学校代号 10532　　学　号 S1918W1337

分 类 号　　密　级

湖南大學
HUNAN UNIVERSITY

专业硕士学位论文

基于多模型融合的 P2P 网贷借款人信用风险评估研究

学位申请人姓名 桑如柳

培 养 单 位 金融与统计学院

导师姓名及职称 谭朵朵（副教授）

吕忠伟（高级经济师）

学 科 专 业 应用统计硕士

研 究 方 向 宏观经济统计

论 文 提 交 日 期 2022 年 5 月 12 日

学校代号：10532

学　　号：S1918W1337

密　　级：

湖南大学专业硕士学位论文

基于多模型融合的 P2P 网贷借款人信用风险评估研究

学位申请人姓名：桑如柳

导师姓名及职称：谭朵朵（副教授）

吕忠伟（高级经济师）

培　养　单　位：金融与统计学院

专　业　名　称：应用统计硕士

论文提交日期：2022 年 5 月 12 日

论文答辩日期：2022 年 5 月 20 日

答辩委员会主席：周四军 教授

Research on Credit Risk Assessment of P2P Online Loan Borrowers based on Multi Model Fusion

by

SANG Ruliu

B.E. (Anhui Agricultural University) 2019

A thesis submitted in partial satisfaction of the requirements for the degree of

Master of Economics

in

Applied Statistics

in the

Graduate school

of

Hunan University

Supervisor

Associate Professor TAN Duoduo

Senior Economist LV Zhongwei

May, 2022

摘　要

P2P 网贷作为一种新型的金融模式，将互联网和传统信贷结合，其凭借有效的融资效率、便捷的交易流程和广泛的覆盖能力迅速发展起来。但是由于缺乏完善的信用监管体系以及信息不对称的问题，近年来 P2P 平台暴雷现象频发，严重损害了投资者的利益，引发网贷行业信任危机。其中信用风险问题是导致平台问题频发的重要原因，因此建立科学有效的信用风险评估体系是促进 P2P 网贷平台持续健康发展的重要举措之一。

针对上述问题，本文在进行信用评估体系的指标筛选时引入随机森林算法。通过随机森林算法计算每个特征的重要性，然后根据特征重要性进行降序排序，将筛选前 64 个特征减少到 46 个特征，极大优化了原始数据集的特征集合，完善了信用评估体系的指标构成。

同时，为了提高信用风险评估体系性能，本文提出了基于多模型融合的 P2P 借款人信用风险评估体系，它是根据投票法思想将 Logistic 回归分类算法、随机森林分类算法和 LightGBM 分类算法按少数服从多数的原则进行融合，修正单模型预测出现的偏差问题，并选择精准率、召回率、f1 分数以及 AUC 作为四个评估指标在 Lending Club 数据集上与三个单模型进行了对比。本文提出的多模型融合算法在四个评估指标上的预测性能都是最佳的，其中精准率为 0.9961、召回率为 0.9765、f1 分数为 0.9862 以及 AUC 值达到 0.9864，充分验证了基于投票法的多模型融合算法的真实可行性和稳健性,可以更加科学的评估借款人的信用风险，降低借贷人和网贷平台在放贷过程中承担的风险，促进网贷平台的良性发展。

关键词：P2P 网贷；信用风险评估；Logistic 回归模型；随机森林分类模型；LightGBM 分类模型

Abstract

As a new financial model, P2P online loan combines Internet with traditional credit. It develops rapidly with effective financing efficiency, convenient transaction process and wide coverage. However, due to the lack of a perfect credit supervision system and the problem of information asymmetry, in recent years, P2P platform thunderbolt phenomenon occurs frequently, which seriously damages the interests of investors and leads to the trust crisis in the online lending industry. Credit risk is an important reason for the frequent occurrence of platform problems, so the establishment of a scientific and effective credit risk assessment system is one of the important measures to promote the sustainable and healthy development of P2P online lending platform.

In view of the above problems, this paper introduces random forest algorithm in the index selection of credit evaluation system. The importance of each feature is calculated by random forest algorithm, and then sorted in descending order according to the importance of each feature. The first 64 features are reduced to 46 features, which greatly optimizes the feature set of the original data set and improves the index composition of the credit evaluation system.

At the same time, in order to improve the performance of the credit risk assessment system, this paper proposes a P2P borrower credit risk assessment system based on multi model fusion. According to the idea of voting method, it fuses logistic regression classification algorithm, random forest classification algorithm and lightgbm classification algorithm according to the principle that the minority is subordinate to the majority, corrects the deviation problem of single model prediction, and selects the accuracy rate, random forest classification algorithm and lightgbm classification algorithm Recall rate, F1 score and AUC were compared with three single models on lending Club dataset. The multi model fusion algorithm proposed in this paper has the best prediction performance in four evaluation indexes, including accuracy rate of 0.9961, recall rate of 0.9765, F1 score of 0.9862 and AUC value of 0.9864, which fully verifies the feasibility and robustness of the multi model fusion algorithm based on voting method, which can more scientifically evaluate the credit risk of borrowers and reduce the credit risk of borrowers and Internet users The risk of online lending platform in the process of lending can promote the healthy development

of online lending platform.

Key words: P2P Network Loan; Credit Risk Assessment; Logistic Regression Model; Random Forest Classification model; LightGBM Classification Model

目 录

插图索引

附表索引

第 1 章 绪论

1.1 研究背景及意义

1.1.1 研究背景

随着科技的进步与发展，互联网金融应运而生。互联网金融作为一种创新型金融业务模式，将传统金融机构和互联网企业结合并利用科学技术去实现资金融通、支付、投资等，极大的促进资金配置效率，便捷了人们的生活。P2P 网贷是网络贷款的简称，是指个体和个体之间通过互联网平台实现的直接借贷。P2P 网贷作为互联网金融的重要组成部分，为个人以及中小企业解决融资难的问题，提高了资金的配置效率。但是由于缺乏完善的信用体系以及规制，P2P 网贷平台暴雷事件频发，不少平台跑路导致部分投资者血本无归。

2007 年拍拍贷的成立标志着中国正式引入 P2P 网络借贷平台。2011 年至 2017 年 P2P 行业快速崛起发展，行业书写了暴富神话，因此大量资本涌入，最高峰曾达到 5000 家 P2P 运营平台。但近年来，大规模跑路和暴雷事件的爆发对该行业造成巨大打击，甚至严重威胁到了金融市场的稳定。截至 2019 年 11 月 30 日，我国 P2P 网贷平台数量累计达 6698 家，其中问题平台 6232 家。针对乱象政府开始出手干预，发表了相关意见，以引导机构退出为主导方向，“能退则退，能关尽关”，故全国各地纷纷展开清退 P2P 网贷业务。P2P 网贷从崛起到蓬勃发展再到问题频发，究其原因一方面是没有很好的监管体系，另一方面是平台风险管理不善，借款人不能还款。其中借款人信用风险过高是平台倒闭的重要原因。P2P 网贷交易全程都是通过网络进行，投资者对借款人信息的真实性无法考量，造成双方处于信息不对称地位，而且部分借款人是传统金融机构不愿提供贷款的次级客户，更容易发生信用违约。所以建立科学的借款人风险评估模型对 P2P 网贷平台持续健康发展十分必要。

1.1.2 研究意义

科学地构建 P2P 网贷平台借款人信用风险评价模型对减少 P2P 网贷平台跑路现象和保障投资者资金安全有重要的作用。传统的信用风险评估往往凭主观经验选取特征建模，本文建模时引入随机森林算法进行特征筛选。模型选择中采用了多模型融合思想，将统计学模型和机器学习算法模型融合，能够显著提高模型的性能，相较于单模型有非常明显的优势。随机森林特征筛选结合多模型融合思想构建的信用风险评估模型，对 P2P 网贷信用风险评估建模方法有参考意义。

P2P 网贷作为互联网金融业重要的组成部分，如果 P2P 网贷平台借款人信用风险的问题不能得到有效的解决，不仅仅会损害投资人的利益，也会威胁整个网

贷行业的发展，不利于互联网金融秩序的稳定。本文通过比对分析单个模型性能的优劣，将单模型融合构建评估效果更好的信用风险评估模型。多模型融合的借款人信用风险评估模型能够更加准确的预测借款人违约的概率，可以帮助平台和借款人筛选违约风险较小的借款人，达到事前控制风险的目的。这也有利于 P2P 网贷行业持续健康的发展，进而促进互联网金融市场秩序的稳定。

1.2 文献综述

P2P 网络借贷作为一种新兴的金融模式将传统民间借贷和互联网结合，一定程度上促进了我国经济发展。但由于信用体系不完善以及一系列风险，P2P 网贷平台频频暴雷,其中借款人信用风险是导致平台倒闭以及借款人跑路的重要原因。所以，国内外学者对 P2P 网贷借款人信用风险评估指标体系、借款人信用风险评估模型等进行了深入的研究。

1.2.1 关于信用风险指标体系研究

从国外研究情况来看，Ravina E（2007）对获贷可能性和影响因素研究中，控制了信用评分、信用记录、收入、就业状况和住房拥有情况之后，发现如美貌、种族会显著影响获得贷款或支付利率的可能性[1]。Jefferson D（2012）指出除了借款人的个人借贷信息，借款人的外貌对于能否获得贷款以及借款人违约的概率也有影响。通过实证分析发现,外表看起来更值得信任的人有较高的概率获得贷款，而且确实他们的违约风险也更低[2]。Carlos（2015）等利用贷款俱乐部 2008-2014 年数据，通过均值检验来分析违约的影响因素，然后运用 logistic 模型回归发现年收入、房车资产状况以及信用历史对信用风险有显著的影响，而且模型的准确性也可以通过加入借款人的债务水平指标得到显著提升[1]。Jin Y（2015）利用贷款俱乐部的数据，研究贷款及其申请人的特征，并在建模阶段利用随机森林进行特征选择，得出贷款期限、年收入、贷款金额、债务收入比、信用等级和周转额度利用率对贷款违约有重要影响[3]。Emekter（2015）对 Lending club 平台数据进行实证研究发现，借款人是否违约主要是受到负债收入比以及借款人的 FICO 得分的影响[4]。Lu-Ming Y（2017）等通过因子分析对 75 家基于互联网的 P2P 借贷平台的数据进行处理，得出网站历史和规模指标对 P2P 借贷平台的贡献最大的结论[5]。Polena M（2018）利用借贷俱乐部的贷款数据研究影响借款人违约的决定性因素,通过定义四个贷款风险类别并测试违约决定变量的显著性,得出债务收入比、过去半年的调查与借款人违约率呈正相关，年收入与以信用卡为贷款用途呈负相关[6]。Canfield C E（2018）利用墨西哥最大的 P2P 在线贷款平台 Prestadero 的数据研究了信用评分以及其他相关指标对借款人信用风险的影响，运用 logistic 回

归模型检验控制贷款质量后，得出性别不会影响投资人的投资决策，贷款的质量与违约行为呈正相关，支付收入比对增加违约几率有较强的影响[7]。

国内文献主要是从借款人个人特征信息、信用状态信用、历史借贷信息这些方面来评估借款人是否违约。严复雷等（2016）利用网贷之家 87 个平台的数据，通过多元有序 Logit 模型回归，分析影响 P2P 网贷平台借款人信用风险的因素，得出人均借款额、平均利率等对是否违约有重要的影响，但是安全系数、平均借款周期以及运营时间等解释作用不强[8]。刘鹏翔（2017）利用拍拍贷平台的数据对借款人信用风险因素研究，采用多元线性模型回归发现年龄与流标次数与信用风险水平呈显著的负相关，信用等级对信用风险影响不显著，性别对信用风险影响显著相关[9]。隋昕（2017）利用拍拍贷和人人贷的数据，运用 excel 宏技术抓取借款人个人信息、借款人历史交易信息以及平台评价信息，用 logit 模型实证分析得出借款人的借款记录以及职业对借款人是否违约呈正向影响[10]。董文奎（2017）利用“新新贷”等 P2P 网贷平台数据实证分析发现借贷者的基本信息、资产情况以及提供的材料数量对借款人信用风险呈正向影响，并以此提出相应的对策建议[11]。雷舰（2019）利用人人贷数据，对客户的基本信息、借款历史信息、信用评级信息以及借款特征和还款能力等进行因子分析，然后用 Logistic 模型回归得出借款人资产情况、婚姻状态以及学历和信用评级对借款人信用风险呈负相关，借款利率以及期限与借款人信用风险呈正相关，并以此结论提出了针对性建议[12]。舒方媛等（2019）利用人人贷平台的数据研究借款人违约的影响因素，通过构建借款人信用评价指标体系，运用 Logistic 二元回归模型分析发现借款人的信用评级、年龄以及逾期次数对借款人是否违约影响显著，借款利率、房产状况以及学历对借款人信用风险也有影响[13]。李昕玮（2020）利用四个 P2P 网贷平台的数据基于借款人的信息特征研究信用风险的影响，通过多元回归模型得出年利率、资产情况以及借款金额与借款人信用风险成反比，与学历成正比[14]。

1.2.2 关于信用风险评估模型研究

信用风险评估作为 P2P 借贷领域中十分重要的环节，直接影响了平台的安全性。对信用风险的评估也随着技术的进步以及数理统计方法的发展由初期的定性分析转化为定量分析，且定量分析也由传统的统计模型发展到机器学习方法，在此基础上有学者通过对比发现组合模型比单个模型有更好的风险预测效果，取得了不少的成果。

从国外研究情况来看，在传统模型方面，Noh H J（2005）提出基于生存的方法预测个人违约概率的能力，可以解决二元分类方法的信用风险模型遗漏重要的时变因素和缺失信息的问题。通过建立生存信用风险模型评估不同变量在违约预测中的相对重要性[15]。Bekhet（2014）提出利用 Logistic 回归模型和径向基函数

模型来支持约旦商业银行贷款决策的信用评分模型，发现 Logistic 回归模型的总体准确率略高于径向基函数模型。然而，径向基函数在识别那些可能违约的客户方面更具优势[16]。Andreas Mild（2015）分析资本市场价格发现其缺乏可靠、可量化的数据，所以很难反映相关的违约风险，于是提出一种线性回归模型以评估借款人的违约风险，并发现基于此模型的决策结果显著获得更高的回报[17]。Sylvester（2017）运用 Logistic 回归和主成分分析处理变量协同效应问题，克服了 Logistic 模型的局限性[18]。

在机器学习方面，Zhang Z（2014）利用银行信贷部门的数据，提出多准则优化分类器模型来评估申请人信用风险。该模型基于核模糊化和惩罚因子，首先用核函数将输入点映射到高维特征空间，然后在 MCOC 中引入适当的模糊隶属度函数并进行分类，可以从不同类别的训练数据中导出一个决策函数，比起支持向量机和模糊支持向量机可以提高分离信用风险的效率得分，泛化能力更优秀[19]。Malekipirbazari（2015）利用 Lending Club 截止到 2015 年的数据来确定与不同平台相关的风险因素。通过比较 SVM、KNN 和随机森林方法，发现随机森林更能够识别优秀借款人，其分类性能更好[20]。Maldonado S（2017）利用治理银行的数据研究其信用评分相关问题，提出一种基于线性支持向量机的分类器构造和变量选择方法，将与业务相关的信息纳入建模过程，最终在业务目标方面取得了优异的效果[21]。Setiawan（2019）提出了一种基于支持向量机的二元粒子群优化算法（BPSOSVM）来对数据集进行特征选择并以极端随机树和随机森林作为分类器来预测一笔借款是否会成为坏账。结果表明，BPSOSVM 可以在不降低性能的前提下生成特征子集，且 ERT 在多个性能指标上都优于 RF[22]。Cai（2020）对 Lending Club 平台 2018 年的贷款数据处理， 对不平衡数据分类得到 4 个评价指标，用 CfsSubsetEval 评价策略和 BestFirst 搜索策略对特征进行搜索，更准确的对贷款人信用进行评估，降低了平台的风险[23]。

在组合模型方面来看，Tran K（2016）利用德国和澳大利亚客户信用数据信息，提出了将深度学习网络的强大功能与综合遗传规划相结合的混合思想，建立了稳健的信用模型，该模型提供了最佳的准确性和可靠的 IF-THEN 规则[24]。Zeng X（2017）利用 Prosper 的真实借贷数据提出 Logistic 分类模型和迭代计算模型组合的模型，发现该模型预测效果优于单模型[25]。A X M（2018）利用 Lending Club 的真实 P2P 数据对贷款违约风险进行预测，提出 LightGBM 和 XGboost 融合的模型，采取“多观察”和“多维”的数据清理方法，发现融合模型的分类预测结果最好[26]。Tong Z（2019） 基于 SPARK 技术引入决策树算法，提出 C4.5 决策树优化的融合模型，发现该模型预测效果更好[27]。

