分类号 C93　　密级 公开
U D C　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

论文题目 基于改进 BalanceCascade 方法的信用评分集成模型研究

研究生姓名：刘 伟

指导教师姓名、职称：韩金仓 教授

学科、专业名称：管理科学与工程

研究方向：信息管理与信息系统

提交日期：2022 年 5 月 29 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：刘伟 签字日期：2022.5.29

导师签名：韩宝彦 签字日期：2022.5.29

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：刘伟 签字日期：2022.5.29

导师签名：韩宝彦 签字日期：2022.5.29

Research on credit scoring ensemble model based on improved BalanceCascade method

Candidate : Liu Wei

Supervisor : Han Jincang

摘 要

近年来随着消费金融的迅猛发展，个人信贷业务也快速发展起来，不仅各种网贷平台增多，而且贷款品种也逐渐丰富，几乎涵盖了个人生产和生活的方方面面。然而来自信用风险的挑战也日趋严峻，通过对申请人信用评分进行风险评估显得尤为重要。

目前，虽然有较多的信用评分模型，但不同模型各有优劣。以往的研究发现单一模型训练速度较快，但预测精度与稳定性差；若能够选取合适的基分类器进行集成，可在一定程度上降低预测误差，提高准确性；并且在实际过程中，由于信用评分数据集自身局限性，正负样本类别差异极大，不平衡问题的处理对模型性能也有重要影响。基于以上问题，本文进行了如下研究：

本文利用随机森林方法进行特征选择，该方法在拟合数据后，能够对所有特征属性进行重要性度量，相较于金融风控中常用的信息值的特征选择方法避免了对每个特征的分箱操作，可直接获得特征重要性排序，实现更为简单，选择特征的速度更加高效；根据特性重要性排名与业务逻辑，最后选择重要性大于 0.1 的特征，一共选取 27 个特征作为入模变量。为检验不同类型模型在实际中的应用情况，选取逻辑回归（LR）、决策树（DT）、朴素贝叶斯（NB）与支持向量机（SVM）四种在信用评分分类预测性能较好、认可度较高的单一模型进行实验；之后分别以 LR、NB、DT、SVM 四种单一模型分别为基分类器进行 Bagging 集成，检验同质集成模型的分类性能；根据不同的基分类器进行集成可以相互补充，提高信用评分模型分类预测的精度与准确性，为检验实际分类效果，以 LR、NB、DT、SVM 四种性能较好的分类算法为基分类器，通过 bootstrap 进行抽样构建数据子集自适应投票选择 AUC 最高的基分类器进行集成，构建一种新的异质集成模型进行实验。

针对信用评分数据集中正负样本类别不平衡性问题，提出了一种改进的 BalanceCascade 方法，该方法通过抽取正类样本与负类样本构成平衡数据集训练 Adaboost 分类器，将分类错误率控制在一定范围内，确保移除正类样本的准确性；之后根据正负样本的不平衡比例，设置一个可调参数，通过不断移除一定比例的正样本，使得剩余正负样本比例接近此参数，对不同正负样本比例下的数据集进行实验，结合新的分层模型进行训练，寻找最优的比例参数。由于 RF 与 XGBoost

在信用评分中准确性方面具有的较大优势，所以选择 RF 与 XGBoost 作为第一层的基分类器，而第二层模型不应太复杂，太过复杂的话可能会导致模型在训练集上过拟合、泛化效果差等问题，所以该层模型选用较为稳定的单一模型逻辑回归为基分类器，通过在阿里天池竞赛上的信用数据集实验结果显示，当正负样本比例设置为 2 时，基于改进 BalanceCascade 方法的信用评分集成模型准确率达到 0.80，精确率 0.90，召回率为 0.84，F1 值为 0.88，AUC 值 0.74，相较于单一分类模型、Bagging 集成模型、自适应选择 AUC 的异质集成模型，基于改进 BalanceCascade 方法的集成模型效果更好，更加稳定。

关键词：信用评分 单一模型 集成模型 不平衡处理

Abstract

With the rapid development of consumer finance in recent years, personal credit business has also developed rapidly, not only with the increase of various online lending platforms, but also with the gradual enrichment of loan varieties, covering almost all aspects of personal production and life. However, the challenge from credit risk is getting more and more serious, and risk assessment through credit score of applicants is especially important.

Currently, there are many credit scoring models, but different models have their own advantages and disadvantages. Previous studies have found that a single model is faster to train but has poor prediction accuracy and stability; if a suitable base classifier can be selected for integration, the prediction error can be reduced to a certain extent and the accuracy can be improved; moreover, in practice, due to the limitations of the credit score dataset itself, the positive and negative sample categories are extremely different, and the handling of the imbalance problem also has an important impact on the model performance. Based on the above issues, the following research is conducted in this paper.

In this paper, we use random forest method for feature selection, which is able to measure the importance of all feature attributes after fitting the data, compared with the feature selection method of information value commonly used in financial risk control, which avoids the operation of

binning each feature and can directly obtain the ranking of feature importance, which is simpler to implement and more efficient in selecting features; according to the ranking of feature importance and business logic, we finally select According to the importance ranking of features and business logic, the features with importance greater than 0.1 are finally selected, and a total of 27 features are selected as entry variables. To test the application of different types of models in practice, four single models of logistic regression (LR), decision tree (DT), simple Bayesian (NB) and support vector machine (SVM) with better performance and higher recognition in credit score classification prediction were selected for experiments; after that, four single models of LR, NB, DT and SVM were used as base classifiers for Bagging integration respectively In order to test the actual classification effect, four classification algorithms with better performance, LR, NB, DT and SVM, were used as base classifiers, and the base classifier with the highest AUC was selected by bootstrap sampling to build a subset of data for adaptive voting. The classifier with the highest AUC is selected by bootstrap sampling to construct a new heterogeneous integration model for experiments.

For the problem of imbalance between positive and negative samples in the credit score dataset, an improved BalanceCascade method is proposed, which trains the Adaboost classifier by extracting positive and negative samples to form a balanced dataset to control the classification error

rate within a certain range and ensure the accuracy of removing positive samples; after that, according to the imbalance ratio of positive and negative samples, an adjustable parameter is set to ensure the accuracy of removing positive samples. After that, an adjustable parameter is set according to the imbalance ratio of positive and negative samples, and by continuously removing a certain proportion of positive samples, the remaining proportion of positive and negative samples is made close to this parameter, and experiments are conducted on data sets with different proportions of positive and negative samples, combined with the new hierarchical model for training to find the optimal proportion parameter. Because of the greater advantage of RF and XGBoost in accuracy in credit scoring, RF and XGBoost are chosen as the base classifier of the first layer, while the second layer model should not be too complex, too complex may lead to problems such as overfitting and poor generalization of the model on the training set, so the layer model is chosen as a more stable single model logistic regression as the base classifier, through the The experimental results of the credit dataset on the Ali Tianchi competition show that when the ratio of positive and negative samples is set to 2, the accuracy of the credit score integration model based on the improved BalanceCascade method reaches 0.80, the accuracy 0.90, the recall 0.84, the F1 value 0.88, and the AUC value 0.74, compared with the single classification model, the Bagging integration model, the Heterogeneous integration model for adaptive

selection of AUC, the integration model based on the improved BalanceCascade method is better and more stable than other models.

Keywords: Credit scoring; Single model; Ensemble model; Imbalance processing

目 录

1 绪 论

1.1 研究背景与意义

1.1.1 研究背景

信用评分虽经历了一个较长时间的发展，然而其发展并不均衡。在西方发达国家，信用评分已经经历了几十年的发展。在十九世纪五十年代，William Fair 和 Earl Isaac 创立的非埃哲公司（简称 FICO）,成为当时信用风险领军企业，而早期美国的征信机构信用评估的指标一般以借款人的婚姻问题、性取向、政治活动和倾向为主，缺乏科学的信贷决策，直到 1956 年该公司提供了一种计分卡产品，通过将借款人的信息填入其中，并统一计算出一个最终结果来看申请者是否超过了一个可接受的风险水平，也就是 FICO 信用评分模型，该模型主要从客户的支付历史、信贷欠款数额、立信时间情况、新开信用账户情况、信用组合类型情况五个方面进行量化评估[1]。之后随着消费信贷业务在美国迅速发展，信用卡对社会人群发放，伴随着人们对信用的熟知，申请人数不断增加，让银行认识到了发卡之前利用信用评分进行发卡决策的重要性，相较于传统的以经验为主的评估技术，信用评分模型的应用能够具有更高的效率。随着信用评分模型在信用卡申请过程中的普遍使用，让银行及第三方机构看到了其应用前景，逐渐开始将其应用于其他产品，比如个人消费贷款、房屋抵押贷款、小企业融资借贷等，进一步推动了信用评分的发展；而且在信用体系的形成和发展的影响下，信用评分模型的开发过程也逐步朝着规范化、标准化、科学化的方向迈进。

对于我国来说，信用一词虽然是历史悠久，但信用评分的发展可以说是一个新生品，我国的信用评分发端于 20 世纪 90 年代末，起源于特定行业发展出现的第三方服务，并且有着特定的模型和应用范围，伴随着行业的发展而不断完善。我国信用评分模型的出现与信用卡业务的发展密切相关，在当时改革开放的大环境下，经济快速发展，居民收入的增加促进了消费的发展，多种利好因素的刺激性，信贷业务有了良好的生长的基础，如房贷、车贷、消费贷等银行的零售业务得以快速推广，信用卡业务也随之进入发展的快车道；更为重要的是，银行自身

在大环境的影响下，开始引入新的外部技术与合作伙伴逐步朝着市场化、专业化的道路发展，将数据化、专业化的金融风控模型与管理体系进行应用，至此我国的信用评分市场才算是随着银行信贷业务的发展真正开展。

随着我国信用经济的发展，我国个人征信体系也逐步发展完善，但主要以中国人民银行为核心的征信体系、政府的公共征信体系、商业征信体系为主。但征信与信用评分却不相同，征信是针对个体历史信贷行为记录收集，或者说是信息的整理，是记录的行为事实；信用评分则是针对贷款资源申请者的历史信贷行为与其他相关数据，通过应用机器学习技术进行评估，给出合理的评定结果，为评估机构进行决策提供科学的依据。近年来互联网技术的飞速发展，新金融行业进入新的发展阶段，新的市场需求开始出现，个人信贷业务随之快速发展起来，不仅各种贷款平台增多，而且贷款品种也逐渐丰富，从住房、教育、消费品等到企业融资贷款等，几乎涵盖了个人生产和生活的方方面面。然而在全球化背景的影响下，金融市场环境波动性加剧的刺激下，违约现象频发，来自信用风险的挑战日趋严峻，推动银行与第三方金融机构等更加严格谨慎地对个人信用状况进行评估，如针对个人的历史记录、行为状况、资产状况等多方面进行分析以确定是否与特定对象开展业务。

而互联网时代，信用评估的新思路开始出现，即基于大数据的信用评估体系，其基本思想是认为在能够获取的大量数据中尽可能地挖掘出有用的信用信息。与传统信用评分体系相比，大数据的信用评分模型具有极大的优势能够高效率地处理大量数据，不仅能够大大提高信用评分模型的准确率，而且减少了人工的干预，最为关键的是，新的信用评分模型能够极大提高对数据的分析能力。在伴随着市场需求的增长和业务模式的发展下，银行与一些独立的第三方机构开始了信用评分模型的研究，建立属于自己的信用评分模型体系。打破传统思维，利用单一分类模型在一定程度上提高了分类效果，但是单一模型各有优劣，很难做到预测准确性与稳定性的统一，限制了对信用风险可用数据的充分挖掘和提取，而对单一模型之间进行集成，可以相互联系、相互补充，充分提取数据中的相关信息，在一定程度上能够提高个人信用评分模型的分类预测准确率，而且各种组合模型能够降低误差变化率，不仅提高预测性能，还能够改善模型稳定性。而在信用评分模型分类过程中，由于数据自身的限制，正负样本类别差异极大，样本类别不平

衡问题，对模型的性能也有着重要影响，因此为提高分类效果，训练出稳定的信用评分分类模型，本文对单一算法、Bagging 集成学习、异质集成学习、不平衡问题处理进行了探索。

1.1.2 研究意义

信用评分是指基于一定的方法，利用过去的经验，对消费者或企业的信用风险进行量化综合评估。通俗来说个人信用评分[2]就是利用一定的方法进行建模，将影响个人信用状况的相关因素进行组合计算，用一定的分数或其他方式（如好中差、高中低）对个人的信用状况进行描述，预测性地评价个人未来信用状况的一种产品（活动）。传统信用评估，判断条件主要以资深授信人员与主管的经验制定，虽是集合所有专家的意见精华，但基本上还是经验给分，在选择风险因子及设定权重时即使反复摸索修改，也很难确定因子间的关系。随着人工智能的发展，信用评分以科学的方法将风险量化，能够直观清晰地了解一个人的信用状况从而使风险评估有所依据。

信用评分通过利用银行或第三方机构收集到的大量冗余信用数据进行分析，利用人工智能技术，对信贷资源进行整合，采用科学的方法挖掘出能够评估个人或企业未来一段时期内的信用表现的信用行为模型，合理评估未来可能出现的风险，相较于传统的基于个人经验进行的评分方式，利用大数据方式整合出的信用资源更具权威性，具有更高的价值，对社会、企业、个人发展都有重要意义。能够在信息不对称的情况下，进行更为合理的评价，避免道德风险和逆向选择的过度影响，推动信贷资源在社会生产过程中的合理应用，促进经济健康发展。对于企业来讲，通过评估能够对企业自身有一个客观评价，避免出现违约现象积累良好的信用资源，有利于稳定融资来源，降低融资成本，还能够通过良好的信用形象赢得忠实的客户群体，增强企业的竞争力。

现代信息技术的快速发展，使得信用评分过程朝着自动化、高效化的方向前进，信用评分能够快速的以准确、简单、通俗易懂的方式展现个人或企业的信用状况，大大提高了评估效率，降低了评估成本，也为评估单位的经济活动能够提供一定的参考，根据客户信用状况评估其风险等级从而制定精准的授信对策，做出科学的决策。而且个人信用评分的应用，也能够让受评对象及时了解自己的信

用状况减少违约行为的出现，并不断强化自己的信用意识，进而规范自身的行为，构建良好的信用体系。

1.2 信用评分模型研究现状

近年来，随着消费金融的迅速发展，信贷业务遍布我们生活的各个方面，而违约带来的损失，推动政府、第三方机构、企业等不断进行研究和探讨，信用评估产业迅速崛起。随着互联网的发展，人们获取数据也越发地迅速、便捷和多元，同时也引发了我们对数据使用的思考，而在大数据与金融风控、人工智能等技术的支持下，推动信用评分技术不断发展，朝着标准化、集中化和自动化的方向迈进。而对于信用评分模型方向的研究主要有以下几种：（1）单一模型研究；（2）集成模型研究；（3）非平衡问题集成模型研究

1.2.1 单一模型研究

早先利用机器学习进行的信用分析研究主要以单一分类模型为主，之后为提高性能又在单一分类算法的基础上进行改进。如 Dushimimana 等人[3]在以往信用评分技术的基础上创建了一个适用于空中借贷的机器学习分类模型，分析了 3 个月还款期超过 4.1 万名客户的 300 万笔贷款，使用交叉验证方法评估了逻辑回归、决策树和随机森林三种算法的分类性能，结果表明随机森林性能表现最好。吴锦华等人[4]利用证据权重计算每个特征的信息值，从 40 多个变量中选择出 10 个有效的特征，之后选择利用决策树方法构建信用评分模型，在真实数据集上进行分析，为评估人员信用决策分析提供参考依据。Nali 等人[5]为从不同算法中选择出高性能的信用评分模型，通过应用四种不同的分类算法：朴素贝叶斯、决策树、支持向量机、广义线性模型进行实验，使用多种指标进行分析，发现广义线性模型具有很高的预测可信度和准确性，最终选择采用广义线性模型算法建立信用评分模型。Ampountolas 等人[6]研究比较了真实小额贷款数据上的不同的机器学习算法，以测试它们在对借款人分类过程中，各种信贷类别方面的有效性，研究证明随机森林能够很好地完成分类任务。李萌等人[7]选择不良贷款率作为衡量信用风险的重要标准，利用主成分分析从 50 多个财务指标中选出 12 个指标结合

商业银行信用风险评估的 Logit 模型进行实验，通过实证分析证明了 Logit 模型的识别、预测能力。蒲峥屹等[8]首先将信用等级评分的样本集合起来建立支持向量机回归模型，而后使用网格搜索法调整支持向量机回归模型的参数，采用五折交叉验证法以训练集的最小均方根误差作为适应度函数进行参数寻优，并验证其实际的应用价值。邓超等人[9]针对小企业信用评分模型演化过程中出现的样本选择偏差问题，引入拒绝推论的思想，利用贝叶斯界定折叠法，选择小企业金融调研数据进行深入研究，实证结果表明贝叶斯界定折叠法能够处理因样本选择性偏差导致的参数估计有偏问题，对样本填补率和模型分类能力均有较大贡献，能够提升模型的预测性能。Li 等人[10]将 Lasso 技术引入个人信用评估，建立了 Lasso-logistic，Lasso-svm 和 Grouplasso-logistic 模型，并进行变量选择与参数估计，从某个贷款平台上提供的数据进行实验得出结论，与全变量 Logistic 模型和逐步 Logistic 模型相比，Grouplasso-logistic 模型的变量选择能力最强，其次为 Lasso-logistic、Lasso-SVM。Chen 等人[11]将逻辑回归算法与加权证据相结合，构建了一种新的信用评分模型，通过经济活动中存在的关系，利用证据权重中的相关正交变换，进一步剖析各经济活动之间的联系，以提高模型的准确性。Munkhdalai 等人[12]提出了一种新的部分可解释自适应 softmax（PIA-Soft）回归模型，以实现预测性能与输入输出之间的边缘解释，通过神经网络增强 softmax 回归，使其适用于每个借款人，该模型主要由两个部分组成：线性 softmax 回归和非线性（神经网络），线性部分解释了输入和输出变量之间的基本关系，非线性部分通过识别每个借款人的特征间的非线性关系来提高预测性能，实验结果表明，该模型不仅优于其他机器学习方法，而且还显示了与现实世界逻辑相关的解释。Wang 等人[13]通过对银行的信用卡数据进行离散化，计算证据权重、信息值和信息散度来选择特征，然后使用逻辑回归进行预测，最后将逻辑回归结果可视化，建立信用评分模型，经验证该模型具有良好的预测效果。Dumitrescu 等人[14]提出了一种高性能和可解释的信用评分方法，称为逻辑树回归（PLTR），通过使用决策树中的信息来提高逻辑回归的性能，从形式上来讲，使用原始预测变量构建的各种短深度决策树中提取的规则被用作惩罚逻辑回归模型的预测因子，PLTR 允许我们捕捉信用评分数据中可能出现的非线性效应，同时保留逻辑回归的内在可解释性，实证应用表明，PLTR 预测信用风险的准确性明显高于逻辑回归与随机森林。Ruiz

