基于逻辑回归算法的客户信用违约分析

王燚坤

中国地质大学（北京）经济管理学院，北京，100083；

摘要：在金融领域，信贷风险评估是一个至关重要的过程，它帮助金融机构确定借款人违约的可能性，并据此做出贷款决策随着数据科学在各行各业的广泛应用，机器学习技术，特别是逻辑回归模型，在信贷评分卡的构建中展现出巨大的潜力本论文详细探讨了如何利用逻辑回归模型构建一个信贷评分卡系统，该系统旨在为金融机构提供一个准确、高效的信贷风险评估工具本研究首先对所使用的数据集进行了细致的预处理，包括处理缺失值和异常值，确保了数据的质量，并采用上采样的方法解决了数据不平衡的问题在模型构建阶段，本文采用逻辑回归算法，利用借款人的个人信息和历史信贷记录来预测其违约的可能性通过在测试集上的评估，分析模型的召回率、精确率等其他标准对模型进行了系统的分析，模型的准确率达到 72.86%，模型展示了良好的性能在系统构建方面，本论文介绍了一个基于 Web 的信贷评分卡系统，该系统采用 Flask 框架作为后端，结合 HTML、CSS 和 JavaScript 构建的前端界面，提供了一个用户友好、响应迅速的平台用户可以通过简单的界面输入借款人的信息，系统将实时返回信贷风险评分此外，系统还提供了结果的图形展示，增强了结果的直观性和可解释性本研究证明了逻辑回归模型在信贷风险评估中的有效性，并展示了如何将这种模型集成到实际的应用系统中在未来的工作中可以对系统进一步的优化更好地服务于金融行业，为借款人提供更加科学和透明的贷款服务。

关键词：逻辑回归；信贷评分卡；信贷风险评估；系统构建

DOI：10.69979/3029-2700.25.06.070

1 引言

1.1 研究背景

信贷市场是金融市场的重要组成部分，它涉及到资金的借贷双方，即债权人和债务人信贷市场的健康运行对于促进资金的有效配置、支持实体经济发展以及维护金融稳定具有至关重要的作用[5]对于整个金融系统而言，良好的信贷风险管理有助于维护金融市场的稳定，防止金融风险的传染和放大，保护投资者和消费者的利益[6]

信贷评分卡模型作为一种有效的风险评估工具，它通过分析借款人的各种财务和非财务信息，预测其违约的可能性逻辑回归算法，作为一种广泛应用的统计方法，因其模型的简洁性、解释性强以及适用性广等特点，在信贷评分卡模型的构建中得到了广泛的应用

随着互联网金融的兴起和大数据技术的快速发展，中国的金融机构和科技公司也开始重视并研发信贷评分卡模型[2]中国的几家大型银行，如工商银行、建设银行等，已经开始使用信贷评分卡模型来优化信贷审批流程和提高风险管理水平中国的互联网公司如蚂蚁金服、京东金融等也在信贷评分领域取得了显著成果，它们利用自身的技术优势和庞大的用户数据，开发出了更为精准的信用评分系统，为用户提供更便捷的金融服务

1.2 研究内容

本论文的研究内容涉及信贷数据的预处理、分析以及逻辑回归模型的应用，具体研究内容如下：

详细介绍数据清洗和预处理的步骤，包括处理缺失值、异常值，以及数据的标准化重点讨论不同数据预处理技术对模型性能的影响

应用统计学方法和可视化工具对预处理后的数据进行深入分析探索数据集中的各种特征分布，及其与目标变量之间的关系

构建逻辑回归模型，分析其参数，并详述模型选择的理论基础[1]讨论模型优化的策略，包括特征选择和超参数调整

使用混淆矩阵、准确率、召回率、精确率和 F1 分数等指标来评估模型性能深入分析模型在预测信贷逾期方面的能力，并对结果进行解释

介绍一个基于逻辑回归模型的信贷评估系统的设计和实现包括系统架构、功能模块、用户界面和技术实现细节

2 数据分析

2.1 数据来源与变量解释

本研究使用的数据集来自 Kaggle 竞赛平台，目标是根据历史数据预测借款人在未来两年内是否有可能经历财务困难[11]

