马文敏. 基于 k-Stratified SMOTE-CV 与 Stacking 集成学习的信贷违约预测[J]. 智能计算机与应用,2024,14(6):145-152.
DOI:10.20169/j.issn.2095-2163.240620

基于 k-Stratified SMOTE-CV 与 Stacking 集成学习的信贷违约预测

马文敏
(福建师范大学 数学与统计学院, 福州 350117)

摘 要: 信贷违约数据通常呈现不平衡分布,会导致模型过拟合和分类效果不佳等问题。为了应对这些问题,提出一种创新方法:将 k-Stratified SMOTE-CV 技术与 Stacking 集成模型相结合。并针对 3 个不同的信贷数据集展开研究。实验结果表明,k-Stratified SMOTE-CV 技术能够有效解决过拟合问题,同时 Stacking 集成模型能够进一步增强正负类别的分类效果。其中, *F*1 得分最大提升 11.7%, *AUC* 最大提升 7.6%。此外,为深入理解模型的决策过程,引入了局部可解释性模型 LIME,增强信贷违约预测的透明性。这些研究结果为金融领域的信贷决策提供了有力支持。

关键词: 不平衡数据; 信贷违约预测; 集成模型; 可解释性
中图分类号: TP391 **文献标志码**: B **文章编号**: 2095-2163(2024)06-0145-08

Credit default prediction based on k-Stratified SMOTE-CV with Stacking integrated learning

MA Wenmin
(School of Mathematics and Statistics, Fujian Normal University, Fuzhou 350117, China)

Abstract: Credit default data usually presents an unbalanced distribution, which can lead to problems such as model overfitting and poor classification. In order to cope with these problems, an innovative approach is proposed: combining the k-Stratified SMOTE-CV technique with the Stacking integration model. The experimental results show that the k-Stratified SMOTE-CV technique can effectively solve the overfitting problem, while the Stacking integration model can further enhance the classification effect of positive and negative categories. Among them, the *F*1 score is maximally enhanced by 11.7% and the *AUC* is maximally enhanced by 7.6%. In addition, a locally interpretable model, LIME, is introduced to enhance the transparency of credit default prediction for a deeper understanding of the model's decision-making process. These findings provide strong support for credit decision making in the financial sector.

Key words: unbalanced data; credit default prediction; integrated modeling; interpretability

0 引 言

随着金融科技的发展,中国个人信贷行业已经进入数字化时代[1]。贷款需求量不断增加,金融产品呈多样化发展,金融机构将面临更加复杂的情况。因此,信用贷款对于金融机构而言也伴随着一定的风险。这种风险可能导致金融机构遭受损失,降低相应的盈利能力。因此,信贷违约风险一直是金融机构和经济决策者关注的问题。

为了降低信贷风险,金融机构通常会依赖大量的信贷数据和机器学习技术进行信贷违约风险预测。一些表现较好的机器学习模型包括逻辑回归[2]、支持向量机[3]等。这些模型已经被广泛用于提高违约风险的识别和管理。Klafft[4]使用逻辑回归模型在借贷领域中发现,借款人的信用评级是对借款利率影响最大的因素之一。Gao 等学者[5]提出使用决策树 C5.0 算法对信贷数据进行分析,提升了信贷风险预测的性能。Fan 等学者[6]提出了一种基于支持向量机的信用评分模型,并通过自适应突变部分蜂群算法进行了优化。实验结果表明,提出的模型达到了较高的预测精度。