等人[15]提出了基于智能手机获取的非传统数据的信用评分模型，用于贷款分类过程，使用逻辑回归（LR）和支持向量机（SVM），对创建布尔指标的训练数据集的转换与使用证据权重 WEE 的分类进行了比较，发现该模型优于传统人工评估方式，能够提高审批率，降低逾期率。Yu 等人[16]将模糊集理论（FST）和近端支持向量机（PSVM）相结合，构建了一种基于信用风险分析的双加权模糊近端支持向量机（FPSVM）模型，该模型在 PSVM 模型的目标函数和约束条件中加入模糊隶属度，充分利用了数据信息，实验结果表明，所提出的 FPSVM 模型优于本研究中列出的其他 SVM 模型。

而从特征选择的角度，研究者也进行了不同的研究。Tripathi 等[17]针对数据可能具有冗余和不相关的信息和特征，研究提出一种基于特征聚类的方法，将选定特征的数据集使用五个机器学习分类算法作为基分类器进行预测，通过加权投票的方法输出最终结果。卞凌志等人[18]借鉴残差学习思想，利用级联残差森林模型提高特征提取的多样性，建立了一种增强的多维多粒度级联森林的模型，实验对比发现，该方法相较于 LightGBM、XGBoost 模型其 AUC 与准确率有了一定的提高。Chornous 等人[19]分析了特征选择在信用评分数据挖掘建模中的重要性，展示了数据预处理、特征创建和特征选择的过程，通过使用 IBM SPSS Modeler 中的节点应用于二进制分类问题的实际业务情况，结果表明应用混合特征选择模型，可以选择出最佳的特征数，有助于提高信用评分模型的分类准确性。Sang 等人[20]使用特征选择方法，构建了一个基于并行的梯度提升、过滤器和包装器的混合信用评分模型，从输入特征对申请人进行信用评分，特征评分表达式由特征重要性（基尼系数）和信息值结合而成，后向顺序方案用于选择相关特征的最佳子集，利用 GBM 分类器评估子集以减少模型运行时间，分析发现对于某些特定数据集，所提出的方法比单一的基分类器方法显示出更高的预测精度。Laborda 等人[21]利用逻辑回归、K 近邻、支持向量机与随机森林，以确定候选对象违约情况，为了缓解分类算法的维数灾难中的过度拟合，使用了三种不同的特征选择方法：一种过滤方法（卡方检验和相关系数）和两种包装方法（正向逐步选择和反向逐步选择），使用平均绝对误差和所选特征数这两个指标进行分析，结果表明，前向逐步选择在所使用的每一种分类算法中都具有优越的性能。Boughaci 等人[22]提出了一种新的变量选择方法 VS-VNS，该方法基于可变邻域搜索元启发式，允

许我们为数据分类任务选择一组重要变量，然后将 VS-VNS 与贝叶斯网络相结合，为信用评分和选择交易对手建立模型，此外还研究了不同变量集上的六种搜索方法，在一些著名的金融数据集上对不同的技术和组合进行了评估，VS-VNS 展现了一定的优势。Van 等人[23]利用随机森林评估特征的重要性，以最佳平均分、中位数和最低标准差三个方面作为特征评分的规则，去除冗余的特征，选择出最佳特征，通过与其他常用特征选择方法相比，使用该方法选取出的特征进行预测能够获得更高的精度。Trivedi 等人[24]通过使用 3 种不同的特征选择技术（例如信息增益、增益比和卡方）和四种不同的机器学习分类算法：朴素贝叶斯、随机森林、决策树、支持向量机，对不同的机器学习分类器之间以及不同的特征选择技术之间进行了比较，通过准确性、F-度量、假阳性率和训练时间进行分析，选择出最佳的特征选择方法与分类算法进行组合。Krishna 等人[25]提出了一种采用自适应差分进化作为包装器的特征子集选择方法，该包装器采用了朴素贝叶斯、逻辑回归、支持向量机、概率神经网络四种独立的分类器，采用马修斯相关系数（MCC）作为适应度函数或评价指标，通过在三个数据集上进行测试，结果表明，提出的方法比文献中的其他标准方法以及差分进化方法具有更好的效果。

对于从数据指标选择的角度进行的研究来说，如刘鹏翔等人[26]参考商业银行对借款人进行信用风险评估使用的指标，通过使用拍拍贷平台上的申请人的相关数据利用多元线性回归模型进行信用风险分析，通过对选择的指标进行回归，分析影响申请人信用风险的主要因素。吴晓昀等人[27]通过选择借款人的学历、年龄、贷款金额、贷款利率等指标来分析 P2P 借款者违约率的影响因素，通过主成分分析减少解释变量的个数，利用逻辑回归模型建立具体的表达式，验证了所提出的猜想。KCA 等[28]利用财务变动指标来建立可靠的信用评分模型，将逻辑回归算法与加权证据结合构建信用评分模型，通过存在于经济活动中的关系利用证据权重中的相关正交变换剖析每个经济活动的联系，提高模型的准确性。陈胜利等人[29]构建了一个三级风险测度评价指标体系，指标体系中包含三个一级指标，七个二级指标，十五个三级指标，采用因子分析法并运用多元线性回归的方法对 50 家网贷平台的相关数据分析验证风险测度评价指标体系模型的有效性，通过实验分析发现借款分散与安全性、知名度、规模、成长能力能够提高风险评价水平，盈利能力的影响力难以确定；借贷余额、代偿金额、累计投资人数、营业时

间、收入、借款申请人数、投资人数、交易金额等 11 个重要指标是影响 P2P 网贷平台风险测度评价的重要因素。

1.2.2 集成模型研究

在机器学习算法分类中，目标是训练一个稳定且各方面性能表现都比较好的模型，但在具体实践中发现，单一模型总会受较多因素影响。于是很多研究者开始进行集成学习的研究，集成学习的思想就是通过对若干个独立的基分类器进行组合，形成一个强学习器。集成模型有两种选择，第一种同质的（基分类器相同）；第二种是异质的（基分类器不同）。

对于同质集成模型（基分类相同）也就是 Bagging 集成模型的研究，Ghodselahi 等人[30]设计了一种信用评分混合模型，将集成学习用于授信决策，混合模型结合聚类和分类技术使用十个支持向量机分类器作为集成模型的基分类器，即使信用评分的微小提高也会显著降低损失，因此在混合模型中应用集成可以带来更好的分类性能。Yao 等人[31]在 SVM 的基础上构建了一种新的混合 RFSVM 集成模型，该模型使用随机森林选择重要变量，并使用集成方法（bagging 和 boosting）聚合单个模型（SVM）作为鲁棒分类器，实验结果表明，RFSVM 在信用评分领域具有广阔的应用前景。Luo 等人[32]建立了一种混合的信用评分方法与机器学习中的四种分类算法进行比较，使用一个 10 年期间的大型信用违约互换数据集来构建分类器并测试其性能，研究结果表明，Bagging 集成方法可以显著改善个体基分类器性能，如决策树、多层感知器和 k-最近邻分类性能显著提高。

对于异质集成模型的研究，Xu 等人[33]建立一个基于广义模糊软集（GFSS）理论的混合数据挖掘模型，该模型使用基于自适应弹性网络的特征选择算法来消除不相关或弱相关的变量，选择合适的变量后将其应用到所提出的集成模型中，利用 GFSS 理论对集成模型中的每个个体根据其绩效建立新的权重分配机制。王宝等人[34]为提高信用评分预测的泛化能力与准确性，建立了一种基于粗糙集的动态异构集成分类模型，该模型将基于粗糙集衍生的特征依赖度的训练集生成、异构集成模型和动态选择基分类器三个方面结合起来，采用七种广泛使用的性能指标度量模型的性能，实验结果表明，基于粗糙集的动态异构集成分类模型能够取得良好的预测性能。Tripathi 等人[35]为提高信用评分模型的预测性能，结合特

征选择方法构建了一种多层集成分类器框架的混合模型，为解决分类器的放置对集成模型的预测性能的影响，设计了基于 Choquet 积分值的分类器放置算法。Zhang 等人[36]以逻辑回归、支持向量机、神经网络、梯度提升树和随机森林为基础，构建了一种新的分类器选择方法，该方法利用遗传算法选择分类器，同时考虑集合的准确性和多样性，此外为了更好地利用数据模式，该模型将无监督聚类方法与模型分配过程进行结合提高数据的效用。刘传哲等人[37]通过学习传统的信用评估方式，建立了一个适合 P2P 网络借贷环境的动态异质集成模型，该模型具备一定的高维数据分析能力，可以完成对冗余特征变量的筛选，同时通过异质集成结构与动态筛选结合的策略能够实现基础模型权重的自适应调节，从而提高信用评估性能。Xiao 等人[38]针对缺失值构建了一种动态分类器集成选择模型，该模型在训练前不需要对缺失值进行预处理，降低了数据缺失机制对模型性能的影响，能够充分提取出数据中包含的有用信息。Kiziloz 等人[39]提出了一个新的多目标选择模型，该模型动态搜索五个分类器的最佳组合，以提取最佳的代表性特征子集进行预测。Hui 等人[40]针对基分类的选择对模型影响，建立了一种基于聚类的信用评分集成模型，该模型使用聚类算法增强基本分类器之间的多样性，然后选择满足准确性要求的基本分类器来投票决定进行最终决策，实验发现该模型可以显著提高基分类器的选择效率和泛化能力，从而为信用风险管理系统提供参考。Chen 等人[41]提出了一种基于广义 Shapley 值和 Choquet 积分的非均匀系统模型，该模型首先使用模糊度量来表达基础学习者任意两个联盟之间的交互特征，基于精度和多样性目标函数建立模糊测度的线性规划模型；为了在训练阶段尽可能地保留原始信息，采用正态模糊数来表示学习者的基本预测值，又定义了广义 Shapley-Choquet 积分（GSCI）聚合算子，计算集合模型的综合预测值基于定义的聚合算子和线性规划模型，提出了一种用于集成信用评分的 GSCI 方法，通过在多个数据集上进行实验，证明了基于 GSCI 集成模型的适应性和模型性能。Parvin 等人[42]选择不同的分类器进行信用评分，通过指标评估每个模型的性能，从而找出预测信用评分的最佳分类器，实验结果表明，随机森林和 Extratree 分类器模型在集成分类器中具有更好的精度，而支持向量机则在基分类器中具有更好的精度。Koutanaei 等人[43]针对多种特征选择算法与分类算法，建立了一个多阶段的混合数据挖掘模型，第一阶段主要进行数据收集与数据预处理相关操作；第

二阶段选择主成分分析、遗传算法、信息增益比和地势属性评价函数四种特征选择算法，利用支持向量机的分类精度为每种特征选择方法选择合适的集成模型；第三阶段将四种特征选择方法筛选出的特征在所有模型上分别进行实验；最终结果表明，主成分分析法选择的特征与人工神经网络自适应提升具有更高的分类精度。Lahmiri 等人[44]评估现有的先进的集成学习和分类系统在企业破产预测和信用评分方面的相对性能，考虑的集成系统包括 AdaBoost、LoditBoost、USBoost、Subspace 和 Bagging；实验所用的三个数据集：一个由定量属性组成、一个包含定性属性，另一个结合了定量和定性属性，通过十折交叉验证方法验证，结果表明 AdaBoost 在分类错误率低、复杂度有限、数据处理时间短等方面是有效的；集成分类系统的性能优于数据库上验证的现有模型。Li 等人[45]对 Rf、AdaBoost、XGBoost、LightGBM 和 Stacking 集成算法的性能进行评估，评估指标选择精度（ACC）、ROC 曲线下面积（AUC）、Kolmogorov–Smirnov 统计（KS）、Brier 评分（BS）和模型运行时间，此外与五种流行的基分类器（神经网络、决策树、逻辑回归、朴素贝叶斯和支持向量机）进行对比实验，发现 AdaBoost 评估优于个体学习者，Random forest 在五个指标上的总体表现最好，XGBoost 和 LightGBM 次之。Tran 等人[46]提出了转换和平衡数据集的方法，然后探索经典分类模型的性能,最后使用集成学习，分别将 XGBoost、RandomForest 和 Logistic Regression 的投票权重值设置为 2、1、1，在一个台湾数据集上进行验证，实验结果表明，集成学习方法非常有前途。Abellán 等人[47]扩展了以前关于选择信用数据集集合中使用的最佳基分类器的工作，结果表明，基于不精确概率和不确定性度量的非常简单的基分类器，在准确度和 AUC 指标上取得了较好的表现，在这种情况下，如果不同类型的错误具有不同的成本或后果，AUC 度量可以被视为更合适的度量，而单一分类单作为集成中的基分类器，个体性能也是集成模型选择基分类的关键点。Feng 等人[48]提出了一种新的基于软概率的信用评分动态集成分类方法，该方法根据分类器的分类能力以及验证集中 I 类型错误和 II 类型错误的相对成本来选择分类器，根据分类结果对测试集中的样本组合不同的分类器，使用软概率得到默认区间概率，通过使用 10 个真实数据集和 7 个性能指标，该方法与一些著名的个体分类器和集成分类方法（包括 5 个选择性集成）进行比较。Xia 等人[49]提出了一种新的异构集成信用模型，该模型集成了 Bagging 算法和 Stacking

算法，该模型在三个方面不同于现有的集成模型，即集合生成、基分类器选择和可训练的融合器；通过四个流行的评估指标：准确性、AUC、AUC-Hmeasure 和 Breier_score 衡量模型的性能，为了验证所提出的 Bstacking 方法的有效性，与单个分类器、同质基础模型和异质基础模型进行了分析。

1.2.3 不平衡问题集成模型研究

数据不平衡问题是指数据中各个类别的样本量极不平衡，从而导致在训练过程中模型偏重多数类别的样本忽视少数类别的样本，使得模型的分类性能下降。在实际应用过程中，数据中类别比例失衡是客观存在的，如在金融风控领域中，好、坏用户的样本类别比例分布一般差异极大，不平衡问题的处理对模型性能有着较大的影响，因此为提高模型性能，不同研究者对不平衡问题也进行了不同角度的研究。如李京泰等人[50]针对不平衡数据进行分类，提出了基于代价敏感激活函数的 XGBoost 算法，通过引入代价敏感激活函数改变样本在不同预测结果下损失函数的梯度变化，解决被误分类的少数类样本因梯度变化小而无法在 XGBoost 迭代过程中有效分类问题。Zhang 等人[51]针对数据分类不平衡问题，提出了一种新的异构集成信用评分模型，该模型以 LSVM、KNN、MDA、DT、LR 等五个标准分类器为基础，根据数据分布投票选择 AUC 最高的基分类器，然后将所用基分类器集成，得到预测结果。Wang 等人[52]针对类别不平衡问题通常会降低大多数标准分类器的分类性能问题，通过考虑测试数据与训练之间的平均欧几里得距离，利用最先进的自适应功能，提出了一种自适应集成方法，其中平均距离是通过 k 近邻算法来计算的，实验结果表明，自适应集成方法比现有的整体方法具有更好的性能。陈启伟等人[53]针对信用评分过程中样本类别不平衡与代价敏感问题，构建了一种基于 Ext-GBDT 集成的类别不平衡信用评分模型，通过欠采样方法对数据进行不平衡处理，利用采样得到的不同训练子集结合特征采样和参数扰动的方法进行训练，从而得到多个差异化的 Ext-GBDT 子模型，然后使用简单平均法整合子模型的预测概率，根据阈值将预测概率转化为分类结果。Engelmann 等人[54]提出了一种基于条件 Wasserstein-GAN 的过采样方法，通过增设辅助分类器损失关注下游的分类任务，可以有效地对含有数值和分类变量的表格数据进行建模，将该方法与标准过采样方法在七个真实数据集上的不平衡处理

