论文题目：基于遗传算法与多模型融

合的信贷风险预测模型研究

学位类别： 工程硕士

学科专业： 软件工程

年 级： 2018级

研 究 生： 王 祥

指导教师： 楼新远

二零二一年 二月

国内图书分类号：TP391.1 密级：公开

国际图书分类号：004

**西 南 交 通 大 学**

**研 究 生 学 位 论 文**

基于遗传算法与多模型融合的

信贷风险预测模型研究

年 级 2018

姓 名 王 祥

申请学位级别 硕士

专 业 软件工程

指 导 教 师 楼新远

二零二一 二月十四日

Classified Index: TP391.1

U.D.C: 004

Southwest Jiaotong University

Master Degree Thesis

Research on Credit Risk Forecasting Model Based on Genetic Algorithm and Multi-model Fusion

Grade: 2018

Candidate: Wang Xiang

Academic Degree Applied for : Master Degree

Speciality: Software Engineering

Supervisor: Lou Xinyuan

Mar 12, 2021

**西南交通大学**

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权西南交通大学可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复印手段保存和汇编本学位论文。

本学位论文属于

1．保密□，在 年解密后适用本授权书；

2．不保密□，使用本授权书。

（请在以上方框内打“√”）

学位论文作者签名： 指导老师签名：

日期： 日期：

西南交通大学硕士学位论文主要工作（贡献）声明

本人在学位论文中所做的主要工作或贡献如下：

1、针对目前信贷风险预测模型存在的问题，围绕国内外研究动态，定量分析了各自存在的缺陷。结合本文研究中使用的用户相关行为数据集，确定了使用集成学习方法和多模型融合的技术来进行信贷风险预测模型的方案。

2、本文的研究主要使用GBDT、XGBoost、LightGBM、RF四种集成学习算法和BP神经网络以及SVM预测信贷风险。围绕当前预测模式存在的问题，根据用户行为挖掘信贷风险相关影响因素，量化具象为相关特征，并使用上述学习算法建立模型。同时，为了提高模型的准确性，引入多模型融合技术，取长补短，确定各模型之间最佳权重系数。通过单模型以及多模型融合之后的模型之间对比，结果表明多模型融合的预测结果明显优于其他几种模型。

3、为了进一步提升模型的泛化能力，本文引入模拟达尔文进化论的遗传算法对四种集成学习方法进行改进，寻找最优的模型超参数组合，减小陷入局部最优的缺陷。结果表明使用遗传算法改进后的集成学习相较于改进前有了一定程度的提升。同时，从算法多样性和参数多样性两方面再次使用多模型融合技术构建加权融合模型，取得了较好的预测效果，证明了该模型的有效性。

4、设计并实现了信贷风险预测模型的应用。通过Django框架和数据的可视化展示了系统的查询模块，使模型结果更加直观地展示。

本人郑重声明：所呈交的学位论文，是在导师指导下独立进行研究工作所得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本论文的研究做出贡献的个人和集体，均已在文中作了明确说明。本人完全了解违反上述声明所引起的一切法律责任将由本人承担。

学位论文作者签名：

日期：

摘要

随着互联网技术的发展以及国家对于互联网金融创新的政策支持，国内互联网金融呈现出蓬勃发展的态势。互联网技术与传统金融领域的结合，相比于传统金融领域商业模式的优势也逐步显现出来。然而在实际贷款过程中，贷款者的宏观因素变化、心理问题、经济环境状况等难以预测，加上央行的征信系统存在结构性缺陷和信息不对称，会导致金融机构的不良贷款余额和不良贷款率。因此精准的信贷风险预测模型成为解决这一问题的有效手段。

本论文就在线信贷风险预测的问题，对相关需求理论进行分析，确定了使用集成学习算法和多模型融合的技术来进行信贷风险预测模型的方案。本文围绕用户相关消费记录、还款账单、浏览行为，挖掘与信贷风险关联特征，并通过集成学习方法和多模型融合技术建立了在线信贷风险预测模型。

为了进一步提升预测模型的泛化能力，本文通过对GBDT、XGBoost、LightGBM、Random Forest的关键超参数进行研究，提出了使用遗传算法（GA）对信贷风险预测模型进行优化，寻找最优的超参数组合。通过仿真实验对比，优化后的单模型各项评价指标均有了一定程度的提升。同时，为了加强个体学习器之间的差异性，引入BP神经网络和支持向量机（SVM），使用Stacking算法分别从算法多样性和参数多样性两个维度进行差异性多模型融合。仿真实验对比显示，利用遗传算法改进后的集成学习做模型融合能够显著提高模型预测的准确性。相较于改进前的单模型，整体评价指标提升超过5%。模型在测试集上AUC值达到0.97以上，精确度超过97%，更适应于解决信贷风险预测的问题。

最后，本文通过Django框架和数据的可视化展示了系统的查询模块，更加直观地展示模型的效果。综上，本文研究的在线信贷风险预测模型具有较好的泛化能力，在信贷风险评估中具备一定的实践价值。

**关键词：**信贷风险；集成学习；多模型融合；遗传算法；Django

**Abstract**

With the development of Internet technology and the country's policy support for Internet financial innovation, domestic Internet finance has shown a booming trend. The combination of Internet technology and traditional financial field, compared with the advantages of the traditional financial field of business model has gradually emerged. However, in the actual lending process, the change of macro-factors, psychological problems and economic environment of borrowers are difficult to predict, as well as the structural defects and information asymmetry in the credit investigation system of the Central Bank of China, which will lead to the non-performing loan balance and non-performing loan ratio of financial institutions. Therefore, accurate credit risk forecasting model has become an effective means to solve this problem.

In this paper, the problem of online credit risk prediction, related demand theory analysis, determine the use of integrated learning algorithm and multi-model fusion technology to carry out credit risk prediction model scheme. Focusing on the consumption record, repayment bill and browsing behavior of users, this paper excavates the characteristics associated with credit risk, and establishes an online credit risk prediction model through integrated learning method and multi-model fusion technology.

In order to further improve the generalization ability of the prediction model, this paper studies the key hyperparameters of GBDT, XGBoost, LightTGBM and Random Forest, and proposes to optimize the credit risk prediction model by using genetic algorithm (GA) to find the optimal combination of hyperparameters. Through comparison of simulation experiments, each evaluation index of the optimized single model has been improved to a certain extent. At the same time, in order to strengthen the difference between individual learners, BP neural network and support vector machine (SVM) are introduced, and Stacking algorithm is used to carry out the difference multi-model fusion from two dimensions of algorithm diversity and parameter diversity respectively. The comparison of simulation experiments shows that model fusion using the improved ensemble learning based on genetic algorithm can significantly improve the accuracy of model prediction. Compared with the single model before improvement, the overall evaluation index increased by more than 5%. The AUC value of the model on the test set reaches more than 0.97, and the accuracy is more than 97%, which is more suitable for solving the problem of credit risk prediction.

Finally, this article uses the Django framework and data visualization to show the query module of the system, and more visually show the effect of the model. In conclusion, the online credit risk prediction model studied in this paper has good generalization ability and has certain practical value in credit risk assessment.

**Key words:** Credit Risk; Ensemble Learning; Multimodel Fusion; Genetic Algorithm; Django

目 录

[第1章 绪论 1](#_Toc64141087)

[1.1 课题研究背景及意义 1](#_Toc64141088)

[1.1.1 背景及意义 1](#_Toc64141089)

[1.1.2 拟解决的问题 2](#_Toc64141090)

[1.2 相关文献综述 2](#_Toc64141091)

[1.2.1 信贷风险研究现状 2](#_Toc64141092)

[1.2.2 互联网金融风险研究现状 3](#_Toc64141093)

[1.2.3 信用风险评估模型研究 4](#_Toc64141094)

[1.3 创新点 6](#_Toc64141095)

[1.4 论文组织结构及章节安排 7](#_Toc64141096)

[第2章 相关知识介绍 9](#_Toc64141097)

[2.1 集成学习算法概述 9](#_Toc64141098)

[2.2 机器学习算法理论 10](#_Toc64141099)

[2.2.1 序列集成算法介绍 10](#_Toc64141100)

[2.2.2 并行集成算法介绍 17](#_Toc64141101)

[2.2.3 CatBoost算法介绍 18](#_Toc64141102)

[2.2.4 SVM算法介绍 20](#_Toc64141103)

[2.3 模型融合理论 22](#_Toc64141104)

[2.4 预测模型评价标准 26](#_Toc64141105)

[2.5 本章小结 29](#_Toc64141106)

[第3章 基于多模型融合的信贷风险预测模型研究 30](#_Toc64141107)

[3.1 数据探索及数据预处理 30](#_Toc64141108)

[3.1.1 数据特征分析 30](#_Toc64141109)

[3.1.2 数据清洗及缺失值处理 32](#_Toc64141110)

[3.2 特征工程构建 35](#_Toc64141111)

[3.2.1 基础特征构建及处理 36](#_Toc64141112)

[3.2.2 显著特征筛选 37](#_Toc64141113)

[3.3 基于机器学习单模型仿真实验及对比分析 38](#_Toc64141114)

[3.3.1 仿真实验环境 38](#_Toc64141115)

[3.3.2 仿真实验及实验结果对比分析 39](#_Toc64141116)

[3.4 差异性多模型融合仿真实验及对比分析 49](#_Toc64141117)

[3.4.1 算法多样性的模型融合 49](#_Toc64141118)

[3.4.2 参数多样性的模型融合 51](#_Toc64141119)

[3.5 本章小结 53](#_Toc64141120)

[第4章 基于遗传算法（GA）改进及多模型融合的信贷风险预测研究 54](#_Toc64141121)

[4.1 遗传算法和网格搜索 54](#_Toc64141122)

[4.1.1 遗传算法GA思想 54](#_Toc64141123)

[4.1.2 遗传算法的设计与实现 54](#_Toc64141124)

[4.1.3 遗传算法实例 57](#_Toc64141125)

[4.1.4 网格搜索 57](#_Toc64141126)

[4.2 单模型算法改进仿真实验及对比分析 58](#_Toc64141127)

[4.2.1 单模型超参数优化设计 58](#_Toc64141128)

[4.2.2 遗传算法改进单模型 58](#_Toc64141129)

[4.3 单模型算法改进后的模型融合 63](#_Toc64141130)

[4.3.1 算法多样性的模型融合 63](#_Toc64141131)

[4.3.2 参数多样性的模型融合 64](#_Toc64141132)

[4.4 信贷风险预测模型的最终建立 65](#_Toc64141133)

[4.5 本章小结 66](#_Toc64141134)

[第5章 基于多模型融合的信贷风险预测模型的应用 67](#_Toc64141135)

[5.1 系统开发环境 67](#_Toc64141136)

[5.2 系统设计 67](#_Toc64141137)

[5.3 预测查询模块及预测设计实现 68](#_Toc64141138)

[5.4 本章小结 70](#_Toc64141139)

[总结与展望 71](#_Toc64141140)

[致 谢 73](#_Toc64141141)

[参考文献 74](#_Toc64141142)

[攻读硕士学位期间发表的论文及科研成果 78](#_Toc64141143)

# 第1章 绪论

## 1.1 课题研究背景及意义

1.1.1 背景及意义

近年来，以大数据、云计算、机器学习等先进技术为依托，国内互联网金融呈现出蓬勃发展的态势，发展形式越来越多元化，正潜移默化影响着我们的生活。互联网技术与传统金融领域的结合，相比于传统金融领域的商业模式，其优势也逐步显现出来[1]。随着互联网技术的不断渗透，两者之间的联系也越来越紧密。如今，以P2P借贷、互联网在线支付等商业模式为代表的互联网金融服务正迅速发展，有效缓解了很多小微企业和个人的资金需求，填补了部分传统金融服务未触及的空白，显著促进了国内电子商务的蓬勃发展以及传统金融领域服务的延伸，显示了其强大的生命力和创造力。早在2015年中国政府已将“互联网+行动计划”写进了工作报告，鼓励政府部门推进先进技术（如互联网、大数据）融入到现代制造业中，把互联网金融行业地持续发展视为国家战略，从而激发这一领域万众创新的热情，加快推进传统金融机构的金融服务互联网化[2]。以P2P网络贷款为例，2018年全年网贷行业成交量达到了17948.01亿元，2019年市场加强监管，清退众多平台，但2019年成交量仍达9649.11亿元[3]。截止到2019年底，P2P网贷行业历史累计成交量约为9万亿元。P2P贷款已经成为个人和小微企业融资不可或缺的渠道。

对于以P2P借贷、在线信贷为代表的互联网金融服务商来说，最大的运营风险源自信贷客户的违约信用风险，即如何评估信贷客户的还款能力等。在对信贷客户的评估环节中，可利用平台自身收集的历史信贷数据作为依据之外，中国人民银行的征信系统也可作为一个重要的补充。目前，央行的征信系统基本为企业和个人都建立了信用电子档案，截止到2019年9月，央行已收集自然人信息量超9.9亿人，覆盖范围和基数已经非常庞大[4]。但其中仅有约3亿多人存在历史往期的信贷记录，存在大量的数据记录缺失，严重限制了对信贷客户评估的应用范围。其次，央行的征信系统的记录主要源自各贷款机构，因此数据的全面性和时效性存在较为明显的短板，较难反映客户的真实信息，对用户信用状况的评估影响很大。如何为金融机构挑选优质贷款用户现已经变成了信用风险管理的核心问题，包括预估客户偿还贷款的能力、了解用户的还款意向、动态的检测客户资质、现金流、偿债能力在贷款发放整个过程中是否会有重大变化发生等[5]。

本文对信贷客户的风险评估，具有重要的理论和现实意义。

1、有助于提高P2P、在线信贷等平台的风险控制能力[6]

如上文所述，由于缺少对客户的征信数据，因此仅凭征信来评估客户的信用风险往往存在较大的偏差。本文的研究方法是从用户的基本信息、银行卡消费记录、信用卡还款记录以及用户的浏览行为记录来挖掘客户的信息，建立模型全面评估用户的信贷风险等级，从而提高金融平台的风险控制水平，降低平台的坏账率，减少损失，达到系统性防范的目的。