近年来,随着科学技术的不断发展,学者们在信

作者简介: 马文敏(1999-),女,硕士研究生,主要研究方向:机器学习,数据挖掘。Email:1076716971@qq.com

收稿日期: 2023-11-28

贷违约预测问题上开始广泛使用集成学习方法。与单一的机器学习模型相比,集成学习方法具有更高的预测能力和更广泛的适用性。Breiman[7] 使用 Bagging 方法改进了决策树,提出随机森林算法,这一代表性的集成学习算法以其高准确性和泛化能力而著称,因此在信贷风险领域中得到了广泛应用。马春文等学者[8]使用随机森林分类模型,筛选出了最重要和最有效的因子组合,对网络借贷的信用风险因素进行了深入分析。此外,Chen 等学者[9]提出了一种强大且高效的机器学习算法—XGBoost。XGBoost 以其在大规模和高维数据集上出色的性能而闻名。周永圣等学者[10]使用 XGBoost-RF 模型评估个人信用风险,取得了可行性的结果。这些研究和方法为信贷违约预测领域提供了有力的工具和框架,有助于金融机构更好地管理和控制风险,既拥有传统的个人风险评价模型,又能够利用大数据技术提升自己的风险管理能力,促进自身不断地发展前进[11]。但是却在解释性方面仍存在一定挑战。

此外,在信贷违约预测研究中,通常存在样本类别不平衡的问题,这可能导致模型在训练和性能评估时面临不小挑战。虽然一些集成算法能够在不平衡数据中获得良好的识别效果,但是过多的基分类器可能导致模型过拟合,准确率下降,并增加了算法的复杂度。

针对上述问题,本文在数据重采样方法和分类算法两个方面进行改进和优化,同时引入了可解释性模型,以增强本文模型的可信度。这些改进主要体现在以下 4 个方面:

(1)在 k 折分层交叉验证的框架下使用 SMOTE 技术的方法(k-Stratified SMOTE-CV),以应对类别不平衡和过拟合问题。这一方法为处理不平衡数据提供了一种创新思路;

(2)构建 Stacking 集成学习模型,旨在增强模型的学习效果,同时避免模型过于复杂。这一改进有助于更准确地识别少数类样本,进一步提高模型的准确性和鲁棒性;

(3)为解决模型的可解释性问题,引入局部可解释性模型 LIME。这个方法能够对复杂模型的预测结果提供可信赖的解释,为深入理解模型决策提供了新的途径;

(4)本研究以 3 个信贷数据集为基础,更客观合理地分析了数据类别不平衡问题。特别关注了不同方法平衡处理的效果对比,以及集成学习与单一模型的比较研究。

1 相关原理及技术

1.1 k-Stratified SMOTE-CV 技术

本文提出 k-Stratified SMOTE-CV 方法,以解决信贷违约预测中的数据类别不平衡问题。通过 k 折分层交叉验证将数据集分为 k 个折叠,并确保每个折叠中的样本类别比例与原始数据集中各个类别的比例分布保持一致,再使用 SMOTE 技术在每个折叠上合成少数类样本,使得每个折叠的正负样本平衡分布,从而能够更加公平评估模型性能并防止过拟合。

1.1.1 SMOTE 技术

SMOTE[12]是一种处理类别不平衡的过采样方法,通过合成新的少数类样本,使其数量达到与多数类样本相等,以实现数据集的平衡化。这一方法有助于改善模型在不平衡数据中的性能表现。SMOTE 算法运行过程如下:

(1)从少数类样本中选取一个样本 x_i;

(2)以欧氏距离为标准,计算 x_i 的 K 近邻样本 x_{zi};

(3)根据不平衡比例确定需要合成的样本个数 n;

(4)根据需合成的样本数 n, 依次在 x_{zi} 和 x_i 之间的连线上任取一点随机合成新样本。

SMOTE 少数类样本合成方法,方法公式如下:

$$x_n = x_i + \lambda \times (x_{zi} - x_i) \tag{1}$$

其中, x_n 表示合成的少数类样本; x_i 表示选取的某一少数类样本; x_{zi} 表示 x_i 的 K 最近邻样本; λ 表示介于 0 到 1 之间的随机数。

1.1.2 k 折分层交叉验证

k 折分层交叉验证(k-fold Stratified Cross-Validation)[13]是一种用于评估机器学习模型性能的常用技术。通过将 k 折交叉验证和分层抽样方法相结合,旨在有效评估模型的泛化性能。通过多次训练和测试模型,有助于减小训练和测试数据的偶然性,从而提供更可靠的性能评估。此外,通过分层抽样,确保了在不同折叠中各个类别的样本都能得到适当的代表,特别适合解决类别不平衡问题。k 折分层交叉验证方法如图 1 所示。

