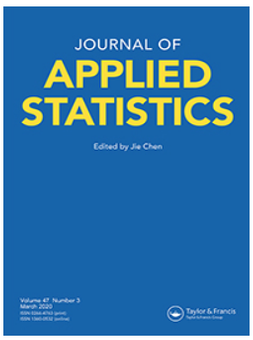

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data

Lili Zhang, Herman Ray, Jennifer Priestley & Soon Tan

To cite this article: Lili Zhang, Herman Ray, Jennifer Priestley & Soon Tan (2020) A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data, Journal of Applied Statistics, 47:3, 568-581, DOI: 10.1080/02664763.2019.1643829

To link to this article: https://doi.org/10.1080/02664763.2019.1643829

Published online: 23 Jul 2019.

JOURNAL OF APPLIED STATISTICS
2020, VOL. 47, NO. 3, 568–581
https://doi.org/10.1080/02664763.2019.1643829

Taylor & Francis
Taylor & Francis Group

APPLICATION NOTE

A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data

Lili Zhang [a], Herman Ray [b], Jennifer Priestley[b] and Soon Tan[c]

[a]Analytics and Data Science Ph.D. Program, Kennesaw State University, Kennesaw, Georgia, USA; [b]Analytics and Data Science Institute, Kennesaw State University, Kennesaw, Georgia, USA; [c]Ermas Consulting Inc., Alpharetta, Georgia, USA

ABSTRACT

Training classification models on imbalanced data tends to result in bias towards the majority class. In this paper, we demonstrate how variable discretization and cost-sensitive logistic regression help mitigate this bias on an imbalanced credit scoring dataset, and further show the application of the variable discretization technique on the data from other domains, demonstrating its potential as a generic technique for classifying imbalanced data beyond credit socring. The performance measurements include ROC curves, Area under ROC Curve (AUC), Type I Error, Type II Error, accuracy, and F1 score. The results show that proper variable discretization and cost-sensitive logistic regression with the best class weights can reduce the model bias and/or variance. From the perspective of the algorithm, cost-sensitive logistic regression is beneficial for increasing the value of predictors even if they are not in their optimized forms while maintaining monotonicity. From the perspective of predictors, the variable discretization performs better than cost-sensitive logistic regression, provides more reasonable coefficient estimates for predictors which have nonlinear relationships against their empirical logit, and is robust to penalty weights on misclassifications of events and non-events determined by their apriori proportions.

ARTICLE HISTORY
Received 13 May 2018
Accepted 10 July 2019

KEYWORDS
Class imbalance; variable discretization; cost-sensitive logistic regression; discrimination ability; credit scoring

1. Introduction

Class imbalance problems refer to a class of problems related to classifying imbalanced data where many more observations are labeled by the majority class than the minority class [1,11]. In practice, the minority class is usually the class of interest, such as fraud in the fraud detection problem [31], malignance in the breast cancer diagnosis problem [23], delinquency in the credit scoring problem [3], sinus bradycardia in the arrhythmia analysis [12], and poor quality in the product quality inspection [5].

However, when trained on imbalanced data, most standard statistics and machine learning models are heavily biased towards the majority class (i.e. non-events) and severely misclassify the minority class (i.e. events) [38], caused by their assumptions of equal target

CONTACT Herman Ray hray8@kennesaw.edu Analytics and Data Science Institute, Kennesaw State University, Kennesaw, Georgia, USA

class distribution [17] and maximizing overall accuracy [33]. Models with poor event discrimination are less useful and generate costs associated with Type II errors (money, reputation, health, etc.).

To solve these problems more efficiently, researchers and practitioners have made efforts from various perspectives, such as data sampling [22], feature selection [25,29], cost-sensitive learning [2,20,24], ensemble learning [4], and kernel-based learning [8], with the considerations of concrete problem characteristics.

Previous research has not considered variable discretization as a generic technique for class imbalance problems. In this paper, we empirically explore the effects of variable discretization on classifying imbalanced data and compare it with cost-sensitive logistic regression models. Variable discretization and cost-sensitive logistic regression are studied for their high interpretability and computational efficiency. A credit scoring dataset is used in the case study. The goal is to predict the probability of a debtor's default or delinquency. The proportion of delinquency observations is only 6.68%. We provide a detailed descriptive study on how variable discretization and cost-sensitive logistic regression help mitigate the model bias and/or variance on an imbalanced credit scoring data. The variable discretization technique is further applied on two datasets from other domains (i.e. biology, business) to demonstrate its potential for use in a wide range of fields.

The paper is structured as follows. In Section 2, related work is reviewed. In Section 3, the data is explored and discretized. In Section 4, the models on the credit scoring dataset are developed, evaluated, and compared. In Section 5, the performance of variable discretization is examined on two datasets from other domains. In Section 6, conclusions and future work are discussed.

2. Related work

A comprehensive review on the foundations, algorithms, and applications of imbalanced learning was conducted by He *et al.* in 2013 [15]. It summarized the previous research in five categories, including sampling methods, cost-sensitive methods, kernel-based learning methods, active learning methods, and one-class learning methods. It also suggested to evaluate models based on both curve-based measures (e.g. ROC curve, AUC) and single-value measures (e.g. Type I Error, Type II Error, F1 score, G-mean), considering that some traditional performance measures (e.g. accuracy) did not serve as a good indicator of discrimination abilities of models [34]. In an imbalanced credit scoring study by Wang *et al.*, AUC and F-measure (i.e. F1 score) were used as model performance metrics [39].

