Incremental Ensemble Learning Model for Imbalanced Data: a Case Study of Credit Scoring

Bui T. T. MY

Faculty of Mathematics and Statistics, College of Technology and Design, UEH University, Vietnam
Mathematical Economics Division, Ho Chi Minh University of Banking, Vietnam

Corresponding Author: Bui T. T. MY (email: mybtt@hub.edu.vn)
(Received: 23-Mar-2023; accepted: 15-May-2023; published: 30-Jun-2023)
DOI: http://dx.doi.org/10.55579/jaec.202372.407

Abstract. *Imbalanced data is a challenge for classification models. It reduces the overall performance of traditional learning algorithms. Besides, the minority class of imbalanced datasets is misclassified with a high ratio even though this is a crucial object of the classification process. In this paper, a new model called Lasso-Logistic ensemble is proposed to deal with imbalanced data by utilizing two popular techniques, random over-sampling and random under-sampling. The model was applied to two real imbalanced credit data sets. The results show that the Lasso-Logistic ensemble model offers better performance than the single traditional methods, such as random over-sampling, random under-sampling, Synthetic Minority Oversampling Technique (SMOTE), and cost-sensitive learning.*

Keywords

Classification, Credit scoring, Imbalanced data, Lasso-Logistic, Resampling techniques.

1. Introduction

In classification, imbalanced data is a big challenge [1]. A data set is considered imbalanced if most of the examples belong to one subset and a few examples belong to the others [2]. In binary classification, if the difference between the quantities of two classes is too large, the true positive rate of conventional models is usually low. The reason is that most classification models are designed to maximize the accuracy metric. It leads to the misclassification of the minority class although this is the crucial object of the classification process. Besides, empirical studies also showed that the overall performance of classification models (measured by AUC metric) was impacted negatively by the imbalanced status of training sets [2]. In short, imbalanced data can be considered as the most factor causing an ineffective performance of classification models, especially in the minority class.

Credit scoring is a typical case of imbalanced classification. The simplest form of credit scoring is to discriminate borrowers into 'bad' or 'good' based on their creditworthiness. Since there are many regulations to screen bad customers, the number of the bad is always far less than the good. Credit scoring is necessary for both the banks and the customers. For

banks, credit scoring provides useful information to make appropriate decisions for credit granting to customers. A misclassification may lead the banks to enormous losses [3]. For customers, credit scoring helps them to access loans with reasonable interest rates and suitable loan periods. Thus, credit scoring always attracts many concerns in financial and banking institutions.

Most of studies on credit scoring focused on utilizing statistical and classification techniques [4], such as Discriminant analysis [5, 6]; Logistic regression (LR) [5–10]; Neural networks (NN) [6, 11–13]; Decision tree (DT) [7, 13, 14]; Support vector machine (SVM) [7, 15–17]. Recently, there has been a shift from single models to ensemble ones in credit scoring, for instance, Random forest (RF) [2, 18, 19]; bagging tree [20]; boosting tree [21]. The ensemble models are the model based on similar learners with different parameters. Empirical studies showed that ensemble models offered much better performance than the single since they could explore more potential information of the training set [20, 22]. However, ensemble models meet a trade-off between accuracy and interpretation which are the two requirements of a credit scoring model. Most ensemble models have a 'black box' computation process, since it is too difficult to determine important predictors for predicted results.

Among single models, SVM and ANN are the representatives of 'black box' type, while LR and DT are the 'transparent' one. Besides, DT usually shows a very high variance in predicted results if there is a small change in the training set. Therefore, LR is still favored in credit scoring [4]. Furthermore, some authors concluded that credit scoring models based LR showed higher accuracy than the Discriminant analysis method [5, 10] while LR is not inferior to the intelligent methods [2, 23]. This fact suggests an ideal of a credit scoring ensemble model based on Logistic regression, which can promote the advantages of the traditional methods and limit the disadvantages of the machine learning ones.

Returning to the imbalanced data in credit scoring, although ensemble models can increase the performance measures, they do not directly deal with this problem. The popular approaches to imbalanced data in credit scoring are resampling techniques and cost-sensitive learning. Despite the fact that these approaches have advantages and disadvantages affecting the response of credit scoring models. For these above reasons, in this paper, a new ensemble model for credit scoring based on LR is proposed with two purposes, including increasing performance measures through handling imbalanced data and showing the important level of predictors.