1.2 Stacking 集成学习模型

Stacking[14]集成学习的核心思想是将多个不同的基本模型(弱模型)的输出作为输入,然后通过训练一个元模型来融合这些模型的预测。这能够充分发挥不同学习器的优势,以提高整个模型的预测准确性。这一方法旨在克服单个模型的局限性,通过综合多个模型的意见来获得更可靠的预测结果。

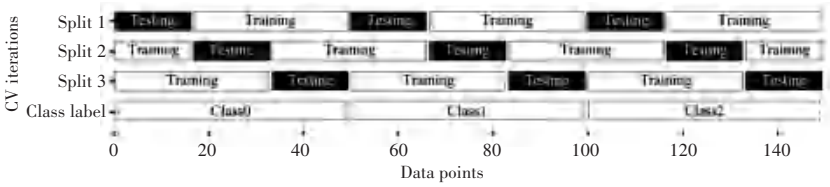

图 1　*k* 折分层交叉验证方法

Fig. 1　*k*-fold stratified cross-validation

本文通过构建 Stacking 集成学习模型，进一步提升模型的预测能力，同时避免模型过于复杂的问题。算法流程如下。

算法 1　Stacking 算法

输入　训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$；初级学习算法 $L_1,L_2,\cdots,L_T$；次级学习算法 L

输出　预测结果 $H(x)=h'(h_1(x),h_2(x),\cdots,h_T(x))$

过程：

1. for $t=1,2,\cdots,T$ do
2. $\quad h_t=L_t(D)$;
3. end for
4. $D'=\varnothing$;
5. for $i=1,2,\cdots,m$ do
6. $\quad$ for $t=1,2,\cdots,T$ do
7. $\quad\quad z_{it}=h_t(x_i)$;
8. $\quad$ end for
9. $\quad D'=D'\cup((z_{i1},z_{i2},\cdots,z_{iT}),y_i)$;
10. end for
11. $h'=L(D')$;

算法中，首先输入包含自变量 x_i 和目标变量 y_i 的训练数据集 D。在此操作后，选择 T 个初级学习算法 $L_1,L_2,\cdots,L_T$，分别用于生成初级学习器 h_t。同时，创建一个空数据集 D'。接下来，将训练数据集分别输入每个初级学习器，获取运算后的预测结果 z_{it}，并将这些结果与真实目标变量标签一起存储在数据集 D' 中。最后，在数据集 D' 上使用次级学习算法 L，以生成次级学习器 h'，并得到最终的预测结果 $H(x)$ 作为输出。

基于以上流程，本文从常见的机器学习模型中筛选一组初级学习模型，包括传统模型（K 近邻（KNN）、逻辑回归（LR）），以及集成算法（随机森林（RF）、梯度提升树（GBDT）、XGBoost、LightGBM 和 CatBoost）。研究中对每个模型进行了独立的训练，并使用准确率、*AUC* 等指标来评估各模型的预测性能。然后，在初级学习层的基础上，将性能稳定、功能强大的 LR 模型作为元模型。LR 模型在各种情况下都表现稳定，适合用于组合不同基学习器的输出，有助于构建一个具有更高实用性的 Stacking 集成模型。这一策略能够充分利用不同模型之间的优势，从而提高模型整体的预测能力。

1.3　LIME 模型解释机制

在信贷违约预测领域，众多模型通常缺乏可解释性，这使得理解模型的决策过程变得困难。LIME 模型[15]为研究者提供了一种新的视角，可以更深入地探究信贷违约预测模型的运作方式。通过 LIME 模型，能够将黑盒模型的预测结果转化为可信赖的解释，从而使用户对预测结果产生充分信任。这一可解释性工具有助于建立金融机构、客户与信贷违约预测模型之间的信任关系，进一步提高了决策的透明度和可信度。

LIME 模型旨在为黑盒模型提供局部可解释性，即解释模型在特定实例上的预测结果。LIME 模型的核心思想是通过局部逼近来解释模型的行为。LIME 模型的核心算法实现如下。