In 2001, King proposed the weighted log-likelihood function in Equation (2) for the logistic regression in rare events data. Compared with the standard log-likelihood function in Equation (1), Class 1 Weight (W_1) and Class 0 Weight (W_0) were added to penalize the misclassifications of events and non-events differently. W_1 and W_0 were determined by the estimated population proportion of events τ and the sample proportion of events $\bar{y}$.

$$\ln L(\beta \mid y) = \sum y_i \ln(\pi_i) + \sum (1 - y_i) \ln(1 - \pi_i) \tag{1}$$

$$\ln L_W(\beta \mid y) = W_1 \sum y_i \ln(\pi_i) + W_0 \sum (1 - y_i) \ln(1 - \pi_i) \tag{2}$$

where

$$\pi_i = \frac{1}{1 + \exp -(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}, \quad W_1 = \frac{\tau}{\bar{y}}, \quad W_0 = \frac{1-\tau}{1-\bar{y}} \tag{3}$$

The weighted logistic regression in Equation (2) is referred to as class-dependent cost-sensitive logistic regression [28]. Bahnsen *et al.* proposed a different version of cost-sensitive logistic regression, called example-dependent cost-sensitive logistic regression [2], where each example (i.e. observation) in the log-likelihood function was associated with a user-defined constant misclassification cost weight based on domain knowledge. Deng and Maher proposed determining each observation's cost weight by Gaussian kernel function [6,26,27], resulting in very high computational complexity $O(n^3)$ and limiting its application on big data.

Different from cost-sensitive logistic regression which has been widely used, the variable discretization method has not been considered for addressing class imbalance problems, although it has been widely used as a domain-specific standard technique in credit scoring. This technique creates more powerful and interpretable predictors from continuous (i.e. interval) data. Dougherty *et al.* reviewed existing variable discretization methods, compared three of them (i.e. equal width interval, entropy-based, and purity-based) in depth on 16 datasets, and found that the global entropy-based one performed the best on average [10]. For entropy-based discretization methods, the evaluation measures include: class information entropy, Gini, dissimilarity, and the Hellinger measure [21]. For the scoring problem, one commonly used variable discretization method is called the optimal binning, which computes the cutoff points based on conditional inference trees and recursive partitioning [18].

To select powerful discretized variables, one common measurement is information value defined in Equation (4) [14], where p_j is the number of non-events (i.e. non-delinquency) in the level j of the variable divided by the total number of non-events, and q_j is the number of events (i.e. delinquency) in the level j of the variable divided by the total number of events. To interpret the information value, the following rule of thumb is proposed [37,40].

- <0.02: useless
- 0.02 to 0.1: weak
- 0.1 to 0.3: medium
- >0.3: strong

$$IV = \sum_j (p_j - q_j) \ln(p_j/q_j) \tag{4}$$

3. Data

Demographic and financial information from 150,000 borrowers is publicly available in a dataset used in a Kaggle 2011 Competition Give Me Some Credit [19]. The characteristics of the individuals in the data are represented by 11 variables, as shown in Table 1. The goal was to predict whether a client will experience financial distress in the next two years or not, indicated by the dependent variable *SeriousDlqin2yrs*. As shown in Table 2, there are

Table 1. Variables for analysis and modeling.

Variable	Type	Description
SeriousDlqin2yrs	Binary	Person experienced 90 days past due delinquency or worse
MonthlyIncome	Interval	Monthly income
DebtRatio	Interval	Monthly debt payments, alimony, living costs divided by monthly gross income
Age	Interval	Age of borrower in years
NumberOfDependents	Interval	Number of dependents in family excluding themselves (spouse, children, etc.)
NumberOfOpenCreditLinesAndLoans	Interval	Number of open loans (installment like car loan or mortgage) and lines of credit (e.g. credit cards)
NumberRealEstateLoansOrLines	Interval	Number of mortgage and real estate loans including home equity lines of credit
RevolvingUtilizationOfUnsecuredLines	Interval	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits
*NumberOfTime*30−−59*DaysPastDueNotWorse*	Interval	Number of times borrower has been 30–59 days past due but no worse in the last 2 years
*NumberOfTime*60−−89*DaysPastDueNotWorse*	Interval	Number of times borrower has been 60–89 days past due but no worse in the last 2 years
*NumberOfTimes*90*DaysLate*	Interval	Number of times borrower has been 90 days or more past due

Table 2. Frequency of dependent variable.

SeriousDlqin2yrs	Frequency	Percent (%)
1	10,026	6.68
0	139,974	93.32

10,026 delinquent observations and 139,937 non-delinquent observations. The proportion of delinquencies is 6.68%.

There are 29,731 observations with missing values either in the variable *MonthlyIncome* or *NumberOfDependents*, which is 19.82% of the total. These missing values are treated as follows.

(1) Missing Completely at Random (MCAR) analysis is conducted, and there is no pattern existing in the missing data. Hence, those observations are dropped to ensure the data accuracy and support the model training computation, when building the model with original variables. After dropping missing data, the proportion of delinquencies is 6.95%, which is very close to the original data.
(2) When building the model with discretized variables, those observations are kept by grouping the missing values separately into a level of a variable.

