A Predictive Analysis of Credit Risk Evaluation and the Quality Decision Making using Different Predictive Models

U Bhuvaneswari, Sharon Sophia

Abstract**: *In the fast growing economy, the role of credit plays a very significant role. Most of the financial institutions offer credit to their customers who are in need to meet their personal requirements and as to pay it back within specific period. Any institution which provides the credit service will predict the credit risk towards an individual which highlights the capability of the person to pay back the amount along with their previous available pay back records. Many predictive models were developed to predict the credit risk with many different variables. In the present work, the different credit risk predictive models were evaluated and compared based on the quality of decision making. The primary data were collected from 151 respondents through various online sources with the structured questionnaire and the secondary data from the previous records. The metrics derived from the predictions reveal high accuracy and precision. From the analysis, the prediction accuracy and time for the linear SVM technique was better than all other methods..***

Index Terms**: *credit risk; predictive model; quality decision; accuracy; linear svm***

I. INTRODUCTION

Providing loans to probable debtors are the main business of banks around the world. Enormous default loss and serious competition over the applicants require accurate and potentially discriminatory to financial intermediaries. So, banks have to decide whether to extend or not credit during application screening. Data are mostly derived from the customer demographics, application tables, and extensive records of previous loans and repayment activities (Xia et al, 2017). Credit risk is most of the often challenged financial risks that can be described as the probability that other borrowers will miscarry to encounter its responsibilities by contracted policies that will budget invested money for investor. Hence the evaluation of the credit score or risk was very significant. (Danenas, P., & Garsva, G. 2015).

Credit assessment is also the most critical procedures of banks recognized as credit management judgments. The procedure comprises Collect, analyze, and categorize various credit variables and elements to evaluate credit judgments. Categorizing the bank's customers is also a portion of credit scoring, reducing the risk of the customer's current and discounted credit. Credit risk is defined as the borrower's failure to repay the debt or the principal or interest within the agreed period of time. The failure to repay the debt is considered as a crime and can be treated as theft or fraud which requires penalty. And there are limits to these penalties. According to the Basel Handbook [The Basel Handbook, 2007], credit risk is the major risk to which banks are exposed, whereas making loans is the primary activity of most banks. The credit risk assessment is important to focus on the future prediction. The model performance is based on the data given to the model for training purpose. And in this data the credit risk analysis should be carried out. In this paper, the different credit risk predictability models that include logistic regression, Bagged trees, Fine Gaussian and Linear SVM, simple and complex trees methods were analyzed based on their decision making quality. The rest of the paper has been organized in the following manner as: the second section details all previous works on the different models that were employed for the credit risk predictions. The subsequent section will provide the insights on research methodology. The fourth section deliberates the assessment of the obtained data through the different predictive models. The next section lists the results obtained along with its discussion. The final section delivers the finding and implication from the comparative study and suggests the scope for future work.

II. LITERATURE REVIEW

A. *Bagging tree methods*

Zhang et al. (2010) proposed the novel method for credit scoring and named it to be the vertical bagging decision tree model that was based on the accumulation of the classifiers through the predictive attributes combination. The presented method was evaluated with the 31 German and 9 Australian credit databases. The evaluated outcome showed that the novel method has very high accuracy than the traditional approach by the weight vote strategy. Paleologo et al (2010) highlighted the drawbacks that existed in the present credit risk prediction models and proposed the usage of the sub bagging techniques which would be appropriate for the data used in credit scoring which were highly unbalanced. The recommended method was used in different classifiers and in their sub bagged versions. Even though being a simple approach it provided the better performance with reasonable interpretability.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author

U Bhuvaneswari, VITBS, VIT University, Chennai, India
Dr Sharon Sophia, VITBS, VIT University, Chennai, India

Dahiya et al. (2017) presented the feature selection-based hybrid-bagging algorithm that was employed to evaluate the credit risk. The method used both the chi square and the principle component analysis for the feature selection and 5 sets of training and testing data sets was used as the input.

The proposed algorithm had performed better than the stand alone classifiers as it uses only the significant features for classifier development and follows the ensemble methodology.

