基于 Logistic 回归和 XGBoost 算法的信贷风险评估研究

黄飞豪
广西大学经济学院，中国 ·广西 南宁 530004

摘要：随着个人信贷业务的不断扩展，如何从海量高维的用户信贷数据中挖掘出潜在的用户模型对用户进行信贷风险评估，以降低信贷业务风险，已经成为研究的主流。由于没有一种模型能够在信贷风险预测中始终优于其他模型，基于各种数据预处理和特征选择方法，将参数优化方法和 Logistic 回归、XGBoost 算法相结合，构建信贷风险评估模型。实验表明，与 Logistic 回归方法相比，XGBoost 算法在参数优化后，能够取得更好的预测效果，在大数据背景下，机器学习和深度学习的方法更具有优势。

关键词：信贷风险评估；特征选择；Logistic 回归；XGBoost

Research on Credit Risk Assessment Based on Logistic Regression and XGBoost Algorithm

Huang Feihao
School of Economics,Guangxi University, China Guangxi Nanning 530004

Abstract: With the continuous expansion of personal credit business,how to mine potential user models from massive high-dimensional user credit data for credit risk assessment of users,in order to reduce credit business risks,has become the mainstream of research.Due to the fact that no model can consistently outperform other models in credit risk prediction,a credit risk assessment model is constructed by combining parameter optimization methods with logistic regression and XGBoost algorithms based on various data preprocessing and feature selection methods.Experiments have shown that compared with logistic regression methods,XGBoost algorithm can achieve better prediction performance after parameter optimization.In the context of big data,machine learning and deep learning methods have more advantages.

Keywords: Credit risk assessment; Feature selection; Logistic regression; XGBoost

0 引言

随着我国经济的快速发展和国民超前消费理念的产生以及金融服务的发展，信贷业务的发展日益加快，个人贷款成为了人民日常生活中的一个重要组成部分。金融服务机构提供的贷款服务衍生出了许多细分内容，如个人购房贷、个人助学贷、个人消费贷等，满足了消费者的各种贷款需求。与此同时，金融机构在发放贷款时，面临着两类风险：一是客户有偿的还款能力，不批准贷款将导致业务流失；二是客户无偿还贷的能力，批准贷款会面临客户可能违约导致经济损失。随着信贷业务的不断扩展，数据量剧增，各种虚假信息混杂，原有的信贷风险评估系统也需要与时俱进的进行更新，避免因使用过时的风险评估方法导致本可避免的经济损失。因此，构建一个更加全面可靠的信贷风险评估系统对保障信贷市场平稳运行具有重要意义。

我国的金融市场发展较晚，对个人信贷风险的研究与金融市场发展早的老牌金融国家还存在一定差距，经过十几年的发展，我国在信贷风险研究方面也取得了显著成果。在信贷风险评估技术的应用上，常用的信贷风险评估模型包括 logistic 回归、XGBoost 模型、LightGBM、神经网络、支持向量机等。不少学者应用 Logistic 回归对信贷风险进行预测研究，并取得了较好的效果。屈忠锋等（2024）利用 Logistic 模型对企业信贷风险进行预测，与 BP 神经网络和 Fisher 判别分析相比，Logistic 回归的效果均优于两类基准模型，且可以计算出具体的违约概率，为银行制定策略提供可靠支持[1]。蔡文学等（2017）将采用 GBDT 模型提取到的组合特征结合 Logistic 回归构建个人信贷风险评估模型，与随机森林、SVM 等基准模型相比，GBDT 和

Logistic 模型的组合取得了最好的预测效果[2]。张国政等（2015）基于商业银行个人消费贷的实际操作数据和稳定性较高的 Logistic 回归模型建立个人信用评分模型，分析各指标对个人信用风险的影响，模型取得的较好的结果。也有许多学者将 XGBoost 算法应用在信贷风险研究中[3]。伍洁等（2024）采用 XGBoost 集成学习方法搭建了一套个人贷款信用评估模型，并使用筛选得到的指标对个人信贷风险进行评估，验证了 XGBoost 算法在信贷风险评估领域的使用具有可靠性[4]。朱丽华和龙海侠（2023）提出了一种改进的麻雀算法（GCOSSA）来优化 XGBoost 参数，优化后的算法在信贷风险预测中的准确率更高[5]。廖文雄等（2019）将特征选择和 XGBoost 结合，对高维数据背景下的个人信贷风险评估进行研究，实验结果表明，基于 Embedded 思想的特征选择方法 XGBFS 能够从高维的数据中选择出重要属性，并且分类器在精确率、召回率等方面具有较为突出的性能[6]。

随着技术的发展，当前对信贷风险评估的研究中，越来越多的机器学习模型和深度学习模型被采用。本文将特征选择方法分别和 Logistic 回归、XGBoost 算法相结合，加入网格调参法，对大样本高维度数据背景下信贷风险预测模型的预测效果，拓展现有研究。