（1）目标函数 $\xi(x)$。可由式（2）表示为：

$$\xi(x)=\arg\min_{g\in G}L(f,g,\pi_x)+\Omega(g) \tag{2}$$

其中，f 表示需要解释的模型；g 表示预测模型；G 表示所有预测模型的集合。

（2）相似度 $\pi_x(z)$，为实例 z 与 x 之间的接近度。可由式（3）表示为：

$$\pi_x(z)=\exp\left\{-\frac{D(x,z)^2}{\sigma^2}\right\} \tag{3}$$

（3）损失函数 $L(f,g,\pi_x)$，为 g 在局部逼近 f 的度量，在保持模型复杂度 $\Omega(g)$ 足够低的情况下，通过最小化损失函数 L 得到目标函数 ξ 的最优解。可由式（4）表示为：

$$L(f,g,\pi_x)=\sum_{z,z'\in Z}\pi_x(z)\{f(z)-g(z')\}^2 \tag{4}$$

其中，z 表示原始数据集中一个被扰动的样本点，z' 表示通过对该样本随机扰动产生的新样本点的集合。

1.4　模型评价指标

混淆矩阵是用于评估二分类模型性能的重要

工具,将模型的预测结果与真实标签进行比较,提供了关于模型性能的详细信息。混淆矩阵的基本结构见表 1。

表 1 混淆矩阵

Table 1 Confusion matrix

		预测	
		正例	负例
实际	正例	True Positive (*TP*)	True Negative (*TN*)
	负例	False Positive (*FP*)	False Negative (*FN*)

然而,混淆矩阵并不能很好地比较不同模型的效果,因此衍生出了其他评价指标。这些指标提供了比混淆矩阵更具一般性的性能度量,适用于各种不同的问题和数据集。评价指标及数学计算公式见表 2。

表 2 评价指标

Table 2 Evaluation indicators

指标	计算公式
准确率	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
精确度	$Precision = \frac{TP}{TP + FP}$
召回率	$Recall = \frac{TP}{TP + FN}$
*F*1 得分	$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$
AUC 值	*ROC* 曲线下的面积

1.5 总结

本文将 k-Stratified SMOTE-CV 技术与 Stacking 集成学习模型相结合对信贷违约识别进行预测,最后使用 LIME 模型对样本实例进行解释。整体模型结构如图 2 所示。

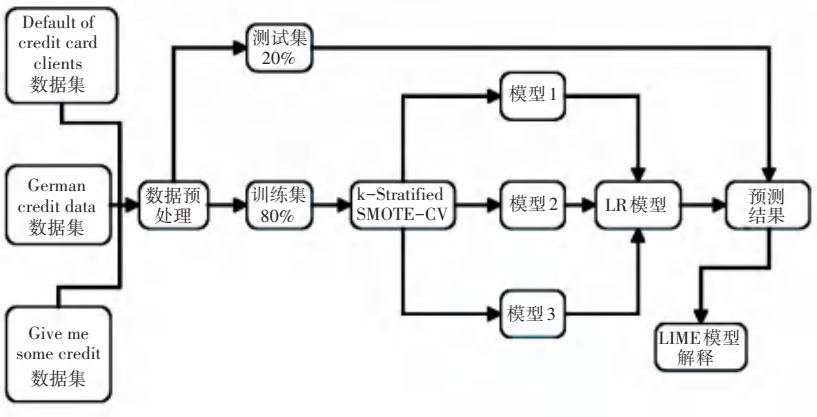

图 2 信贷风险识别预测模型结构图

Fig. 2 Credit risk identification prediction model structure

2 实验数据分析

2.1 数据获取

为了验证本文所提出模型的有效性,选取了来自于 UCI 和 Kaggle 网站上的 3 个信贷数据集,分别是 German Credit Data (GCD)[16]、Default of Credit Card Clients(DCCC)[17]和 Give Me Some Credit(GMSC)[18]。这 3 个数据集是来自不同国家的信用贷款数据,并且具有不同的特征和训练难度。因此,通过使用多样化的数据集,可以更好地验证本文提出的算法在不同国家的信贷数据上的适用性。