B. *Logistic regression model*

Since there was a possibility of expressing the details of the loan in the form of fuzzy numbers, Sohn et al. (2016) recommended the Fuzzy Logistic Regression model to predict the credit scoring. The fuzzy least square estimation technique was used in this proposed methodology and the performance on credit scoring was improved. The major drawback was that this method did not have any proper diagnostic methodology. Kral et al. (2016) evaluated the predictability model for the bankruptcy with the application of logistic regression. The study revealed that the most bankruptcy models in the world having the predictable ability that belongs to the concept of logistic regression. The main aspect to be reflected in this model was that the number of variables used for evaluation was not exhibited clearly. Caigny et al. (2018) proposed the novel model that was an integrated form of the logistic regression and tree model and named it to be Logit Leaf model. The two stage methodology was used to segment the data with the decision rule and construct them using the tree model. Being the hybrid model, the performance of the novel method was superior to that of the logistic regression or the decision tree.

C. *SVM*

Ribeiro et al. (2012) proceeded with dividing the assorted dataset into numerous categorized clusters regarding the scope and yearly income of organizations, a commercial agony expectation model centered on SVM accomplished leading expectedness enactment associated to the standard SVM and SVM with multiple learning tasks. Harris (2015) uses SVM Linear and Linear Classification with Clustered SVM for the development of model with 20,000 entries from Barbados lending unions. Results suggest that the performance of straight SVM from SVM used with RBF kernel has not significantly varies. Zhong et al. (2014) employed MLP and SVM with another two algorithms for credit score assessment, and showed that SVM ratings were perfectly distributed and the SNM was well implemented for NNs methods to better reliability. Zhang, et al, (2014), has reported that more than 6000 US cases have been employed, cases showed that linear SVM (result 75% accuracy) does not show a significant increase in linear SVM kernel's application classification for general SVM, fuzzy SVM and hybrid fuzzy SVMs. Another mechanism that relate to their expansion on greater data’s employ SVM with the RBF kernel function, the most popular choice in such research. Vladimir Vapnik proposed the Supported vector machine as controlling machinery used to categorize huge size data. The method employs the nonlinear planning to alter the innovative training database to a greater angle. This allows the SVM to pursuit for the optimum centrifuge hyper-aircraft that acts as an operative fixed border, splitting cluster groups with the major probable margins among them.

2.4 Tree models

Wu and Hsu (2012) implemented a support decision model centered on the DT for choice creation on credit risk evaluation. The performance suggested that the expected performance, overall application, and descriptive power was promising, as it created different instructions with the significance of the vector machine and the DT. Though, the rottenness of the acquired knowledge acquisition is unproductive due to the receptor vector machine is complicated and expensive.

Stochastic Gradient Boost (SGB) is an adaptable and dynamic data mining instrument that generates several small deciduous tree-making error-correcting processes. SGB's adaptability aids to erase the tainted information with false target tickets. Such data is frequently more difficult to increase conservative and is defy in maintaining while employing traditional data mining instruments; SGB is low due to such deficiencies (Mukkammala et al, 2006). Random Forests (RF) is a collective practice technique for taxonomy and retreat. In taxonomy, outcome is the method of taxonomy of individual trees. In regression, the output is average from each discrete tree. According to their inner atmoshpere, the RF models bring an odds standard in observation. An RF inequality standard can also be defined as there is no data available. The awareness is to construct an RF analyst, which separates the observed data from the synthetic data that is properly prepared (Chandra et al. 2009). Delen, et al, (2013) to investigate the effect of financial ratios on applied the CHAID, C5.0, Quest and Cart company performance. In four CHAID, DT algorithms, and C5.0 produced the best estimate accuracy.

III. METHODOLOGY

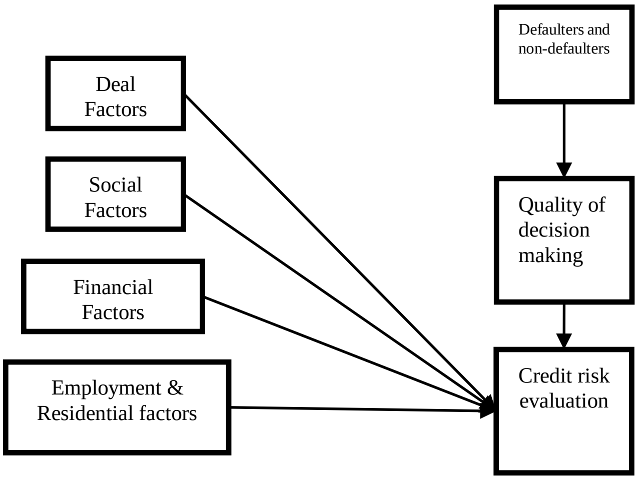
Defaulters and non-defaulters
Deal Factors
Social Factors
Financial Factors
Employment & Residential factors
Quality of decision making
Credit risk evaluation