The rest of this article is organized as follows. Section 2 provides the related works of the methods for balancing data in credit scoring and the basic knowledge of LR. Section 3 introduces the algorithm for the new credit scoring ensemble model and the empirical results. Finally, section 4 is the summary and conclusions of the paper.

2. Related works

In binary imbalanced classification, the minority class is named the 'positive class' with the label denoted '1'. This is the crucial object that needs to identify in the classification task. The majority class is called the 'negative class' with the label '0'. In credit scoring, the bad and the good customers form the positive and the negative class, respectively.

2.1. Methods for imbalanced data in credit scoring

1) Cost-sensitive learning

Cost-sensitive learning (CSL) is the most favored method for imbalanced data in credit scoring [24–26]. The basic idea of CSL is that every misclassified result causes a loss. Furthermore, the misclassification of the bad into the good is more serious than the one of the good into the bad [25]. Therefore, each misclassified result is assigned a level of loss and the credit scoring model should be constructed so that the total loss of the classification process is minimal. If symbols $C(+,-)$ and $C(-,+)$ are the loss when predicting a bad customer to the good one and the good to be the bad, respectively, then $C(+,-)$ is greater than $C(-,+)$. Previous stud-

Tab. 1: Cost matrix in credit scoring

	Predicted Positive	Predicted Negative
Actual Positive	0	C
Actual Negative	1	0

ies set $C(-,+)$ a unit and $C(+,-)$ a constant number C greater than 1 (see the cost-matrix in Table 1). However, this approach is subject to some controversy because the difference between the losses $C(+,-)$ and $C(-,+)$ is set by the researchers' subjective intention [27]. Besides, the real difference is not a constant since it depends on the banks and the customer attributes. These arguments reduce the practical values of CSL.

2) Re-sampling methods

This approach proposes a series of steps of re-sampling techniques to balance the original data set. These techniques are easy to handle and do not depend on the learners training the classification models. The popular techniques in credit scoring are random under-sampling (RUS), random over-sampling (ROS), and Synthetic Minority Over-sampling (SMOTE).

- RUS decreases the quantity of the majority class by removing randomly its examples. Thus, RUS reduces sample size and shortens the computation time of learners. However, important information in the original data set may be omitted by RUS.
- ROS increases the quantity of the minority class by randomly repeating its examples. This technique leads to a data set with a greater sample size and longer computation time. Both ROS and RUS can be employed at an arbitrary ratio to lighten or clear the imbalanced status. Experimental credit scoring studies showed that ROS was more effective than RUS [2,23,28]. However, ROS can duplicate outliers, which might lead to overfitting models [22, 29].
- SMOTE [30] is an innovation of ROS. Instead of purely repeating the examples of the minority class, SMOTE generates synthetic positive examples which are in the neighborhood of each original positive example. Empirical studies showed that SMOTE was better than RUS and ROS [22, 31]. However, some authors criticized SMOTE due to the possibility of a more seriously overlapping status between classes which reduces the performance of classification models [23, 32].

2.2. Logistic regression

LR models the relationship between the probability of belonging to the positive class and the predictors via the following formula

$$P(Y=1|X) = \frac{e^{\beta_0+\beta X'}}{1+e^{\beta_0+\beta X'}} = f(\beta_0+\beta X') \quad (1)$$

where, X' is the column matrix of P predictors; $\beta = (\beta_1, ..., \beta_p) \in \mathbb{R}^p$ and β_0 are the parameters, β shows the impacts of the predictors X on the conditional probability $P(Y=1|X)$; the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$ has the range [0, 1].

Suppose that the data set consists of n independent examples, the parameters in (1) can be estimated by Maximum Likelihood method with the objective function:

$$\begin{aligned} &\log\left(P(Y|\,X,\beta)\right) = \log\left(\prod_{i=1}^{n} P(Y_i|X_i,\beta)\right) \\ &= \sum_{i=1}^{n}\left(\log(1+e^{\beta_0+\beta X'}) - Y_i\left(\beta_0+\beta X'\right)\right) \\ &:= l\left(Y|X,\beta\right) \end{aligned} \quad (2)$$

With a new sample X^*, it is classified into the positive class if and only if the conditional probability $P(Y=1|X^*)$ is not less than a given threshold.