2、对我国信贷风险模型的开发具有一定的借鉴意义

当前，我国政府正在逐步放开征信管制，各个互联网金融平台和民营征信机构创新个人信用风险评估模型，背靠大数据充分挖掘用户的相关特征，并与之业务紧密结合。本文研究中包括的关键特征、选取的模型算法等方面对该类模型的开发具备一定的借鉴意义。

3、对于完善国内金融服务的模式有积极意义[6]

目前国内现有的商业模式中，发展较好的有京东白条、花呗借呗等小贷服务，因为其有平台的交易、支付等众多数据作为支撑。而在传统的金融服务模式下，通常需要抵押实物，难以有效评估其信用风险。在本文的研究中，可以通过客户的基本信息以及消费行为等重新评估贷款客户的信用风险状况，拓宽个人和企业的融资渠道，拓展和完善国内金融的服务模式。

1.1.2 拟解决的问题

信贷风险管理一直是金融服务行业重要的组成部分，尤其对于互联网金融服务尤为重要。而在当前传统的个人信用风险评估模式下，主要依赖于中国人民银行的征信系统。如前所述，央行的征信系统存在结构性缺陷，会大大限制评估客户的范围。对于互联网金融企业来说，信贷业务可能产生的风险以及其他问题一直是需要面对并解决的核心问题。本文通过分析国内外的研究动态，从用户的基本信息、银行卡消费记录、信用卡还款记录以及用户的浏览行为出发，通过数据的探索、预处理、挖掘出相关特征并建立信贷风险评估模型，解决信贷风险评估范围小及准确性低的问题。同时为了进一步优化模型的效果，将遗传算法与单模型结合，寻找最优的模型超参数，并运用多模型融合的技术，构建加权融合模型，以最优权重参数确定最终模型，取得较佳的预测效果。

## 1.2 相关文献综述

1.2.1 信贷风险研究现状

小额担保贷款最早可追溯到孟加拉国。20世纪70年代，穆罕穆德·尤努斯在孟加拉国创办了孟加拉农业银行格莱珉（Grameen，意为乡村）试验分行，此后该模式不断发展并逐渐成形，成为个人、企业甚至发展中国家减轻贫困的有效途径[7]。21世纪以来，互联网的普及和互联网技术的发展，一种依靠网络的在线信贷开始逐步发展起来。

关于信贷风险和互联网金融，王汉君（2013）、闫真宇（2013）[8]等认为互联网金融既需要解决传统金融所面临的风险，如信用风险、市场风险、操作风险、流动性风险、系统性风险、和技术风险等方面，也面临与互联网技术相关联的一些特定风险之下。王裕粟[9]认为信用风险是目前金融机构面临的最主要风险之一，信用风险以及由此产生的巨额不良贷款已经成为制约我国金融业发展的一个重要问题，金融机构或平台对信用风险的控制和管理能力直接关系到金融体系的稳定与健康发展，建立内部信用评级是适应新的监管要求的必然趋势，并成为提高金融机构或平台核心竞争力的重要手段之一。洪娟、曹彬、李鑫[10]等认为互联网与金融的结合拓宽了金融服务的边界，然而，由于虚拟化、信息技术、安全体系以及法律和监管问题，互联网金融的风险管理要比传统金融更加复杂。

对于贷款客户的信用风险评估，需借助统计学、运筹学和计算机科学等理论，挖掘该用户的数据中所蕴涵的关于信用风险的特征和行为特点，建立相应的预测模型。周贤[11]在新形势下银行信贷风险管理问题研究中分析了现阶段了我国信贷风险管理工作中存在的一些问题，主要体现在信贷对象与行业领域的问题。同时，针对这两个问题提出了新形势下提高信贷风险管理的有效措施。刘佳蒙[12]在商业信贷风险管理存在的问题与对策研究中分析了强化商业信贷风险管理的意义，并从社会环境和认知因素这两大方面的因素下分析我国商业信贷风险管理中存在的问题，同时给出了解决商业信贷风险管理的对策。徐溪蔓[13]在基于时间序列模型的商行信贷规模与风险管理分析中运用时间序列模型测定了全国商业近10年的信贷/GDP比率，这个比率能够反映出我国商业信贷相对规模变化的长期趋势。崔璇在基于贝叶斯模型的信贷风险研究中，利用大数据样本分析，建立贝叶斯信贷风险评估模型，通过直观性的预测结果制定各类客户风险评分值，并根据风险评分值确定是否发放贷款。谢昌文、李黎[14]在基于层次分析法的客户信贷风险预警模型的研究中站在商业银行内部信贷风险管理的角度，分析了影响信贷风险的因素，并运用层次分析法构建了预警模型，以提升对信贷风险的管控能力。

1.2.2 互联网金融风险研究现状

二十一世纪以来，随着互联网的发展，互联网+金融成为金融领域的一个最新发展趋势。基于智能移动设备的普及，用户可直接完成金融服务的注册并使用其金融产品，如蚂蚁花呗、京东白条、360借条等。当前互联网金融提供的服务总共有两种模式，一种是线下金融机构的传统业务，同时提供线上金融服务的销售等；另一种是新兴的互联网金融业务类型，只针对互联网用户在线上进行销售及服务[15]。互联网金融的出现拓宽了个人和企业的融资渠道，拓展和完善了国内金融的服务模式。互联网金融给社会生产生活带来极大便利的同时，表现出了比传统金融服务更大的风险，且这种风险往往是不可逆的，更容易引起金融恐慌。

（1）互联网背景下的金融传统风险

互联网背景下的金融传统风险，首当其冲的是市场风险。金融市场本身就具备较大的市场风险，还包括许多外部风险。我国的金融市场虽处于国家的监管之下，但主要依赖于市场经济的管控。其次是流动性风险。互联网金融流动性风险主要是指资金流动性差，供给存在不足时，可能会导致金融机构缺乏资金来偿还给用户的情况，这种情况下，用户可能会发生恐慌，那么金融机构也可能遭受重创[16]。在央行等十部委发布的《关于促进互联网金融健康发展的指导意见》中提出，在整治互联网金融风险时，应特别关注其流动性风险，互联网金融流动性是传统流动性风险在互联网环境下的延伸和变种。最后是信用风险。对于互联网金融的信贷业务，用户只需要注册并提供身份信息证明以及历史消费记录，即可从平台借贷到一定数量的资金，而忽视了对借贷用户的可用风险的严格评估。若平台没有充足的资金链，一旦资金链断裂，小型互联网公司难以承受宣布破产，则对应的投资资金便无法收回。

（2）互联网背景下的金融新风险

在互联网发展的新形势下，新的金融风险也应运而生，这些新的风险正在制约着互联网金融的发展。新形势下的金融新风险，第一个是技术风险。线上的互联网金融依托于智能移动设备上的APP软件来提供相应的服务，尤其对于高并发量的APP来说，技术可能会成为一个制约发展的瓶颈[17]。另外，互联网金融的交易均是数字交易并保存在后台数据库中，若数据库遭受攻击，有可能会造成无法挽回的损失。此外，对于提供信贷服务的平台来说，如何建立精准的评估贷款客户违约风险的模型是一个核心的技术难题。第二个是数据风险。互联网金融交易均是通过对数据信息的处理来实现对资金的管理。在交易过程中，互联网内部会收集大量的交易数据保存在数据库中，若在交易过程中或事后出现数据库数据的丢失或者出错，则相应的资金往来交易状况便会出现混乱，造成无法想象的严重后果。

综上所述，由于我国一些法律法规和规章监管制度尚未完善，互联网金融仍存在较多的风险和危机诱导因素，需要政府和互联网金融机构的共同努力，进一步完善我国互联网金融的监管制度，强化金融监管部门的职责，加强互联网金融机构风险控制指导，稳定我国互联网金融市场的秩序[18]。

1.2.3 信用风险评估模型研究

信用风险又称违约风险，主要是指债务人与金融机构约定的还款期限到期后，债务人主观上不愿意或者客观上没有能力履行双方签订的还本付息协议约定内容，导致金融机构出现资金受到损失的可能性[19]。对于贷款客户的信用风险评估，需借助统计学、运筹学和计算机科学等理论，挖掘该用户的数据中所蕴涵的关于信用风险的特征和行为特点，建立相应的预测模型。

目前国内外关于信用风险评估模型主要有如下几种：

1. 专家模型

专家模型是一种传统的信贷风险评估方法，在互联网时代未到来前，贷款用户的数据和历史资料很难全面获取到，对客户的信用状况评估依赖于信贷专家的主观判断，逐渐形成了一套完善的信用风险评估管理体系[20]。例如，应用最为广泛的专家模式“5C”要素分析法[20]，从机构中挑选出拥有丰富信贷风险评估的专家，对借款人的Character（品格）、Capital（资本）、Capacity（偿付能力）、Collateral（抵押品）以及Cycle Conditions（周期的或经济的形式）五个方面进行评估，并赋予其合理的指标权重，最终做出判断。此外，常用的专家模式还包括LAPP法，通过评估Liquidity、Activity、Profitability和Potentialities，即借款人的流动性、活动性、营利性和潜力这四个方面进行综合评估；CAMEL体系—Capital adequacy、Asset quality、Management、Earnings、Liquidity，即资本充足率、资产质量、管理水平、收益状况、流动性等。专家模型在指标处理上具有突出优势，且灵活性较好，但此模型过多依赖信贷风险评估专家的主观判断，且评估效率低下。

1. Logistic模型

Logistic模型是目前应用较多信贷风险评估模型，是一种广义线性模型，它的模型简单鲁棒性强，可解释性好，并且可直接预测出借款人发生违约的概率。陈晓兰、任萍[21]在对企业信用风险评价研究中，构建了AHP和Logistic这两种混合模型，并运用该模型对国内部分样本企业的信用风险进行了预测评估，结果表明该模型对于商业银行视角的企业信用风险评价具有良好的适用性，为商业银行进行信用评级提供了参考。光丽[22]基于Logistic回归模型对P2P网贷平台借款人的信用进行评估，参考了银行风险评估体系的指标，通过数据处理、量化筛选出最终特征指标，继而使用Logistic算法进行建模，得到新的信用风险评估模型。呼振凯[23]在P2P网络借贷中借款人的信用风险评估研究中选取了18项与借款人信用风险相关的指标，并运用K-Means聚类分析，挑选出其中最具代表性的正负样本作为训练集，使用Logistic算法进行训练，结果证明了模型的预测能力和有效性。

1. 基于决策树算法的信贷风险评估模型

决策树算法目标就是根据给定的训练数据集构建一个决策树模型，这个模型能够拟合训练数据，使之正确分类。决策树学习的本质就是从训练集中归纳出一组分类的规则以此来预测未知的数据。徐晓霞、李金林[24]等结合我国当前商业银行的实际状况，基于机器学习和数据挖掘理论中的决策树算法建立了商业银行的信用风险评估体系和评估模型。通过这个模型可以预测出企业是否会违约以及会违约的概率，对于商业银行的信贷风险控制具有很好的指导和借鉴意义。张红梅[25]在分析了房贷信用风险评估的基础上，进一步分析了在房贷信用评估中存在的问题，并重新设计了影响房贷信用风险的指标，选择了决策树算法训练模型来评估房贷的信用风险。马文勤[26]在基于决策树在小额农户信贷风险评估研究中着重描述了C4.5决策树算法相较于其他数据挖掘技术方法对农村小额贷款的客户信用风险评估存在的优势，根据农村信用社信贷业务风险控制的特点提出了一套适合农村信用社的信贷管理的技术框架，该成果给农村小额农户贷款信用风险提供了一定的参考意义。

1. 神经网络模型

神经网络模型以“神经元”为基本处理单元，经过广泛的互相连接形成复杂的网络结构，是一个复杂的非线性学习函数，可以达到较好的预测效果。神经网络最初应用于金融方面是Dutta等在债券信用评级引入的。1994年Altman[27]在公司的财务危机中引入了神经网络作为预测模型，取得了比较好的效果。方先明、熊鹏和张谊浩[28]在基于Hopfiled神经网络的信用风险评价模型研究中为克服商业银行信用风险评估中所遇到的评判失效的难题，通过Hopfield神经网络建立了信用风险评价模型，模型的网络运行结果反映了当前信用风险的状态。实验证明，该模型能在一定程度上反映样本数据的数字特征，相较于其他模型更适合于信用风险的评价。张佳维[29]在分析了国内外模糊神经网络的研究现状，根据现有的指标评估体系，对国内部分银行的真实数据进行预处理，并提出了将模糊神经网络应用于个人信用风险的评估中，最终实验证明模型具有很高的准确率和精确率，为商业银行的个人信用风险评估提供了更多的参考价值。

## 1.3 创新点

本论文研究重点在于分析国内互联网金融企业或平台快速发展中存在的一些问题及原因，围绕当前评估模型可能导致信贷风险评估出现偏差，客户评级不准确，提出了基于集成学习方法的机器学习技术建立个人信用风险评估模型。为了提高模型预测的准确性，运用Stacking模型融合方法，构建加权融合模型，以最优权重参数确定最终模型。同时，为了进一步优化模型，使用遗传算法对GBDT、XGBoost、LightGBM、Random Forest模型的超参数进行优化，完成了基于用户基本信息、银行卡消费记录、信用卡还款记录和用户浏览行为的个人信贷风险评估模型的设计与实现。最后将该模型在基于Django框架进行了系统化的可视化应用。本论文研究的创新点主要包括如下三个方面：

1、针对单模型预测精度有限的问题，使用Stacking算法对GBDT、XGBoost、LightGBM、Random Forest、BP神经网络，SVM（支持向量机）构建加权融合模型，从特征、算法、参数这三个层面加大模型之间的差异性，以最优权重参数确定最终模型，模型的效果大大提升。

2、为了解决模型的超参数优化问题，使用遗传算法对GBDT、XGBoost、LightGBM、Random Forest进行改进和超参数的寻优，该算法是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程中的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。由此建立的模型的准确率有了显著提升。