3.1. Exploratory analysis

Because the dependent variable is binary and all independent variables are interval, the empirical logit plot is used to examine the linearity of the relationship between the dependent variable and independent variables. If the relationship is linear, it is reasonable to use the interval form of an independent variable. Otherwise, a transformation is required. Moreover, through the empirical logit plots, we can check the univariate effects, positive or negative.

The empirical logit plot is created in the following steps.

(1) For each interval variable, generate percentile ranks from 1 to 100 [35].
(2) For each rank i of each interval variable, calculate the total number of observations N_i, the number of delinquency observations Y_i, and the mean of the interval variable $\bar{x}_i$.
(3) For each rank i of each interval variable, compute the empirical logit using the formula $elogit_i = \log((Y_i + 0.5)/(N_i - Y_i + 0.5))$ [9].
(4) For each interval variable, plot the empirical logit *elogit* against the mean in each rank $\bar{x}$ and their linear regression line. Each point in the plot represents N_i data points from the dataset by their mean.
(5) For each interval variable, plot the empirical logit *elogit* against the rank i and their linear regression line. Each point in the plot represents N_i data points from the dataset by their rank index.

For example, consider the predictor variable *RevolvingUtilizationOfUnsecuredLines*. Percentile ranks can be found in Table 3. Ranks 1–8 are merged together because their respective minimum and maximum points are the same. As shown in Figure 1(a), there is a nonlinear relationship between *RevolvingUtilizationOfUnsecuredLines* and its empirical logit, mainly caused by extreme values. These extreme values in the empiricial logit plot cannot be simply removed, considering they represent several hundred data points in the dataset. However, the relationship between its rank and its empiricial logit is approximately linear as shown in Figure 1(b). In this case, its rank, the discretized form of its original interval values, is preferred to be used in the modeling.

3.2. Variable discretization

Four variable discretization methods (i.e. distance, quantile, Gini, optimal binning) are compared. On the credit scoring dataset, the quantile discretization produces the highest AUC on the test data with the logistic regression model trained on the training data, where the ratio of training data and test data is 70% vs. 30%. Each variable is ranked and discretized into 20 bins maximally based on the quantile, with the threshold value 20 selected by the same procedure above.

Information value is used as the measurement of the discrimination power of each individual variable after discretization, as shown in Table 4. Note that for some variables, the resulting number of bins is less than 20 because the bins with non-significant differences are merged together. For the variable *MonthlyIncome*, an additional bin has been included

Table 3. Percentile ranks of *RevolvingUtilizationOfUnsecuredLines*.

Rank i	Min	Max	Mean $\bar{x}_i$	Count N_i	Event Y_i	$elogit_i$
1–8	0	0.000707	0.000034	12,000	335	−3.5488
9	0.000708	0.001733	0.001210	1501	18	−4.3844
10	0.001735	0.002969	0.002334	1499	25	−4.0574
…	…	…	…	…	…	…
99	1.0062	1.092954	1.036357	1500	556	−0.5290
100	1.093178	50708	573.887190	1500	589	−0.4358

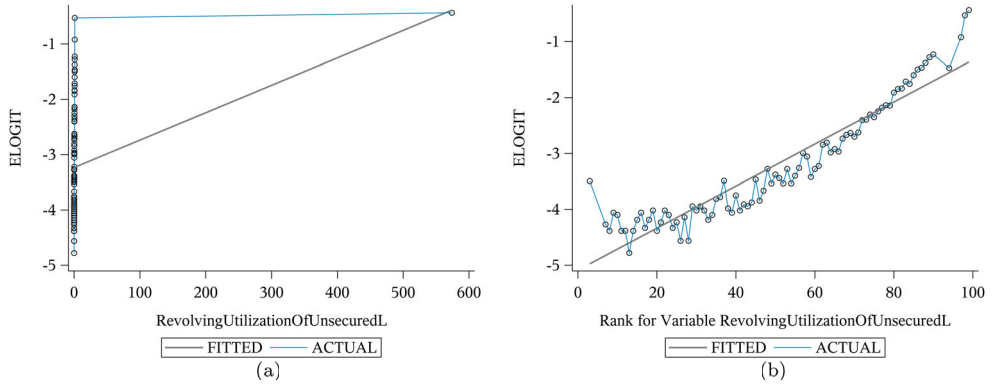

Figure 1. Empirical logit plot against *RevolvingUtilizationOfUnsecuredLines* and its rank.

Table 4. Information values.

Variables	Bins	Information value
RevolvingUtilizationOfUnsecuredLines	19	1.1635
NumberOfTime30−−59DaysPastDueNotWorse	3	0.4865
NumberOfTimes90DaysLate	2	0.4842
NumberOfTime60−−89DaysPastDueNotWorse	2	0.2648
Age	20	0.2620
NumberOfOpenCreditLinesAndLoans	15	0.0852
MonthlyIncome	21	0.0813
DebtRatio	20	0.0795
NumberOfDependents	5	0.0279
NumberRealEstateLoansOrLines	4	0.0184

to accomodate missing values. By following the rule suggested by Hand *et al.* [14], the variables with the information value over 0.1 will be studied.

To prepare the discretized variables for the modeling, they are further transformed by one-hot encoding. A one-hot encoder converts a discretized variable into multiple binary dummy variables with each bin represented by one binary dummy variable [30,36].