从国内研究情况来看，在传统模型方面，徐慧婷等（2018）利用美国借贷平台的数据研究借款人信用风险评估效果，通过建立 Logistic 回归模型,发现借款利

率、房产资产情况对借款人是否违约有显著的影响[28]。李淑锦等（2018）利用微贷网平台的数据评估 P2P 网贷平台信用风险，采用 Logistic 回归模型发现预测的准确度比较高且模型有较强的适用性[29]。马瑞（2019）利用广东省 925 家网贷平台数据研究网贷平台违约风险的影响因素，发现平台评价、注册资本、平均收益率对平台违约概率有显著的影响[30]。王浩明（2019）利用 Lending Club 平台的数据评估借款人信用风险，采取 Logit 模型实证分析风险、FICO 分数、旋转线利用率以及信用等级对是否违约有显著影响[31]。陈雪莲等（2019）利用人人贷平台的数据，运用 Logistic 模型逐步回归建立借款人信用风险评估模型，发现该模型预测效果比较好[32]。井浩杰等（2019）利用 Lending Club 平台的数据，对特征变量采取主成分分析然后赋权重，建立 Logistic 回归模型评估借款人信用风险[33]。

在机器学习方面，柳向东等（2016）利用人人贷平台的数据，使用 SMOTE 算法处理不平衡数据来提高模型评估性能，发现随机森林更适合用于信用风险评估，其次是 CART、ANN、C4.5[34]。操玮等（2018）利用人人贷平台的数据，采取随机森林算法筛选特征构建信用风险评价模型。通过和其他 4 种集成算法对比发现 Rotation Forest 集成模型不仅可以显著提高信用风险的预测率还能和识别风险的重要指标相结合[35]。刘传哲等（2018）在传统信用评分的基础上提出有高维数据处理能力的动态异质集成分类模型 DSHE，发现该模型预测准确率更高且指标评价下的平均秩更优[36]。阮素梅等（2018）利用拍拍贷的数据研究影响网贷信用违约的核心指标，采取 L1 惩罚 Logit 模型，发现该模型比普通 Logit 模型、支持向量机等有更好的预测效果，不仅能准确预测违约状态还可以分析影响的关键因素[37]。谭中明等（2018）利用人人贷平台的数据，在使用 Logistic 条件回归方程式筛选与目标变量相关度较高的特征基础上，构建基于梯度提升决策树（GBDT）的 P2P 网贷借款人信用风险预测模型，发现该模型各方面性能优于传统统计模型[37]。李汛等（2019）利用人人贷平台的数据，采用了 KNN 模型、SVM 模型以及 CART 模型预测借款人信用风险，将三种模型对比发现 KNN 模型预测效果优于另外两种[39]。邱伟栋（2020）使用 Lending Club 平台的数据，提出通过改进平均数编码方法，编码地址数据之后用 LightGBM 进行搭建网贷平台风险控制模型，发现该方法泛化能力更强，预测更准确[40]。黄建琼等（2020）采用支持向量机模型对网贷借款人信用风险评估，选取影响信用风险的指标变量并通过交叉验证确定最优参数，发现该模型的稳定性和泛化能力都比较好[41]。

在组合模型方面，王文怡等（2018）利用 HLCT 平台的数据，提出 Logistic 和 ID3 决策树融合的模型，从借款人的借款信息、借款人的信用等级以及历史表现三个指标来研究影响借款人是否违约的信用风险因素，发现融合模型的效果优于单个模型的效果[42]。任静（2019）利用 Lending Club 平台的数据对借款人违约的信用风险评估，通过 Lasso 算法对数据进行清洗，然后将 logistic 回归、随机森

林、支持向量机以及 XGBoost 进行对比，根据 ROC 曲线、精准率、召回率等相关指标对比，结果发现集成模型预测效果更好，在此基础上将 Lasso 算法和 XGBoost 模型结合，发现该模型可以显著提高风险预测的精度[43]。李淑锦等（2019）利用人人贷平台的数据信息建立了一套新的信用风险评估指标体系，提出 LightGBM 模型和 Bagging 模型结合的组合模型来度量借款人的信用风险，结果发现融合模型的预测效果高于单个模型的预测效果，能够显著的提高风险预测的效率[44]。姜晨等（2021）运用对初始权值与阈值进行优化的 BP 神经网络和机器学习相结合的 GA-BP 神经网络模型对借款人信用风险进行识别，发现该融合模型预测准确率高于 BP 神经网络模型[45]。

1.2.3 文献评述

通过对国内外的文献研究可以发现，在信用风险指标体系研究中主要是从两个方面来筛选指标：一方面是借鉴以往的指标体系或者银行的信用评估指标体系，从借款人个人信息、借款历史、信用评分等筛选指标。另一方面对借款人信用风险指标选取是基于大数据等进行的定量研究，如用 Lasso 算法或者 WOE 等定量分析来筛选指标。在信用风险评估模型研究中多是用单个模型建模，如传统的 logistic 回归模型或者机器学习下的 LightGBM 模型等。近今年来也有学者通过将单个模型对比，然后融合两个模型建立信用风险评估模型。总体而言，在信用风险指标体系现有研究中，或局限在定性研究层面，又或只局限于一种算法，不能更加精准的筛选指标体系。在 P2P 网贷借款人信用风险评估模型选择中，鲜有将三个模型进行对比并在此基础上融合的。如何更加完善科学的建立借款人信用风险评估模型，存在诸多的改进空间。

1.3 研究内容及方法

1.3.1 研究内容

本文在借鉴国内外学者对 P2P 网贷平台借款人信用风险评估研究基础上，利用 Lending Club 平台的借贷数据，首先引入随机森林算法筛选评估指标，然后分别用 Logistic 回归、随机森林以及 LightGBM 方法构建借款人信用风险评估模型，通过精准率、召回率、f1 分数和 AUC 指标发现上述单模型预测性能和效果有待提升。在此基础上，将三个模型组合起来建立融合模型来评估借款人信用风险，发现其预测效果相对单模型都显著提高，融合模型更有利于科学地预测借款人违约的风险。

本文的研究内容包括：

（1）绪论。本章主要介绍文章研究的背景及意义、国内外文献综述研究和评述，提出文章的研究内容和思路以及存在的创新点。

（2）P2P 网贷信用风险评估方法概述。本章主要介绍了 P2P 网贷信用风险的定义、成因、控制以及相关的评估方法研究。

（3）信用评估模型的理论基础。本章主要是对论文所需要的理论知识做了初步介绍。首先是介绍数据预处理的操作步骤；然后，介绍了随机森林如何选择因素构建信用评估体系；紧接着对搭建借款人信用评估模型所涉及到的理论知识及原理进行了详细描述；最后，对借款人信用风险评估模型的各个指标进行了详细的阐述。

（4） P2P 网贷信用风险建模实证分析。本章主要构建信用风险评估模型，进而证明借款人信用评估模型的有效性和可行性。首先对借贷数据集进行了数据探索分析操作，了解数据集的整体分布情况；然后基于数据探索分析的基础上，对 P2P 借贷数据进行预处理操作；为了构建预测性能更加优越的借款人信用评估模型，使用了随机森林算法对数据集的变量进行选择，选择表现良好、贡献较大的变量作为信用评估模型的指标体系；最后在评估指标精准率、召回率、f1 分数以及 AUC 上验证本论文提出的多模型融合算法在预测借款人是否违约上性能的有效性以及稳健性。

（5）P2P 网贷信用风险防控建议。本章主要根据第 4 章实证结果，分别从网贷平台运营、政府监督以及投资者平台选择方面提出建议。

（6）结论。本章主要介绍了本文研究的结论，最终构建了多模型融合算法构建借款人信用风险模型，并指出了文章可改进的空间以及对未来的展望。

1.3.2 研究方法

本文主要采用文献研究、对比分析的方法来评估 P2P 网贷平台中借款人信用风险问题。

（1）文献研究法。本文主要通过大量阅读国内外学者对借款人信用风险评价的指标筛选方法以及建模方法为文章撰写提供思路。然后对信用风险相关的理论以及建模相关的理论进行概括总结，为后文进行实证分析提供基础。

（2）对比分析法。本文首先通过介绍随机森林算法筛选指标的优势，发现随机森林算法筛选的指标更加科学。然后在随机森林指标筛选的基础上分别构建 Logistic 回归模型、随机森林分类模型、LightGBM 分类模型将三者对比分析其评估效果。借鉴以往学者两模型融合思想，本文将三个单模性融合构建借款人信用风险评估模型，对比发现融合模型评价效果更好。

1.4 研究思路及技术路线

1.4.1 研究思路

首先，通过背景分析和现状分析发现借款人信用风险问题是导致网贷平台频繁暴雷的重要原因，因此构建科学的借款人信用风险评估模型对投资人资金安全和网贷行业持续健康发展有重要意义。因而，将研究问题聚焦于构建科学的借款人信用风险评估体系，根据相关文献和理论的研究确定了构建模型的具体思路和方法。

然后，选取 Lending Club 平台的数据进行模型构建。首先对数据进行初步的处理。其次引入随机森林算法筛选评估指标，然后分别选取 Logistic 回归模型、随机森林分类模型、LightGBM 分类模型构建信用风险评估模型，并分别对三个单模型的预测效果进行比较分析，提出了将三者融合构建借款人信用风险评估模型。

最后对融合模型的预测效果进行分析，发现该模型各方面的性能都显著好于单模型效果。说明基于多模型融合的借款人信用风险模型对借款人违约预测更准确。同时结合研究过程中的数据处理、模型构建融合问题对文章进行归纳总结，提出结论和展望。

1.4.2 技术路线

本文技术路线如图 1.1 所示：

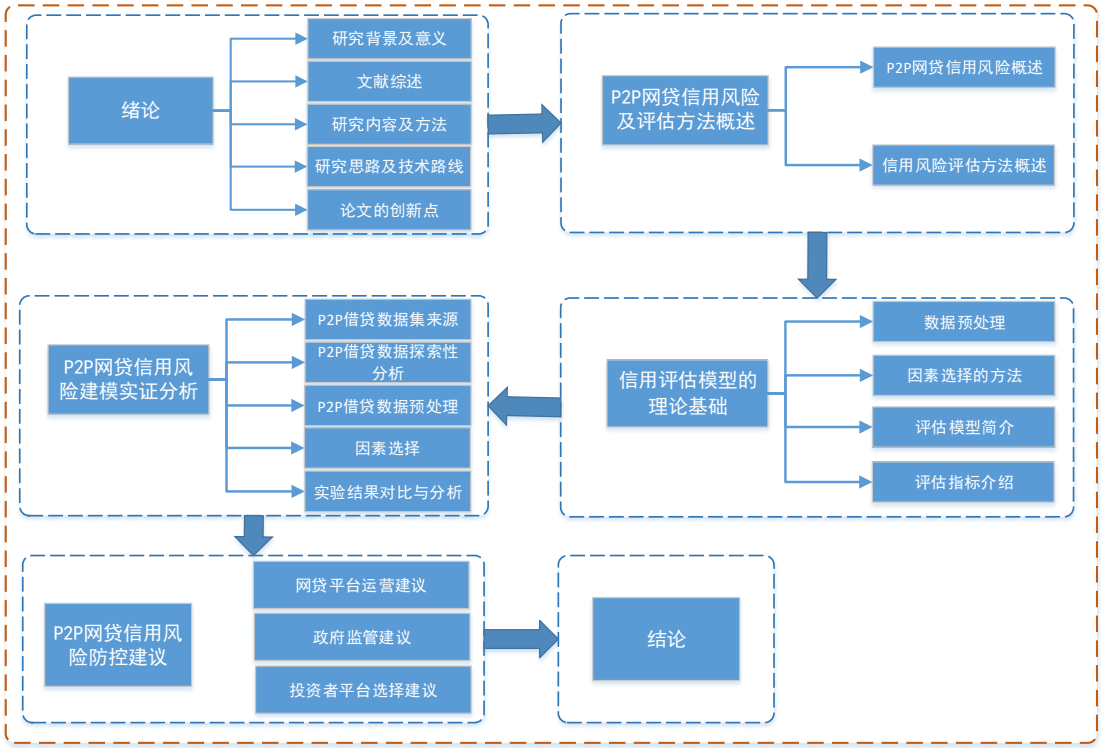

图 1. 1 技术路线图

1.5 论文的创新点

本文的创新点主要有两点：

（1） 提出基于 Logistic 回归模型、随机森林分类模型和 LightGBM 分类模型的融合分类模型，该融合分类模型是基于投票法思想将上述三个单模型的结果根据少数服从多数的原则进行融合，修正单模型出现的预测偏差问题。利用 Lending Club 数据集，选择精准率、召回率、f1 分数以及 AUC 四个指标，将融合模型与 Logistic 回归模型、随机森林分类模型以及 LightGBM 分类模型进行了对比。发现融合模型在四个评估指标上的结果都是最好的，其中精准率为 0.9961、召回率为 0.9765、f1 分数为 0.9862 以及 AUC 值达到 0.9864。

（2）提出使用随机森林算法来进行指标筛选，该方法通过有放回的随机选择样本策略来生成袋外数据集，利用袋外数据集的误差来计算特征重要性，不仅可以提高模型的泛化性能，也不需要额外使用外部数据集来验证模型的性能。

第 2 章 P2P 网贷信用风险及评估方法概述

2.1 P2P 网贷信用风险概述

2.1.1 P2P 网贷信用风险定义

信用风险是指约定双方签订了协议，但是其中一方没能按照协议履行承诺导致另一方承受了风险，也就是违约风险。P2P 网贷信用风险可以分为 P2P 网贷平台信用风险和借款人信用风险，它的划分依据是根据主体来划分。平台信用风险是指网络借贷平台因为自己经营不善导致坏账超出预期，平台资金链条断裂等，最终情况就是平台破产或者跑路。由于 P2P 网贷平台吸收资金主要是来自社会广大的投资者，一旦破产跑路，数千亿元的资金窟窿都是投资者承担，严重的威胁了金融市场的稳定。借款人信用风险是指借款人在 P2P 贷款平台上提交自己的信息，投资者根据信息选择投资，但是借款人逾期未能及时偿还本金和利息，从而使投资人承受了经济损失。本文主要研究的就是因借款人未能及时还款导致的信用风险问题。

2.1.2 网贷信用风险的成因

借款人因为未能及时还款导致的信用风险也有多种原因：一种是借款人有能力按期偿还贷款和利息，但是贷款人处于主观上不想还；另一种借款人主观上想按期偿还贷款且也有一定能力偿还，但是由于一些没有预料到的突发情况，没能按时偿还。其根本是由于信息不对称导致的逆向选择和道德风险的问题以及 P2P 网贷平台的特性导致。

信息不对称是一种现象，是因为不能掌握全部交易信息，使得掌握信息比较充分的一方处于优势地位，掌握信息不充分的一方处于劣势地位。P2P 网贷平台是一种典型的信息不对称市场，借款人通过平台发布借款需求并提交自己的个人信息以及信用状况等寻求资金的融通，但是信息的真实性不能完全确认，再加上一切都是在虚拟的网络平台完成，不能确认借出的资金被借款人用到什么地方，投资者处于劣势地位其投资的资金不能保证安全回收。这是基于信息不对称导致的逆向选择从而造成了借款人信用风险。同时借款人为了以更低的利率获得更高额的资金可能会填写虚假信息，这是基于信息不对称导致的道德风险从而造成了借款人信用风险。

P2P 网贷平台的特性也导致了借款人信用风险问题较为严重。首先 P2P 网贷平台吸纳了很多在传统金融机构借不到贷款的借款人，从根本上他们的还款能力或者征信状况可能就有一定问题，所以一定程度上加剧了 P2P 网贷平台的违约风

险。其次平台的借款人大部分都是个人或者小微企业，两者抵御风险的能力都比较弱。影响借款者个人的还款能力的因素主要是：月收入、资产情况、公司规模或工作年限等信息。如果借款人有较为稳定的收入，则能如期偿还贷款的概率很大，违约风险很低。但是个人的财务状况极易遭受风险，比如失去工作、缺乏稳定收入、资产贬值或者健康状况等都容易导致借款人无力偿还贷款，出现违约风险。影响小微企业还款能力的主要就是市场环境以及自身经营问题，一旦经营不善或者市场环境变化，可能导致企业的资金断裂不能按时还款，出现信用风险。

