文章编号:1674-2869(2025)03-0343-06

基于SMOTE-XGBoost算法的信用卡违约预测模型研究

赵 阳,张杰萌,严国义*
武汉工程大学光电信息与能源工程学院、数理学院,湖北 武汉 430205

摘 要:针对信用卡违约现象,提出了一种基于SMOTE-XGBoost算法的预测模型。该模型采用合成少数类过采样技术(SMOTE)对数据集进行处理,选择极限梯度提升树(XGBoost)模型作为学习器,提升模型整体的预测效果。为验证SMOTE的有效性以及XGBoost算法的最优性,本文首先采用随机森林、神经网络、梯度提升决策树、逻辑回归、k近邻、XGBoost和LightGBM模型对原数据集进行数据建模分析和预测,之后使用Regular-SMOTE、Borderline-SMOTE和SVM-SMOTE采样方式对数据集做相对平衡处理,然后再使用7种模型分别对平衡处理后的数据集进行建模分析和预测,并引入准确率、精确率、F1指数、曲线下面积作为模型好坏的评价指标。不同采样方式和模型之间的对比分析结果表明,在经过SMOTE采样后,各模型的预测效果显著提升,其中使用XGBoost模型对经过SVM-SMOTE采样后的数据进行建模分析,该方法的预测效果最好,采用此模型可为金融行业制定放贷策略和降低企业风险提供决策支持。

关键词:SMOTE;XGBoost;违约预测;不平衡数据
中图分类号:TP39 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202312031

A credit card default prediction model based on SMOTE-XGBoost algorithm

ZHAO Yang, *ZHANG Jiemeng*, *YAN Guoyi**
School of Optical Information and Energy Engineering, School of Mathematics and Physics,
Wuhan Institute of Technology, Wuhan 430205, China

Abstract: To address credit card defaults, a prediction model based on SMOTE-XGBoost algorithm was proposed, which uses the synthetic minority oversampling technique (SMOTE) to process the dataset, and selects the extreme gradient boosting tree (XGBoost) model as the learner to improve its overall predictive performance. In order to verify the validity of SMOTE and the optimality of the XGBoost algorithm, this study first used random forest, neural network, gradient boosting decision tree, logistic regression, k-nearest neighbor, XGBoost and LightGBM models to model, analyze and predict the original dataset, and then used Regular-SMOTE, Borderline-SMOTE and SVM-SMOTE sampling methods to perform relative balance processing on the dataset, and later used seven models to model, analyze and predict the balanced dataset respectively, and introduced accuracy, precision, F1 measure, and area under the curve as evaluation indexes of the model. The results of comparative analysis between different sampling methods and models showed that after SMOTE sampling, the prediction performance of each model is significantly improved. The XGBoost model was used to model and analyze the data after SVM-SMOTE sampling, and it had the best predictive power. This prediction model provides decision support for the financial industry to formulate lending strategies and reduce enterprise risks.

Keywords: SMOTE; XGBoost; default prediction; imbalanced data

收稿日期:2024-01-19
基金项目:国家自然科学基金(12101469)
作者简介:赵 阳,硕士研究生。Email:601911304@qq.com
***通信作者**:严国义,博士,副教授。Email:yanguoyi@wit.edu.cn
引文格式:赵阳,张杰萌,严国义.基于SMOTE-XGBoost算法的信用卡违约预测模型研究[J].武汉工程大学学报,2025,47(3):343-348.

信用卡因其安全便捷、利息低等特点被广泛使用于日常生活中。然而这一现象可能会引发诸多问题,其中信用卡的风险管控对于银行等金融机构而言是一项具有挑战性的任务,金融机构需要权衡信用卡业务带来的利益与风险。预测用户信用卡的消费额度、还款时效性、违约可能性等已成为监控信用卡风险的重要内容。

国外对于信用卡违约问题的研究较早,1999年,Brause等[1]将关联规则和神经网络(neural network, NN)融合起来预测信用卡欺诈。国内也有对这方面问题的研究,范巍强等[2]使用反向传播NN模型预测信用卡使用者的违约风险,郭建山等[3]通过改进随机森林(random forest, RF)预测模型提高违约风险预测效果。