结果进行对比，证明了 GAN-based 过采样方法的可行性。Junior 等人[55]评估了动态选择技术对信用评分问题的适用性，提出了一种减少少数 k-最近邻（RMKNN）方法，该方法是一种提高定义不平衡信用评分数据集动态选择技术局部区域的技术，与现有技术相比该方法在不平衡信用评分数据集上的预测性能具有一定的优越性，而且该方法不需要任何预处理或采样方法来生成动态选择数据集。向欣等人[56]针对数据样本类别不平衡和代价敏感问题，提出一种基于 DESMID-AD 动态选择的信用评估集成模型，根据每一个测试样本的特点动态地选择合适的基分类器，为提高模型对信用差的客户（小类）的识别能力，使用过采样方法对训练数据进行类别平衡，采用元学习的方式基于多个指标进行基分类器的性能评估并在此阶段设计权重机制增强小类的影响。Adolfo 等人[57]研究了不平衡数据集预处理对关联分类器性能的影响，参考信用评分的 13 个不平衡数据集，评估了 6 种欠采样算法和 4 种混合算法，然后分析四种关联分类器的影响；研究发现带平移的混合关联分类器、扩展伽马关联分类器和天真关联分类器使用信用平衡的采用算法并不能提高其性能，采用过采样和混合算法对最小归一化差联想记忆分类器则进行了优化。Petrides 等人[58]评估了成本敏感学习方法，针对错误成本研究出一种新的估算方法，并对信用局信用评分常用的数据缺失处理方法进行了评估，研究发现，表现最好的成本敏感型模型能够提高三种业务渠道的利润率，其中两个渠道受违约率的影响仅有略微提高，另一个渠道的盈利能力提高较为明显。He 等人[59]根据信用评分数据不平衡问题，扩展了 BalanceCascade 方法，根据训练数据的不平衡性生成可调整的数据子集，减少数据不平衡性的影响，提高预测模型的综合性能。罗雅晨等人[60]为解决数据样本比例失衡问题，通过欠采样方法抽取多个比例下的样本子集，再融合随机子空间添加属性扰动，利用随机森林构建集成的分类器模型进行预测，通过在拍拍贷上的真实借贷数据集对机器学习单模型、集成模型、平衡的集成模型三种类型的模型进行对比实验，验证了不平衡处理后的集成模型的有效性。黄静等人[61]将半监督学习技术与 Bagging 集成模型相结合，在类别分布不平衡环境下，构建了一种基于 Bagging 的半监督集成模型（SSEBI）,将有标签样本与无标签样本结合来提高模型的性能，该模型分为三个阶段：第一阶段首先从无类别标签数据集中选择性标记一部分样本，训练多个基

分类器；第二阶段利用测试样本测试模型性能；第三阶段是对分类结果进行集成得到最终结果。

从以上文献来看，对于个人信用评分模型的研究，从单一模型研究、同质集成模型、异质基础模型的研究都取得了显著成效，但忽视了信用评分数据的特点，数据中样本类别不平衡问题，虽然对不平衡问题有一定的研究，但仍需深入，寻找普遍适用的方法，解决现实信用评分业务中样本类别不平衡问题，提高分类准确率，有效识别用户降低潜在的信用风险。

1.3 研究内容及结构

1.3.1 研究内容

为提高信用评分模型的模型预测准确率与稳定性，本文主要进行的研究内容如下：

(1)单一模型研究

对于单一模型本文选取了逻辑回归（LR）、朴素贝叶斯（NB）、决策树（DT）、支持向量机（SVM）四种在信用评分分类预测性能较好的模型。在使用单一模型进行预测之前，首先对数据集进行预处理，了解数据集的大小与原始特征维度，进行探索性分析，之后对数据中的缺失值进行处理，类别型特征、时间格式特征进行转换，对异常值进行删除等一系列处理后，利用随机森林进行特征筛选，特性变量筛选一般选取 10-20 个左右，探索各变量之间的相关程度，判断各变量之间是否存在多重共线性，去除相关性过高的变量；选取特征后，通过 LR、NB、DT、SVM 进行使用，以 Accurary、Precision、F1-score、Recall、AUC 等指标来评估模型性能。

(2)集成模型研究

集成模型以两种形式进行实验对比，一是同质集成，也就是 Bagging 集成。分别对 LR、NB、DT、SVM 四种单一模型进行 Bagging 集成，对比同质集成后的分类效果是否有所提高。之后又以 LR、NB、DT、SVM 四种分类算法为基分类器，构建一种新的异质集成模型，该模型基于 bootstrap 进行抽样，构建

数据子集，通过模型分别进行训练，之后自适应投票选择 AUC 最高的基分类器，然后将所有基分类器集成，得到预测结果。

(3)基于改进 BalanceCascade 方法的信用评分集成模型研究

信用评分所用数据集大都属于极度不平衡，而不平衡问题对最终的预测有一定的影响，基于此，本文提出了基于改进 BalanceCascade 方法的多层信用评分集成模型，改进的 BalanceCascade 方通过抽取正类样本与负类样本组成平衡数据集训练 Adaboost 分类器，将分类错误率控制在一定范围内，确保移除正类样本的准确性；之后根据正负样本的不平衡比例，设置一个可调参数（参数设置为 1，2，3，4，5），通过不断移除一定比例的正样本，使得剩余正负样本比例接近此参数，构建不同正负样本比例下的数据集，寻找最优的比例参数结合新的分层模型进行训练。

1.3.2 创新点

(1)本文利用随机森林进行特征选择，与金融风控中常用的信息值的特征选择方法相比，避免了对每个特征的分箱操作，可直接获得特征重要性排序，实现更为简单、选择特征的速度更加高效；选择信用评估分类预测性能较好的四种单一模型、分别进行 Bagging 集成、并以这四种单一模型为基分类器，自适应选择 AUC 较高的分类器，构建新的异质集成模型，对不同类型模型的性能进行总体分析。

(2)从不平衡问题角度出发，提出了一种改进的 BalanceCascade 处理方法，该方法通过抽取正类样本与负类样本组成平衡数据集训练 Adaboost 分类器，确保移除正类样本的准确性；之后根据正负样本的不平衡比例，增加一个可调参数，通过不断移除一定比例的正样本，使得剩余正负样本比例接近此参数，对不同正负样本比例下的数据集结合新的分层模型进行训练，寻找最优的比例参数。通过与单一模型、同质、异质集成模型进行对比，证明了该模型的有效性。

1.3.3 论文研究结构图

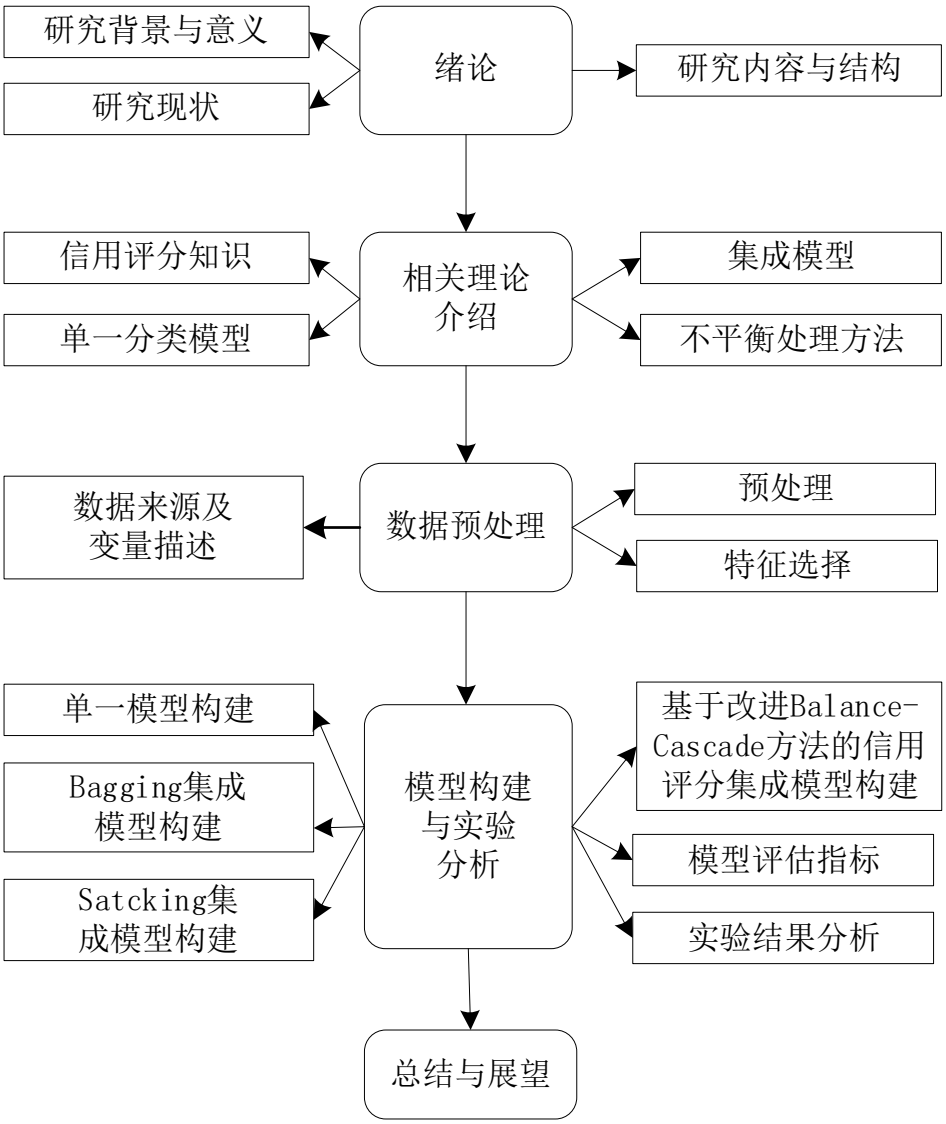

图 1.1 论文结构图

1.3.4 论文结构

第一章：绪论。本章主要介绍了论文研究背景及意义，并对现有信用评分模型研究方法进行了综述，然后论述了研究内容，并对研究路线进行了说明。

第二章：信用评估相关理论概述。本章主要对信用评分过程中涉及到的相关理论进行了阐述，为下文的实验研究提供理论支持。

第三章：数据预处理。对数据来源及变量进行介绍，分析特征类型，之后对数据中存在的缺失值、异常值进行处理，并对特征进行转换，利用随机森林进行特征的重要性度量，选择重要性较高的特征。首先人工选择，删除无实际参考意义的特征，然后利用随机森林对剩余特征进行筛选，进行相关性分析，对相关性大于 0.8 的特征进行选择，计算每个特征的重要性，并按降序排序，选择特征重要性值大于 0.01 的作为特征进行训练，最终选择 27 个特征作为入模变量。

第四章：信用评分模型构建与对比分析。以 LR、NB、DT、SVM 四种模型进行实验，对四种单一算法分别进行 Bagging 集成、并以这四种单一模型为基分类器，构建一种新的异质集成模型，对不同类型模型的性能进行总体分析。由于考虑到不平衡问题对信用评分模型的影响，为减少误分类成本，提出了一种改进的 BalanceCascade 方法的多层信用评分集成模型，通过改进的 BalanceCascade 进行不平衡处理，选用 RF 与 XGBoost 作为第二层基分类器，之后将 RF 与 XGBoost 的输出结果作为新的特征，送入下一层模型，通过逻辑回归进行训练，构建多层信用评分模型，与以上实验结果进行对比，验证模型的有效性。

第五章：总结与展望。本章对全文进行了总结，分析了新构建模型的优势以及存在的问题，并对未来可能的研究方向进行了探索。

2 信用评分相关理论介绍

2.1 信用评分知识

2.1.1 信用与个人信用

信用一词自古就有。中国古语称：人之道德，有诚笃不欺，有约必践，夙为人所信任者，为之信用。不同的角度有着不同的内涵界定，从道德层面来看，信用主要是社会活动主体，在各种经济活动过程中所坚守的诚信守约行为；从法律的角度来讲，主要是指参与者之间根据协议或约定涉及到的权利与义务之间的相互关系，并且不能够及时践行，都会考虑信用问题；经济学上的信用指的是授信人对其经济活动对象充分信任的基础上，通过签订契约向申请者发放贷款资源，并且尽量保障在获得一定收益的基础上使本金也能够回流。

个人信用则是指通过特定协议，或者是基于信任的基础上使得自然人无需立刻付现便可以获得所需的商品或服务，不仅可以用作个人消费用途的信用交易，也可用其他经济活动中，如个人投资、生产经营等。个人信用主要有两种表现形式：一是个人消费信用，二是个人经营信用；随着社会的快速发展，如今可以说是网络经济和信用经济时代，个人信用与人们的日常生活、生产经营等多方面息息相关，如房贷、车贷、住宿，出行、企业融资等，信用可以说是一个人的“经济身份证”，越来越为人所关注。

2.1.2 个人信用评分

个人信用评分是通过使用机器学习相关技术和统计分析方法，对个人的历史信用行为数据进行分析，从中提取出能够反映出未来一段时间内个人信用状况的相关行为特征，利用相关技术将相关特征信息转化为能够未来代表个人信用风险的刻度值。主要应用在信用卡发放审核、贷前审批等授信额度小、成本高的领域，可以直观反映个人信用的状况。其类别按照数据来源可分为三种：信用局信用评分、行业共享评分、定制评分；按预测可分为：风险评分、收益评分、流失倾向评分、转账倾向评分、循环信贷倾向评分、欺诈评分。

信用评分模型主要可分为四个类别：A 卡（申请评分卡），根据客户申请融资类业务时提交的数据进行评级；B 卡（行为评分卡）则是在贷中业务经营期间对存量客户进行风控管理，预测未来一段时间客户可能出现的违约现象；C 卡（催收评分卡）适用于贷后的催收管理；F 卡（反欺诈评分卡）主要应用于一些融资类业务中可能存在的欺诈行为的检测。由于金融业的发展，人们对贷款的需求越来越多，通过建立个人信用评分模型进行评估，能给为信贷管理人员贷前评估提供重要的参考依据，对申请者进行精准划分，选取优质客户，避免坏账的出现，通过事前控制降低风险、控制成本，实现消费信贷业务的高效益。

2.2 单一分类模型介绍

2.2.1 逻辑回归

逻辑回归(Logistic Regression,LR)是信用评分中最常用的经典技术，它被用于计算客户违约概率，与其他回归模型类似，LR 模型利用一些特性分类,然而不同于其他线性回归模型的是，LR 模型是一个非线性模型通过引入分对数函数和 LR 模型的因变量使 LR 具有可以获得实例默认概率和不需要正态分布输入数据的优点。然后根据预先设定的截断点和概率来预测实例的类别。例如，当截断点设置为 0.5 时，如果概率大于 0.5，该实例将被划分为“坏用户”，否则将被划分为“好用户”。

对于分类问题，我们考虑将线性回归的输出与分类任务的真实标签 y 联系起来，即映射函数。我们采用一个 sigmoid 函数（也叫对数几率）

$$f(z)=\frac{1}{1+e^{-z}} \tag{2.1}$$

公式表达，也就是在 sigmoid 函数内嵌套一个线性回归：

$$f(x)=\frac{1}{1+e^{-(w^{T}x+b)}} \tag{2.2}$$

再将其变换得到逻辑回归的另一种常见形式：

$$\ln\frac{p(y=1|x)}{1-p(y=1|x)}=w^{T}x+b \tag{2.3}$$

右边就是线性回归，而左边则引入了几率公式的概念，即事件发生概率相对于不发生概率的比值。可以得到正负样例的概率表达式：

$$p(y=1|x)=\frac{e^{w^T x+b}}{1+e^{w^T x+b}} \tag{2.4}$$

$$p(y=0|x)=\frac{1}{1+e^{w^T x+b}} \tag{2.5}$$

逻辑回归数学原理和求解过程易于理解，实现较为便捷，计算占用内存低，而且可以通过对特征进行离散化，每个变量有相应的权重相当于引入非线性，处理非线性问题，具有更好的解释性，是一个非常强大的分类器。

2.2.2 朴素贝叶斯

朴素贝叶斯分类(Naive Bayesian,NB)朴素贝叶斯分类器假设特征向量的分量之间是相互独立的，给定样本的特征向量 x，该样本属于某一类c_i的概率为:

$$p(y=c_i|x)=\frac{p(y=c_i)p(x|y=c_i)}{p(x)} \tag{2.6}$$

假设特性向量各个分量相互独立，则有：

$$p(y=c_i|x)=\frac{p(y=c_i)\prod_{j=1}^{n}p(x_j|y=c_i)}{Z} \tag{2.7}$$

训练流程如下：首先获取训练样本，之后对每个样本类别出现的概率进行计算，以及在每个类别条件下的每个属性值出现的概率，然后计算每个属性组合属于每个类别的概率，最终选择最大的概率值作为输出结果。

对于不同类型的数据，朴素贝叶斯的训练过程有所不同，若在训练过程中对预测速度要求比较高，对于给定的训练集，可以把朴素贝叶斯涉及的各种概率估值预先计算好保存起来，以便在做出预测时可以随时查询并进行判断；若训练数据变化频繁，则采用懒惰学习方式，事先不进行任何操作，等到待收到预测请求时再针对当前数据集进行概率估计。

2.2.3 决策树

决策树(Decision Tree,DT)是机器学习中常用的分类算法。有两种类型：一种是分类树，另一种是回归树，在我们的研究中使用的是分类树。分类树是一种基于训练数据集构建的树状模型，用于预测新实例的标签。分类树是由节点和有向边组成的树形图，通常只有一个根节点，其他的是内部节点和叶节点。分类树从一个根节点开始，然后根据所选特征的一定规则将训练数据集分成两个子集。决策树算法搜索所有可能的特征分割，以最低的总体错误率为目标找到最优的特征。其优点是可解释性好，易于实现。典型的决策树生成算法有：ID3、C4.5 和 CART。

ID3:根据信息熵衡量数据信息量的大小，利用信息增益来进行特征划分，计算数据集中各特征的信息增益，选择信息增益最大的作为树的根节点，通过递归的方式继续构建子节点，其中熵、条件熵、信息增益计算公式如下：

$$H(D)=-\sum_{i} p_i \log_2 p_i \tag{2.8}$$

$$H(D|A)=\sum_{i}^{n} p_i H(D|A=a_i) \tag{2.9}$$

$$\text{gain}(D,A)=H(D)-H(D|A) \tag{2.10}$$

其中 $H(D)$ 为信息熵； $H(D|A)$ 为条件熵，A 特征属性， a_i 特征 A 的所有取值； $\text{gain}(D|A)$ 为信息增益。

C4.5 在 ID3 的基础上进行了改进，由于数据中离散属性是个别的，于是对连续属性进行离散化，利用信息增益率作为选择节点的标准，避免了 ID3 算法倾向取值多的属性问题，并且进行了相应的决策树剪枝处理，提高决策树的训练速度与分类准确性。

