**	***0.8362***	**0.3823**	**1.6617**	**0.2438**	**0.6431**	**0.4595**

Finally, at the Taiwan dataset, in Table 8, TN has exposed superior performances in seven measures out of ten. TN has the highest accuracy (75.85%), followed by MARS (75.75%) and CART (75.16%). The other top result-oriented measures by TN are AUC (67.64%), type II error (53.33%), kappa (38.23%), P+ (1.6617), P– (0.2438), G-mean (64.31%), and DP (0.4595). MARS also secures first position in F-score (83.76%) and a second position in following aspects, accuracy (75.75%), AUC (66.92%), type II error (53.72%), P+ (1.6297), G-mean (63.65%). The other classifiers also demonstrate better performance except for RF.

The mean and standard deviation of all performance measures is to provide quantitative experimental results and supplementary precise, are correspondingly calculated and listed in Table 9, in which the most favourable outcome for the row is in bold and the second best is market with italic and bold fonts.

For the mean and standard deviation combined with all credit dataset, the experimental results presented in Table 9 reveal that in all performance criterions, except a few, LR has shown an excellent prediction capability. On the other hand, TN also proved as a consistent classifier by outstanding results with LR. Beside this two, MARS and CART also display reasonable performance, but RF and LASSO failed to establish

themselves as a reliable classifier in the present study even though RF is considered as a new and dependable approach for credit classification.

Table 9 The mean and standard deviation of model evaluation measures

Name		*CART*	*LASSO*	*LR*	*MARS*	*RF*	*TreeNet*
Accuracy	Mean	0.8248	0.6536	**0.8383**	0.8191	0.4549	***0.8334***
	Std.	0.0866	0.3203	0.1043	**0.0754**	0.3441	***0.0837***
AUC	Mean	0.6993	0.6713	**0.7171**	***0.7163***	0.6136	0.7062
	Std.	***0.1342***	0.1665	0.1518	0.1466	0.2173	**0.1299**
Type I error	Mean	0.1200	0.1334	**0.0925**	***0.0989***	0.2012	0.1073
	Std.	0.0959	**0.0766**	***0.0775***	0.0786	0.1902	0.0864
Type II error	Mean	0.48135	0.5240	**0.3176**	***0.3383***	0.5715	0.4802
	Std.	0.3564	0.3970	**0.3179**	0.3593	0.4161	***0.3401***
F-score	Mean	0.8642	0.6910	**0.8757**	0.8640	0.4678	***0.8674***
	Std.	***0.0706***	0.3373	0.0781	0.0595	0.3413	0.0715
Kappa	Mean	0.4278	0.3704	**0.4706**	0.4302	0.2678	***0.4458***
	Std.	***0.2469***	0.2984	0.2817	0.3163	0.2960	**0.2414**
G-mean	Mean	0.6180	0.5594	**0.6412**	0.6025	0.4974	***0.6278***
	Std.	***0.2393***	0.3054	0.2440	0.2814	0.3219	***0.2367***
P+	Mean	4.6927	4.5749	**6.3827**	***6.2882***	4.2107	3.7929
	Std.	5.0866	5.0880	8.4638	7.0806	***4.8601***	**3.6177**
P–	Mean	0.1964	1.7463	**0.1554**	***0.1717***	7.2731	0.1780
	Std.	0.1142	3.7432	0.1278	**0.0877**	17.2067	***0.0947***
DP	Mean	0.7131	0.7287	**1.9387**	***1.5375***	0.4633	1.0085
	Std.	**0.2917**	0.9325	2.8895	2.0479	0.8227	***0.6188***

4.1 Significance test results

According to García et al. (2015), as the different approach utilised unlike splitting methods, it is not sufficient to validate one model achieves results more significant to other. To evaluate performance thoroughly, it would appear appropriate to use a few hypotheses testing to highlight that the investigational differences in results are statistically significant.

In this research, we have measured the value of Friedman's ranks for all measures. Those values express the performance and make rank for of all models. Figures 1–10 show average rank of all approaches over six different datasets on ten different performances measures. For space constraints, we have put only two (accuracy, AUC) statistical significance test results here. With $\alpha = 0.1$ significance level, it has shown TN is the best model according to accuracy and LR is the best model from the perspective of AUC. The post-hoc Holm test in Table 10 has expressed the level of differences with the best model to other models. If the P-value of the test is lower than 10%, then the null hypothesis is rejected. Therefore, TN is significantly better than corresponding models

regarding accuracy. It has a significant statistical difference between RF and LASSO but not with other models. In the case of AUC with 10% significant, the result indicates that no significant difference with CART, MARS, and TN and has a significant difference with RF and LASSO.

