基于集成学习的 P2P 网贷用户的违约预测研究

重庆大学硕士学位论文

（专业学位）

学生姓名：陈旭然

指导教师：张良才　教授

专业学位类别：应用统计

研究方向：机器学习

答辩委员会主席：何传江　教授

授位时间：2022 年 6 月

Research on Default Prediction of P2P Online Loan Users Based on Ensemble Learning

A Dissertation Submitted to Chongqing University

In Partial Fulfillment of the Requirement for the

Master of Applied Statistics

By

Xuran Chen

Supervised by Prof.Liangcai Zhang

June , 2022

摘　要

大数据时代的到来，不断推动着互联网金融的飞速发展，进一步促使了各类P2P网贷平台的数量与日俱增，P2P网贷平台一方面降低了借贷者的融资门槛，更加高效便捷，极大地提高了资金利用效率。另一方面，由于互联网金融的虚拟性、隐蔽性，互联网欺诈的事件层出不穷，如何控制P2P网贷平台用户的信用风险成为当前风险领域的研究难题。同时，对于高维度且海量的用户数据，集成学习算法凭借着高精准性、高稳定性，为识别违约用户、降低平台损失提供了新的解决方法。

首先，本文基于Lending Club平台数据，包含115677条信贷用户数据以及144个变量，包括借款人的资质信息、历史信用情况、贷款利率等信息。对数据进行预处理，包含处理无关变量、缺失值、异常值等，进行变量衍生。

其次，对清洗后的数据进行WOE分箱，通过IV值筛选、相关系数及方差膨胀因子检验、随机森林重要性排序的方法来进行特征筛选，最终选择了11个自变量，以此作为模型的输入变量，构建借款人信用风险的违约预测模型。

最后，分别基于逻辑回归的单一分类算法和随机森林、AdaBoost、GBDT、CatBoost的四种集成学习算法构建用户违约预测模型，通过精确率、召回率、f1得分、AUC值等指标进行评估，结果表明随机森林受不平衡的影响较大，基于AdaBoost、GBDT和CatBoost三种集成算法的分类预测结果均优于逻辑回归的分类预测结果，同时CatBoost的预测精度最高，对平衡后样本的违约用户召回率达到了95%。故在此基础上建立以AdaBoost、GBDT、CatBoost为第一层的基模型，逻辑回归为第二层模型的Stacking融合模型，基于Stacking算法模型的违约用户召回率提升至97%，分类预测效果最好。

本文的创新点在于采用WOE分箱进行有监督的编码来提高变量的解释能力；通过多种方法进行特征筛选，加强了特征筛选的准确性；将CatBoost算法应用于违约预测中，经过与其他集成模型的对比，证明了该算法的可行性及优越性。

关键词：P2P网络借贷；信用风险；逻辑回归；集成学习；WOE分箱

Abstract

The arrival of the era of big data has continuously promoted the rapid development of Internet finance, and further promoted the increasing number of various P2P lending platforms. On the one hand, P2P lending platforms have reduced the financing threshold of borrowers, made them more efficient and convenient, and greatly improved the efficiency of capital utilization. On the other hand, owing to the fictitious and invisibility of Internet finance, Internet fraud events emerge in endlessly. How to control the credit risk of P2P lending platform users has become a research problem in the current risk field. At the same time, for high-dimensional and massive user data, the ensemble learning algorithm provides a new solution for identifying default users and reducing platform losses by virtue of high accuracy and high stability.

Firstly, based on Lending Club platform data, this paper contains 115677 credit user data and 144 variables, including borrower's qualification information, historical credit situation, loan interest rate and so on. The data are preprocessed, including processing irrelevant variables, missing values, abnormal values, etc., and variable derivation is carried out.

Secondly, after cleaning the data by WOE box, through IV value screening, correlation coefficient and variance expansion factor test, random forest importance sorting method for feature selection, finally selected 11 independent variables, as the input variables of the model, build default prediction model of borrower credit risk.

Finally, the user default prediction models are constructed based on single classification algorithm of logistic regression and four ensemble learning algorithms of random forest, AdaBoost, GBDT and CatBoost, respectively. The accuracy, recall, f1 score and AUC value are evaluated. The results show that random forest is greatly affected by imbalance. The classification prediction results based on AdaBoost, GBDT and CatBoost are better than those based on logistic regression, and CatBoost has the highest prediction accuracy. The default user recall for the balanced sample reached 95 %. Therefore, on this basis, the Stacking fusion model is established with AdaBoost, GBDT and CatBoost as the basic model of the first layer and logistic regression as the second layer model. The recall rate of default users based on the Stacking algorithm model is increased to 97 %, and the classification prediction effect is the best.

The innovation of this paper is that WOE bins are used for supervised coding to

improve the explanatory power of variables ; the accuracy of feature selection was enhanced by a variety of methods for feature selection ; CatBoost algorithm is applied to default prediction. Compared with other ensemble models, the feasibility and superiority of the algorithm are proved.

Keywords: P2P Online Lending; Credit Risk; Logical Regression; Ensemble Learning; WOE Box

目 录

1 绪 论

1.1 研究背景与意义

1.1.1 研究背景

互联网技术的快速发展，促进了互联网技术与传统金融的进一步高效结合，在该背景下，互联网金融由此而生，持续推动传统金融实现创新型转型升级。2014年以来，以云计算、大数据、区块链、人工智能为主要特征的第四代互联网金融更是出现井喷式发展。现有的金融模式已无法满足当前经济发展的要求，P2P网贷、众筹、第三方支付平台、大数据金融等各种新的金融发展模式呈现百花齐放、百家争鸣的态势。此外，由于中小企业自身担保能力差、财务管理问题以及银行利率普遍较高、上市过程中资本市场的门槛高等问题，过去中小企业深陷融资难的困境，互联网金融具备交易成本低、大数据技术对贷款人的分析比较精准等优势，为中小企业提供了新的融资模式，推动了普惠金融的健康、绿色发展。

P2P网贷作为互联网金融中蓬勃发展的一种创新金融方式，也遇到一些发展瓶颈，信息不对称、社会信用评估体系不健全、法律法规不完善等问题的出现，无形提高了信用风险，给众多网络金融公司造成了巨大的损失。同时由于互联网技术的虚拟性，P2P网贷平台面临的主要风险，除了传统金融领域的信用风险外，还存在互联网欺诈。其中，互联网欺诈呈现以下几个显著特征[1]：一是欺诈专业化。不再是过去类似于电话欺诈、盗号、盗卡等低级的欺诈方式，如今的欺诈手段更复杂专业，难以识别，如大规模的营销欺诈，进行抽奖、返现奖励等；抵押、借贷欺诈。二是欺诈系统化，不再是个人，而是有明确的组织分工的团队，甚至形成了欺诈灰色产业链，具备完整配套的金融欺诈体系。三是欺诈隐蔽化，利用互联网虚拟这一特点，采用区块链加密技术来保护隐私，给识别及打击欺诈案带来了巨大挑战。因此，如何建立准确有效的借款人信用的的风险评估体系，从而更加科学有效的识别用户的信用风险甚至欺诈行为，是各个网络借贷平台面临的迫在眉睫的难题。

1.1.2 研究意义

① 理论意义

近年来，随着移动互联网技术的飞速发展，传统金融机构及网贷平台逐渐积累了更高维度、更海量的用户数据，使得金融机构在贷前通过定量分析来预测借款人的未来信用情况提供了可能。在当前的研究中，可分为两个阶段。第一个阶段主要采用简单的预测模型进行预测，如线性回归、判别分析、逻辑回归，但这些模型需要满足线性关系的假设。此外，虽然FICO信用评分以逻辑回归为核心，

具有很长的历史，但在高维度数据下表现较差。然而大数据的推动使客户的信用数据数量更大、维度更高，第二个阶段下的集成学习方法应运而生，不断应用于贷款违约预测之中。凭借着其更高的精确程度、更快的运行速度，成为预测信贷违约的重要方法，该方法主要关注提升预测违约用户的精准度和增强变量的可解释性。

本文将多种集成模型应用到违约预测中，将逻辑回归这单一模型，与随机森林、AdaBoost、GBDT、CatBoost、Stacking 的集成模型进行对比，以此来得到表现最优的模型。

② 实践意义

同时，对信贷违约预测的研究也具有重要的实践意义。首先，通过尽可能多地识别出违约用户，能够最大限度地避免金融机构的利益损失。其次，P2P 网络借贷平台作为我国信贷市场的重要组成部分，完善个人信用评估体系有助于信贷市场实现健康、稳定发展。最后，信贷消费能够拉动经济增长，消费型信贷带动商品及服务的消费需求，投资型信贷增强企业的发展动能。

本文基于 Lending Club 网贷平台的借款人数据，采用集成学习算法构建违约预测模型，旨在能够更加高效地识别正常用户与违约用户，从而降低网络借贷用户的信用风险，为中国 P2P 网络借贷平台如何控制信用风险提供理论支持。

1.2 国内外研究现状

1.2.1 国外研究现状

从 20 世纪三四十年代开始，国外就开始了对于违约预测问题的大量相关研究，从违约预测问题研究发展的进程上来看，可分为三个阶段，分别为判别分析法、信用评分法、建立违约预测模型。

判别分析法指在分类明确的情况下，通过属性值判断实例所属某个类别的统计方法。Fisher 在分类问题的基础上，首次提出了线性判别分析（LDA）[2]。David Durand 将线性判别分析应用于信用风险评估，对于贷款申请人，根据个人资质信息、财务状况、历史信用情况等指标将贷款分为“好”和“坏”两类，是最早运用统计模型开展对个人信用风险评估的研究，具有前瞻性，这标志着基于个人的信用风险评估逐渐从定性分析向定量分析的过渡[3]。

FICO 信用评分法，由 Bill Fair 和 Earl Isaac 提出，指基于偿还历史、信用账户数、使用信用的年限等五大影响因素，并赋予相应权重，计算客户的信用得分的客观评分方法，排除了主观因素干扰、缩短授信时间，主要通过借款人的历史信用情况，与该系统内的所有借款人或者经常发生违约、甚至破产等各种深受经济问题困扰的借款人，将信用习惯、发展趋势等进行比较，来预测该借款人未来还

款的可能性[4]。其中典型的例子是 Altman 提出 Z-score 评分模型，首先使用判别分析构建简单违约预测模型，后通过产生信用评分进行违约风险的有序排名[5]。

20 世纪 80 年代以来，随着统计分析模型及机器学习的不断深化与发展，大量机器学习模型也开始应用于信用风险评估领域。Wiginton 最早基于 logistic 回归模型进行违约信用风险的评估，得到了 0-1 之间的概率值，克服了传统线性回归模型取值在正负无穷之间的弊端[6]。Kokkinaki 基于布尔逻辑函数及决策树模型对用户的消费行为进行分析，通过聚类来判断客户是正常客户还是欺诈客户[7]。Baesens 等人提出将最小二乘支持向量机(LS-SVM)应用于真实信用评分数据集，同时与逻辑回归、判别分析、决策树、K 近邻、神经网络分类器进行对比，根据分类精度和 ROC 曲线指标，LS-SVM 和神经网络模型表现出很好的效果，同时逻辑回归和判别分析也表现良好[8]。Oreski 提出了神经网络混合遗传算法(HGA-NN)，通过 UCI 信用数据集建立模型，验证出该算法显著提高了分类准确性[9]。Ala'raj 等人结合多元自适应回归样条(MARS)和 Gabrel 邻域图编辑(GNG)提出了一种新的混合集成模型，与常见的朴素贝叶斯、决策树、人工神经网络、支持向量机 SVM、随机森林五种基分类器进行对比，发现混合集成模型效果更好[10]。Ma 等人采用 LightGBM 模型和 XGBoost 模型对 Lending Club 平台用户的违约情况进行预测，在多种预测指标的对比下，得到 LightGBM 模型预测效果最好[11]。

1.2.2 国内研究现状

相较于国外对违约信用风险的研究，我国对该领域的研究相对较晚、文献较少。

我国初期采用的信用评估模型主要使用美国的 FICO 信用评分法，其主要依赖于逻辑回归模型。于立勇和詹捷辉运用正向逐步选择法筛选指标，同时使用逻辑回归构建违约预测模型，将贷款企业分为正常类与违约类，发现对违约组的正确判别率较高[12]。涂伟华针对商业银行信用卡数据，构建判别模型与逻辑回归模型，最终两者大致相同，判断准确率高于 80%[13]。

由于面对高维度的数据，逻辑回归表现不佳的问题，很多学者也提出了改进方法。方匡南等人提出了 Lasso-logistic，与全变量逻辑回归、逐步回归相比，总体用户的预测准确率与违约用户的预测准确率均达到最高，同时克服了多重共线性问题，也显著增强了模型的稳健程度和可解释程度[14]。过新伟对于上市中小企业的信用风险，构建稳健 logistic 与随机效应 logistic 模型；对于未上市中小企业的信用风险，构建多元有序 logistic 模型，改进后的模型准确率更高、稳健性更强[15]。周丽峰基于数据平衡的思想，提出加权的并行平衡随机森林(WPBRF)，根据 OOB 数据的错误率对每个决策树加权，不仅准确率高于 SVM、C4.5、KNN 等算法，还通过并行运算减少了训练时间[16]。杨斌(2013)在分析制造业企业违约影响因素时，

建立了附带宏观经济变量的生存分析中 COX 回归方程[17]。夏雨霏等人基于 P2P 平台用户数据，提出了聚类支持向量机，先后进行聚类、分类，该模型提高了预测精度、大幅降低了误判成本[18]。刘铭等人将模糊神经网络与狼群搜索算法融合，提出了改进型模糊神经网络(IFNN)，准确度高、泛化性强，有效解决了信用卡违约预测问题[19]。

除了单一模型，近些年来大量学者也尝试利用集成模型来进行贷款违约预测。沙靖岚基于 Lending Club 平台交易数据，采用“多维度”和“多观测”两种角度分别建立样本数据集，同时对两个数据集分别基于 LightGBM 和 XGBoost 两种算法构建模型，研究得到在同种算法下，多观测数据集的分类预测能力更强；同时在同种样本训练集下，LightGBM 算法的分类效果更优[20]。王嘉以拍拍贷的借款人数据为例，首先建立 XGBoost、SVM、随机森林等违约风险模型，后筛选出分类最佳的 XGBoost 算法输出作为第一层的基模型，逻辑回归为第二层的训练模型，通过 Stacking 方法集成，得到 Stacking 集成模型的判别违约用户的能力优于单一模型[21]。杨盛辉提出了加权 Stacking 算法，第一层分类器根据其预测违约客户的分类错误率加权，再以逻辑回归为第二层的次级分类器，提高了违约客户的召回率[22]。马晓君等人将 CatBoost 模型应用于 P2P 平台用户的违约预测中，预测准确率达到 96%[23]。逯瑶瑶采用 Smote 采样进行平衡处理，之后以 XGBoost 模型和 LightGBM 模型为初级分类器，逻辑回归为次级分类器，构建 Stacking 融合模型，对借款人的违约情况预测，得到 Stacking 模型的识别效果最佳[24]。刘美伶将基于 AHP 网络层次分析法计算出的信用评分作为新的变量加入到 LightGBM 模型中，在准确性、适用性、功效性等各方面有更好的表现[25]。