CART 树：又称分类回归树，既能是分类是也可以是回归树，根据目标任务来决定，分类时采用 GINI 作为选择节点的依据，一般来说基尼系数越小，选择的特征越好，其剪枝策略与 C4.5 也不同，通过代价复杂度进行剪枝，建立决策树。

2.2.4 支持向量机

支持向量机(Support Vector Machine,SVM)是一种二分类模型性，其目标是寻找能够对所有数据正确分类的几何间隔最大的分类超平面，尽量能够正确地分类每一个样本，并且使离超平面最近的样本的距离尽可能地远。假设方程$w^T x + b$=0，首先要保证每个样本都被正确分类，对于正样本$w^T x + b \geq 0$，对于负样本$w^T x + b < 0$,其次要求超平面离两类样本的距离 d 尽可能大。

$$d = \frac{\left| w^T x_i + b \right|}{\|w\|} \tag{2.11}$$

对于线性不可分问题，需要加入松弛变量 C(当 C 无穷大时，意味着可以严格分类，当 C 很小时意味着可以有错误容忍)。

2.3 集成模型介绍

2.3.1 随机森林

随机森林是由多棵决策树组成的集成学习算法。用多棵决策树联合预测提高模型的精度。简单来说就是通过 Bootstrap 进行多次抽样，利用多棵决策树训练，对所有分类结果进行投票，将得票最多的结果作为最终预测结果输出。

其训练流程如下：

表 2.1 随机森林算法流程

随机森林算法：
输入样本集 $D = \{(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)\}$，弱分类器迭代次数 T
输出分类器 $f(x)$
（1）对于 t=1，2……T
对训练集进行 t 次随机采样，共采集 m 次，得到包含 m 个样本的采样集 D_t
用采样集 D_t 训练第 t 个决策树模型 $G_t(x)$
（2）对于分类算法预测，则 T 个弱学习器投票最终类别

2.3.2 Xgboost

Xgboost[62]是基于梯度提升树的一种集成算法，梯度提升树是让新的基模型，拟合前面模型的偏差使加法模型的偏差降低。XGBoost 的基本思想和 GBDT 相同，也是利用前向分步算法进行学习，但有一定的区别，XGBoost 添加正则项来控制模型的复杂度，利用二阶泰勒展开去除常数项优化损失函数。

XGBoost 学习主要是定义目标函数与优化目标函数。目标函数由损失函数与正则项组成，假设第 i 棵树，第 i 个样本第 t 次的预测值为 $\hat{y}_i^{(t)}$，则模型的预测值为：$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$ 进一步，可以得到原始目标函数：$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{t} \Omega(f_j)$，之后泰勒公式对目标函数近似展开，目标函数展开公式为如下：$Obj^{(t)} \approx \sum [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + c$。

而 XGBoost 如何求解最优目标？假设树的表达式 $f(x)$，模型优化的过程实际上是求解参数的过程，所以通过对树模型参数化来进行优化。树的参数化分为树模型参数化与树的复杂度参数化两种。树的复杂度的主要内容是确定每棵树的叶子节点的个数与叶子节点的权重向量两部分。树的训练过程中，重要问题就是如何选择出最好的节点划分结构，而节点选择的优劣关乎到树的模型结构的优劣，对此常用节点选择方法有基本精确的贪心算法、近似算法、带权重的分位数草图。

2.3.3 Bagging 集成模型

Bagging 集成模型通过并行化的方式，将若干个弱分类器进行组合训练。其主要思想是基于自助采样法(bootstrap sampling)进行有放回抽样，经过 n 次随机采样操作，我们得到含 n 个样本的采样集，之后样本集进行预测，需要解决的是分类问题，那么我们可以对前面得到的 n 个模型采用投票的方式得到分类的结果，对于回归问题来说，我们可以采用计算模型均值的方法来作为最终预测的结果。

在本文中采用 bagging 方法中分别对 LR、NB、DT、SVM 五种单一模型，通过 boostrap 进行抽样训练，对比集成后的分类效果是否有所提高。

2.3.4 Stacking 集成模型

Stacking 又称堆叠法，实际上是一种分层模型集成结构。以两层为例进行介绍，第一层以五折交叉验证划分训练集，训练第一层的基模型；然后第一层模型的预测结果组成新的数据集，当作新的特征输入第二层模型进行训练；第二层模型为避免过拟合一般不应太过复杂，金融风控过程中一般选取逻辑回归作为次级模型。

Stacking 集成了不同的算法，充分利用不同算法从不同的数据空间角度和数据结构角度的对数据观测，来取长补短，优化结果。所以尽可能地选择合适的基模型，这样可能会使集成结果会更加稳健，更加精确。

2.4 不平衡处理方法介绍

一、抽样方法。抽样方法一般是通过一系列采样算法来调整原始数据中不同类别的样本数量，解决数据中样本类别不平衡问题。常用方法主要有欠采样与过采样。欠采样方法是从多类别样本出发，去除训练集内一定数量的多数样本，使得两类数据样本量大致均衡，然后再进行学习，该方法有一定的局限性，若在极度不平衡条件下，删除大量多数类样本，将会丢失大量有用信息，影响分类器的泛化能力。过采样方法是从少数类样本出发，对训练集内的少数样本进行扩充，通过增加一定数量的少数类样本使样本类别接近均衡，然后再进行学习，该方法简单粗暴地复制少数样本，虽然引入了额外的训练数据，但只是单纯地辅助并没有给少数类样本增加新的信息，容易造成过拟合现象。

二、代价敏感学习。代价敏感学习是根据不同类别的样本被错误划分将会产生不同的成本，从而让机器学习模型进行学习的一种方法。在通常的学习任务中，所有样本的权重一般都是相等的，但在信用评分过程中，数据存在严重的不平衡，将一个违约用户当作不违约用户所造成的损失远远大于将一个不违约用户划分到为违约用户中去所造成的损失，因此可以为坏用户样本设置更高的学习权重，从而让算法更加专注于坏用户的分类情况。但准确确定误分类成本不仅需要充足的经验，而且确定代价参数的确定需要耗费大量时间学习，同时数据内在特征也为该方法的应用带来巨大挑战。

三、BalanceCascade 方法。BalanceCascade 算法[63]基于 Adaboost 分类器进行训练，其每一轮训练中都使用样本类别均衡的训练集训练，然后使用该分类器对全体多数类进行预测，通过控制分类阈值（即概率超过多少判定为少数类）来控制 FP（False Positive）率，将所有判断正确的类删除，然后进入下一轮迭代继续降低多数类数量。

主要过程如下：

（1） 初始化：迭代次数：i；最大迭代次数：T；H_i 每一次训练得到的分类器，s_i 为 H_i 中弱分类器数量，误报率（把一个多数类的样本分类成少数类）$f=\sqrt[T-1]{\frac{S_{\min}}{S_{\mathrm{maj}}}}$。

（2） 从多数类 S_{maj} 中随机选择一个子集 E，$E=S_{\min}$。

（3） 使用子集 E 和 $S_{\min}$，训练分类器 H_i，$H(x)=\mathrm{sgn}(\sum_{j=1}^{S_i}\alpha_{i,j}h_{i,j}(x)-\theta_i)$。其中 $\alpha_{i,j}$ 为弱分类器权重，θ_i 为阈值

（4） 重新计算误报率，调整阈值

（5） 从 S_{maj} 中删除可以被正确分类的样本

（6） 重复步骤 2-6，直到 i=T

（7） 最后，输出一个集成学习器 $H(x)=\mathrm{sgn}(\sum_{i=1}^{T}\sum_{j=1}^{S_i}\alpha_{i,j}h_{i,j}(x)-\sum_{i=1}^{T}\theta_i)$

2.5 本章小结

本章基于所研究问题，对信用评分过程中涉及到的相关理论知识进行介绍，主要有信用评分相关知识、单一分类算法，Bagging 集成、Stacking 集成、不平衡处理方法等，为下面的研究提供理论基础。

3 数据预处理

3.1 数据来源及变量描述

本文数据来源 https://tianchi.aliyun.com/。一共有 538880 条数据 45 个变量，其中 70%为训练集，30%为测试集。其中“isdefault”为目标变量，其他为特征变量。变量名称，中文解释如下：

表 3.1 变量解释

Field	Description
id	为贷款清单分配的唯一信用证标识
loanAmnt	贷款金额
term	贷款期限
interestRate	贷款利率
installment	分期付款金额
grade	贷款等级
subGrade	贷款等级之子级
employmentTitle	就业职称
employmentLength	就业年限（年）
homeOwnership	借款人在登记时提供的房屋所有权状况
annualIncome	年收入
verificationStatus	验证状态
issueDate	贷款发放的月份
purpose	借款人在贷款申请时的贷款用途类别
postCode	借款人在贷款申请中提供的邮政编码的前 3 位数字
regionCode	地区编码
dti	债务收入比

Field	Description
delinquency_2years	借款人过去 2 年信用档案中逾期 30 天以上的违约事件数
ficoRangeLow	借款人在贷款发放时的 fico 所属的下限范围
ficoRangeHigh	借款人在贷款发放时的 fico 所属的上限范围
openAcc	借款人信用档案中未结信用额度的数量
pubRec	贬损公共记录的数量
pubRecBankruptcies	公开记录清除的数量
revolBal	信贷周转余额合计
revolUtil	循环额度利用率，或借款人使用的相对于所有可用循环信贷的信贷金额
totalAcc	借款人信用档案中当前的信用额度总数
initialListStatus	贷款的初始列表状态
applicationType	表明贷款是个人申请还是与两个共同借款人的联合申请
earliesCreditLine	借款人最早报告的信用额度开立的月份
title	借款人提供的贷款名称
policyCode	公开可用的策略_代码=1 新产品不公开可用的策略_代码=2
n 系列匿名特征	匿名特征 n0-n14，为一些贷款人行为计数特征的处理

其中数值型特征有：loanAmnt、term、interestRate、installment、employmentTitle、homeOwnership、annualIncome、verificationStatus、isDefault、purpose、postCode、regionCode、Dti、delinquency_2years、ficoRangeLow、ficoRangeHigh、openAcc、pubRec、pubRecBankruptcies、revolBal、revolUtil、totalAcc、initialListStatus、applicationType、title、policyCode、n0、n1、n2、n3、n4、n5、n6、n7、n8、n9、n10、n11、n12、n13、n14。

单一值：policyCode

类别型特征：grade、subGrade、employmentLength、issueDate、earliesCreditLine

数值型特征：term、homeOwnership、verificationStatus、isDefuult initialListStatus、applicationType、policyCode、n11、n12

3.2 预处理

首先对缺失值进行处理。在了解数据过程中，发现数据中存在一定的缺失值，缺失值分布如下：

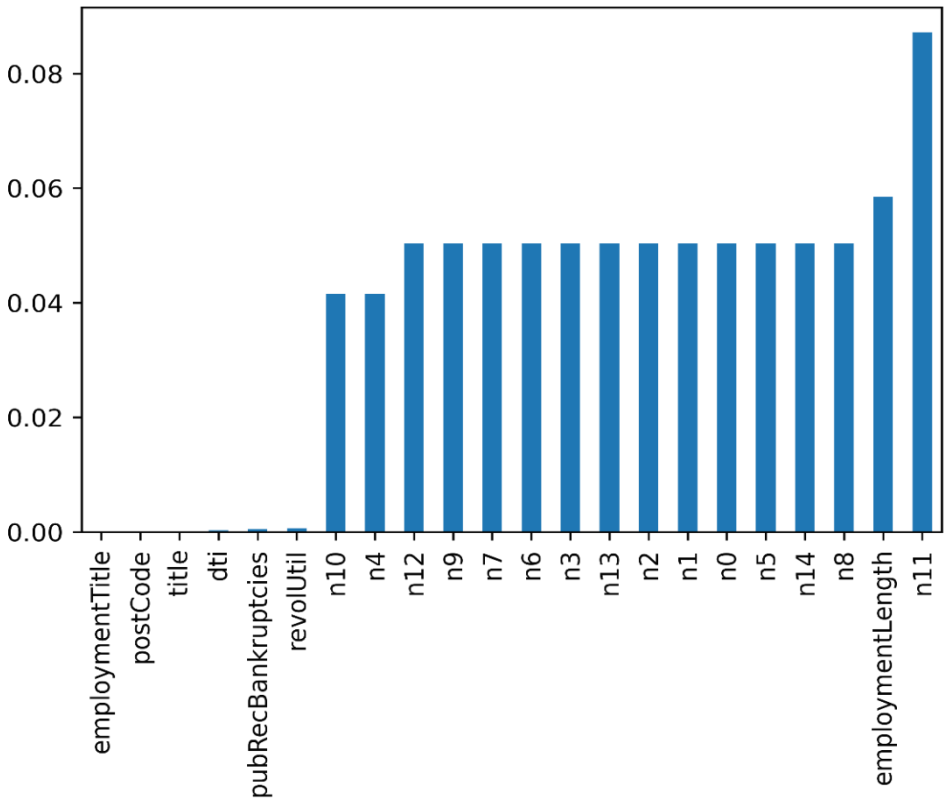

图 3.1 缺失值及缺失率分布图

对缺失值进行了以下处理，employmentTitle、postCode、title 都只有一个缺失值，做删除处理，dti（债务收入比）缺失 300 个；pubRecBankruptcies（公开记录清除的数量）缺失 521 个；revolUtil（循环额度利用率）缺失 658 个，这三个属性本身缺失数量相对总量较少，利用众数进行填充；employmentLength（就业年限）也使用众数方式进行填充；n0 到 n14 每个特征缺失值都比较多，其含义为行为计数，缺失值用 0 填补；对于异常值的处理，一般选择进行删除处理。

其次进行特征转换，类别型特征处理：将 grade、subgrade 等级字母转换成数值；employmentLength 提取数字年份作为数值型特征；issueDate（转变年月舍弃日）,转变为 00.00 格式,并转为数值格式；earliesCreditLine 转变年、月将英文数字化，如（“Jan”，“01”，“Feb”，“02”）。

3.3 特征选择

特征选择是特征工程里的必不可缺的一步，主要是为了去除数据中的冗余特征，寻找最优特征子集，减少训练时间，提高模型精确度；另一方面，选取出真正有用特征，能够协助理解数据产生的过程。

本文选择利用随机森林进行特征选择，是因为随机森林模型在拟合数据后，会对数据属性列，有一个变量重要性的度量，在 sklearn 中即为随机森林模型的 features_importances 参数，这个参数将返回一个 numpy 数组对象，数组里的元素对应为随机森林模型在拟合后认为的所给训练属性列的重要程度，变量重要性度量数组中，数值越大的属性列对于预测的准确性更加重要，而且利用随机森林进行特征选择相较于其他其他方法实现更为简单、对样本的训练速度更加高效。

利用随机森林进行特征选择，首先进行特征重要性度量，具体步骤如下：

表 3.1 随机森林特征选择步骤

具体流程：	
（1）	使用相应的袋外数据来计算每一棵树的袋外数据误差，记为 error1
（2）	然后随机地改变袋外数据特征，再次计算每一棵树的袋外数据误差，记为 error2
（3）	假设森林又 N 棵树，则特征重要性 $= \Sigma(\text{error2}) - \text{error1}/N$, 如果加入随机噪声后，袋外数据准确率大幅度下降（即 error2 上升）则说明这个特征对于样本的分类结果影响很大，也就是说它的重要程度比较
（4）	计算每个特征的重要性，并按降序排序
（5）	根据特征重要性确定要剔除一定比例的特征，得到一个新的特征集

特征选择前，人工判断与目标无关联特征 id 进行删除，以及 postCode、regionCode 特征对预测无实际参考意义进行删除，之后对剩余特征进行筛选，进行相关性分析，对相关性大于 0.8 的特征进行相关性处理。

表 3.2 高相关性特征

feature_x	feature_y	Corr
loanAmnt	installment	0.9534
interestRate	grade	0.9533
ficoRangeLow	ficoRangeHigh	1
openAcc	n7	0.8178
openAcc	n10	0.9839
n1	n2	0.8083
n1	n3	0.8083
n1	n4	0.8268
n1	n9	0.8012
n2	n3	1
n2	n9	0.9816
n3	n9	0.9816
n5	n8	0.8378
n7	n10	0.8271

从表中可以看出，loanAmnt（贷款金额），installment（分期付款金额）两个特征间相关系数为 0.95；intersetRate（利率）与 grade（等级）两个特征间相关系数为 0.95；ficoRangeLow（所属的下限范围），ficoRangeHigh（所属的上限范围）两个特征间相关系数为 1；openAcc（未结信用额度的数量）与 n7、n10 特征的相关系数分别是 0.82、0.98；n1 与 n2、n3、n4、n9 之间的相关系数分别是 0.81、0.81、0.83、0.80；n2 与 n3、n9 之间的相关系数分别是 1、0.98；n3 与 n9 的相关系数为 0.98；n5 与 n8 的相关系数 0.84；n7 与 n10 的相关系数为 0.83。对于这些特征两两之间相关性大于 0.8 的根据其特征重要性大小，选取重要性值大的，删除值较小的，最终选择删除 installment、grade、ficoRangeHigh、n7、n2 n3、n4、n8、n10。