A modification of LR is Lasso-Logistic regression (LLR), in which the main problem is finding $\left(\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_p}\right)$ satisfying:

$$\begin{cases} \max\limits_{\beta}\left(l\left(Y|X,\beta\right)\right) \\ \sum\limits_{j=1}^{p}|\beta_j| \le t \end{cases} \quad (3)$$

where, $t > 0$ is a tuning parameter.

If t is sufficiently large, the constraint imposing on the parameters is not strict, the solution of (3), $\widehat{\beta}_j\ (j \in \overline{1,p})$, are the same as the one of (2). On the contrary, if t is very small, the magnitude of $\widehat{\beta}_j\ (j \in \overline{1,p})$ is shrunk. Then, due to the property of the absolute function, some of $\widehat{\beta}_j$ are zero. Therefore, the constraint on $\beta_j\ (i \in \overline{1,p})$ in (3) plays a role of a feature selection method: only the predictors relevant to the response, which are corresponding to non-zero $\widehat{\beta}_j$, are retained in the fitted model.

Based on the theory of convex optimization, problem (3) is equivalent to:

$$\min_{\beta} \left[-\left(l(Y|X,\beta)\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right] \qquad (4)$$

where, λ is a penalty level, corresponding 1-1 to the tuning parameter t in (3). If λ is zero, the solution of LLR is exactly equal to LR's solution in (2). Otherwise, if λ is sufficiently large, the solution of LLR is zero. For values of λ between the two extremes, LLR gives a solution with some of $\widehat{\beta}_j$ zero, thus some predictors are excluded from the model. The values of λ are surveyed on a grid search to select the best based on criteria AIC, BIC, or cross-validation procedure. With a given λ, problem (4) is solved by the coordinate descent algorithm and proximal – Newton interaction(see [33] for more details).

Besides being a feature selection method, the predictive power of LLR is better than LR in empirical studies [20, 34].

3. The proposed ensemble credit scoring model

The paper proposes an ensemble credit scoring model expected to solve the imbalanced issue and offer the importance of the predictors. With the first expectation, the combination of ROS and RUS generates a family of balanced data sets with increasing quantities. They are the training sets of an ensemble model based on the LLR learner. With the second expectation, both LLR and LR are good choices. However, LLR has the ability to shrinkage without using p-values: the irrelevant predictors will be removed from the model. Meanwhile, LR shows the significant level of predictors through p-value which has been criticized for misunderstandings and misusing [35].

3.1. Algorithm for ensemble model

The proposed ensemble model called Lasso-Logistic ensemble (LLE) comprises several steps which are shown in Table 2. Some explanations of the algorithm for LLE are following.

- Consider a training data T consisting of p predictors. MI and MA are the minority and majority class of T, respectively. Besides, consider a threshold α to distinguish the two classes and B, the number of sub-models in LLE.
- Firstly, calculate D, the difference between the quantities of MA and MI; and $S_i, (i = 1, ..., B)$, the number of more positive examples duplicated in each iteration (Step 1).
- For every value of i which varies from 1 to B, combine RUS and ROS to generate a balanced data set T_i (Steps 3-5). Then, apply Lasso-Logistic learner on T_i to get the fitted model LL_i (Step 6). Another output in this stage is the binary vector $v_i = (v_{ij})_{j=\overline{1,p}}$ where v_{ij} is zero if the j^{th} predictor is excluded from LL_i; otherwise v_{ij} is 1 (Step 7).
- Finally, the overall predicted status of a new example is the majority of predicted results of B sub-models. Besides, the important level of a predictor is the number of sub-models in which this predictor is present. For convention, LLE consists of B sub-models denoted LLE(B).