3.3. Datasets from other domains

Beyond the credit scoring data, two public datasets from other domains (i.e. biology, business) are collected. They include 206 and 11 interval variables respectively, as shown in Table 5. The goal of the arrhythmia data is to predict sinus bradycardia [12], and the goal of the wine_quality data is to predict poor quality [5]. The process illustrated in Sections 3.1 and 3.2 is performed on these two datasets. Among all variable discretization methods, the optimal binning method produces the best performance. The resulting discretized variables will be modeled using logistic regression in Section 5.

Table 5. Basic characteristics of datasets.

Dataset	Repository	Target	Event rate	Observations	Variables	Domain
arrhythmia	UCI	06	5.55%	452	206C, 73N	Biology
wine_quality	UCI	score<= 4	3.70%	4898	11C	Business

4. Modeling

Logistic regression and class-dependent cost-sensitive logistic regression are used as classifiers for their high interpretability. The models are evaluated by 10-fold cross-validation. The performance measurements include ROC curve, AUC, Type I Error, Type II Error, accuracy, and F1 Score. The mean of AUCs of 10-fold cross-validation is used to measure the model bias, while the standard deviation of AUCs of 10-fold cross-validation is used to measure the model variance. They are reasonable measurements, considering that the model bias refers to the error introduced by approximating the true model, and the model variance refers to the amount of the change of the estimated model if using a different training dataset [16].

To evaluate and compare the performance of variable discretization and class-dependent cost-sensitive logistic regression, the following five models are built.

- Model 1: Logistic regression model on all original interval form of independent variables in Table 1.
- Model 2: Logistic regression model on original interval form of variables with the information value over 0.1 in Table 4.
- Model 3: Class-dependent cost-sensitive logistic regression model on the same independent variables in Model 2. The class weights (i.e. W_0, W_1) that produce the highest mean of AUCs of 10-fold cross-validation are used in the modeling, indicated by the dash line in Figure 2(b). The search for the best class weights will be discussed below.
- Model 4: Logistic Regression model on discretized form of independent variables used in Model 2. The discretized variables are transformed by the one-hot encoder. In total, 48 binary dummy variables are created.
- Model 5: Class-dependent cost-sensitive logistic regression model on the same discretized independent variables in Model 4. The class weights (i.e. W_0, W_1) that produce the highest mean of AUCs of 10-fold cross-validation are used in the modeling, indicated by the solid line in Figure 2(b).

For the class weights (i.e. W_0, W_1) in Model 3 and Model 5, they are determined by the population proportion of events τ and the sample proportion of events $\bar{y}$ in Equation (3). $\bar{y}$ is known from the data. τ is typically unknown and hard to obtain accurate estimation [7].

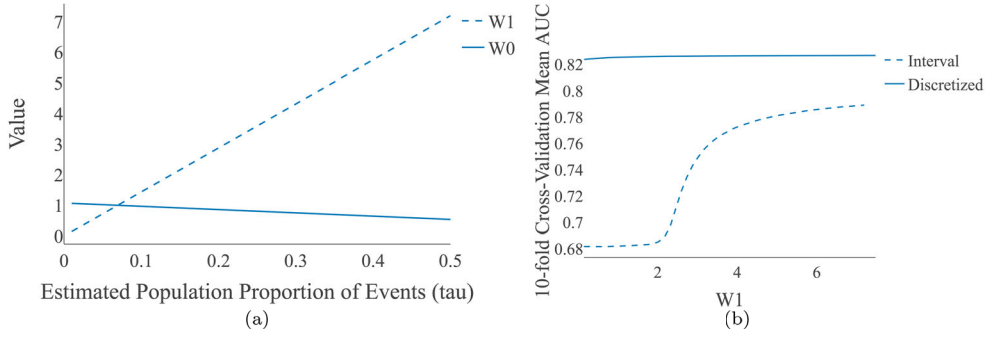

Figure 2. The result of tuning τ. (a) τ vs. Class Weights and (b) AUROC vs. $W1$.

Here τ is tuned as a hyperparameter from 0 to 0.5. As shown in Figure 2(a), as τ increases, W_1 increases and W_0 decreases linearly. Figure 2(b) shows how the mean of AUCs on the 10-fold cross-validation changes as $W1$ increases. When modeling on interval variables in Model 3, the best occurs at $\tau = 0.5$, resulting in $W_1 = 7.19$ and $W_0 = 0.54$. The changes of the class weights have minimal influence on the modeling of discretized variables used in Model 5, implying that good variable discretization is robust to penalty weights determined by proportions of events and non-events. Hence, for Model 5, we take $W_1 = 1$ and $W_0 = 1$, leading Model 5 the same as Model 4. Because of this, we will only compare Model 4 with other models in the following section.