1 模型与算法

1.1 逻辑回归模型介绍

逻辑回归又叫对数几率回归，是一种广义的线性回归分析模型，常用于监督学习的二分类问题上。其原理是用 Sigmoid 函数把线性回归的结果映射到（0，1）之间，此时 P 值表示预测值取 0 或 1 的概率，即预测为正常用户和违约用户的概率。逻辑回归在 0 附近敏感，在远离 0 点的位置不那么敏感，使得逻辑回归关注的重点在分类边界，模型的鲁棒性得到增强，在面对分类问题和做预测时，线性概率模型的适用性不如逻辑回归模型[7]。

逻辑回归的累积分布函数如式 2.1 所示，其中函数 $\Lambda(\cdot)$ 的定义为 $\Lambda(z) \equiv \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$。

$$P(y=1|x) = F(x,\beta) = \Lambda(x'\beta) \equiv \frac{\exp(x'\beta)}{1+\exp(x'\beta)} = \frac{1}{1+\exp(-x'\beta)} \tag{2.1}$$

其中，$P(y=1|x)$ 为 x 用户是违约用户的概率，$P(y=0|x)$ 为 x 用户是未违约用户的概率，参数 β 通过极大似然法求取。

1.2 XGBoost 算法介绍

Gradient Boosting Decision Tree（以下简称 GBDT）机器学习算法采用了 Boosting 的原理，基于第一次计算的结果进行第二次计算以减少上一次的偏差，不断迭代决策树，以实现残差最优。陈天奇等[8] 在 GBDT 的基础上进行了工程改进，开发出 XGBoost 算法，二者都属于 Boosting。二者最大的区别在于目标函数不同，XGBoost 即加入了正则项防止过拟合，又考虑了二阶导数对代价函数进行二阶泰勒展开，以实现在训练集上更快地得到收敛。

损失函数可由模型预测输出 $\bar{y}_i$ 和样本真实标签进行表示：

$$L = \sum_{i=1}^{n} l\left(y_i, \bar{y}_i\right) \tag{2.2}$$

其中 n 为样本数量。

为了使方差最小，需要在目标函数中加入正则项，以防止过拟合。XGBoost 目标函数的定义如下：

$$Obj = \sum_{i=1}^{n} l\left(y_i, \bar{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{2.3}$$

K 表示树的数量，f_k 表示第 k 棵树，Ω 是模型复杂度函数，$\sum_{k=1}^{K} \Omega(f_k)$ 是将全部 K 棵树的复杂度进行求和，添加到目标函数作为正则项。模型的任务是找到一组树，使得 Obj 最小。第 t 棵树可以表示为：

$$\bar{y}^{(t)} = \sum_{k=1}^{t} f_k(x) = \bar{y}^{(t-1)} + f_t(x) \tag{2.4}$$

此时目标函数为：

$$\begin{aligned} Obj^t &= \sum_{i=1}^{n} l\left(y_i, \bar{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \\ &= \sum_{i=1}^{n} l\left(y_i, \bar{y}_i^{(t-1)} + f_t\left(x_i\right)\right) + \sum_{k=1}^{t-1} \Omega\left(f_k\right) + \Omega\left(f_t\right) \end{aligned} \tag{2.5}$$

在经过二阶泰勒展开后，最终的目标函数可以表示为：

$$Obj^t \approx \sum_{i=1}^{n} \left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega\left(f_t\right) \tag{2.6}$$

2 描述性统计

本节对数据来源、数据探索性分析和数据预处理进行阐述，使用来源可靠、真实的数据进行研究是研究结果真实可靠的关键组成部分，对于数据的探索性分析可以对数据的结构和分布等有初步的了解，对数据的预处理则是为后续的建模做前期准备，对于数据的预处理操作主要包括：数据清洗、通过特征工程构建和筛选特征、平衡数据。

2.1 数据来源

建立信用评估模型对信贷违约情况进行预测评估、模型的训练和目标特征的选择、模型的预测性能和最终得到的结论都需要依赖于真实来源的原始数据。要建立有效的信贷违约预测模型必须要基于我国个人信贷业务中实际发生的业务数据。

本文数据来源于阿里云天池大赛学习赛信贷违约预测板块的数据（https://tianchi.aliyun.com/），并对数据进行了脱敏处理，总原始数据量为 80 万条，包含用户背景信息、贷款行为等共 47 列字段信息，其中包含 15 列匿名变量，使用字母 n 加数字 0-14 进行命名，1 列为目标特征，标签为“default”，标签为 0 表示未违约客户，标签 1 表示违约客户。