用户信用卡违约预测任务包含信用卡消费预测、信用卡还款预测等。本质上来说,用户信用卡违约预测属于分类任务。用户信用卡历史相关的各项数据可以作为数据集特征,如在不同月份的消费金额、还款金额等,经过监督学习或无监督学习的方式可以预测未来一段时期内用户是否会发生违约行为。

在信用卡违约预测中,信用卡数据存在标签获取难度大和数据分布不均衡的问题,使得经典算法的分类效果不佳[4]。Douzas等[5]利用条件生成对抗网络生成少数类样本以平衡正负样本之间的数量;Yi等[6]提出了结合聚类的合成少数类过采样技术(synthetic minority over-sampling technique, SMOTE),并分析不同样本分布的均衡问题。过采样技术的优点是不会导致信息损失,缺点是其复制了原始数据集中的观测值,因此加大了过拟合的可能性。

本文将采用3种SMOTE对数据集进行预处理,随后分别使用RF、NN、梯度提升决策树(gradient boosting decision tree, GBDT)、逻辑回归(logistic regression, LR)、k近邻(k-nearest neighbor, KNN)、极限梯度提升树(eXtreme gradient boosting, XGBoost)和轻量梯度提升机(light gradient boosting machine, LightGBM)模型对初始数据集以及预处理后的数据集进行训练和预测,根据不同模型之间的比对,研究3种SMOTE与各种模型的结合,并分析它们在信用卡违约预测中的表现。

1 数据处理

1.1 数据来源及预处理

本文所采用的数据集来自Kaggle官网,为国外某银行信用卡数据集,该数据集包含从2015年到2017年间信用卡用户的属性特征及其是否违约的相关数据。该数据共有15万条样本,包含10个属性特征及1个标签变量。用户特征有借款人的月收入、未结贷款数量、借款人逾期30~59 d的次数等10个变量,标签变量"SeriousDlqin2yrs"表示2年内是否有超过90 d的严重逾期行为,如果有逾期行为,则取值为1,否则取值为0。

查看该数据集发现有些变量的值存在严重缺失。比如,用户"月收入"变量缺失值较多(缺失比例达20%),为了增强填充后数据的稳健性,采用中位数法填补该缺失值,以确保数据建模的准确性。其中变量"家庭数量"缺失值较少,故采取直接删除的方式去掉该变量含有缺失值的观测值。

1.2 SMOTE

在模型训练时,数据的不平衡会使模型倾向于将待判样本预测为样本数偏多的类别,导致预测误差出现偏向性。不平衡数据的研究当前主要集中在数据预处理、特征选取和算法3个层面,目的是确保分类器在多数类或少数类的情况下都具有较高的分类精度。数据预处理的主要思想是改变原始数据集的分布,增加少数类样本或减少多数类样本,以降低或者消除数据不平衡现象,使非平衡数据集趋于平衡[7]。

SMOTE是由Chawla等[8]提出的一种用于解决分类问题中类别不平衡的采样方法,其通过生成合成的样本来平衡训练数据集中的正负样本数量。SMOTE可使分类器对少数类样本的关注度增加,从而对样本具有更大的泛化空间,改善少数类样本的分类效果,提升算法整体的分类精度。SMOTE作为一种解决数据非平衡问题的采样技术,可以简洁、高效地获取大量少数类样本,故而在金融、数据挖掘、医学等领域得到了广泛运用[9]。

1.2.1 Regular-SMOTE Regular-SMOTE表示对少数类别的样本进行随机采样,具体步骤如下:

(1)对少数类别中的每个样本X,以欧氏距离为标准计算它到少数类样本集F中所有样本的距离,得到其K个近邻。

(2)根据样本不平衡比例设置采样比例以确定采样倍率P,对于少数类样本中的每个样本X,从其K个近邻样本中随机选择若干个样本进行随机线性插值。

(3)将该样本与原数据合成,产生新的训练集。

1.2.2 Borderline-SMOTE Borderline-SMOTE是Regular-SMOTE的改进。SMOTE会导致少数类样

本被错误分类，并且这些样本位于决策边界的边缘或边界区域，因此Han等[10]提出了Borderline-SMOTE算法，该算法是对位于分类边界的少数类样本进行过采样，加强对分类边界样本的学习。Borderline-SMOTE算法流程如图1所示。