3、本论文研究中的数据均为脱敏数据，主要通过对脱敏数据的统计加工特征建立模型，摆脱了对用户真实数据的依赖，有助于保护用户隐私数据。

## 1.4 论文组织结构及章节安排

本论文围绕国内外互联网金融企业或平台信贷风险模型评估中存在的问题，结合客户信贷业务流程，利用集成学习、多模型融合、遗传算法等机器学习技术建立行之有效的信贷风险评估预测模型，用以衡量贷款客户的风险评分和等级，减少不良贷款率，增加优质客户的收益，降低平台的风险。此外，本文通过对GBDT、XGBoost、LightGBM、Random Forest这四个集成学习算法的超参数进行研究，利用遗传算法对其进行改进，寻找最优超参数并再次进行多模型融合，对客户的信贷风险进行了精确的评估，并最后进行了系统的应用展示。具体地，论文的结构框架安排如图1-1：

本论文共五章，主要内容如下：

第一章绪论首先介绍了当前互联网金融平台的发展现状，并引出信贷风险评估模型研究的背景和意义，目的是为了解决互联网金融平台对贷款客户信用风险管理的难题。在国内外研究综述中，简要介绍了信贷风险模型和集成学习算法在国内外研究的现状。最后介绍了论文的三个主要创新点和组织结构框架。

第二章主要是本论文研究的相关知识介绍。首先是对集成学习思想的概述，主要包含了Boosting和Bagging，分析对比了它们的不同点。其次是机器学习算法的一些理论研究介绍，主要包含了GBDT、XGBoost、LightGBM、Random Forest、Catboost、SVM。再次是模型融合的理论介绍，重点介绍了Stacking算法。最后提出了预测模型的评价指标。

第三章主要为多模型融合的信贷风险预测模型研究的实验部分，包含数据的探索和预处理、特征工程构建、机器学习单模型的仿真实验对比分析以及差异性多模型融合的仿真实验，多模型融合主要从算法多样性和参数多样性加大个体学习器之间的差异性，提升模型融合的整体效果。

第四章重点介绍使用了遗传算法和网格搜索算法优化GBDT、XGBoost、LightGBM、Random Forest这四个单模型超参数的过程。针对四个模型确定不同的搜索空间，优化不同的超参数组合，有效提升了单模型的泛化能力。同时，再次使用Stacking从算法多样性和参数多样性提升模型融合的效果，并确定最终的信贷风险预测模型。

第五章主要介绍了基于多模型融合的信贷风险预测模型的应用实现过程。首先介绍了开发预测系统的硬件和软件环境，并对Django的框架原理做了重点阐述。紧接着结合Django框架和数据库相关技术完成了信贷风险预测模型查询模块的可视化工作。



图1-1 论文结构框架图

# 第2章 相关知识介绍

本章节是关于信贷风险预测理论、机器学习算法预测模型、多模型融合理论以及预测模型的相关评价指标的内容介绍。

## 2.1 集成学习算法概述

在有监督学习任务中，机器学习算法的目标是拟合训练集学习出一个稳定的且在各个方面表现较好的模型。但由于模型在训练时存在偏好，所以实际情况有时并不理想，这时可通过结合并构建多个学习器来完成分类或回归的学习任务，集成学习的思想正是基于这一点。集成学习就是先产生一组个体学习器，再用某种策略将他们结合起来，获得比单一学习器更好的泛化能力，同时也可降低陷入局部极小点的风险[30]。其中，个体学习器可以是同种类型的，如只包含“决策树集成”，这样的集成称为同质的；个体学习器也可以是异质的，如同时包含决策树和神经网络的集成。集成学习算法思想的示意如图2-1所示：



图2-1 集成学习算法示意图

集成学习方法根据个体学习器的生成方式不同，可大致分为两类，一种是序列集成方法，该集成学习中，个体学习器之间存在强依赖关系，第N个学习器的生成依赖于前N-1个学习器的学习效果串行生成，通过对之前训练中错误标记的样本赋予较高的权重提高整体的预测效果。该学习方式的代表是Boosting，包括梯度提升树GBDT、XGBoost等。另一种是个体学习器之间不存在强依赖关系，参与训练的基础学习器独立并行生成，并行方法的原理是利用基础学习器之间的独立性，通过平均可以显著降低错误。该种集成学习算法的代表是Bagging和随机森林。

表2-1为两类集成学习算法Boosting和Bagging的细节对比：

表2-1 Boosting和Bagging的细节对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 集成学习算法 | 样本选择 | 样例权重 | 预测函数 | 并行计算 |
| Boosting | 每一轮训练集不变 | 根据错误率调整样例的权值，错误率越大权重越大 | 每个弱分类器都有相应的权重，分类误差小的分类器有更大的权重 | 各预测函数只能顺序生成，后一模型参数需要等前一轮模型的结果 |
| Bagging | 训练集在原始集中有放回选取 | 使用均匀取样，每个样例的权重相等 | 所有预测函数的权重相等 | 各预测函数可以并行生成 |

## 2.2 机器学习算法理论

2.2.1 序列集成算法介绍

集成学习中的序列集成算法属于迭代算法，在学习过程中，通过不断地使用一个弱学习器弥补之前学习器的不足来串行地构造一个较强的学习器，代表是Boosting算法。Boosting算法的基本思想是先赋予每个训练样本相同的概率，之后进行N次迭代，在每次迭代中，对分类错误的样本加大权重，使得学习器在下一次的迭代中更加关注这些样本[31]。Boosting系列算法中最著名的有AdaBoost和提升树系列算法，包括GBDT、XGBoost、LightGBM等。Boosting算法的思想如图2-2所示：



图2-2 Boosting算法流程

（一）GBDT算法

梯度提升树GBDT（Gradient Boosting Decision Tree）算法可分为两部分，分别是梯度提升算法（Gradient Boosting）和决策树算法（Decision Tree）。GBDT在训练时梯度提升每一次的计算是为了减小上一次的残差，也就是说每个新模型的建立是为了使得之前模型的残差往梯度方向减少[32]。GBDT中的基学习器均为回归树，利用Boosting算法思想将多个弱学习器组合成强学习器，可做分类或回归任务。

（1）梯度提升算法

梯度提升算法是对Boosting算法的延伸，在模型训练迭代过程中，通过增加基学习器降低损失函数值[32]。在GBDT的迭代中，假设前一轮迭代得到的强学习器为

，损失函数为 ，本次迭代的弱学习器为 ，则本轮的损失函数为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-1） |

基本思路如下：

强学习器的公式如式2-2，其中为各基学习器。

 （2-2）

利用梯度下降的方式（损失函数的负梯度）进行残差拟合。

 （2-3）

 （2-4）

（2）决策树算法

决策树算法是一个类似于流程图的树形结构，树的内部节点代表对特征的测试划分，树的每个分支代表该特征的结果，叶子节点代表一个类别[33]。决策树本质即为划分数据集，树的根节点包含全部样本，每个叶子节点包含的是同种类别的样本。

决策树的学习过程主要分为三个部分，分别是特征选择、决策树生成和剪枝。关键点在于如何确定最优划分属性和最优切分点。特征选择的准则有信息增益（ID3使用）、信息增益比（C4.5使用）和基尼系数（CART使用）三种。其中常用的是根据信息增益准则选取信息增益最大的特征作为分裂节点从树的根节点自顶向下进行划分，递归生成完整的树结构。

信息熵*Ent*(*D*)定义如式2-5，其中表示样例集合*D*中第*k*类样本所占比例。

 （2-5）

信息增益公式，表示在*V*个分支节点中第*v*个属性*A*包含的取值样本，公式如2-6所示：

 （2-6）

GBDT算法中决策树的分裂方法可分为两种，深度优先和广度优先。深度优先是采用递归思想按最大收益叶子生长的方式（Leaf-wise），广度优先是按照层次来构建树（Level-wise）。

按深度优先的方式可以花费较小的代价构建最终的决策树。这种方式的优点是精准度高，拟合数据的能力强，且可以在较短的时间内完成树的生长，但此种方式生成的树结构复杂，容易造成过拟合[34]。另外，树节点构造的过程是顺序的，无法直接进行并行计算。图2-3是决策树按叶子生长的过程。



图2-3 按叶子生长分裂的决策树

按广度优先的方式（按层生长的方式）指每一层的每一个结点都要进行分裂，因此以这种构造方式可以并行加速计算，但在训练时会产生多余的分裂节点，计算代价较高[35]。此外，模型在训练过程中每次迭代都需要遍历整个数据集，对运算设备的运行内存要求较高。图2-4是按层分裂的决策树生长过程。



图2-4 按层分裂的决策树

（二）XGBoost算法

（1）理论推导

XGBoost（eXtreme Gradient Boosting）全称为极端提升树，是由陈天奇等人开发的一个开源机器学习项目，高效地实现了GBDT算法并行计算和工程上的许多改进[36]。XGBoost同样是一种基于梯度提升树的改进算法，相较于GBDT，XGBoost对损失函数做了泰勒二阶展开并且支持损失函数的自定义[36]。对于目标函数的优化是XGBoost的核心，下面以目标函数为例：

XGBoost的目标函数由训练损失和正则化项两部分组成，定义如式2-7：

|  |  |  |
| --- | --- | --- |
|  |  | （2-7） |

其中，前一部分为训练损失，后一部分为树的复杂度作为正则化项。代表损失函数，常见的损失函数有平方损失函数、逻辑回归损失函数， 是第i个样本的预测值。

假设第t次迭代要训练的模型是 ，则有：

|  |  |  |
| --- | --- | --- |
|  |  | （2-8） |

带入目标函数Obj，可以得到：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | | | （2-9） |
|  | |  | （2-10） | |

将训练损失按泰勒公式展开并去除常数项得到：

|  |  |  |
| --- | --- | --- |
|  |  | （2-11） |
|  |  | （2-12） |

定义一棵树的复杂度，如式2-13所示，树的复杂度由两部分组成，叶子节点的数量和叶子节点权重向量的L2范数，即：

|  |  |  |
| --- | --- | --- |
|  |  | （2-13） |

对叶子节点归组，将属于第j个叶子节点的所有样本划入到一个叶子节点样本集中，数学表达式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-14） |

代入目标函数得到：



其中

|  |  |  |
| --- | --- | --- |
|  |  | （2-15） |

最后，给树结构打分，套用一元二次函数的最值公式，可以得到每个叶子节点的权重 ，以及达到最优的Obj的目标值：

|  |  |  |
| --- | --- | --- |
|  |  | （2-16） |
|  |  | （2-17） |

（2）寻找最佳分裂点

在构造树结构时，最主要的学习任务是寻找最佳分裂点，一般是采用贪心策略的全局扫描法，每次进行分裂时遍历一遍全部候选分裂点，步骤如下：

* 遍历每个节点的每个特征；
* 对于每个特征，将特征值按大小排序；
* 线性扫描，计算每个分裂点的分裂增益，找出每个特征的最佳分裂特征值；
* 在全局所有特征中找出分裂增益最大的特征及特征值。

若数据量较大导致内存无法一次加载全部数据，贪心算法的效率就会大大降低。基于此问题，XGBoost采取了下列措施加快最佳分裂点的查找：

* 特征预排序+缓存：在XGBoost训练之前，预先对每个特征按照特征值大小排序，并保存为block结构，方便之后在迭代过程中可重复使用此结构，大大减少计算量[37]。
* 分位点近似法：对每个特征按特征值排序后，采用类似分位点选取的方式，仅选出常数个特征值作为候选分割点，并从中选出最优一个作为最佳分裂点[37]。
* 并行查找：各特征值已预存为block结构，XGBoost可利用多线程并行计算每个特征的最佳分裂点，可在大规模训练集上适应性扩展[37]。

（三）LightGBM算法

LightGBM是微软开源的一个实现GBDT算法的框架，它和XGBoost一样是对GBDT的高效实现，都采用的是损失函数的负梯度作为当前决策树的残差近似值去拟合新的决策树[38]。LightGBM使用的是histogram算法，训练时使用的内存更低，数据分隔的复杂度更低。其思想是将连续的浮点特征离散成k个离散值，并构造宽度为k的histogram，再遍历训练数据，统计每个离散值在直方图中的累计统计量，在进行特征选择时，只需要根据直方图的离散值，遍历寻找最优的分割点[39]。

为了降低算法在训练时内存的占用以及计算的代价，LightGBM在XGBoost的基础上进行了进一步的优化，使用了直方图优化算法、带有深度限制的按叶子生长（leaf-wise）算法、直方图差加速等。使用特征并行、数据并行和投票并行这三种方式，从特征、数据、通信IO三个方面并行加速。相较于其他提升树算法，在内存占用、训练速度和准确性方面均有较大幅度的提升。

（1）直方图算法优化

相比于其他提升树算法遍历每一个分割点计算分裂增益，直方图算法首先将连续的浮点数据转换为bin数据，确定每个特征需要的桶bin的数量，再均分，将属于该桶的样本数据更新为bin的值，最后用直方图表示，整个过程将大规模的数据映射到了直方图上[39]。训练时只需保存特征离散后的值，桶bin值的存储比原本训练样本所占内存大大减少，内存消耗可以降低为原来的八分之一。



图2-5 LightGBM直方图算法内存优化图

此外，使用bin替代原始数据相当于增加了正则化，bin数量的选择决定了正则化的程度，bin越少惩罚越严重，欠拟合的风险越高。

（2）直方图差加速

LightGBM另一个优化是利用直方图做差加速。由直方图的构造方法可知，一个叶子的直方图可以由它的父节点的直方图与它兄弟的直方图做差得到。我们正常构造直方图时，需要遍历该叶子上的所有数据，但直方图做差加速仅需要遍历直方图的k个桶，利用此方法可以用很小的代价得到兄弟叶子的直方图，进一步提升训练速度[40]。