表 3 展示了 3 个信贷数据集的详细信息,包括数据集名称、样本数量、特征数量、违约样本数量和履约样本数量。

表 3 数据集描述

Table 3 Description of the data set

数据集名称	样本数量	特征数量	违约样本数量	履约样本数量
GCD	1 000	21	300	700
DCCC	30 000	25	23 364	6 636
GMSC	150 000	11	139 974	10 026

2.2　数据预处理

首先,进行数据预处理,处理数据中的缺失值和重复值。对于缺失值,采用均值填充的方法,以保持数据整体的分布特性。同时,检测并删除重复数据,以避免对模型产生不良影响。其次,对分类变量进行处理,通过独热编码将其转化为数字形式,以便于模型的训练和分析。然后,采用 Z-score 标准化,将连续性数据转换为具有均值为 0 和标准差为 1 的标准正态分布,以确保数据特征归一化在相同的尺度上。最后,将数据集划分为训练集和测试集,按照 8:2 的比例从多数类样本和少数类样本中随机划分,以确保训练集和测试集都包含足够的样本类别,有助于模型能够进行有效的学习和评估。

2.3　实证检验

2.3.1　平衡处理效果对比

为了评估平衡处理方法的效果,分别对使用 SMOTE 技术处理的数据和使用 k-Stratified SMOTE-CV 技术处理的数据进行 K 近邻和随机森林模型的训练。随后,计算处理后结果的准确率和 *AUC* 值,有利于评估各方法的性能和效果,其结果见表 4。

表 4　平衡处理效果比对

Table 4　Comparison of equilibrium treatment effects

模型	平衡处理方法	数据集	GCD		DCCC		GMSC	
			准确率	*AUC*	准确率	*AUC*	准确率	*AUC*
KNN	SMOTE	训练集	0.738	0.851	0.786	0.871	0.895	0.954
		测试集	0.640	0.671	0.654	0.633	0.828	0.662
RF		训练集	0.830	0.920	0.873	0.939	0.950	0.989
		测试集	0.795	0.730	**0.802**	0.670	0.921	0.642
KNN	k-Stratified	训练集	0.620	0.697	0.653	0.678	0.826	0.711
	SMOTE-CV	测试集	0.665	0.755	0.655	0.674	0.827	0.700
RF		训练集	0.746	0.792	0.806	0.756	0.923	0.829
		测试集	**0.820**	**0.812**	0.800	**0.757**	**0.921**	**0.831**

从表 4 中可以看出,使用 SMOTE 技术处理后的 K 近邻和随机森林模型在训练集上表现较好,但在测试集上效果明显下降,这表明在训练集上可能发生了过拟合。而本文所提出的 k-Stratified SMOTE-CV 技术对训练集进行平衡处理后,K 近邻和随机森林模型测试集上的准确率和 *AUC* 值都有显著提升。这表明通过采用 k-Stratified SMOTE-CV 方法,成功地处理了类别不平衡问题,有助于提高模型的性能。更为重要的是,这些模型没有出现过拟合问题,表明将能够更好地泛化到未见数据。因此,k-Stratified SMOTE-CV 是一个强大的工具,现如今则更适用于处理各个类别的不平衡情况。

此外,在处理类别不平衡问题时,随机森林模型的性能明显优于 K 近邻模型。随机森林具有更好的抗过拟合能力和更强的泛化能力,能够更有效地应对复杂的数据模式和特征。因此,考虑使用随机森林等性能优越的集成模型可能会带来更好的结果。这些改进可以增强模型在信贷违约识别等领域的实际应用性能。

2.3.2　Stacking 集成模型与其他模型比对

为了提升 Stacking 集成模型的性能,本文采用了多样化的策略。首先,在选择基学习器时,优先考虑具有较强学习能力和差异性的基分类器;其次,在选择元模型时,使用更稳定的 LR 模型。

基于上述策略,本文选择训练效果最佳的前 3 个模型作为基学习器,并使用 LR 模型作为元模型构建 Stacking 集成模型。通过在 3 个不同数据集上进行实验,可以看出,本文所采用的 Stacking 集成学习模型在综合性能评价指标上均取得了显著的提升,具体数据见表 5~表 7。