2.2 数据预处理

利用 Python 库扫描，发现月收入和家庭成员数量是存在缺失值，并采用了中位数填充法处理计算每个变量的第一四分位数 Q1 和第三四分位数 Q3，然后确定四分位数范围 IQR = Q3 - Q1 将低于 Q1 - 1.5 * IQR 或高于 Q3 + 1.5 * IQR 的数值作为异常值移除

2.3 数据描述

年龄数据反映出一个相对年轻的借款人群体；大多数借款人保持着较低的信用卡余额，但也有一些借款人的使用率超过了 100%；月收入在中低收入区间，收入分布的尾部延伸得较长，表明尽管有一些较高收入的个体，但大多数人的收入较为集中；信贷产品数量是适中的，既不太多也不太少，这表明大多数借款人可能在财务管理上比较理智；大部分借款人可能拥有一两笔关于住房的贷款，是借款人最大的负债之一

3 信贷评分卡模型

3.1 模型的构建与训练

本文首先进行了特征选择，即将目标变量（SeriousDlqin2yrs）从特征集中分离出来代码中使用 drop 方法从数据集中移除了目标变量，以便本文可以将其作为模型的标签（y）

数据集使用 sklearn 的 train_test_split 方法进行随机划分，被分为 80%的训练集和 20%的测试集，具体代码如下

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

在逻辑回归中，特征的规模可以极大影响模型的性能为了避免变量的不同量纲对模型产生影响，本文采用了标准化（Standard Scaler）来预处理数据，标准化处理后，各特征的均值为 0，标准差为 1，利用训练集数据对模型进行了训练，代码如下

```
lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
```

在本研究中，模型的目标是预测借款人在未来两年内是否会出现逾期行为（SeriousDlqin2yrs）模型的输出是一个介于 0 和 1 之间的概率值，表示借款人逾期的可能性通过设定一个阈值（通常为 0.5），本文可以将概率值转换为二元类别，从而完成分类[12]

3.2 模型结果评估

表 3- 评估指标

类别	精确度	召回率	F1 分数
0	0.68	0.88	0.77
1	0.83	0.57	0.68
总体	0.73	0.73	0.72
宏平均	0.75	0.73	0.72
加权平均	0.75	0.73	0.72

总体准确率为 72.86%，有近四分之三的预测是正确的非逾期客户（标签 0）的精确度为 68%，而逾期客户（标签 1）的精确度则高达 83%非逾期客户的召回率为 88%，远高于逾期客户的 57%非逾期客户的 F1 分数为 77%，而对逾期客户则是 68%表明当预测客户会逾期时效果更好 F1 分数的差异进一步揭示了模型在识别两类客户时的不平衡性可以看到模型在识别非逾期客户（TN）方面表现较好，但在识别实际逾期客户（TP）方面还有提升空间，尤其是模型存在一定数量的假负例（FN）

3.3 信贷评分卡系统的构建

前端负责与用户的交互，利用 HTML、CSS 和 JavaScript 进行网页构建并设计了指引和校验机制，确保用户输入的数据格式正确后端的核心任务是处理前端发来的请求，执行数据处理，并将结果返回给前端首先进行数据的预处理，随后调用事先训练好的逻辑回归模型，对结果进行预测

4 结论

本论文详细探讨了利用逻辑回归模型构建信贷评分卡系统的整个过程，包括数据的预处理、模型的构建和评估，以及信贷评分卡系统的实现通过对每一步骤的细致讨论，本文不仅提供了一种有效的信贷风险评估方法，还展示了如何将这种方法应用于实际的系统开发中