3.2. Empirical data

Two data sets, Vietnamese (VN) and German (GER) data, are used to perform and verify the effectiveness of LLE. The Vietnamese data set is

Tab. 2: Algorithm for Lasso-Logistic ensemble - LLE algorithm

Inputs:	B: the number of sub-models in LLE; α: the threshold; T: training data set with p predictors, $T = MI \cup MA, \|MI\| < \|MA\|, MI \cap MA = \emptyset$.
1.	$D = \|MA\| - \|MI\|$, $S_i = \text{round}(\frac{iD}{B} \times 100\%), (i = 1, ..., B)$
2.	$i = 1$ do:
3.	Apply ROS to get a new positive class MI_i with $\|MI_i\| = \|MI\| + S_i$.
4.	Apply RUS to get a new negative class MA_i, satisfying $\|MA_i\| = \|MI_i\|$.
5.	$T_i = MI_i \cup MA_i$
6.	$LL_i \leftarrow$ Lasso-Logistic regression(T_i), where $LL_i(x) = 1 \Leftrightarrow P(Y = 1\|x) \geq \alpha$.
7.	$v_i = (v_{ij})_{j=\overline{1,p}}$, where $v_{ij} = I\left(\beta_j^{(i)} \neq 0\right)$ ($\beta_j^{(i)}$ is the parameter corresponding to j^{th} predictor in LL_i).
8.	$i \leftarrow i + 1$
9.	Repeat from Step 2 to Step 8 until $i = B + 1$.
Outputs:	LLE model: $LLE(x) = \begin{cases} 1, \text{ if } \sum_{i=1}^{B} LL_i(x) \geq 0.5B \\ 0, \text{otherwise} \end{cases}$ The important level of j^{th} predictor: $IP_j = \sum_{i=1}^{B} v_{ij}$, $j \in \overline{1,p}$.

Tab. 3: Characteristics of empirical data sets

Data sets	Size	# pos[a]	IR	# feat[b]
GER	1,000	300	2.33	20
VN	3,232	454	6.12	10

[a]: The quantities of positive class
[b]: The number of features.

from a commercial bank in Vietnam. It consists of 3232 observations, including 454 default and 2778 non-default customers. The input features are expressed by 10 nominal predictors, such as Total asset, Borrower type, Loan type, Duration, Credit history, Draw-down amount, Interest rate, Purposes, Married status, and Liquidity of collateral. Besides, the Vietnamese data suffered from a quite high imbalanced issue with the imbalanced ratio is:

$$IR = \frac{|MA|}{|MI|} = \frac{2778}{454} = 6.12$$

The German data set, public in the UCI machine learning repository, is very popular with the credit scoring literature [36]. For convenience, the input features of German data set are denoted $A_1, A_2, ..., A_{20}$. The details of German data set can be found in the UCI website [1].

[1] https://archive.ics.uci.edu/ml/datasets

Table 3 summarizes some characteristics of the two empirical data sets.

All numerical features of data sets are standardized to have zero mean and deviation unit.

3.3. Performance metrics

The area under the ROC curve (AUC), Kolmogorov-Smirnov statistic (KS), F-measure (F_1 and F_2), and G-mean are utilized to evaluate the performance of considered models. These criteria are suitable for imbalanced classification [2, 23, 37]. The values of these metrics are expected as high as possible.

AUC is the area under the ROC curve (Receiver Operating Characteristics) which shows the relationship between the series of FPR and TPR across thresholds. AUC is the expected true positive rate, averaged over all false positive rates with all possible thresholds [38].

KS measures the degree of separation between the positive and negative predicted results. The KS metric is defined as follows:

$$KS = \max_{\alpha} \left(TPR(\alpha) - FPR(\alpha)\right) \qquad (5)$$

where $TPR(\alpha)$ and $FPR(\alpha)$ are the true positive rate and false positive rate corresponding to

threshold α. AUC and KS do not depend on the threshold of distinguishing the two classes, thus they express the overall performance of classifiers.