Table 10 The result of statistical significance test (Friedman and Holm test)

Accuracy			*AUC*		
Method	*P-value*	*Hypothesis ($\alpha = 0.1$)*	*Method*	*P-value*	*Hypothesis ($\alpha = 0.1$)*
TreeNet vs. RF	0.0034	Rejected	LR vs. RF	0.0043	Rejected
TreeNet vs. LASSO	0.0896	Rejected	LR vs. LASSO	0.0760	Rejected
TreeNet vs. CART	0.1228	Not rejected	LR vs. CART	0.1468	Not rejected
TreeNet vs. MARS	0.5892	Not rejected	LR vs. TreeNet	0.4875	Not rejected
TreeNet vs. LR	0.8170	Not rejected	LR vs. MARS	0.8774	Not rejected

Figure 1 Average rank of overall accuracy (see online version for colours)

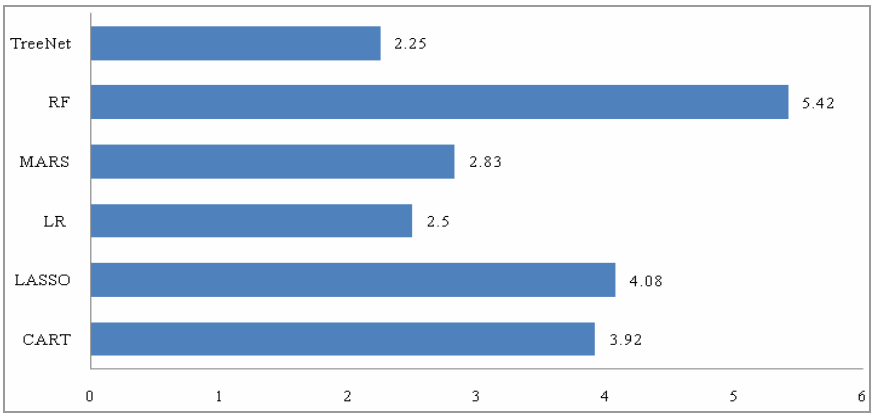


Figure 2 Average rank of AUC (see online version for colours)

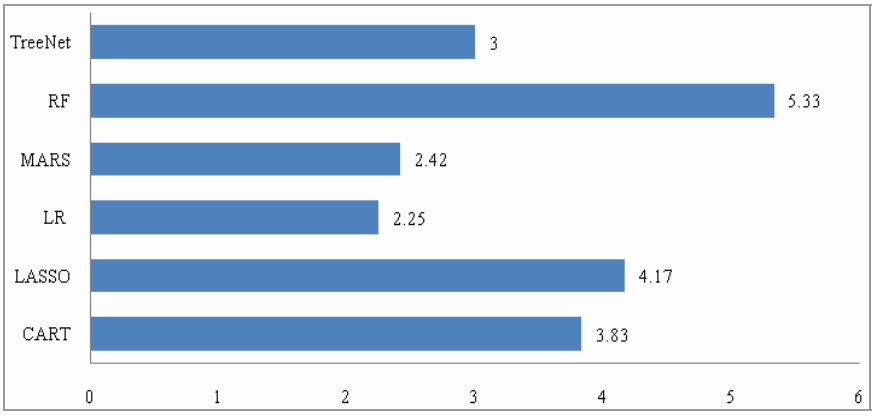


Figure 3 Average rank of type I error (see online version for colours)

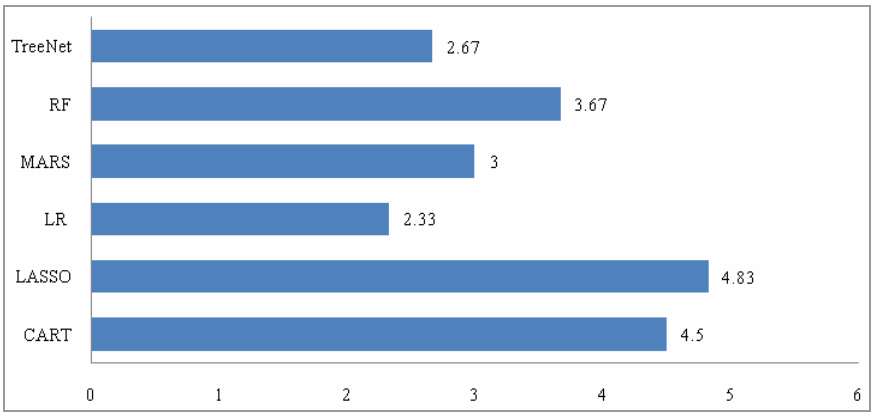


Figure 4 Average rank of type II error (see online version for colours)

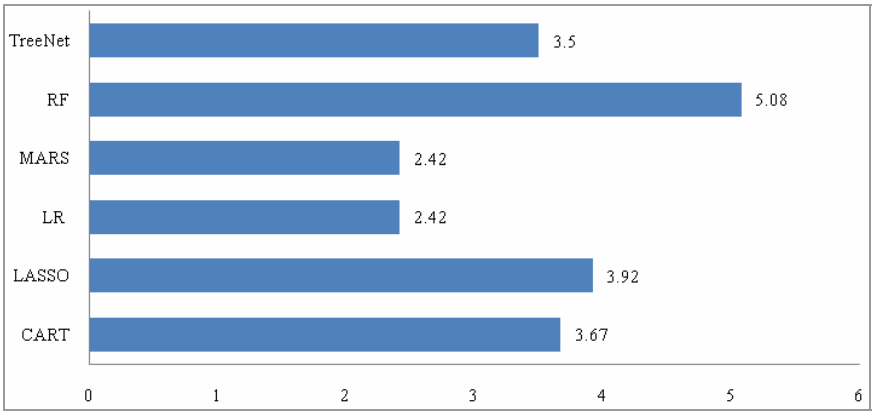


Figure 5 Average rank of F-score (see online version for colours)