部分数据的字段属性说明如表 3.1 所示。

2.2 数据可视化分析

2.2.1 变量类型可视化分析

首先对数据类型进行分析，数据集包含 33 个连续型变量，8 个离散型变量（将变量值的取值个数小于 10 的变量定义为离散型变量）和 1 个单值变量，在后续的处理中，将单值变量进行了剔除。其中离散型变量包括：term、homeOwnership、verificationStatus、isDefault、initialListStatus、applicationType、n11 和 n12；单值变量为 policyCode；其余为连续型变量。图 3.1 和图 3.2 展示了连续型变量（部分变量）和离散型变量的分布图。

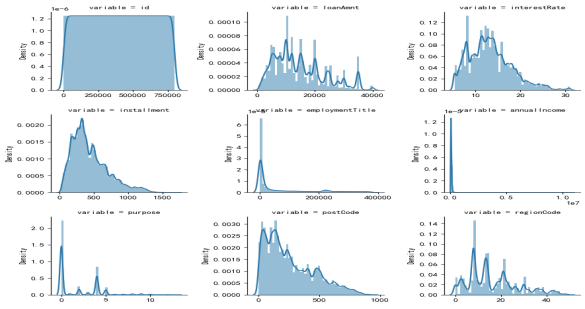

图3.1 连续型变量分布图

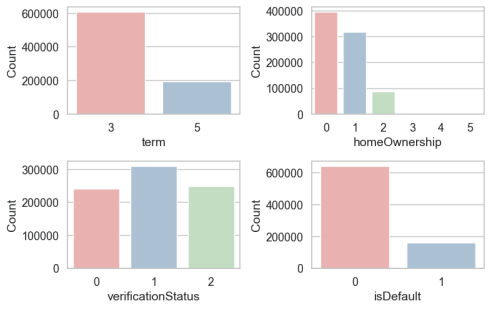

图3.2 离散型变量分布图

2.2.2 个人信贷违约状态统计

是否违约变量的条形图如图 3.3 所示，isDefault=0 的频数为 640390，占比为 80.05%，isDefault=1 的频数为 159610，占比为 19.95%。通过对 80 万条信贷数据中个人违约状态的统计可以看出，未违约的数据量是违约数据量的 4 倍多，类别比例相差悬殊，在后续部分对数据集进行了数据平衡处理。

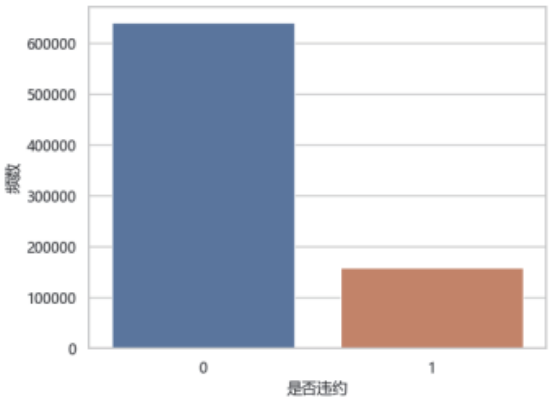

图3.3 信用违约状态分布情况

2.2.3 分类变量与信贷违约状态

数据集中包含许多分类变量，如贷款等级 grade（取值 A-G）、贷款等级之子集 subGrade（取值 A1-A5,B1-B5,…,G1-G5）等。这些变量与违约状态的统计数据如图 3.4、图 3.5、图 3.6、图 3.7 所示。

用户贷款等级的七个等级中，A 级表示用户的贷款等级最低，贷款频率不高；G 级表示用户的贷款等级最高，贷款频率高。

表3.1 字段信息表

字段名	说明	字段名	说明
id	为贷款清单分配的唯一信用证标识	delinquency_2years	借款人过去2年信用档案中逾期30天以上的违约事件数
loanAmnt	贷款金额	ficoRangeLow	借款人在贷款发放时的fico所属的下限范围
term	贷款年限（year）	ficoRangeHigh	借款人在贷款发放时的fico所属的上限范围
interestRate	贷款利率	openAcc	借款人信用档案中未结信用额度的数量
installment	分期付款金额	pubRec	贬损公共记录的数量

ISSN：3060-9437(Print)；3060-9445(Online)

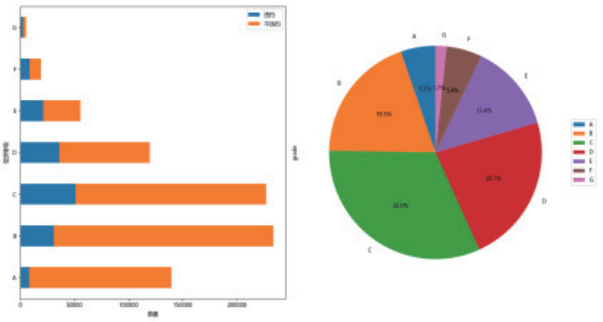

图3.4 用户贷款等级与违约状态以及违约比例的统计图

如图 3.4 所示，贷款等级为 B 的客户占比最高，贷款等级为 G 的用户最少。未违约客户中，贷款等级为 B 的用户占比也是最高的，其次是贷款等级为 C 的客户。而在违约客户中，信用等级为 C 的客户占比最高，第二第三位分别是等级为 D 和 B 的用户，这三个等级的违约用户占违约总额的比例达到 74.2%。