Contrary to AUC and KS, F-measure, and G-mean are related to a specific threshold. F-measure denoted F_β, is the weighted harmonic mean of the Recall (another name for the true positive rate) and the Precision. If β is 1, the interest of classification model in Recall and Precision is equal. If β is greater than 1, the concern is toward Recall. F-measure is usually utilized in credit scoring [39, 40]. The popular terms of F-measures are F_1 and F_2.

$$F_\beta = \frac{(1+\beta^2)Precision.Recall}{\beta^2 Precision + Recall} \tag{6}$$

G-mean is the geometry mean of the true positive rate (TPR) and true negative rate (TNR).

$$G - mean = \sqrt{TPR.TNR} \tag{7}$$

3.4. Computation process

On each data set, LLE(B)s are employed with many values of B to find the optimal value B^* which may be different values corresponding to data sets. The process of finding the optimal B^* is carried out according to the procedure in Table 4. After that, LLE(B^*) is conducted on German and Vietnamese data sets, while the Lasso-Logistic regression with popular balanced methods, such as CSL, RUS, ROS, and SMOTE are employed. For the CSL method, constant C is considered with some values, such as 4, 8, and 12 and the corresponding models are denoted CSL(4), CSL(8), and CSL(12), respectively. The performance of these models is evaluated based on AUC, KS, F_1, F_2, and G-mean from the testing data. The general computation process of all considered models is shown in Table 5.

Tab. 4: Process of finding the optimal B

Steps	Contents
1.	Consider data set S and a value of B.
2.	Divide randomly S into training data T (70%) and testing data TE (30%).
3.	Randomly split the training data T into ten equal-sized parts: $T_1, ..., T_{10}$.
4.	Set $k = 1$.
5.	On $T \backslash T_k$, construct the ensemble classifier LLR(B).
6.	On T_k, fix LLR(B) to get the performance measures AUC_k, KS_k, and F_{1k}.
7.	$k \leftarrow k + 1$.
8.	Repeat from Step 5 to 7 until $k = 11$.
9.	Calculate the cross-validation values of performance measures: $AUC_{cv} = \frac{1}{10}\sum_{k=1}^{10} AUC_k$, $KS_{cv} = \frac{1}{10}\sum_{k=1}^{10} KS_k$, and $F_{1cv} = \frac{1}{10}\sum_{k=1}^{10} F_{1k}$.
10.	Repeat from Step 2 to Step 9 twenty times and record the averaged values of $AUC_{cv}, KS_{cv}, F_{1cv}$, denoted $\overline{AUC}_{cv}^{(B)}, \overline{KS}_{cv}^{(B)}, \overline{F_1}_{cv}^{(B)}$.
11.	Repeat the steps from 1 to 10 with other values of B.
12.	Compare $\overline{AUC}_{cv}^{(B)}, \overline{KS}_{cv}^{(B)}, \overline{F_1}_{cv}^{(B)}$ corresponding to several B to find out the optimal B^* which offers the highest ones.

Tab. 5: Computation process of the considered models

Steps	Contents
1.	Divide randomly the data set into training (70%) and testing data (30%).
2.	On the training data, construct the opitmal ensemble LLE(B*), LLR without any balanced method, LLR with the re-sampling techniques RUS, ROS, SMOTE, and CSL(4), CSL(8), CSL(12) models.
3.	On the testing set, calculate AUC, KS, F_1, F_2, and G-mean of the above fitted models.
4.	Repeat the steps from 1 to 3 twenty times and average all the performance metrics.

3.5. Empirical results

1) The optimal proposed ensemble classifier

The performance of LLE(B) depends on B, which is the number of sub-models of the ensemble. For the proposed classifier algorithm, the upper bound of B is D, the difference between the quantities of the majority and minority classes. When B is too large, the values of S_i determined in Step 1 of Table 2 are very similar to each other. Therefore, the differences between $T_i, (i = 1, ..., B)$ which are new balanced data generated in Step 5 of Table 2 are not significant. Then, the sub-models of the LLE(B) are homogeneous. That makes the predicted results from sub-models similar to each other and similar to the output of the ensemble. It means the ensemble classifiers cannot leverage the collective power of sub-classifiers shown through the diversity of sub-classifiers. Meanwhile, the larger B is, the longer the computation time is. Thus, large B only wastes time and does not raise the effectiveness of the classification process. In summary, the process of finding the optimal B^* does not focus on large Bs.

According to the averaged cross-validation values of AUC, KS, and F_1 of LLE(B)s, which are shown in Figure 1, the optimal value B^* of each data set is determined as follows:

- On the German data set, at $B = 11$, LLE(11) has the highest cross-validation of KS and F_1. Thus, the optimal B^* is 11.