Figure 1 Framework of the study

The main objective of the study is to analyze the different predictive models based on credit risk evaluations for quality decision making. For this research we have selected 6 models for evaluation of the decision making. In this evaluation, totally we have considered 4 factors like Deal factors, Social factors, financial factors and Employment factors. This 4 factors are considered for the assessing the model in the prediction of credit risk portfolio. Factors influencing the individuals about the decision of the car purchasing and the credit risk evaluating in the 6 selected models are studied, and the performance evaluation is done. The 6 models are Bagged Trees model, Logistic Regression, Fine Gaussian SVM, Linear SVM, Simple tree and Complex tree. This study utilizes quantitative methods for data collection for analyzing the 6 models. For quantitative research, the data of the 151 samples are collected for analysis of credit risk analysis which is divided into defaulters and non-defaulters. The present study, three common measurements in the credit risk assessment field was designated to measure the quality of the samples. The three indicators comprise the average accurateness (average), the first type error (type I error) and the second type error (type II error). Among them, the first and second types of classification errors in the credit risk assessment system are errors are commonly considered.

A. Performance evaluation:

In order to assess the suitability of the aforementioned classification and regression ensemble models, we have employed the before applied standard classification and regression performance steps of credit risk modeling (Florez-Lopez and Ramon-Jeronimo, 2015; Witten, et al, 2016; Yao, et al, 2017). Standard performance standards for accuracy (Acc) and region under the receiver operative inherent line (AUC) considered for analyzing PD models. Accuracy based on the confusion matrix (Table 1) is exactly the percentage of categorized loans:

$$\text{Acc} = \frac{TP+TN}{TP+FP+FN+TN}$$

where TP, TN, FP and FN are the numbers of instances classified as true positive, true negative, false positive and false negative, respectively.

TABLE 1: CONFUSION MATRIX FOR MODELLING PD

Prediction/Target	Positive	Negative
Positive	TP	FP (type I error)
Negative	FN (type II error)	TN

IV. RESULTS

A. *Bagged trees*

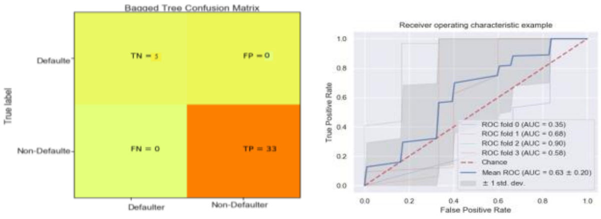

Figure 2: Confusion matrix and ROC curve for Bagged trees

From the above Figure.2, ROC curve explains the area under the curve. Roc curve area for fold 0 is 0.35. That means 35 percent of the area is covered under Roc fold 0. And the Roc curve for fold 1 is 0.68. That is 68 percent of the area is under the fold 1. Roc curve for fold 2 is 0.90. That means, fold 2 is covered almost 90 percent of the total area. Roc for fold 3 is 0.58. That means, area under the curve for fold 3 is 58 percent of the total area. Thus it is concluded that perfect ROC fold 1 is 0.68, Roc fold 2 is 0.90, and Roc fold 5 is 0.58. These three are the True positive rate and the Roc curve graph which increases gradually for the highest level. This is concluded to say that accuracy of the results is high as the graph of Roc curve increases.

B. *4.2. Logistic regression*

From the above Figure 3, ROC curve explains the area under the curve. Roc curve area for fold 0 is 0.44. That means 45 percent of the area is covered under Roc fold 0. And the Roc curve for fold 1 is 0.26. That is 26 percent of the area is under the fold 1. Roc curve for fold 2 is 0.51. That means, fold 2 is covered almost 51 percent of the total area. Roc for fold 3 is 0.56. That means, area under the curve for fold 3 is 56 percent of the total area. Thus it is concluded that perfect roc curve: Roc fold 2 is 0.51 and Roc fold 3 is 0.56. This two are the True positive rate and Roc curve is increasing gradually to the highest point. This is concluded to say that the accuracy also increases for the corresponding results.

