

A Better Comparison Summary of Credit Scoring Classification

Sharjeel Imtiaz
School of Computing
Doctorate Data Science Student
University of East London (UEL)
London, United Kingdom

Professor Allan J. Brimicombe
School of Computing Head,
Centre for Geo-Information Studies
University of East London
London, United Kingdom

Abstract—The credit scoring aim is to classify the customer credit as defaulter or non-defaulter. The credit risk analysis is more effective with further boosting and smoothing of the parameters of models. The objective of this paper is to explore the credit score classification models with an imputation technique and without imputation technique. However, data availability is low in case of without imputation because of missing values depletion from the large dataset. On the other hand, imputation based dataset classification accuracy with linear method of ANN is better than other models. The comparison of models with boosting and smoothing shows that error rate is better metric than area under curve (AUC) ratio. It is concluded that artificial neural network (ANN) is better alternative than decision tree and logistic regression when data availability is high in dataset.

Keywords—Credit score data mining; classification; artificial neural network; imputation

I. INTRODUCTION

In 2006 Taiwan faced credit flow crises as matter of fact to propel in market the bank has over issued cash and credit regardless of their ability to repayments, overconsumptions history. This situation blows consumer and finance rapport badly. The well down financial system crisis on downstream and risk management on upstream. The purpose of risk management is to enforce checks on consumer ability to repay bills thus reduce damage and consumer's credit repayment uncertainty.

A lender commonly makes two types of decisions: first, whether to grant credit to a new application or not, and second, how to deal with existing application, including whether to increase their credit limits or not [1]. Scoring model is better alternative for traditional model but model is not perfect-sometimes a bad application will receive high score and accepted, therefore, model is needed. There are various methods testing in past and in this literature that group into parameterized and non-parameterized methods. In this direction, [2]-[3] highlighted the importance of artificial neural network (ANN) and support vector machine (SVM) towards credit scoring as a better alternative of conventional approaches. Further from analysis point of view this literature is not focusing on ensemble methods or hybrid models' due to complexity of design. Although, admired the complexity of ensemble models and utilized for multiple tasks [4].

The non-parameterize model logistic regression is not perfectly suitable for the classification than parameterized models. These models are decision trees [5]-[7] and artificial neural networks [8]-[10]. Here, ultimate focus is to elaborate the importance of parameters of models with smoothing and boosting with the dataset imputation. This research will further interrogate the existing approaches to find the better smoothing and boosting model with respect to imputation technique.

II. RELATED WORK

The work of author [11] has proven that area ratio (AUC) is better metric in accuracy than error rate compared with decision tree, logistic and ANN. It was concluded that ANN is better in performance in area ratio metric regardless of over consumed computation. Our approach further analyzed the existing work in depth to prove that boosting and smoothing of models worked well for these core approaches especially for ANN, which is better determiner of metric accuracy. That is the reason, why we contradict the existing approach because existing approach undertake the same dataset without treatment of missing values with imputation technique. Our work is continuation of the existing work to prove that neural network on similar dataset performed better when imputation technique applied over smoothing and boosting of models.

III. OVERVIEW OF EXISTING WORK

A. Decision Tree

Decision trees allow creating a tree-based classification model. Decision trees can graphically illustrate other choices that can be made and enable the decision maker is to identify the best situation in a circumstance. Common algorithms for decision tree induction include ID3, C4.5, CART, CHAID and QUEST [12]. In [13], author says that the decision rules should maximize a divergence measure of the difference in default risk between the two subsets. The splitting is repeated until no group can be split into two subgroups which are statistically different. According to [14], there are three major tasks of a classification tree: 1) how to partition the data at each step; 2) when to stop partitioning; and 3) how to predict the value of y for each x in partition.

B. Logistic Regression

Logistic regression analysis is the multivariate technique, which allow to estimate the probability that an event occurs or

not, by predicting a binary dependent outcome from a set of independent variables. The logit model is a widely used statistical parametric model for modelling binary dependent variable. The logit model for credit scoring is presented with comparisons with other models including conventional one [15].

C. Artificial Neural Network (ANN)

ANN is an information processing model resembling connections structure in the synapses. It consists of many nodes (also called neurons or units) by links. The feed-forward neural network with back propagation (BP) is widely used for credit scoring, where the neurons receive signals from pre-layer and output them to the next layers without feedback.

According to [16], made a comparison of neural networks and linear scoring models in the credit union environment and the results indicated that neural network had better performance for correctly classifying bad loans than LR model. Besides, ANN need many training samples and long learning time. In [17], found that ANN has a higher accuracy rate by comparing with Logistic regression and discriminate analysis.