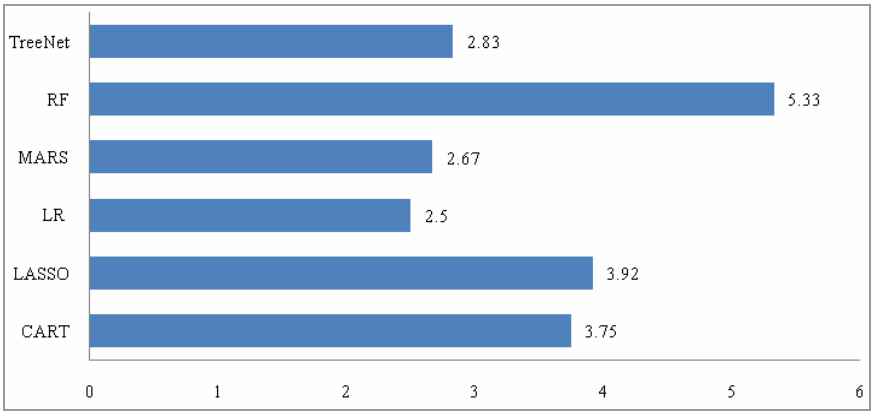


Figure 6 Average rank of kappa (see online version for colours)

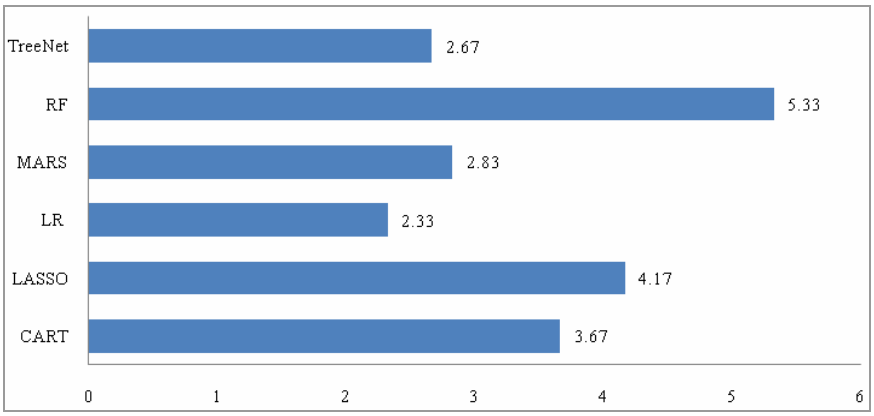


Figure 7 Average rank of G-mean (see online version for colours)

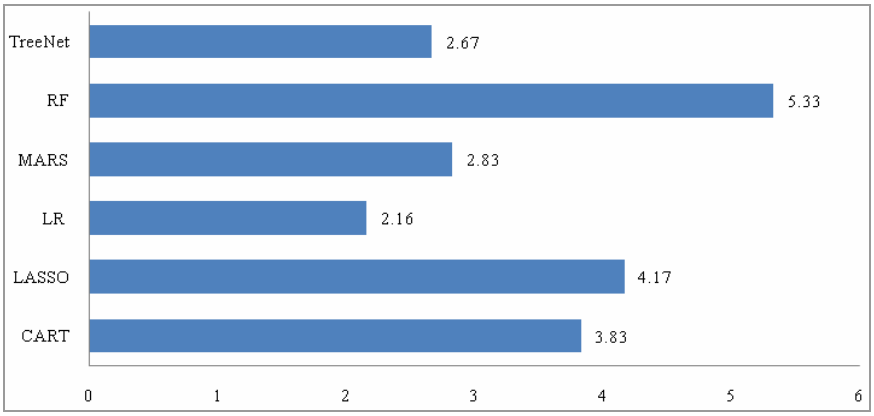


Figure 8 Average rank of P+ (see online version for colours)

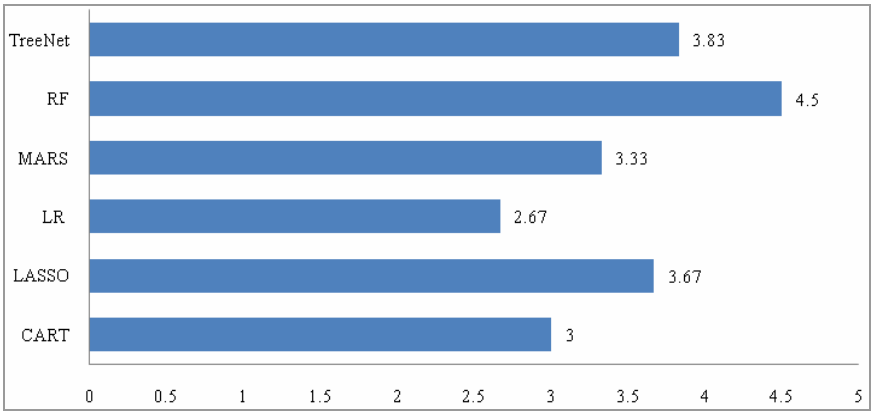


Figure 9 Average rank of P– (see online version for colours)