我国信用监管体系不完善也是借款人信用风险严重的重要原因。不同于发达国家有着完善的信用体系，在信用体系监管下一方面可以共享查看不良信用者的记录，另一方面正是因为监管体系下违约成本较高，所以借款人可能违约的概率也比较小。中国人民银行有一套信用体系记录，但是并未公开。所以 P2P 网贷平台只能靠自己去审核信用的真实性，由于借款人较多且缺乏信用体系的支撑，所以获取借款人的真实信用成本大，准确性也有待考量。再加上 P2P 网贷平台的信用数据也不共享,这就导致信用人在一个平台违约后还可以去其他平台继续借款，违约可能性更大。

2.2 信用风险评估方法概述

总结以往学者对信用风险评估的研究，发现信用风险评估的方法比较丰富。主要可以归纳为两类：一种是定性分析的方法，主要代表就是专家系统法；另一种是定量分析法，主要代表有多元分析统计方法和数据挖掘技术的模型法。

1．定性分析法

定性分析以专家系统法为代表，专家根据贷款人的个人信息、资产状况、历史信用记录等资料，对这些信息进行详细的分析评估，找出影响信用风险的关键因素,并根据影响大小赋予分值,从而得出总体得分来评价借款人信用风险大小。5C 法是其中的典型代表，5C 是指专家从品质（Character）、能力（Capacity）、资本（Capital）、抵押（Collateral）、条件（conditions）五个方面来考量借款人违约的概率。主要是在五个指标基础上结合平台评估者的主观经验进行赋权，最终计算出借款人的信用分。但是，该方法更大的依靠评估人的经验，缺乏数据支撑的基础。

2. 定量分析方法

定量分析方法是根据平台收集到的海量借款人数据，构建合适的模型去评估借款人违约的概率。一类是多元统计分析方法,另一类是基于数据挖掘的模型法。

多元统计分析方法主要分为判别分析法、聚类分析法、回归分析三大方法。借款人信用风险评估常用的多元统计方法就是判别分析法和 Logistic 回归。判别分析法是按照一定的判别规则，建立一个或多个判别函数，用研究对象的大量信

息确定判别函数中的待定系数并计算判别指标，根据结果确定某一类样本数据分类。Logistic 回归是一种多变量统计分析方法，适用于研究被解释变量为二分类或多分类的情况，属于概率型非线性回归。

数据挖掘法主要包括神经网络模型、随机森林模型以及 LightGBM 模型等。以神经网络模型为例，神经网络模型是以神经元的数学模型为基础来描述的，其由网络拓扑、节点特点和学习规则来表示。神经网络模型能够并行分布处理、分布储存及学习，还具有比较高的鲁棒性和容错能力，克服了计量经济模型对假设条件的依赖。但是神经网络容易产生过度拟合问题，对样本的训练时间比较长，可解释性也较差，因此神经网络模型并未在信用评估中得到广泛的应用。随机森林模型作为并行性的集成学习算法，模型训练速度较快且能够处理高维数据，故被广泛应用于信用评估领域的建模问题。LightGBM 模型凭借直方图算法、单边梯度采样、互斥特征捆绑以及限制树的深度，有效地提高了模型的训练速度和精度。故本文选取了多元统计分析中常用的 Logistic 回归模型、数据挖掘方法中常用的随机森林模型和 LightGBM 模型来构建 P2P 网贷借款人信用风险评估模型。

第 3 章 信用评估模型的理论基础

随着人工智能技术的快速发展，信用借贷产业也日趋兴盛，借贷人信用评估在借贷过程中占据越来越重要的地位，因此需要建立一种良好的信用借贷评估体系。若要建立良好的信用评估模型，关键之处在于两点，其一是选择有效的特征因素来构建指标体系；其二是基于选取的指标体系来搭建泛化性能最好的借贷人信用评估模型。本章将介绍上述两点所需要的理论知识。

3.1 数据预处理

在真实的世界里，原始数据或多或少都存在一定的问题，若不对原始数据做任何的预处理操作，会影响模型的最终效果。如图 3.1 所示，一般数据预处理主要包括以下两个部分，即数据清洗和数据转换。

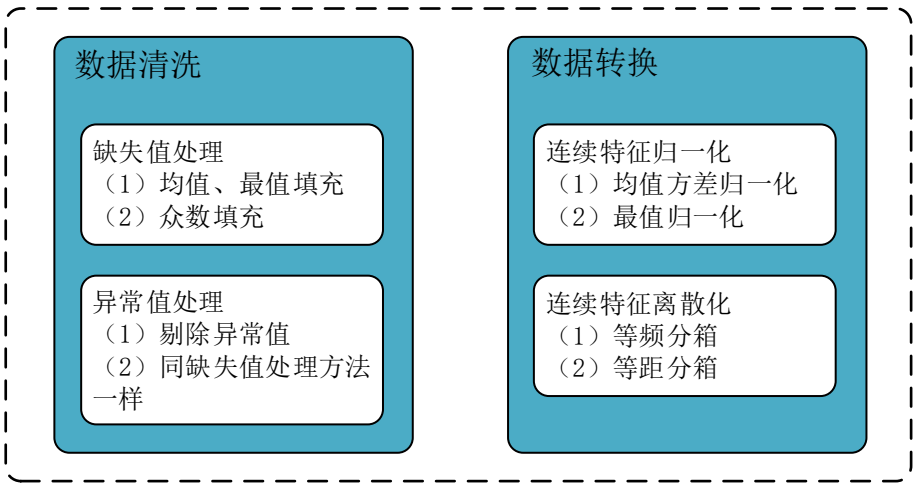

图 3. 1 数据预处理组成部分图

3.1.1 数据清洗

在数据建模过程中，数据清洗是最简单但也是最关键步骤之一，若数据清洗方式不合理，数据的质量反而没有未进行治理的数据质量好，这对于数据模型的性能是负面影响。数据清洗主要包括数据缺失值的处理、数据异常值的检测与处理。数据缺失值的处理主要采取填充法和剔除缺失值的方法。数据异常值的检测运用了常用的箱线图和 3δ 原则。对于检测到的异常值，一般会删除含有异常值的记录，或者将其视为缺失值进行处理。当离散特征存在较多的异常值时，将这类异常值单独作为一种类型数据进行处理和分析。

3.1.2 数据转换

在数据建模过程中，需要通过数据集中特征与特征之间的关联性去挖掘潜在的数据价值问题。主要包括：特征构建、连续特征归一化、连续特征离散化。

特征构建也叫特征衍生，是指基于原始数据集的基础变量来构建新的变量，因为在原始数据集中，很多基础变量没有实际含义，不适合直接建模。属性分割和结合是特征构建时常用的方法，通过将属性分割成多个属性，通过对分割之后的属性进行分析提取更多有用的信息。同时，对同一数据集中具有一定关联的属性进行交叉结合来生成新的特征。

连续特征归一化有两种。一种是对均值方差归一化，将所有数据归一到均值为 0，方差为 1 的分布中。另一种是最值归一化，它是将所有数据映射到区间[0,1]内，该种归一化仅使用于分布有明显边界的数据。

连续型特征进行离散化是通过某些方法将连续的区间划分为多个小的区间，并将这些连续的小区间与离散值关联起来，其本质是决定选择多少个分割点和确定分割点的位置。连续特征离散化可以降低异常数据对其干扰性，可以简化信用评估模型的复杂性以及降低信用评估模型过拟合风险。连续型特征离散化主要通过等距或等宽箱法和等频分箱法。

3.2 因素选择的方法

在进行信用评估模型构建过程中，选择合理有效的特征集合构建信用评估体系是非常有必要的，常见的方法是基于证据权重和信息价值。在本文中，使用随机森林算法来选择特征构建借款人信用评估体系。这里仅介绍随机森林进行特征选择的原理，与下一节中提到的随机森林的算法原理侧重点不同。

随机森林是有放回地从包含 M 个样本的数据集中抽取样本为每个决策树构建新的数据集，则每个样本被抽到的概率为$\frac{1}{M}$，不被抽到的概率为$1-\frac{1}{M}$，则样本在 M 次之后都不被抽到的概率为$(1-\frac{1}{M})^{M}$，对该式取极限可知$\lim_{M\to+\infty}(1-\frac{1}{M})^{M}=0.368$，即 36.8%的样本将从始至终都不会被抽到，这些样本被称为袋外数据（Out of Bag, OOB），常被当作测试集来使用。对于每棵决策树，选择相应的 OOB 来计算袋外数据误差，记为 eOOB1；随机对 OOB 数据所有样本的某个特征 G 加入噪声数据，重新计算袋外数据误差，记为 eOOB2。假设随机森林共有 N 棵决策树，则特征 G 的重要性为$\frac{1}{N}\sum_{i=1}^{N}$(eOOB2-eOOB1)，当这个值越大时，说明特征 G 越重要，原因是对特征 G 加入随机噪声后，OOB 准确率大幅下降，说明特征 G 对于样本的预测结果有很大的影响，进一步说明该特征重要性较高。

基于特征重要性，随机森林进一步进行特征选择，其步骤如下：

（1）计算每个特征的重要性，将每个特征的重要性进行降序排序；

（2）根据要剔除特征的比例来剔除相应的特征，得到新的特征集合；

（3）用新的特征集合重复上述过程，直到剩下预先设置好的特征个数；

（4）经过步骤（1）（2）（3）的计算，可以得到每一轮的特征集以及特征集对应的袋外数据误差率，选择袋外误差率最小的特征集合。最终，随机森林完成了特征选择。

3.3 评估模型简介

在进行借款人信用评估模型搭建时，分别选择随机森林分类算法、Logistic 回归模型算法和 LightGBM 分类算法，以及本文基于投票法机制的多模型融合算法，接下来介绍上述前三个算法的原理以及优缺点，并着重介绍最后一个算法。

3.3.1 Logistic 回归模型

1．Logistic 算法原理

逻辑斯蒂回归(Logistic Regression)算法是目前信用评估模型中运用比较广泛的算法，虽然 Logistic 回归模型中名字带有回归，但却是分类算法。通过 Logistic 回归模型，可以构建自变量与因变量之间的映射关系。Logistic 回归模型是一种广义线性模型，以线性回归作为理论基础，通过 sigmoid 函数为模型引入了非线性因素，因此可以处理二分类问题，即输出类别为两种。接下来具体介绍 Logistic 回归模型的起源以及它的原理。

（1）线性回归模型

Logistic 回归模型是以线性回归模型作为理论支持，线性回归是指利用称为线性回归方程的最小平方函数对一个或者多个自变量与因变量之间的关系进行建模的一种回归分析。假设需要分析的数据集的特征个数为 N，即影响因变量 y 的自变量有$x_1, x_2, \cdots, x_N$，则 y 与x之间的映射函数称为线性回归方程，其表达式见式(3.3)：

$$y = H_v(x) = v_0 + v_1 x_1 + v_2 x_2 + \cdots + v_N x_N \tag{3.3}$$

其中，$v_1, \cdots, v_N$分别是线性回归方程中自变量$x_1, x_2, \cdots, x_N$对应的权重系数，v_0是常量系数。线性回归方程需要求得上述权重系数的最优解，使得自变量可以更好地去拟合因变量，即因变量的值与通过线性回归方程得到的预测值之间的均方误差最小，这里选择均方误差函数作为损失函数其表达形式见式(3.4)：

$$loss(v) = \frac{1}{2M}\sum_{i=1}^{M}(H_v - y^{(i)})^2 \tag{3.4}$$

其中，M 表示总的样本数，$y^{(i)}$表示第 i 个样本的真实值，$\mathrm{H}_v\left(x^{(i)}\right)$表示第 i 个样本的预测值。上述问题就是当式(3.4)的值最小时，求对应的v的值，是一个优化问题，在 Logistic 回归模型中会介绍该问题的优化。

（2）Logistic 回归模型

Logistic 回归模型是在线性回归的基础上添加了 sigmoid 函数操作，使得线性回归模型中引入了非线性因素，sigmoid 函数形式见式(3.5)：

$$\mathrm{f(g)} = \frac{1}{1+\mathrm{e}^{-\mathrm{g}}} \tag{3.5}$$

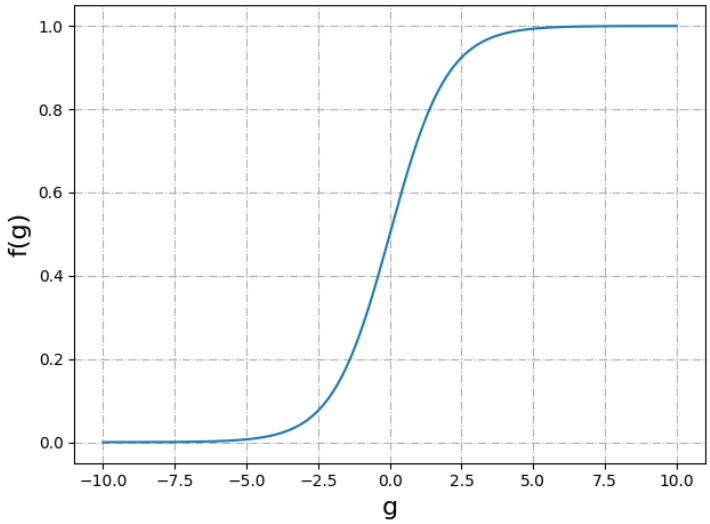

图 3. 2 sigmoid 函数图

该函数可以将任意的一个数映射到区间[0,1]上，如图 3.2 所示。根据线性回归表达式以及 sigmoid 函数可以得到 Logistic 回归表达式，见式(3.6)：

$$H_v(x) = \frac{1}{1+e^{-g_v(x)}} \tag{3.6}$$

其中$g_v(x) = v_0 + v_1x_1 + v_2x_2 + \cdots + v_Nx_N$，通过 Logistic 回归表达式可以将预测的结果映射到[0,1]区间上，可以理解映射到[0,1]区间上的值为取 1 的概率，将该值与所设的概率阈值进行比较，来判断该值是否为 1。例如，在进行借款人是否违约的场景中，假设设置的违约概率阈值为 0.5，则当 Logistic 回归模型对某个借款人的预测结果大于或者等于 0.5 时，则判定该借款人违约，即因变量的结果为 1，若输出的预测结果小于 0.5，则不违约，即因变量的结果为 0。

Logistic 回归模型通过优化求解参数v，最大限度提高 Logistic 回归模型正确分类的概率，这里采用交叉熵损失函数作为损失函数，其表达形式见式(3.7)：

$$\mathrm{cost}(\mathrm{H}_v(x),\mathrm{y}) = \begin{cases} -\log\left(\mathrm{H}_v(x)\right) & \text{if } y = 1 \\ -\log\left(1-\mathrm{H}_v(x)\right) & \text{if } y = 0 \end{cases} \tag{3.7}$$

根据式(3.6)和式(3.7)得到最终需要优化的目标函数：

$$\mathrm{J}(v) = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\mathrm{cost}\left(\mathrm{H}_v\left(x^{(i)}\right), y^{(i)}\right)$$
$$= -\frac{1}{\mathrm{M}}\left[\sum_{i=1}^{\mathrm{M}} y^{(i)}\log\mathrm{H}_v\left(x^{(i)}\right) + (1-y^{(i)})\log\left(1-\mathrm{H}_v\left(x^{(i)}\right)\right)\right] \quad (3.8)$$

这里使用梯度下降法来对参数v进行优化更新，同时需要设置一个学习率常量γ，表示每次更新的幅度，参数v基于梯度下降法的更新如式(3.9)：

$$v_{j(t+1)} = v_{jt} - \gamma\frac{\partial \mathrm{J}(v)}{\partial v_j} = v_{jt} - \gamma\frac{1}{M}\sum_{i=1}^{M}\left(\mathrm{H}_v\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)} \quad (j = 1,2,\cdots,N) \quad (3.9)$$