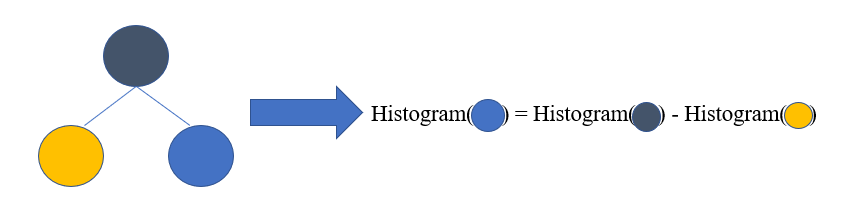


图2-6 差加速优化

（3）基于梯度的单边采样算法Gradient-based One-Side Sampling（GOSS）

GOSS算法的思想是一种在减少数据量和保证精度上平衡的算法。通过区分不同梯度的实例，保留较大梯度实例的同时对较小梯度随机采样的方式减少计算量，从而达到提升效率的目的[41]。采取该方法会影响部分数据的分布，为了抵消对数据分布的影响，在计算信息增益的时候，GOSS会给予小梯度的数据一定的权重。具体地，GOSS在计算时首先依据数据的梯度绝对值排序，选取前a个实例，然后在剩余的数据中随机采样b个实例样本。在计算小梯度数据的信息增益时，会乘以(1-a)/b的权重系数[41]。

假设O表示某个固定节点上的训练集，样本损失函数的梯度表示为{*g*1, *g*2, ···, *gn*}。则分割特征j的分割点d定义为：

 （2-18）

其中，

 （2-19）

较大的梯度*a*×100%组成实例数据集合*A*，之后对由数据集合组成的具有较小梯度的其余集合*Ac*采样为大小为的子集*B*。最后，根据上述公式得出子集*A*∪*B*计算得到增益为：

 （2-20）

其中各子项定义如公式2-21所示。

 （2-21）

通过较小的数据集估计信息增益将大大地减小计算量，GOSS不会丢失许多训练精度，胜过随机采样。

（4）LightGBM模型决策树的查找最优分裂点原理

在LightGBM模型中决策树的最优分裂点查找伪代码如下所示：

从LightGBM的伪代码可看出，LightGBM在寻找最优分裂点时，先为每个特征构造直方图，对每一种离散值出现的次数进行统计，并降序排列，之后查找划分出bin桶的个数，对符合条件的bin使用一阶梯度和二阶梯度计算分裂增益，利用计算出的分裂增益从大到小进步排序，并在k个bin上进行搜索，查找出最优分裂特征和分裂点[42]。

（四）三种Boosting算法细节对比

表2-2 GBDT、 XGBoost和LightGBM模型细节对比

|  |  |  |  |
| --- | --- | --- | --- |
| 模型细节 | GBDT | XGBoost | LightGBM |
| 树生长模式 | 按层生长 | 按层生长 | 受深度限制的按叶子生长 |
| 分裂点搜索方式 | 直接遍历 | 特征预排序（默认） | 直方图算法 |
| 分裂时收益计算方式 | 数据特征 | 数据特征 | #*bin*容器特征 |
| 内存开销 | 大 | 大 | 小 |
| 类别特征 | 需做特殊处理 | One-hot编码 | 直方图数据处理 |
| 并行计算 | 不可并行 | 特征并行 | 特征、数据并行 |
| 是否支持缺失值处理 | 不支持 | 支出 | 支持 |

|  |
| --- |
|  |

2.2.2 并行集成算法介绍

集成学习中的并行集成算法，个体学习器之间不存在强依赖关系，可同时并行生成个体学习器。最主要的代表是Bagging和随机森林。

Bagging（Bootstrap Aggreating自助聚集）是并行式集成算法最著名的代表。Bgging的基本思想是基于自助采样法（bootstrap sampling）采样出T个训练集，然后基于每个采样出的数据集合训练出一个基学习器，再将这些基学习器进行组合得到最终的结果。对基学习器的预测输出进行结合时，分类任务通常使用简单投票法，回归任务使用简单平均法[43]。算法的流程如图2-7所示：



图2-7 Bagging算法流程

以随机森林为例，随机森林（Random Forest）由LeoBreiman（2001）提出。RF的基本思想是从原始训练样本集N中有放回地重复随机抽取k个样本生成新的训练样本集合，然后根据自助样本集生成k个分类树组成随机森林[44]。其本质是对决策树算法的改进，将多个决策树的结果合并融合在一起，每棵树的建立依赖于独立抽取的样本，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和他们之间的相关性，因此，随机森林更注重降低模型的方差[44]。

2.2.3 CatBoost算法介绍

CatBoost是由Yandex的研究人员和工程师于2017年推出的基于梯度提升决策树的能够很好处理类别特征的机器学习算法。当前流行的大多数梯度增强实现算法都将决策树用作基础预测变量，在用于数据集全部都是数字型特征的情况下很方便，但实际上，许多数据集中不仅包含数字型特征，同样也包括类别特征，这些类别特征对于预测也很重要[45]。所谓类别特征，是指这类特征是以离散的集合形式存在，例如，学历（高中、本科、硕士等）。在现有的梯度提升算法中，对类别特征的常见处理方式是将这些类别特征转化为数值型，一般类别特征会转化为一个或多个数值型特征，如XGBoost对类别特征采用One-Hot编码。该算法可以很好地处理各种类型的数据，如文本、音频、图像等，被广泛用于各种类型的商业挑战，如欺诈检测、推荐项目、预测等等。

1. 类别特征处理方式

常见的处理方式中，将类别特征加工成目标变量统计值TS（Target Statistics），但这种方法主要缺陷在于类别型特征中低频次特征会带来一定的噪声[46]。为了解决这一问题，CatBoost算法采用了Greedy TS，在公式中添加先验分布项。TS值的计算是通过估算每个类别目标变量的平均期望，再计算样本标签编码值的统计总量作为类别型特征[46]，具体公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-22） |

其中， 是样本k的特征j， 是样本k中的特征p， 是样本j的标签编码值，m是先验项，a是权重系数。

CatBoost在类别型特征处理上，另一重要的优化是将不同类别型特征的组合作为新的特征，以获得高阶依赖（high-order dependencies）。为当前树构造新的分割点时，CatBoost会采用贪婪的策略组合。在树的第一次分割时，不考虑任何特征组合。对于下一次分割，CatBoost将当前树的所有组合、类别型特征与数据集中的所有类别型特征相结合，并将新的组合类别型特征动态地转换为数值型特征[47]。

1. 梯度偏差

CatBoost和所有标准梯度提升算法一样，通过构建新的树来拟合当前模型的梯度，这种方法天然存在由有偏的点态梯度估计引起的过拟合问题[47]。同时，使用的梯度都是用当前模型中相同的数据来估计，导致梯度在特征空间的任何域中的分布与该域中梯度的真实分布相比发生了偏移。在CatBoost中，对经典的梯度提升算法进行了一些改进，主要采用了梯度步长的无偏估计。

假设 为构建i棵树的模型， 为构建i棵树后的第k个训练样本上面的梯度值。对于每一个样本 ，我们训练一个单独的模型 ，该模型不使用该样本的梯度估计进行更新，使用 估计上 的梯度（即叶子节点的值），并使用这个估计对结果树进行评分。算法流程如下：

其中，Loss( , a)是需要优化的损失函数，y是标签值，a是公式计算值。这里为了进一步减少过拟合的影响，会使用若干不同扰动来加强算法鲁棒性 。对用于计算的类别特征，获取s个随机扰动排列，对于每个排列，训练n个不同的模型。



2.2.4 SVM算法介绍

支持向量机（Support Vector Machine,SVM）的基本原理就是在特征空间中找到最佳的分离超平面，使间隔最大，这样的超平面可以使得训练集上正负样本间隔最大[48]。SVM是可以用来解决二分类问题的有监督算法，在引入核方法之后可用于解决非线性问题。

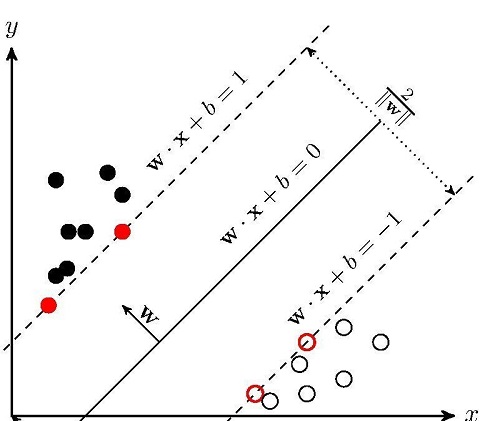


图2-8 SVM示意图

对于图2-8中的三个分割超平面均可以使得数据分类正确，但最好的平面应该是平面C，因为使用超平面C进行划分对训练样本的局部扰动的“容忍”度最好，分类的鲁棒性最强。

1. 硬间隔支持向量机

给定训练样本集D= ， ，i标识第i个样本，n标识样本容量。分类学习器的任务就是在特征空间中找到最佳划分超平面将正负样本分开，SVM的超平面和分类决策函数如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-23） |
|  |  | （2-24） |

为了找到具有最大间隔的最佳超平面，需满足如下的参数约束：

|  |  |  |
| --- | --- | --- |
|  |  | （2-25） |

式2-25等价于：

|  |  |  |
| --- | --- | --- |
|  |  | （2-26） |

该问题为一个凸二次规划问题，转化为拉格朗日对偶问题来解决，拉格朗日函数可写为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-27） |

其中， 分别是每个样本对应的拉格朗日乘子。对拉格朗日函数求偏导可得：

|  |  |  |
| --- | --- | --- |
|  |  | （2-28） |

将式2-28带入对偶问题得：

|  |  |  |
| --- | --- | --- |
|  |  | （2-29） |

求解出之后，再求解出和b即可得到SVM决策模型：

|  |  |  |
| --- | --- | --- |
|  |  | （2-30） |

（2）核函数

现实任务中原始的样本空间D很可能并不存在一个完全将正负样本分开的分隔超平面。当训练数据线性不可分时，可通过核方法以及软间隔最大化学得一个非线性支持向量机[49]。核函数实现了计算在低维计算，返回在高维的形式：

|  |  |  |
| --- | --- | --- |
|  |  | （2-31） |

原本SVM的优化目标是：

|  |  |  |
| --- | --- | --- |
|  |  | （2-32） |

假设 和 都是低维非线性函数，我们定义映射到高维的函数为 ，则上面的式子可写为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-33） |

常见的核函数有：

表2-3 常见得核函数

|  |  |  |
| --- | --- | --- |
| 名称 | 表达式 | 参数 |
| 线性核 |  | 无 |
| 多项式核 |  | N为多项式的次数 |
| 高斯核（RBF） |  | 为高斯核的带宽 |
| 拉普拉斯核 |  | >0 |
| Sigmod核 |  | tanh为双曲正切函数 |

## 2.3 模型融合理论

在机器学习中存在一个普适定理—没有免费的午餐（No Free Lunch Theorem，NFL定理）。该定理可描述为，对所有可能的目标函数求平均，得到的所有算法的“非训练集误差”的期望值相同，NFL定理表明没有一个学习算法可以在任何领域总是产生最准确的学习器[50]。因此我们可以以适当的方式将多个模型融合，发挥各个模型的长处，扬长避短，提升模型结果的准确度。模型融合就是训练多个模型，按照某种策略集成成为一个更复杂、泛化性能更好的模型。

可以通过数学证明，随着集成中个体分类器数量T的增大，集成的错误率将以指数级下降，最终趋向于零[51]。同时，模型融合要求个体学习器的准确性越高，多样性越大，则模型融合的效果会越好[51]。以下运用误差-分歧分解以回归任务为例证明。

假设有T个个体学习器为 ，通过加权平均法结合生成H完成回归任务，

,对样本x，定义学习器 的分歧为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-34） |

则集成H的分歧为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-35） |

这里的分歧代表了个体学习器的多样性。个体学习器 的平方误差为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-36） |

同理，H的平方误差为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-37） |

个体学习器误差的加权均值为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-38） |
|  |  | （2-39） |

令 表示样本的概率密度，则在全样本上有：

|  |  |  |
| --- | --- | --- |
|  |  | （2-40） |

类似的，对于个体学习器 ，在全部样本上的泛化误差和分歧为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-41） |
|  |  | （2-42） |

集成的泛化误差为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-43） |

代入个体学习器的泛化误差和分歧得到：

|  |  |  |
| --- | --- | --- |
|  |  | （2-44） |
|  |  | （2-45） |

其中，表示个体学习器泛化误差的加权均值，表示个体学习器的加权分歧值。上式说明，个体学习器准确性越高，多样性越大，则融合的效果越好。

模型融合的策略有如下几种：

1. 投票法

这是一种最简单的模型融合方法，对于要融合的个体学习器采取投票机制的方法，投票多者确定为最终的分类结果。投票法包括绝对多数投票法、相对多数投票法、加权投票法等。

1. 平均法

对于回归任务来说，最常见的融合策略就是使用平均法。对于每个个体学习器，最终的结果为所有个体学习结果的平均。主要包括，简单平均法、加权平均法。

简单平均法公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-46） |

加权平均法公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-47） |

其中， 是个体学习器 的权重。

1. Stacking算法

除了投票法和平均法，还可以使用一个学习器将个体学习器进行结合，称之为学习法，典型的代表之一是Stacking[52]。在学习法中，个体学习器称为初级学习器，用于结合个体学习器的称之为次级学习器或元学习器。

Stacking算法的步骤如下：

1. 对于初始训练数据集D，运用交叉验证法，先从训练集中运用T个初级学习算法训练出T个初级学习器。
2. 针对测试集中的每一个样本，用T个初级学习器分别预测得到对应的输出，类别标记为它原来的标记，并用次级学习器进行训练。