高相关性特征处理后，生成相关性热力图如下：

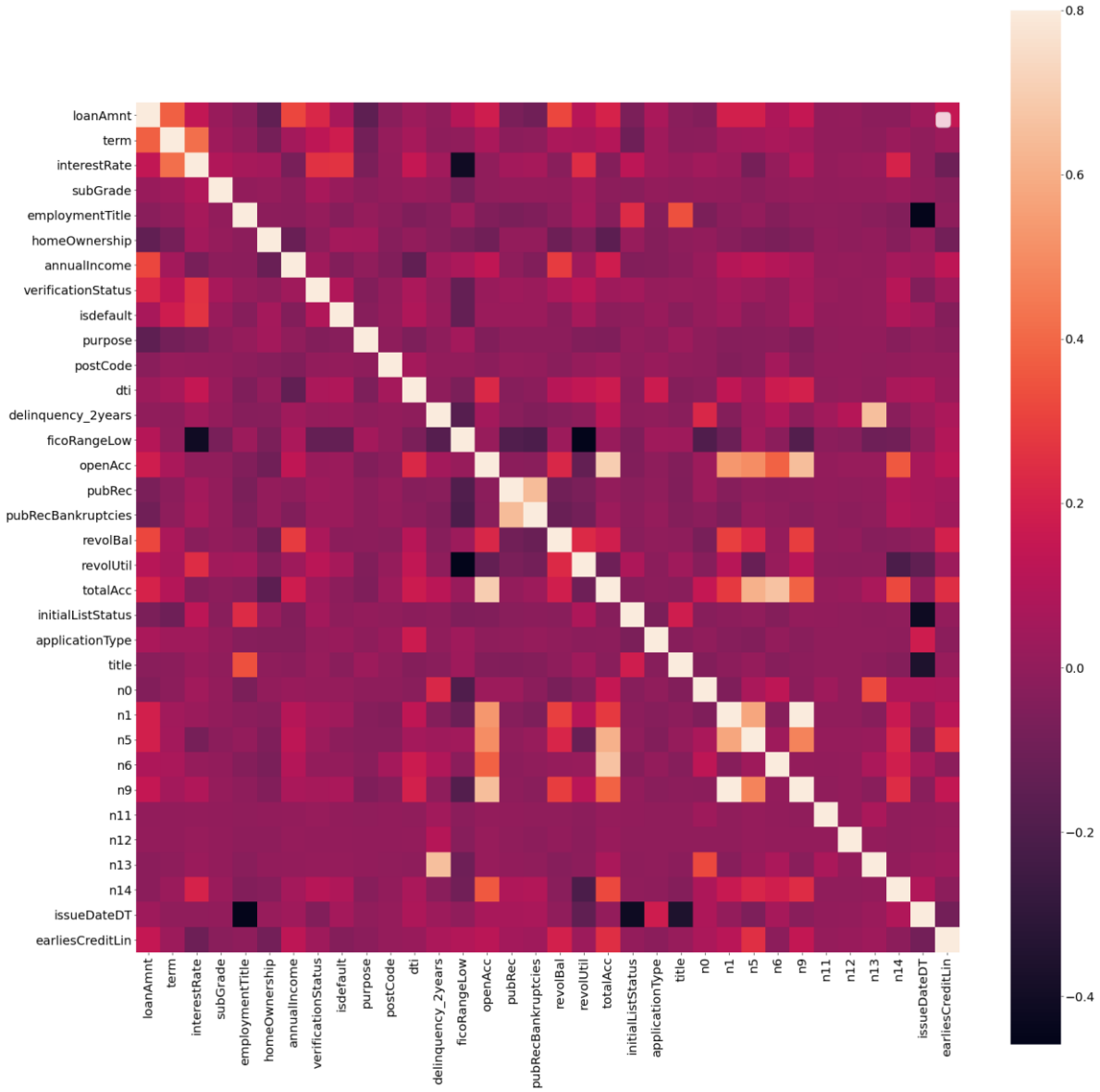

图 3.2 特征相关性热力图

由图 3.2 可知，颜色越深负相关性越强，颜色越浅正相关性越强，在进行高相关性特征处理之后，正相关性在 0.7 左右，负相关性在 0.4 左右，特征之间的正负相关性均低于 0.8，无高相关性特征。之后计算每个特征的重要性，并按降序排序，选择特征重要性值大于 0.1 的作为特征进行训练。选取特征如下：

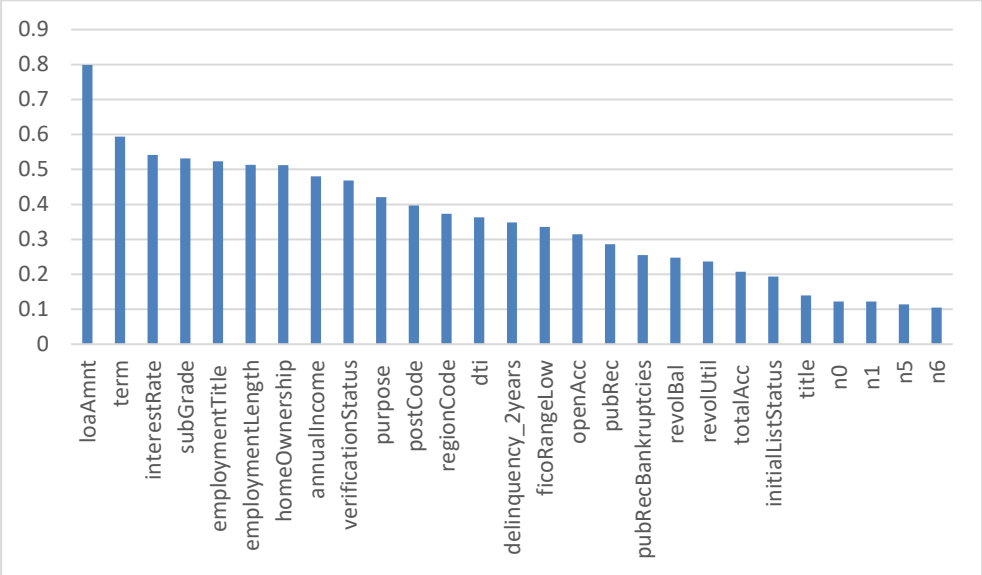

图 3.3 特征重要性排序图

3.4 本章小结

本章主要研究内容为数据预处理相关操作，及随机森林特征选择。首先对数据来源进行介绍、数据相关特征变量进行表述，分析特征类型，之后对数据中存在的缺失值、异常值进行处理，并对特征进行转换。特征选择，首先人工选择，删除无实际参考意义的特征，然后利用随机森林对剩余特征进行筛选，并进行相关性分析，对相关性大于 0.8 的特征进行选择，计算每个特征的重要性，并按降序排序，选择特征重要性值大于 0.1 的作为特征进行训练，最终选择 27 个特征作为入模变量。

4 信用评分模型构建与对比分析

4.1 单一模型构建

对于单一模型本文选取了逻辑回归（LR）、朴素贝叶斯（NB）、决策树（DT）、支持向量机（SVM）四种在信用评分分类预测性能较好的模型。逻辑回归通过使用 sigmoid 函数，将预测映射到[0,1]之间获得概率值，判断类别，建模过程简单易懂，其准确率在实际应用过程中效果较佳，在金融风控分类中认可度较高；朴素贝叶斯算法在属性相关性较小时效果较好且逻辑简单易于实现；决策树通过二叉树的形式对变量进行判断，根据树状分叉的判断规则一层层达到结果，应用在信用评分过程中容易理解判别规则；支持向量机有着严格的数学理论支持，可解释性较强，简化了通常的分类，在信用评分建模中较受欢迎。

将随机森林筛选出的特征作为入模变量，使用逻辑回归、朴素贝叶斯、决策树、支持向量机分别进行训练，最后以准确率（Accurary）、精确率（Precision）召回率（Recall 等）、F1-score、AUC 等指标来评估模型性能。而在实验过程中发现 class_weight（用于标示分类模型中各种类型的权重）这一参数的调整，对单一模型的预测结果有着极大的影响，当 class_weight={0:0.2,1:0.8}时，也就是说这样不违约样本类别（0）的权重为 20%，而违约样本类别 1 的权重为 80%时，单一模型的预测结果可得到有效改善。由此可以看出样本类别的不平衡性对模型的性能影响。

4.2 Bagging 集成模型构建

在本文中采用 bagging 方法中分别对 LR、NB、DT、SVM 四种单一模型进行集成，通过 boostrap 方法进行抽样训练，每次随机抽取 m（m=1000）个样本，共抽取 n 次，得到 n 个数据集，共得到 n（n=100）个模型的预测结果，采用投票方式输出最终结果。

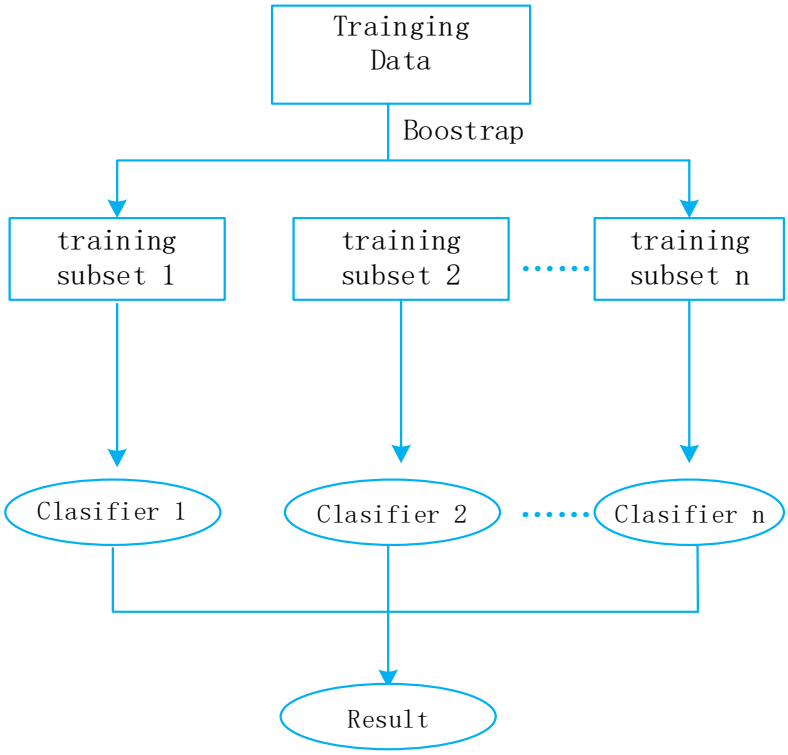

图 4.1 Bagging 集成模型图

4.3 异质集成模型构建

由于不同分类算法分类效果有所不同，为对比不同集成模型在信用评分中的优劣，本文以最开始选择的逻辑回归（LR）、朴素贝叶斯（NB）、决策树（DT）、支持向量机（SVM）四种分类算法为基分类器，通过 bootstrap 进行抽样，构建数据子集，通过模型分别进行训练，之后自适应选择 AUC 最高的基分类器，然后将所有基分类器集成。

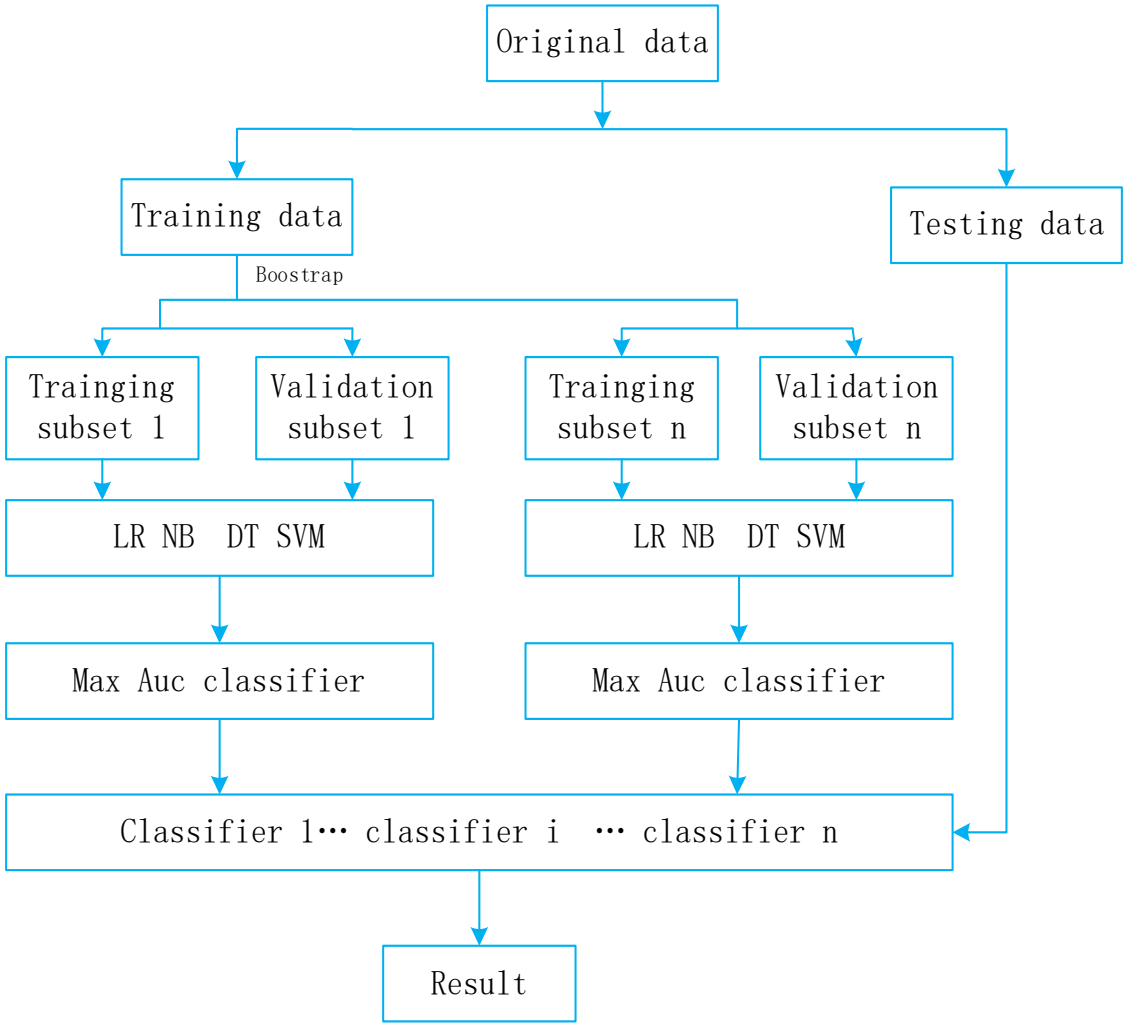

图 4.2 异质集成模型

表 4.1 异质集成模型训练过程

其训练过程如下：
（1）从训练数据集 $S=\{(x_i,y_i),i=1,2...n\}$，中，随机挑选 n 个样本作为训练集，记为 $S^a=\{(x_i^a,y_i^a),i=1,2,...n\}$，剩余的样本作为验证集，记为 $S^b=\{(x_i^b,y_i^b),i=1,2...n^b\}$，其中 y_i 为类别标签
（2）利用 S^a 分别训练 LR、NB、DT、SVM 四个模型，然后分别使用 4 个模型预测验证数据集 S^b 的标签
$T_{LR}=\text{train}(LR,x^a,y^a),yp_{LR}^b=\text{validation}(T_{LR},x^b)$

$$T_{NB}=train(NB,x^a,y^a),yp_{NB}^b=validation(T_{NB},x^b)$$

$$T_{DT}=train(DT,x^a,y^a),yp_{DT}^b=validation(T_{DT},x^b)$$

$$T_{SVM}=train(SVM,x^a,y^a),yp_{SVM}^b=validation(T_{SVM},x^b)$$

（3）比较验证数据集的预测标签与真实标签，计算每个模型在 S^b 数据集上的 AUC 值

$$AUC_{LR}=AUC(yp_{LR},y^b) \qquad AUC_{NB}=AUC(yp_{NB},y^b)$$

$$AUC_{DT}=AUC(yp_{DT},y^b) \qquad AUC_{SVM}=AUC(yp_{SVM},y^b)$$

（4）选择 AUC 值最大的模型进行集成，利用测试集进行测试

4.4 基于改进 BalanceCascade 方法的信用评分集成模型构建

4.4.1 改进的 BalanceCascade 方法介绍

BalanceCascade 主要思想是在每一轮训练时，对两类别样本数量均衡的数据集进行训练，然后利用训练好的分类器，对全体多数类进行预测，通过控制分类阈值来控制 FP 率，将所有判断正确的类从正样本中移除，错误分类的样本放回到原样本空间中，然后进入下一轮迭代继续降低多数类数量。改进 BalanceCascade 方法通过抽取正类样本与负类样本构成平衡数据集训练 Ada-boost 分类器，将分类错误率控制在一定范围内，确保移除正类样本的准确性；之后根据正负样本的不平衡比例，设置一个可调参数，通过不断移除一定比例的正样本，调整正负样本比例，避免获得非代表性数据。具体流程如下：

（1） 输入数据集 D，数据集中包含不违约 P,和违约样本 N

（2） 从正类样本中随机选取，得到样本P_i，然后与负类样本组成一个新的数据子集$S_i = P_i \cup N$

（3） 然后利用新的数据子集，训练一个 Adaboost 分类器C_i，C_i是一些不同权重弱分类器的集合，设置阈值将分类器集成的误报率控制在一定范围内（$f=_{T-1}\sqrt{\frac{N}{P}}$，T 为抽取子集的次数）