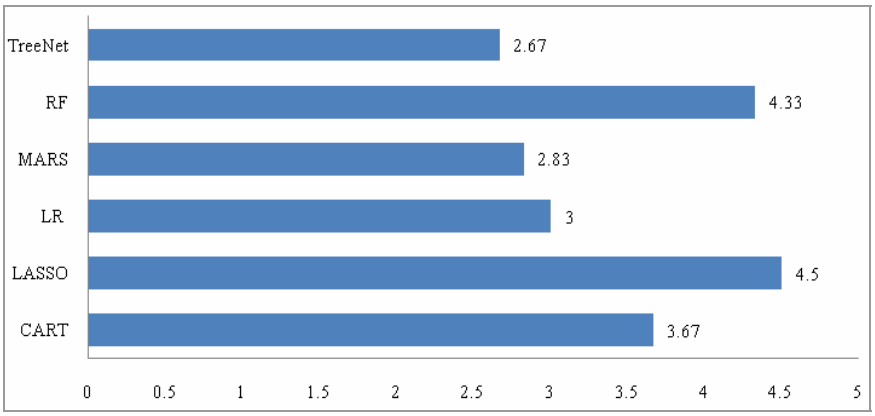
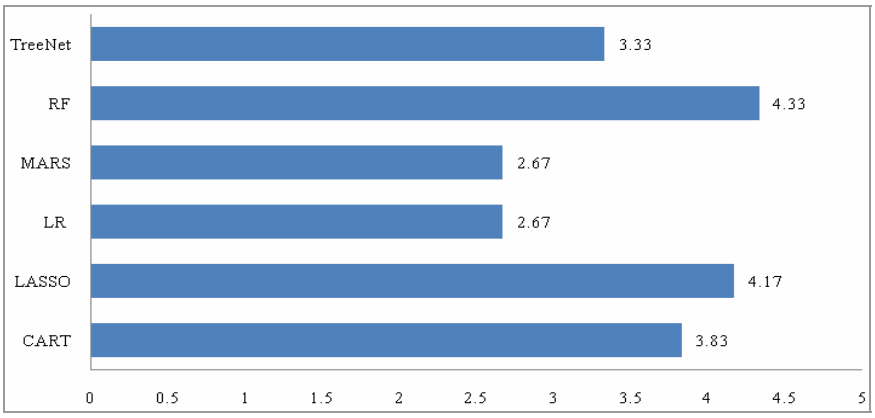


Figure 10 Average rank of DP (see online version for colours)



4.2 Cost of credit prediction errors

In this part, we sum up the costs of credit prediction errors, type I and type II errors and their influence on classifier assortment. Some previous study (Lee and Chen, 2005; Nanda and Pendharkar, 2001; West, 2000) give an opinion that accumulating these costs into the prediction models can direct to improved and more specific results. It is noticeable that the costs correlated to type I errors (a creditor being non-default is misclassified as default) and type II errors (a creditor being default is misclassified as non-default) are notably different. Generally, the misclassification costs associated with type II errors, P_{12}, are much higher and more unfavourable than those related type I errors, P_{21}. In this aspect, It is essential, to evaluate the credit prediction algorithms with their related cost, equation (15) rather than relying on the only overall accuracy of the respective model.

$$\text{EMCC} = P_{12} * \pi_2 * (x_2/X_2) + P_{21} * \pi_1 * (x_1/X_1) \tag{15}$$

To settle on the cost of the credit prediction models, the proportion of misclassification (MC) costs proposed by Dr. Hofman, connected with type II and type I, is 5:1 (West, 2000). It is not only on this relative cost ratio at 5:1, but also it presents a sensitivity analysis using higher cost ratios, e.g., 7:1, 10:1, 12:1 and 15:1. In the present study, accordingly, we reflect on four different levels of MC cost for each database. For the turmoil financial circumstances, mainly, it is predictable that the higher cost ratio might be more suitable, on the other hand, recommended that the relative cost ratio can vary from 5:1 to 20:1. To determine the cost function, it also requires an assessment of prior probabilities of non-default credit, π_1, and default credit, π_2, in the application pool of the credit prediction model. These prior probabilities are anticipated from real ratios of non-default and default credit in the empirical databases. The ratios x_2/X_2 and x_1/X_1 in equation (15) calculate the probability of making type II errors and type I errors, in that order.