The ROC curve of each model can be found in Figure 3. Model 1 and Model 2 have similar AUCs, indicating that the variables with the information value below 0.1 provide minimal contribution. The ROC curves of Model 3 and Model 4 demonstrate stronger

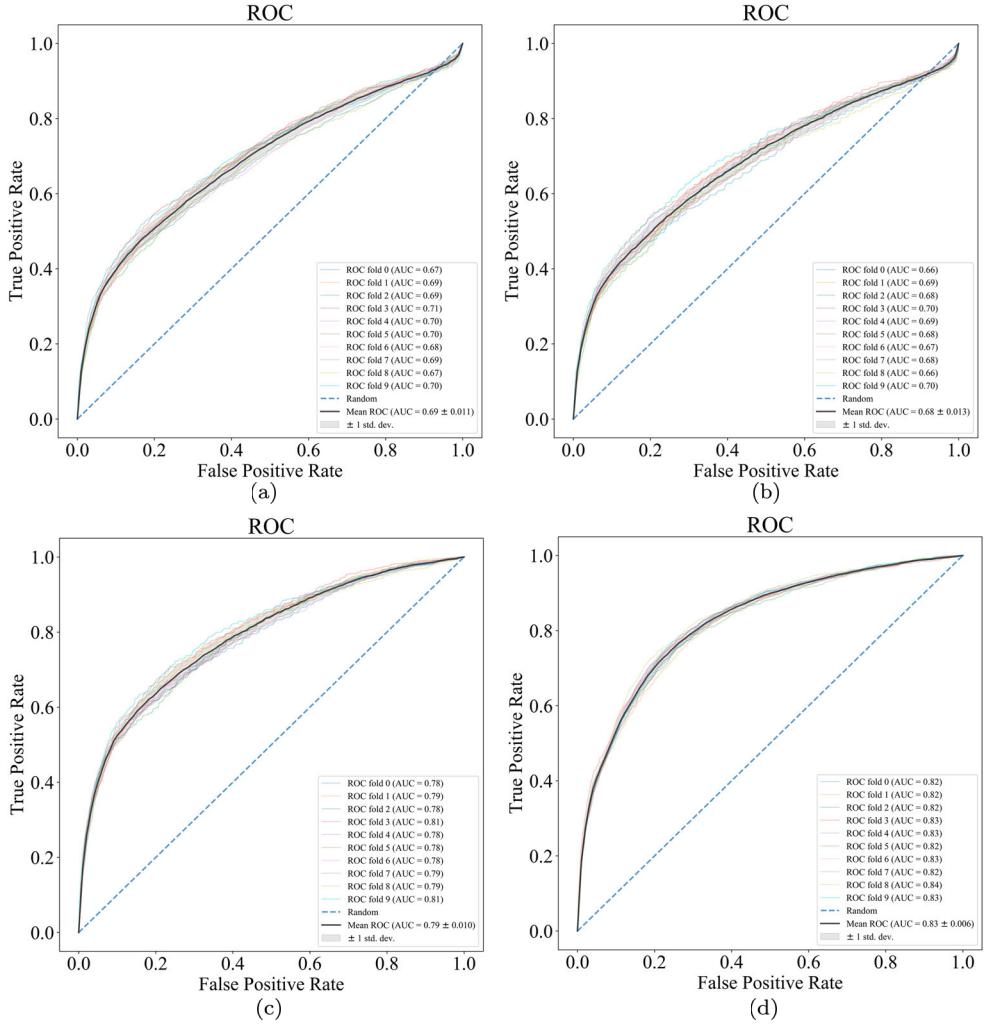

Figure 3. 10-fold cross-validation ROC curves of credit scoring data. (a) Model 1. (b) Model 2. (c) Model 3 and (d) Model 4.

Table 6. 10-fold cross-validation AUC of models.

Model	Mean	Std.
Model 1	0.69	0.011
Model 2	0.68	0.013
Model 3	0.79	0.010
Model 4	0.83	0.006

Table 7. Estimated parameters of Model 2 and Model 3.

Parameter	Model 2 Estimate	Model 3 Estimate
Intercept	−1.45644	2.69671
RevolvingUtilizationOfUnsecuredLines	−0.000048	−0.000053
*NumberOfTime*30−59*DaysPastDueNotWorse*	0.50255	0.67117
*NumberOfTimes*90*DaysLate*	0.45629	0.79821
*NumberOfTime*60−89*DaysPastDueNotWorse*	−0.92206	0.47276
Age	−0.02791	−0.02809

results than Model 2. Moreover, for Model 4, the ROC curves on 10-fold cross-validation are closer to each other, indicating lower model variance. This can be further confirmed by the mean and standard deviation of AUCs on 10-fold cross-validation in Table 6. Model 4 produces the highest mean and the lowest standard deviation of AUCs, demonstrating the power of variable discretization.

The estimated coefficients of the models are also examined. As shown in Table 7, Model 2 and Model 3 produce different estimates for every independent variable, as well as the sign of the variable *NumberOfTime*60−89*DaysPastDueNotWorse*. Its sign is negative in Model 2, while its sign is positive in Model 3. Its empirical logit plot in Figure 4(c) shows the positive relationship. Based on its variance inflation factor (VIF) in Table 8, its sign change in Model 2 is caused by its multicollinearity with the variables *NumberOfTime*30−59*DaysPastDueNotWorse* and *NumberOfTimes*90*DaysLate*. None of them can be dropped in the modeling because of their information values presented in Table 4. Model 3 specificly guarantees a positive estimate, which is consistent with the univariate effect. For other variables, the signs of estimated parameters are consistent with their univariate effect shown in their empirical logit plots in Figures 4(a,b,d). The estimated parameters of Model 4 are not presented here because of space limitation. Considering these dummy variables are binary indicators transformed by one-hot encoder, their estimated coefficients are more interpretable.