（4） 移除一定比例且属于P_i中不违约的样本R_i（移除比例为 0.1）

（5） 判断 P/N 是否接近设置的不平衡比率（不平衡率设置为 1、2、3、4、5），若是保存数据子集，若否返回第 2 步继续循环

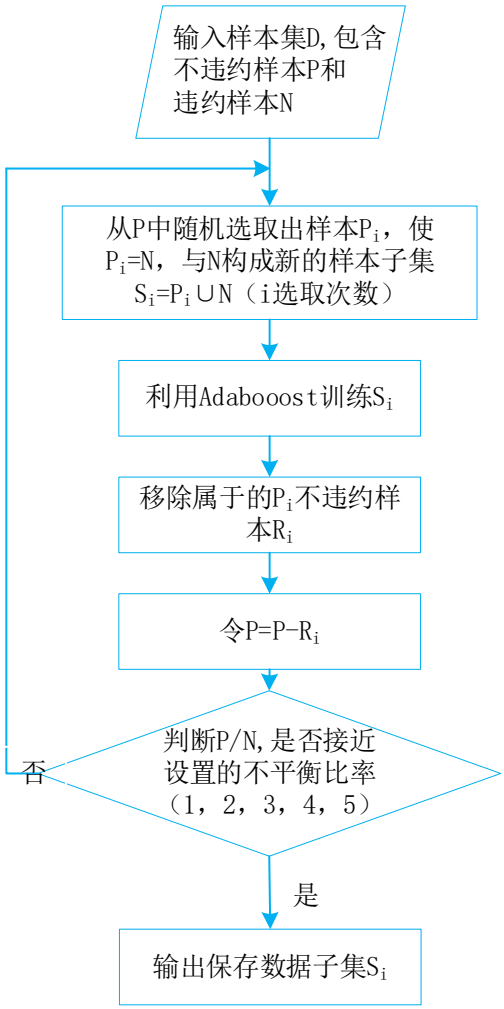

图 4.3 改进 BalanceCascade 方法的处理流程

4.4.2 基于改进 BalanceCascade 方法的信用评分集成模型构建

该模型首先进行使用改进的 BalanceCascade 方法进行不平衡处理，其中数据正负样本比例大致为 8：1，从正负样本随机抽取样本P_i，抽取数量与负样本数量相同，与负样本组成新的数据子集S_i，利用 Adaboost 进行训练，分类错误率为 f，根据分类结果移除一定比例的分类正确的正样本（移除比例为 0.1），判断移除样本之后的正负样本比例是否接近设置的不平衡比率（1，2，3，4，5），若否，则继续进行迭代，若接近，则保存数据子集进行下一阶段的训练。该阶段以 RF

与 XGBoost 为基分类器，首先采样五折交叉验证划分数据集，四折进行训练，一折进行预测，之后将 RF 与 XGBoost 的预测结果 predict1、predict2 合并为新的数据集作为新的特征输入下一层模型进行训练，该层模型不易太复杂，这样会导致模型在训练集上过拟合，测试集泛化效果差，根据对单一模型的分析，该层模型选用逻辑回归为基分类器。之后进行贝叶斯调参，选取最佳参数后，利用测试集进行测试。

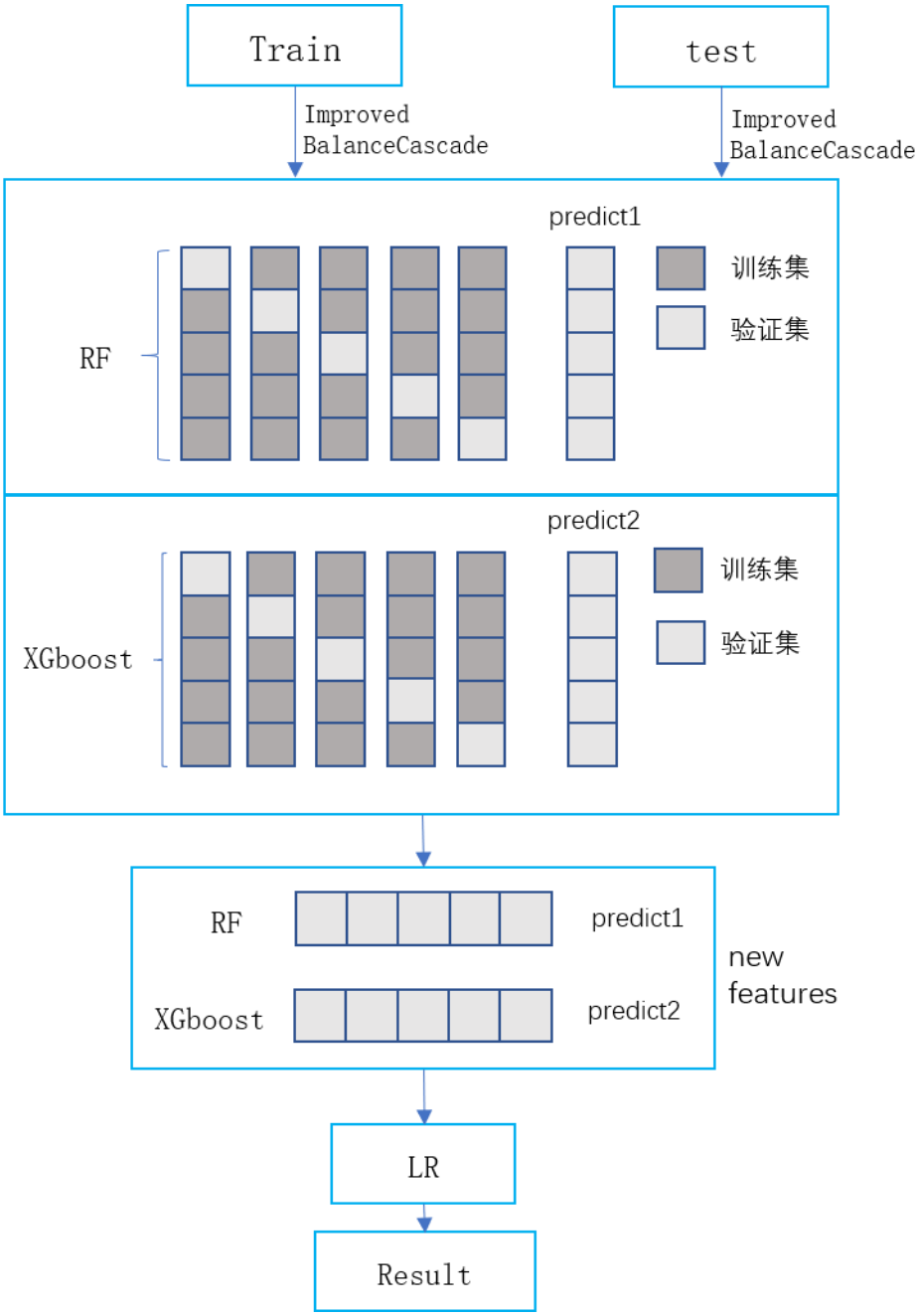

图 4.4 基于改进 BalanceCascade 方法的信用评分集成模型图

4.5 模型评估指标

4.5.1 混淆矩阵

评估指标是检验模型性能优劣的重要参考，如果指标选择的不合理，有可能会得出错误的结论，因此应针对具体的数据、模型选取不同的评价指标进行分析。

混淆矩阵是分类评估中常用的评估指标，在二分类中，可以将样本根据其真实结果和模型的预测结果的组合可分为四类：划分为真正例（true positive，TP）、真反例（true negative，TN）、假正例（false positive，FP）、假反例（false negative，FN）。

表 4.1 分类混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

其中，准确率表示的是分类正确的样本（真正例与真反例）占总样本数的比例，是针对所有样本的统计量。能够清晰地反映出模型的预测性能，准确率虽能反映总体的分类情况，但也有一定的局限性：在正负样本类别数量差异较大时，多数类的样本往往会成为准确率的主要影响因素；金融风控中，在确保不违约样本正确分类的前提下，提高违约样本的分类准确性，准确性也是重要参考。其计算公式为：

$$Accurary = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

精确率又称为查准率，是针对预测结果而言的一个评价指标，是针对为真正例的样本的统计量，在借贷预测中，反映的是模型预测结果为不违约的样本中真正是不违约的样本的比例。其计算公式为：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

召回率：反映的是在实际的正样本的样本量中，模型正确分类的正样本数占实际的正样本量的比例。其计算公式为：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

F1_Score 是精准率和召回率的调和平均值，其计算公式为：

$$F_1\text{-Score} = 2\frac{\text{Precision*Recall}}{\text{Precision+Recall}} \tag{4.4}$$

4.5.2 ROC 曲线

在分类任务中，测试样本通常是用一个概率表示当前样本属于正例的概率，往往会采取一个阈值，大于该阈值的为正例，小于该阈值的为负例。如果降低阈值，分类为正类的样本将会增多，会提高对正类样本的识别率，但模型对负类样本的识别率就会降低。为了形象地描述上述的这种变化，引入 ROC 曲线来评价一个分类器的好坏，中文名为“受试者工作特征曲线”。

ROC 曲线的横坐标为假阳性率（False Positive Rate，FPR），表示的是负样本的预测错误量占所有负样本的比例。其计算公式为：

$$FPR = \frac{FP}{FP + TN} \tag{4.5}$$

纵坐标为真阳性率（True Positive Rate，TPR），表示的是正样本的预测正确量占所有正样本的比例。其计算公式为：

$$TPR = \frac{TP}{TP + FN} \tag{4.6}$$

4.5.3 AUC-ROC 曲线下的面积

AUC 为 ROC 曲线下的那部分面积，这个面积的大小一般是处于 0 到 1 之间，主要用来评估模型性能的好坏，AUC 的值越大，模型性能越好。

AUC=1：完美的分类器，采用该模型，不管设定什么阈值都能得出完美预测（绝大多数时候不存在）；0.5<AUC<1：优于随机猜测，分类器好好设定阈值的

话，有预测价值；AUC=0.5：跟随机猜测一样，模型没有预测价值；AUC<0.5：比随机猜测还差，但是如果反着预测，就优于随机猜测。

4.6 实验结果分析

4.6.1 实验环境

本文实验环境配置如下表：

表 4.2 实验环境配置

实验配置说明	参数
CPU	Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz 3.70 GHz
内存	32.0GB
操作系统	Windows 10 64 位
语言	Python3.8
开发平台环境	Anaconda
开发工具	Spyder、Jupyter Noteboook

4.6.1 单一模型实验结果

首先对逻辑回归（LR）、朴素贝叶斯（NB）、决策树（DT）、支持向量机（SVM）四种单一模型进行实验，在该实验中，实验使用的是阿里云天池竞赛数据集，其中 70%为训练集，30%为测试集。其中“isdefault”为目标变量，其他为特征变量

实验结果如下：

表 4.3 逻辑回归混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约	86214	43046
违约	11979	20425

表 4.4 决策树混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	95822	33438
违约（1）	19635	12769

表 4.5 朴素贝叶斯混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	121046	8214
违约（1）	26078	6326

表 4.6 支持向量机混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	94982	34278
违约（1）	14752	17652

从表 4.3 逻辑回归混淆矩阵中可看出，129260 个不违约用户数据中，正确分类的有 86214 个用户，错误分类的有 43046 个用户；在 32404 个违约用户中，有 20425 个用户分类正确，11979 个用户分类错误，模型对负样本的识别率为 63%。表 4.4 决策树模型，不违约用户数据中，正确分类的有 95822 个用户，错误分类的有 33438 个用户；违约用户数据中，正确分类的 12769 个用户，错误分类的有 19635 个用户，模型对负样本的识别率为 39%。表 4.5 朴素贝叶斯模型，不违约用户数据中，正确分类的有 121046 个用户，错误分类的有 8214 个用户；违约用户中，正确分类的有 6326 个用户，错误分类的有 26078 个用户，模型对违约的识别率仅为 20%。表 4.6 支持向量机中，不违约用户数据中，正确分类的

有 94982 个用户，错误分类的有 34278 个用户；违约用户数据中，正确分类的有 17652 个用户，错误分类的有 14752，模型对违约用户的识别率为 54%。

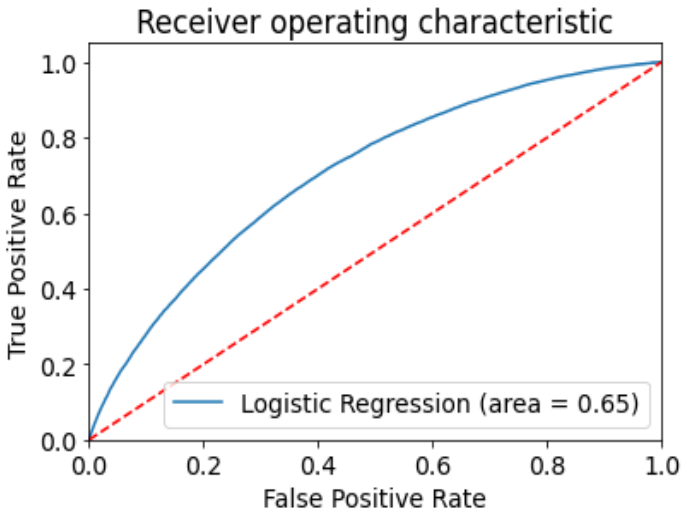

图 4.5 逻辑回归_ROC 曲线图

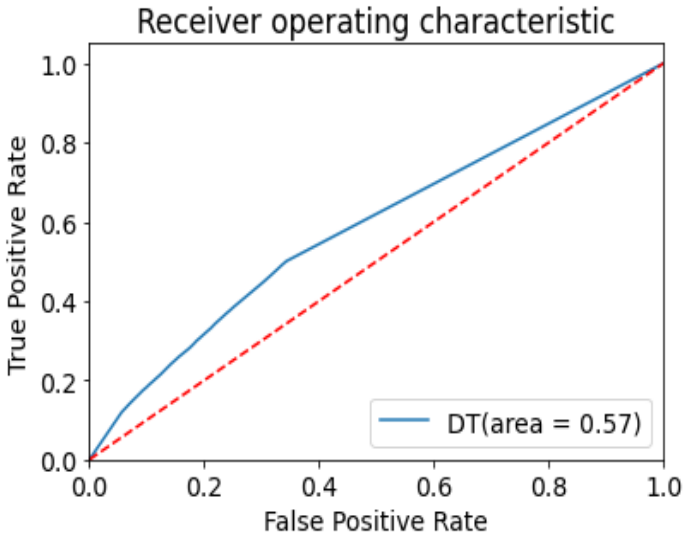

图 4.6 决策树 ROC 曲线图

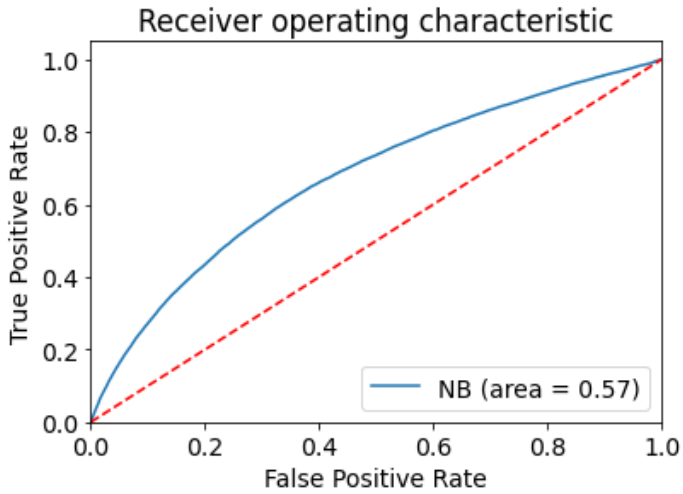

图 4.7 朴素贝叶斯 ROC 曲线图

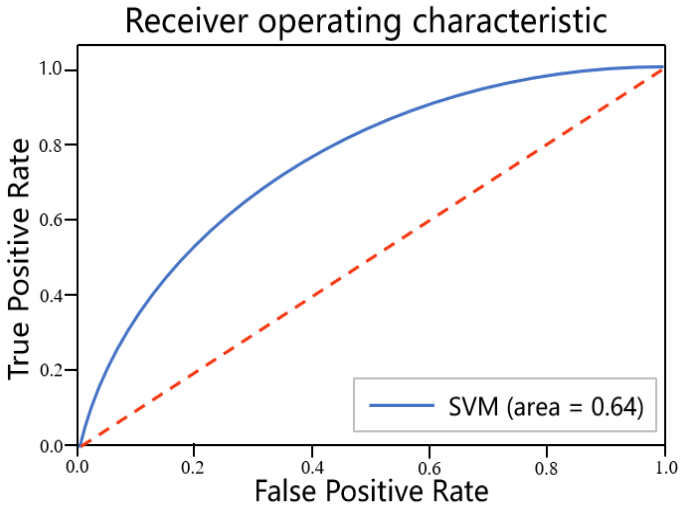

图 4.8 支持向量机 ROC 曲线图

表 4.7 各模型实验结果对比表

Model	Accurary	Precision	Recall	F1-score	AUC
LR	0.66	0.88	0.67	0.76	0.65
DT	0.67	0.83	0.74	0.78	0.57
NB	0.78	0.82	0.94	0.88	0.57
SVM	0.70	0.87	0.74	0.80	0.64

从表 4.7 中可以看出，四种单一模型中，朴素贝叶斯准确率最高为 78%，其次是支持向量机准确率为 70%，决策树准确率为 67%，逻辑回归准确率为 66%。在不为违约用户类别下，逻辑回归与支持向量机的精确率较高，分别为 88%与 87%，决策树与朴素贝叶斯的精确率为 83%与 82%；在召回率指标下，则是朴素贝叶斯较高为 94%，其次是支持向量机与决策树召回率为 74%，逻辑回归召回率为 67%；在 F1 值中，朴素贝叶斯较高为 88%，其次支持向量机为 80%，逻辑回归与决策树相差为 2%。而从模型稳定性角度来说，由 ROC 曲线下面积可看出，逻辑回归的 AUC 值最高为 0.65，其次是支持向量机 AUC 值为 0.64，决策树与朴素贝叶斯 AUC 值相同为 0.57，相比较之下四种单一模型中逻辑回归模型最稳定。

4.6.2 集成模型实验结果

对四种单一算法进行 Bagging 集成，对比同质集成后模型是否能够提高；之后又以四种算法为基分类器，建立异质集成模型，基于 boostrap 进行抽样，构建数据子集，通过模型分别进行训练，自适应选择 AUC 最高的基分类器，然后将所有基分类器集成，进行实验。实验结果如下：

表 4.8 逻辑回归 Bagging 集成混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	112880	16380
违约（1）	21301	11103