Table 11 Misclassification costs achieved from Chinese SME and farmer credit datasets

Credit prediction model	*Chinese SME credit*				*Chinese farmer credit*			
	Expected misclassification cost				*Expected misclassification cost*			
	5:1	*7:1*	*10:1*	*15:1*	*5:1*	*7:1*	*10:1*	*15:1*
CART	0.0149	0.0154	0.0159	0.0162	0.0102	0.0105	0.0107	0.0108
LASSO	0.0179	0.0186	0.0192	0.0197	0.0280	0.0239	0.0206	0.0178
LR	0.0137	0.0143	0.0149	0.0153	0.0005	0.0004	0.0003	0.0002
MARS	0.0174	0.0181	0.0187	0.0192	0.0010	0.0007	0.0005	0.0004
RF	0.0193	0.0200	0.0206	0.0211	0.0921	0.0720	0.0555	0.0418
TreeNet	0.0142	0.0147	0.0152	0.0156	0.0104	0.0106	0.0108	0.0109

Tables 11–13 sum up expected misclassification (EMC) cost of the six respected models over six real-world credit databases. According to the presented tables, LR credit prediction approach is the best for the Chinese SME, a Chinese farmer, Australian and German credit databases. LR is the most excellent among all classifiers for the four datasets at all MC ratio 5:1, 7:1, 10:1, and 15:1. For the Japanese, between all credit prediction classifiers, the lowest EMC in all MC ratios, 5:1 to 15:1, is for MARS; for the Taiwan credit, it is TN. These outcomes are significant for the decision makers, to choose the appropriate balance involving error types so as not to drop prospective non-default creditors.

Table 12 Misclassification costs achieved Australian and German credit datasets

Credit prediction model	*Australian credit*				*German credit*			
	Expected misclassification cost				*Expected misclassification cost*			
	5:1	*7:1*	*10:1*	*15:1*	*5:1*	*7:1*	*10:1*	*15:1*
CART	0.0482	0.0458	0.0439	0.0423	0.1455	0.1470	0.1482	0.1492
LASSO	0.0492	0.0470	0.0452	0.0436	0.1359	0.1378	0.1393	0.1407
LR	0.0263	0.0254	0.0247	0.0241	0.1270	0.1288	0.1303	0.1315
MARS	0.0313	0.0298	0.0286	0.0275	0.1385	0.1412	0.1434	0.1453
RF	0.0552	0.0501	0.0459	0.0425	0.1602	0.1652	0.1693	0.1726
TreeNet	0.0611	0.0595	0.0582	0.0571	0.1368	0.1389	0.1407	0.1422

Table 13 Misclassification costs achieved Japanese and Taiwan credit datasets

Credit prediction model	*Japanese credit*				*Taiwan credit*			
	Expected misclassification cost				*Expected misclassification cost*			
	5:1	*7:1*	*10:1*	*15:1*	*5:1*	*7:1*	*10:1*	*15:1*
CART	0.0482	0.0458	0.0439	0.0423	0.1159	0.1170	0.1180	0.1187
LASSO	0.0476	0.0454	0.0435	0.0420	0.1314	0.1330	0.1344	0.1355
LR	0.0500	0.0481	0.0466	0.0453	0.1316	0.1333	0.1347	0.1358
MARS	0.0428	0.0405	0.0386	0.0370	0.1152	0.1161	0.1168	0.1175
RF	0.0494	0.0476	0.0461	0.0449	0.1431	0.1475	0.1510	0.1540
TreeNet	0.0562	0.0550	0.0540	0.0532	0.1131	0.1143	0.1153	0.1161

To summarise, before the conclusion, in combination with research data of Tables 10–13 and result involved are highlighted font, the following ending can be drawn.

1 The LR approach has confirmed to be a reliable and efficient approach across several performance indicator measures and most of the dataset distributions after empirical analysis, a statistical significance test, and misclassification cost estimation.

2 TN is also proficient classifier in credit scoring, attaining better results than other classifiers with CART and MARS. CART and MARS also showed competitive performances with LR and TN.

3 RF is a new but well-established classifier in credit scoring, but in this study, it could not attain competitive results with LASSO. Maybe the nature of the dataset heavily influences the performance of the models.

4 Using only one database and few performance measures is not enough to make a reasonable comparison and constant conclusion. For more authentication and discover hidden features current study employed a mixture of datasets like SME credit, farmer credit, and four public credit datasets.

5 In order to develop a comprehensive credit scoring model, this study utilises a set of performance measures combining traditional and new criterions. It will help future study to develop a more robust credit risk prediction model.

5 Conclusions

Credit scoring is an essential tool for screening loan application whose importance is increasing day by day because of worldwide financial instability in the banking sector. In contemporary banking risk management practices, the bank can correctly and effectively quantify the level of credit risk by credit scoring. Because of the importance, it is considered as a continuous contemporary research issue and due to shortfalls and gaps are still not investigated.

Therefore, the existing study is motivated to explore the capacity of six statistical and machine learning classifiers for prediction of the credit risk of potential customers. The industry standard LR and TN model have been proved as the excellent capacity for credit risk prediction. These two models of performance are also compared with four

well-known classifiers namely CART, LASSO, MARS, and RF. It is mentioned that CART and MARS also provide competitive classification ability in the credit scoring process. All of the models have been investigated by two unique datasets (Chinese SME and farmer credit) and four well known real-world public credit datasets (Australian, German, Japanese and Taiwan credit datasets). In addition, this study utilised multiple sets of representative performance measures, combining with traditional and new criterions for evaluating several algorithms. Out of ten performance metrics few are well established in credit scoring, i.e., accuracy, AUC, F-score, etc., and few are not common in credit scoring but very effective in other sectors. For example, likelihood ratios P+, P– and discriminant power DP are originated in the medical sector for clinical diagnosis.