IV. PROPOSED SOLUTION AND DATASET

This section is discussing about boosting and smoothing criteria for each model with parameters. The purpose of the approach parametrization is to evaluate the different conventional models to improve the accuracy of classification. The dataset chosen from UCI website as shown below, it is about 2006 Taiwan faced credit flow crises thus result of repayments evidences are required. All the dependent and independent variables are given in Table 1. Later the models feed by opted strong and moderately correlated variables after data insight analysis technique, because, these variables are not necessarily correlated to each other thus prune from dataset eventually after correlational test. The categorical variables sex, education, marital status and age and continues variable Limit_BAL eventually have weak correlation with other independent variables, therefore, not included as input to the models. The final set of 19 independent variables has given as input to the models from Table 1.

A. Artificial Neural Network (ANN)

According to [18], it has presented credit scoring by integrating back propagation neural networks with linear method. Linear model (1) and non-linear model (2) definitions are as follows:

$$y = h(x_i; w) = w^T x \tag{1}$$

$$a = \sigma \left(\sum_k \omega_{jk}^l a_k^{l-1} + b_j^l \right) \tag{2}$$

TABLE I. CREDIT SCORE RELETED VARIABLES

Variable	Description	Scale of variable
Limit_BAL	Amount of the given credit (NT dollar)	Continuous Interval
Sex	Gender (1 = male; 2 = female).	Categorical Nominal
Education	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).	Categorical Nominal
Marital Status	Marital status (1 = married; 2 = single; 3 = others).	Categorical Nominal
Age	Age (year).	Continues Interval
Pay_0 to Pay_6	April to September	Categorical
Bill_AMT1 to BILL_AMT6	Amount of bill statement (NT dollar)	Continues Interval
Pay_AMT1 to PAY_AMT6	Amount of previous payment (NT dollar)	Continues Interval
Y	Default payment (Yes = 1, No = 0)	Dichotomous

To find the best model, the gradual increment of hidden nodes with settings of 5 hidden followed by 2 hidden layers are involving better for the result accuracy. There are 19 input parameters of all three models, where, ANN-l is single layer default model, ANN-H model train with minimum hidden layers and ANN-L model is train with linear model without any activation function. The results are shown as in Table 2, where, ANN-L performed slightly better than all other models ANN-l and ANN-H with accuracy metrics error rate and area under curve (AUC).

B. Decision Tree (DT)

In [19] author says, the decision rules should maximize a divergence measure of the difference in default risk between the two subsets. The splitting is repeated until no group can be split into two subgroups which are statistically different. According to [20], there are three major tasks of a classification tree: (i) how to partition the data at each step, (ii) when to stop partitioning and (iii) how to predict the value of y for each x in partition. The decision tree [21] purpose is to find the optimal sub-tree that gives bad and good credit based on overall accuracy and error rate. This paper evaluated result using C4.5 classification, which formulate classification tree based on principle of entropy (1) and information gain principle (2).

$$E(S) = \sum_{i=1}^c -p \log_2 p_i \tag{1}$$

$$E(T, X) = \sum_{c \in X} P(c) E(c) \tag{2}$$

TABLE II. STATISTICS OF ANN

Model	Error Rate		AUC	
	Training	Testing	Training	Testing
ANN - 1	75.22	76.0	69.5	70.0
ANN -H	78.11	80.22	73.6	76.33
ANN -L	78.81	81.46	73.6	78.03

TABLE III. STATISTICS OF DT

Model	Error Rate		AUC	
	Training	Testing	Training	Testing
Default	79.75	78.48	75.89	73.25
Pruning	79.58	78.52	75.33	72.83
Boosting 10%	80.57	78.79	77.31	73.84
Boosting 100%	80.0	78.69	76.27	73.09

Table III: Default DT, Pruning DT, Boosting 10% , Boosting 100%

TABLE IV. STATISTICS OF LOGISTIC REGRESSION

Model	Error Rate		AUC	
	Training	Testing	Training	Testing
Default	77.76	81.22	73.4	77.91

Table IV: Regression Classification

The split of tree based on pure values evaluated by measures entropy and information gain. The result evaluated and compared among four model configurations as in Table 3. It was presented that boosting 10% of decision tree model is better in terms of accuracy than all other configuration such as booting 100%, pruning and default DT model.

C. Logistic Regression

According to [22], it was the first paper published investigates the logistic regression (LR) with discriminant analysis applied to credit scoring. Its results shows LR exhibiting higher accuracy rates, however, neither method was found to be sufficiently good to be cost effective for his problem. LR was also applied by [23] to a commercial loan evaluation process (exploring several models using random effects for bank branches).

$$P=(\alpha+\beta x) \quad \text{Eq. Logistic Regression}$$

In Table 4, logistic regression error rate metric considerably evaluated results better than AUC for bad and good credit. The area under curve (AUC) metric accuracy is 77.91, on the opposite, error rate metric test accuracy 81.22 is far better than AUC.