其中，v_{jt}表示第 j 个权重系数在第 t 轮迭代后的值，$v_{j(t+1)}$表示第 j 个权重系数在第 t+1 轮时的值。

Logistic 回归模型根据式(3.9)不断迭代更新权重参数v的值，最终得到一组较优的权重系数$v_1, v_2, \cdots, v_N$和常量系数v_0，最终 Logistic 回归模型的表达式为

$$\mathrm{y} = \frac{1}{1+e^{-(v_0+v_1x_1+v_2x_2+\cdots+v_Nx_N)}}$$

这里设置阈值为 0.5，当 y≥0.5 时，借款人违约，当 y<0.5 时，借款人不违约。

2. Logistic 回归算法的优缺点

Logistic 回归算法与其他的统计学习方法相比，有着自身的优势，同时也存在诸多不足：

（1）优点：Logistic 回归模型在进行二分类预测时，计算量仅仅与数据集的特征数目相关，因此具有训练速度快的优势；Logistic 回归模型实现简单并且容易理解，可解释性非常好；在进行二分类预测时，不需要对输入的特征进行缩放处理；Logistic 回归模型由于只需要存储各个维度的特征值，因此占用内存空间资源较少。

（2）缺点：Logistic 回归模型不能解决非线性问题，；对于多重共线性数据较为敏感，减少数据之间的多重共线性；当数据集中数据类别不均衡时，Logistic 回归模型效果较差；由于 Logistic 回归模型形式简单，很难去拟合数据的真实分布，准确率不高；

3.3.2 随机森林分类模型

1．随机森林算法原理

随机森林算法是由 Leo Breiman(2001)等人提出的一种集成学习算法，由于其简单，容易实现并且计算开销小，使得它在很多领域得到广泛使用[46]。随机森林使用随机的方法建立一个森林，森林由很多的分类与回归树（Classidfication And

Regression Tree, CART）组成，CART 树不仅可以用于分类预测也可以用于回归预测，并且每个 CART 决策树之间是相互独立的。在建立每一棵 CART 决策树时，使用了随机采样和完全分裂的方式。

随机采样是指随机森林对输入的数据集随机进行行、列采样。行采样是通过自助法重采样技术从原始数据集 M 中有放回地重复随机抽取 m 个样本数据构成每个 CART 决策树的训练数据集，通过随机抽取的方式可以使得每棵 CART 决策树的数据集都不是一样的，这可以有效的防止随机森林模型出现过拟合现象。列采样是指从原始数据集 N 个特征中随机选择 n（n<<N）个特征作为每个 CART 决策树的输入特征集。

完全分裂是指基于前文随机采样后的训练数据集和输入特征集使用完全分裂的方式来构建 CART 决策树模型，对于每个 CART 决策树的叶子节点要么是无法继续分裂，要么该叶子节点下的所有样本都属于同一个类别。在进行 CART 决策树模型构造过程中，根据基尼系数指标来选取某一特征作为 CART 树的最佳分裂点，通过离散穷举的方法来计算出每个特征作为分裂点时，分裂前基尼系数与分裂后基尼系数减少了多少，选取基尼系数值减少最多所对应的特征作为最佳分裂点，基尼系数的值表示数据集的纯度，当基尼系数值越小，表示数据的纯度越高，反之，则越低，计算公式见式(3.1)和式(3.2)。由于本文是对借款人是否违约进行预测，属于分类问题，因此仅介绍回归树的计算公式。

$$\mathrm{GINI}(M) = 1 - \sum_{i=0}^{d} \left(\frac{M_i}{M}\right)^2 \tag{3.1}$$

$$\mathrm{GINI}_{\mathrm{cut}}(M) = \sum_{i=0}^{d} \frac{M_i}{M} \mathrm{GINI}(M_i) \tag{3.2}$$

其中，d表示取值类别数，M_i表示第 i 个类别的数量，M 表示数据集，$\mathrm{GINI}(M)$表示数据集 M 的基尼系数，$\mathrm{GINI}_{\mathrm{cut}}(M)$表示以$M_i$作为分裂点基尼系数的值。

2．随机森林算法的优缺点

随机森林算法相对于其它的机器学习算法以及传统的统计学方法，有着很大的优势，但同时也存在不足之处：

（1）优点：随机森林能够很好的处理高维度数据，并且不需要做特征选择；随机森林算法模型训练完成之后，可以给出每个特征对于结果输出的重要性；由于随机森林中弱学习器之间是相互独立的，所以容易做成并行化方法，加速模型训练速度；随机森林是随机选择训练样本集，所以训练出来的模型具有方差小和泛化能力强的优势；在原始数据集中某些特征值缺失的情况下仍可以保证模型预测的精确度。

（2） 缺点：随机森林在原始数据集含有较多噪声数据时，模型容易出现过拟合情况；取值划分过多的特征容易对随机森林产生重要的影响，从而影响模型拟合的效果。

3.3.3 LightGBM 分类模型

1. LightGBM 算法原理

轻量级梯度提升机器（Light Gradient Boosting Machine，LightGBM）算法是微软公司 2017 年提出来的一种实现梯度提升决策树（Gradient Boosting Decision Tree, GBDT）算法的框架，被广泛应用于多分类、点击率预测等场景任务中[47][48]。LightGBM 算法是集成学习中 Boosting 簇的算法，它主要是为了解决 GBDT 算法在处理海量数据方面遇到的问题，使得 GBDT 可以更好、更快地用于海量数据的业务场景中。LightGBM 模型在传统的 GBDT 算法上主要进行了以下优化：

（1）基于直方图算法寻找决策树节点的最优分割点。如图 3.3 所示，直方图算法是将连续的数值特征离散成 l 个整数，同时构造一个宽度为 l 的直方图。此时在遍历数据时，可以在直方图中将离散后的 l 个整数作为索引累积统计量。当

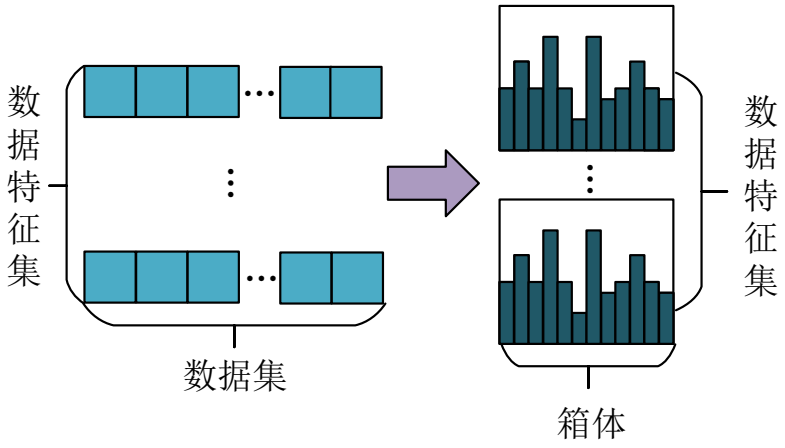

图 3. 3 直方图算法

遍历完一遍数据后，根据直方图的离散值来遍历寻找最优的分割点。

（2）基于梯度的单边采样算法来减少样本的数量，提高模型运行的速度。根据信息增益的定义，梯度大的样本对信息增益有更大的影响，因此基于梯度的单边采样算法的目的是丢弃对信息增益影响较小的样本，同时，为了不影响数据的总体分布，首先将要分裂的特征的所有取值按照绝对值进行降序排序。假设数据集的总样本数为 M，选取每个特征绝对值最大的前$a \times M$个数据，然后在剩下的梯度较小的数据中随机选取$b \times M$个数据。为这些梯度较小的数据乘上一个常数$\frac{1-a}{b}$，

使得梯度较小的数据可以得到更多的关注。最后只需要计算$(a+b)\times M$个数据的信息增益。

（3）基于互斥特征捆绑算法进行特征的选择，降低数据集的维度，进而降低模型的内存消耗和和时间消耗。互斥特征指的是两个特征不会同时为非零值；当两个特征不完全互斥时，即部分情况两个特征都是非零值，此时通过使用冲突比率来计算两个特征之间的不互斥程度，当值较小时，说明两个特征的互斥程度较大。互斥特征捆绑算法主要是将完全互斥的两个特征或者冲突比率较小的两个特征进行融合绑定，进而减少特征的数量，减少构建直方图的时间。

（4）基于带有深度限制的叶子生长（leaf-wise）的决策树生长策略。如图 3.4 所示，该策略每次都会从当前所有的叶子节点中，找到分裂增益最大的一个叶子节点进行分裂，如此循环进行下去，直到树的深度达到最大深度的限制结束分裂。该策略相对于其它按层生长的策略，可以在分裂次数相同的情况下，获取更低的误差和更好的精度。为了防止决策树过深，因此加上树最大深度的限制，可以防止过拟合现象的发生。

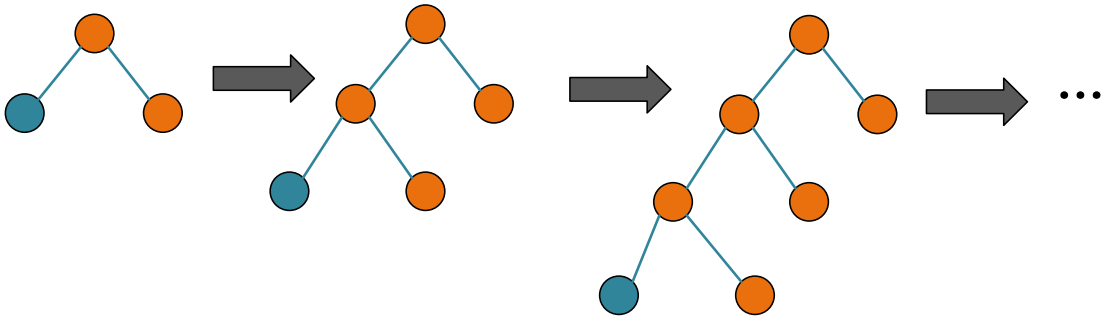

图 3. 4 基于叶子生长的决策树

2. LightGBM 算法优缺点

LightGBM 算法由于其自身的优点使得该算法被广泛应用于各种各样的情景任务中，相对于其它机器学习算法，其优点：

（1）LightGBM 采用了直方图算法，将连续特征离散化，加速模型的训练速度，降低了模型训练的时间消耗；

（2）LightGBM 通过采用直方图算法将连续特征离散化后，只需存储分箱值，降低了内存消耗；

（3）LightGBM 在训练过程中，采用单边梯度算法来过滤梯度较小的样本，减少了模型的计算量；

（4）LightGBM 算法通过使用互斥特征捆绑减少特征数量，加速运行速度和内存消耗。

每一种算法都有优点，同时也存在缺点，LightGBM 算法的缺点：

LightGBM 每一次迭代都是基于上一次迭代预测的结果进行权重调整，随着迭代次数增加，误差减少，模型的偏差也在降低，但 LightGBM 是基于偏差的算法，对数据集中的噪点较为敏感。

3.3.4 基于投票法的多模型融合

上述三个模型除了 Logistic 回归分类模型是单模型之外，另外两个分类模型都是集成学习的代表性算法，虽然算法精度和预测能力相对于一般模型比较优秀，但是为了发挥各个模型的优势，最大化地提高多个模型的预测性能，使用了集成学习 bagging 算法的投票法思想。

1. 算法原理

bagging 算法是由 Leo Breiman 于 1996 年提出的一种将各个弱学习器结合起来，并且每个弱学习器之间是没有依赖关系，相互独立的。bagging 算法通过与机器学习领域内的其他分类算法结合，在提高其准确率、稳定性的同时，降低了结果的方差，从而降低了模型出现过拟合现象的概率。图 3.5 为基于投票法机制的多模型融合算法原理流程图。首先，从原始数据集有放回地抽取部分数据，构成第一个分类器的数据集 M_1，并利用该数据集训练和测试第一个分类器得到预测结果 R_1；然后再从原始数据集有放回地抽取部分数据构成第二个分类器的数据集 M_2，利用同样的方法得到第二个分类器的预测结果 R_2，以此类推，T 轮之后，可以得到 T 个分类模型的预测结果，分别为 $R_1, R_2, \ldots, R_T$。最终，根据少数服从多数的投票法融合策略，将 T 个分类模型预测结果最多的类别作为多模型融合算法的最终预测类别。

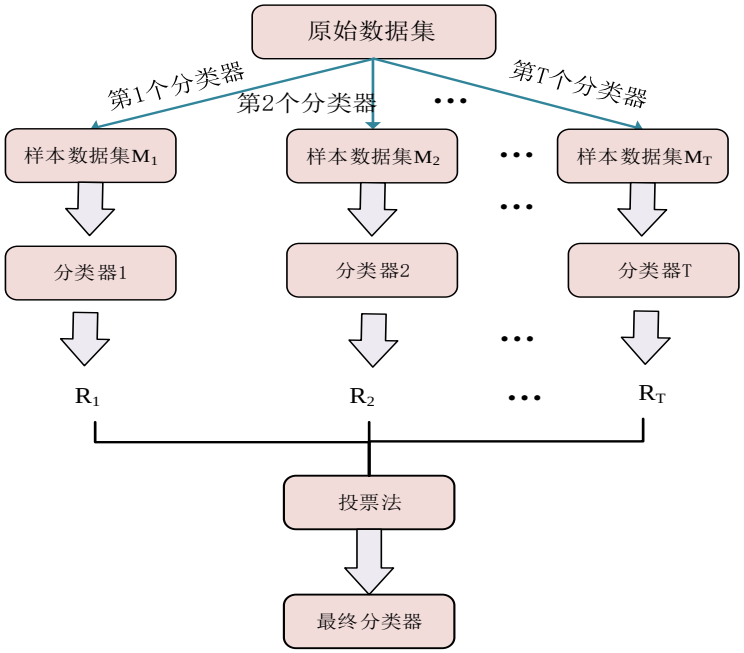

图 3. 5 基于投票法的多模型融合算法示意图

2．基于投票法的多模型融合算法流程

假设多模型融合算法是由 T 个不同的分类器构成：

（1）假设样本训练集总数为 M，则从样本训练集总数为 M 中随机抽取 m 个训练样本作为第 1 个分类器的样本训练集M_1；

（2）利用样本训练集M_1训练第 1 个分类器，并利用测试集测试其性能，得到分类预测结果R_1；

（3）利用步骤（1）到步骤（2）来构造剩下的 T-1 个分类器模型；

（4）根据步骤（1）到步骤（3）循环构造得到 T 的分类器模型以及 T 个分类预测结果$R_1, R_2, \cdots, R_T$。

（5）根据少数服从多数的投票法融合策略，将上述 T 个分类器预测结果最多的类别作为多模型融合的结果输出。

3. 基于投票法的多模型融合算法优点

（1）基于投票法思想的多模型融合算法可以提升单个分类器的拟合能力和泛化能力。

（2）通过投票法对多模型进行融合可以降低预测结果的方差，对异常点的敏感度降低，提升模型的稳定性。

（3）基于投票法思想融合的多模型充分汲取了每个单模型的优点，不仅具有单模型 LightGBM 的训练速度加快、内存消耗降低等优点；同时，利用少数服从多数的投票法可以降低 LightGBM 对数据集中的噪点敏感度；

（4）基于投票法思想融合的多模型不仅具有随机森林可以处理高维数据的优点，同时也可以减少噪声数据的干扰；多模型不仅具备 Logistic 回归模型易操作，实现简单的优点，同时也可以处理多重共线问题。

3.4 评估指标介绍

在机器学习分类模型研究中，普遍采用精准率（Precision）、召回率（Recall）、f_β分数（$f_\beta\ score$）和 AUC（Area Under the Curve）作为评估指标，这些指标的值越大，一般说明该分类模型的性能越好。在本小节将详细介绍这些评估指标的理论知识。

为了更好的理解上述几个指标，首先利用混淆矩阵来解释需要用到的四个结果含义，如表 3.1 所示。横轴方向表示真实值，纵轴方向表示预测值，通过表 3.1 的混淆矩阵可知，TP（True Positive），又叫真正类，表示将正类的样本预测为正类样本的数目；FP（False Positive），又叫假正类，表示将负类的样本预测为正类的样本的数目；FN（False Negative），又叫假负类，表示将正类的样本预测为负