- On the Vietnamese data set, LLE(B) gets the highest AUC at $B = 3$, the highest KS at $B = 15$, and the highest F_1 at $B = 5$. Considering the trade-off between the effectiveness and the computation time, $B = 15$ is not a good choice. On the other hand, at $B = 5$, all three metrics are quite good, while at $B = 3$, the KS and F_1 are really lower than the ones at the other B. Thus, the optimal B^* is 5.

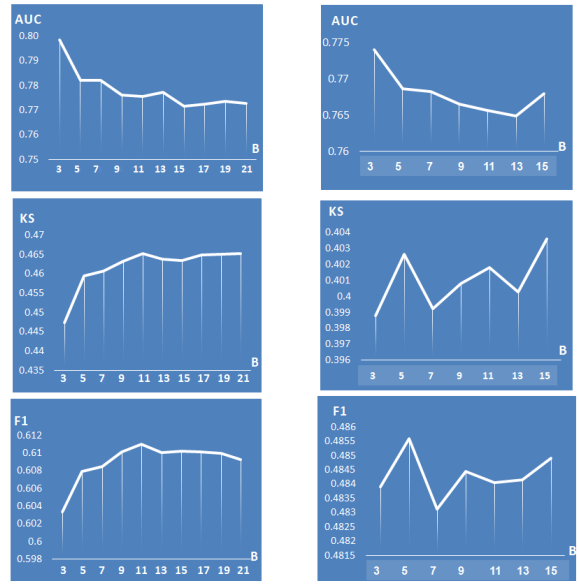

Fig. 1: Mean cross-validation performance metrics of LLE(B)s.

2) Important level of predictors

A minor output of the algorithm for LLE is the important level of predictors. In this experiment, the computation protocol of LLE(B^*) is applied 20 times, so the important level of each predictor is summed up 20 times. For the German data set, if a predictor is always present in all sub-models of LLB(11) through 20 times, its important level will be $IP = 220$. For the Vietnamese data set, the value is $IP = 100$ for the predictor always included in all sub-models of LLE(5) after 20 times applied. We scale the important level to the maximum value of 100 for two data sets in order to evaluate advantageously.

In German data set, A_1 (Status of existing checking account), A_3 (Credit history), and A_{12} (Property) are the most crucial predictors which are always present in every sub-classifier of LLE(11). Then, the features such as A_6 (Savings account/bonds), A_8 (Installment rate in percentage of disposable income), A_7 (Present employment since), A_{20} (foreign worker), A_{14} (Other installment plans), and A_{13} (Age in

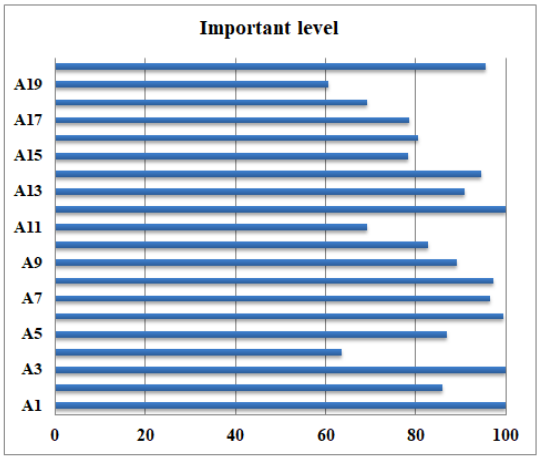

Fig. 2: The important level of predictors of German data.

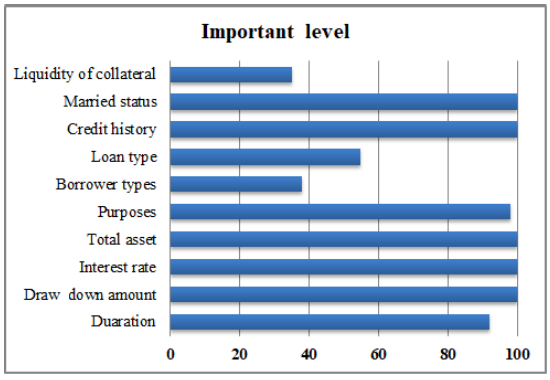

Fig. 3: The important level of predictors of Vietnamese data.