以5折交叉验证为例说明。

Stacking的思想如图2-9，图中所采用的为五折交叉验证：

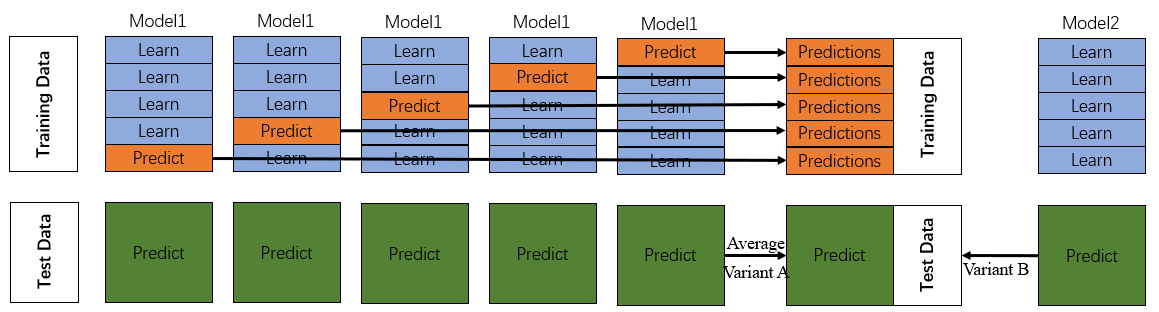


图2-9 Stacking五折训练图

假设整个样本集包含12000条数据，分为训练集training data 10000条和测试集testing data 2000条。

第一层采用四个模型（假设为RF、ET、GBDT、XGB），分别对训练集进行五折交叉训练，然后将预测的结果作为下一层的输入。

Step1：将训练集分为5折

1. 分别用第2、3、4、5折训练一个RF，用训练好的RF直接预测第1折训练数据；
2. 分别用第1、3、4、5折训练一个新的RF，用训练好的RF直接预测第2折训练数据；
3. 分别用第1、2、4、5折训练一个新的RF，用训练好的RF直接预测第3折训练数据；
4. 分别用第1、2、3、5折训练一个新的RF，用训练好的RF直接预测第4折训练数据；
5. 分别用第1、2、3、4折训练一个新的RF，用训练好的RF直接预测第5折训练数据。

第一步训练完成，可以得到10000\*1的RF对training data的预测结果，对于testing data，用上面得到的五个RF，预测出2000\*5维的预测结果，然后对其取平均，得到最终的2000\*1的RF预测结果。

Step2: ET、GBDT、XGB三个模型同理。

最终第一层中，training data会输出10000\*4的预测结果，将这个结果作为第二层训练集的输入。Testing data会输出2000\*4的结果，将这个结果作为第二层预测集的输入。

第二层将上一层的结果带入新的模型中，进行训练再预测，第二层的模型一般为了防止过拟合会采用比较简单的模型，如LR等。

1. Blending算法

Blending思想与Stacking算法思想类似。Blending的第一层通过将训练集留出一部分作为holdout set，之后通过剩下的数据生成模型对holdout set进行预测[53]。Blending的第二层直接对预测结果建模生成第二层的模型，不同于Stacking使用的交叉验证法。如下图所示：

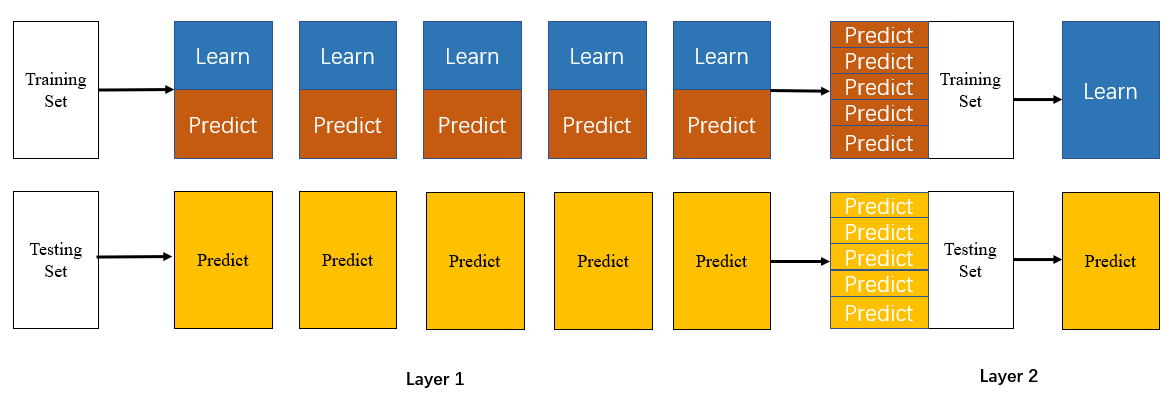


图2-10 Blending算法训练预测图

## 2.4 预测模型评价标准

对于预测模型的评价，即性能度量，是衡量模型泛化性能的标准，对于不同的模型，性能度量反映了任务的需求，不同的评价标准可能会导致不同的评判结果[54]。在分类任务中，最常用的性能度量指标是错误率和精度，还包括查准率、查全率、AUC等。

1. 错误率和精度

精度是指分类正确的样本占样本总数的比例，错误率是指分类错误的样本数占样本总数的比例。具体公式如下：

错误率定义为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-48） |

精度定义为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-49） |

更一般的，使用概率密度函数 来描述错误率和精度：

|  |  |  |
| --- | --- | --- |
|  |  | （2-50） |
|  |  | （2-51） |

（2）查准率和查全率

精度和错误率并不能满足所有任务的需求，比如在判断是否为好瓜时，我们关心有多少比例的好瓜被挑选出来，挑出的好瓜当中有多少是真实的好瓜。在二分类任务中，分类结果的混淆矩阵如下：

表2-4 分类结果的混淆矩阵

|  |  |  |
| --- | --- | --- |
| 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） |

查准率P定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-52） |

查全率R的定义为：

|  |  |  |
| --- | --- | --- |
|  |  | （2-53） |

（3）P-R曲线和F1

P-R曲线是描述查准率和查全率变化的曲线，P-R曲线的绘制是对学习器的预测结果概率值进行排序，将最可能是正例的样本放在前面，最不可能是正例的放在后面，按此顺序依次将样本作为正例预测，依次计算出当前的查准率和查全率，绘制图示例如2-11所示：

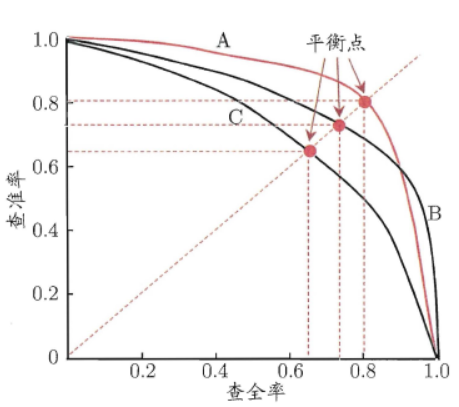


图2-11 P-R图示例

P和R有时会出现矛盾的情况，这时我们需要综合考虑查准率和查全率，最常见的方法就是F-Measure，又称为F-Score，F-Measure是查全率和差准率的加权调和平均，其表达式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-54） |

特别地，当 为1时，就是常用的F1度量，公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-55） |

（4）ROC与AUC

ROC曲线与P-R曲线类似，按照预测概率的大小排列并逐一按照正例预测，不同的是ROC曲线以“真正例率”（TPR）为纵轴，“假正例率”（FPR）为横轴绘制曲线，ROC主要侧重于研究样本判断概率排序的优劣[55]。ROC曲线的示意图如下：

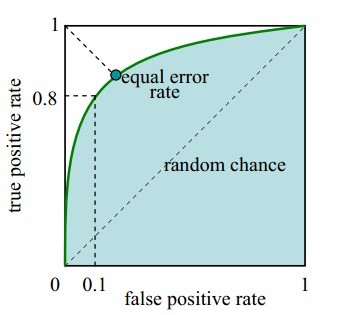


图2-12 ROC图示例

TPR和FPR的定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-56） |

ROC曲线下的面积定义为AUC（Area Uder ROC Curve），AUC的计算公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | （2-57） |

AUC值越大，预测概率排序的质量越好，模型的效果越好。

## 2.5 本章小结

本章主要是相关知识的介绍。首先是对集成学习思想的概述，主要包含了Boosting和Bagging，分析对比了它们的不同点。其次是机器学习算法的一些理论研究介绍，主要包含了GBDT、XGBoost、LightGBM、Random Forest、Catboost、SVM以及模型融合的理论介绍。最后提出了信贷风险预测模型的评价指标。

# 第3章 **基于多模型融合的信贷风险预测模型研究**

## 3.1 数据探索及数据预处理

3.1.1 数据特征分析

本论文研究的基于多模型融合的信贷风险预测模型的数据来源于数据挖掘竞赛—用户贷款风险预测。用于训练的数据包括用户的基本属性、用户银行卡流水记录、用户浏览行为、信用卡账单记录、放款时间以及这些顾客是否发生逾期行为的记录。相应地，还有用于测试的用户基本属性、银行流水、信用卡账单记录、浏览行为、放款时间以及待预测用户的ID列表。训练数据和测试数据都做了相应的脱敏处理，隐藏了用户的ID信息，所有金额均做了函数变换。

数据探索EDA（Exploratory Data Analysis）是指对原始数据通过统计、作图、计算相关性等手段初步探索数据的结构和规律的数据分析方法，是数据挖掘过程中的重要一环[56]。数据探索的主要价值在于快速熟悉了解和把握数据集，了解变量之间的相关关系以及变量与预测值之间存在的关系等，为数据挖掘后续过程的数据结构和特征集做准备[56]。

（1）用户基本信息表

用户基本信息表中包括职业、性别、教育程度、婚姻状态和户口类型，对于每个属性的每一个离散值逾期的比例如图3-1所示：

图3-1 每个属性的离散值逾期比例

图3-1的柱状图描述的是用户基本属性的属性值与是否逾期的比例关系，以用性别为例，用户性别为0的用户逾期占比高达近40%，高于性别为1和2的用户，说明用户性别为0的逾期信贷风险可能较大，与预测值的相关性较大。

1. 银行卡流水记录表

银行卡流水记录主要包括用户标识、流水时间、交易类型、交易金额和工资收入标记。放款前/后用户收入和支出笔数、以及放款前/后用户的收入占比分布如下：

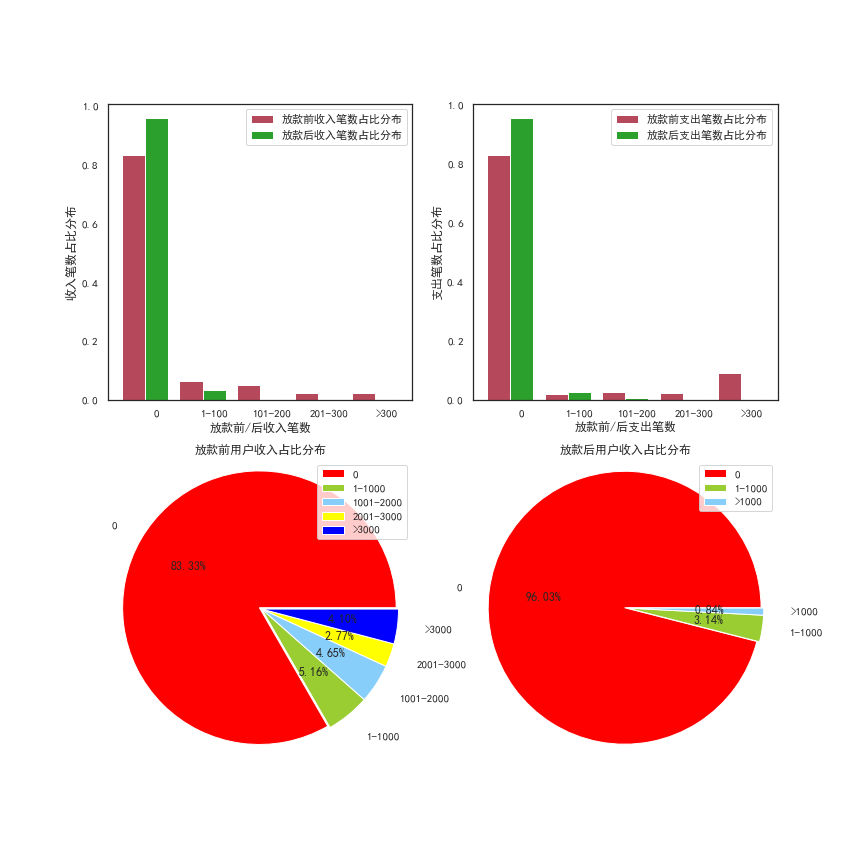


图3-2 用户银行卡流水记录探索

从图3-2中的柱状图可以明显看出放款前后收入和支出的笔数占比分布发生较为明显的变化，放款前用户的支出笔数为0的占比约0.83，放款后约为0.96，也就是说，放款后的消费变低。从饼图可以看出，放款前后的收入占比为0的均达80%以上，占据主体地位。

1. 用户浏览行为表

用户浏览行为包括用户标识、浏览时间、浏览行为数据和浏览子行为编号。通过放款时间将用户的浏览行为分为放款前和放款后，观察放款前后用户行为的变化。

从散点图的对比可以看出，用户在放款前和放款后的浏览行为数据的分布发生较为明显的变化，浏览行为逐渐向放款后聚集。箱图显示的是放款前后浏览行为数据mean的分布变化，可以看出，放款后浏览行为平均值较放款前增多。