Further, these models are compared based on Type I Error, Type II Error, accuracy, and F1 score on the test data after splitting the original dataset into training data (70%) and test data (30%), which can be found in Table 9. The probability cutoff is chosen as the intersection point of the specificity plot and sensitivity plot, one of the most frequently used criterion [13,32]. We have the following findings.

- There is no improvement from Model 1 to Model 2, indicating that variables with information value below 0.1 provide limited contribution.
- Compared with Model 2, Model 3 decreases Type I Error by 8.23%, decreases Type II Error by 8.3%, increases accuracy by 8.24%, and increases F1 score by 0.0656, indicating the contribution of penalizing the misclassifications of events and non-events in different scales by running the class-dependent logistic regression.

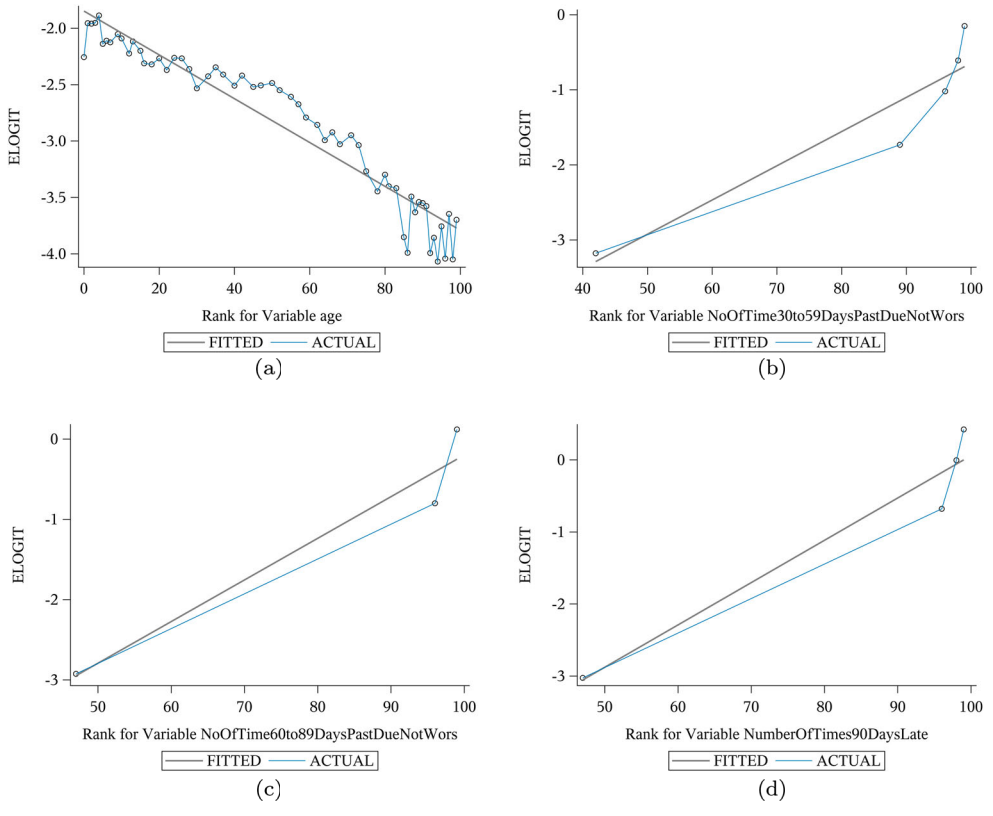

Figure 4. Empirical logit plots against ranks. (a) *Age*. (b) *NumberOfTime*30*to*59*DaysPastDueNotWorse*. (c) *NumberOfTime*60*to*89*DaysPastDueNotWorse* and (d) *NumberOfTimes*90*DaysLate*.

Table 8. VIF for *NumberOfTime*60−89*DaysPast DueNotWorse*.

Parameter	VIF Factor
RevolvingUtilizationOfUnsecuredLines	1
*NumberOfTime*30−59*DaysPastDueNotWorse*	20.5
*NumberOfTimes*90*DaysLate*	20.5
Age	1

Table 9. Performance measures under the best probability cutoff.

Model	Type I Error	Type II Error	Accuracy	F1 score	Probability cutoff
Model 1	36.61%	36.54%	63.39%	0.1941	0.0666
Model 2	36.73%	36.74%	63.27%	0.1931	0.0654
Model 3	28.50%	28.44%	71.51%	0.2587	0.4486
Model 4	24.88%	24.83%	75.12%	0.2877	0.0646

- Compared with Model 2, Model 4 decreases Type I Error by 11.85%, decreases Type II Error by 11.91%, increases accuracy by 11.85%, and increases F1 score by 0.0946, indicating the contribution of variable discretization.

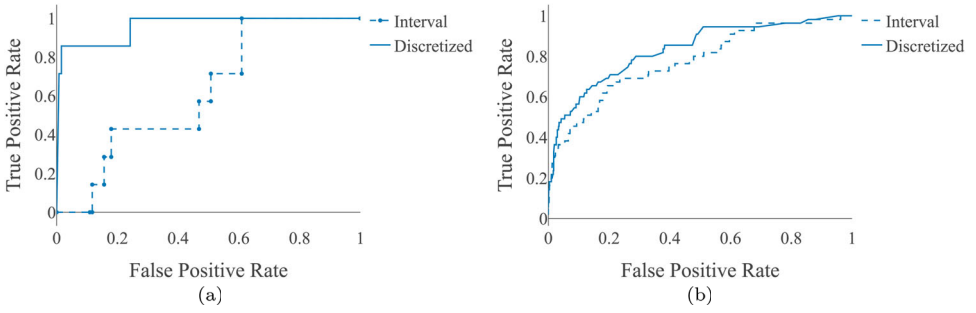

Figure 5. ROC curves of wine_quality and arrhythmia data. (a) arrhythmia and (b) wine_quality.