类的样本的数目；TN（True Negative），又叫真负类，表示将负类的样本预测为负类的样本的数目。

表 3. 1 混淆矩阵

	正类（违约）	负类（不违约）
正类（违约）	TP	FP
负类（不违约）	FN	TN

3.4.1 精准率、召回率和f_β分数

精准率，也叫查准率，指正确预测为正类的样本个数占全部预测为正类的样本个数的比例，它的值越高，表示研究的分类模型的性能越好，取值区间为[0,1]。在本论文中，精准率是指构建信用评估体系预测借款人违约的样本数占全部预测借款人违约的样本数的比例，计算公式见式(3.10):

$$\mathrm{P}=\frac{\mathrm{TP}}{TP+\mathrm{FP}} \tag{3.10}$$

其中 P 表示精准率，TP+FP 表示全部预测为违约的样本的数目，TP 表示将违约的样本预测为违约的样本的数目。

召回率，也叫查全率，指正确预测为正类的样本个数占全部样本中实际为正类的样本个数的比例，它的值越高，表示分类模型的性能越好，其取值区间为[0,1]。在本论文中，召回率是指信用评估体系中预测借款人违约的样本数占全部借款人实际违约的样本数的比例，计算公式见式(3.11)：

$$\mathrm{R}=\frac{\mathrm{TP}}{TP+\mathrm{FN}} \tag{3.11}$$

其中 R 表示召回率，TP+FN 表示全部借款人实际违约的样本数，TP 表示将违约的样本预测为违约的样本的数目。一般情况下，都希望二者的值越高越好，但是二者之间是相互制约的，存在此消彼涨的关系，因此需要一个指标兼顾召回率和精准率的值。

f_β分数，是统计学中用来衡量二分类模型精确度的一种指标，它同时兼顾了分类模型的精准率和召回率，其公式见式(3.12)：

$$f_\beta=\frac{(1+\beta^2)\cdot \mathrm{P}\cdot \mathrm{R}}{\beta^2\cdot \mathrm{P}+\mathrm{R}} \tag{3.12}$$

其中，P 表示精准率，R 表示召回率，β表示召回率的权重是精准率的倍数。当$\beta>1$时，表示分类模型评估更加关注召回率的结果，即召回率的权重比精准率的权重大；当$\beta=1$时，表示分类模型既关注召回率也关注精准率，即精准率与召回率同等重要；当$\beta<1$时，表示分类模型评估更加关注精准率的结果，即精准率的权重比召回率的权重大。在本论文中对借款人信用进行评估，给予精准率和召回率同等重要的地位，因此β取值为 1。

3.4.2 ROC 曲线和 AUC 值

ROC 曲线（Receiver Operating Characteristic Curve），即接收者操作特征曲线，是反映敏感性和特异性连续变量的综合指标，曲线上的每个点都反映着对同一信号刺激的感受性。图 3.6 所示是 ROC 曲线图，纵坐标表示真正类率（True positive rate, TPR），即预测为正并且实际为正的样本数目占全部正类样本数目的比例，其计算公式与召回率计算公式相同。横坐标表示伪正类率（False positive rate, FPR），即预测为正类但实际为负类样本的数目占全部负样本数目的比例，其计算公式见式(3.13)：

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{3.13}$$

其中，FP,TN 表达的含义同前文所述。理想情况下，TPR 的值应接近 1，FPR 的值接近 0，即图 3.6 中的(0,1)点。ROC 曲线越偏离 45°，越靠近(0,1)点，表示该分类模型的性能越好。

AUC（Area Under the Curve）是指 ROC 曲线下的面积，将它作为二分类模型评价的指标，是因为 ROC 曲线大多数情况下不能很清晰的说明哪个分类模型性能更好，通过 AUC 值可知，AUC 值越大的分类模型性能越好。根据 AUC 值对分类模型的优劣标准进行以下划分：

（1）AUC = 1，表示该分类模型的性能是完美的。但绝大多数场景下都不存在完美分类模型；

（2）0.5<AUC<1，表示该分类模型优于随机猜测。该分类模型在设置较好的阈值情况下，具有良好的预测价值；

（3）AUC = 0.5，表示该分类模型跟随机猜测一样，模型没有预测价值；

（4）AUC < 0.5，表示该分类模型比随机猜测还差，但可以逆着预测，就优于随机猜测。

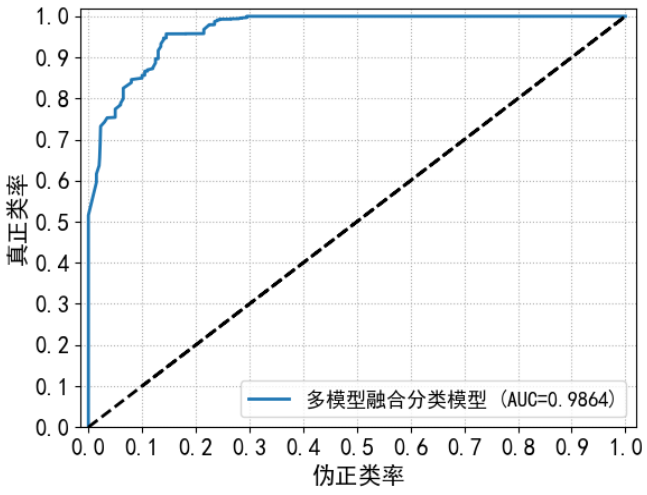

图 3. 6 ROC 曲线图

第 4 章 P2P 网贷信用风险建模实证分析

本章以 Lending Club 公司的数据为例验证本文提出的多模型融合的借款人信用评估体系架构的可行性以及有效性。

4.1 P2P 借贷数据集来源

4.1.1 Lending Club 介绍

Lending Club 是全球最大的 P2P 借贷平台公司，成立于 2006 年，总部位于美国加州旧金山市，其主营业务是为市场提供 P2P 贷款的平台中介服务。与传统的借贷机构不同的是 Lending Club 利用网络技术打造可以直接连接个人投资者与个人借贷者的平台，缩短了资金流通的细节。Lending Club 最大的竞争优势体现在全面网上运营，不用设立实体分支机构，利用互联网技术可以自动处理大量业务，并将投资人与借款人进行合理配对，降低运营成本。但是，Lending Club 最大的特点是它不提供任何资金和风险保障，这使得借贷人在投资时承担很大的风险性。

为降低借贷人的风险，Lending Club 希望通过建立有效的信用评估体系筛选优质借款人、保留一般借款人、拒绝风险较高借款人，并根据不同信用等级划分，实现借款利率差异化定价。为了解决上述问题，本文提出基于多模型融合的借款人信用评估体系架构，进一步提高对借款人信用评估的精确度，最大程度上规避坏账风险，降低借贷人的风险。

4.1.2 数据集的介绍

选择 Lending Club 公司 2016 年第二季度到 2018 年第二季度的数据，共包含 955136 条样本数据，以及借款人借款状态信息、贷款周期、贷款利率、借款人评估等级等 145 条特征信息，表 4.1 是数据集的部分特征描述。

在进行借款人信用评估模型研究时，需要对数据集做一些前期准备工作，一般包括：数据的探索性分析、数据的预处理以及数据特征工程建设。

4.2 P2P 借贷数据探索性分析

在进行数据的预处理操作之前，为了解整个数据集的分布情况以及一般特征与目标特征之间的相关性，对 P2P 借贷数据集进行了数据探索性分析。

4.2.1 借款人借款状态分析

loan_status 字段表示借款人的借款状态信息，是本次构建多模型融合的借款人信用评估模型的目标特征，通过数据集中的其它一般特征来预测 loan_status 的

值。通过使用 Python 编程语言以及第三方库 Pandas 对 loan_status 字段进行分析与处理，如表 4.2 所示，loan_status 字段共有 7 种取值状态，其中取值为 Current 的数量最多，它表示借款人目前处于还款中状态。

表 4. 1 部分特征描述

变量名	含义	类型
loan_status	借款人借款状态	字符型
Term	借款周期	字符型
loan_amnt	借款人申请贷款金额	数值型
home_ownership	住房性质	字符型
verification_status	借款人收入来源是否证实	字符型
purpose	借款人贷款目的	字符型
annual_inc	借款人的年收入	数值型
dti	借款人的负债比	数值型
emp_length	借款人工作年限	字符型
grade	借款人风险等级	字符型
int_rate	借款人的贷款利率	字符型
total_pymnt	借款人已还金额	数值型
out_prncp	借款人未还金额	数值型
total_acc	借款人信用额度总和	数值型
emp_title	借款人职业	字符型

表 4. 2 loan_status 字段取值分析

取值	含义	样本数
Current	借款人处于还款中	706123
Fully Paid	借款人已还清借款	174630
Charged Off	借款人违约	46883
Late(31-120 days)	借款人已经逾期 31 到 120 天	14271
In Grace Period	逾期后处于观察期(逾期 15 天以内)	7878
Late(16-30 days)	借款人逾期 16 到 30 天	4468
Default	借款人违约	883

为了后续模型预测借款人是否违约，需要对 loan_status 字段的值进行转换，转换规则如下：

（1）将 Fully Paid, Late(16-30days)，Current 定义为良好信用并在数据集中统一用 0 表示，即无违约并且信用良好的用户，样本数量为 885221 条。

（2）将 Charged off，Late(31-120days)， Default 定义为违约信用并在数据集中统一用 1 表示，即违约的信用较差用户，样本数量为 62037 条。

转换后的 loan_status 字段的值如表 4.3 所示，不违约样本与违约样本的比例分别为 93.45%和 6.55%，这也符合现实情况，违约的人数比不违约的人数要少很多。

表 4. 3 loan_status 处理后取值分析

取值	样本数量	含义	所占比例
0	885221	无违约，信用良好	93.45%
1	62037	违约，信用较差	6.55%

4.2.2 借款状态与一般变量相关分析

由于数据集中变量数目较多，这里仅选取部分有代表性的变量分析其与目标变量借款状态之间的关系，为了使结构更加清晰，这里分为两大部分进行分析：部分连续型特征与借款状态之间的关系分析，以及离散特征与借款状态之间的关系分析。

1．连续型变量与借款状态之间的相关性

（1）借款人的年收入与借款状态的相关性分析

本文通过对借款人的年收入进行分区间段处理，从图 4.1 可以看出，随着借款人的年收入越来越高，违约的比例在逐步降低，即年收入与违约概率成反比。这符合实际情况，年收入越低，偿还贷款压力越大，违约的概率也就越大。

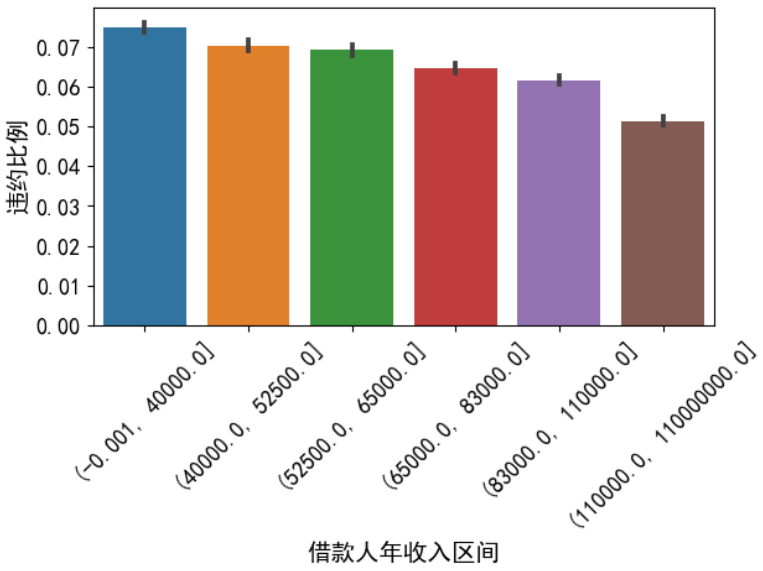

图 4. 1 借款人年收入与违约情况分析图

（2）借款人的借款利率与借款状态的相关性分析

由图 4.2 可知，随着借款利率越来越高，借款人违约的比例也会逐步增加，即借款利率与借款人的违约比例呈正相关。

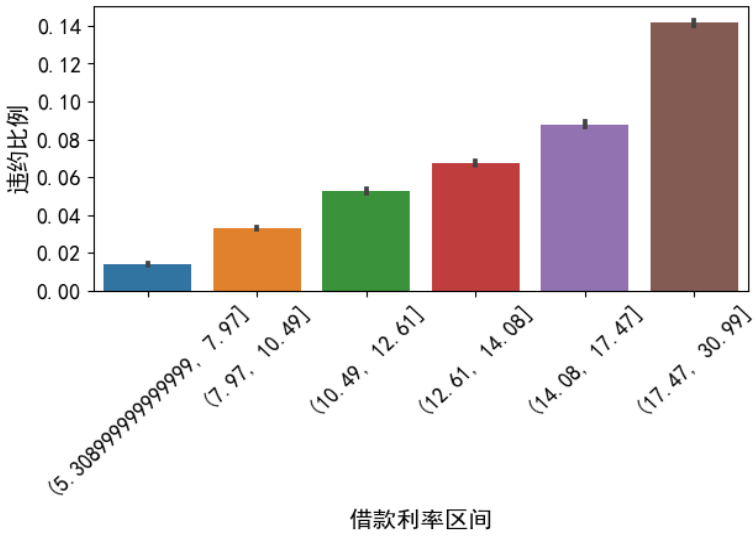

图 4.2 借款利率与借款人违约情况分析图

（3）借款人的负债率与借款状态的相关性分析

负债率表明借款人的还款压力，负债率越高，还款压力越大，则相应的违约比例也会增加。如图 4.3 所示，负债率与借款人违约比例呈正相关性。

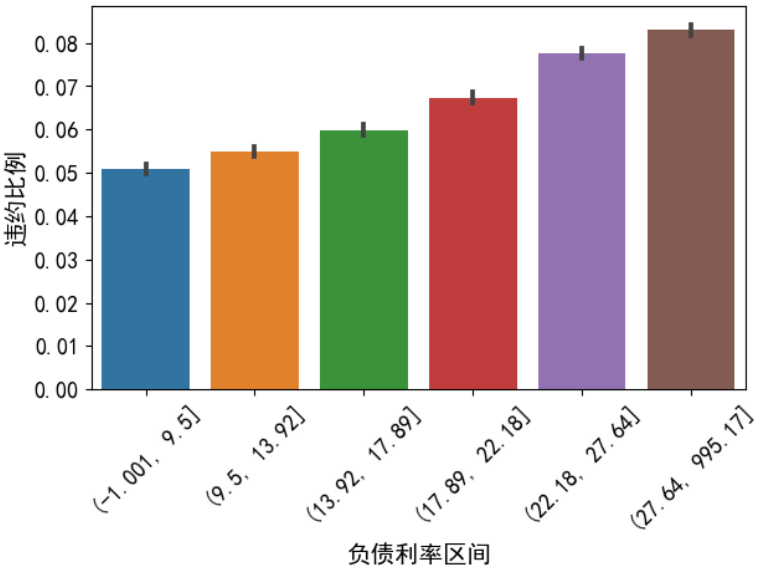

图 4.3 负债率与借款人违约情况分析图

2．离散型变量与借款状态之间的相关性

（1）借款期限与借款状态之间的相关性分析

Lending Club 公司提供给借款人的借款期限一般为 3 年或者 5 年。如图 4.4 所示，可以看出大多数人主要选择短期（3 年）借款，少部分人选择长期借款。长

期借款的违约率为 7.5%比短期借款的 6.2%要高一点，说明随着借款时间越长，借款人违约的风险会增加，即借款期限与借款人违约的风险呈正相关关系。

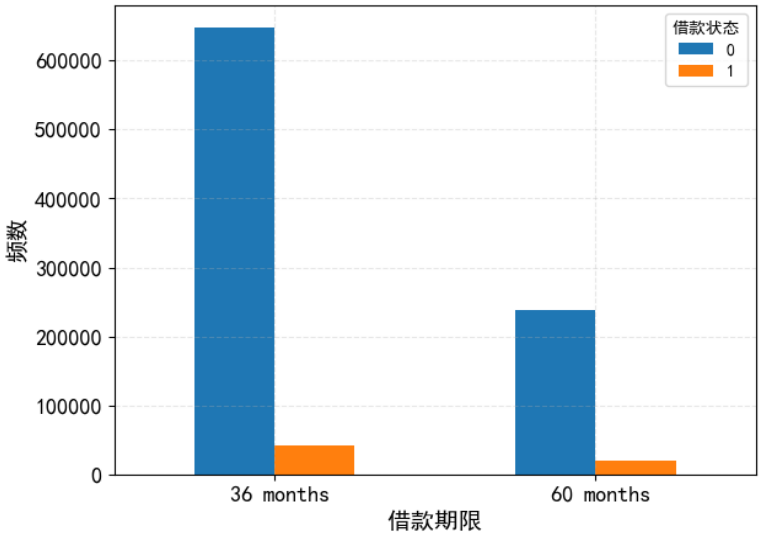

图 4. 4 借款期限与借款状态关系图

（2）借款人的住房性质与借款状态之间相关性分析

从 Lending Club 公司提供的数据集可以了解到，借款人的住房拥有权状况主要分为 4 类，分别是按揭（MORTAGE）、自有（OWN）、租赁（RENT）以及没有（NONE 和 ANY）。如图 4.5 所示，按揭和租赁的违约率最高，分别为 5.53%和 7.72%，这也符合一般家庭租房的情况。