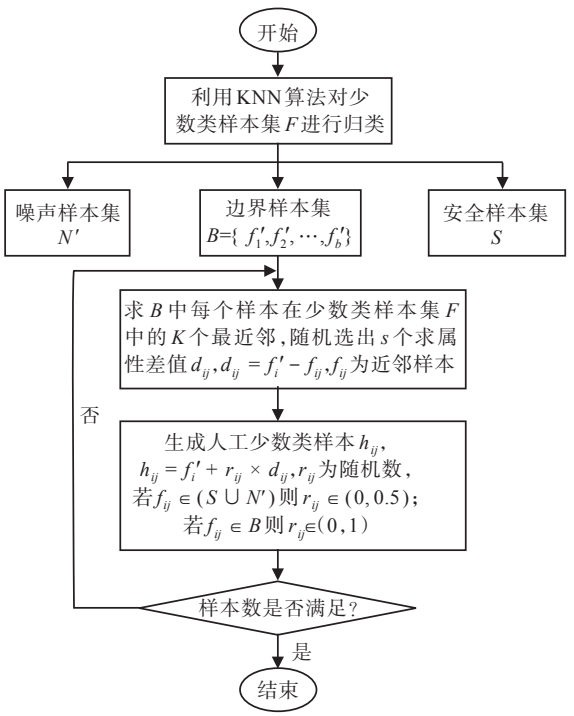

图1　Borderline-SMOTE算法流程图

Fig. 1　Borderline-SMOTE algorithm

1.2.3　SVM-SMOTE　Nguyen等[11]提出了Borderline-SMOTE的替代方案SVM-SMOTE，其使用支持向量机（support vector machine，SVM）算法而非k近邻算法来识别决策边界上的错误分类样本。除此之外，SVM-SMOTE还尝试选择少数类样本较少的区域，并尝试向类边界外推。SVM-SMOTE算法流程如图2所示。

2　XGBoost算法与模型

XGBoost算法是由Chen等[12]提出的一种高效的集成学习算法，属于加法模型，由多个基学习器构成，每个基学习器为一棵回归树，每棵回归树的输出值即为叶节点的权重。当对一个样本进行预测时，根据内部结点的决策条件进行划分，最终被划分到的叶节点的权重即为该样本的预测输出值。XGBoost因其高预测精度、处理缺失值和稀疏数据的能力、灵活性以及支持并行化操作，在机器学习领域中得到了广泛的应用[13-15]。XGBoost算法基本原理及流程如下：

给定一个具有N个样本和m个特征的数据集

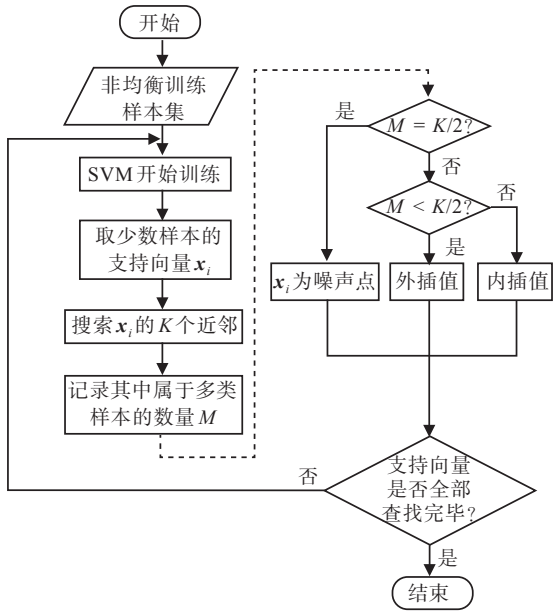

图2　SVM-SMOTE算法流程图

Fig. 2　SVM-SMOTE algorithm

D，$D=\{\left(x_i,y_i\right)\}(|D|=N,x_i\in R^m,y_i\in R)$，其中，$x_i$为第$i$个特征值，$y_i$为第$i$个目标值，$R$为实数集。假定满足模型迭代终止规则时，XGBoost模型共建立κ棵回归树，则第i个样本的预测值$\hat{y}_i$的表达式为：