表 5　GCD 数据集实验结果对照表

Table 5　Comparison of experimental results for the GCD dataset

指标	LR	RF	GBDT	XGBoost	LightGBM	CatBoost	Stacking
准确率	0.705	**0.820**	0.755	0.775	0.765	0.770	0.815
*F*1	0.683	0.774	0.717	0.728	0.719	0.729	**0.777**
AUC	0.805	0.812	0.800	0.810	0.794	0.812	**0.832**

表 6 DCCC 数据集实验结果对照表

Table 6 Comparison of experimental results for the DCCC dataset

指标	LR	RF	GBDT	XGBoost	LightGBM	CatBoost	Stacking
准确率	0.770	0.800	0.808	0.806	0.810	**0.812**	0.809
*F*1	0.687	0.682	0.687	0.674	0.682	0.681	**0.695**
AUC	0.760	0.757	0.765	0.744	0.762	0.766	**0.767**

表 7 GMSC 数据集实验结果对照表

Table 7 Comparison of experimental results for the GMSC dataset

指标	LR	RF	GBDT	XGBoost	LightGBM	CatBoost	Stacking
准确率	0.763	0.921	0.901	0.930	0.926	**0.932**	0.923
*F*1	0.562	0.650	0.674	0.645	0.668	0.638	**0.679**
AUC	0.779	0.831	0.850	0.841	0.847	0.849	**0.855**

从表 5~表 7 可看出,与单一模型相比,Stacking 集成模型在 *F*1 得分和 *AUC* 值上都取得了显著的改善,其中 *F*1 得分最大提升 11.7%, *AUC* 最高提升 7.6%。虽然 Stacking 集成模型准确率略微降低,但 *F*1 得分和 *AUC* 指标的显著提升表明在这种情况下要更多关注避免第二类错误的发生,即错过正例的情况。考虑到信贷数据通常存在类别不平衡问题,使得第二类错误(cost-II)所导致的损失显著大于第一类错误(cost-I),证明了采用 Stacking 集成模型的合理性。此外, *AUC* 的提升还表明模型在区分正负类别方面的性能得到了明显改善,证实了本文所提出的模型有效提升了预测的准确度。

2.3.3 预测结果解释

构建信贷违约预测模型的目的是评估客户未来违约的潜在风险。根据上述实验分析,通过结合 k-Stratified SMOTE-CV 技术和 Stacking 集成学习方法,有效提升了模型预测的准确性。然而,对于被预测为违约的客户,需要了解为何该客户被预测为违约、而其他客户没有被预测为违约同样具有重要意义。同时,风险管理部门也需要对模型的预测结果进行可信度评估。因此,本文采用 LIME 模型来深入研究不同用户特征对预测结果的影响,以解释模型预测的结果,并进一步证明该模型是值得信任的。这一方法解决了复杂模型解释性不足的问题。

利用事后解释机制 LIME 模型进行可视化解释,输出了模型对这 3 个数据集中单个样本的解释结果。这些结果详细显示了影响样本预测结果的前 10 个最重要的特征,以及这些特征对预测的贡献度和相应特征值的关联。GCD、DCCC 和 GMSC 中某单个样本预测结果解释分别如图 3~图 5 所示。

由图 3 可看到,LIME 对本文所采用的预测模型进行事后解释,GCD 中某一客户有 83%的概率认为其会违约,这个高概率的违约预测是基于客户现有的支票账户状态、借款目的、月持续时间等多个因素的综合考虑。这意味着模型认为,根据这些特征的组合,客户可能存在违约的风险。