Tab. 6: Performance metrics on German data set

Models	AUC	KS	F_1	F_2	G-mean
LLE(11)	**.8113**	.4424	**.6101**	**.7068**	**.7173**
LLR	.7712	.4463	.5931	.6642	.6874
RUS-LLR	.7654	.4349	.5845	.6597	.7023
ROS-LLR	.7702	.4453	.5940	.6622	.7111
SMOTE-LLR	.7746	**.4554**	.5934	.6579	.7109
CSL(4)	.7712	.4463	.5894	.6960	.6938
CSL(8)	.7712	.4463	.5672	.6974	.6606
CSL(12)	.7712	.4463	.5568	.6942	.6458

The bold is the highest in each column.

Tab. 7: Performance metrics on Vietnamese data set

Models	AUC	KS	F_1	F_2	G-mean
LLE(5)	**.7678**	.4034	**.4889**	**.6611**	**.6893**
LLR	.7282	**.4274**	.3078	.1943	.5026
RUS-LLR	.7173	.3888	.4381	.6188	.6664
ROS-LLR	.7246	.3971	.4445	.6281	.6726
SMOTE-LLR	.7245	.4001	.4432	.6268	.6713
CSL(4)	.7282	**.4274**	.4445	.4389	.5658
CSL(8)	.7282	**.4274**	.4888	.6254	.6874
CSL(12)	.7282	**.4274**	.4760	.6278	.6854

The bold is the highest in each column.

years) in descending order are also important since their important level are greater than 90.

In the Vietnamese data set, Married status, Credit history, Total assets, Interest rate, and Draw-down amount are the most crucial predictors. Figure 2 and 3 shows the whole predictors' importance of German and Vietnamese data set.

3) The effectiveness of the proposed ensemble classifier

LLE(11) and LLE(5) are the optimal ensemble model for German and Vietnamese data set, respectively. In comparison to other balanced methods combined with Lasso-Logistic, LLE(B^*) completely outperformed in terms of AUC, F_1, F_2, and G-mean on experimental data. The testing performance metrics of LLE(B^*) were shown in Tables 6 and 7.

By AUC, LLE(B*) completely won the others on two data sets. The differences in AUC between LLE($_B^*$) and the other were significant. Furthermore, by threshold-based metrics, LLE(B^*) impressively boosted F_1 and F_2 in comparison with RUS, ROS, and SMOTE.

In contrast, the re-sampling techniques did not improve the overall performance metrics (AUC and KS) on two data sets. On German data, RUS, ROS, and SMOTE did not raise F_1, F_2 of the original version of LLR. Meanwhile, CSL method even pulled F_1 and G-mean down. On Vietnamese data, the balanced meth-

ods seemed to be successful in boosting F_1, F_2, and G-mean although they were not as effective as the proposed ensemble classifiers.

With the CSL method, the three last models remained the original algorithm of LLR but changed the threshold to get the minimal total cost of loss. Hence, the AUC and KS of these models were the same as LLR's, but other threshold-based metrics had a great innovation. Despite the fact that the effectiveness of CSL was still less than LLE(B^*).

In general, the outperformance of LLE(B^*) in F-measure means that LLE(B^*) showed the best trade-off between precision and recall, which are interested in credit scoring application. Similarly, the greatest G-mean of LLE(B^*) suggested that LLE(B^*) had the most ability to balance the true positive rate and true negative rate.

4. Conclusions

The paper proposed a credit scoring ensemble model based on LLR called LLE. This model can solve the imbalanced status in the training data to improve the performance metrics of LLR while can show the important level of predictors to the final response. The combination of ROS and RUS has proved to prevail over these methods when applied individually. Not only that, the operation of an ensemble model based on the combination of ROS and RUS has triumphed over other classical balanced methods, such as SMOTE and CSL in the performance measures AUC, F_1, F_2, and G-means. The proposed ensemble model should perform on more data sets to have a robust conclusion on its effectiveness. Besides, the empirical result suggests a further study on the choice of the optimal B^* of LLE.

References

[1] Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

[2] Brown, I. & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453.

[3] Abdou, H.A. & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, *18*(2-3), 59–88.

[4] Onay, C. & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*.