$$\hat{y}_i=\sum_{k=1}^{\kappa}f_k\left(x_i\right),f_k\in F^* \tag{1}$$

其中：$f_k(x_i)$为第i个样本在第k棵回归树的预测值；F^*为回归树组成的函数空间，$F^*=W_{q(x_i)},W_{q(x_i)}$表示$q(x_i)$对应的叶子节点的权重，$q(x_i)$表示特征值$x_i$所在的叶子节点。

构造样本集合I_j与第j个叶子节点之间的映射关系为$I_j=\{i|q(x_i)=j\}$，其含义为给定一个样本，根据特征找到其对应的叶子节点。

XGBoost对目标函数的优化是其区别于GBDT算法的关键，XGBoost模型第t次迭代后的优化目标函数L_t为：

$$L_t=\sum_{i=1}^{n}L(y_i,\hat{y}_{i,t})+\sum_{k=1}^{t}\Omega\left(f_k\right) \tag{2}$$

其中：$\hat{y}_{i,t}$为第i个样本在第t次迭代后的预测值；$L(y_i,\hat{y}_{i,t})$为预测值与真实值的损失；n表示训练样本的容量，由于测试样本的预留，选择$n\leqslant N$；$\Omega\left(f_k\right)$为建立第k棵回归树的复杂度惩罚项，为了防止决策树过拟合，其常用的表示形式为：

$$\Omega\left(f_k\right)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}W_j^2 \tag{3}$$

其中：γ和λ为超参数，用来控制惩罚力度；T为树

结构中的叶子节点数；W_j表示第j个叶子节点的权重。

XGBoost采用前向分步算法优化目标函数，在第t次迭代中，$\hat{y}_{i,t}$可表示为：

$$\hat{y}_{i,t}=\hat{y}_{i,t-1}+f_t(x_i)=\hat{y}_{i,t-1}+W_{q(x_i)} \tag{4}$$

其中：$f_t(x_i)$表示第t次迭代模型的预测值。

XGBoost目标函数可以等价转化为对树结构中各个叶子节点的计算并遍历求和，表示为

$$L_t=\sum_{j=1}^{T}\sum_{i\in I_j}L\left(y_i,\hat{y}_{i,t}\right)+\sum_{k=1}^{t-1}\Omega\left(f_k\right)+\Omega\left(f_t\right)=$$

$$\sum_{j=1}^{T}\sum_{i\in I_j}L\left(y_i,\hat{y}_{i,t-1}+W_{q(x_i)}\right)+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}W_j^2+C=$$

$$\gamma T+\sum_{j=1}^{T}\left[\sum_{i\in I_j}L\left(y_i,\hat{y}_{i,t-1}+W_j\right)+\frac{1}{2}\lambda W_j^2\right]+C \tag{5}$$

其中C为常数项。

对式(5)所示的目标函数进行二阶泰勒展开后得到：

$$L\left(y_i,\hat{y}_{i,t-1}+W_j\right)\approx L\left(y_i,\hat{y}_{i,t-1}\right)+L'\left(y_i,\hat{y}_{i,t-1}\right)W_j+\frac{1}{2}L''\left(y_i,\hat{y}_{i,t-1}\right)W_j^2 \tag{6}$$