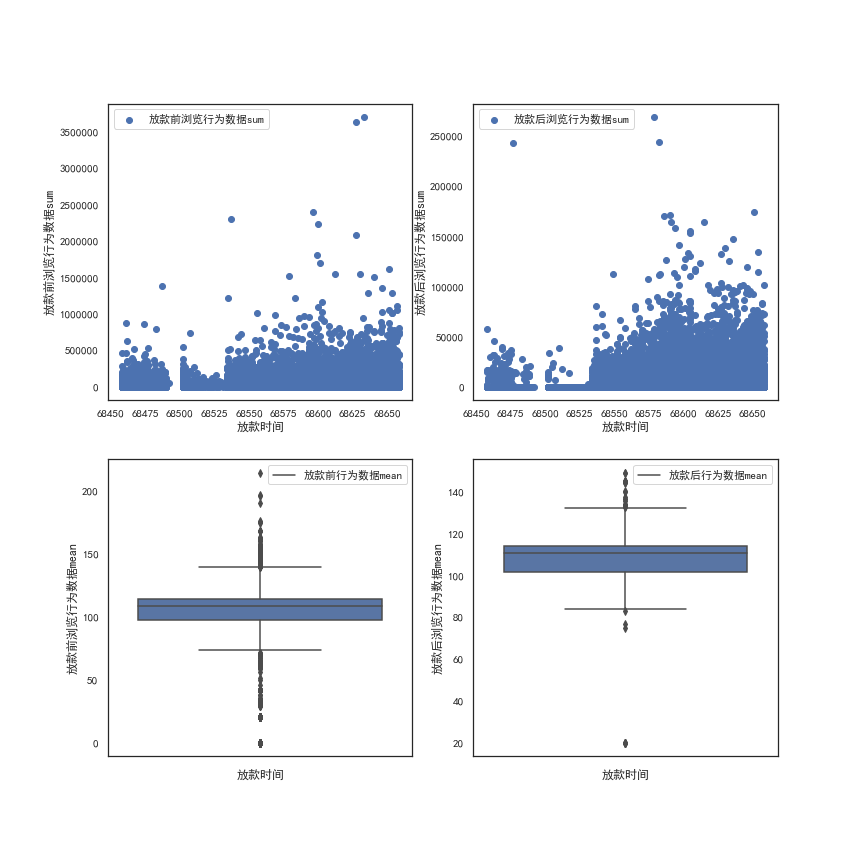


图3-3 放款前后用户浏览行为变化图

3.1.2 数据清洗及缺失值处理

数据是整个建模过程中的基础，数据的质量很大程度上决定了模型最后的效果。在数据挖掘过程中，数据清洗的主要目的是基于数据探索的一些结论，对异常值、缺失值、重复值、噪音数据等进行处理，提高数据的质量，为后续特征工程打好基础[57]。

解决数据质量问题主要包含五类问题[57]：

* 数据的完整性，在本论文研究的数据中，存在大量的数据缺失，如用户的银行消费记录等。
* 数据的唯一性，如不同来源的数据出现重复的情况。
* 数据的权威性，如同一属性不同来源的数据不相同。
* 数据的合法性，违背常识的数据，如人的年龄为1000岁。
* 数据的一致性，如同一指标含义不一致。

1. 异常值处理

数据中的离群点，它是显著不同于其他数据对象，在数据分布上与其他数据有显著的不同，离群点亦称异常数据。离群点的存在对于模型的训练会有一定的影响[58]。离群点大致来源于三种情况，第一种是数据来自不同的类，第二种是数据的自然变异，第三种是数据测量或收集误差。通常，离群点会显著违反数据或模型的假设分布，因此，我们需要对离群点进行检测并剔除，减少后续建模的影响[58]。

常用的离群点检测方法有：

* 基于统计模型的方法，即建立一个数据模型来预测相对远离的对象。
* 基于邻近度的方法，通过在对象之间定义邻近性度量识别异常点。
* 基于密度的方法，当一个点的局部密度显著低于它的大部分近邻时判定为异常点。
* 基于聚类的方法。聚类可以识别相同数据类型的点，用聚类分析的方法就可以识别远离聚类簇的那些离群点。

DBSCAN是一种基于密度聚类的算法。密度聚类可以通过样本分布的紧密程度来做聚类分析。密度聚类的假设是同一类别的样本，它们之间是紧密相连的，将紧密相连的样本归为一类，得到一个聚类类别[59]。DBSCAN的思想很简单，首先任意选择一个没有类别的核心对象作为种子，然后找到所有这个核心对象能够密度可达的样本集合形成一个聚类簇，接着继续选择另一个没有类别的核心对象去寻找密度可达的样本集合，直到所有核心对象都有类别停止[59]。

DBSCAN的算法流程如下：



通过TSNE降维和DBSCAN的密度聚类异常点检测效果如图3-4：

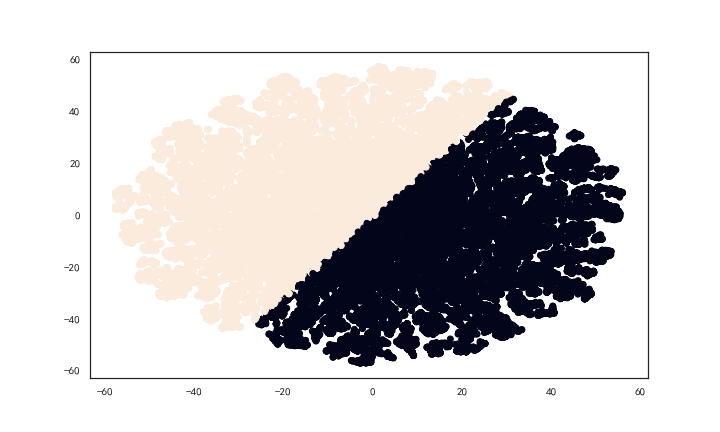


图3-4 密度聚类异常点检测效果图

从图3-4来看，本论文研究的数据中异常点较少，对于离群点予以剔除。

（二）缺失值处理

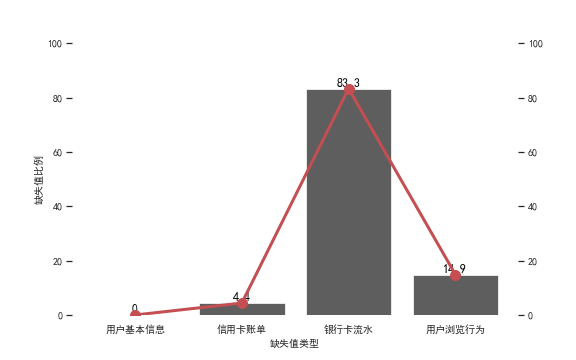
在现实的数据中，数据的缺失往往是不可避免的。信息的遗漏或者是特征属性不存在均会导致数据的部分缺失。在本文研究的数据中存在大量的缺失，各部分数据缺失的比例如图3-5所示：

图3-5 各部分缺失值比例

从图3-5中可以看出，用户基本信息无缺失，缺失值比例最多的是银行卡流水记录，仅包含16.7%的用户记录，大部分用户的银行卡流水记录缺失不存在。其次是用户浏览行为记录，存在14.9%的缺失，最后是信用卡账单的缺失比例为4.4%。

对于数据缺失值的处理，主要有三种方式：

1. 直接删除记录

该方法简单易操作，在整体数据量很大且缺失值数据较少的情况下比较有效，但若数据量较少或者缺失值比例较大的话不建议使用，会牺牲大量的数据，丢失很多影藏的信息，甚至导致数据的分布发生变化。

1. 缺失值填补

对缺失值按照一定规律或方法进行填充，主要的方法如表3-1所示：

表3-1 缺失值填补的方法

|  |  |
| --- | --- |
| 插补方法 | 方法描述 |
| 平均值/中位数/众数 | 用属性值的平均值/中位数/众数进行插补 |
| 固定值 | 使用某一固定值插补，如-1或0等 |
| 最临近值插补 | 与缺失值最接近的样本值进行插补 |
| 回归预测 | 通过已有数据与其他变量建立回归模型预测缺失值 |
| 插值法 | 利用已知点建立合适的插值函数并计算出缺失值 |

（3）不处理

无论是删除缺失值或者是插值，都可能引入新的噪声数据，不符合原本数据的分布和客观事实。因此，在许多情况下，希望保持原始信息不变的前提下进行处理。在建模过程中，有些算法无法直接处理带有缺失值的数据，也有一些模型本身就可以处理带有缺失值的样本，如XGBoost、LightGBM、CatBoost等。

综上，对于缺失值的处理，需要综合考虑数据缺失的原因、数据缺失的类型、缺失比例等因素，选取最适合该数据分布的缺失值处理方法。

## 3.2 特征工程构建

特征工程是利用数据领域的相关知识来创建能够使得机器学习算法到最佳性能的特征的过程[60]。在数据挖掘的整个过程中，数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限。特征工程在机器学习中占有相当重要的地位，是机器学习效果成功的关键。特征工程主要可分为三部分，第一部分为特征构建，特征构建是指从最原始的数据中人工构建新的特征，需要我们花费大量的时间去研究真实的数据样本，思考数据的结构和潜在的形式[60]。第二部分是特征提取，特征提取的目的是自动地构建新的特征，将原始特征转换为一组具体明显物理意义或者统计意义的特征，如通过变换特征取值来减少原始数据中某个特征的取值个数等[60]。特征工程第三部分是特征选择，特征选择是从一个特征集合中挑选出一组最具统计意义的特征子集，达到降维的效果[60]。

3.2.1 基础特征构建及处理

（1）类别特征--用户基本信息表

用户基本信息表中的用户性别、用户职业、用户教育程度、用户婚姻状态和用户户口类型均为类别型变量，是离散和无序的，对这些属性采取One-Hot独热编码。One-Hot编码，又称为一位有效编码，主要采用N位状态寄存器来对N个状态进行编码，每个状态都由它独立的寄存器位，并且在任意时候只有一位有效。在独热编码之后，对特征进行组合构造，进一步增加特征的数量。

（2）非类别特征

基于信用卡账单表、银行卡流水记录表、用户浏览行为表中的数据均为脱敏数据，所以对该表做特征主要从统计特征出发。具体地，根据用户的账单时间和放款时间，将用户的所有账单分为放款前、放款后以及不区分放款前后这三种情况。统计特征主要包括求和sum，计数count，最大值max，最小值min，平均值mean，中位数median，方差std，标准差var。图3-6为放款前信用卡额度sum、放款后本期账单余额sum、放款前可用余额sum三个特征的变量关系组图：

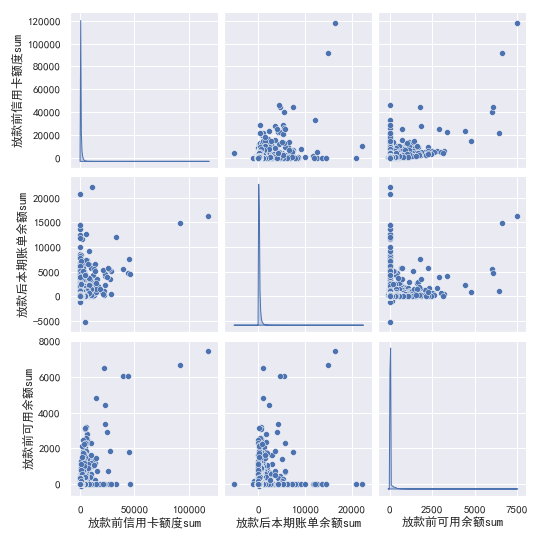


图3-6 特征变量关系组图

3.2.2 显著特征筛选

在特征工程中，特征选择是一个非常重要的过程。特征选择就是从众多特征的特征集合中选出对分类识别最有效的特征子集，实现特征空间维数的压缩，缓解维数灾难问题，同时去除不相关特征，降低学习器的学习任务难度[61]。特征选择的本质是使在最小维数特征空间中异类模式点相距较远，即类间距离较大，而同类模式点相距较近，即类内距离较小。特征选择的要求，首先要求所选择的特征具有很大的识别信息量，即所选择出的特征应具有很好的可分性，使学习器容易判别，其次，特征应具有可靠性，对于那些模棱两可的特征应筛掉，再次，特征之间应具有尽可能强的独立性、重复冗余或相关性强的特征只选一个，因为强的相关性并没有增加更多的分类信息。最后要求所选择的特征数量应尽可能少，同时损失的信息量尽量小[62]。

特征选择可分为两个阶段的任务，首先是子集搜索，可通过前向搜索或后向搜索，第二阶段是对选出的子集进行评价，如可采用信息增益对子集评价。常用的特征选择方法有三种，分别为Filter过滤法、Wrapper包裹法和Embedded嵌入法。

（1）Filter过滤法

Filter过滤法按照发散性或者相关性对各个特征进行评分，设定阈值或者选择阈值的个数选择特征。如Relief（Relevant Features），该方法假设特征子集的重要性是由子集中的每个特征所对应的相关统计量分量之和所决定的[63]。因此，只需要选出前k个相关统计量对应的特征即可。

 Relief在 的同类样本中寻找其最近邻 ，称为“猜中近邻”，再从 的异类样本中寻找其最近邻 ，称为“猜错近邻”，则属性j的相关统计量定义为：

|  |  |  |
| --- | --- | --- |
|  |  | （3-1） |

对于多分类可采用Relief-F：

|  |  |  |
| --- | --- | --- |
|  |  | （3-2） |

其中， 为第*l*类样本的比例。

过滤法做特征选择计算简单，计算量小，但往往会选择出冗余特征，效果一般不如Wrapper包裹法和Embedded嵌入法。

（2）Wrapper包裹法

包裹法特征选择直接将最终使用的学习器的性能指标作为特征选择时子集的评价准则，如递归消除特征法，它使用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集合进行下一轮的训练[64]。再如LVM（Las Vegas Wrapper），是典型的包裹式特征选择方法，每次随机选取一定数量的特征训练，并对模型效果评价，最终选择模型效果最好的那一组特征子集。包裹法从性能上看，比过滤式选择更好，但计算开销要大的多。

（3）Embedded嵌入法

Embedded嵌入法主要包含两种，一种是基于惩罚项的特征选择法，如基于L1正则化。该方法实现特征选择的原理是保留了多个对目标值具有同等相关性的特征中的一个。另一种常用的Embedded嵌入法是基于树模型的特征选择法，如GBDT、XGBoost等均可作为基模型用作特征选择，学习器在训练过程中自动进行了特征选择。

在使用嵌入法时，使用机器学习算法进行训练，得到每个特征的权重系数，即重要性系数，并根据权重系数的大小选取特征，这些权值系数实际上代表了该特征对于模型的贡献值或者重要性[65]。基于这种贡献值评估，筛选出对建立模型最有用的特征。