This study can show a new way to financial institutions in their credit assessment process by multiple sets of criterion for comprehensive credit risk prediction. This capacity will assist in the decision-making process and increase profitability by enhancing credit refund, decreasing credit losses. The present study also constructs credit scoring models by utilising comprehensive SME and farmer credit data from 28 major cities in China. Furthermore, this study also will help the academicians and researchers to explore a new means of research. The present study is not beyond the limitation, most of the datasets are imbalanced, and we cannot consider all modern AI techniques like SVM, NN. In the future reaches, we will employ the data sampling technique and new AI methods to compare our current findings.

Acknowledgements

This study has been supported by the Key Projects of National Natural Science Foundation of China (the grant number 71731003 and 71431002), the General Projects of National Natural Science Foundation of China (the grant number 71471027 and 71873103), the National Social Science Foundation of China (the grant number 16BTJ017), the Youth Project of National Natural Science Foundation of China (the grant number 71601041 and 71503199). The project has also been supported by the Bank of Dalian and Postal Savings Bank of China. We thank the organisations mentioned above.

The authors' contributions are equal.

References

Ala'raj, M. and Abbod, M.F. (2016) 'Classifier consensus system approach for credit scoring', *Knowledge-Based Systems*, Vol. 104, pp.89–105.

Altman, E. and Sabato, G. (2007) 'Modeling credit risk for SMEs: evidence from the US market', *Abacus*, Vol. 43, No. 3, pp.323–357.

Altman, E.I. (1968) 'Financial ratios, discriminant analysis and the prediction of corporate bankruptcy', *The Journal of Finance*, Vol. 23, No. 4, pp.589–609.

Angelini, E., di Tollo, G. and Roli, A. (2008) 'A neural network approach for credit risk evaluation', *The Quarterly Review of Economics and Finance*, Vol. 48, No. 4, pp.733–755.

Arminger, G., Enache, D. and Bonne, T. (1997) 'Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward networks', *Computational Statistics*, Vol. 12, No. 2, pp.293–310.

Atiya, A.F. and Parlos, A.G. (2000) 'New results on recurrent network training: unifying the algorithms and accelerating convergence', *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, pp.697–709.

Banasik, J., Crook, J. and Thomas, L. (2001) 'Scoring by usage', *Journal of the Operational Research Society*, Vol. 52, No. 9, pp.997–1006.

Biggerstaff, B.J. (2000) 'Comparing diagnostic tests: a simple graphic using likelihood ratios', *Statistics in Medicine*, Vol. 19, No. 5, pp.649–663.

Blakeley, D. and Oddone, E. (1995) 'Non-invasive carotid artery testing', *Annals of Internal Medicine*, Vol. 122, No. 5, pp.360–367.

Boyes, W.J., Hoffman, D.L. and Low, S.A. (1989) 'An econometric analysis of the bank credit scoring problem', *Journal of Econometrics*, Vol. 40, No. 1, pp.3–14.

Breiman, L. (2001) 'RFs', *Machine-Learning*, Vol. 45, No. 1, pp.5–32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA, ISBN: 978-0-412-04841-8.

Briand, L.C., Freimut, B. and Vollei, F. (2004) 'Using multiple adaptive regression splines to support decision making in code inspections', *J. Syst. Softw.*, Vol. 73, No. 2, pp.205–217.

Chatterjee, S. and Barcun, S. (1970) 'A non-parametric approach to credit screening', *Journal of the American Statistical Association*, Vol. 65, No. 329, pp.150–154.

Chen, M.Y. (2011a) 'Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches', *Computers and Mathematics with Applications*, Vol. 62, pp.4514–4524, DOI: 10.1016/j.camwa.2011.10.030.

Chen, M.Y. (2011b) 'Predicting corporate financial distress based on the integration of decision tree classification and logistic regression', *Expert Systems with Applications*, Vol. 38, No. 9, pp.11261–11272.

Chen, N., Ribeiro, B. and Chen, A. (2016) 'Financial credit risk assessment: a recent review', *Artificial Intelligence Review*, Vol. 45, No. 1, pp.1–23.

Chi, G., Abedin, M.Z. and Moula F.E. (2017) 'Modeling credit approval data with neural networks: an experimental investigation and optimization', *Journal of Business Economics and Management*, Vol. 18 No. 2, pp.224–240, DOI: 10.3846/16111699.2017.1280844.

Cox, D.R. (1958) 'The regression analysis of binary sequences (with discussion)', *J. Roy. Stat. Soc. B.*, Vol. 20, No. 2, pp.215–242.

Crone, S.F. and Finlay, S. (2012) 'Instance sampling in credit scoring: an empirical study of sample size and balancing', *International Journal of Forecasting*, Vol. 28, No. 1, pp.224–238.