图 3.5 直观展示了贷款等级之子集与违约状态的统计数据。从贷款申请人数量的信贷等级分布情况来看，最多申请贷款的群体主要是贷款等级为 C 和 B 的客户，信贷主要客户是贷款等级处于中等区间的人，从违约客户贷款等级的分布情况来看，违约客户主要集中在贷款等级处于中等区间的群体，因此需要重点关注信贷等级处于中等区间的用户。

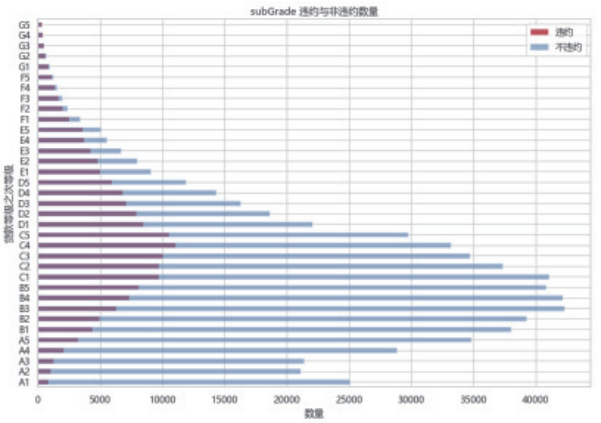

图3.5 用户贷款等级之子集与违约状态的统计图

从就业年限字段类别来看，就业年限大于 10 年的客户群体占比最高，这一类别群体的违约人数和未违约人数同样是占各自所在违约状态类型比例中最高的，需要重点关注这部分主体。其他就业年限的贷款人数分布比较均衡。

从贷款期限划分来看，贷款期限有 3 年期和 5 年期两类，如图 3.7 所示，3 年期贷款的客户占比最大，约为 5 年期贷款客户总数的 4 倍。在违约客户这一类别中，贷款期限为 3 年的人数占该类别总人数的比例达到 60.9%，5 年期贷款的人数占比为 39.1%。从各贷款期限的违约率来看，5 年期贷款的违约率要高于 3 年期。在违约率上，长期借款的违约率会更高。

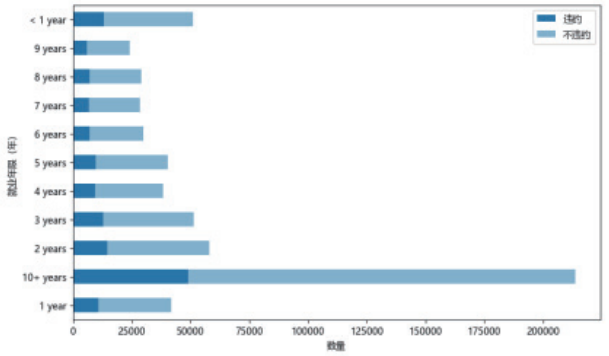

图3.6 就业年限与违约状态统计图

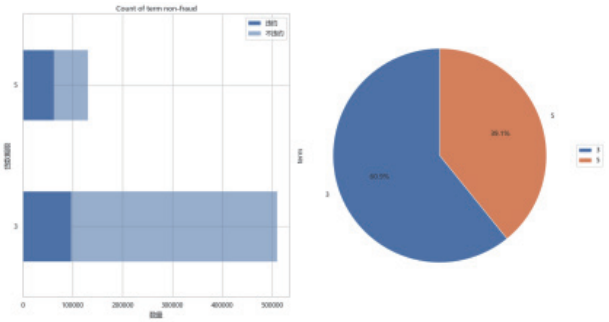

图3.7 贷款年限与违约状态以及违约比例的统计图

2.3 数据清洗

2.3.1 缺失值处理

图 3.8 展示了含有缺失值的特征及其缺失比例。本文使用到的原始数据样本量大，采用删除法对包含缺失值的行进行删除，其次是因为含有缺失值的特征，缺失比例很小，最大缺失比例小于 0.1，对缺失值进行删除对结果的影响不大。