1.3 研究内容及框架

本文主要研究的是采用集成学习的方法来提高对于 P2P 网络借贷平台用户的违约贷款的预测能力，主要分为六大部分，分别为绪论、P2P 网络借贷理论、相关理论及模型、探索性数据分析、网贷用户违约预测模型、总结与展望。

第一章为绪论。首先从介绍 P2P 网贷平台出现的背景以及信贷违约预测的实践意义与理论意义开始展开，接下来梳理了国内外对于信用风险违约预测的研究，最后说明了本文研究的内容及框架。

第二章为 P2P 网络借贷理论。主要阐述了 P2P 网络借贷的含义、优势和弊端、面临的风险，以及提出了如何控制风险的几点措施。

第三章为相关理论及模型。介绍了非平衡数据处理、WOE 分箱、模型评估指标的理论知识，然后对逻辑回归、随机森林、AdaBoost 算法、GBDT 算法、CatBoost 的算法原理和优缺点进行了介绍，阐述了各种集成策略，为后文的实证分析奠定

基础。

第四章为探索性数据分析。主要概述了本文的数据来源，接着对数据的预处理，其中包括：(1)删除泄露信息的贷后变量、无关变量。(2)计算变量的缺失值比重，对变量缺失超过一定阈值的进行删除，低于该阈值的变量采用均值插补法填充。(3)为了后续的分箱，删除异常值。(4)根据业务知识进行变量衍生，增强自变量的解释能力。然后进行描述性分析，寻找与因变量的相关性。最终对清洗后的数据通过 IV 值筛选、相关系数检验、方差膨胀因子筛选、随机森林变量筛选的方式筛选特征，最终筛选出 11 个变量用于建模。

第五章为网贷用户违约预测模型。这是本文研究的主体部分，对上一章经数据清洗及特征筛选后的数据依据 8:2 的比例划分得到初步的训练集与测试集，由于原始数据集具有高度不平衡性，直接建模容易忽视少数类样本，本文使用 SMOTE 采样法分别对训练集与测试集进行平衡处理。首先构建逻辑回归模型，接下来基于集成学习中的随机森林算法、AdaBoost 算法、GBDT 算法、CatBoost 算法分别建立违约预测模型，后将分类效果较好的模型作为初级学习器，逻辑回归为次级学习分类器搭建 Stacking 融合模型，然后根据评价指标对模型进行对比，找出表现最优的模型。

第六章为总结与展望。对本文总体的研究结果予以总结，在此基础上，归纳出了本文的创新点，也阐明了后续研究需要改进的方向及内容。

本文的重要研究框架如下图所示：

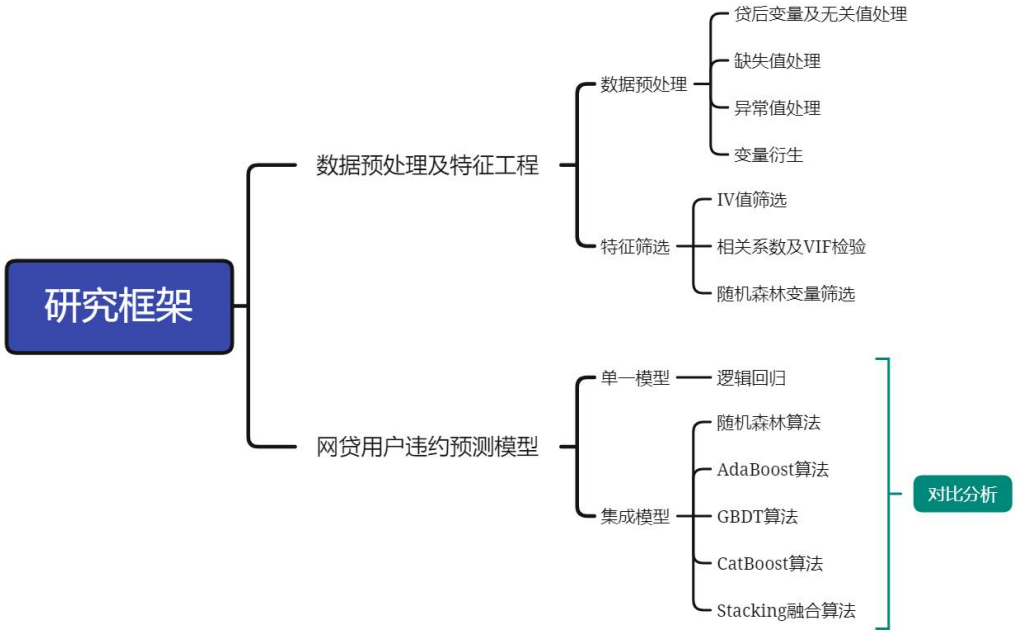

图 1.1 整体研究框架

Fig.1.1 Overall research framework

2 P2P 网络借贷理论

2.1 P2P 网络借贷的含义

P2P 网络借贷即“Peer to Peer”，表示个人对个人，是互联网金融快速发展下的新生业态。它通过互联网平台，同时面向欲将闲置资金进行投资的投资者与缺乏资金且遇到财务难题的融资者，将投资需求与融资需求有效对接，实现资金的运用与融通。从国外来看，英国于 2005 年率先成立 Zopa，美国于 2006 年成立 Prosper，网络借贷开始发展，成为个人临时资金中转的重要渠道，此后 Lending Club 平台较 Prosper 晚一年创立，却发展为当地最大的网贷平台，在目前也不断成长为全球最大的 P2P 网贷上市公司。从国内来看，2007 年上海建立了我国首家 P2P 网络借贷平台——“拍拍贷”，随后 P2P 网络借贷模式也在我国逐渐发展起来。与传统金融机构的借贷相比，P2P 网贷平台依托互联网运营，经营成本更低，因此利率较低，深受融资者的偏爱。

P2P 网络借贷的优势体现在以下几个方面：一、拓宽市场投融资渠道，有效提高社会民众闲散资金的有效利用率。对于有闲置资金的民众来说，相比于买国债、理财产品、炒股等投资渠道，P2P 平台无疑是一个更便捷、收益更高的投资平台。二、降低交易成本，满足个性化需求。用户可根据自身的资金需求，灵活地选择贷款期限。三、融资主体的门槛降低。相比于传统机构对贷款用户的严苛申请条件，P2P 平台要求较低，很大程度上地破解了中小微企业融资难的困境，满足了中小微企业迫切的资金需求，同时融资成本也大幅降低，推动了普惠金融的健康发展。四、操作便捷，效率高。用户足不出户，便可在网上进行投资及融资，审批手续较少，申请时间较短，节约了时间成本，大大提高了办事效率，具有显著的优势。五、推动经济发展。根据鲶鱼效应，P2P 平台如同鲶鱼，冲击着传统金融机构的这群沙丁鱼，凭借着其高效便捷及门槛低的显著优势，倒逼着传统金融机构加强科技创新，为客户提供更优质的服务，在新竞争者、新模式的竞相出现下愈发高效与活跃，从而推动了整个金融体系的快速高质发展。

但 P2P 网络借贷平台也面临着很多弊端，如征信体系不完善、信用评估方法不完备、信息不对称、缺乏有效监管。这些问题也严重制约了 P2P 平台的发展，使其面临着严峻的态势。

2.2 P2P 网络借贷面临的风险

① 信用风险

信用风险指由于还款意愿的变动或者不可抗力因素引起的还款能力的下降，导

致借款人无法按期履约甚至发生违约，从而使平台遭受损失的风险。信用风险主要是由于缺乏对借款人的统一评估标准、信息不对称、界限不清等，导致 P2P 平台无法充分了解借款人的还款意愿以及目前真实的财务状况。

更严重地，体现在网络诈骗上。一方面，网贷平台参与非法集资，诈骗投资者的资金，首先通过虚假高利润诱惑以及采用资金池包装理财产品出售两种方式获得资金，之后卷取大量钱财抽身；另一方面，借款人恶意欺诈，起初就没有还款意愿，虚假捏造个人信息进行骗贷，当前互联网技术的发展导致诈骗成本随之降低，信息的可复制性及传播性进一步加强，该风险不断增大。

② 监管风险

监管风险指由于相关法律体系的不完善、滞后性引起 P2P 平台存在损失的风险。为规范 P2P 网贷平台，央行对 P2P 网贷平台的规定，从“四条红线”到“十项监管原则”，再到新七条[26]，这些规定从本质上仍然没有划分明确的界限，如 P2P 平台的准入门槛、经营方式、非法集资的判断标准等，很多问题还没有得到有效解决，这些问题可能导致不良用户钻法律的漏洞，从中谋取不当利益。

③ 流动性风险

流动性风险指由于 P2P 平台对流动性计划的不科学合理分配，造成无法满足投资者从账户正常提取资金甚至本金的需求。在借贷过程中，投资者偏爱周期短、高收益的理财产品，而融资者渴望周期长、低成本的借贷资金，这两者的差异引起了资金在时间上、数量上的错位，使 P2P 平台的预期损失减少。此外，民众的投资意识保守、监管不完善等原因也难以保证获得长期稳定的资金。当 P2P 平台现金流紧张甚至资金链有断裂风险时，风险规避型的投资者也会纷纷出售自身的产品，若平台没有足够强大的风控体系，可能会引起破产[27]。

2.3 P2P 网络借贷对风险的控制

① 加强对 P2P 平台的监管

监管部门指定统一的标准、完善法律法规，包括以下方面：1) 合理制定 P2P 平台公司准入标准，严格进行资质审查，评估风控能力，将有问题、不合规的平台放到黑名单。2) 审查 P2P 资金来源，以防出现非法集资。3) 规范 P2P 放贷行为，确定利率的波动空间，打击高利贷，同时确保借款利息不得预先扣除。4) 完善 P2P 贷款回收，创新贷款回收方式，禁止暴力催收[28]。

② 提高信息披露水平，促进信息公开

我国 P2P 平台起步较晚、中间遇到很多发展难题，收集到的用户信息及历史交易数据不完善，因此需要对于企业和个人建立系统、公开的征信体系，进一步完善信用风险评估体系，提高透明度，各平台将黑名单进行公开。加强对借款人申

请信息的真实性、有效性的审核，对有风险的客户持续跟踪，保证按期还款。此外，也要注重信息的保护，以防信息泄露，给用户带来不良影响。

③ 加强风险控制能力

确定统一的借款人评价指标，建立合理的评估模型，对借款人的资质、收入来源、借款目的等综合评估，提高平台的风险判别能力。同时，对交易过程进行实时监测，当借款人出现还款危机时，及时预警，采用合理方案，如冻结资产、追加担保等方式，降低违约事件发生的频率，控制平台损失。

④ 合理把控流动性风险

面对资金流紧张的情况，需要提高应对风险的能力，可采取以下措施。一是创新金融产品，科学调整资金结构。将长期的金融产品适当切割为短期的金融产品，将大额的金融产品适当切割为小额的金融产品，同时鼓励引导投资者适当延长投资期限。二是建立完善全方位的预警机制，对投资者的主体类型、风险偏好程度、资金来源渠道等全面分析，对未来撤出资金的时间、金额进行科学预判，从而能够及时的转移风险，提前应对风险、合理规划，控制住流动性风险。

3 相关理论及模型

3.1 相关理论

3.1.1 非平衡数据的处理

非平衡数据指不同类别的样本数量差别特别大，其中多数类的训练样本数量过于多，少数类的训练样本数量过于少，数据分布十分不均衡。对此类数据构建分类器时，分类器不能平等地学习各个类别的信息，少数样本由于数量少且包含样本信息少，致使传统分类器的预测结果更倾向多数类，忽略了少数类。然而在反欺诈、疾病诊断、保险索赔等领域，少数类是重要的研究对象，具有更大的应用价值。

对于非平衡数据，处理方法包括样本处理及算法处理。算法处理通过代价敏感算法及集成分类算法，即改进原有的分类器算法[29]。这里主要介绍样本处理方法，其大致分为过采样、欠采样、SMOTE 采样。这些方法各有利弊，其中，过采样指增加少数类样本，简单的方法是基于少数类样本进行简单地随机复制来弥补少数类样本，但容易导致模糊边界、过拟合。欠采样指减少多数类样本，这可能导致丢失重要信息。SMOTE 采样是这两种方式的统一结合，基于 KNN 算法，通过线性插值的方式来人工不断生成新样本，既一定程度上减少了过拟合，又保留了原始样本的信息，增强了泛化能力。

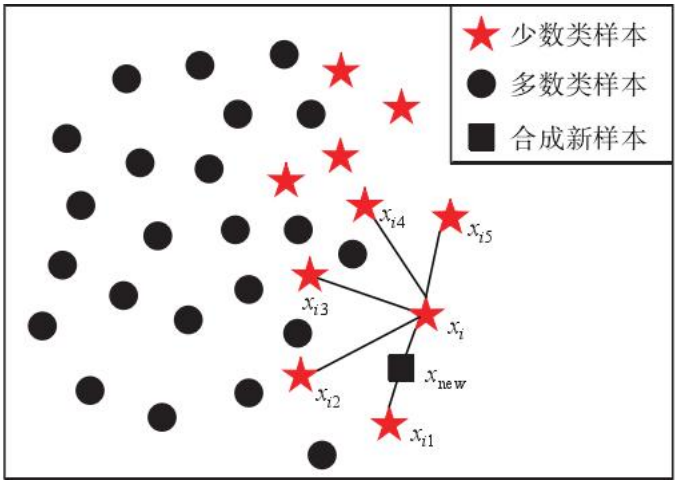

图 3.1 SMOTE 采样示例图

Fig.3.1 Smote sampling example diagram

根据图 3.1，SMOTE 采样的基本步骤[30]如下：

① 依次选取少数类样本 x_i，作为新样本的根样本；

② 采用欧式距离法，找到x_i的 K 近邻，并依次随机选取 N 个少数类样本，作为产生下一个新样本的辅助样本；

③ 基于线性插值的方法，在采集得到的样本x_i与辅助样本间来不断生成新的少数类样本。

$$x_{new} = x_i + rand(0,1) * \left| x_i - y_j \right|, j = 1,2,3\ldots\ldots N \tag{3.1}$$