Cuadros-Rodríguez, L., Pérez-Castaño, E. and Ruiz-Samblás, C. (2016) 'Quality performance metrics in multivariate classification methods for qualitative analysis', *Trends in Analytical Chemistry*, Vol. 80, pp.612–624, DOI: 10.1016/j.trac.2016.04.021.

Demšar, J. (2006) 'Statistical comparisons of classifiers over multiple datasets', *The Journal of Machine Learning Research*, Vol. 7, pp.1–30.

Derelioglu, G. and Gurgen, F. (2011) 'Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey', *Expert Systems with Applications*, Vol. 38, No. 8, pp.9313–9318.

Desai, V.S., Crook, J.N. and Overstreet Jr., G.A. (1996) 'A comparison of neural networks and linear scoring models in the credit union environment', *European Journal of Operational Research*, Vol. 95, No. 1, pp.24–37.

Dimitras, A.I., Zanakis, S.H. and Zopounidis, C. (1996) 'A survey of business failure with an emphasis on prediction methods and industrial applications', *European Journal of Operational Research*, Vol. 90, No. 3, pp.487–513.

Durand, D. (1941) *Risk Elements in Consumer Installment Lending. Studies in Consumer Installment Financing*, National Bureau of Economic Research, New York.

Ewert, D.C. (1968) 'Trade-credit management: selection of accounts receivable using a statistical model', *The Journal of Finance*, Vol. 23, No. 5, pp.891–892.

Friedman, J.H. (1991) 'Multivariate adaptive regression splines', *Annals of Statistics*, Vol. 19, No. 1, pp.1–67.

Friedman, J.H. (1999a) 'Multivariate adaptive regression splines (with discussion)', *Annals of Statistics*, Vol. 19, No. 1, pp.1–141.

Friedman, J.H. (1999b) 'Stochastic gradient boosting', *Computational Statistics and Data Analysis*, Vol. 38, No. 4, pp.367–378.

Friedman, M. (1940) 'A comparison of alternative tests of significance for the problem of m rankings', *The Annals of Mathematical Statistics*, Vol. 11, pp.86–92.

García, V., Marqués, A.I. and Sánchez, J.S. (2015) 'An insight into the experimental design for credit risk and corporate bankruptcy prediction systems', *J. Intell. Inf. Syst.*, Vol. 44, No. 1, pp.159–189.

Hand, D.J. (2009) 'Measuring classifier performance: a coherent alternative to the area under the ROC curve', *Mach. Learn*., Vol. 77, No. 1, pp.103–123.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2005) 'The elements of statistical learning: data mining, inference, and prediction', *Math. Intell.*, Vol. 27, No. 2, pp.83–85.

He, H. and Garcia, E. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, pp.1263–1284, DOI: 10.1109/TKDE.2008.239.

He, H., Zhang, W. and Zhang, S. (2018) 'A novel ensemble method for credit scoring: adaption of different imbalance ratios', *Expert Systems with Applications*, Vol. 98, pp.105–117, DOI: doi.org/10.1016/j.eswa.2018.01.012.

Huang, Y., Murphey, Y.L. and Ge, Y. (2015) 'Intelligent typo correction for text mining through machine learning', *Int. J. of Knowledge Engineering and Data Mining*, Vol. 3, No. 2, pp.115–142.

Johnsen, T. and Melicher, R. (1994) 'Predicting corporate bankruptcy and financial distress: information value added by multinominal logit models', *Journal of Economics and Business*, Vol. 46, No. 4, pp.269–286.

Jones, S., Johnstone, D. and Wilson, R. (2015) 'An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes', *Journal of Banking & Finance*, Vol. 56, pp.72–85, DOI: 10.1016/j.jbankfin.2015.02.006.

Khoraskani, M.M., Kheradmand, F. and Khamesh, A.A. (2017) 'Application and comparison of neural network, C5.0, and classification and regression trees (CART) algorithms in the credit risk evaluation problem (case study: a standard German credit dataset)', *Int. J. of Knowledge Engineering and Data Mining*, Vol. 4, Nos. 3–4, pp.259–276.

Kim, K-J. and Ahn, H. (2012) 'A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach', *Computers and Operations Research*, Vol. 39, No. 8, pp.1800–1811.

Kumar, P.R. and Ravi, V. (2007) 'Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review', *European Journal of Operational Research*, Vol. 180, No. 1, pp.1–28.

Laitinen, E.K. and Laitinen, T. (2000) 'Bankruptcy prediction application of the Taylor's expansion in logistic regression', *International Review of Financial Analysis*, Vol. 9, No. 4, pp.327–349.

Lee, K.C., Han, I. and Kwon, Y. (1996) 'Hybrid neural network models for bankruptcy predictions', *Decision Support Systems*, Vol. 18, No. 1, pp.63–72.

Lee, T.S. and Chen, I.F. (2005) 'A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines', *Expert Systems with Applications*, Vol. 28, pp.743–752, DOI: 10.1016/j.eswa.2004.12.031.