[5] Desai, V.S., Crook, J.N., & Overstreet Jr, G.A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European journal of operational research*, *95*(1), 24–37.

[6] Etheridge, H.L. & Sriram, R.S. (1997). A comparison of the relative costs of financial distress models: artificial neural networks, logit and multivariate discriminant analysis. *Intelligent Systems in Accounting, Finance & Management*, *6*(3), 235–248.

[7] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*, 627–635.

[8] Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(3), 133–150.

[9] Chen, K., Yadav, A., Khan, A., & Zhu, K. (2020). Credit Fraud Detection Based on Hybrid Credit Scoring Model. *Procedia Computer Science*, *167*, 2–8.

[10] 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*(3), 757–770.

[11] Altman, E.I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of banking & finance*, *18*(3), 505–529.

[12] West, D. (2000). Neural network credit scoring models. *Computers & operations research*, *27*(11-12), 1131–1152.

[13] Yobas, M.B., Crook, J.N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, *11*(2), 111–125.

[14] Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications*, *78*, 225–241.

[15] Bellotti, T. & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert systems with applications*, *36*(2), 3302–3308.

[16] Huang, Z., Chen, H., Hsu, C.J., Chen, W.H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, *37*(4), 543–558.

[17] Huang, C.L., Chen, M.C., & Wang, C.J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, *33*(4), 847–856.

[18] Van Sang, H., Nam, N.H., & Nhan, N.D. (2016). A novel credit scoring prediction model based on Feature Selection approach and parallel random forest. *Indian Journal of Science and Technology*, *9*(20), 1–6.

[19] Zhang, H., He, H., & Zhang, W. (2018). Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. *Neurocomputing*, *316*, 210–221.

[20] Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, *10*(2), e0117844.

[21] Roncalli, T. (2020). *Handbook of Financial Risk Management*. CRC Press.

[22] Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91.

[23] Marqués, A.I., García, V., & Sánchez, J.S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, *64*(7), 1060–1070.

[24] Xiao, J., Xie, L., He, C., & Jiang, X. (2012). Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Systems with Applications*, *39*(3), 3668–3675.

[25] Xiao, J., Zhou, X., Zhong, Y., Xie, L., Gu, X., & Liu, D. (2020). Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. *Knowledge-Based Systems*, *189*, 105118.

[26] Zhang, L., Ray, H., Priestley, J., & Tan, S. (2020). A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. *Journal of Applied Statistics*, *47*(3), 568–581.

[27] Moepya, S.O., Akhoury, S.S., & Nelwamondo, F.V. (2014). Applying cost-sensitive classification for financial fraud detection under high class-imbalance. In *2014 IEEE international conference on data mining workshop*, IEEE, 183–192.

[28] He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105–117.

[29] Barandela, R., Valdovinos, R.M., Sánchez, J.S., & Ferri, F.J. (2004). The imbalanced training sample problem: Under or over sampling? In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural*

and syntactic pattern recognition (SSPR), Springer, 806–814.

[30] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

[31] Shen, F., Zhao, X., Li, Z., Li, K., & Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*, *526*, 121073.

[32] Kaur, P. & Gosain, A. (2018). Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. In *ICT Based Innovations*, Springer, 23–30.

[33] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

[34] Li, Q. *et al.* (2019). Logistic and SVM credit score models based on lasso variable selection. *Journal of Applied Mathematics and Physics*, *7*(05), 1131.

[35] Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, volume 45, Elsevier, 135–140.

[36] Hofmann, H., Statlog (German Credit Data) Data Set.

[37] Batista, G.E., Prati, R.C., & Monard, M.C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

[38] Ferri, C., Hernández-Orallo, J., & Flach, P.A. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 657–664.

[39] Abdoli, M., Akbari, M., & Shahrabi, J. (2023). Bagging Supervised Autoencoder Classifier for credit scoring. *Expert Systems with Applications*, *213*, 118991.

[40] Akay, M.F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, *36*(2), 3240–3247.

About Authors

Bui T. T. MY is now a Ph.D. Candidate in Statistics major at the Statistical Mathematics Faculty, UEH University. She is a lecturer at the Department of Mathematical Economics, Ho Chi Minh City University of Banking. Her research interests are classification models and imbalanced data in classification.