Table 10. The performance of variable discretization on other datasets.

Dataset	Model	AUC	Type I Error	Type II Error	Accuracy	F1 score	Probability cutoff
arrhythmia	Interval	0.6216	46.87%	42.86%	53.33%	0.1127	3.64e−11
	Discretized	0.9603	14.06%	14.28%	85.92%	0.3871	2.09e−17
wine_quality	Interval	0.7757	30.74%	30.91%	69.25%	0.1439	0.0317
	Discretized	0.8327	26.08%	25.45%	73.95%	0.1763	0.0289

- Compared with Model 3, Model 4 decreases Type I Error by 3.62%, decreases Type II Error by 3.61%, increases accuracy by 3.61%, and increases F1 score by 0.0290, indicating that variable discretization performs better than the inclusion of class-dependent costs in the logistic regression.

5. Application of variable discretization in other domains

To further examine the power of variable discretization, logistic regression models with original interval variables and discretized variables in the datasets arrhythmia and wine_quality are built and compared. The original datasets are split into training data (70%) and test data (30%). Logistic regression models are trained on the training data and then evaluated on the test data.

Their resulting ROC curves on the test data can be found in Figure 5. For both datasets, the ROC curve by discretized variables moves closer to the upper-left corner than the one by interval variables. The improvement can be further checked by other performance measures (i.e. Type I Error, Type II Error, accuracy, F1 score) in Table 10, where the probability cutoff is chosen as the intersection point of the sensitivity plot and specificity plot. For example, on the dataset arrhythmia, Type I Error decreases by 32.81%, Type II Error decreases by 27.58%, accuracy increases by 32.59%, and F1 score increases by 0.2744. Note that the probability cutoff on this dataset is very small, but it is reasonable that some estimated probabilities are very close to 0, considering the facts that they are direct outputs of a sigmoid function ranging from 0 to 1 and target classes (i.e. non-event, event) are represented by 0 and 1 in the data.

6. Discussions and conclusions

To improve the model performance on imbalanced data, efforts have been made from the perspective of the predictors and the modeling algorithm, respectively, in this study.

Through the detailed study on the credit scoring dataset, we show that the proper variable discretization and class-dependent cost-sensitive logistic regression with the best class weights help reduce the model bias and/or variance, based on the ROC curves and AUC on 10-fold cross-validation, Type I Error, Type II Error, accuracy, and F1 score. Moreover, class-dependent cost-sensitive logistic regression is beneficial for increasing the prediction power of predictors during the training phase even if those predictors are not transformed in their best forms and keeping the multivariate effect and univariate effect of predictors consistent.

On the other hand, the logistic regression model with proper discretized variables performs better than class-dependent cost-sensitive logistic regression, provides more reasonable coefficient estimates, and is robust to penalty scales of misclassification costs of events and non-events determined by their proportions. This indicates that we should always discretize the variables showing nonlinear relationships against their empirical logits.

In this study, logistic regression and its variant (i.e. class-dependent cost sensitive logistic regression) are used as classifiers. In the future, we will study the performance of variable discretization with other classifiers such as decision tree, support vector machine, and neural network.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Lili Zhang http://orcid.org/0000-0003-1935-4223
Herman Ray http://orcid.org/0000-0003-3444-098X

References

[1] A. Ali, S.M. Shamsuddin, and A.L. Ralescu, *Classification with class imbalance problem: A review*, Int. J. Adv. Soft Comput. Appl. 7 (2015), pp. 176–204.
[2] A.C. Bahnsen, D. Aouada, and B. Ottersten, *Example-dependent cost-sensitive logistic regression for credit scoring*, 2014 13th International Conference on Machine Learning and Applications (ICMLA), IEEE, 2014, pp. 263–269.
[3] I. Brown and C. Mues, *An experimental comparison of classification algorithms for imbalanced credit scoring data sets*, Expert Syst. Appl. 39 (2012), pp. 3446–3453.
[4] G. Collell, D. Prelec, and K.R. Patil, *A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data*, Neurocomputing 275 (2018), pp. 330–340.
[5] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, *Modeling wine preferences by data mining from physicochemical properties*, Decis. Support Syst. 47 (2009), pp. 547–553.
[6] K. Deng, *Omega: On-line memory-based general purpose system classifier*, Ph.D. diss., Carnegie Mellon University, 1998.
[7] J. Ding and W. Xiong, *A new estimator for a population proportion using group testing*, Commun. Stat. Simul. Comput. 45 (2016), pp. 101–114.
[8] S. Ding, B. Mirza, Z. Lin, J. Cao, X. Lai, T.V. Nguyen, and J. Sepulveda, *Kernel based online learning for imbalance multiclass classification*, Neurocomputing 277 (2018), pp. 139–148.
[9] S. Donnelly and J. Verkuilen, *Empirical logit analysis is not logistic regression*, J. Mem. Lang. 94 (2017), pp. 28–42.