在本文研究的特征中，特征重要性排序如图3-7：

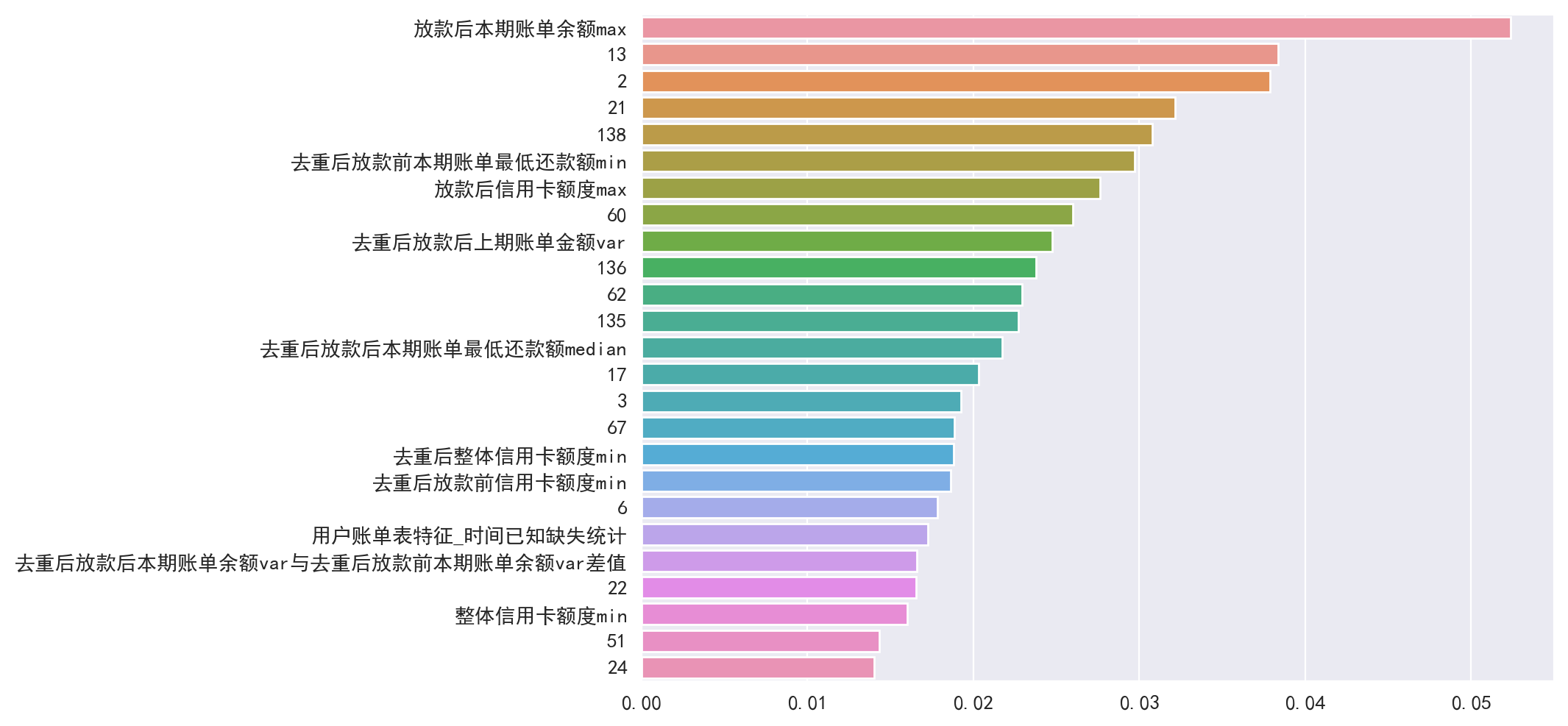


图3-7 特征重要性排序柱状图

从图中可以看出，特征重要性排名靠前的基本是信用卡账单和用户基本信息以及用户的浏览行为。其中，放款后本期账单预测max的贡献度最大，逾0.05。图中的数字特征为用户基本信息通过特征构造形成。在本文的研究中，选取了贡献度超过0.001的特征，共计209个用于后续模型的训练学习。

## 3.3 基于机器学习单模型仿真实验及对比分析

将本文研究的数据集随机分为两部分，80%用于模型的训练，剩余的20%用于模型的结果测试，并给出具体的模型评估指标结果及可视化。

3.3.1 仿真实验环境

关于基于机器学习的信贷风险预测研究仿真实验所使用的硬件和软件配置环境如表3-2和表3-3所示。

表3-2 实验硬件配置环境

|  |  |
| --- | --- |
| 硬件 | 详细信息 |
| CPU | Intel(R) Core(TM) i7-10710U CPU @ 1.10GHz |
| RAM | 16G |
| 硬盘 | 500GSSD |

表3-3 实验软件配置环境

|  |  |
| --- | --- |
| 软件 | 详细信息 |
| 操作系统 | Win10 64位 |
| 实验平台 | Python 3.6.1和Anaconda |
| Scikit-learn | 0.23.2 |
| lightgbm | 3.1.0 |
| Keras | 2.2.4 |

3.3.2 仿真实验及实验结果对比分析

在本文的研究中，单模型主要基于集成学习方法建立模型，集成学习方法比单个学习器具有更优的学习性能，主要包括两种，分别是序列化集成学习和并行化集成学习方法。本文采用了四种集成学习方法，包括GBDT、XGBoost、LightGBM、Random Forest。为了使得模型融合的个体学习器具有较大的差异性，加入了支持向量机SVM和BP神经网络，提升模型融合的整体效果。

1. 基于GBDT建立模型

对于GBDT的结构和参数设置，由于是进行分类任务的学习，使用GBDT的GradientBoostingClassifier分类类。对于分类模型，GBDT的损失函数有对数似然损失函数deviance和指数损失函数exponential两者，本文选用deviance。除此之外，还需要设置弱学习器的最大迭代次数以及学习率等。GBDT具体参数如表3-4：

表3-4 GDBT单模型参数设置

|  |  |
| --- | --- |
| 参数名称 | 设置值 |
| *loss* | deviance |
| *learning\_rate* | 0.01 |
| *n\_estimators* | 500 |
| *subsample* | 0.8 |
| *max\_features* | sqrt |
| *max\_depth* | 7 |

GBDT模型的训练和预测结果如图3-8和3-9所示：

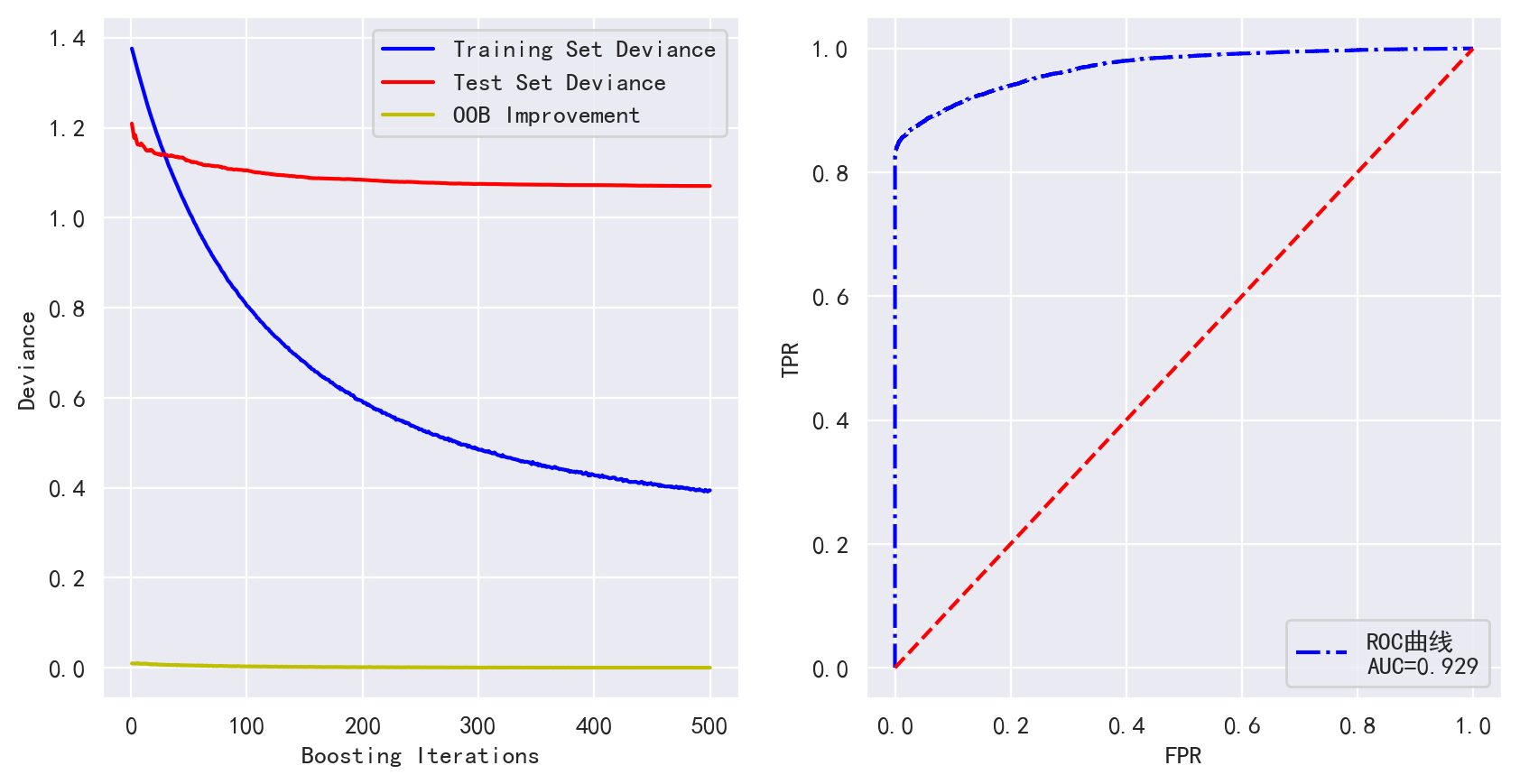


图3-8 GBDT模型预测损失变化和ROC曲线

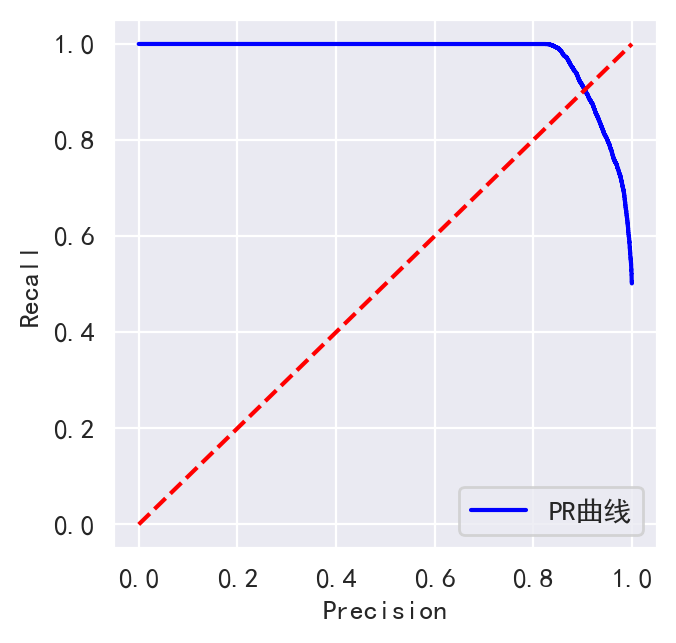


图3-9 GBDT模型预测结果的P-R曲线

图3-8左图为GBDT在训练时随着个体学习器的增加，训练集损失和测试集损失的变化图以及OOB Improvement的趋势变化。图3-8右图为模型的ROC曲线，AUC的值，即ROC曲线与坐标轴围成的面积为0.929。图3-9为模型结果的P-R曲线，虚红线与PR曲线的交点即为平衡点。

GBDT模型结果的混淆矩阵如表3-5：

表3-5 GBDT单模型预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.99 | 0.87 | 0.93 | 10986 |
| 1 | 0.86 | 0.98 | 0.92 | 8380 |
| accuracy |  |  | 0.92 | 19366 |
| macro avg | 0.92 | 0.93 | 0.92 | 19366 |
| wight avg | 0.93 | 0.92 | 0.92 | 19366 |

（2）基于XGBoost建立模型

XGBoost是GBDT的升级版，加入了正则化项、支持损失函数自定义、缺失值处理等。XGBoost的参数分为三类，分别是通用参数、Booster参数和学习目标参数。Booster参数一般可以调控模型的效果和计算代价。XGBoost模型的具体参数如表3-6所示：

表3-6 XGBoost单模型参数设置

|  |  |
| --- | --- |
| 参数名称 | 设置值 |
| *booster* | gbtree |
| *gamma* | 0.1 |
| *max\_depth* | 7 |
| *nthread* | 12 |
| *lambda* | 10 |
| *Eval\_metric* | auc |
| *min\_child\_weight* | 1.1 |
| *colsample\_bytree* | 0.7 |
| *Colsample\_bylevel* | 0.7 |

XGBoost在训练时损失值变化过程如图3-10所示：

图3-10的左图为XGBoost在训练集和测试集上的对数损失函数值变化曲线图，图3-10的右图为XGBoost分类器的错误率变化曲线图。从图中可以看出，随着训练迭代次数的增加，训练集和测试集上的误差都下降到较低位置并保持平缓。