令$L'\left(y_i,\hat{y}_{i,t-1}\right)=g_i$，$L''\left(y_i,\hat{y}_{i,t-1}\right)=h_i$，最终的目标函数为：

$$L_t=\gamma T+\sum_{j=1}^{T}\left[\sum_{i\in I_j}\left(g_iW_j+\frac{1}{2}h_iW_j^2\right)+\frac{1}{2}\lambda W_j^2\right]=$$

$$\gamma T+\sum_{j=1}^{T}\left[W_j\sum_{i\in I_j}g_i+\frac{1}{2}W_j^2\left(\lambda+\sum_{i\in I_j}h_i\right)\right] \tag{7}$$

令$\sum_{i\in I_j}g_i=G_j$，$\sum_{i\in I_j}h_i=H_j$，当$W_j=-G_j/(\lambda+H_j)$时，得到的目标函数为最小值，表示为：

$$\min L_t=\gamma T+\sum_{j=1}^{T}\left[-\frac{G_j^{\ 2}}{\lambda+H_j}+\frac{1}{2}\frac{G_j^{\ 2}}{H_j^{\ 2}}\left(\lambda+H_j\right)\right]=$$

$$\gamma T-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^{\ 2}}{\lambda+H_j} \tag{8}$$

关于最优特征以及最优划分点的选取，本文采用精确贪心算法，从深度为0的回归树开始，假设进行一次分枝，该节点在划分前的G_j和H_j是已知的，因此其最优节点值及其对应的损失值也是已知的。在划分节点时，将G_j、H_j分别划分为左节点G_L、H_L和右节点G_R、H_R，从而求出节点值W^*和分裂前后的损失值O_f^*、O_b^*，表示为：

$$W^*=-\frac{G_j}{H_j+\lambda} \tag{9}$$

$$O_f^*=\gamma-\frac{1}{2}\frac{G_j^2}{H_j+\lambda}=\gamma-\frac{1}{2}\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda} \tag{10}$$

$$O_b^*=\gamma T-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^{\ 2}}{\lambda+H_j}=2\gamma-\frac{1}{2}\left(\frac{G_L^{\ 2}}{\lambda+H_L}+\frac{G_R^{\ 2}}{\lambda+H_R}\right) \tag{11}$$

根据不同的划分条件所得到的不同节点，其节点值是不同的，特征数m较多或样本量N较大时还需要采用列采样技术和分桶技术。精确贪心算法通过遍历每个特征下每个可能的切分点取值，计算切分后的增益，选择增益最大的特征及切分点。最优特征划分为

$$L_{\text{split}}=O_f^*-O_b^*=\frac{1}{2}\left[\frac{G_L^{\ 2}}{H_L+\lambda}+\frac{G_R^{\ 2}}{H_R+\lambda}-\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{12}$$

另外，本算法终止规则同其他树模型一样可考虑最大信息增益、叶子节点包含的样本数、限制叶子节点个数等实现。

3 结果与讨论

3.1 模型评判指标

分类模型常用的评价指标为预测结果的准确率，对于不平衡样本的机器学习，抽样技术优劣的评价指标为准确率提升度，分类结果偏向多数类。本文所选信用卡用户违约数据中多数类为无逾期行为，这样会出现更多有逾期行为的用户被划分为无逾期行为用户的现象，对银行来说会造成更大的损失。因此本文引入准确率、精确率、F1指数、曲线下面积(area under curve，AUC)4个指标以评价模型的优劣。

3.2 模型对比及结果分析

首先考虑不使用SMOTE处理数据的情形，选择其他6种模型(RF、GBDT、LR、KNN、LightGBM、NN)与XGBoost模型进行比较。实验结果如表1所示。

然后，分别选择3种SMOTE对数据进行预处理，并使用上述模型进行建模，分析其耦合效果，结果如表2~表4所示。

对比模型预测的整体性能，经SMOTE采样后，数据达到相对均衡，大多数实验效果有了显著提升。RF、XGBoost、LightGBM模型的预测准确率有了大幅度提升，其中提升最大的为XGBoost模型，从0.813 474最高提升到0.877 119，提升近6%。除了LR模型有所下降外，其他模型的精确