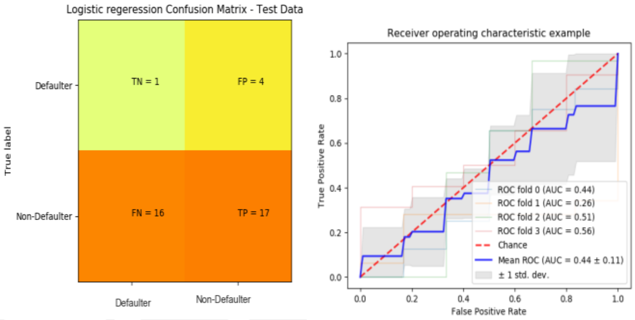

Figure 3: Confusion matrix and ROC curve for Logistic Regression

C. *Fine Gussian SVM*

From the above Figure 4, it shows six AUC scores. The score is 1.0 for the classifier with the perfect performance level (P) and 0.5. ROC curves clearly shows classifier ROC fold 0, ROC fold 1, ROC fold 2, ROC fold 3, ROC fold 4, ROC fold 5. The highest area covered is fold 2 i.e. 52% of the total are is cover under fold 2. The perfect roc curve-Roc fold 1 is 0.52, Roc fold 2 is 0.54, and Roc fold 5 is 0.52. These three are the True positive rate and Roc curve tend to increase the graph gradually to the highest point. That means the area under the curve is large. This is concluded to say that results are more accuracy.

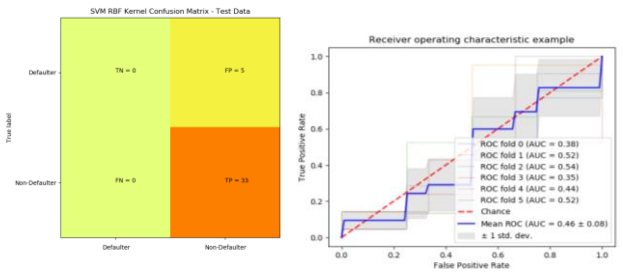

Figure 4: Confusion matrix and ROC curve for Fine Gaussian SVM

D. *4.4. Linear SVM*

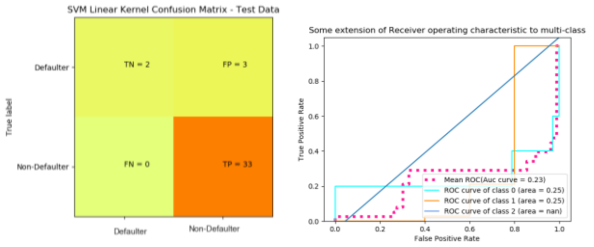

Figure 5: Confusion matrix and ROC curve for Linear SVM

ROC curve explains the area under the curve which is plotted in the above figure 5. Micro-average ROC curves is 0.23. That means 23 percent of the area is covered under Roc micro-average curve. And the Roc curve of class 0 is 0.25. That is 25 percent of the area is under the ROC curve for class 0. Roc curve for class 1 is 0.25. That means, class 1 is covered almost 25 percent of the total area. Thus it is concluded that all the curves are imperfect roc curve and the Roc curve is tends to be increasing gradually for all the ranges.

E. *4.5. Simple tree*

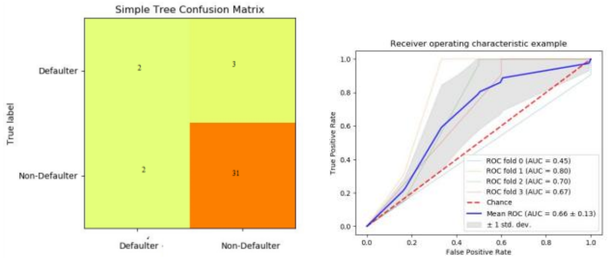

Figure 6: Confusion matrix and ROC for Simple tree

ROC curve explains the area under the curve which is plotted in the above figure 6. Roc curve area for fold 0 is 0.45. That means 45 percent of the area is covered under Roc fold 0. And the Roc curve for fold 1 is 0.80. That is 80 percent of the area is under the fold 1. Roc curve for fold 2 is 0.70. That means, fold 2 is covered almost 70 percent of the total area. Roc for fold 3 is 0.67. That means, area under the curve for fold 3 is 67 percent of the total area. Thus it is concluded that perfect roc curve is Roc fold 1 is 0.80, Roc fold 2 is 0.70 and Roc fold 3 is 0.67. This three are the True positive rate.