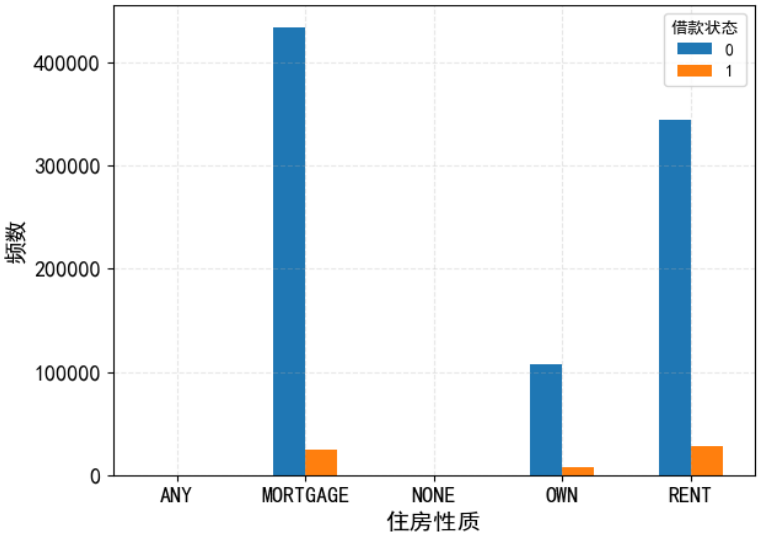

图 4. 5 住房性质与借款状态关系图

（3）收入来源是否证实与借款状态之间的相关性分析

借款人是否违约也可以从借款人的收入来源是否核实这一变量中找到相关性，从 Lending Club 公司所给的数据集以及图 4.6 可以看出，收入经过核实的借款人

违约率却是最高的，达到了 9.33%，而未经核实和收入来源经过核实的违约率分别为 4.03%和 6.86%。

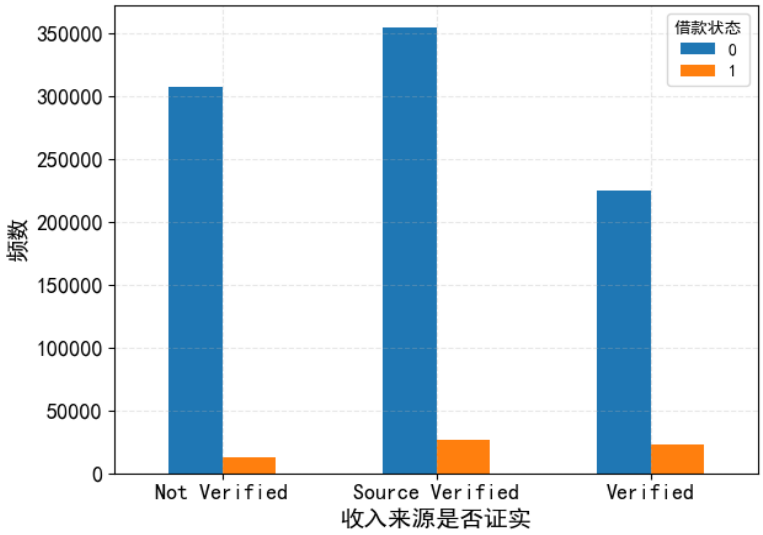

图 4. 6 收入来源是否证实与借款状态关系图

4.3 P2P 借贷数据预处理

4.3.1 数据缺失值处理

Lending Club 公司在进行借款人信息登记时，难免会造成部分数据缺失，如果不对缺失的数据进行处理，则会直接影响最终模型的性能。通过合理有效的处理这些缺失数据，为后续信用评估模型的搭建以及预测性能优化做准备。在第 3 章已经详细阐述缺失值的处理方法，本节主要将这些方法运用到真实的数据集中。

1．直接删除缺失比例过高的特征

对于缺失值较大比例的特征，采取直接删除的办法。表 4.4 为缺失比例超过 25%的部分特征，在本论文中，设置缺失比例阈值为 25%，当某列特征缺失比例超过或者等于这个阈值时，则直接将该列特征删除。缺失比例过高时，特征所含的信息价值不大，反而会对模型性能产生干扰。

2．采用均值或者 unknow 填充缺失值

经过一系列特征筛选操作之后，在剩下的变量中，表 4.5 所示为缺失比例小于 25%的连续性变量，对于缺失的连续型变量采取均值的方法填充；缺失比例低于 25%的离散型变量的有 emp_length 和 next_pymnt_d，分别表示借款人的工作年限和借款人下次还款日期，选择使用 unknow 来填充缺失值。

表 4. 4 缺失比例超过 25%的部分特征

变量名	缺失比例
url	1.0
desc	0.9999
dti_joint	0.9174
hardship_type	0.9964
hardship_status	0.9964
settlement_date	0.9951
mths_since_last_delinq	0.5042
mths_since_last_major_derog	0.7331
revol_bal_joint	0.9274
…	…

表 4. 5 缺失比例小于 25%的连续型变量

变量名	缺失比例
revol_util	0.0009
il_util	0.1432
all_util	0.0001
avg_cur_bal	0.0001
bc_open_to_buy	0.0127
bc_util	0.0131
dti	0.0011
mo_sin_old_il_acct	0.0309
percent_bc_gt_75	0.0127

4.3.2 数据异常值处理

在进行数据预处理的时候，需要对数据集中的各个特征进行异常值的检测与处理，Lending Club 网贷平台出现异常值的原因主要有两个，第一可能是数据采集时出现错误数据采集；第二可能是申请人在填写数据时，为了隐瞒真实信息，随便填写导致的异常。在进行异常值处理的时候，主要分两种情况进行处理，即连续型变量和离散型变量。对于连续型变量，若异常值的个数较少，并且不影响数据集的整体分布时，直接删除异常值所在的样本，以字段 annual_inc 为例，它表示借贷人的年收入，如表 4.6 所示，可以看出字段 annual_inc 的平均值为 79047，

但其最大值为 9522972，比平均值高很多，根据第 3 章描述的两种方法都能检测出最大值为该字段的异常值，本论文采取直接将这个异常值删除。对于离散型变量，出现异常值的时候，若出现的异常样本数量过多的时候，会将这些异常值单独作为一个类别处理，而不直接删除异常样本；若出现的异常样本数量较少的时候，则会直接将这些异常样本直接删除。

表 4. 6 借款人年收入描述性分析

最小值	25%分位值	50%分位值	75%分位值	最大值	平均值	标准差
0	48000	66700	95000	9522972	79047	75723

4.3.3 数据转换

1．离散变量的处理

（1）时间格式的转换

经过前面数据处理的步骤后，在剩下的变量中，包含一些具有时间格式的变量，例如，字段 issue_d、earliest_cr_line 分别表示贷款发放时间和借款人首次征信时间，需要将其格式进行转换，不是为了算法模型更好的理解这些数据，同时也是为了后续特征衍生部分的顺利进行。本文通过使用 python 语言以及第三方库 datetime 对这些字段内容进行转换，转换前后的上述两个字段的形式如表 4.7 所示。

表 4. 7 时间字段转换前后的形式

转换前	转换后
May-16	2016-05-01
18-Jan	2018-01-01

（2）文本类型的变量处理

除了上述时间格式的离散型变量之外，更多的离散型变量的值是以文本形式为主。以字段 home_ownership 为例，它表示借款人在登记时提供的或从信用报告中取得的房屋拥有权状况，其取值分别为 MORTGAGE（按揭）、RENT（租房）、OWN（拥有）、ANY（其它），由于算法模型不接受这种文本形式的数据，同时为了提高模型的性能，需要将这类变量值的形式转换为数字形式。这里以字段 home_ownership 为例展示转换前后的形式，如表 4.8 所示，通过转换成数值型数据之后，就可以很方便的被模型算法理解并使用。

表 4.8 房屋所有权状况转换前后的形式

转换前	转换后
MORTGAGE	0
RENT	1
OWN	2
ANY	3

2．连续型变量的处理

通过对原始数据集的各个连续型变量进行描述性分析，可以发现部分变量内部之间数据值相差甚远，导致数据两极分化显著，如果不对数据进行处理，则会对信用评估模型的预测性能产生干扰。根据第 3 章所述，这里采用最值归一化的思想对变量内部值差距太大的变量进行处理，这里以申请借款金额 loan_amnt 为例，转换前后该变量整体情况如表 4.9 所示。可以看出该变量经过最值归一化处理之后，所有的值被缩放到了区间[0,1]上，减少了数据差距过大带来的影响，对其它连续型变量采取同样的操作。

表 4.9 申请借款金额变化前后概况

描述性分析指标	转换前	转换后
最小值	1000	0.0000
25%分位值	7500	0.1667
50%分位值	12000	0.2821
75%分位值	20000	0.4872
最大值	40000	1.0000
平均值	14897.6	0.3564
标准差	9557.5	0.2451

3．特征衍生

基于原始数据集的变量以及每个变量所代表的意义，构建新的变量进而挖掘数据更深层次的信息价值是数据建模过程中不可或缺的一部分。在本文中，通过对原有变量本身的考量，进行了下述特征的构建：

（1）通过借款人分期付款金额 installment 和借款人年收入 annual_inc 构建借款人每个月的还款金额占月收入的比重 installment_per_rate，比重越大，借款人的违约可能性就越大。构建新特征的逻辑是将 annual_inc 除以 12 得到借款人每个月的收入，然后将 installment 除以借款人每月收入，即可得到 installment_per_rate。

（2）前一小节描述了对时间格式的数据转换，时间字段经过转换后，可以根据字段贷款发放时间 issue_d 与首次使用信用卡的时间 earliest_cr_line 的差值作为一个新的变量 cre_hist，表示借贷人首次借款与本次贷款发放时间的间隔，单位是月份，并将原来的两个时间字段进行删除。

4.4 因素选择

经过前面对借贷数据进行预处理等一系列操作之后，目前借贷数据集中还剩 63 个一般变量和 1 个目标变量借款状态 loan_status，如表 4.10 所示。

由于网贷数据集中变量个数较多，如果不经过筛选就直接建模，不仅会增加模型的复杂度，而且也会延长模型的训练时间，这样的模型既复杂也无法在现实中落地。因此，为了构建一个性能优越的借款人信用评估体系，需要对这 63 个变量进行筛选，选择其中更加重要、对模型性能贡献较大的部分变量构建本文提出的基于多模型融合算法的借款人信用评估模型，本文使用随机森林算法来筛选变量。

表 4. 10 预处理操作后剩下的 64 个变量

变量名	变量名	变量名	变量名
mo_sin_old_il_acct	out_prncp	all_util	loan_amnt
funded_amnt_inv	installment	grade	num_bc_tl
total_rev_hi_lim	total_pymnt	inq_fi	num_il_tl
total_pymnt_inv	emp_length	total_cu_tl	num_op_rev_tl
acc_open_past_24mths	avg_cur_bal	inq_last_12m	num_rev_accts
total_rec_prncp	num_bc_sats	term	inq_last_6mths
bc_open_to_buy	tot_cur_bal	int_rate	num_sats
home_ownership	open_acc_6m	purpose	out_prncp_inv
mo_sin_rcnt_tl	open_act_il	dti	pct_tl_nvr_dlq
verification_status	open_il_12m	open_acc	percent_bc_gt_75
total_il_high_cred_limit	open_il_24m	revol_bal	tot_hi_cred_lim
num_rev_tl_bal_gt_0	bc_util	total_acc	total_bal_ex_mort
num_tl_op_past_12m	il_util	annual_inc	total_bc_limit
num_actv_rev_tl	total_bal_il	mort_acc	open_rv_12m
mo_sin_old_rev_tl_op	open_rv_24m	num_actv_bc_tl	cre_hist
mo_sin_rcnt_rev_tl_op	max_bal_bc	total_rec_int	loan_status

随机森林算法通过有放回的随机选择样本的策略，使得在进行变量筛选时可以使用袋外数据集来计算每个特征的重要性。本文采取 Python 语言以及第三方库 sklearn 中 RandomForestClassifier 函数来输出每个特征的重要性。

如图 4.7 所示。以特征重要性 0.0035 为阈值进行筛选，将特征重要性低于 0.0035 的特征进行剔除，如表 4.11 所示，此时共剩下 46 个变量用于最终信用评估模型的搭建。

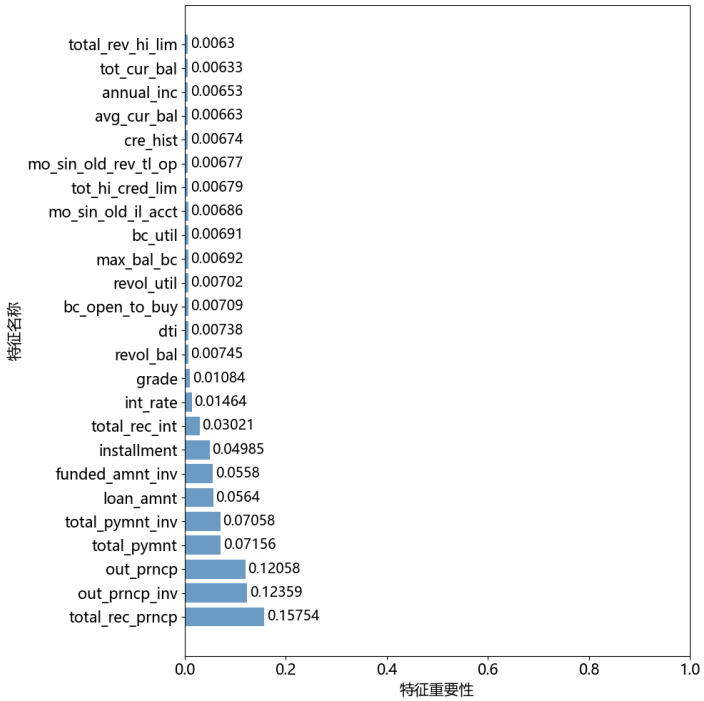

图 4. 7 特征重要性图

4.5 实验结果对比与分析

在上一节通过随机森林算法选出了对模型贡献较大、重要性高的变量，这些变量构成了借款人信用评估模型的指标体系。完成变量筛选工作之后，将进入模型搭建和模型评估阶段。本文在评估分类模型性能时，选用了信用风险常用四个评估指标精准率、召回率、f1 分数以及 AUC，即能有效的区分违约用户与不违约用户。其中，ROC 曲线被用于检验模型区分违约与不违约用客户的效果，混淆矩阵主要用于检验信用风险模型的精确程度。同时，选取了 Logistic 回归模型、随机森林分类模型和 LightGBM 分类模型与本文提出的多模型融合算法进行性能比较，具体过程将在后续章节叙述。

表 4. 11 最终用于模型搭建的变量

变量名	变量名	变量名	变量名
total_bal_ex_mort	out_prncp	revol_bal	total_bal_il
total_rev_hi_lim	total_pymnt	il_util	total_acc
total_pymnt_inv	loan_amnt	all_util	bc_util
funded_amnt_inv	installment	dti	term
num_rev_tl_bal_gt_0	inq_last_12m	total_bc_limit	num_rev_accts
total_il_high_credit_limit	out_prncp_inv	grade	mo_sin_rcnt_tl
bc_open_to_buy	total_rec_prncp	tot_cur_bal	num_il_tl
mo_sin_old_rev_tl_op	max_bal_bc	annual_inc	num_bc_tl
mo_sin_rcnt_rev_tl_op	tot_hi_cred_lim	avg_cur_bal	int_rate
mo_sin_old_il_acct	revol_util	cre_hist	pct_tl_nvr_dlq
num_actv_rev_tl	open_acc	total_rec_int	num_sats
acc_open_past_24mths	num_op_rev_tl	loan_status	

4.5.1 借贷数据集的划分

在进行信用评估模型搭建时，需要将原始借贷数据集划分为训练集和测试集。训练集用于训练模型和寻找模型的最优参数；测试集用于验证已经训练好的模型分类性能。在划分数据集之前，本文进行了平衡样本的操作。从章节 4.2.1 可知，借贷数据集是严重不平衡样本，即违约比例与不违约的比例相差 15 倍，虽然这很符合实际情况，但是为了构建一个更好的信用评估模型，本文采用合成少数类过采样技术（Synthetic Minority Oversampling Technique，SMOTE）对该不平衡样本中的少数类样本进行过采样[49]。最终实现了违约样本和不违约样本比例为 1:1，过采样之后数据量总计为 1760480 条。