Leshno, M. and Spector, Y. (1996) 'Neural network prediction analysis: the bankruptcy case', *Neurocomputing*, Vol. 10, No. 2, pp.125–147.

Lessmann, S., Baesens, B., Seow, H-V. and Thomas, L.C. (2015) 'Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research', *European Journal of Operational Research*, Vol. 247, No. 1, pp.124–136.

Li, K., Niskanen, J., Kolehmainen, M. and Niskanen, M. (2016) 'Financial innovation: credit default hybrid model for SME lending', *Expert System with Applications*, Vol. 61, pp.343–355, DOI: 10106/j.eswa.2016.05.029.

Li, Z., Tian, Y., Li, K., Zhou, F. and Yang, W. (2017) 'Reject inference in credit scoring using semi-supervised support vector machine', *Expert Systems with Applications*, Vol. 74, pp.105–114, DOI: doi.org/10.1016/j.eswa.2017.01.011.

Makowski, P. (1985) 'Credit scoring branches out', *Credit World*, Vol. 75, No. 1, pp.30–37.

Mishra, A. and Srivastava, V. (2012) 'Extracting a knowledge from source code comprehension using data mining methods', *Int. J. of Knowledge Engineering and Data Mining*, Vol. 2, Nos. 2–3, pp.174–199.

Moula, F.E., Chi, G. and Abedin, M.Z. (2017) 'Credit default prediction modeling: an application of support vector machine', *Risk Manag.*, Vol. 19, pp.158–187, DOI: 10.1057/s41283-017-0016-x.

Myers, J.H. and Forgy, E.W. (1963) 'The development of numerical credit evaluation systems', *Journal of the American Statistical Association*, Vol. 58, No. 303, pp.799–806.

Nanda, S. and Pendharkar, P. (2001) 'Linear models for minimizing misclassification costs in bankruptcy prediction', *International Journal of Intelligent Systems in Accounting, Finance & Management*, Vol. 10, pp.155–168, DOI: 10.1002/isaf.203.

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D. and Joksimovic, I. (2013) 'The application of brute force logistic regression to corporate credit scoring models: evidence from Serbian financial statements', *Expert Systems with Applications*, Vol. 40, No. 15, pp.5932–5944.

Ohlson, J. (1980) 'Financial ratios and the probabilistic of bankruptcy', *Journal of Accounting Research*, Vol. 18, No. 1, pp.109–131.

Orgler, Y.E. (1970) 'A credit scoring model for commercial loans', *Journal of Money, Credit and Banking*, Vol. 2, No. 4, pp.435–445.

Piramuthu, S. (1999) 'Financial credit-risk evaluation with neural and neuro-fuzzy systems', *European Journal of Operational Research*, Vol. 112, No. 2, pp.310–321.

Pisharody, A.S., Pargaonkar, S. and Kulkarni, V.Y. (2015) 'Fingerprint classification and building a gender prediction model using random forest algorithm', *Int. J. of Knowledge Engineering and Data Mining*, Vol. 4, Nos. 3–4, pp.286–298.

Schaefer, S.M. and Strebulaev, I.A. (2008) 'Structural models of credit risk are useful: evidence from hedge ratios on corporate bonds', *Journal of Financial Economics*, Vol. 90, pp.1–19, DOI: 10.1016/j.jfineco.2007.10.006.

Siddiky, F.A., Kabir, F. and Rahman, S.M.M. (2012) 'Data generation with K-means for scalable data mining', *Int. J. of Knowledge Engineering and Data Mining*, Vol. 2, Nos. 2–3, pp.215–235.

Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing and Management*, Vol. 45, pp.427–437, DOI: 10.1016/j.ipm.2009.03.002.

Tibshirani, R. (1996) 'Regression Shrinkage and Selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological), JSTOR*, Vol. 58, No. 1, pp.267–288, Wiley.

Tinoco, M.H. and Wilson, N. (2013) 'Financial distress and bankruptcy prediction among listed companies using accounting, market, and macroeconomic variables', *International Review of Financial Analysis*, Vol. 30, pp.394–419, DOI: 10.1016/j.irfa.2013.02.013.

Tsai, C-F. and Wu, J-W. (2008) 'Using neural network ensembles for bankruptcy prediction and credit scoring', *Expert Systems with Applications*, Vol. 34, No. 4, pp.2639–2649.

Vukovic, S., Delibasic, B., Uzelac, A. and Suknovic, M. (2012) 'A case-based reasoning model that uses preference theory functions for credit scoring', *Expert Systems with Applications*, Vol. 39, No. 9, pp.8389–8395.

West, D. (2000) 'Neural network credit scoring models', *Computers & Operations Research*, Vol. 27, Nos. 11–12, pp.1131–1152.

Wiginton, J.C. (1980) 'A note on the comparison of logit and discriminant models of consumer credit behavior', *Journal of Financial and Quantitative Analysis*, Vol. 15, No. 3, pp.757–770.

Zhong, H., Miao, C., Shen, Z. and Feng, Y. (2014) 'Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings', *Neurocomputing*, Vol. 128, pp.285–295, DOI: 10.1016/j.neucom.2013.02.054.