F. *Complex tree*

ROC curve explains the area under the curve which is plotted in the above figure. Roc curve area for fold 0 is 0.55. That means 55 percent of the area is covered under Roc fold 0. And the Roc curve for fold 1 is 0.74. That is 74 percent of the area is under the fold 1. Roc curve for fold 2 is 0.77. That means, fold 2 is covered almost 77 percent of the total area. Roc for fold 3 is 0.68. That means, area under the curve for fold 3 is 68 percent of the total area.

Thus it is concluded that perfect roc curve is Roc fold 0 is 0.55, Roc fold 1 is 0.74, Roc fold 2 is 0.77 and Roc fold 3 is 0.68. This four are the True positive rate.

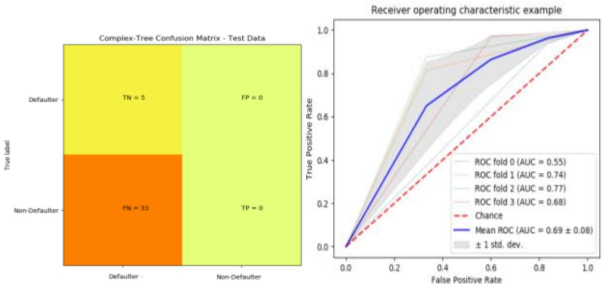

Figure 7: Confusion matrix and ROC curve for Complex tree

G. *Comparison:*

So as to additionally to prove the authenticity of proposed model proposed in the present research, the running results of AUC value of each model was also chosen to match the estimated accuracy of various models. Likewise, with the purpose to decrease the effectiveness of the options, to reduce the limit of the test set and tests set on the initial weight and results, the selected evaluation indicators are 4 times after the program's 4 results. The results of each group's AUC value have been entered and the average value is considered.

TABLE 2 THE RECORDED AUC VALUES

Model	**0**	**1**	**2**	**3**	**AVG**
Bagged trees	0.35	0.68	0.90	0.58	0.63
Logistic regression	0.44	0.26	0.51	0.56	0.44
Fine Gaussian SVM	0.38	0.52	0.54	0.35	0.46
Linear SVM	0.23	0.25	0.25	0.23	0.24
Simple tree	0.45	0.80	0.70	0.67	0.66
Complex tree	0.55	.74	.77	.68	0.69

Table 2 shows the AUC and average value each model of the results of the 4 corresponding results of the various credit risk evaluation models. The observation displays that

TABLE 3 MODEL COMPARISON CHART

	Bagged Trees	Logistic Regression	Fine Gaussian SVM	Linear SVM	Simple Tree	Complex Tree
Accuracy	90%	47%	87 %	92%	87 %	13 %
~ **Speed (obs/sec)**	0.007s	0.05	0.004	0.003	0.006	0.004
Training Time (secs)	0.018	0.005	0.007	0.007	0.015	0.013

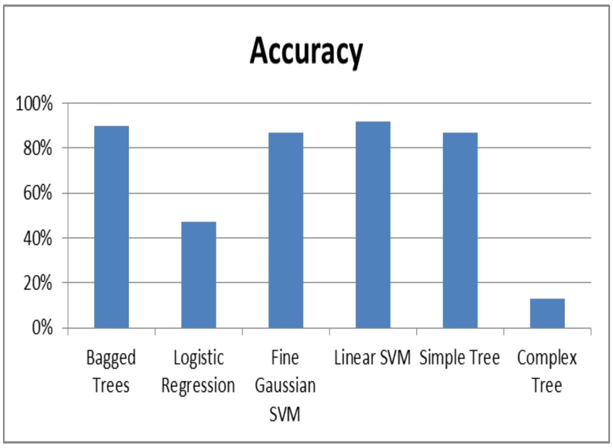

Figure 8: Comparison of Accuracy of Each model

Thus from the above Figure comparison chart of all the methods in order to evaluate the quality of portfolio, it is concluded that from the analysis made by using bagged trees having the maximum accuracy of 90% with speed of 0.007seconds. And linear SVM is observed to have 92% of the accuracy with 0.003 seconds speed. Complex trees which is having less accuracy rate among all other methods (13%) having speed of 0.004 seconds and training time as 0.013. Therefore from the overall analysis of portfolio evaluation linear Support Vector Machine possesses high accuracy results in classification models of 92%. From the above comparison of the models for the credit risk evaluation, the Linear SVM has shown a best accuracy with less time. Because the evaluation of model, the responses of non-defaulters have impacted over all the factors that was presented in this model. So the accuracy of this model is high when compared to other model which is described in above table.