使用第三方库 sklearn 中的 train_test_split 函数将过采样后的数据集按照 7:3 比例随机分割成训练数据集和测试数据集，如表 4.12 所示。从训练集和测试集可以看出，违约样本和不违约样本的数量分布很平衡，并且训练集和测试集的分布相似。

表 4. 12 训练集和测试集样本分布情况

	训练集	测试集
违约样本数	615538	264702
不违约样本数	616798	263442
总样本数	1232336	528144

4.5.2 Logistic 回归模型

首先，选 Logistic 回归模型进行建模。在搭建 Logistic 回归模型时，使用第三方库 sklearn 中的 LogisticRegression 函数。在构建过程中，有以下几个参数会影响 Logistic 回归模型的分类性能，即正则化系数的倒数 C 以及惩罚项 penalty。通过使用网格搜索算法对这两个参数的取值进行寻优，得到 C 的最优值为 10 以及 penalty 取 l2 范式。将 4.5.1 节中借贷数据集的训练集部分用于 Logistic 回归模型的训练，然后使用测试集来验证已经训练完成的 Logistic 回归模型的泛化性能。在测试集上验证的结果如表 4.13 所示，可以看出 Logistic 回归模型在四个评估指标上的结果都比较差，也正因为 Logistic 回归模型本身实现简单，导致其无法构建分类性能较好的信用风险评估模型。从图 4.8 可知，该模型的 AUC 值为 0.6459，说明其不能很好的区分违约用户与不违约用户。

表 4. 13 Logistic 回归模型结果表

精准率	召回率	f1 分数	AUC
0.6368	0.6836	0.6593	0.6459

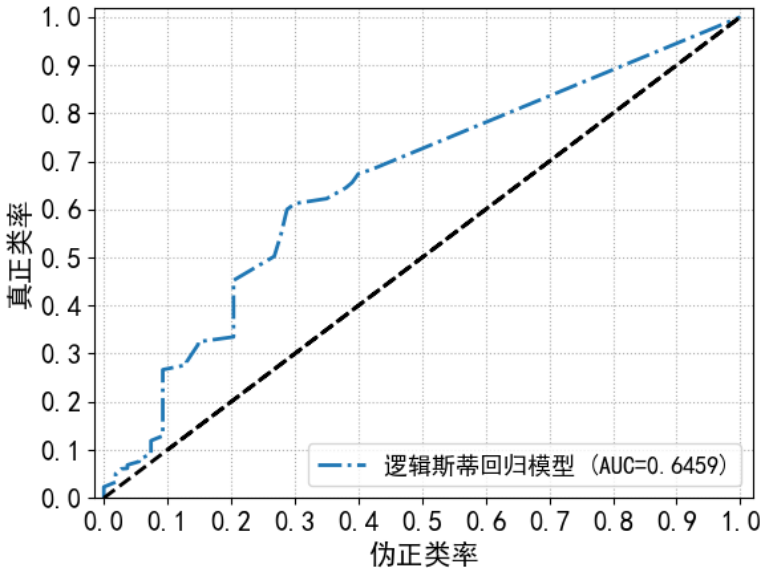

图 4. 8 基于 Logistic 回归模型的 ROC 曲线图

从表 4.14 可以看出，该模型对违约用户进行预测时，预测为不违约用户的数量为 83762 条，即把违约用户预测为不违约用户的数量，同时，也可以看出，把不违约用户预测为违约用户的数量为 103192 条，这已经是一个很大的错误量。因为在实际业务中，把不违约用户预测为违约用户或者把违约用户预测为不违约用户都会给借贷人带来较大的风险。从该表可以看出，该模型的精确度不好，不能满足较高的预测性能需求。

表 4. 14 基于 Logistic 回归模型的混淆矩阵

	违约	不违约
违约	160250	103192
不违约	83762	180940

4.5.3 随机森林分类模型

Logistic 回归模型虽然实现简单、可解释性好、便于理解以及运行速度快，但是经过验证发现，该模型的泛化性能差，无法满足实际的业务需求。接下来选择使用集成学习中的随机森林分类模型进行建模。本文采用的随机森林分类模型如前面章节介绍的那样，采用多个 CART 决策树作为弱学习器。构建随机森林分类模型使用了 Python 语言和第三方库 sklearn 中的 RandomForestClassifier 函数。根据随机森林的原理可知，在构建信用风险模型时，有 3 个参数比较重要，即弱学习器的个数 n_estimators、是否采用袋外样本来评估模型的好坏 oob_score 以及决策树的最大深度 max_depth，如果值设置不合适，则会影响随机森林分类模型的性能。当 n_estimators 值设置过小时，容易造成模型的欠拟合，当 n_estimators 设置过大时，会导致模型运行速度慢，一般设置为 100 的整数倍数。max_depth 表示每个 CART 树的最大深度，在本文中，将值设置为 10。为了反应模型的泛化性能，这里将 oob_score 设置为 true。设置好随机森林分类模型参数之后，即可使用前面章节所述的借贷数据集的训练集来训练模型，此时模型搭建完成，紧接着使用借贷数据集的测试集来验证模型泛化性能。

从表 4.15 可以看出，随机森林分类模型在四个评估指标上的性能完胜于 Logistic 回归模型，分类效果完全优于 Logistic 回归模型。通过表 4.16 的混淆矩阵，表明基于集成学习的随机森林分类模型能够更好地利用样本信息，降低对用户是否违约的错分个数。与 Logistic 回归模型相比，随机森林分类模型将错分个数量从 186954 减少到 17226，极大地降低了借贷人所承受的风险和促进网贷平台的良性发展。通过图 4.9 可知，随机森林分类模型的 AUC 的值为 0.9674，表明该模型已经能够很好的区分违约用户和不违约用户。

表 4. 15 随机森林分类模型结果表

精准率	召回率	f1 分数	AUC
0.9955	0.9391	0.9665	0.9674

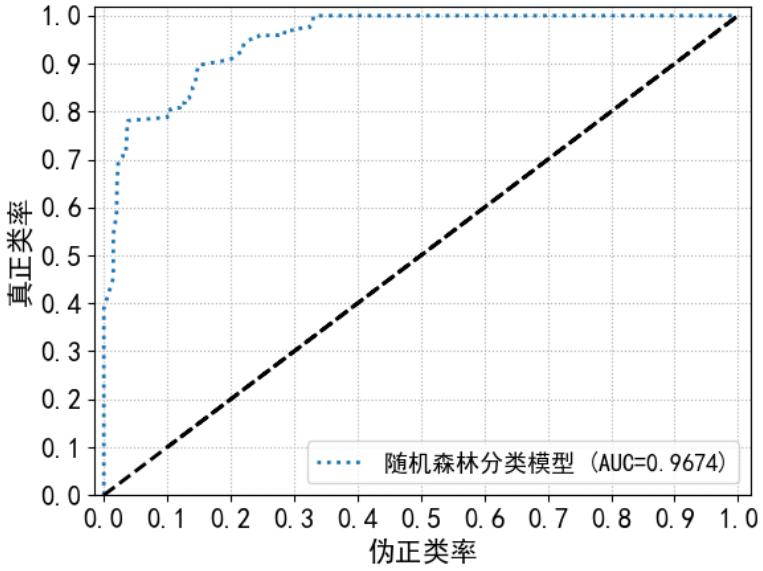

图 4.9 基于随机森林分类模型的 ROC 曲线图

表 4.16 基于随机森林分类模型的混淆矩阵

	违约	不违约
违约	262338	1104
不违约	16122	248580

4.5.4 LightGBM 分类模型

在构建 LightGBM 分类模型时，通过使用 K 折交叉验证的方法来判断模型的效果，其中 K 折交叉验证就是将前面章节中已经划分的训练集随机分成 K 份，每次取其中的 K-1 份数据集用于训练模型，剩下的 1 份数据集用于验证模型，提高模型的泛化性能，在本文中，将 K 值设置为 6。

本文构建 LightGBM 分类模型的实验环境是 Windows 10 系统，16GB 内存，程序环境是 python，使用微软公司开源的 LightGBM 库来搭建 LightGBM 分类模型。搭建模型过程中，通过网格搜索函数寻找以下几个参数的最优值：训练方式 boosting_type 设置为 gbdt、训练目标 objective 设置为 binary、评价函数 metric 设置为 auc、树的最大深度 max_depth 设置为 6 以及学习率 learning_rate 设置为 0.2。模型参数设置好之后，接下来就是模型的训练阶段和验证阶段。

结合表 4.17 可以看出，LightGBM 分类模型在四个评估指标上的结果都比前两个分类模型要好，其中，召回率指标从随机森林分类模型的 0.9391 提升到了 0.9692，说明 LightGBM 分类模型能够更加精准地预测出借款人违约的概率，进一步降低了借贷人与网贷平台所承受的风险。同时，根据图 4.10 可知，LightGBM 分类模型 AUC 的值为 0.9772，相比于随机森林分类模型，它更加能够很好的区分违约用户与不违约用户。

表 4. 17 LightGBM 分类模型结果表

精准率	召回率	f1 分数	AUC
0.9852	0.9692	0.9771	0.9772

结合表 4.18 也可以看出，LightGBM 分类模型大大减少了预测借款人是否违约错分个数，从随机森林的 17226 减少到了 12029 个，LightGBM 分类模型极大提升了预测借款人是否违约的精确度。

表 4. 18 基于 LightGBM 分类模型的混淆矩阵

	违约	不违约
违约	259578	3864
不违约	8165	256537

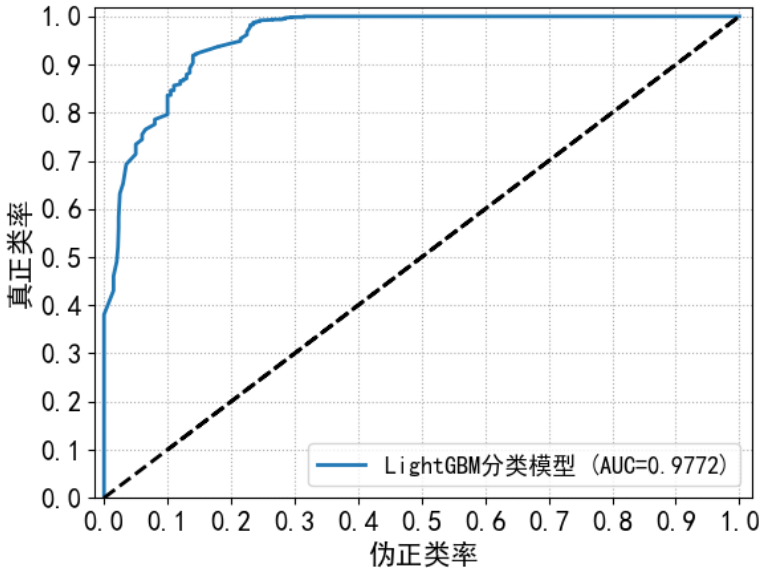

图 4. 10 基于 LightGBM 分类模型的 ROC 曲线图

4.5.5 多模型融合分类模型

Logistic 回归模型、随机森林分类模型被广泛应用于构建 P2P 网贷信用风险评估模型，主要原因是 Logistic 回归模型具有实现简单，可解释性好，运算速度快等；随机森林能够很好的处理高维度数据，并且在特征缺失的情况下还可以保证预测的精度。但同时，由于 Logistic 回归模型形式简单，很难去拟合数据的真实分布，准确率不高；随机森林在噪声数据较多的时候，模型容易出现过拟合现象。基于以上考虑，为了充分利用二者的优势，互相弥补缺点，将二者模型进行融合，提高单个模型的预测性能。为了极大化地提高网贷平台信用风险评估模型

的性能，同时，数据竞赛中常用的算法 LightGBM 分类模型由于其在保证较高精度的前提下能够减少运行时内存和时间的消耗，也将作为融合模型的一部分。为了提高网贷平台对借款人的信用风险合理预测的效果，减少网贷平台和借贷人在放贷过程中的损失，本文提出了一种多模型融合算法的分类模型以及基于多模型融合的借款人信用评估模型的完整架构体系。如图 4.11 所示，该架构体系主要由三个部分组成，数据准备阶段、因素选择以及多模型融合算法，其中，数据准备阶段主要包括数据探索分析、数据预处理以及数据特征工程，该部分是将原始数据处理成模型需要的干净数据；因素选择部分是使用随机森林算法进行了特征选择，通过剔除无意义的特征，提高最终模型的性能。

多模型融合算法是本文创新之处，其基本思想是通过将三个模型的结果利用投票法思想融合在一起，可以对单模型的结果进行修正，提高模型对信用风险评估能力，具体实现过程如下：

（1）首先是利用因素选择部分得到的特征选择集合 featureSet；

（2）将包含 featureSet 的数据集分别输入到 Logistic 回归模型、随机森林分类模型和 LightGBM 分类模型中进行训练与测试，进而得到三个结果，分别为 result1，result2，result3。

（3）利用投票法思想将上述三个模型的结果进行融合，例如，对于某客户，现在利用三个模型对其进行是否违约的预测，Logistic 回归模型预测的结果为违约客户、随机森林分类模型的预测结果为不违约客户和 LightGBM 分类模型的预测结果为不违约客户，此时，根据少数服从多数的投票法思想，可以判定该客户是不违约客户。

（4）最后通过精准率、召回率、f1 分数以及 AUC 作为评估指标来评估多模型融合的分类模型性能。

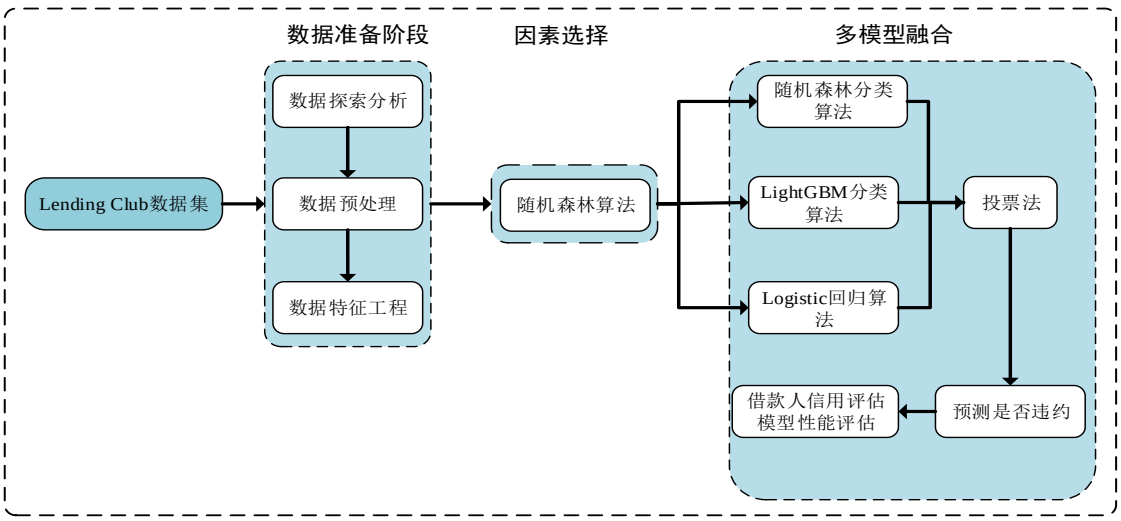

图 4. 11 基于多模型融合的借款人信用评估模型架构图

通过上述步骤，可以将不同模型的带有差异性的结果融合在一起，修正单模性的预测偏差，克服每个单模性的不足，充分利用每个单模型的优势，实现信用风险评估性能的最大提升，促进网贷平台的良性发展。

表 4. 19 四种模型结果对比表

模型	精准率	召回率	f1 分数	AUC
Logistic 回归模型	0.6368	0.6836	0.6593	0.6459
随机森林分类模型	0.9955	0.9391	0.9665	0.9674
LightGBM 分类模型	0.9852	0.9692	0.9771	0.9772
多模型融合算法	0.9961	0.9765	0.9862	0.9864

由表 4.19 可知，Logistic 回归模型作为单模型，在四个评估指标上的结果都是最弱的，而基于集成思想的随机森林分类模型和 LightGBM 分类模型表现良好。同时，可以看出本文提出的多模型融合分类模型几乎在四个指标上都是最好的，说明多模型融合分类模型的分类预测能力最好。多模型融合算法的优势在于充分考虑了单模型的不足，通过其它多个单模型的结果来修正某一个单模型分类预测方面的偏差，彼此之间进行优势互补，进而提升多个单模型组成的多模型分类预测的性能。