在数据预处理阶段，本文通过处理缺失值和异常值，确保了数据的质量和一致性，为模型的训练奠定了坚实的基础在模型构建和评估阶段，通过逻辑回归模型，本文成功地将借款人的个人信息和历史信贷记录转化为其信贷风险的量化评估模型在测试集上的表现证明了其在信贷风险评估方面的有效性，尽管仍有提升空间，

特别是在提高对逾期客户的识别准确性方面

在信贷评分卡系统的构建中，本文采用了现代 Web 技术，创建了一个应用程序，使非技术用户也能轻松地进行信贷风险评估系统的设计确保了高效的数据处理和直观的结果展示，为用户提供了实时的、基于数据驱动的决策支持

通过不断的探索和改进，信贷评分卡系统将能够更准确、更有效地帮助金融机构评估和管理信贷风险，同时为客户提供更透明、更公正的借贷服务

参考文献

[1]张利斌,吴宗文.基于XGBoost机器学习模型的信用评分卡与基于逻辑回归模型的对比[J].中南民族大学学报(自然科学版),2023,42(06):

[2]谢瑞.基于逻辑回归的个人信用评分卡设计与分析[D].北京交通大学,2022.

[3]李先航.基于机器学习的可解释信贷风险评分和违约预测研究[D].西南财经大学,2022.

[4]唐春玲.基于机器学习的信用评分卡研究与设计[D].湖南大学,2022.

[5]卢悦冉,芮英健,袁芳.基于评分卡模型下中小微企业的信贷决策[J].中国市场,2021,(27):53-54.

[6]林炜.信用评分方法在信贷风险管理中的应用[J].信息系统工程,2021,(08):143-145.

[7]王方春.信用评分模型在信贷审批决策中的应用探讨[J].甘肃金融,2021,(03):35-38.

[8]朱岚.基于机器学习的信用评分的研究与应用[D].上海财经大学,2020.

[9]陈秋华,杨慧荣,崔恒建.变量筛选后的个人信贷评分模型与统计学习[J].数理统计与管理,2020,39(02):368-380.

[10]徐英浩.信用评分系统的设计与实现[D].浙江工业大学,2020.

[11]张俊丽,郭双颜,任翠萍,等.基于逻辑回归的个人信用评分卡模型研究[J].现代信息科技,2024,8(05):12-16.DOI:10.19850/j.cnki.2096-4706.2024.05.003.

[12]杨蓉.基于商业银行信贷视角的财务报表分析[J].福建金融,2024, (02):75-79.

[13]郑亚明.金融经济风险及防范措施[J].投资与合作, 2024,(02):28-30.

[14]张馨月.基于多任学习的冷启动下行为评分模型研究[D].西南财经大学,2023.DOI:10.27412/d.cnki.gxncu.2023.001827.

[15]李佼洋.基于消费信贷的个人信用风险评估模型研究[D].对外经济贸易大学,2022.DOI:10.27015/d.cnki.gdwju.2022.001047.

[16]Jammalamadaka K R, Itapu S. Responsible AI in automated credit scoring systems[J]. AI and Ethics, 2023, 3(2): 485-495.

[17]Liu W, Fan H, Xia M. Tree-based heterogeneous cascade ensemble model for credit scoring [J]. International Journal of Forecasting, 2023, 39(4): 1593-1614

[18]Gero S . An Overview on the Landscape of R Packages for Open Source Scorecard Modelling [J]. Risks, 2022, 10 (3): 67-67.

[19] Jalil E ,Abderrahim Q E ,Mehdi B,et al. Modeling with ontologies design patterns: credit scorecard
as a case study[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2020, 17 (1): 429-429.

[20]Paula DVAD ,Artes R ,Ayres F , et al. Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques[J].RAUSP Management Journal, 2019, 54(3):321-336.

作者简介：王燚坤（2005 年 7 月 25 日—）性别男 民族汉 籍贯山东省临沂市平邑县 职称无 学历本科在读 单位中国地质大学（北京）经济管理学院 研究方向会计学