$rand(0,1)$表示(0,1)之间的随机数，x_{new}表示生成的新样本，y_j表示随机选取的一个 K 近邻样本。

3.1.2 WOE 分箱及 IV 值

分箱指对连续变量进行离散化处理，以及对取值稀疏的离散变量进行合并处理，以组间差距尽可能大、组内差距尽可能小的原则进行分箱，能够避免异常值的扰动，增强模型稳定性，从而提高了变量的预测能力及解释能力。分箱方法包括无监督分箱及有监督分箱，前一种方法使用更多。无监督分箱一般包括聚类分箱、等频分箱与等距分箱，对因变量的解释能力提高的程度有限。有监督的分箱根据采用优化函数的不同，可分为 Chi-merge 卡方分箱、Best-KS 分箱、最优 IV 分箱等[31]。本文主要采用卡方分箱进行变量筛选。

卡方分箱属于自底向上的分箱方法，首先初始化将数据划分多个区间，基于计算出的卡方值对相邻区间进行合并。卡方值取值越小，说明相邻区间的类分布越相似，则将区间合并，否则分开。分箱后，对各个变量的各箱进行 WOE 映射，同时计算变量的 IV 值，可以进一步衡量变量对标签的预测能力。

表 3.1 某变量的 WOE 值

Table 3.1 Woe value of a variable

	Good	Bad	Good Percent	Bad Percent
Group 1	G_1	B_1	G_1/G_{Total}	B_1/B_{Total}
Group 2	G_2	B_2	G_2/G_{Total}	B_2/B_{Total}
……	……	……	……	……
Group N	G_N	B_N	G_N/G_{Total}	B_N/B_{Total}
Total	$G_{Total} = \sum G_i$	$B_{Total} = \sum B_i$		

表 3.1[32]表示计算某变量的各组好坏样本占总体的比重，则当前变量的每个分组的WOE_i值为

$$WOE_{\mathrm{i}} = -\ln\frac{P_{good}}{P_{bad}} = -\ln\frac{G_i / G_{Total}}{B_i / B_{Total}} \qquad i = 1,2\ldots\ldots N \tag{3.2}$$

根据上述公式，可以看到，WOE 表示的是“当前分组中违约客户数量占样本涵盖的所有违约客户数量的比例”和“当前分组中未违约客户数量占样本涵盖的所有未违约客户的比例”的差异，该差异通过计算这两个比例的比值，后取对数表示。WOE 越大，表示当前分组的违约客户越多，则该分组内样本发生违约可能性越大，WOE 可被概括为自变量的取值对因变量的影响。IV 值体现自变量对于解释因变量所拥有的信息总量，可以判断出自变量对于标签的预测能力，从而可根据该指标进行变量筛选。IV 取值越高，说明自变量的预测能力越强，计算某变量的每个分组的 IV 值的公式如下：

$$IV_i = (P_{bad} - P_{good}) * WOE_i = (\frac{B_i}{B_{Total}} - \frac{G_i}{G_{Total}}) * \ln\frac{B_i / B_{Total}}{G_i / G_{Total}} \qquad i = 1,2\ldots\ldots N \tag{3.3}$$

则该变量所拥有的信息量为

$$IV = \sum_{i=1}^{N} IV_i \tag{3.4}$$

如下表 3.2 给出了 IV 的取值与预测能力的关系：

表 3.2 IV 值的预测能力

Table 3.2 Prediction ability of IV value

IV 值	预测能力
IV<0.02	无预测能力
0.02~0.10	低
0.10~0.30	中
IV>0.30	高

3.1.3 评估指标

① 准确率、召回率、精确率及 F1 值

在二分类的预测问题中，实例可划分为正类(T)、负类(P)两类。基于真实类别与预测类别的差异可划分为四种组合：

1) 真正率（TP）：实际为正例且被预测为正例
2) 假负类（FN）：实际为正例但被预测为负例
3) 假正类（FP）：实际为负例但被预测为正例
4) 真负类（TN）：实际为负例且被预测为负例

根据分类结果得到的混淆矩阵如下表：

表 3.3 混淆矩阵

Table 3.3 Confusion matrix

真实\预测	正例	负例
正例	TP	FN
负例	FP	TN

进一步可以计算出以下指标：

1) 准确率（Accuracy）：预测正确的样本数与总样本数之比

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{3.5}$$

2) 精确率（Precision）：在预测类别为正类的样本中，被正确分类的样本占比

$$\frac{TP}{TP+FP} \tag{3.6}$$

3) 召回率（Recall）：在真实类别为正类的样本中，被正确分类的样本占比

$$\frac{TP}{TP+FN} \tag{3.7}$$

4) F1 值：精确率与召回率的调和平均值，衡量模型整体预测能力

$$\frac{2*\mathrm{Pr}ecision*\mathrm{Re}call}{\mathrm{Pr}ecision+\mathrm{Re}call} \tag{3.8}$$

② ROC 曲线及 AUC 值

ROC 曲线以真正率（TPR）为横轴，以假正率（FPR）为纵轴，在不同的判断阈值下绘制而成的曲线。TPR 表示预测类别为正类且真实类别为正类的样本与所有真实正类样本的比值，FPR 表示预测类别为正类但真实类别为负类的样本与所有真实负类样本的比值。

$$TPR=\frac{TP}{TP+FN} \tag{3.9}$$

$$FPR=\frac{FP}{TN+FP} \tag{3.10}$$

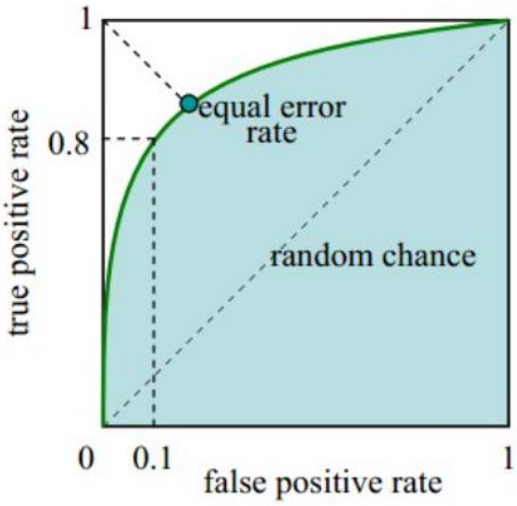

图 3.2 ROC 曲线示例图

Fig.3.2 ROC curve example

如上图 3.2 为 ROC 曲线的示例图，曲线上的各个点分别代表不同的阈值，代表一对不同的 FPR 和 TPR。为了将分类的能力用具体的数值量化，引入 AUC 值，定义是 ROC 曲线下的覆盖面积，一般取值在 0.5-1 之间，AUC 取值越大，则曲线的图像效果越接近坐标左上角，说明模型效果越优于随机预测，分类能力越强。

③ KS 值

KS 曲线，又称洛伦兹曲线，以阈值为横轴，以 TPR 和 FPR 为纵轴。KS 值是 TPR 和 FPR 相差最大的距离，即 KS=max(TPR-FPR)。取值越大，说明越能正确区分正类样本与负类样本，预测效果越好。一般 KS>0.2 时，说明模型预测较好。但它只能体现部分区别最大分段的情况，不能很好地体现整个过程的区分能力。

3.2 相关模型

3.2.1 逻辑回归

逻辑回归模型常在分类问题的预测上被广泛使用，是一种广义线性模型，有别于线性回归输出没有界限的连续数值，逻辑回归输出的是取值在(0,1)之间的概率值，表示因变量属于某种类别的可能性。线性回归的公式为

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots + \theta_n x_n = \theta^T x \tag{3.11}$$

逻辑回归则是将线性回归的结果映射到激活函数 sigmoid 函数中，从而能够将其无界限的数值结果转换为具有统计意义的概率值，sigmoid 函数的公式为

$$y = \frac{1}{1+e^{-z}} \tag{3.12}$$

sigmoid 函数的曲线类似于 S 型，最大值为 1，最小值为 0，以(0,1/2)为对称中

心,在左右两端增长较慢，在中心 0 处增长较快，从而能够较好区分中心附近的数值，将连续变量离散化，应用于 0/1 的分类问题上。

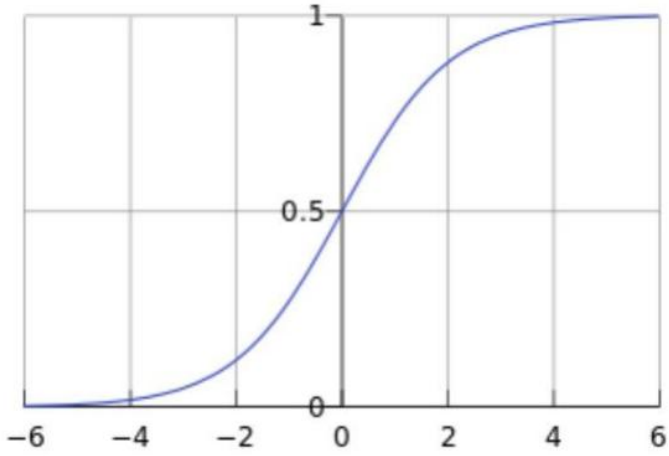

图 3.3 sigmoid 函数曲线

Fig.3.3 Sigmoid function curve

将上述公式(3.11)代入到公式(3.12)可得， $y=\frac{1}{1+e^{-\theta^T x}}$ ，可转化为 $\ln\frac{y}{1-y}=\theta^T x$，该模型称为逻辑回归模型。此外，基于极大似然函数思想建立方程，来进一步对未知参数向量 $[\theta^0,\theta^1,\theta^2\ldots\ldots\theta^n]^T$ 进行估计，采用梯度下降法、牛顿迭代法等求得最优解。

与此同时，将 y 视为样本为正类 1 的概率，则 1-y 为负类 0 的概率，称两者比值 $\frac{y}{1-y}$ 为几率，对数化后的比值 $\ln\frac{y}{1-y}$ 为对数几率，因此输出 y=1 的对数几率可以通过输入 x 的线性函数[33]解释。同时可以得到 y=1 及 y=0 的概率，计算公式为

$$p(y=1\mid x)=\frac{e^{\theta x}}{1+e^{\theta x}} \tag{3.13}$$

$$p(y=0\mid x)=\frac{1}{1+e^{\theta x}} \tag{3.14}$$

① 逻辑回归的优点：

1) 前提条件简单，不用事先对数据分布假设，故几乎不会引起由于假设分布错误导致预测结果较差的问题；

2) 以(0,1)间的概率输出，具有概率意义，可以根据实际设置阈值，辅助科学决策；

3) 在预测分类问题中，精确性和稳定性较高。

② 逻辑回归的缺点：

当变量过多时，易引起多重共线性现象，从而导致对测试集的分类预测能力

远低于训练集的预测能力，即引起过拟合现象，解决办法包括最优变量子集选择、向前选择法、向后选择法、逐步回归法。数据需满足线性可分，遇到非线性可分，则需要进行转换或采用其他模型。

3.2.2 集成策略

集成学习指综合并学习多个基学习器的预测结果，旨在提升预测效果、增强泛化能力，根据学习的基学习器是否同质，可概括为同质集成和异质集成，同质集成包括没有依赖关系、可并行生成的 bagging 算法，以及有依赖关系、必须串行生成的 boosting 算法，异质集成为 Stacking 融合算法及 Blending 融合算法。

Bagging 算法原理是对样本数据使用 Bootstrap 自助采样法得到若干个几乎不一致的采样集，之后用采样集分别对基分类器训练，最终根据组合策略构建的集成分类器，在分类问题上进行投票，在回归问题上计算均值。由于 Bagging 算法中基分类器通过在不完全相同的采样集上学习会引起差异化，从而泛化能力强，故在降低分类器模型拟合的方差上表现较好。

Boosting 算法原理是用弱分类器集成为强分类器，首先初始化赋予平等权重，后根据每轮对样本的分类预测结果，调整各个样本的权重，使错分类的样本在下一轮学习器的学习中能够引起更多关注，最后对若干个弱分类器依据误差率加权组合得到一个强分类器。由于 Boosting 在持续提高误差率小的基分类器的权重、降低误差率大的基分类器的权重，故在降低模型的整体误差上表现优秀。

这两种算法的对比[34]如下图：

表 3.4 Bagging 算法及 Boosting 算法对比

Table 3.4 Comparison between Bagging algorithm and Boosting algorithm

算法	样本选择	样本权重	计算方式
Bagging	Bootstrap 采样，每轮训练集独立	相等	可并行
Boosting	每轮训练集不变	根据错误率调整，错误率越大则权重越大	必须串行

Stacking 算法作为一种分层模型集成框架，可以使用不同质的学习器组成。Stacking 的优秀学习效果并非由于采用了多层堆叠，而是有效地结合不同种类的学习器对不同变量的学习能力差异，为减少过拟合，通常设置为两层模型。第一层通过多个基学习器对原始训练集进行训练，第二层对第一层得到的预测结果进行再训练，构建 Stacking 模型。为防止过拟合，并不是直接采用基学习器的训练集来作为下一层的训练集，一般采用交叉验证法，用验证集即第一层学习器未使用的样本来构成第二层的训练样本。因此，由于 Stacking 算法采用交叉验证，故稳

健性更强，同时准确率较高，是目前提升机器学习效果最有效率的方法。

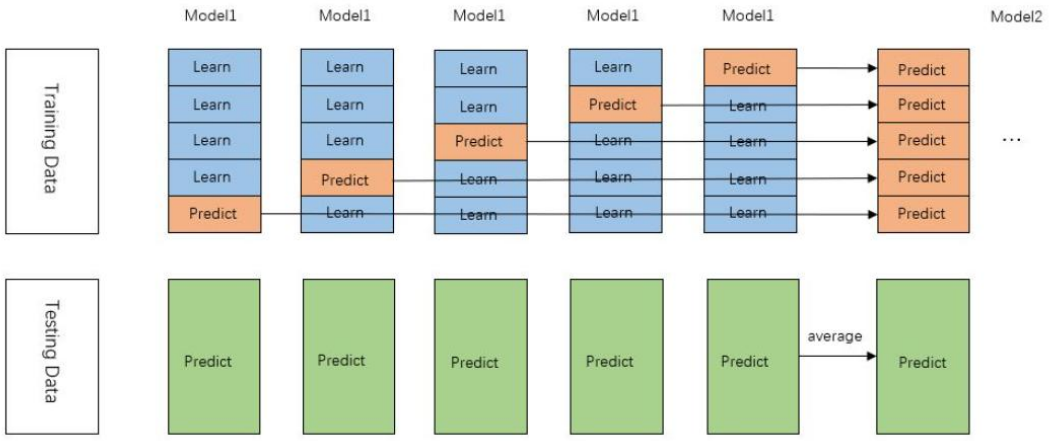

图 3.4 Stacking 算法过程

Fig.3.4 Stacking algorithm process

上述图 3.4 是五折交叉检验的 Stacking 算法过程，以两层为例，具体步骤如下：

① 将原始数据集按一定比例划分为训练集与测试集，并通过五折交叉检验划分训练集，假设有 n 个基模型;