表 4.9 决策树 Bagging 集成混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	115205	14055
违约（1）	22667	9737

表 4.10 朴素贝叶斯 Bagging 集成混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	110041	19219
违约（1）	20904	11500

表 4.11 支持向量机 Bagging 集成混淆矩阵表

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	97968	31292
违约（1）	14372	18032

表 4.12 异质集成模型混淆矩阵

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	113385	15875
违约（1）	21807	10597

在集成模型的实验中，首先是四种单一算法进行 Bagging 集成，由表 4.8 中可以看出，LR_Bagging 模型预测在不违约用户数据中，正确分类的有 112880 个用户，错误分类的 16380 个用户，在违约用户数据中，正确分类的有 11103 个用户，错误分类的有 21301 个用户，该模型对违约用户的识别率为 34%。从表 4.9，DT_Bagging 混淆矩阵中可以看出，在不违约用户数据中，正确分类的有 115205 个用户，错误分类的有 14055 个用户，在违约用户数据中，正确分类的有 9737 个用户，错误分类有 22667 个用户，该对违约用户的识别率为 30%。从表 4.10 中，NB_Bagging 混淆矩阵可以看出，不违约用户数据中，正确分类的有 110041 个用户，错误分类的有 19219 个用户，在违约用户数据中，正确分类的有 11500 个用户，错误分类有 20904 个用户，该模型对违约用户的识别率为 36%。从表 4.11

中，SVM_Bagging 混淆矩阵表可以看出，不违约用户数据中，正确分类的有 97968 个用户，错误分类的有 31292 个用户，在违约用户数据中，正确分类的有 18032，错误分类有 14372 个用户，该模型对违约用户的识别率为 56%。而在表 4.12 中，异质集成模型混淆矩阵，不违约用户数据中，正确分类的有 113385 个用户，错误分类的有 15875 个用户，在违约用户数据中，正确分类的有 10597，错误分类有 21807 个用户，异质集成模型对违约用户的识别率为 33%。

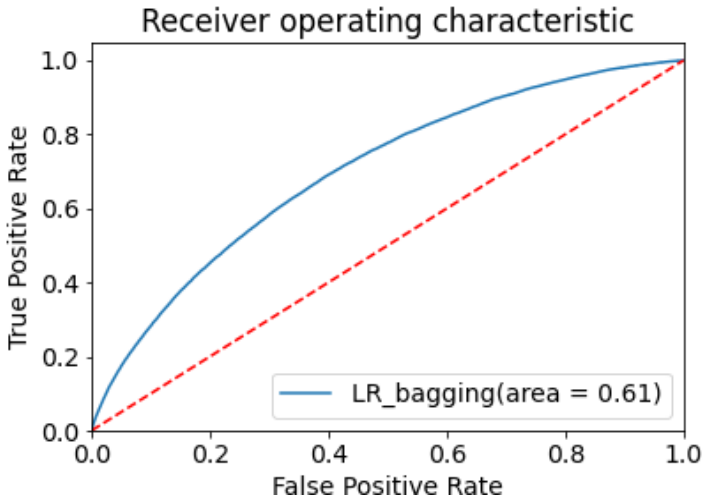

图 4.9 逻辑回归 Bagging 集成 ROC 曲线图

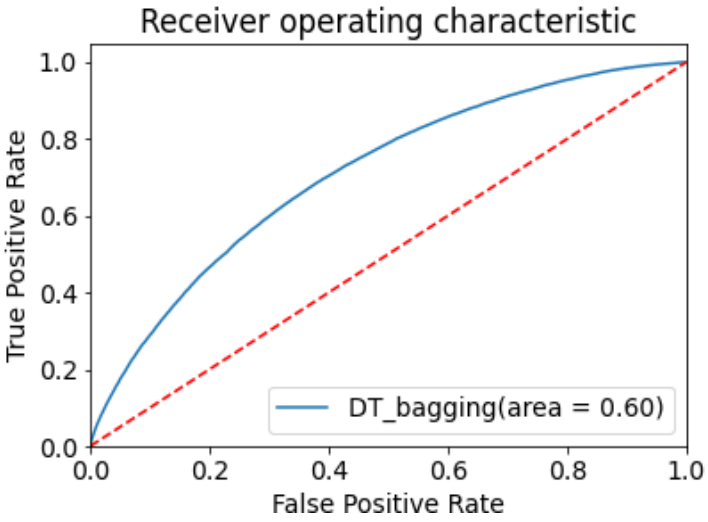

图 4.10 决策树 Bagging 集成 ROC 曲线图

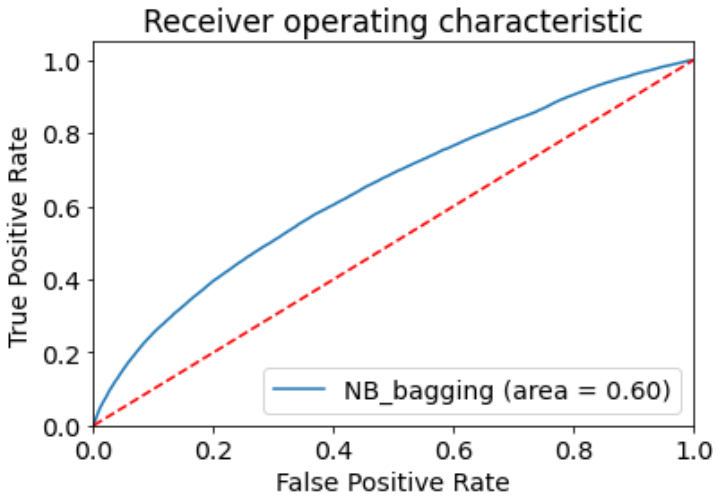

图 4. 11 朴素贝叶斯 Bagging 集成 ROC 曲线图

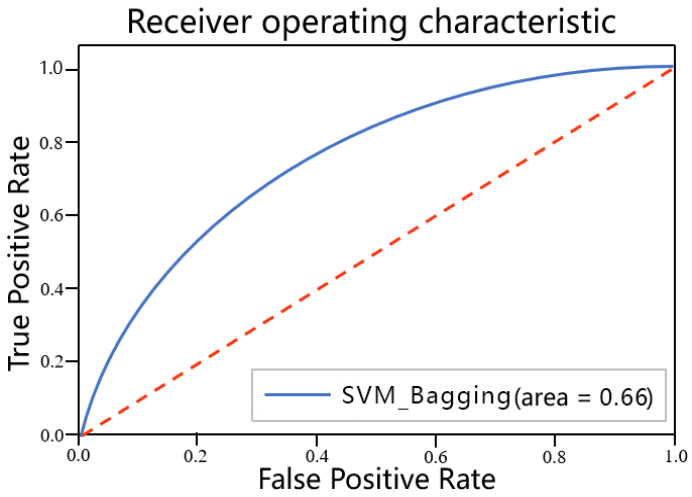

图 4. 12 支持向量机 Bagging 集成 ROC 曲线图

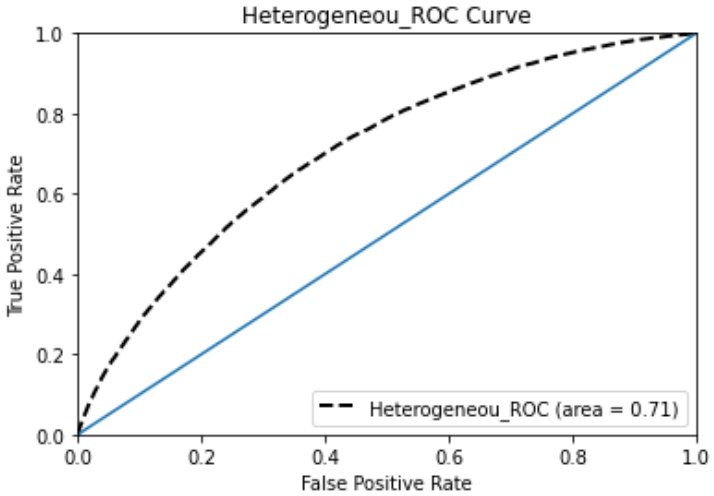

图 4. 13 异质集成模型 ROC 曲线图

表 4.13 集成模型实验结果对比表

Model	Accurary	Precision	Recall	F1-score	AUC
LR_Bagging	0.76	0.84	0.87	0.86	0.61
DT_Bagging	0.77	0.84	0.89	0.86	0.60
NB_Bagging	0.75	0.84	0.85	0.85	0.60
SVM_Bagging	0.72	0.87	0.76	0.81	0.66
heterogeneou _ensemble	0.767	0.84	0.88	0.86	0.71

从表 4.13 中可以看出，在对逻辑回归、决策树、朴素贝叶斯、支持向量机四种单一模型进行 Bagging 集成后，从准确率（Accurary）上来看，DT_Bagging 决策树准确率最高为 77%，其次是 LR_Bagging 准确率为 76%，之后是 NB_Bagging 准确率为 75%，SVM_Bagging 准确率为 72%。在不为违约用户类别下，从精确率（Precision）指标来看，SVM_Bagging 精确率 87%，DT_Bagging 精确率 84%，DT_Bagging 与 NB_Bagging 精确率都为 84%。从召回率（Recall）指标来看，DT_Bagging 召回率为 88%，LR_Bagging 召回率为 87%，NB_Bagging 召回率为 85%，SVM_Bagging 召回率最低为 76%。从 F1 值来看，LR_Bagging 与 DT_Bagging 的 F1 值相同为 86%，LR_Bagging 与 SVM_Bagging 略低。从模型稳定性来看，根据 ROC 曲线下面积可看出，SVM_Bagging 的 AUC 值最高为 0.66，其次是 LR_Bagging，AUC 值为 0.61，DT_Bagging 与 NB_Bagging 的 AUC 值最低为 0.60。

从异质集成模型来说，其准确率 76.7%，在不违约用户类别下，精确率为 84%，召回率为 88%，F1 值为 86%，模型稳定性评估指标 AUC 值 0.71，相比较其他四种 Bagging 集成模型，以逻辑回归、朴素贝叶斯、决策树、支持向量机四种单一算法为基分类器的异质集成模型分类性能较好。

4.6.3 基于改进 BalanceCascade 方法的信用评分集成模型实验结果

改进的 BalanceCascade 方法通过抽取正类样本与负类样本构成平衡数据集训练 Adaboost 分类器，将分类错误率控制在一定范围内，确保移除正类样本的

准确性；之后根据正负样本的不平衡比例，设置一个可调参数，通过不断移除一定比例的正样本，使得剩余正负样本比例接近此参数，对不同正负样本比例下的数据集使用分层模型进行实验。模型选用 RF 与 XGBoost 作为第一层基分类器，将 RF 与 XGBoost 的输出结果作为新的特征，送入下一层模型，通过逻辑回归进行训练。通过实验发现，当正负样本比例为 2 时，模型预测结果最优，实验结果如下：

表 4.14 基于改进 BalanceCascade 方法的信用评分集成模型混淆矩阵

真实情况	预测结果	
	不违约（0）	违约（1）
不违约（0）	108593	20667
违约（1）	11665	20739

从表 4.14 中可以看出，不违约用户的数据中，正确分类的有 108593 个用户，错误分类的 20667 个用户，在违约用户数据中，正确分类的有 20739 个用户，错误分类的有 11665 个用户，该模型对违约用户的识别率为 64%。

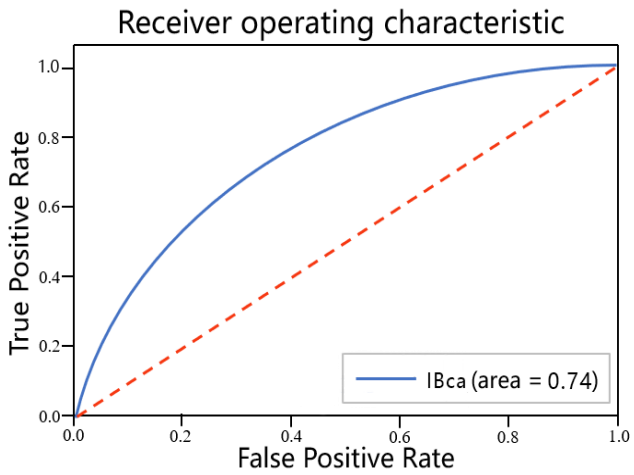

图 4.14 基于改进 BalanceCascade 方法的集成模型 ROC 曲线图

表 4.15 基于改进 BalanceCascade 方法的信用评分集成模型实验结果

Model	Accurary	Precision	Recall	F1-score	AUC
Improved BalanceCascade	0.80	0.90	0.84	0.87	0.74

从表 4.15 中可以看出改进 BalanceCascade 的信用评分集成模型分类准确率 80%，在不违约用户类别下精确率为 90%，召回率为 84%，F1 值为 88%。从模型稳定性指标来看，模型性能较好 AUC 值达到 0.74。

4.6.4 总体比较

表 4.16 总体指标比较图

Model	Accurary	Precision	Recall	F1-score	AUC
LR	0.66	0.88	0.67	0.76	0.65
DT	0.67	0.83	0.74	0.78	0.57
NB	0.78	0.82	0.94	0.88	0.57
SVM	0.70	0.87	0.74	0.80	0.64
LR_Bagging	0.76	0.84	0.87	0.86	0.61
DT_Bagging	0.77	0.84	0.89	0.86	0.60
NB_Bagging	0.75	0.84	0.85	0.85	0.60
SVM_Bagging	0.72	0.87	0.76	0.81	0.66
heterogeneou_ensemble	0.77	0.84	0.88	0.86	0.71
Improved BalanceCascade	0.80	0.90	0.84	0.87	0.74

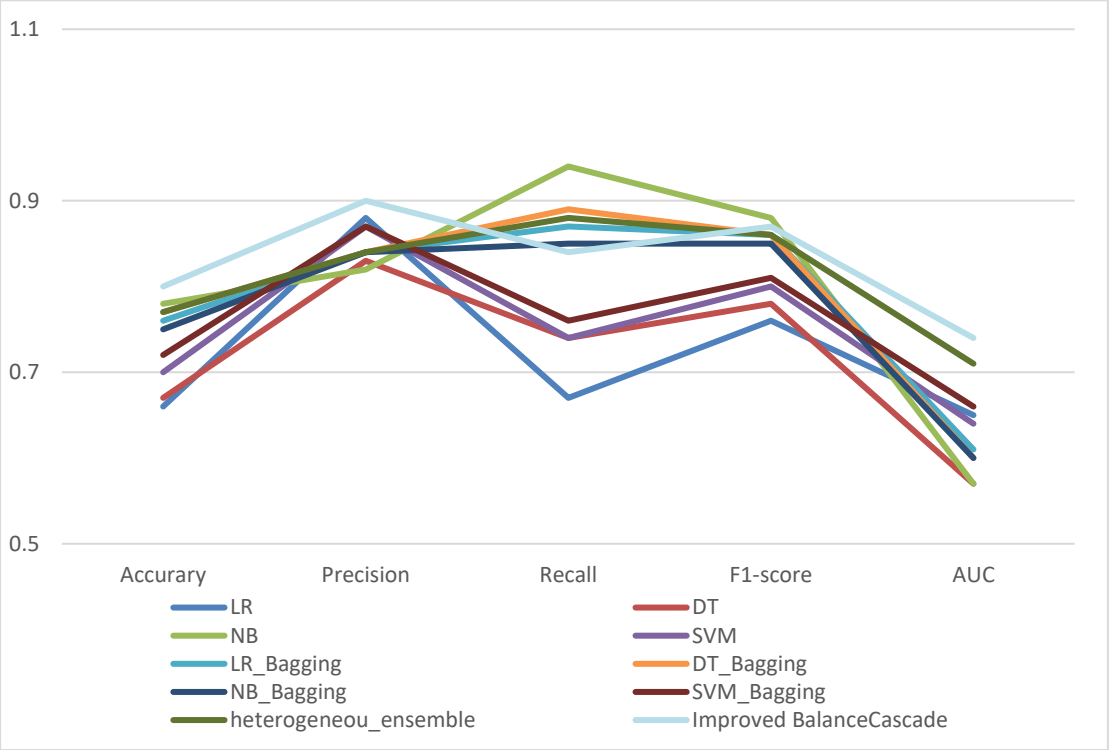

图 4.15 总体指标折线图

从表 4.16 中可以看出，LR、DT、NB、SVM 四种单一模型，LR_Bagging、DT_Bagging、NB_Bagging、SVM_Bagging 及 heterogeneou_ensemble、Improved_BalanceCascade，共 10 种模型，从评估指标准确率（Accurary）、精确率（Precision）召回率（recall）、F1-score、AUC 等五个指标上看，四种单一模型虽训练过程简单、速度较快，但预测准确率与模型稳定性效果并不理想，其中朴素贝叶斯准确率最高为 78%，但 AUC 值却较低为 0.57；而对于 Bagging 集成模型，四种 Bagging 集成模型实验结果有不同程度的改善，其中 LR_Bagging、DT_Bagging 的准确率提升较为明显，提升 10%,但模型稳定性 AUC 值无明显提升，而且对于稳定的分类器，同质集成的改进很小，如 SVM_Bagging 准确率提升了 2%，AUC 值提升了 2%；关于自适应选择 AUC 的异质集成，异质集成进一步增强了集成模型的多样性，五个评估指标的结果有了明显提高，其中 AUC 值相较于四种同质集成模型更优；而改进 BalanceCascade 方法的集成模型，通过对数据进行不平衡处理，以 RF 与 XGBoost 为基分类器，之后将两者的输出作为新的特征通过逻辑回归进行训练，基于改进 BalanceCascade 方法的集成模型在召回率指标低于其他一