V. CONCLUSION

In the present study, R is used for such data mining functions available in the package and this database is considered from the UCI catalouge. The pre-processing phase is one of the most significant and time considering, classification and clustering methods in R is employed to prepare data for additional usage.The Linear SVM model is the one which got the highest score of the accuracy in analysing customer behaviours. It is evident that in the linear SVM model the years of experience, asset price, approximate total income and disposable income to EMI ratio has made a significant influence over the default of the customers. Thus, this research was effective in identifying the substantial elements and their permutations which categorize credit risk of the portfolio management for defaulters or non-defaulters.

Predictive model can be employed to evaluate the probability of credit nonpayment. Receiving such approaches can help improve customer knowledge and enhance the retrieval amount for a company. Consumer can be categorized by two clusters to award the credit or not to consumer based on model of credit scoring. Employing the modern taxonomy method enhances credit risk evaluation over the traditional taxonomy model. Bearing in mind a profit centered tactic in the taxonomy framework raises the effectiveness of the loan-granting conclusion for the company, which increases productivity. Performance measurement measures the welfares created by vigorous credits and aids to remove debt default costs. Consequently, profit-oriented method permits for a profit-based model assortment and consents to recognize the credit scoring model.

REFERENCES

1. Chandra, Karthik, Vadlamani Ravi, and Indranil Bose Failure prediction of dotcom companies using hybrid intelligent techniques. Expert Systems with Applications 2009, 36: 4831–37
2. Dahiya, S., Handa, S. S., and Singh, N. P. *A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. Expert Systems, 2017, 34(6), e12217.* doi:10.1111/exsy.12217
3. Danenas, P., and Garsva, G. *Selection of Support Vector Machines based classifiers for credit risk domain. Expert Systems with Applications, 2015, 42(6), 3194–3204.*doi:10.1016/j.eswa.2014.12.001
4. De Caigny, A., Coussement, K., and De Bock, K. W. *A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 2018, 269(2), 760–772.*doi:10.1016/j.ejor.2018.02.009
5. Delen D, Kuzey C, Uyar A Measuring firm performance using financial ratios: a decision tree approach. Expert Syst Appl 2013, 40(10):3970–3983
6. Harris, T. Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 2015, *42*(2), 741-750.
7. Mukkamala, Srinivas, Armando Vieira, and Andrew H. Sung. Model selection and feature ranking for financial distress classification. Paper presented at the 8th International Conference on Enterprise Information Systems (ICEIS 2006), Paphos, Cyprus, May 2006, 23–27.
8. Ong, M. K. (Ed.). *The Basel Handbook: A guide for financial practitioners*. Risk Books, 2007.
9. Paleologo, G., Elisseeff, A., and Antonini, G. *Subagging for credit scoring models. European Journal of Operational Research, 2007, 201(2), 490–499.*doi:10.1016/j.ejor.2009.03.008
10. Ribeiro, B., Silva, C., Chen, N., Vieira, A., and das Neves, J. C. Enhanced default risk models with SVM+. *Expert Systems with Applications*, 2012, *39*(11), 10140-10152.
11. Sohn, S. Y., Kim, D. H., and Yoon, J. H. *Technology credit scoring model with fuzzy logistic regression. Applied Soft Computing, 43, 150–158.*doi:10.1016/j.asoc.2016.02.025
12. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
13. Wu, T.-C., and Hsu, M.-F. Credit risk assessment and decision making by a fusion approach. Knowledge-Based Systems, 2012, 35, 102–110.
14. Xia, Y., Liu, C., Li, Y., and Liu, N. *A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 2017, 78, 225–241.*doi:10.1016/j.eswa.2017.02.017
15. Zhang, D., Zhou, X., Leung, S. C., and Zheng, J. Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, 2010, *37*(12), 7838-7843.
16. Zhang, Z., Gao, G., and Shi, Y. Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. European Journal of Operational Research, 2014, 237, 335–348.
17. Zhong, H., Miao, C., Shen, Z., and Feng, Y. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. Neurocomputing, 2014, 128, 285–295. https://doi.org/10.1016/j.neucom.2013.02.054
18. Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2016.
19. Florez-Lopez, R., and Ramon-Jeronimo, J. M. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. Expert Systems with Applications, 2015, 42(13), 5737-5753.
20. Yao, X., Crook, J., and Andreeva, G. Enhancing two-stage modelling methodology for loss given default with support vector machines. European Journal of Operational Research, 2017, 263(2), 679-689.