② 对于其中的一个基模型，记训练集划分后的数据分别为V_1,V_2,V_3,V_4,V_5，依次将其中的一折作为验证集，同时基于其余的四折数据作为训练集来训练模型，对验证集、原始的测试集进行预测，将五次的验证集结果按顺

③ 序组合为一个矩阵，记为$train_1$，将五次的测试集结果求平均，记得到的矩阵为$test_1$;

④ 对其余的基模型重复上述的过程，得到其余模型的验证集结果$train_2,\cdots\cdots train_n$，及测试集$test_2,\cdots\cdots test_n$;

⑤ 将所有基模型的验证集结果$train_1,\cdots\cdots train_n$拼接为一个矩阵，作为新的训练集；多个测试集结果$test_1,\cdots\cdots test_n$同样拼接为一个矩阵，作为新的测试集。最终使用第二层学习器对新的训练集和测试集进行再训练。

Blending 算法同样作为一种模型融合算法，大致相同，不同的是它没有如同 Stacking 算法采用 K-Fold 的 CV 方法得到预测值，而是建立 Holdout 集，将 K-Fold CV 替换为 Holdout CV。简单地来说，Stacking 算法的各个基模型使用全部训练集数据进行训练，而 Blending 算法的各个基模型采用训练集互不相交的子集进行训练。

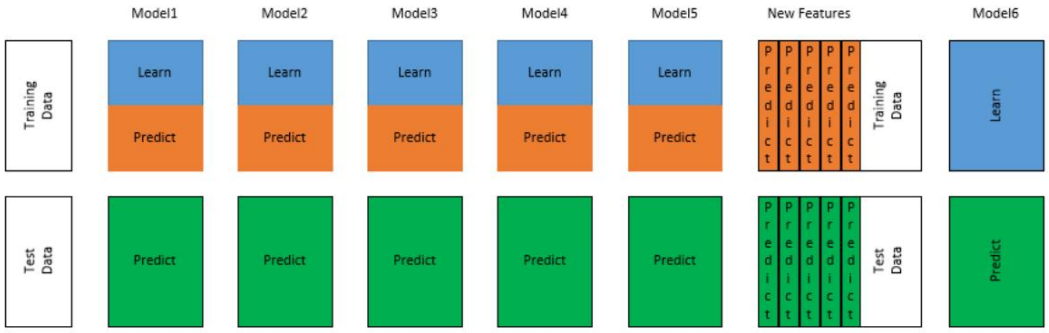

图 3.5 Blending 算法过程

Fig.3.5 Blending algorithm process

上述图 3.5 是 Blending 算法过程，同样以两层为例，具体步骤如下：

① 将原始数据集按一定比例划分为训练集与测试集，再对训练集一般按 7:3 的比例划分为训练集V_1与验证集V_2;

② 以其中的一个基模型为例，使用训练集V_1训练出的模型，对验证集V_2来预测，其结果生成新的训练集；同时对原始数据集中的测试集上来预测，其结果生成新的测试集;

③ 对多个基模型重复上述过程，将所有基模型的训练集与测试集分别拼接得到第二层的训练集与测试集来再训练。

Stacking 算法与 Blending 算法的对比：

① Stacking 采用 K 折交叉验证对全部训练集数据训练，Blending 采用留出法训练了部分数据，因此 Stacking 降低了过拟合，稳健性更强;

② 不过，由于没有使用交叉验证，Blending 算法更简单，在运行时间上具有显著优势;

③ 在进行模型融合时，Stacking 算法将第一层基分类器的分类预测结果作为训练集输入到第二层，为非线性融合，Blending 算法则对于第一层基分类器可以按照一定的权重得到预测结果输入到下一层，为线性融合。

3.2.3 决策树

决策树指基于树的模型进行多分类决策，包含根节点、内部节点及叶节点，其中根节点涵盖样本全集，内部结点用于通过属性测试将样本划分到子节点，叶节点表示最终所属的类别。

① 划分特征

决策树学习的关键是如何筛选出最优特征对节点分裂，实现其分支节点所涵盖的样本尽可能多地属于一个类别的目标，即分裂后的节点“纯度”最高。常用的分裂标准包括信息增益、信息增益比和基尼指数。根据这三种指标，分别形成了

三种不同的决策树算法，分别是 ID3 算法、C4.5 算法、CART 算法。

1) 信息增益

首先引入信息熵，假设样本数据集合为 D，每个样本包含 n 个类别，则 D 的信息熵为

$$Ent(D) = -\sum_{k=1}^{n} p_k \log_2(p_k) = -\sum_{k=1}^{n} \frac{|D_k|}{|D|} \log_2(\frac{|D_k|}{|D|}) \tag{3.15}$$

p_k 表示第 k 类样本在样本中所占的比例，$|D_k|$ 表示属于第 k 类的样本数，$|D|$ 表示样本总数。信息熵用来展示样本整体的不确定性程度，取值越大，则不确定性越高。基于此计算出的信息增益为

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{3.16}$$

其中，a 表示为一离散特征，具备 V 个可能性取值 $\{a^1, a^2, \ldots\ldots a^V\}$。$D^V$ 则表示在特征 a 中取值为 a^V 的样本。信息增益表示基于特征 a 进行分裂所得到的纯度提升，取值越大，则纯度提升的越多。

2) 信息增益比

由于多取值的变量易导致信息增益较高，为缓解该问题，提出信息增益比这一准则。

$$Gain_{ratio}(D,a) = \frac{Gain(D,a)}{IV(a)} \tag{3.17}$$

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \tag{3.18}$$

$IV(a)$ 取值越大，则说明 a 的可能取值种类数越多，一定程度上削弱了多取值变量的影响。

3) 基尼指数

$$Gini(D) = \sum_{k=1}^{n} \sum_{t \neq k} p_t p_k = 1 - \sum_{k=1}^{n} p_k^2 \tag{3.19}$$

基尼指数表示随机抽取的两个样本，它们的所属类别不相同的概率，反映了数据集 D 的纯度大小，取值越小，则说明样本的纯度越高。

② 决策树生成

从根节点出发，由上至下依据分裂准则递归地划分子集，直到剩余样本同属于一个类别或者没有剩余特征，则递归停止，产生叶节点。

③ 决策树剪枝

为在保持一定的分类准确率下，减少过拟合、降低模型复杂度，往往有必要剪枝。主要可分为预剪枝和后剪枝两种方式，前者一边递归生成决策树，一边判断

是否需要剪枝，可选取阈值判断是否有必要进一步划分节点；后者在生成好的决策树上采取剪枝，计算测试数据样本在不同候选决策树上的分类正确率，以预测错误率尽可能小的原则选取。

3.2.4 随机森林

随机森林算法由 Breiman[35]于 2001 年提出，是 Bagging 框架下的更加关注决策树的一种集成学习算法，既可以解决回归问题，也可以解决分类问题。随机森林算法与 bagging 算法同属于 bagging 框架，不同之处在于：随机森林的基模型为 CART 决策树；同时在 Bootstrap 自助采样的基础上，又进行了特征采样。此外，由于训练样本与特征筛选的双随机性，克服了过拟合问题，也减少了由于异常数据或者噪声数据存在而带来的波动，故无需进行剪枝。假设训练集共有 n 个，每个样本包含 M 个特征，则随机森林的大致构建步骤如下：

① 采用 Bootstrap 采样法有放回地抽取 n 个样本，采样 k 次，得到 k 个训练样本；

② 对于每个训练样本，筛选 m 个特征，共训练出 k 个决策树；

③ 将生成的 k 个决策树进行组合，产生得到随机森林，将每一个决策树的预测结果作为一次投票，将汇总后投票数最多的类别作为随机森林预测出的最终所属类别。

随机森林算法流程示意图如下：

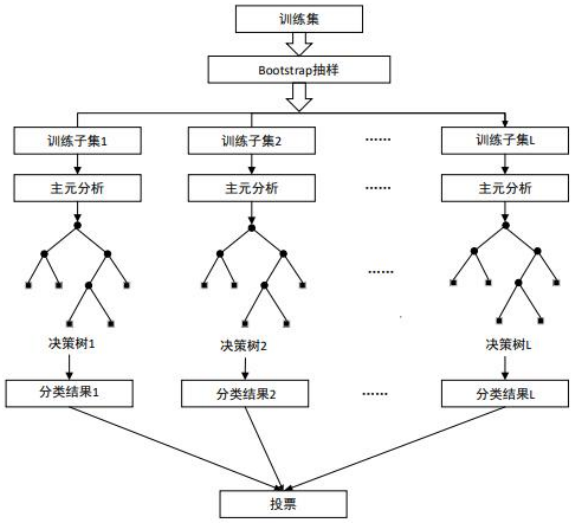

图 3.6 随机森林算法流程

Fig.3.6 Random Forest algorithm steps

随机森林也常用于变量重要性排序。将进行 Bootstrap 自助采样后从来没有被抽中的样本称为包外样本(OOB)，这些样本约占初始训练集的 36.8%。使用包外样

本来测试样本数据中变量的性能。当一个对因变量预测结果非常重要的变量发生异常、有污染时，则会引起特别大的预测误差，大幅降低预测准确率。基于这个思想，将对某特征进行随机扰动后计算出来的包外数据的预测准确率与原始包外数据的预测准确率的减少率[36]，作为该特征的重要性，由此对变量进行排序。可以设定相应的阈值，如果变量的重要性低于该阈值，则有理由将该变量删除。

① 随机森林的优点：

1) 能处理高维数据，不用事先做特征选择；

2) 依靠自身内部独立的决策树高度并行化运行，分类效率较高，时间复杂度较低，尤其对于大样本有显著的速度优势；

3) 能够对特征进行重要性排序；

4) 对于缺失数据不敏感，能容忍不过于大的噪声及异常值。

② 随机森林的缺点：

1) 当变量取值较多，对随机森林模型的影响较大；

2) 噪声异常大时，易导致过拟合，降低预测性能。

3.2.5 AdaBoost

Adaboost 算法是 Boosting 框架下的一种代表性集成算法。它的原理是首先将样本初始化赋予同等权重，训练出一个弱学习器，依据分类误差率，不断提高被错分样本的权重，基于调整好权重的样本来训练下一层弱学习器，直到学习的弱学习器数量达到事先预定的 T 值，则将这 T 个学习器加权组合为一个强学习器。学习弱学习器及更新权重的具体流程及示意图如下：

假设训练集为$T=\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_m,y_m)\}$，其中$x_i$表示实例，$y_i$={-1，+1}表示标签。

第一步：初始化权重

$$D_1=(w_{11},w_{12},\cdots\cdots，w_{1m}),w_{1i}=\frac{1}{m},i=1,2,\cdots\cdots;m$$

第二步：对于 k=1,2……,K

① 对样本权重分布D_k的样本，学习第 k 个弱分类器$G_k(x)$，并计算分类误差率

$$e_k=P(G_k(x_i)\neq y_i)=\sum_{i=1}^{m}w_{mi}I(G_k(x_i)\neq y_i) \tag{3.20}$$

② 计算第 k 个弱分类器$G_k(x)$的权重

$$\alpha_k=\frac{1}{2}\log\frac{1-e_k}{e_k} \tag{3.21}$$

③ 更新训练样本的权重

$$D_{k+1}=(w_{k+1,1},w_{k+1,2},\cdots\cdots，w_{k+1,m})$$

$$w_{m+1,i}=\frac{w_{mi}*\exp(-\alpha_k y_i G_m(x_i))}{\sum_{i=1}^{m} w_{mi}*\exp(-\alpha_k y_i G_m(x_i))} \tag{3.22}$$

第三步：最终得到强分类器

$$G(x)=sign(\sum_{k=1}^{K}\alpha_k G_k(x)) \tag{3.23}$$

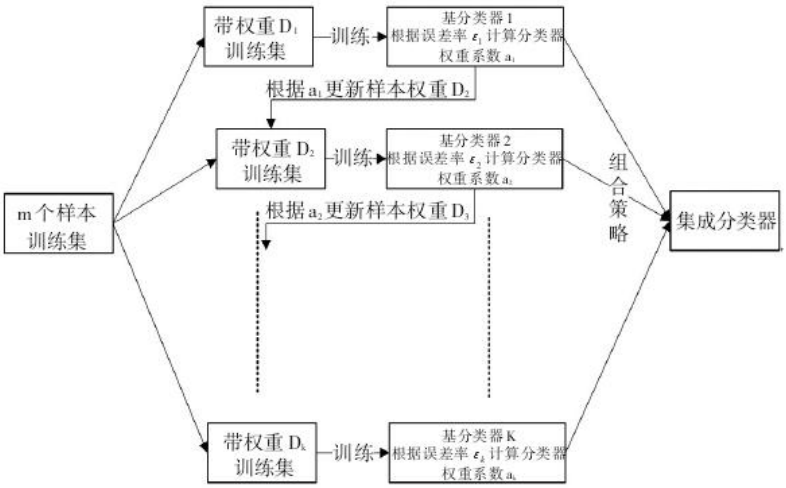

图 3.7 AdaBoost 算法流程

Fig.3.7 AdaBoost algorithm steps

① Adaboost 算法的优点：

1) 可以基于多种分类模型构建弱分类器，使用灵活；

2) 分类精度较高；

3) 不易引起过拟合。

② Adaboost 算法的缺点：

1) 运行时间较长；

2) 需事先设定的弱学习器的数量不好确定；

3) 对异常样本数据敏感，可能影响较大。

3.2.6 GBDT

GBDT 算法全称为梯度提升树，是基于 CART 决策树的 Boosting 框架下的另外一种集成学习算法。与 Adaboost 算法不同的是，GBDT 模型直接采用当前决策树损失函数的负梯度作为残差的近似值来拟合新的决策树，其核心便是在于采用最速下降的近似方法，即让损失函数沿梯度方向下降，尽快地使损失函数不断降低。由于拟合的是梯度值，为连续数据，故基模型是 CART 回归树。首先介绍回

归问题的提升树算法。

假设训练集$T=\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_N,y_N)\},x_i\in X\subseteq R^n,y_i\in Y\subseteq R$