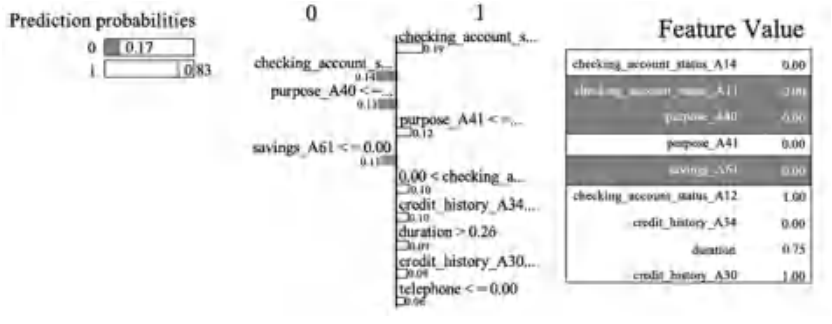

图 3 GCD 中某单个样本预测结果解释

Fig. 3 Interpretation of the prediction results for a single sample of GCD

由图 4 可看到,预测该客户会有 68%的概率出现违约行为。通过解释结果,可以得到该客户的受教育程度被标记为未知,这可能表明目前仍缺乏关于客户教育背景的信息,而这通常是信用评估的一

个重要因素。其次，客户曾经有过 6 个月的延迟还款记录，这可能被视为不良信用历史，对信用评分产生了负面影响。此外，客户的婚姻状况被标记为已婚，这可能在模型中产生了一些额外的负面权重。以上几个因素对于这一预测结果产生了负面影响。

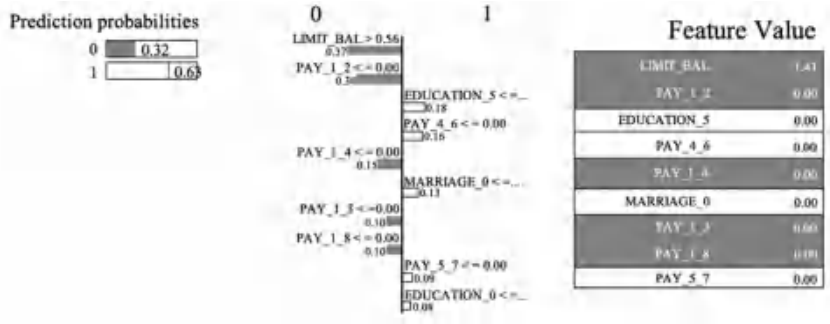

图 4　DCCC 中某单个样本预测结果解释

Fig. 4　Interpretation of the prediction results for a single sample of DCCC

由图 5 可看到，根据本文模型判断该客户不会出现违约行为的概率高达 98%。通过解释结果，可以看出该客户逾期笔数为 0，意味着该客户没有任何未按时还款的记录，这通常被视为良好的信用历史。客户的可用信贷额度比例相对较高，这表明客户的信用额度使用率较低，这通常被视为良好的信用健康状况。最后，客户的家属数量为 1，这可能表明客户的家庭状况稳定。由此可看出可解释性的结果使客户能够更好地了解自身的信用状况，并提供了信心，表明客户的信用风险较低。

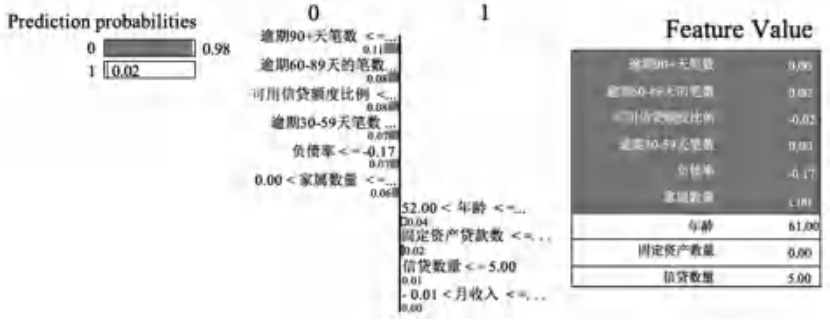

图 5　GMSC 中某单个样本预测结果解释

Fig. 5　Interpretation of the prediction results for a single sample of GMSC

3　结束语

本文的实验结果基于 3 个公开数据集，结果表明所提出的 k-Stratified SMOTE-CV 技术成功地解决了过拟合问题。同时，构建 Stacking 集成学习模型在预测准确度和泛化性能等方面表现出色。这一系列改进不仅提升了分类准确率，还大幅降低了犯第二类错误的风险。值得注意的是，引入事后解释机制 LIME 模型，显著提升了模型的可解释性。这意味着研究者不仅拥有一个强大的分类工具，还能够理解模型的决策过程，为决策提供合理的解释。尤其在需要透明决策和法规遵从性方面，这一方法对于实际应用具有至关重要的意义。