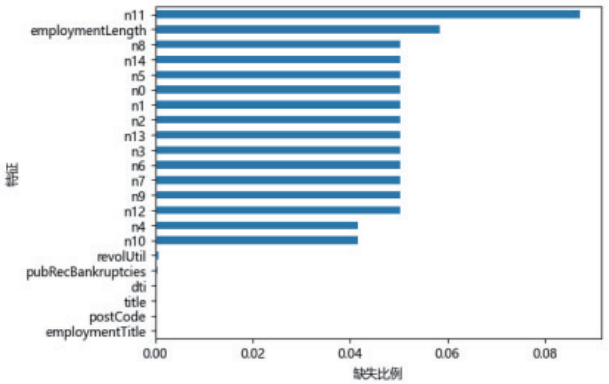

图3.8 含有缺失值的变量的缺失比例

2.3.2 异常值处理

本文通过绘制特征变量的箱线图，将超出上四分位数加上 1.5 倍四分位距的值和低于下四分位数减去 1.5 倍四分位距的值的数据点确定为异常值[9]，对异常值进行剪除。

图 3.9 展示了部分自变量的箱线图。

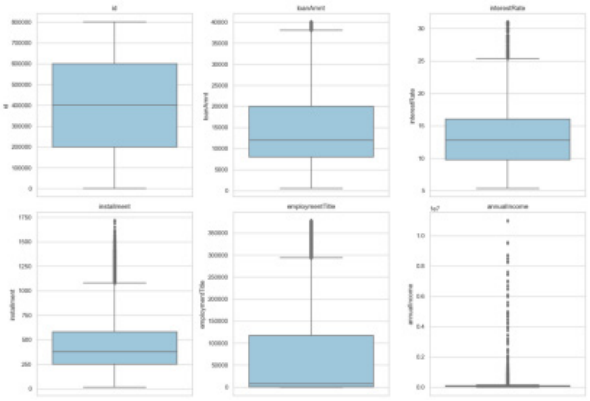

图3.9 自变量箱线图

2.3.3 单值变量的处理

由于单值变量通常不包含太多信息，本文对单值变量进行删除，即对单值特征变量进行删除。

2.3.4 类别特征转化

由于算法无法识别文字，因此需要对部分特征进行哑变量处理，转化为机器能够识别的类型。需要进行编码的特征变量有 employmentLength、grade 和 subGrade。对于就业年限，当就业年限小于 1 时，取值为 0，就业年限取值为 1-9 时，取值为 1-9，就业年限为 10+ 的取值为 10。贷款等级 A-G 分别取值 1-7，对应的子集 A1-A5 取值 11-15，B1-B5 取值 21-25，…，G1-G5 取值 71-75。

2.4 特征工程

机器学习可以高效地在特征工程处理阶段提取到有用的特征信息以节省模型的收敛时间。为了充分挖掘数据中包含的信息，需要对各特征变量进行分析，从而能够获得真正与问题本质最相关的信息[10]。特征工程就是在机器学习和数据挖掘任务中，对原始数据进行转换、创建、选择或提取特征的过程。

2.4.1 特征衍生

特征衍生是指在现有特征的基础上进行组合或加工生成新的特征的过程，其目的在于提升模型的预测能力和可解释性[11]。本文将 issuDate 和 earliesCreditLine 两个特征进行组合，衍生出新特征 CreditLine（开卡年限）；将特征变量 ficoRangeLow 和 ficoRangeHigh 进行加权平均，得到新特征 fico_average（平均信用评分）；将地区编码相同的各 interestRate（利率）取平均值，得到新的特征 region_interest_average（地区平均贷款利率）。保留新特征后，将使用到的原始特征进行删除，最终特征数量从 47 个减少为 44 个。

开卡年限这一新特征可以进一步研究开卡年限对信贷违约的影响。此外，通过构造借款人平均信用评分这一新特征来研究借款人信用评分对违约状态的影响。为了避免因地区利率差异过大对违约状态的判断产生大的偏误，构造地区平均贷款利率指标，减小利率差异对违约状态判断的影响。

2.4.2 特征选择

为了获取区分度更好的特征，需要对数据集进行降维处理，以得到最优子集，这个过程被称为特征选择。本文采用皮尔逊相关系数法和最小方差法对特征进行选择。皮尔逊相关系数是衡量特征与响应变量关系的方法，它反映的是两个变量之间的线性相关性，取值区间为[-1,1][12]。其中，1 表示完全正相关，0 表示完全没有线性关系，-1 表示完全的负相关。相关系数越接近 0，相关性越弱。最小方差法基于每个特征的方差来评估其重要性，基本思想是方差较小的特征对目标变量的影响较小，可以考虑将其从特征中剔除。相关系数法和最小方差法操作简单且具有很好的鲁棒性。

首先将特征中的单值特征变量删除，因为这类型的变量通常不包含太多有用的信息，且对模型会产生不利影响。

接着计算各特征变量与目标变量（isDefault）的相关系数，经过综合评价，本文将相关系数绝对值小于 0.01 的特征进行删除，将不符合要求的特征删除后，特征变量减少至 29 个。