V. RESULT DISCUSSION

Researchers either consciously or by default in a statistical analysis drop the variables that have in-complete data. As an alternative to complete-case analysis, researchers may in a plausible value for the missing observations, such as using the mean of the observed cases on that variable [24]. But, here this research is focus on nearest distance based imputation technique. Besides, k-mean there are many statisticians recently advocated methods that are based on distributional models for the data (such as maximum likelihood and multiple imputation). More literature has been published in the statistical literature on missing data [25]-[27].

In [28], propose a new approach to clustering that divides the data features into observed features, which are known for all objects, and constraining features, which contain missing values. We generate a set of constraints based on the known values for the constraining feature. Based on our observation, we found high percentage of missing values in our dataset, therefore, we implemented similar technique of k-mean clustering for imputation to diminish the missing percentage in dataset to gain accuracy. We evaluated the result with all best models chosen after boosting and smoothing.

Fig. 1 below shows that k-mean imputation with ANN-linear model outperformed all other models in accuracy of error rate metric. Similarly, logistic regression slightly performed less ANN-Linear but, better than other model DT in error metric and AUC metric. Here, it was clearly notable that error metric is better metric for accuracy gain of the model in all comparisons. On the same note, we evaluated all models comparatively without imputation, which means contain more empty values.

No imputation results are shown in Fig. 2 below which comprehends the result for training dataset, it clearly reveals that DT shown significance over ANN. It was also evident that in test dataset ANN-Linear performed better than DT with no imputation technique; further below given table shows the test dataset comparison between models (Table 5).

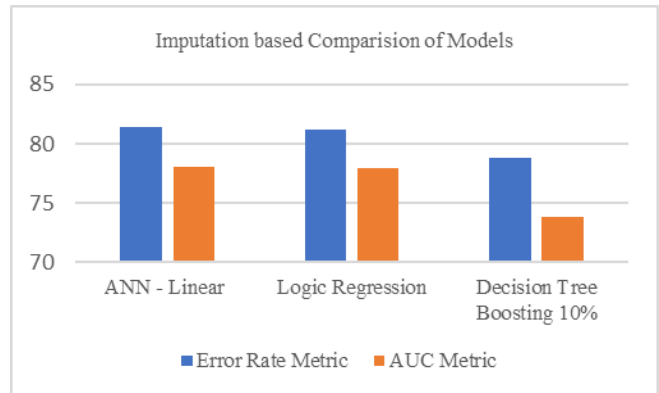


Fig. 1. The graph shows the comparisons between models with Error Rate Metric and AUC with imputation.

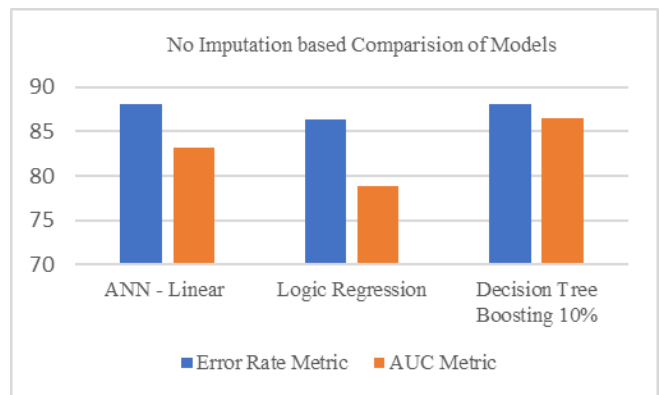


Fig. 2. The graph shows the comparisons between models with Error Rate Metric and AUC without Imputation.

TABLE V. NO IMPUTATION COMPARISONS

Model without Imputation	Error Rate		AUC	
	Training	Testing	Training	Testing
Decision Tree Boosting 10%	88.14	89.57	86.53	79.09
Logistic model	86.36	90.29	78.90	86.18
ANN-LINEAR	88.04	90.99	83.13	87.90

It was evident by all comparisons with imputation and without imputation technique that ANN performed better when data is in piece fully available for analysis into depth of model such as in case of imputation technique of K-mean. K-mean fill the values with nearest neighbor thus increase data availability that increases the accuracy rate of classification of ANN. But, in case of without imputation dataset lacks the volume in the form of missing values thus DT performed better than other model in training dataset but in test dataset ANN still performed better over low volume of dataset.