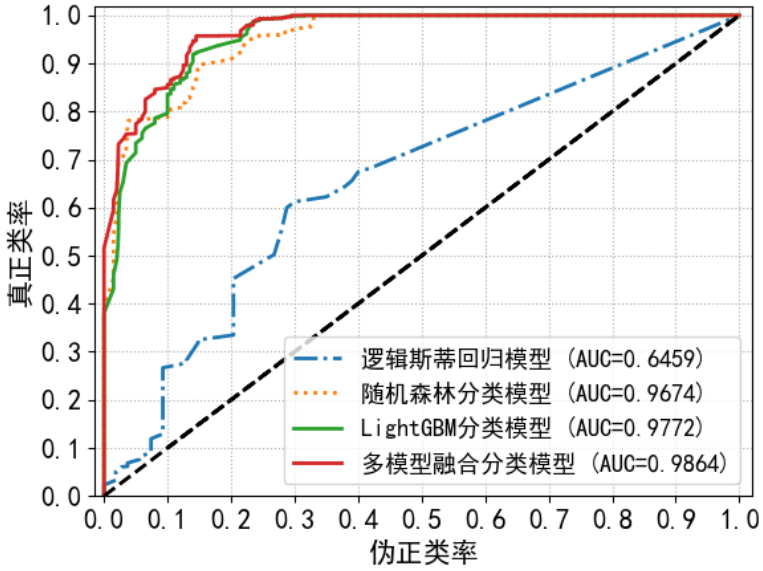

图 4. 12 基于四种模型的 ROC 曲线对比图

通过图 4.12 可以看出，多模型融合分类模型的 ROC 曲线下的面积 AUC 值最大，为 0.9864，相对于其他三个分类模型，它是最可靠的模型，可以用来判断借款人是否违约。根据表 4.20 可知，多模型融合分类模型进一步减少了对借款人是否违约预测错误的个数，从 LightGBM 分类模型的 12029 减少到了 7209，极大地提升了模型预测借款人是否违约的精确度。

根据以上实验结果可知，本文提出的多模型融合分类算法可以很好的预测借款人是否违约，充分证明了此模型是真实有效的。同时也说明了基于多模型融合的借款人信用评估模型整体架构的可行性和有效性。

表 4. 20 基于多模型融合算法的混淆矩阵

	违约	不违约
违约	262442	1000
不违约	6209	258493

第 5 章 P2P 网贷信用风险防控建议

信息不对称、网贷平台自身特性以及我国信用监管体系不完善导致的借款人信用风险问题，对投资者个人和整个行业都有不利的影响。首先借款人违约严重地损害了投资者的利益；其次大规模的借款人违约导致网贷平台频频暴雷，投资人不愿对 P2P 网贷行业进行投资，从而致使一些发展规范、风险水平比较低的优良平台也无法吸纳资金，整个行业的发展受到阻碍；最后 P2P 网贷行业作为资金融通的平台，为各行各业提供资金融通。一旦 P2P 网贷行业出现大规模的违约风险，在互联网环境下风险会加速扩散，从而波及到其他企业的经营状况，造成其他行业的危机，不利于市场稳定。因此，加强 P2P 网贷平台信用风险的防范与控制是至关重要的。

5.1 网贷平台运营建议

研究先进的信用风险评估模型。从本文构建的 P2P 网贷平台信用风险评估模型性能分析可知，引入随机森林筛选特征然后基于多模型融合构建的信用评估模型能够更准确的预测借款人是否违约，有助于帮助 P2P 平台更好的筛选借款人，保障投资人的利益。因此，本文建议应当加大对技术的投入，利用机器学习、深度学习等领域的前沿技术挖掘更加丰富的借款人信息，完善借款人信息特征。选取人工智能领域性能较佳的模型算法对借款人信息进行筛选。同时，对于筛选后的借款人信息利用该领域性能较好的多个模型进行融合，并且不断迭代优化，最终构建性能更佳的借款人信用评估模型。

建立多样化的风险分散机制。我国的 P2P 网贷平台主要靠自己单一的力量运营，所以在遇到市场或者信用风险时抵御风险能力较弱，容易造成资金断流从而破产。像国外学习，和一些实力雄厚的金融机构或者银行紧密合作，不仅仅是在业务资金上，更重要的是在信用体系库上信用资源共享，通过风险分散来进行信用风险控制，以促进平台长久性良性发展。

5.2 政府监管建议

完善信用评价体系建设。信息不对称引起的逆向选择和道德问题是产生 P2P 网贷借款人信用风险问题的根源。正是由于我国信用风险体系建设不完善才导致信息不对称现象的存在，所以要完善国内信用环境，像国外学习建立统一完善的信用数据库系统。在建立数据库系统的基础上既要注意设计信息保护机制，也要

有一定的信息披露机制，对一些信用黑名单的人进行公示，加强大家对信用的重视，减少信用违约风险。

建立网贷行业监管委员会。政府要加强对网贷行业的管控，提高网贷平台的准入门槛。严格审查网贷平台是否有足够的资质经营，对网贷平台的业务经营状况有合理的监管，确保整个行业健康有序的运转。需要建立层次分明、分工明确的政府监管部门。对违法违规经营的 P2P 平台，给予严厉的惩罚，并对处罚结果进行公示，以构建良好、有序的 P2P 网贷行业的经营环境。

5.3 投资者平台选择建议

多角度评估选择网贷平台。投资者在选择网贷平台时，尽量优先选择知名度较高、用户体量较大的平台。知名度较高的平台往往都有比较强大的背景，其背后有优雄厚的资金、人才、技术的支持，而且这类平台拥有较多的用户和标的，这有利于平台内资金的流通，减少平台因借款人信用违约导致的跑路风险。平台的知名度高、用户体量大，对社会的影响力也比较大，所以当平台遇到困难时，政府考虑社会稳定往往对平台进行扶持，尽量保护投资者的利益。再者要理智的投资，不要盲目追求高利率平台。利率和风险成正比，利率越高，借款人信用违约的风险也越高，要根据个人的风险偏好选择适合自己的平台和标的。同时也可以多关注网贷平台以及门户网站的相关评价和动态信息，用户的真实反馈一定程度上可以反映平台的经营运行状况，为投资者提供参考。

结　论

本文针对 P2P 网贷平台借款人信用风险评价建立模型，通过随机森林算法筛选指标，之后建立不同的分类模型以预测借款人信用风险，通过比较分析最终将三个模型进行融合构建多模型融合算法评估借款人信用风险。文章的主要结论如下：

（1）对于建模来说，首要的任务是要对数据进行探索性分析，了解目标变量的界定以及目标变量与一般变量的相关关系，再此基础上对数据进行清洗处理，主要是数据缺失值、异常值的处理以及数据转换，一系列处理后能建立比较准确的分类模型。

（2）指标的筛选对建立模型以及模型的评价效果有很大的影响作用，影响实际的应用。由于借贷信息包含的指标比较多，如果不经过筛选直接建模会增加建模的复杂度也不宜实际情况的应用。因此本文选取了随机森林算法筛选有效的变量，简化模型，发现随机森林可以通过有放回的随机选择样本策略生成袋外数据集，利用袋外数据集来计算特征重要性，不仅可以提高模型的泛化性能，也不需要额外使用外部数据集来验证模型的性能。

（3）本文主要通过将多模型融合分类模型与 logistic 回归模型、随机森林分类模型和 LightGBM 分类模型在召回率、准确率、AUC、F1 分数四个指标上进行对比，验证了基于多模型融合的分类模型在各个指标上的性能优越性和稳健性；同时，也可以看出 LightGBM 分类模型在各个指标上表现良好，表现最差的是 Logistic 回归模型。

总体来说，多模型融合的借款人信用风险评估模型在预测 P2P 网贷平台借款人是否违约上准确性更高，能够帮助平台更好的筛选借款人，保障投资者和平台的利益。为了减少借款人信用违约带来的损失，本文从 P2P 网贷平台、政府监管还有投资人角度提出了针对性的意见来防范控制。但鉴于本人知识和精力的局限，本文存在不足，之后研究者也可进行如下改进：

（1）本文选取的是美国最大 P2P 网贷平台 Lending Club 的贷款数据进行指标筛选和建模，和国内的环境或者数据有一定的差异，之后的研究者可以利用国内 P2P 网贷平台数据按照本文提出的方法建立借款人信用风险评估模型。

（2）本文在筛选变量的时候用的是随机森林算法，之后的研究者可以考虑其他的指标筛选方法，在对变量筛选方法的比较之上可以将指标筛选方法进行融合，找出更加科学有效的变量。

（3）本文仅选取了三种模型进行建模，然后在三个模型的基础上将模型直接融合并没有将模型两两融合和三个模型融合的结果进行对比，在精确度的论证上有待完善的空间。

参考文献

[1] Ravina E. Beauty, Personal Characteristics, and Trust in Credit Markets[J]. SSRN Electronic Journal, 2007.

[2] Jefferson D, Stephan S, Lance Y. Trust and Credit: The Role of Appearance in Peer-to-peer Lending[J]. Review of Financial Studies, 2012(8):2455-2483.Carlos, Serrano-Cinca, Begoña. Determinants of Default in P2P Lending[J]. Plos One, 2015.

[3] Jin Y, Zhu Y. A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending[C]. Fifth International Conference on Communication Systems & Network Technologies. IEEE, 2015.

[4] Emekter R, Tu Y, Jirasakuldech B. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending[J]. Applied Economics, 2015, 47(1-3):54-70.

[5] Lu-Ming Y, Ya-Na W U, Shu-Zheng H. Study on the Operation Quality Assessment of P2P Platform Based on Factor Analysis[J]. Academic Exploration, 2017.

[6] Polena M, Regner T. Determinants of Borrowers' Default in P2P Lending under Consideration of the Loan Risk Class[J]. Games, 2018, 9(4).

[7] Canfield C E. Determinants of default in p2p lending: the Mexican case[J]. Independent Journal of Management & Production, 2018, 9(1):001.

[8] 严复雷,李浩然.P2P 网贷平台信用风险影响因素分析[J].西南金融,2016(10):13-17.

[9] 刘鹏翔. P2P 网贷平台借款人信用风险的影响因素分析——以拍拍贷平台为例[J].征信,2017,35(03):71-76.

[10] 隋昕. 基于 Logit 模型的 P2P 网络借贷平台借款人信用风险影响因素研究[D]. 2017.

[11] 董文奎. P2P 网贷平台借款人信用风险影响因素研究[J]. 时代金融(中旬), 2017(02):311-312.

[12] 雷舰. P2P 网贷借款人信用风险因素分析与对策[J]. 金融理论与实践, 2019, 000(012):31-39.

[13] 舒方媛,赵公民,武勇杰.P2P 网贷借款人违约风险影响因素研究——基于 Logistic 模型的实证分析[J].湖北农业科学,2019,58(04):103-107+119.

[14] 李昕玮.P2P 网络信贷信用风险影响因素研究——基于借款人的信息特征[J].北方经贸,2020(04):85-88.

[15] Noh H J, Roh T H, Han I. Prognostic personal credit risk model considering censored information[J]. Expert Systems with Applications, 2005, 28(4):753-762.

[16] Bekhet H A, Eletter S F K. Credit risk assessment model for Jordanian commercial banks: Neural scoring approach[J]. Review of Development Finance, 2014, 4(1):20-28.

[17] Andreas Mild, Martin Waitz, Jtirgen Wockl. How low can you go? -Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending maxkets[J], Journal of Business Rearch 2015, 68(6)

[18] Sylvester Walusala W, Dr. Richard Rimiru, Dr. Calvin Otieno.A Hybrid Machine Learning Approach for Credit Scoring Using PCA and logistic Regresson [J]. International Journal of Computer,2017,27(1) :84-10

[19] Zhang Z, Gao G, Shi Y. Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors[J]. European Journal of Operational Research, 2014, 237(1):335-348.

[20] Malekipirbazari M, Aksakalli V. Risk Assessment in Social Lending via Random Forests[J]. Expert Systems with Applications, 2015.

[21] Maldonado S, Bravo C, Lopez J. Integrated framework for profit-based feature selection and SVM classification in credit scoring[J]. Decision Support Systems, 2017, 104(dec.):113-121.

[22] Netty Setiawan. A comparison of prediction methods for credit default on peer to peer lending using machine learning. Procedia Computer Science,157:38-45,2019.

[23] Cai S, Zhang J. Exploration of credit risk of P2P platform based on data mining technology[J]. Journal of Computational and Applied Mathematics, 2020, 372:112718.

[24] Tran K, Duong T , Ho Q . Credit scoring model: A combination of genetic programming and deep learning[C]// Future Technologies Conference. IEEE, 2016.

[25] Zeng X, Liu L, Leung S. A decision support model for investment on P2P lending platform[J]. Plos One, 2017, 12(9):e0184242.

[26] A X M, A J S, B D W. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning[J]. Electronic Commerce Research and

Applications, 2018, 31:24-39.

[27] Tong Z, Chen X. P2P net loan default risk based on Spark and complex network analysis based on wireless network element data environment[J]. EURASIP Journal on Wireless Communications and Networking, 2019, 2019(1).

[28] 徐慧婷. 基于 Logistic 的 P2P 网贷借款人信用风险评估研究[J].中国石油大学学报(社会科学版),2017,33(06):16-20.

[29] 李淑锦,詹子涵. 基于逻辑回归的 P2P 网贷信用风险评估研究——以微贷网为例[J].生产力研究,2018(08):29-34.

[30] 马瑞. 基于 Logistic 模型的 P2P 网贷平台违约风险问题研究——以广东省为例[J].特区经济,2019(04):111-114.

[31] 王浩名,马树才. 互联网金融 P2P 贷款违约风险评估、贷款期限和风险溢价[J].财经论丛,2019(07):44-53

[32] 陈雪莲,潘美芹. 基于 Logistic 回归模型的 P2P 借款人信用违约风险评估模型研究[J].上海管理科学,2019,41(03):7-10.

[33] 井浩杰,彭江艳. P2P 网贷平台借款人信用风险评估[J].厦门理工学院学报,2019,27(06):51-56.

[34] 柳向东，李凤．大数据背景下网络借贷的信用风险评估-以人人贷为例[J].统计与信息论坛,2016,31(05):41-48.

[35] 操玮,李灿,贺婷婷,朱卫东.基于集成学习的中国 P2P 网络借贷信用风险预警模型的对比研究[J].数据分析与知识发现.2018,2(10):65-76.

[36] 刘传哲,马达亮,夏雨霏.动态异质集成信用评分模型在 P2P 网络借贷中的应用[J].金融发展研究,2018(09):24-31.

[37] 阮素梅,周泽林. 基于 L1 惩罚 Logit 模型的 P2P 网络借贷信用违约识别与预测[J].财贸研究,2018,29(02):54-63.

[38] 谭中明,谢坤,彭耀鹏. 基于梯度提升决策树模型的 P2P 网贷借款人信用风险评测研究[J].软科学,2018,32(12):136-140.

[39] 李汛，龙真，付怀宇，刘品璐．基于机器学习的 P2P 违约预测算法比较一以“人人贷”为例[J].统计与管理，2019(06):104-109.

[40] 邱伟栋. 基于 LightGBM 模型的 P2P 网贷平台违约预测研究 D]. 2020.

[41] 黄建琼,郭文龙,陈晓峰.基于支持向量机的网贷借款人违约风险评估[J].科技和产业,2020,20(04):40-44.

[42] 王文怡,程平.基于 Logistic 和决策树模型的 P2P 网络借贷信用风险研究——以 HLCT 为例[J].上海立信会计金融学院学报,2018(03):42-55.

[43] 任静. 基于 Lasso-XGboost 模型的 P2P 网贷违约信用风险评估[D].兰州大学，

2019.

[44] 李淑锦,嵇晓佳. LGB-BAG 在 P2P 网贷借款者信用风险评估中的应用[J].技术经济,2019,38(11):117-124.

[45] 姜晨,刘喜波. 基于 GA-BP 神经网络模型的 P2P 网贷借款人信用风险预测研究[J].商展经济,2021(01):49-51.

[46] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[47] LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

[48] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232

[49] Hui H, Wang W, Mao B. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[C]. International Conference on Intelligent Computing. Springer, Berlin, Heidelberg, 2005.