随后，再通过计算各特征变量之间的相关系数，进一步筛选特征变量。将相关系数大于 0.6 的特征变量认定为高度相关的变量并剔除。原因在于，模型中引入高度相关的自变量可能会导致多重共线性，在自变量数量很多的情况下尤为严重，且使用高度相关的自变量会对模型的稳定性和解释性产生不利影响。经过计算，将特征变量 installment、subGrade、title、totalAcc、n2、n3、n5、n7、n9 和 n10 删除。特征变量减少至 19 个，其中包含一个被解释变量。

最后，采用最小方差法对特征进行筛选。将经过前 3 个步骤筛选后余下的特征变量计算其方差值。最终计算得到，特征 applicationType 具有最小方差，为 0.02，决定将该特征变量剔除。最终，经过特征选择后，保留下了包括响应变量在内的 18 个变量。

2.5 平衡数据

通过对数据的平衡性进行检验，isDefault = 0 的数据共有 172155 条，占比为 0.81，isDefault = 1 的数据有 39776 条，占比为 0.19，类别比例相差大，需要对数据集进行平

衡处理。

平衡数据集常用的方法有三种，分别是欠采样、过采样和综合采样[7]。本文采用 BorderlineSMOTE 采样法对数据集进行平衡。SMOTE 采样方法及其各类变体是被广泛应用的过采样方法，它通过将人造样本随机的插入少数类来平衡数据，使得少数类的样本空间扩张，达到平衡数据的目的[13]。传统的 SMOTE 采样方法容易产生过拟合的问题，而 BorderlineSMOTE 采样方法在合成样本的生成过程中通常会限制生成的数量，以避免对数据集的过度合成，降低过拟合的风险。此外，BorderlineSMOTE 采样方法能够更好地保留少数类别样本的分布特征，提高分类器性能。

经过使用 BorderlineSMOTE 采样方法对数据进行平衡，最终违约样本和未违约样本的数量均为 172155 条。

3 模型构建与评估比较

网格调参法是一种被广泛使用在机器学习中的参数优化方法，其基本思想是穷举搜索，通过自动化和并行技术搜索并确定最优的超参数[14]。此外，交叉验证的方法也常被用于最优超参数的选择。上述两种方法在本文中被用于最佳超参数的选择。在评估指标的选取上，参照现有研究的主流操作，将 AUC 值、F1 值和 KS 值等指标作为模型性能的评价标准。在训练集和测试集的划分上，按照七比三的比例进行划分。通过构建逻辑回归模型和 XGBoost 模型对信贷违约情况进行识别，并评判各模型预测能力。

3.1 逻辑回归

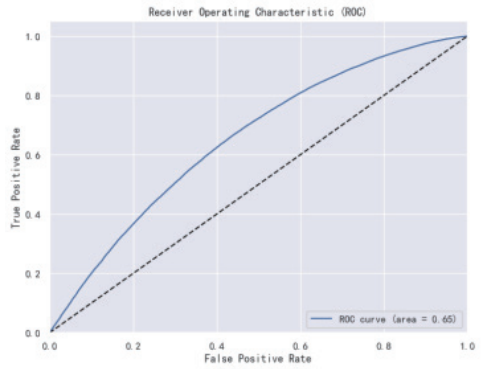

图4.1 逻辑回归ROC曲线图

使用 sklearn.linear_model 库中的 LogisticRegression 构建逻辑回归模型，将划分好的数据使用训练集对模型进行训练，随后将解释变量的测试集加入到训练好的模型中进行回归，输出预测值。同时输出混淆矩阵和评价指标。混淆矩阵及评价指标值如表 4.1、表 4.2 所示，对应的 ROC 曲线如图 4.1 所示。

由表 4.2 展示的结果，逻辑回归模型下的准确率为 0.61，说明能被准确预测为违约的客户和未违约客户的数量占总数的 61%；召回率为 0.62，说明实际违约客户中能被模型准确识别出来的客户占比约为 62%；查准率为 0.61，说明被模型预测为违约的客户中真正属于违约的客户占比约 61%；F1 值为 0.62，反映出模型在查准率和召回率上的综合性能达到 62%；KS 值为 0.23，说明逻辑回归模型在区分正例和反例方面的能力较弱；AUC 值为 0.65，模型的性能较好。

在参数的调整上，本文添加了参数 penalty 和 solver，并通过网格调参法和交叉验证对参数进行优化，并在训练集上拟合 GridSearchCV 对象，将得到的最优参数和最优模型使用测试集数据进行预测，并计算相关的评价指标。经过优化的模型，其准确率为 0.88，查准率为 0.96，F1 值为 0.87，AUC 值达到了 0.93。虽然优化后的模型其评价指标有了显著提高，但可能存在过拟合的问题，导致结果可信度不高，故未将调参后模型的结果作为最终结果。