第一步：初始化 $f_0(x)=0$

第二步：对于$m=1,2,\cdots\cdots,M$

① 计算残差

$$r_{mi}=y_i-f_{m-1}(x_i),i=1,2,\cdots\cdots,N \tag{3.24}$$

② 拟合残差学习回归树，可得$T(x;\Theta_m)$

③ 更新

$$f_m(x)=f_{m-1}(x)+T(x;\Theta_m) \tag{3.25}$$

第三步：得到回归问题的提升树

$$f_M(x)=\sum_{m=1}^{M}T(x;\Theta_m) \tag{3.26}$$

注：对于模型$f_{m-1}(x)$，需求解第 m 棵树的参数$\hat{\Theta}_m$

$$\hat{\Theta}_m=\arg\min_{\Theta^m}\sum_{i=1}^{N}L(y_i,f_{m-1}(x_i)+T(x_i;\Theta_m)) \tag{3.27}$$

将常用的平方误差函数应用于损失函数时，由

$$L(y,f(x))=(y-f(x))^2 \tag{3.28}$$

可得损失函数：

$$L(y,f_{m-1}(x)+T(x;\Theta_m))=[y-f_{m-1}(x)-T(x;\Theta_m)]^2=[r-T(x;\Theta_m)]^2 \tag{3.29}$$

其中$r=y-f_{m-1}(x)$表示$f_{m-1}(x)$模型拟合数据的残差

提升树算法采用向前算法及加法模型来不断实现学习的优化，但它也面临着问题，若采用其他的损失函数，后期很难优化，因此，Freidman[37]针对该难题提出了梯度提升算法，使用负梯度来近似求解残差。具体流程如下：

第一步：初始化，首先求解令损失函数达到最小的常数值

$$f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c) \tag{3.30}$$

第二步：对于每个模型 m=1,2,……,M

① 对于每个样本 i=1,2,……,N，求解损失函数的负梯度值来近似估计残差

$$r_{mi}=-[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \tag{3.31}$$

② 根据残差r_{mi}拟合回归树，作为新的回归树$f_m(x)$，生成的叶子结点区域为R_{mj}，j=1,2,……,J 表示叶节点的数量

③ 对于每个叶节点 j=1,2,……,J，求解最优拟合值

$$r_{mj}=\arg\min_{c}\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x_i)+r) \tag{3.32}$$

④ 更新回归树

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}r_{\mathrm{mj}}I(x\in R_{mj}) \tag{3.33}$$

第三步：输出最终回归树

$$f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}r_{mj}I(x\in R_{mj}) \tag{3.34}$$

因此，根据以上步骤可知，对于每个输入的训练样本，首先会输出一个初始值F_0，后通过遍历各个决策树计算出各自的预测值 T，从而不断修正原始值，将 M 个决策树的预测值依次累加后输出最终的模型预测结果[38]，同时采用加法模型，将 M 个决策树的线性组合构成的分类器称为一个强分类器，具体的流程图如下：

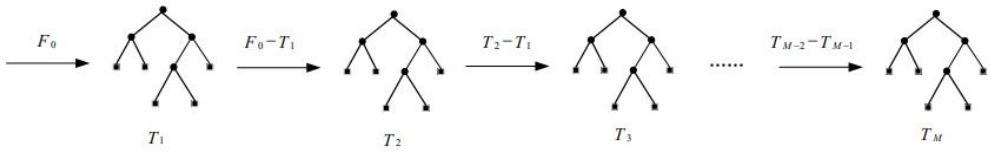

图 3.8 GBDT 算法流程

Fig.3.8 GBDT algorithm steps

① GBDT 算法的优点：

1) 对不同类型的数据均可灵活、有效地处理，如离散数据和连续数据；

2) 对于特征空间存在的异常值，鲁棒性强、稳健度高。同时可以使用更平滑的损失函数来提高模型的稳健程度，如 Huber 损失函数；

3) 调参时间短，且预测的准确率也较高。

② GBDT 算法的缺点：

弱学习器存在较强的依赖关系，难以对样本数据进行并行操作，故面临大量数据时过于耗时。该解决方法是通过自采样的 SGBT 实现部分并行，降低运行时间。

3.2.7 CatBoost

CatBoost 是 Categorical 与 GBDT 的组合，即支持类别特征的 GBDT 算法。GBDT 算法框架下有 CatBoost、XGBoost、LightGBM 三个代表性算法。XGBoost 常用于工业生产中，LightGBM 显著提高 GBDT 的算法效率，CatBoost 的算法准确性一般优于 XGBoost 与 LightGBM。CatBoost 算法不仅准确性高，也可以处理各种类型的数据，如文本、音频、图像。在训练模型时无需大量的数据、调参则可以得到较好的结果。此外，克服了预测偏移和梯度偏差两个问题，一定程度上避免了过拟合，泛化能力得到提高。

① 将类别特征转换为数值特征

处理类别特征的一般方法是采用训练数据集的平均标签值来进行替换样本数据，根据标签平均值对节点分裂，该方式称为 Greedy TS。假设观测数据集 $S=\{(\mathrm{X}_1,\mathrm{Y}_1),(\mathrm{X}_2,\mathrm{Y}_2),\cdots\cdots,(\mathrm{X}_n,\mathrm{Y}_n)\}$，$\mathrm{X}_i$ 为 m 维的特征向量，$Y_i \in R$ 为标签值。则表达式为：

$$x_{ik}=\frac{\sum_{j=1}^{n}[x_{jk}=x_{ik}]\cdot Y_j}{\sum_{j=1}^{n}[x_{jk}=x_{ik}]} \tag{3.35}$$

其中，若 $x_{jk}=x_{ik}$，则$[x_{jk}=x_{ik}]$取值为 1，否则取值为 0。

然而这种方式在遇到训练与测试数据集的分布不一致时，可能出现条件偏移，从而引起过拟合。因此，CatBoost 算法对其进行改进，对数据集随机排列，产生随机序列 $\sigma=(\sigma_1,\cdots\cdots,\sigma_{\mathrm{n}})$，然后计算平均标签值，同时引入先验值 P 及其权重 a，从而减少低频类别带来的噪声，同时降低过拟合，允许使用整个数据集训练。改进后的表达式为[39]：

$$x_{\sigma_p k}=\frac{\sum_{j=1}^{p-1}[x_{\sigma_j k}=x_{\sigma_p k}]\cdot Y_{\sigma_j}+\alpha\cdot P}{\sum_{j=1}^{p-1}[x_{\sigma_j k}=x_{\sigma_p k}]+\alpha} \tag{3.36}$$

② 特征组合

在特征组合中，几个分类特征经任何组合可产生新特征，然而组合的数量随分类特征的数量呈指数型增长，且无法考虑到所有特征。CatBoost 采用贪婪的方式进行组合，构建拆分。第一次拆分，不进行组合；之后的拆分，将当前树的分类特征和组合与训练数据集的全部分类特征相结合，且将组合值转换为数值。

③ 计算叶子节点值——对称树

采用对称树的方式降低在上式直接同时计算多个样本数据集排列导致的过拟合。

CatBoost 算法的优缺点

① CatBoost 算法的优点:

1) 分类精度高，具有 python 接口及 R 接口，易于使用；

2) 鲁棒性强，不用进行过多的超参数寻优，不易过拟合；

3) 可以处理各种类型数据，建模前不用处理类别特征。

② CatBoost 算法的缺点:

在处理分类特征时，需要消耗的内存和时间较多。

4 探索性数据分析

4.1 原始数据介绍

Lending Club 于 2007 年成立，作为面向市场提供网络贷款、进行资金融通的平台中介，于 2014 年 12 月 12 日在纽交所上市。它利用移动互联网打造了比传统银行更有效率、方便在贷款人和投资人之间自有配置资本的机制，撮合贷款人和投资人的资金交易，已经成长为全球最大的 P2P 平台。因此，本文将 Lending Club 官网的公开数据进行研究分析具有有效性、代表性。

本文的数据来源于 Lending Club 官网的 2019 年第一季度的信用贷款用户数据，共计 115677 条记录，144 个变量，包括贷款金额(loan_amnt)、分期付款金额(installment)、年收入(annual_inc)、债务收入比(dti)等 114 个数值型变量及就业职称(emp_title)、贷款用途(purpose)、验证状态(veritification_status)等 30 个字符型变量。以下表 4.1 对部分变量指标的含义进行说明：

表 4.1 部分变量及含义

Table 4.1 Partial variables and their meanings

变量名称	变量含义	变量类型
loan_amnt	贷款金额	数值
term	贷款周期	字符
int_rate	贷款利率	字符
grade	LC 指定的贷款等级	字符
home_ownership	住房性质	字符
annual_inc	年收入	数值
purpose	贷款用途	字符
dti	债务收入比	数值
delinq_2yrs	过去 2 年逾期 30 天以上次数	数值
delinq_amnt	逾期总金额	数值
inq_last_6mths	近 6 个月征信查询次数	数值
open_acc	信用贷款额度	数值
revol_bal	尚未结清信贷总额	数值
total_acc	总信用额度	数值
tot_cur_bal	所有账户当前余额	数值
acc_open_past_24mths	历史 24 个月的交易量	数值

同时，由于该数据对好坏样本的处理比较模糊，需要重新定义，表 4.2 展示目标变量 loan_status 在数据中的分布情况以及各个状态的含义：

表 4.2 目标变量分布

Table 4.2 Target variable distribution

状态取值	数量	状态含义
Current	110918	正常还款
Fully Paid	3608	完全结清
In Grace Period	327	处于宽限期
Late (16-30 days)	256	逾期 16-30 天
Late (31-120 days)	468	逾期 31-120 天
Charged Off	98	坏账

通过对贷后状态的映射，将正常还款(Current)、完全结清(Fully Paid)定义为好样本，取值为 0；将逾期 16-30 天(Late (16-30 days))、逾期 31-120 天(Late (31-120 days))及坏账(Charged Off)定义为坏样本，取值为 1；将处于宽限期(In Grace Period)定义为不确定样本，取值为 2。为方便后续的处理，仅保留好样本及坏样本的数据。

4.2 数据预处理

4.2.1 贷后变量及无关变量处理

在 Lending Club 数据集中，变量可分为贷前变量和贷后变量，而贷后数据会泄露信息，不利于模型的构建，如 LC 公司信用评估的结果：等级(grade)、次等级(sub_grade)，同时也需要删除无关指标，如工作岗级(emp_title)、用户编号(id)，这些指标应予以剔除。

4.2.2 缺失值处理

在采集数据时，由于部分指标难以获取、被采集人员的刻意隐瞒、操作人员失误等，无可避免地造成了缺失值的存在。根据缺失原因的不同，可分为完全随机缺失、随机缺失和完全非随机缺失三种情况。

① 完全随机缺失：数据的缺失完全随机；

② 随机缺失：数据的缺失不完全随机，与某些变量有关；

③ 完全非随机缺失：数据的缺失与自身变量有关。

大量的缺失值可能会丢失部分有效信息，从而影响模型构建的稳定性，因此需要处理缺失值，提高预测的稳健性。常用的处理方法可分为简单的删除法，基

于已有数据的均值插补法和拟合缺失值法。以下表 4.3 可以看到部分变量的缺失情况。

表 4.3 部分变量的缺失情况

Table 4.3 Absence of partial variables

变量名称	缺失数量	缺失比率
desc	115348	100%
hardship_dpd	115348	100%
hardship_length	115348	100%
settlement_term	115347	100%
debt_settlement_flag_date	115347	100%
settlement_status	115347	100%
sec_app_mths_since_last_major_derog	110468	95.77%
mths_since_last_record	101958	88.39%
verification_status_joint	100783	87.37%
sec_app_revol_util	98955	85.79%

经统计，在删除贷后指标、无关指标后保留的的 120 个变量中，共 53 个变量存在缺失，有 38 个变量的缺失比率高于 70%，以 70%作为缺失值的阈值，删除缺失率超过 70%的变量，同时删除行全为缺失值的样本。由于均值插补法能够同时处理数值型数据和字符型数据，采用均值法对之后的数据进行填充。

4.2.3 异常值处理

异常值指明显偏于其他正常数值的数值。因为后续进行 WOE 分箱处理，故需要剔除同值性数据，同时采用箱线图及“3σ”原则检测异常值，如债务收入比(dti)的变量值存在大于 100%的，对这些异常值用其均值进行填充。

4.2.4 变量衍生

变量衍生指依托现有的变量衍生出新的变量。它能够充分挖掘原始指标的有效信息，提高后续建模的稳定性。因此，根据贷款业务上的理解以及部分现有的指标，衍生以下 5 个新的变量：信用借款账户数与总的账户数比、周转余额与所有账户余额比、欠款总额和本次借款比、银行卡状态较好的个数与总银行卡数的比、余额大于零的循环账户数与所有循环账户数的比。同时，由于采用比值的方法来引入新变量，并不会引起共线性。

4.3 描述性分析

4.3.1 单变量分析

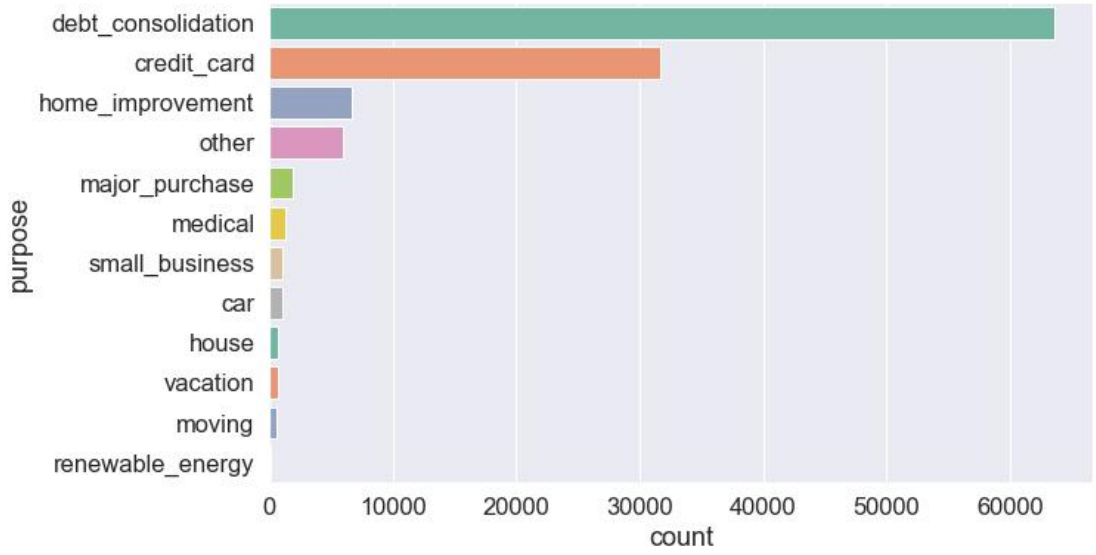

图 4.1 贷款用途

Fig.4.1 Loan purpose

图 4.1 为申请人的贷款用途的分布情况，可以看到 P2P 平台用户贷款用途最多的是债务重组，即用新债偿还旧债，其次是信用卡还款。然而，这些用户一般在传统银行贷款无法满足需求的情况下，向 P2P 平台贷款，故偿还能力较弱，违约风险较高。