VI. CONCLUSION

This paper examines the major classification techniques in data mining and compares the performance of classification. The novel imputation method of k-mean improvised to avoid the data loss, for the first time, is presented for the similar dataset and its comparison with no imputation technique. Obviously, error rate is more sensitive than AUC, because, it is more appropriate criterion to measure the classification accuracy of models. Artificial neural networks linear model performs classification accurately than the other models in comparison to imputation and without imputation. Artificial neural networks model is also shown the best performance in no imputation test dataset but performed second last in case of training dataset. It shown more accuracy in case of availability of data like in imputation based dataset accuracy is better than all models. From the perspective of risk control, estimating the client risks without imputation is more meaningful than imputation on classification. It was also concluded that, artificial neural networks model is more reliable to be employed for credit scoring for bad and good clients' awareness. In future, big scoring and its impact can be tested with larger dataset using ANN ensemble or hybrid approach to cater the multiple tasks 1) feature selection; 2) classification. Because conventional credit score techniques inherit with narrow scope that is not perfect model because it only analyzes customer payment history but unable to justify customer characters, nature and credibility by the help of external source for instance, social media.

REFERENCES

[1] Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics.

[2] Atiya, A. F., & Parlos, A. G. (2000). New results on recurrent network training: unifying the algorithms and accelerating convergence. *IEEE transactions on neural networks*, 11(3), 697-709.

[3] Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36, 3302-3308.

[4] Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications*, 64, 36-55.

[5] DAVIS, Rober Hunter, D. B. Edelman, and A. J. Gammerman. "Machine-learning algorithms for credit-card applications." *IMA Journal of Management Mathematics* 4.1 (1992): 43-51.

[6] Frydman, Halina, Edward I. Altman, and DUEN-LI KAO. "Introducing recursive partitioning for financial classification: the case of financial distress." *The Journal of Finance* 40.1 (1985): 269-291.

[7] Zhang, D., Zhou, X., Leung, S. C., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, 37(12), 7838-7843.

[8] Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial finance*, 18(6), 15-26.

[9] West, David. "Neural network credit scoring models." *Computers & Operations Research* 27.11 (2000): 1131-1152.

[10] West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*, 32(10), 2543-2559.

[11] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480.

[12] Rokach, Lior, and Oded Maimon. *Data mining with decision trees: theory and applications*. World scientific, 2014, pp. 168.

[13] Witzany, Jiří. *Credit risk management and modeling*. Praha: Oeconomica, 2010, pp. 46.

[14] Loh, Wei-Yin. "Classification and regression tree methods." *Encyclopedia of statistics in quality and reliability* (2008), pp. 315-323

[15] Huang, S. C., & Day, M. Y. (2013, August). A comparative study of data mining techniques for credit scoring in banking. In *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on* (pp. 684-691). IEEE.

[16] V. S. Desai, J. N. Crook and G. A. Overstreet, "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment," *European Journal of Operational Research*, Vol. 95, No. 1, 1996, pp. 24-47. doi:10.1016/0377-2217(95)00246-4.

[17] H. Abdou, J. Pointon and A. E. Masry, "Neural Nets Versus Conventional Techniques in Credit Scoring in Egyptian Banking," *Expert Systems with Applications*, Vol. 35, No. 2, 2008, pp. 1275-1292. doi: 10.1016/j.eswa.2007.08.030.

[18] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3),1.

[19] Witzany, Jiří. *Credit risk management and modeling*. Praha: Oeconomica, 2010, pp. 46.

[20] Loh, Wei-Yin. "Classification and regression tree methods." *Encyclopedia of statistics in quality and reliability* (2008), pp. 315-323.

[21] Breiman, L., Friedman, J. H., & Olshen, R. A. (2009). *Stone, cj (1984) classification and regression trees*. Wadsworth, Belmont, California.

[22] Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15(03), 757-770.

[23] McFadden, D. and G. Leonard (1993). *Issues in the contingent valuation of environmental goods: Methodologies for data collection and analysis*. In *Contingent Valuation: a critical Assessment*, pp. 165-215. New York: North-Holland: Hausman.

[24] Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, 7(4), 353-383.

[25] Little, Roderick JA. "Regression with missing X's: a review." *Journal of the American Statistical Association* 87.420 (1992): 1227-1237.

[26] Little, Roderick JA, and Donald B. Rubin. "Statistical Analysis with Missing Data." (1987).

[27] Graham, John W., et al. "Analysis with missing data in prevention research." *The science of prevention: Methodological advances from alcohol and substance abuse research* 1 (1997): 325-366.

[28] Wagstaff, Kiri. "Clustering with missing values: No imputation required." *Classification, Clustering, and Data Mining Applications* (2004): 649-658.