在未来的研究中，希望能够继续探索其他解释性工具和方法，以进一步提高模型的可解释性。此外，还会陆续将这种方法扩展到更广泛的领域，以解决更多不同领域中的类似问题。这个研究的成果不仅有望在当前问题领域产生影响，还为未来的数据科学和机器学习研究提供了有价值的方法和借鉴。

参考文献

[1] 秦洪涛. 数字化时代的中国个人信贷[J]. 清华金融评论，2018(1)：47-48.

[2] WANG Hong，XU Qingsong，ZHOU Lifeng. Large unbalanced credit scoring using Lasso-logistic regression ensemble[J]. PLoS One，2017，10(2)：e0117844.

[3] WEI Liwei，ZHANG Ying，LIU Mochen，et al. Credit risk evaluation using ES based SVM-MK[C]// Proceedings of the 2016 5th International Conference on Measurement，Instrumentation and Automation (ICMIA 2016). Istanbul，Türkiye：Atlantis Press，

2016:679-684.
[4] KLAFFT M. Online peer-to-peer lending: A lenders' perspective [C]//The International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government. Las Vegas, USA: dblp, 2008: 371-375.
[5] GAO Shang, WANG Changbao. Personal credit scoring based on decision tree C5.0 algorithm [C]//Proceedings of the 2017 7th International Conference on Education, Management, Computer and Society (EMCS 2017). Mesa, USA: Atlantis Press, 2017: 1729-1734.
[6] FAN Qinglan, LIU Xinxin, ZHANG Yunfeng, et al. Adaptive mutation PSO based SVM model for credit scoring [C]// The Second International Conference on Computational Science and Applications. Hohhot, China: ACM, 2018: 1-7.
[7] BREIMAN L. Random forest[J]. Machine Learning, 2001, 45: 5-32.
[8] 马春文, 赵慧, 李琪. 基于随机森林分类模型的 P2P 网络借贷标的信用风险因子研究[J]. 吉林大学社会科学学报, 2019, 59(3): 39-48, 231-232.
[9] CHEN Tianqi, HE Tong, BENESTY M. xgboost: Extreme Gradient Boosting[J]. R Package Version 0.4-2 1.4, 2015(4): 1-4.
[10] 周永圣, 崔佳丽, 周琳云, 等. 基于改进的随机森林模型的个人信用风险评估研究[J]. 征信, 2020, 38(1): 28-32.
[11] 石菲. 大数据背景下商业银行个人信贷风险管理的完善思考[J]. 农村经济与科技, 2019, 30(20): 143-144.
[12] CHAWLA V N, BOWYER W K, HALL O L, et al. SMOTE: Synthetic minority over-sampling technique [J]. arXiv preprint arXiv: 1106.1813, 2011.
[13] STONE M A. Cross-validatory choice and assessment of statistical predictions[J]. Journal of the Royal Statistical Society, 1974, 36(2): 111-147.
[14] 丁岚, 骆品亮. 基于 Stacking 集成策略的 P2P 网贷违约风险预警研究[J]. 投资研究, 2017, 36(4): 41-54.
[15] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?": Explaining the predictions of any classifier[J]. arXiv preprint arXiv: 1602.04938, 2016.
[16] HOFMANN H. Statlog(German Credit Data) [EB/OL]. [1994-02-06]. https://doi.org/10.24432/C5NC77.
[17] YEH I C. Default of credit card clients[EB/OL]. [2016-01-25]. https://doi.org/10.24432/C55S3H.
[18] FUSION C, CUKIERSKI W. Give me some credit[EB/OL]. [2011-01-05]. https://kaggle.com/competitions/GiveMeSomeCredit.