表1　不使用SMOTE采样时数据模型指标评价对比

Tab. 1　Comparison of data model metric evaluation without SMOTE sampling

模型	准确率	精确率	F1指数	AUC值
RF	0.813 697	0.661 247	0.466 688	0.653 317
GBDT	0.821 715	0.706 052	0.478 671	0.659 020
LR	0.809 020	0.736 132	0.364 108	0.608 274
KNN	0.789 087	0.555 285	0.419 018	0.628 874
XGBoost	0.813 474	0.661 805	0.464 343	0.652 127
LightGBM	0.818 263	0.684 944	0.474 565	0.657 139
NN	0.809 465	0.635 976	0.465 813	0.653 024

表2 使用Regular-SMOTE采样时数据模型指标评价对比

Tab. 2 Comparison of data model metrics evaluation using Regular-SMOTE sampling

模型	准确率	精确率	F1指数	AUC值
RF	0.845 075	0.868 936	0.841 921	0.845 379
GBDT	0.797 225	0.855 199	0.782 208	0.798 038
LR	0.670 696	0.683 036	0.665 990	0.670 918
KNN	0.775 409	0.731 805	0.797 785	0.774 332
XGBoost	0.872 041	0.919 917	0.865 942	0.872 615
LightGBM	0.863 458	0.910 888	0.856 864	0.864 037
NN	0.718 046	0.740 450	0.709 206	0.718 445

表3　使用Borderline-SMOTE采样时数据模型指标评价对比

Tab. 3　Comparison of data model metrics evaluation using Borderline-SMOTE sampling

模型	准确率	精确率	F1指数	AUC值
RF	0.864 030	0.858 407	0.863 522	0.864 074
GBDT	0.795 937	0.807 495	0.789 307	0.795 708
LR	0.635 648	0.622 702	0.645 659	0.635 980
KNN	0.779 773	0.717 815	0.804 471	0.781 065
XGBoost	0.875 402	0.911 517	0.868 210	0.874 957
LightGBM	0.867 606	0.901 520	0.860 186	0.867 174
NN	0.719 119	0.717 669	0.715 538	0.719 064

表4　使用SVM-SMOTE采样时数据模型指标评价对比

Tab. 4　Comparison of data model metrics evaluation using SVM-SMOTE sampling

模型	准确率	精确率	F1指数	AUC值
RF	0.857 020	0.875 076	0.851 783	0.856 759
GBDT	0.814 462	0.851 551	0.801 682	0.813 916
LR	0.724 269	0.731 024	0.715 771	0.724 047
KNN	0.798 441	0.755 572	0.811 555	0.799 187
XGBoost	0.877 119	0.910 424	0.870 476	0.876 705
LightGBM	0.869 823	0.904 807	0.862 392	0.869 383
NN	0.761 534	0.795 878	0.743 302	0.760 919

率都有了极大提升，XGBoost模型的精确率最高，可达0.919 917，提高了近25%。F1指数和AUC值则对于所有模型都具有很大提升。因此，进一步说明了对类别不平衡数据利用采样策略进行数据平衡的重要性。

此外，从不同模型的实验结果可知，XGBoost模型在准确率、精确率、F1指数和AUC值上都取得了最优值，各项指标几乎都优于其他模型。而在3种SMOTE方法中，采用SVM-SMOTE方法的XGBoost模型结果要优于其余两种采样方法，该模型能够作为解决信用卡违约预测问题的优良模型。同时也发现LR、NN模型表现欠佳，说明SMOTE在对数据进行均衡处理时并不是对任何算法都有效。

3.3　对模型与算法的思考

由于SMOTE是随机线性插值，因此通常会存在以下问题：

（1）在取样过程中，k近邻选择存在一定的盲目性，当数据集极度不平衡时，新合成的少数类样本会造成类之间出现重叠，从而影响分类的效果。

（2）在插值过程中，合成的新样本基于所有少数类样本，包括噪声样本，因此容易产生模糊边界、分布边缘化等问题。当少数类样本处于分界线边缘或少数类样本为噪声样本时，合成的新样本也将处于分类边界，进而模糊分类边界。