3.2 XGBoost

XGBoost 模型的拟合主要通过 xgboost 库中的 XGBClassifier 分类器进行。首先，将模型中的参数 learning_rate 设为 0.3，待调整参数为 max_depth 和 min_child_weight，其他主要参数说明及最终取值如表 4.3 所示。通过网格搜索和 5 折交叉验证对待调参数进行选择。随后将选择出的最优参数建立模型，使用训练集数据对模型进行训练，再将测试集数据加入到训练好的模型中进行预测。通过输出混淆矩阵和计算准确率、召回率、AUC 值等指标，对模型的预测性能进行初步评价，并通过画 AUC

表4.1逻辑回归的混淆矩阵

预测未违约		预测类别	
		预测违约	
真实类别	真实未违约	30892	20730
	真实违约	19301	32370

表4.2逻辑回归模型评价指标

评估指标	准确率	召回率	查准率	F1	KS值	AUC
指标值	0.61	0.62	0.61	0.62	0.23	0.65

表4.3 XGBoost参数设置

参数名称	参数说明	取值
mind_child_weight	叶子节点的最小数量，避免过拟合	3
max_depth	决策树深度	6
gamma	在节点分裂时需要满足的最小损失下降值	0
n_estimators	决策树的数量	100
learning_rate	学习率，控制每次迭代对权重的更新幅度	0.3
subsample	每次迭代中对样本的抽样比例	0.8
colsample_bytree	每次迭代中对特征的采样比例	0.8
seed	控制随机数的种子	0
reg_alpha	L1正则化项的权重	0
reg_lambda	L2正则化项的权重	1

表4.4 XGBoost的混淆矩阵

		预测类别	
		预测未违约	预测违约
真实类别	真实未违约	50367	1255
	真实违约	11021	40650

表4.5 XGBoost模型评价指标

评估指标	准确率	召回率	查准率	F1	KS值	AUC
指标值	0.88	0.79	0.97	0.87	0.76	0.88

表4.6模型评价指标对比

	准确率	召回率	查准率	F1	KS值	AUC
逻辑回归	0.61	0.62	0.61	0.62	0.23	0.65
XGBoost	0.88	0.79	0.97	0.87	0.76	0.88

曲线图的方式直观展示结果。混淆矩阵及评价指标值如表 4.4、表 4.5 所示，对应的 ROC 曲线如图 4.2 所示。

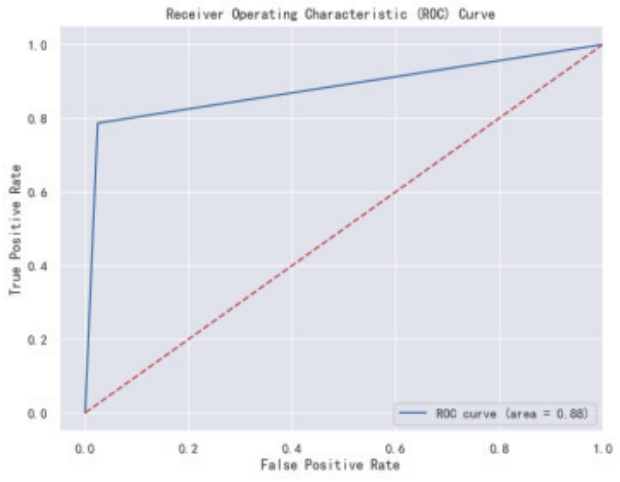

图4.2 XGBoost ROC曲线图

由表 4.5 展示的结果，XGBoost 模型的准确率为 0.88，说明能被准确预测为违约的客户和未违约客户的数量占总数的 88%；召回率为 0.79，说明实际违约客户中能被模型准确识别出来的客户占比约为 79%；查准率为 0.97，说明被模型预测为违约的客户中真正属于违约的客户占比约 97%；F1 值为 0.87，反映出模型在查准率和召回率上的综合性能达到 87%；KS 值达到 0.76，说明模型在区分正例和负例方面的能力强；AUC 值达到 0.88，说明模型的预测性能很好。

3.3 模型对比

如表 4.6 所示，通过与逻辑回归的结果进行比较可知，XGBoost 模型的准确率、查准率、F1 值和 AUC 值均大于逻辑回归模型的对应值，XGBoost 模型的预测效果要优于逻辑回归模型，可能原因在于：（1）XGBoost 采用了一些新的提升方法，可以更好地拟合数据；（2）XGBoost 可以通过特征重要程度对特征进行排序来选择更加重要的特征，从而减少噪声数据的影响；（3）XGBoost 采用梯度提升机制，利用多个树的集成来提高性能，而逻辑回归本质上是单一模型，缺乏集成的优势；（4）面对复杂数据集，在实践中，XGBoost 通常比逻辑回归更准确，因为它可以自动