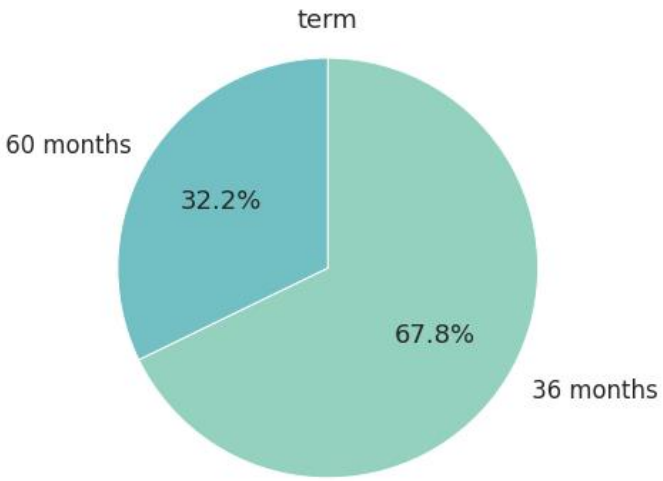

图 4.2 贷款周期

Fig.4.2 Loan cycle

图 4.2 为申请人的贷款周期的占比情况，可分为六十个月和三十六个月，其中三十六个月占比达到 67.8%。说明该平台大部分为短期贷款，但也有部分为长期贷款，对于平台来说，放贷期限越长，风险也就越高，但一般的收益率及利率也较高。

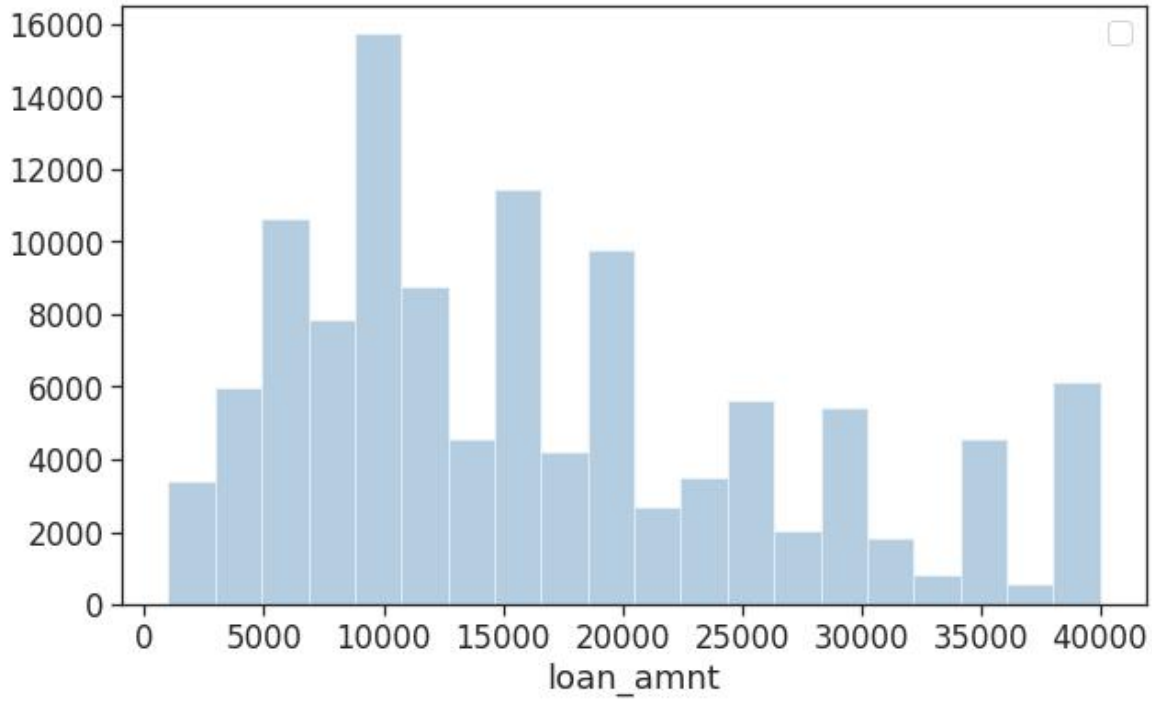

图 4.3 贷款金额

Fig.4.3 Loan amount

图 4.3 为申请人的贷款金额的分布情况，可以发现单笔贷款金额在 10000 美元左右范围的占比最高，这也进一步地证明了 Lending Club 主要经营中小额度的贷款发放。

4.3.2 多变量分析

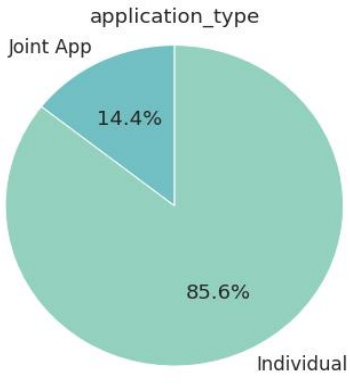

图 4.4 申请人类型

Fig.4.4 Applicant type

图 4.4 为申请人类型的分布情况。其中单独申请占 85.6%，联合申请占 14.4%。根据好坏样本计算违约率得到，单独申请的申请人违约率为 0.72%，联合申请的申请人的违约率为 0.64%。因此，可以得到结论：贷款申请大多为单独的个人申请，并且其违约率高于联合申请，风险相对较高。

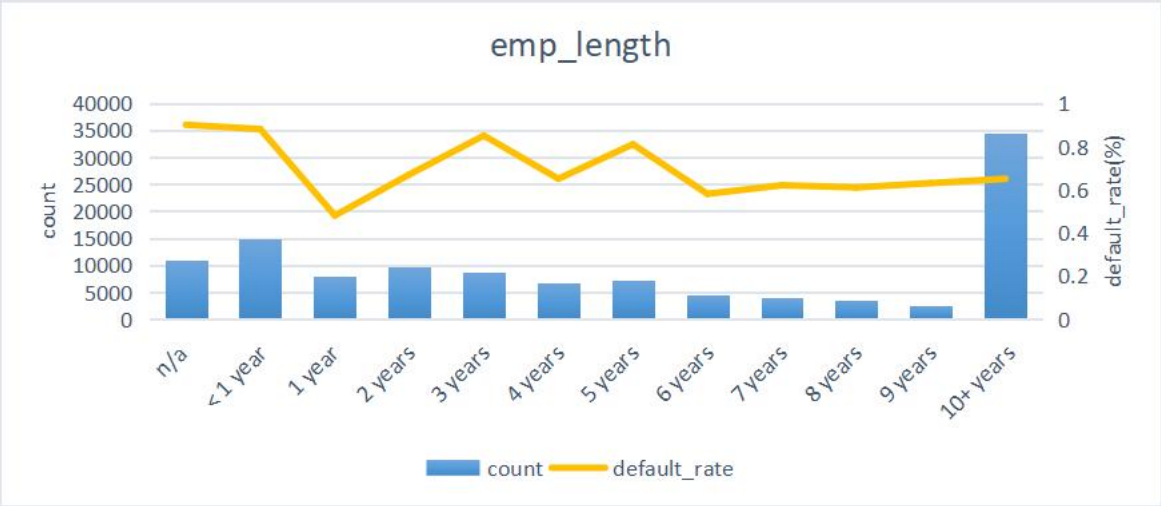

图 4.5 就业年限与违约率

Fig.4.5 Years of employment and default rate

图 4.5 为申请人的就业年限与违约的关系情况，整体来看，无就业经验的申请人发生违约的可能性最高；就业年限越长，违约率越低也越稳定。此外，就业年限超过 10 年的客户占比最高，这也恰恰是该 P2P 平台的主要服务对象。

4.4 特征筛选

在经过上述的数据清洗之后，143 个自变量已经减少到 81 个自变量。为了避免出现过拟合以及增强模型的泛化性，将原始数据分类为训练集与测试集，训练集占比 80%，测试集占比 20%。由于数据清洗后，变量依旧较多，且可能包含着对因变量解释能力较弱的变量，直接建模的效果较差，因此仍然需要进行变量筛选，剔除解释能力差或强相关的变量。

4.4.1 WOE 分箱及 IV 值

首先查看数据集的数据类型。若变量的数据类型为整数型或者浮点型，记为连续变量，否则离散变量。初步得到，连续变量为 73 个，离散变量为 8 个，包括

贷款期限、就业年限、贷款用途等。

为了便于之后的分箱，将连续变量取值少于 10 种的，重新定义为离散变量，以下表 4.4 展示了这 7 个变量。故最终得到 66 个连续变量，15 个离散变量。

表 4.4 取值种类小于 10 的连续变量

Table 4.4 Continuous variable with value type less than 10

变量名称	取值的种类
collections_12_mths_ex_med	7
pub_rec_bankruptcies	5
pub_rec	5
issue_m	3
open_il_12m	7
chargeoff_within_12_mths	6
inq_last_6mths	6

接下来分别对连续变量和离散变量进行分箱，分箱的数量不多于 5 箱，分箱能够减少数据的无谓波动及极端值的影响。对卡方分箱后的变量进行 WOE 编码，同时对数据映射，形成新的训练集与测试集。部分连续变量的分箱结果展示如下表 4.5，部分离散变量的分箱结果展示如下表 4.6：

表 4.5 部分连续变量的分箱展示

Table 4.5 Box display of partial continuous variables

	Bin	Bin_low	Bin_up	WOE
delinq_2yrs	1	-inf	3.15	-0.006544172
	2	3.15	5.04	0.750340265
	3	5.04	9.03	-0.479900466
	4	9.03	11.13	1.80070572
	5	11.13	inf	-3.345617136
inq_last_12m	1	-inf	0.07	-0.25641392
	2	0.07	2.03	-0.009946402
	3	2.03	3.01	0.2580603
	4	3.01	6.02	0.133783885
	5	6.02	inf	0.489236263

表 4.6 部分离散变量的分箱展示

Table 4.6 Box display of partial discrete variables

	Bin	Var_name	WOE
verification_status	1	Not Verified	-0.086317699
	2	Source Verified	-0.014338839
	3	Verified	0.245262928
term	1	60 months	-0.149336666
	2	36 months	0.064060719

分箱完成后可以得到各个变量的 IV 信息值，筛选 IV 值预测强度较高的变量，即满足 IV>=0.2 的变量，初步筛选出了 61 个自变量，下图展示 IV 值处于前 15 名的变量。

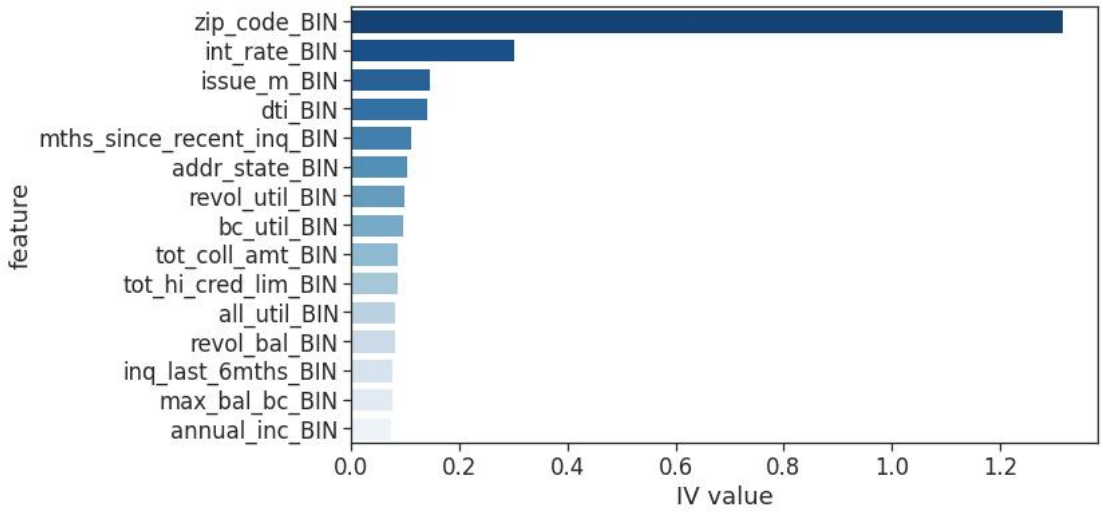

图 4.6 前 15 名变量的 IV 值

Fig.4.6 IV value of the top 15 variables

4.4.2 相关系数检验

多重共线性指自变量之间高度相关，严重的多重共线性会降低自变量对因变量的预测能力、分析结果不稳定，因此需要检测并剔除存在高度相关性的特征，判定方法包括相关系数检验以及方差膨胀因子检验。

设置相关系数即 Pearson 相关系数的阈值为 0.7，若两个变量间的相关系数大于 0.7，则剔除 IV 值较小的那个变量，不断循环，直到均大于 0.7 为止。通过相关系数检验，删除了 5 个变量，剩余变量的相关系数图如下图所示：

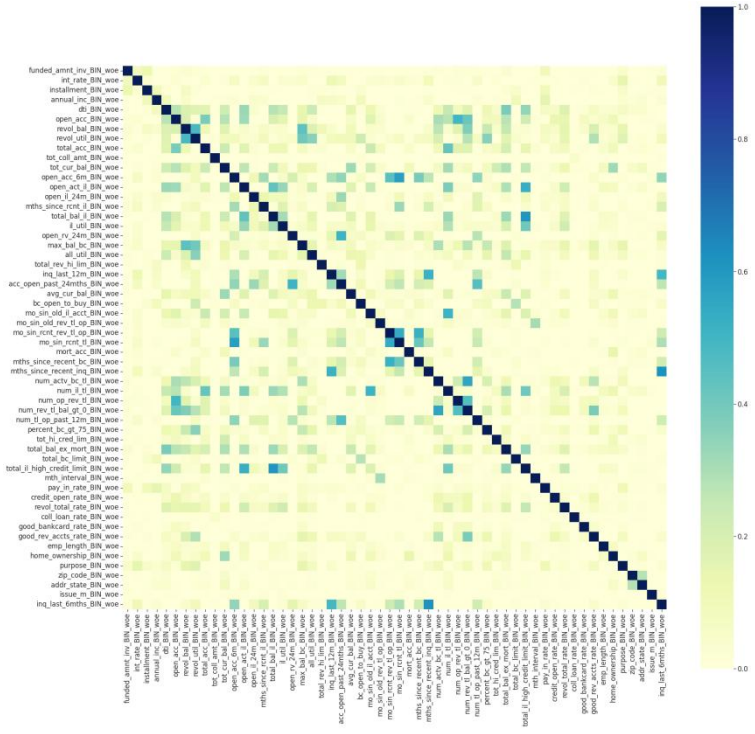

图 4.7 相关系数热力图

Fig.4.7 Correlation coefficient thermodynamic diagram

4.4.3 方差膨胀因子筛选

由于较小的相关系数并不意味着没有多重共线性问题，故在剔除不显著的变量后，需要通过计算方差膨胀因子 VIF 进一步检验。

$$VIF = 1/(1-R^2) \tag{4.1}$$

VIF 为容忍度的倒数，R^2 表示判决系数，VIF 取值越小，则说明多重共线性越轻，反之越重。当 VIF<10 时，则认为几乎不存在多重共线性。上述计算得出所有变量中的 VIF 的最大取值为 2.17988<10，因此，样本的不同特征之间的多重共线性较弱，继续保留上一环节的变量。