[10] J. Dougherty, R. Kohavi, and M. Sahami, *Supervised and unsupervised discretization of continuous features*, in *Machine Learning Proceedings 1995*, Elsevier, San Francisco, CA, 1995, pp. 194–202.
[11] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, *On the class imbalance problem*, 2008 Fourth International Conference on Natural Computation, Vol. 4. IEEE, 2008, pp. 192–201.
[12] H.A. Guvenir, B. Acar, G. Demiroz, and A. Cekin, *A supervised machine learning algorithm for arrhythmia analysis*, in *Computers in Cardiology 1997*. IEEE, Lund, Sweden, 1997, pp. 433–436.
[13] F. Habibzadeh, P. Habibzadeh, and M. Yadollahie, *On determining the most appropriate test cut-off value: The case of tests with continuous results*, Biochem. Med. 26 (2016), pp. 297–307.
[14] D.J. Hand and W.E. Henley, *Statistical classification methods in consumer credit scoring: a review*, J. R. Statist. Soc. Ser. A (Stat. Soc.) 160 (1997), pp. 523–541.
[15] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, John Wiley & Sons, Hoboken, NJ, 2013.
[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Vol. 112, Springer, Berlin, Germany, 2013.
[17] N. Japkowicz, *The class imbalance problem: Significance and strategies*, Proc. of the Int'l Conf. on Artificial Intelligence, Las Vegas, NV, 2000.
[18] H. Jopia, *Scoring Modeling and Optimal Binning* (2018). Available at https://cran.r-project.org/web/packages/smbinning/smbinning.pdf, Accessed: 2018-10-11.
[19] Kaggle, *Give Me Some Credit*. Available at https://www.kaggle.com/c/GiveMeSomeCredit/data, Accessed: 2018-02-01.
[20] G. King and L. Zeng, *Logistic regression in rare events data*, Polit. Anal. 9 (2001), pp. 137–163.
[21] S. Kotsiantis and D. Kanellopoulos, *Discretization techniques: A recent survey*, GESTS Int. Trans. Comput. Sci. Eng. 32 (2006), pp. 47–58.
[22] B. Krawczyk, *Learning from imbalanced data: Open challenges and future directions*, Progress Artif. Intell. 5 (2016), pp. 221–232.
[23] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, *Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy*, Appl. Soft Comput. 38 (2016), pp. 714–726.
[24] B. Krawczyk and M. Woźniak, *Cost-sensitive neural network with roc-based moving threshold for imbalanced classification*, International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2015, pp. 45–52.
[25] J.L. Leevy, T.M. Khoshgoftaar, R.A. Bauder, and N. Seliya, *A survey on addressing high-class imbalance in big data*, J. Big Data 5 (2018), p. 42.
[26] M. Maalouf and T.B. Trafalis, *Robust weighted kernel logistic regression in imbalanced and rare events data*, Comput. Stat. Data Anal. 55 (2011), pp. 168–183.
[27] M. Maalouf, T.B. Trafalis, and I. Adrianto, *Kernel logistic regression using truncated newton method*, Comput. Manag. Sci. 8 (2011), pp. 415–428.
[28] mlr-org, *Cost-Sensitive Classification*. Available at https://mlr-org.github.io/mlr-tutorial/release/html/cost_sen sitive_classif/index.html, Accessed: 2018-04-27.
[29] A. Moayedikia, K.L. Ong, Y.L. Boo, W.G. Yeoh, and R. Jensen, *Feature selection for high dimensional imbalanced class data using harmony search*, Eng. Appl. Artif. Intell. 57 (2017), pp. 38–49.
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res. 12 (2011), pp. 2825–2830.
[31] C. Phua, D. Alahakoon, and V. Lee, *Minority report in fraud detection: Classification of skewed data*, ACM SIGKDD Explor. Newsl. 6 (2004), pp. 50–59.
[32] L.A. Pramono, S. Setiati, P. Soewondo, I. Subekti, A. Adisasmita, N. Kodim, and B. Sutrisna, *Prevalence and predictors of undiagnosed diabetes mellitus in indonesia*, Age 46 (2010), pp. 100–100.
[33] F. Provost, *Machine learning from imbalanced data sets 101*, Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets, Austin, TX, 2000, pp. 1–3.

[34] M.M. Rahman and D. Davis, *Addressing the class imbalance problem in medical datasets*, Int. J. Mach. Learn. Comput. 3 (2013), pp. 224.
[35] R. Rousseau, *Basic properties of both percentile rank scores and the i3 indicator*, J. Am. Soc. Inf. Sci. Technol. 63 (2012), pp. 416–420.
[36] Scikit-learn, *One Hot Encoder*. Available at https://scikit-learn.org/stable/modules/generated/sklearn.pre processing.OneHotEncoder.html, Accessed: 2018-02-01.
[37] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Vol. 3, John Wiley & Sons, Hoboken, NJ, 2012.
[38] S. Visa and A. Ralescu, *Issues in mining imbalanced data sets-a review paper*, Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, Vol. 2005. SN, 2005, pp. 67–73.
[39] H. Wang, Q. Xu, and L. Zhou, *Large unbalanced credit scoring using lasso-logistic regression ensemble*, PLoS ONE 10 (2015), pp. e0117844.
[40] G. Zeng, *Metric divergence measures and information value in credit scoring*, J. Math. 2013 (2013), Article ID 848271.