些模型，其他指标均优于其他模型，而且对违约样本的识别率也高于其他模型，其准确率为 80%，精确率为 90%，F1-score 值为 87%，AUC 值为 0.74。综上，虽然单一模型得益于其简单性，但分类性能并不理想，bagging 集成虽有所提升但不如异质集成模型具有竞争力。从图 4.15 总体指标折线图，可以清晰看出各模型的指标分布，基于改进 BalanceCascade 方法的信用评分集成模型总体分布优于其他模型，而在信用评分模型分析过程中，分类准确率与模型稳定对在实际应用过程中具有重要的参考意义。综上，可以看出提出的基于改进 BalanceCascade 方法的集成模型在实际应用过程中具有重要的参考意义。

4.7 本章小结

本章主要内容是信用评分模型的构建，并对实验结果进行分析。本文选取 LR、NB、DT、SVM 四种经典的分类模型，建立单一评分模型，对四种单一算法分别进行 Bagging 集成、并以这四种单一模型为基分类器，通过 boostrap 进行抽样，构建数据子集，通过模型分别进行训练，自适应选择 AUC 最高的基分类器，然后将所有基分类器集成构建一种新的异质集成模型，对不同类型模型的性能进行总体分析；然后针对信用评分不平衡问题的影响，为减少误分类成本，提高模型性能，本章在 BalanceCascade 方法的基础上进行改进，构建分层集成模型，提出了基于改进 BalanceCascade 方法的信用评分集成模型，并验证了该模型的有效性。

5 总结与展望

5.1 总结

随着我国消费金融的快速发展，个人信贷业务迅速扩张，随之带来信用风险问题引起了社会各界的广泛关注。本文在国内外学者的基础上，着眼于个人信贷业务进行个人信用评分模型的研究。

在本文中，在第一章节中，主要阐述了信用评分的研究背景及意义，根据国内外研究现状对信用评分模型进行介绍，主要包括单一信用评分模型、集成信用评分模型、非平衡处理方法的集成模型研究。第二章节主要对信用评分相关理论进行了介绍，主要包括信用评分知识、单一模型算法逻辑回归（LR）、朴素贝叶斯（NB）、决策树（DT）、支持向量机（SVM），集成模型（Bagging 集成与 Stacking 集成）、不平衡处理方法等。

第三章节中，主要是数据预处理，数据来源介绍，对数据相关特征变量进行表述，分析特征类型，之后对数据中存在的缺失值、异常值进行处理，并对特征进行转换。特征选择部分，本文利用随机森林进行特征的重要性度量，选择重要性较高的特征。首先人工选择，删除无实际参考意义的特征，然后利用随机森林对剩余特征进行筛选，进行相关性分析，对两两相关性大于 0.8 的特征进行选择，计算每个特征的重要性，并按降序排序，选择特征重要性值大于 0.01 的作为特征进行训练，最终选择 27 个特征作为入模变量。

第四章节中，主要是信用评分模型的构建，本文选取 LR、NB、DT、SVM 四种经典的分类算法，建立单一评分模型，对四种单一算法分别进行 Bagging 集成、并以这四种单一模型为基分类器，通过 bootstrap 进行抽样，构建数据子集，自适应选择 AUC 最高的基分类器，然后将所有基分类器集成构建一种新的异质集成模型；而后针对信用评分不平衡问题的影响，为减少误分类成本，提高模型性能，本章在 BalanceCascade 方法的基础上进行改进，构建分层集成模型，提出了基于改进 BalanceCascade 方法的集成模型。改进的 BalanceCascade 方法主要是通过抽取正类样本与负类样本构成平衡数据集训练 Adaboost 分类器，将分类错误率控制在一定范围内，确保移除正类样本的准确性；根据正负样本比例，设

置一个可调参数，根据参数获得不平衡率在一定范围的平衡数据子集，构建可调整数据子集的集合，避免获得非代表性数据。并对单一模型、Bagging 集成模型、异质集成模型、改进 BalanceCascade 方法的集成模型的实验结果进行了全面的分析，验证了改进的 BalanceCascade 方法的集成模型的有效性。

5.2 展望

虽然本文对四种单一模型、Bagging 集成模型、Stacking 集成模型进行分析，验证了集成模型在信用评分方面的有效性，并且针对不平衡问题，提出了一种改进 BalanceCascade 方法的信用评分集成模型，相比较其他模型有一定的优势，但也存在着一些局限性，之后的研究可以从以下方面进行探索：

（1）不同的特征选择方式，选取的特征不同，对模型最终的预测结果有着不同的影响，而且要考虑到不同变量之间可能会出现衍生变量，衍生变量对模型性能的影响需要进一步探究。

（2）对于集成模型的研究，虽然异质集成模型较 Bagging 集成有很大优势，但本文中的异质集成模型选取四种算法作为基分类器，改进 BalanceCascade 方法的集成模型选取两种算法作为基分类器，但是如何能够高效地选取合适是基分类器，以及基分类器数量的设置，建立模型，之后的研究可以考虑如何根据数据自身特点，动态地选取最佳数量并根据数据自身特征动态地选取相适应的基分类器进行集成。

（3）本文构建的针对不平衡问题的处理方法，虽有一定的成效，但对违约样本的识别率还不是令人满意，可以尝试对多种不平衡处理方法进行探索，尝试改进其他方法，如改进的采样方法、代价敏感学习等，进行实验。

参考文献

[1] 陈彩霞, 石春, 程明雄. 基于FICO信用评分模型的电商小贷信用评价分析研究[J]. 现代商业, 2015(26):2.

[2] 王俊山,王玥.对我国个人信用评分及监管的分析与思考[J].金融发展研究,2021(01):86-89.

[3] Dushimimana B, Wambui Y, Lubega T, et al. Use of machine learning techniques to create a credit score model for airtime loans[J]. Journal of Risk and Financial Management, 2020, 13(8): 180.

[4] 吴锦华,王志生,刘重阳,胡龙彪.基于决策树的用户信用评分模型的构建[J].无线互联科技,2019,16(08):45-46.

[5] Nali J , Martinovic G . Building a Credit Scoring Model Based on Data Mining Approaches[J]. International Journal of Software Engineering and Knowledge Engineering, 2020, 30(2):147-169.

[6] Ampountolas A , Nde T N , Date P , et al. A Machine Learning Approach for Micro-Credit Scoring[J]. Risks, 2021, 9(3):50.

[7] 李萌.Logit 模型在商业银行信用风险评估中的应用研究[J].管理科学,2005(02):33-38.

[8] 蒲峥屹,李云飞.基于网格搜索支持向量机的个人信用等级评分预测[J].市场研究,2020(03):33-36.

[9] 邓超, 胡梅梅, 曾文潮. 基于贝叶斯界定折叠法的小企业信用评分模型研究[J]. 管理工程学报, 2015(4):9.

[10] Li Q. Logistic and SVM credit score models based on lasso variable selection[J]. Journal of Applied Mathematics and Physics, 2019, 7(05): 1131.

[11] Chen K , Yadav A , If A , et al. Credit Fraud Detection Based on Hybrid Credit Scoring Model[J]. Procedia Computer Science, 2020, 167(2020):2-8.

[12] Munkhdalai L , Ryu K H , Namsrai O E , et al. A Partially Interpretable Adaptive Softmax Regression for Credit Scoring[J]. Applied Sciences, 2021, 11(7):3227.

[13] Wang M, Yang H. Research on Customer Credit Scoring Model Based on Bank

Credit Card[C]//International Conference on Intelligent Information Processing. Switzerland: Springer, Cham, 2020: 232-243.

[14] Dumitrescu E, Hue S, Hurlin C, et al. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects[J]. European Journal of Operational Research, 2022, 297(3): 1178-1192.

[15] Ruiz S, Gomes P, Rodrigues L, et al. Credit scoring for microfinance using behavioral data in emerging markets[J]. Intelligent Data Analysis, 2019, 23(6): 1355-1378.

[16] Yu L, Yao X, Zhang X, et al. A novel dual-weighted fuzzy proximal support vector machine with application to credit risk analysis[J]. International Review of Financial Analysis, 2020, 71: 101577.

[17] Tripathi D, Edla D R, Kuppili V, et al. Credit scoring model based on weighted voting and cluster based feature selection[J]. Procedia computer science, 2018, 132: 22-31.

[18] 卞凌志,王直杰.基于增强多维多粒度级联森林的信用评分模型[J].计算机应用,2021,41(09):2539-2544.

[19] Chornous G, Pysanets K, Yakovenko N. A Hybrid Approach for Feature Selection in Data Mining Modeling of Credit Scoring[C]//ICTERI Workshops. Kharkiv, Ukraine: CEUR WS, 2020: 256-269.

[20] Sang H V,Ha N N ,Bao H.A hybrid feature selection method for credit scoring[J]. EAI Endorsed Transactions on Context-aware Systems and Applications, 2017, 4(11):152335.

[21] Laborda J, Ryoo S. Feature Selection in a Credit Scoring Model[J]. Mathematics, 2021, 9(7): 746.

[22] Boughaci D,Alkhawaldeh A . A new variable selection method applied to credit coring[J]. Algorithmic Finance, 2018, 7(2):1-10.

[23] Van Sang H, Nam N H, Nhan N D. A novel credit scoring prediction model based on Feature Selection approach and parallel random forest[J]. Indian Journal of Science and Technology, 2016, 9(20): 1-6.

[24] Trivedi S K. A study on credit scoring modeling with different feature selection and machine learning approaches[J]. Technology in Society, 2020, 63: 101413.

[25] Krishna G J, Ravi V. Feature subset selection using adaptive differential evolution: an application to banking[C]//Proceedings of the ACM India Joint International Conference on Data Science and Management of Data. New York, NY: ACM, 2019: 157-163.

[26] 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.

[27] 吴晓昀. P2P 借款者违约率影响因素分析[D].湖南:湖南大学,2017.

[28] Kca B , Ay B , Akc D , et al. Credit Fraud Detection Based on Hybrid Credit Scoring Model[J]. Procedia Computer Science, 2020, 167:2-8.

[29] 陈胜利, 张璇. P2P 网贷平台风险测度评价[J]. 统计与决策, 2020(19):5.

[30] Ghodselahi A. A hybrid support vector machine ensemble model for credit scoring[J]. International Journal of Computer Applications, 2011, 17(5): 1-5.

[31] Yao J R, Chen J R. A New Hybrid Support Vector Machine Ensemble Classification Model for Credit Scoring[J]. Journal of Information Technology Research (JITR), 2019, 12(1): 77-88.

[32] Luo C. A comparison analysis for credit scoring using bagging ensembles[J]. Expert Systems, 2022, 39(2): e12297.

[33] Xu D , Zhang X , Feng H . Generalized fuzzy soft sets theory‐based novel hybrid ensemble credit scoring model[J]. International Journal of Finance & Economics, 2019, 24.

[34] 王宝,宁连举.基于粗糙集的动态异构集成信用评分模型[J].经济统计学(季刊),2018(02):213-225.

[35] Tripathi D , Edla D R , Cheruku R , et al. A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification[J]. Computational Intelligence, 2019, 35(2):371-394.

[36] Zhang H,He H, Zhang W .Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring[J].Neurocomputing,2018, 316:210-221.

[37]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, 000(009):24-31.

[38]Xiao J , Liu D H , Xin G U , et al. Dynamic classifier ensemble selection model for bank customer's credit scoring[J]. Journal of Management Sciences in China, 2015, 18:114-126.

[39]Kiziloz H E, Deniz A. Feature selection with dynamic classifier ensembles [C]//2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Piscataway, NJ: IEEE, 2020: 2038-2043.

[40]Hui X, Gang Y S. Using clustering-based bagging ensemble for credit scoring[C]//2011 International Conference on Business Management and Electronic Information. Piscataway, NJ: IEEE, 2011, 3: 369-371.

[41]Chen X, Li S, Xu X, et al. A novel GSCI-based ensemble approach for credit scoring[J].IEEE Access, 2020, 8: 222449-222465.

[42]Parvin A S, Saleena B. An ensemble classifier model to predict credit scoring-comparative analysis[C]//2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS). Piscataway, NJ: IEEE, 2020: 27-30.

[43]Koutanaei F N, Sajedi H, Khanbabaei M. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring[J]. Journal of Retailing and Consumer Services, 2015, 27: 11-23.

[44]Lahmiri S, Bekiros S, Giakoumelou A,et al. Performance assessment of ensemble learning systems in financial data classification[J].Intelligent Systems in Accounting, Finance and Management, 2020, 27(1): 3-9.

[45]Li Y, Chen W. A comparative performance assessment of ensemble learning for credit scoring[J]. Mathematics, 2020, 8(10): 1756.

[46]Tran D Q, Nguyen D D, Nguyen H H, et al. An Ensemble Learning Approach for Credit Scoring Problem:A Case Study of Taiwan Default Credit Card Dataset[C].International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences. Switzerland: Springer, Cham, 2021: 283-292.

[47] Abellán J, Castellano J G. A comparative study on base classifiers in ensemble methods for credit scoring[J]. Expert systems with applications, 2017, 73: 1-10.

[48] Feng X, Xiao Z, Zhong B, et al. Dynamic ensemble classification for credit scoring using soft probability[J]. Applied Soft Computing, 2018, 65: 139-151.

[49] Xia Y, Liu C, Da B, et al. A novel heterogeneous ensemble credit scoring model based on bstacking approach[J]. Expert Systems with Applications, 2018, 93: 182-199.

[50] 李京泰,王晓丹.基于代价敏感激活函数 XGBoost 的不平衡数据分类方法[J]. 计算机科学,2022,49(05):135-143.

[51] Zhang T, Chi G. A heterogeneous ensemble credit scoring model based on adaptive classifier selection: An application on imbalanced data[J]. International Journal of Finance & Economics, 2021, 26(3): 4372-4385.

[52] Wang L , Lei Z , Guan G , et al. Adaptive Ensemble Method Based on Spatial Characteristics for Classifying Imbalanced Data[J]. Scientific Programming, 2017, 2017(1):1-8.

[53] 陈启伟,王伟,马迪,毛伟.基于 Ext-GBDT 集成的类别不平衡信用评分模型[J]. 计算机应用研究,2018,35(02):421-427.

[54] Engelmann J, Lessmann S. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning[J]. Expert Systems with Applications, 2021, 174: 114582.

[55] Junior L M , Nardini F M , Renso C , et al. A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems[J]. Expert Systems with Applications, 2020, 152:113351.

[56] 向欣, 陆歌皓. 基于 DESMID-AD 动态选择的类别不平衡信用评估模型[J]. 计算机应用研究, 2021, 38(12):7.

[57] Adolfo Rangel-Díaz-de-la-Vega, Villuendas-Rey Y , Cornelio Yáez-Márquez, et al. Impact of Imbalanced Datasets Preprocessing in the Performance of Associative Classifiers[J]. Applied Sciences, 2020, 10(8):2779.

[58] Petrides G, Moldovan D, Coenen L, et al. Cost-sensitive learning for profit-driven

credit scoring[J]. Journal of the Operational Research Society, 2020: 1-13.

[59]He H, Zhang W, Zhang S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios[J]. Expert Systems with Applications, 2018, 98: 105-117.

[60]罗雅晨.类别不平衡的集成学习预测 P2P 网贷信用风险[J].科技与创新,2018(24):1-4.

[61]黄静, 薛书田, 肖进. 基于半监督学习的客户信用评估集成模型研究[J]. 软科学, 2017, 31(7):4.

[62]Li Zhiqiang Li Shouyan,Li Zhilong, Hu Hanlin. Application of XGBoost in P2P Default Prediction[J]. Journal of Physics Conference Series,2021,1871(1).

[63]陶朝杰,杨进.基于 BalanceCascade-GBDT 算法的类别不平衡虚假评论识别方法[J]. 经济数学, 2020, 37(3):7.

致 谢

时光流逝着青春，岁月沉淀着情感，转眼间，三年兰财的悠悠岁月已如同流沙，无声流逝。然岁月流年尘封的往事值得细细品味，逝去的青春依然值得回忆，走过的路子，留下的一系列印痕，蕴含着成长的足迹。回首硕士三年兰州财经大学的点滴生活，得良师指导、益友同行是我这三年来最大的收获，也将会是我人生中重要的财富。毕业之际终会不舍，但想借此向他们表示衷心的感谢。

我要感谢我的导师韩金仓教授，韩老师在学术上对我们严格要求，经常分享一些前言的科研材料、研究工具供我学习使用，在我的论文撰写过程中，从开题、中期、预答辩直到最终完成，韩老师都严格监督我、并且提出指导意见，给我提供了很大帮助。不光在学术上，在生活中亦是如此。三年的研究生生活，韩老师对我的学术上的教导、做人做事的道理将会是我人生中宝贵的财富，打心里感谢韩老师三年的辛苦栽培。

感谢何老师的指导，何老师在学术上的严谨、一丝不苟是非常让人敬佩的，并且也以此来严格要求我们，引导我们思维方式由本科生向研究生不断转变，使得我们的科研素质不断提高，自身能力不断增强。何老师的教导让我们受益匪浅，真心地感谢何老师的谆谆教导。

感谢张克宏老师在我论文写作过程中给予的巨大帮助，张老师严格督促我，时刻关心我的论文进度，并对研究过程中遇到的难题进行具体指导，给出合理的建议，我的论文才得以顺利完成。同时也非常感谢李强教授，三年来在李强老师的实验室进行学习，正是李老师的大力支持，才有了如此优越的资源，才有了我们良好的学习环境；同时也感谢其他任课老师，感谢他们的辛苦付出。

在此也感谢我的家人们，正是有他们的支持，我才能顺利完成学业。同时也特别感谢一下我的各位益友，三位优秀的室友磊哥、文哥、顾总，以及可爱善良的王木木、王雪绒、李聪聪，帅气的学弟黄建民、赵金雨和已经毕业的学长多哥、强哥，还有其他同学们，正是有了你们的相伴相助，我的三年硕士生活，才更加精彩。

同时也对所有评审我论文的老师及出席答辩的其他老师，感谢你们抽出宝贵的时间对我的论文进行指导，谢谢你们的批评指正。