4.4.4 随机森林变量筛选

筛选重要程度大于等于 0.02 的变量，最终筛选出建立后续模型的 11 个变量。

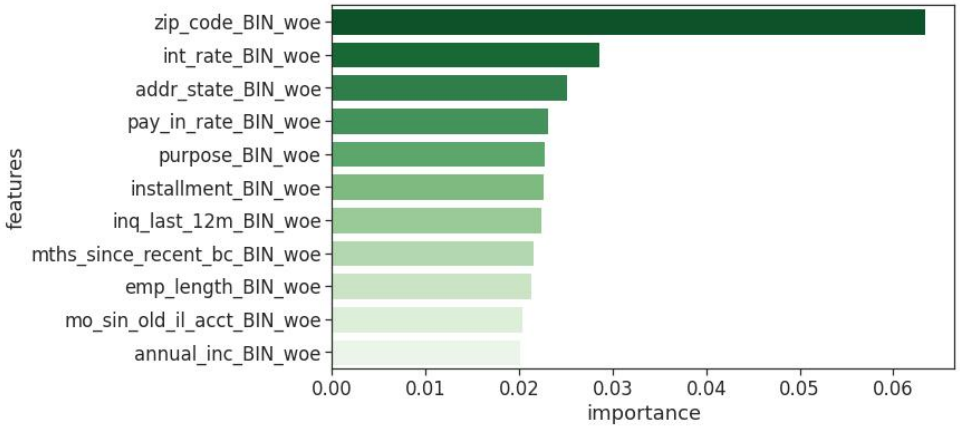

图 4.8 随机森林重要性排序

Fig.4.8 Random forest importance ranking

表 4.7 最终的变量

Table4.7 Final variables

变量	变量说明	重要性程度
zip_code_BIN_woe	邮编前三位	0.06337
int_rate_BIN_woe	贷款利率	0.02856
addr_state_BIN_woe	申请地址	0.02510
pay_in_rate_BIN_woe	年还款总额占年收入之比	0.02310
purpose_BIN_woe	贷款用途	0.02269
installment_BIN_woe	月还款额	0.02258
inq_last_12m_BIN_woe	过去 12 个月的查询次数	0.02239
mths_since_recent_bc_BIN_woe	自最近开设银行卡账户以来的月份	0.02155
emp_length_BIN_woe	就业年限	0.02136
mo_sin_old_il_acct_BIN_woe	自最早的分期付款账户开立以来的月份	0.02040
annual_inc_BIN_woe	年收入	0.02014

最终筛选后的变量共 11 个，大致可分为个人资质信息类、申请信息类、信用信息类。个人资质信息类指贷款人的资质、还款的稳定性及能力，包括所在地址、就业年限、年收入；信用信息类指贷款人的历史信用情况，包括过去 12 个月的查询次数、距最近开设银行卡账户的时间、距最早开设分期付款账户的时间；申请信息类指贷款人申请贷款的信息，包括贷款利率、月还款额、贷款用途。这些指标对于申请人能否成功申请贷款以及发生违约的风险都具有很大的参考价值，可见最终的这些变量对是否违约的目标变量 y 具有较强的解释性与预测性。

5 网贷用户违约预测模型

根据上一章节划分的 80%的训练集以及 20%的测试集数据，由于原始数据的不平衡比例约为 139:1，属于不平衡数据，分别对训练集与测试集进行 SMOTE 平衡处理，使好样本与坏样本的比例达到 1:1，基于平衡后的训练集数据进行建模，之后对平衡后的测试集数据进行预测，最后对比各模型的测试集结果。

5.1 逻辑回归结果分析

通过逻辑回归模型对测试集的数据进行预测，得到混淆矩阵以及部分算法指标如下所示：

表 5.1 逻辑回归结果混淆矩阵

Table5.1 Logistic regression result confusion matrix

实际\预测	0(好样本)	1(坏样本)
0(好样本)	17286	5620
1(坏样本)	6115	16791

表 5.2 逻辑回归算法指标

Table5.2 Logistic regression algorithm index

	precision	recall	f1-score	support
0	0.74	0.75	0.75	22906
1	0.75	0.73	0.74	22906
Avg	0.740	0.740	0.745	45812

由表 5.1 及 5.2 可得，逻辑回归模型对好样本的召回率为 75%，对坏样本的召回率为 73%。模型整体的平均精确率及平均召回率为 74%，两者水平相当。因此，逻辑回归模型预测能力一般，不能很好地区分好坏样本。

5.2 随机森林结果分析

表 5.3 随机森林结果混淆矩阵

Table5.3 Random forest result confusion matrix

实际\预测	0(好样本)	1(坏样本)
0(好样本)	22850	56
1(坏样本)	8897	14009

表 5.4 随机森林算法指标

Table5.4 Random forest algorithm index

	precision	recall	f1-score	support
0	0.72	1.00	0.84	22906
1	1.00	0.61	0.76	22906
Avg/total	0.860	0.805	0.800	45812

由表 5.3 及 5.4 可得，基于 Bagging 算法构建的随机森林模型，对好样本的判断能力很强，召回率近乎高达 100%，远远优于逻辑回归模型；但是对坏样本的判断能力仅为 61%，识别能力较差。整体的平均精确率为 86%，F1 得分为 0.80，较难识别坏样本，说明随机森林模型在处理非平衡数据上的能力有待进一步地提升。

5.3 AdaBoost 结果分析

表 5.5 AdaBoost 结果混淆矩阵

Table5.5 AdaBoost result confusion matrix

实际\预测	0(好样本)	1(坏样本)
0(好样本)	20017	2889
1(坏样本)	4799	18107

表 5.6 AdaBoost 算法指标

Table5.6 AdaBoost algorithm index

	precision	recall	f1-score	support
0	0.81	0.87	0.84	22906
1	0.86	0.79	0.82	22906
Avg/total	0.835	0.830	0.830	45812

由表 5.5 及 5.6 可以发现，AdaBoost 模型对预测好坏样本的结果比较理想，对好样本的召回率虽略低于随机森林模型，但坏样本的召回率为 79%。在构建网络

借贷的违约预测模型中，我们也更加关注坏样本是否能够正确地被识别。模型的整体平均精确率为 83%，预测能力较强。故通过不断更新观测数据的权重，采用自适应加强算法的 AdaBoost 模型能较好地处理非平衡数据，对预测客户是否会发生违约的效果显著。

5.4 GBDT 结果分析

表 5.7 GBDT 结果混淆矩阵

Table5.7 GBDT result confusion matrix

实际\预测	0(好样本)	1(坏样本)
0(好样本)	21661	1245
1(坏样本)	3639	19267

表 5.8 GBDT 算法指标

Table5.8 GBDT algorithm index

	precision	recall	f1-score	support
0	0.86	0.95	0.90	22906
1	0.94	0.84	0.89	22906
Avg/total	0.900	0.895	0.895	45812

由表 5.7 及 5.8 可得，在 GDBT 模型的预测下，好样本的召回率为 95%，几乎全被召回；同时，在 22906 个坏样本中，正确识别了 19267 个样本，坏样本的召回率也达到 84%。整体的平均召回率和精确率都在 90%左右，F1 得分为 0.89，效果优于前几个模型。

5.5 CatBoost 结果分析

表 5.9 CatBoost 结果混淆矩阵

Table5.9 CatBoost result confusion matrix

实际\预测	0(好样本)	1(坏样本)
0(好样本)	22903	3
1(坏样本)	1219	21687

表 5.10 CatBoost 算法指标

Table5.10 CatBoost algorithm index

	precision	recall	f1-score	support
0	0.95	1.00	0.97	22906
1	1.00	0.95	0.97	22906
Avg/total	0.975	0.975	0.970	45812

由表 5.9 及 5.10 可得，CatBoost 算法下的模型对好样本的召回接近于 1，只有 3 个好样本没有被正确识别，此外，对于坏样本的召回率也达到了 95%，整体的效果优于逻辑回归单一模型、随机森林模型、Adaboost 模型与 GBDT 模型，预测能力最强。

5.6 Stacking 结果分析

在构建 Stacking 融合的模型中，第一层选取对坏样本的召回率较高的模型，即 GBDT 模型、AdaBoost 模型、CatBoost 模型，第二层为了防止出现过拟合，采用逻辑回归模型。因此，第一层采用 GBDT 模型、AdaBoost 模型及 CatBoost 模型对训练集数据进行预测，然后将其预测结果作为第二层的训练样本，采用逻辑回归模型再预测，后输出最终预测结果。以下表 5.9 及 5.10 为得到的预测结果，可以发现对好坏样本的召回率都较高，整体的平均精确率及平均召回率为 98%，模型预测效果较好。

表 5.11 Stacking 算法融合模型结果混淆矩阵

Table5.11 Stacking algorithm fusion model result confusion matrix

实际\预测	0(好样本)	1(坏样本)
0(好样本)	22876	30
1(坏样本)	683	22223

表 5.12 Stacking 算法融合模型指标

Table5.12 stacking algorithm fusion model indicator

	precision	recall	f1-score	support
0	0.97	1.00	0.98	22906
1	1.00	0.97	0.98	22906
Avg	0.985	0.985	0.980	45812

5.7 模型对比

表 5.13 各模型准确度对比

Table5.13 Comparison of the accuracy of each model

模型	实际\预测	0(好样本)	1(坏样本)
逻辑回归	0(好样本)	17286	5620
	1(坏样本)	6115	16791
随机森林	0(好样本)	22850	56
	1(坏样本)	8897	14009
AdaBoost	0(好样本)	20017	2889
	1(坏样本)	4799	18107
GBDT	0(好样本)	21661	1245
	1(坏样本)	3639	19267
CatBoost	0(好样本)	22903	3
	1(坏样本)	1219	21687
Stacking	0(好样本)	22876	30
	1(坏样本)	683	22223

根据实际违约结果与预测违约结果来看，实际违约样本共 22906 个，在单一模型中，逻辑回归模型正确预测 16791 个样本，随机森林模型正确预测 14009 个样本，AdaBoost 模型正确预测 18107 个样本，GBDT 模型正确预测 19267 个样本，CatBoost 模型正确预测 21687 个样本。可以得出结论：1) 逻辑回归模型及随机森林模型对坏样本的判别能力较弱，受数据不平衡影响大。2) AdaBoost 模型、GBDT 模型及 CatBoost 模型对坏样本的召回率较高，明显高于前两个模型，一定程度上削弱了样本不均衡的影响，预测能力较强。3) 对于 Stacking 算法的融合模型来说，与单一模型预测效果最好的 CatBoost 模型相比，虽识别的好样本略有减少，但多了 356 个坏样本被正确预测，模型综合表现更好。

表 5.14 各模型精度指标对比

Table5.14 Comparison of the accuracy indicators of each model

模型	precision	recall	f1-score
逻辑回归	0.749	0.733	0.741
随机森林	0.996	0.612	0.758

模型	precision	recall	f1-score
AdaBoost	0.862	0.790	0.825
GBDT	0.939	0.841	0.888
CatBoost	0.999	0.946	0.973
Stacking	0.998	0.970	0.984

表 5.14 是各个模型整体的精准率、召回率、f1 得分的对比情况，可以得到：1) 逻辑回归模型整体的精准率及召回率都表现一般，对好坏样本的区分能力不高。2) 随机森林模型对坏样本的召回率表现甚至低于逻辑回归模型，可见受不均衡影响较大，从而影响到了模型的整体效果，整体 f1 得分为 0.758。3) AdaBoost 模型和 GBDT 模型在预测的准确程度及召回程度上表现较好，相差不大。4) CatBoost 模型在各方面表现更加优秀。5) Stacking 算法的融合模型的召回率和 f1 得分均高于 CatBoost 模型。因此，综合来看，Stacking 算法的融合模型在这三个指标的表现效果上更加显著。

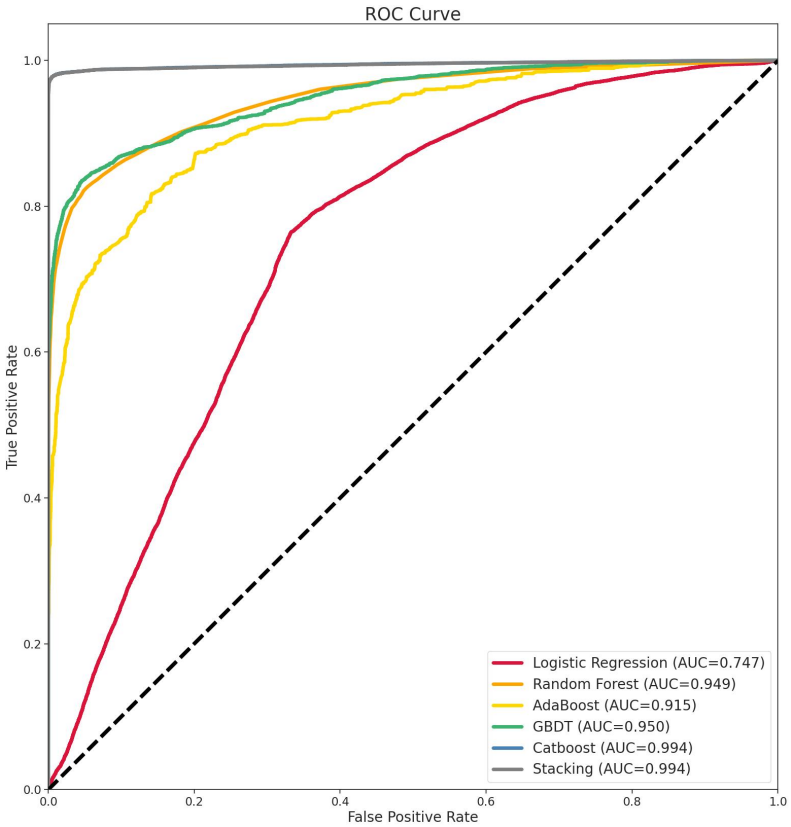

图 5.1 各模型 ROC 曲线

Fig.5.1 ROC curve of each model

表 5.15 各模型 AUC 值及 KS 值对比

Table5.15 Comparison of AUC and KS values of each model

模型	AUC 值	KS 值
逻辑回归	0.74720	0.43167
随机森林	0.94884	0.77294
AdaBoost	0.91488	0.67567
GBDT	0.95035	0.78905
CatBoost	0.99447	0.97236
Stacking	0.99430	0.97232