寻找最优的分割点，并综合多个树的预测；（5）在处理非线性问题方面，XGBoost 比逻辑回归更具优势。

4 结语

前文对信贷违约预测系统的必要性和可行性进行了详细的论证，本部分将对本文的研究结果进行总结，以及对未来研究的展望。

4.1 结论

第一，本文使用的预测模型包括逻辑回归模型和 XGBoost 模型，通过实证研究，证实了机器学习算法在信贷违约预测问题的应用能够取得较好的结果。从模型预测效果来看，XGBoost 模型的预测效果要优于逻辑回归模型，在所有的模型评价指标中，XGBoost 模型的评价指标均优于逻辑回归，在经过参数调试后，XGBoost 模型的 KS 值达到 0.76，AUC 值能达到 0.88，逻辑回归的 KS 值只有 0.23，AUC 值为 0.65，两者的预测效果存在较大的差距。

第二，本文同时也证明了将特征选择和机器学习相结合的做法，有助于对信贷违约风险的预测。在大数据时代，面对巨量的数据，传统的计量经济方法在对信贷风险进行预测时，会面临过拟合、样本量过大导致模型复杂程度加大、预测效果差等问题，而机器学习方法在解决大样本数据方面表现出优势性，在解决过拟合和预测精度方面，具有很大的优势。

4.2 不足与展望

本文在研究过程中仍存在着不足之处。主要不足如下：

第一，本文使用到的模型较少，只使用到了逻辑回归和 XGBoost 模型，在信贷违约预测方面，还有很多性能优越的模型可以使用，如 CatBoost、LightGBM 模型等。在后续的研究中，可以引入更多的模型进行比较分析。

第二，本文使用到的特征选择方法较少，在特征选择方法上，过滤法、嵌入法和包装法在特征选择方面都能取得好的结果，在面对相关系数法和最小方差法这些简单高效的方法时，也能取得较好的预期效果，在后续研究中可以比较不同特征选择方法选取的特征在预测精度上的差异，拓展现有研究。

第三，在数据的使用上，本文使用到的原始数据含有 80 万条，在对数据进行清洗过后，仅保留了 211931 条，大部分数据被剔除，可能会对结果产生不利影响，而且其中可能包含虚假数据，且未被识别出来。真实可靠的数据是使得研究结果可信赖的前提之一，在后续研究中应对数据的可靠性更加重视，对数据的处理采用更加科学合理的方法，使得数据中包含的真实有用信息发挥最大的作用。

参考文献：

[1] 屈忠锋，吴鸿华，李凡军 . 基于 Logistic 回归的中小企业信贷风险评估与信贷策略优化建模[J/OL]. 山东大学学报 (理学版)，1-9[2024-05-10].

[2] 蔡文学，罗永豪，张冠湘等 . 基于 GBDT 与 Logistic 回归融合的个人信贷风险评估模型及实证分析[J]. 管理现代化，2017,37(02):1-4.

[3] 张国政，陈维煌，刘呈辉 . 基于 Logistic 模型的商业银行个人消费信贷风险评估研究[J]. 金融理论与实践，2015,(03):53-57.

[4] 伍洁，陈迪芳，李瑞彤等 . 基于 XGBoost 和 SHAP 方法的个人信贷风险评估研究[J]. 现代信息科技，2024,8(08):146-150+155.

[5] 朱丽华，龙海侠 . 群智能算法优化 XGBoost 的信贷风险预测[J]. 计算机工程与应用，2023,59(23):305-310.

[6] 廖文雄，曾碧，梁天恺等 . 面向高维数据的个人信贷风险评估方法[J]. 计算机工程与应用，2020,56(04):219-224.

[7] 徐敏洁 . 基于集成学习的网络信贷违约预测[D]. 西南大学，2023.

[8] Chen,Tianqi,and Carlos Guestrin.”Xgboost:A scalable tree boosting system.”Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.2016.

[9] 马强，高雅，王红等 . 基于电价形成机制与 XGBoost 的单轨制电力现货市场的电价预测[J/OL]. 系统科学与数学，1-25[2024-05-10].

[10] 徐智超 . 基于特征工程和改进 SMOTE 算法的信贷风险预测研究[D]. 南京邮电大学，2023.

[11] 刘晓川 . 基于 Blending 融合算法的贷后信用风险预测[D]. 西南大学，2023.

[12] 罗文，厚峰 . 大语言模型评测综述[J]. 中文信息学

报，2024,38 (01):1-23.

[13] 苏逸，李晓军，姚俊萍等 . 不平衡数据分类数据层面方法：现状及研究进展[J]. 计算机应用研究，2023,40 (01):11-19.

[14] 张玥 . 基于算法融合的电商客户流失预警研究[D]. 东北财经大学，2022.

作者简介：黄飞豪（1998.02-），男，汉族，广西扶绥，硕士在读，研究方向：信贷风险评估。