鉴于SMOTE的上述问题，首先，采样规模必须得到有效控制，可以将采样比例作为超参数并用交叉验证方法得到。其次，宜采用“内插”的方法获取新样本，减少获取“边界”样本数量，甚至人为地过滤掉部分边界样本数据，本文实验结果说明SVM-SMOTE方法更加稳健。再次，由于SMOTE增加了新样本，并且不可避免地产生失真的样本，这样会使训练的模型相对于真实数据产生过拟合，使得在测试集上的预测准确度下降，并且这种准确度的下降程度还可能与采用的模型优化算法密切相关。本文结果显示，对于NN算法，当其隐藏层层数较多时，由于该算法天然具有过拟合的“特质”，而SMOTE又会添加新的样本导致该算法预测精度下降，该结果说明NN算法在本数据集上不宜与SMOTE结合使用。

4　结　论

本文采用Kaggle国外某银行信用卡数据集，为验证SMOTE在处理不平衡数据时的有效性，首先对原数据集用RF、GBDT、LR、KNN、XGBoost、LightGBM、NN等7种模型进行建模分析和预测，为评价模型的优劣引入准确率、精确率、F1指数、AUC值，之后分别使用Regular-SMOTE、

Borderline-SMOTE和SVM-SMOTE等3种采样方式对数据集做相对平衡处理，然后再使用上述7种模型进行数据建模分析和预测。结果表明数据经SMOTE采样后达到相对均衡，大部分实验效果均具有显著提升。3种采样方式中SVM-SMOTE的提升最大。而在7种模型中XGBoost模型准确度及精度表现最优，模型具有较强的稳定性，该模型在研究信用卡违约预测领域具有一定可行性。

参考文献

[1] BRAUSE R, LANGSDORF T, HEPP M. Neural data mining for credit card fraud detection[C]//Proceedings 11th International Conference on Tools with Artificial Intelligence. Piscataway, NJ: IEEE, 1999: 103-106.

[2] 范巍强，刘暾东. 基于BP神经网络的信用卡违约风险预测[J]. 电脑知识与技术，2011，7(10)：2348-2349.

[3] 郭建山，钱军浩. 基于随机森林的信用卡违约预测研究[J]. 现代信息科技，2020，4(3)：1-4，9.

[4] 杨磊，姚汝婧. 基于Transformer的信用卡违约预测模型研究[J]. 计算机仿真，2021，38(8)：440-444.

[5] DOUZAS G, BACAO F. Effective data generation for imbalanced learning using conditional generative adversarial networks [J]. Expert Systems with Applications, 2018, 91: 464-471.

[6] YI H K, JIANG Q C, YAN X F, et al. Imbalanced classification based on minority clustering synthetic minority oversampling technique with wind turbine fault detection application [J]. IEEE Transactions on Industrial Informatics, 2021, 17(9): 5867-5875.

[7] 刘志函，张忠林，赵磊. 面向不平衡数据分类的DPC-SMOTE过采样算法[J]. 哈尔滨理工大学学报，2024，29(6)：45-60.

[8] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[9] 张梦，陈旭勇，彭元林，等. 基于改进合成少数类过采样技术的非概率可靠性指标解[J]. 武汉工程大学学报，2024，46(2)：231-236.

[10] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C]//International Conference on Intelligent Computing. Berlin, German: Springer, 2005: 878-887.

[11] NGUYEN H M, COOPER E W, KAMEI K. Borderline over-sampling for imbalanced data classification [J]. International Journal of Knowledge Engineering and Soft Data Paradigms, 2011, 3(1): 4-21.

[12] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016: 785-794.

[13] 刘佩. 基于XGBoost算法的医保信息系统入侵安全风险监测方法[J]. 中国医疗设备，2024，39(5)：61-65.

[14] 朱小平，张丽英，刘静，等. 基于XGBoost的自动驾驶汽车事故风险预测研究[J]. 时代汽车，2024(6)：187-189.

[15] 胡晓东，吕铭春，阿克弘，等. 基于优化XGBoost算法的电信用户流失预测[J]. 科技与创新，2024(10)：36-39，44.

本文编辑：霍淑利