图 5.1 绘制了各模型 ROC 曲线，表 5.15 展示了各模型 AUC 值及 KS 值对比。通过这些指标的对比，可以得到逻辑回归模型的 AUC 值略高于 0.7，对正负样例的判别效果一般。随机森林模型由于对好样本的召回率过于高，导致 AUC 值和 KS 值较高，但忽略了对坏样本的召回率较低这一弊端，故整体来看，表现不佳。AdaBoost 模型及 GBDT 模型在这些指标的表现上都比较优秀，CatBoost 模型及基于 Stacking 算法的融合模型两者预测能力最高，有效性最强，更加适合构建违约预测模型。

6 总结与展望

6.1 总结

互联网金融的飞速发展衍生出了 P2P 网贷的新生业态，P2P 网贷更快捷、更高效，也缓解了中小企业融资难的难题、有效提高了资金的利用率，但也出现了互联网欺诈，信用风险成为金融企业面临的巨大挑战，因此金融企业有必要构建更加精准的借贷违约预测模型，最大程度地降低企业损失。

本文基于以 Lending Club 平台为代表的 P2P 网贷平台的客户数据，分别使用了单一模型逻辑回归和集成模型随机森林算法、AdaBoost 算法、GBDT 算法、CatBoost 算法以及 Stacking 融合算法对 P2P 网贷数据建立了违约预测模型。

第一部分主要阐述了本文研究的背景与意义、国内外关于违约预测的研究、研究的内容及框架。第二部分阐述了 P2P 网贷的含义、面临的风险及相应举措。第三部分对涉及到的模型与理论进行介绍。建模最重要的是对数据的处理，对杂乱无章的数据建模没有重要意义。故第四部分通过处理贷后指标、缺失值、异常值等方式清洗数据，然后采用描述性分析对部分变量有大概了解，最后对高维数据进行变量筛选。第五部分建立了逻辑回归和五种集成模型对客户分类，使用好坏样本的召回率、AUC 值、KS 值等指标对单一模型和集成模型对比。以下是本文研究得到的结论：

从单一模型的预测能力来看，随机森林、AdaBoost、GBDT、CatBoost 算法的效果整体均优于逻辑回归模型，逻辑回归的精确度较低，其余集成模型的预测能力较强，基于 CatBoost 算法建立的 P2P 违约预测模型的整体效能最好，充分显现出了 CatBoost 算法应用于个人信用风险评估的优越性与可行性。

从数据的不平衡角度来看，逻辑回归模型与随机森林模型受数据不平衡的影响较大，对坏样本的召回率较低。

从融合模型的预测能力来看，将精确率较高、受样本不平衡的影响程度较小的模型，AdaBoost、GBDT、CatBoost 算法下的模型作为第一层的基模型，逻辑回归模型作为第二层的训练模型，采用 Stacking 算法来融合模型，与 CatBoost 的分类效果相比，增强了对坏样本的识别能力，同时根据 AUC 及 KS 值，预测的准确率达到 99.4%，同时区分能力极强。

从违约影响因素来看，本文最终筛选出的变量包括邮编前三位、贷款利率、申请地址、年还款总额占年收入之比、贷款用途、月还款额、过去 12 个月的查询次数等指标，从而可以发现借款人的位置信息、财务收入、历史信用情况、贷款利率对借款人是否为发生违约的影响较大。

6.2 创新及不足

本文研究的创新性体现在以下几个方面：

① 编码形式：本文没有采用传统的独热编码、哑变量编码的无监督的编码形式，而是采用了 WOE 分箱编码的有监督的编码形式，降低了数据的干扰，增强了变量的可解释性。

② 特征筛选：不同的特征筛选方式利用的指标不同，从而对变量的敏感程度有所差异，本文采用了 IV 值筛选、相关系数及方差膨胀因子检验、随机森林重要性排序的方法，提高了特征筛选的准确性，充分挖掘了变量的有效信息，将最终清洗好的数据使用到模型中。

③ 模型：CatBoost 属于近年来集成学习的前沿算法，本文创新性地基于 CatBoost 算法建立模型来控制违约风险，经过与其他的集成模型进行对比，论证了 CatBoost 算法应用到违约风险控制领域的可行性。最后建立了 Stacking 融合模型，其分类预测结果略优于 CatBoost 算法。

本文虽然对于用户违约预测问题建立了集成学习模型并且取得了优秀的预测结果，但仍存在不足，需在之后的研究予以改进，大致可概括为以下几点：

① 数据应用：由于 Lending Club 官网只提供了截止到 2019 年的数据，故本文采用的数据为 2019 年 Lending Club 的客户数据，有一定的滞后性；同时，由于国内数据的不公开、没有建立系统的评估体系，本文采用美国 P2P 平台为例进行研究，可能会对国内 P2P 网贷平台的应用有局限性。

② 模型层次：本文使用的传统逻辑回归模型的准确性不高，模型的非线性以及前期平衡处理、数据分箱、特征筛选引起的信息流失，所以该模型的分类结果受到了影响，因此后期对于数据的预处理、特征筛选的过程需要进一步完善，提高逻辑回归模型的预测精度。

③ 非平衡处理：本文使用 SMOTE 分别对训练集、测试集平衡，好坏样本比例达到 1:1，但是 SMOTE 方法通过对附近的少数类样本线性差值来扩充少数类样本，仍然是依据原有数据生成新数据，并没有真正解决非平衡的数据分布问题，同时还出现了分布边缘化、近邻选择盲目性、数据量扩大等问题，有待采用其他平衡方法进行改进。

④ 模型调参：本文使用三折交叉验证和网格调参，若效果较好则停止对参数的寻优，容易出现局部最优；并且当参数的取值范围较大时，易出现维度灾难。此外，本文的数据量较大，参数寻优的时间过长。最近几年，也出现了一些优秀的调参方法，如贝叶斯调参、贪心调参，能够有效提高模型的整体效能，有待进一步的实践与研究。

参 考 文 献

[1] 桂琴. 互联网金融平台中欺诈型用户的识别及防控研究[D]. 哈尔滨工业大学,2020.

[2] Thomas L. A Survey of Credit and Behavioural Scoring: Forecasting financial risk of lending to consumers[J]. International Journal of Forecasting, 2000,16(2):149-172.

[3] Durand D. AppendixB:Application of the Method of Discriminant Functions to the Good-andBad-Loan Samples[J]. Ecological Entomology,1941,30(6):692-699.

[4] 姜琳. 美国 FICO 评分系统述评[J]. 商业研究,2006(20):81-84.

[5] Altman. Discriminant analysis and the prediction of corporate bankruptcy[J]. The Journal of Finance,1968,23(4):5889-609.

[6] Wiginton,John C. A Note on the Comparsion of Logit and Discriminant Models of Consumer Credit Behavior[J]. The Journal of Financial and Quantitative Analysis,1980,15(3):757.

[7] A. I. Kokkinaki. On A typical Database Transactions: Identification of Probable Frauds using Machine Learning for User Profiling[C]. Proceedings of IEEE Knowledge and Data Engineering Exchange Workshop, 1997, 107-113.

[8] Baesens B.,Van Gestel T.,Viaene S. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring[J]. The Journal of the Operational Research Society,2003,54(6):627-635.

[9] Stjepan Oreski, Goran Oreski. Genetic algorithm-based heuristic for feature selection in credit risk assessment[J]. Expert System with Applications,2014,41:2052-2064.

[10] Maher Ala'raj,Maysam F. Abbod. A new hybrid ensemble credit scoring model based on classifiers consensus system approach[J]. Expert Systems With Applications,2016,64(C):36-45.

[11] Xiaojun Ma,Jinglan Sha,Dehua Wang,Yuanbo Yu,Qian Yang,Xueqi Niu. 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.

[12] 于立勇,詹捷辉. 基于 Logistic 回归分析的违约概率预测研究[J]. 财经研究,2004(09):15-23.

[13] 涂伟华,王索漫. 基于数据挖掘方法对商业银行信用卡违约预测模型的研究[J]. 中国证券期货,2011(09):146-147.

[14] 方匡南,章贵军,张惠颖. 基于 Lasso-logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济究,2014,31(02):125-136.

[15] 过新伟. 我国中小企业信用风险度量研究[D]. 南开大学,2012.

[16] 周丽峰. 基于非平衡数据分类的贷款违约预测研究[D]. 中南大学,2013.

[17] 杨斌. 基于 COX 回归的企业违约风险研究[D]. 浙江财经学院,2013.

[18] 夏雨霏,刘传哲,徐嘉辰. 聚类支持向量机在 P2P 网络借贷违约预测中的应用[C]. 第十届（2015）中国管理学年会论文集,2015:507-514.

[19] 刘铭,张双全,何禹德. 基于改进型模糊神经网络的信用卡客户违约预测[J]. 模糊系统与数学,2017,31(01):143-148.

[20] 沙靖岚. 基于 LightGBM 与 XGBoost 算法的 P2P 网络借贷违约预测模型的比较研究[D]. 东北财经大学,2017.

[21] 王嘉琪. 基于数据挖掘技术的 P2P 借贷违约风险识别模型研究[D]. 浙江工商大学,2018.

[22] 杨盛辉. 基于加权 Stacking 集成学习的信用卡违约预测[D]. 桂林理工大学,2019.

[23] 马晓君,宋嫣琦,常百舒,袁铭忆,苏衡. 基于 CatBoost 算法的 P2P 违约预测模型应用研究[J]. 统计与信息论坛,2020,35(07):9-17.

[24] 逯瑶瑶. 基于机器学习分类算法的贷款违约预测研究[D]. 兰州大学,2021.

[25] 刘美伶. 基于 ANP-LightGBM 算法的信用卡用户违约预测模型研究[D]. 重庆工商大学,2021.

[26] 郑欣彤. 中国互联网金融下的 P2P 网贷发展现状和风险控制[J]. 经济管理文摘,2020(19):27-29.

[27] 鲍菲. P2P 网贷平台信用风险评价研究[D]. 东北石油大学,2019.

[28] 易楚钧. 互联网金融背景下 P2P 网络借贷平台的风险及规制[J]. 韶关学院学报,2020,41(04):65-69.

[29] 杨悦. 面向不平衡数据的分类方法研究[D]. 桂林电子科技大学,2021.

[30] 石洪波,陈雨文,陈鑫. SMOTE 过采样及其改进算法研究综述[J]. 智能系统学报,2019,14(06):1073-1083.

[31] 王青天,孔越,李华君. Python 金融大数据风控建模实战:基于机器学习[M]. 机械工业出版社,2020:91-95.

[32] 白婧怡. 基于经典评分卡与机器学习的金融风险识别模型及其应用[D]. 天津商业大学,2019.

[33] 周志华. 机器学习[M]. 清华大学出版社,2016:57-59.

[34] 杨剑锋,乔佩蕊,李永梅,王宁. 机器学习分类问题及算法研究综述[J]. 统计与决策,2019,35(06):36-40.

[35] Breiman L. Random Forests[J]. Machine Learning,2001,45(1):5-32.

[36] 尹华,胡玉平. 基于随机森林的不平衡特征选择算法[J]. 中山大学学报(自然科学版),2014,53(05):59-65.

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

[38] 陈钰. 基于数据挖掘方法 P2P 平台借贷违约预测模型研究[D]. 重庆大学,2020.

[39] 秦婉怡. 基于 CatBoost 算法的信用卡用户信用风险预测模型应用研究[D]. 重庆工商大学,2021.

附 录

A 学位论文数据集

关键词	密级	中图分类号
P2P 网络借贷；信用风险；逻辑回归；集成学习；WOE 分箱	公开	O1

学位授予单位名称	学位授予单位代码	学位类别	学位级别
重庆大学	10611	专业学位	硕士

论文题名	并列题名	论文语种
基于集成学习的 P2P 网贷用户的违约预测研究	无	中文

作者姓名	陈旭然	学号	202006131072
培养单位名称		培养单位代码	
重庆大学		10611	
学科专业	研究方向	学制	学位授予年
应用统计	机器学习	2	2022
论文提交日期	2022 年 6 月	论文总页数	55
导师姓名	张良才	职称	教授
答辩委员会主席		何传江	

电子版论文提交格式

文本（√） 图像（） 视频（） 音频（） 多媒体（） 其他（）

致 谢

在重庆大学读研究生的两年生活，如同流星般短暂而绚丽，回首过往，有欢笑也有汗水，心中百感万千，留下的满满都是感激。

首先，感谢我的导师——张良才老师，初见老师，便觉得待人亲切平和，渐渐地，老师严谨的治学态度、认真履责的作风也一直在感染着我，成为我努力的方向。张老师对我的学业给予了很多的指导，在我对未来迷茫时，帮助我分析我的性格特点以及未来的职业规划，让我明确了自己的初心所在，曾经老师赠送的栀子花，点缀了我的整个研究生生活，给予我勇气和力量去面对生活中的困难，不断突破自己。遇到张老师，何其有幸，非常感激老师两年来的指导。

其次，感谢我的辅导员李欣老师、学院秘书罗宥余老师，将学院的各项工作整理地井井有条，为我们的生活和学习提供了很多的帮助。同时感谢数学与统计学院的所有老师，知识是无价的，你们的谆谆教导才成就了今天的我们，我会牢记你们的殷切期盼和教诲，走好未来的人生之路。

此外，还需要感谢应用统计班的每一位同学，我们并肩作战走过绚烂的两年，互帮互助，一起留下了最美好的回忆，愿我们聚是一团火、散是满天星。尤其感激我的室友康欣、吴雨濛同学，谢谢你们一直包容着我的不足，倾听我的烦恼，在我怀疑自己时鼓励我，想方设法地帮助我，让我更加自信、能够从容地面对得与失。其次，感谢我的家人，无论我走多远，你们永远都坚定地站在我的身后，支持我的决定，为我无私的付出，在任何的风雨中都是最温暖的依靠，我会用一生来感恩和报答。

最后，非常感谢母校重庆大学，我会牢记学校对我们的教导，努力向社会贡献自己的力量，也希望更多的重庆大学学子在社会中能够一展锋芒！提起重庆，重庆是一座极具魅力又包容、冒着热气的城市，我的两年青春也留在了这片土地上。

祝愿我们一路走来，未来可期，谨以此文献给你们！

陈旭然

二O二二年三月 于重庆