图3-11展示的是XGBoost模型预测结果的ROC曲线和PR曲线，由图中可以看出，预测效果较好，AUC的值达到了0.936。

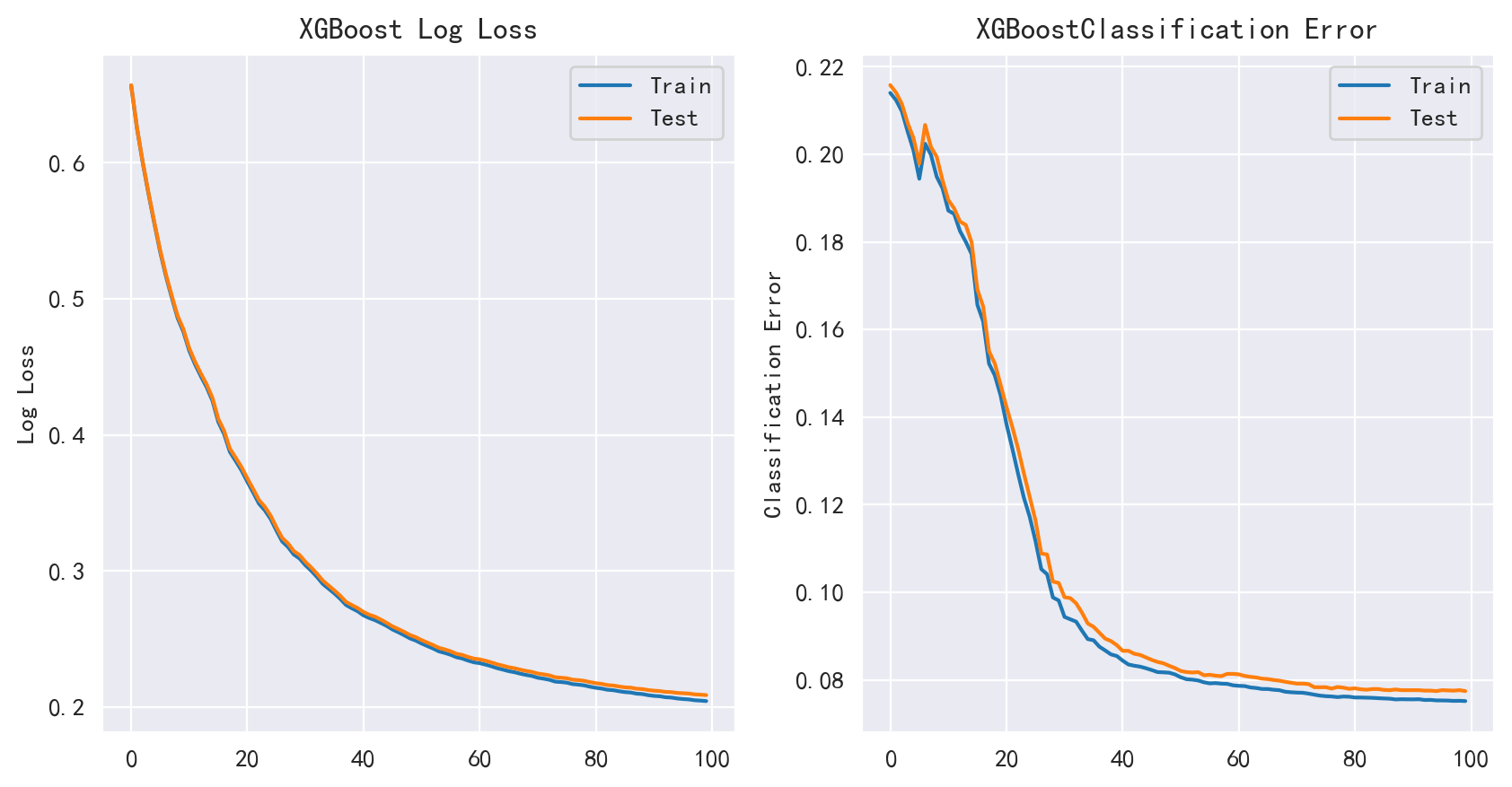


图3-10 XGBoost模型训练对数损失和错误率

XGBoost模型的ROC曲线和PR曲线图如图3-11：

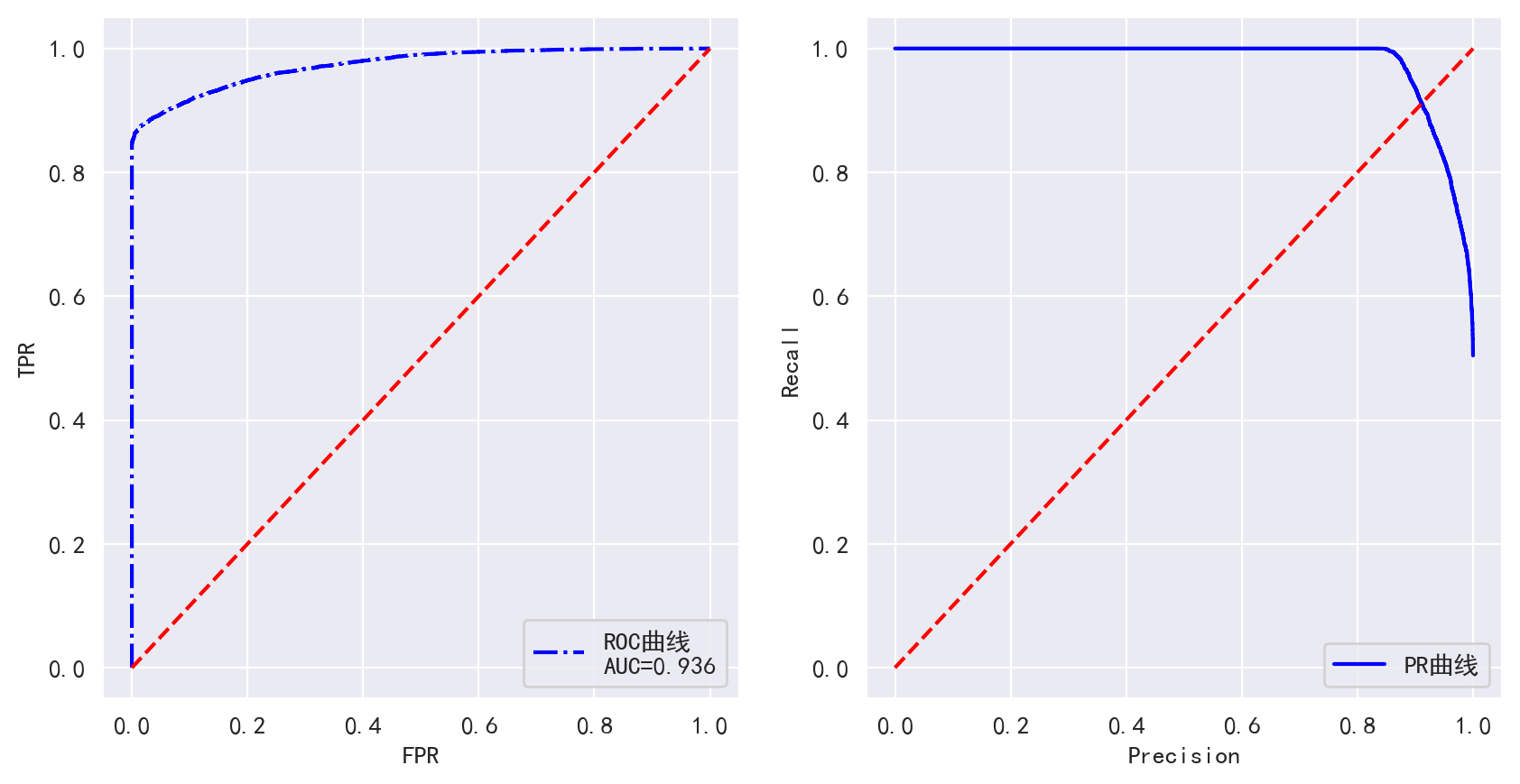


图3-11 XGBoost模型预测结果的ROC曲线和P-R曲线

XGBoost模型结果的混淆矩阵如下：

表3-7 XGBoost单模型预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.99 | 0.87 | 0.93 | 11051 |
| 1 | 0.85 | 0.99 | 0.92 | 8315 |
| accuracy |  |  | 0.92 | 19366 |
| macro avg | 0.92 | 0.93 | 0.92 | 19366 |
| wight avg | 0.93 | 0.92 | 0.92 | 19366 |

（3）基于LightGBM建立模型

LightGBM是在XGBoost基础上采用直方图算法做了进一步的优化。LightGBM的参数优化主要从两个方面出发，一个是从提高准确率角度，可对学习率、叶子节点数进行调优。另一种是从降低过拟合角度，调整工具箱数、L1正则和L2正则、执行切分的最小增益等。本论文研究的LightGBM模型参数具体如下：

表3-8 LightGBM单模型参数设置

|  |  |
| --- | --- |
| 参数名称 | 设置值 |
| *boostering\_type* | gbtree |
| *learning\_rate* | 0.1 |
| *max\_depth* | 10 |
| *num\_leaves* | 25 |
| *lambda\_l1* | 1 |
| *lambda\_l2* | 1 |
| *feature\_fraction* | 0.8 |
| *bagging\_fraction* | 1 |
| *min\_split\_gain* | 0.5 |
| *metric* | auc |

LightGBM模型预测结果如图3-12：

如图3-12所示，前两张图片是所有测试集上用户违约概率的分布图，违约概率0.5以上的明显少于违约概率为0.5以下的，且向两端靠拢。后两张分别为ROC曲线和P-R曲线，ROC曲线显示LightGBM模型结果的AUC值为0.931，与GBDT和XGBoost基本持平。

LightGBM模型结果的混淆矩阵如下：

表3-9 LightGBM单模型预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.99 | 0.87 | 0.93 | 11071 |
| 1 | 0.85 | 0.99 | 0.92 | 8295 |
| accuracy |  |  | 0.92 | 19366 |
| macro avg | 0.92 | 0.93 | 0.92 | 19366 |
| wight avg | 0.93 | 0.92 | 0.92 | 19366 |

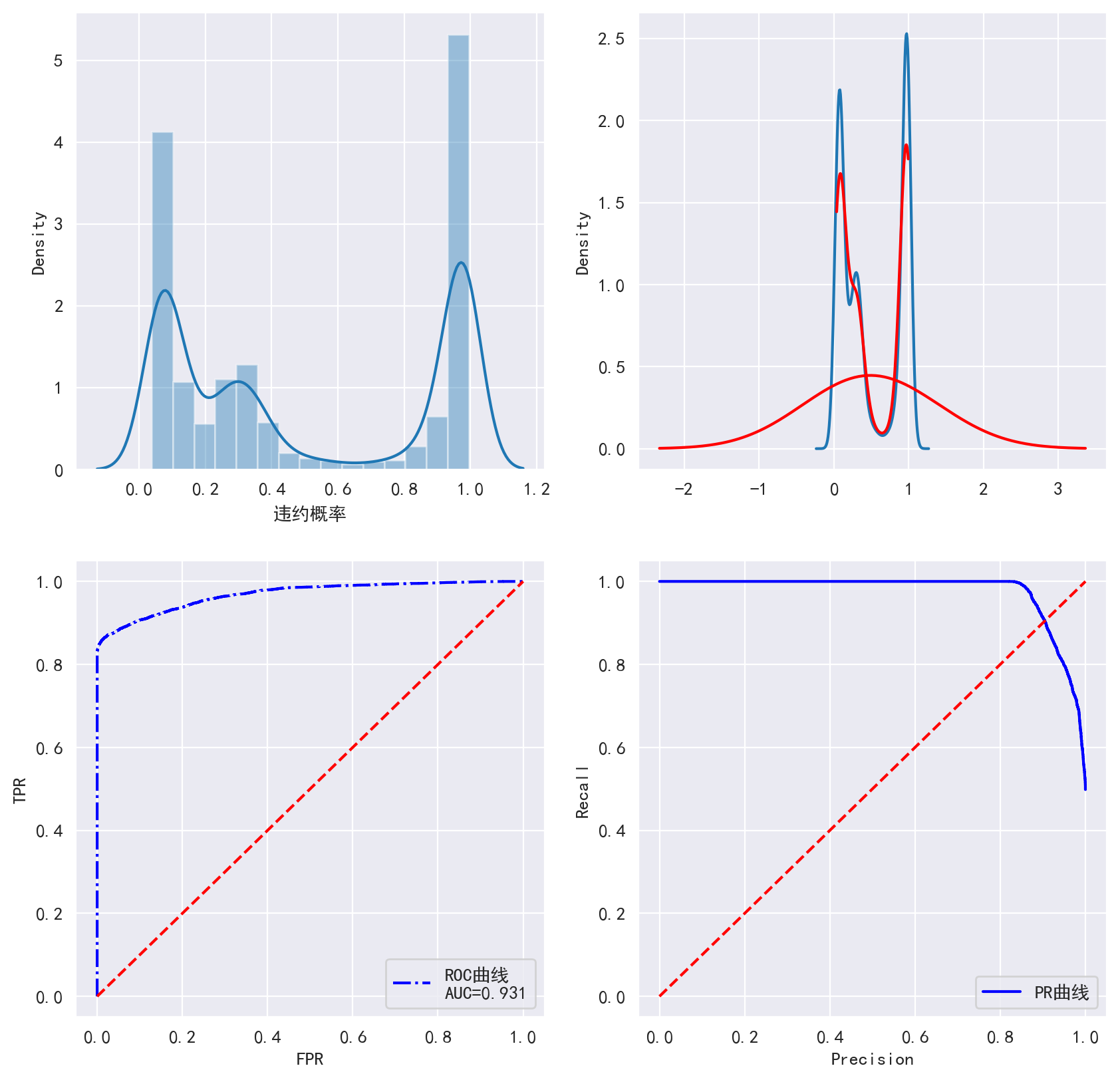


图3-12 LightGBM单模型预测概率分布及预测结果评价

（4）基于Random Forest建立模型

随机森林Random Forest属于并行化集成学习方法，是Bagging的扩展变体。在以决策树为基学习器构建Bagging基础上，进一步在训练过程中引入随机属性选择，降低模型的方差。关于参数，随机森林的参数可分为两类，一类是关于决策树的参数，如max\_depth等，另一类是随机森林特有的参数，如bootstrap等。本文研究中随机森林模型的参数如下：

表3-10 RF单模型参数设置

|  |  |
| --- | --- |
| 参数名称 | 设置值 |
| *n\_estimators* | 100 |
| *criterion* | gini |
| *max\_depth* | 10 |
| *min\_samples\_split* | 2 |
| *min\_samples\_leaf* | 1 |
| *min\_weight\_fraction\_leaf* | 1 |
| *min\_impurity\_split* | 1e-07 |
| *max\_features* | auto |
| *bootstrap* | True |

Random Forest模型预测结果如图3-13：

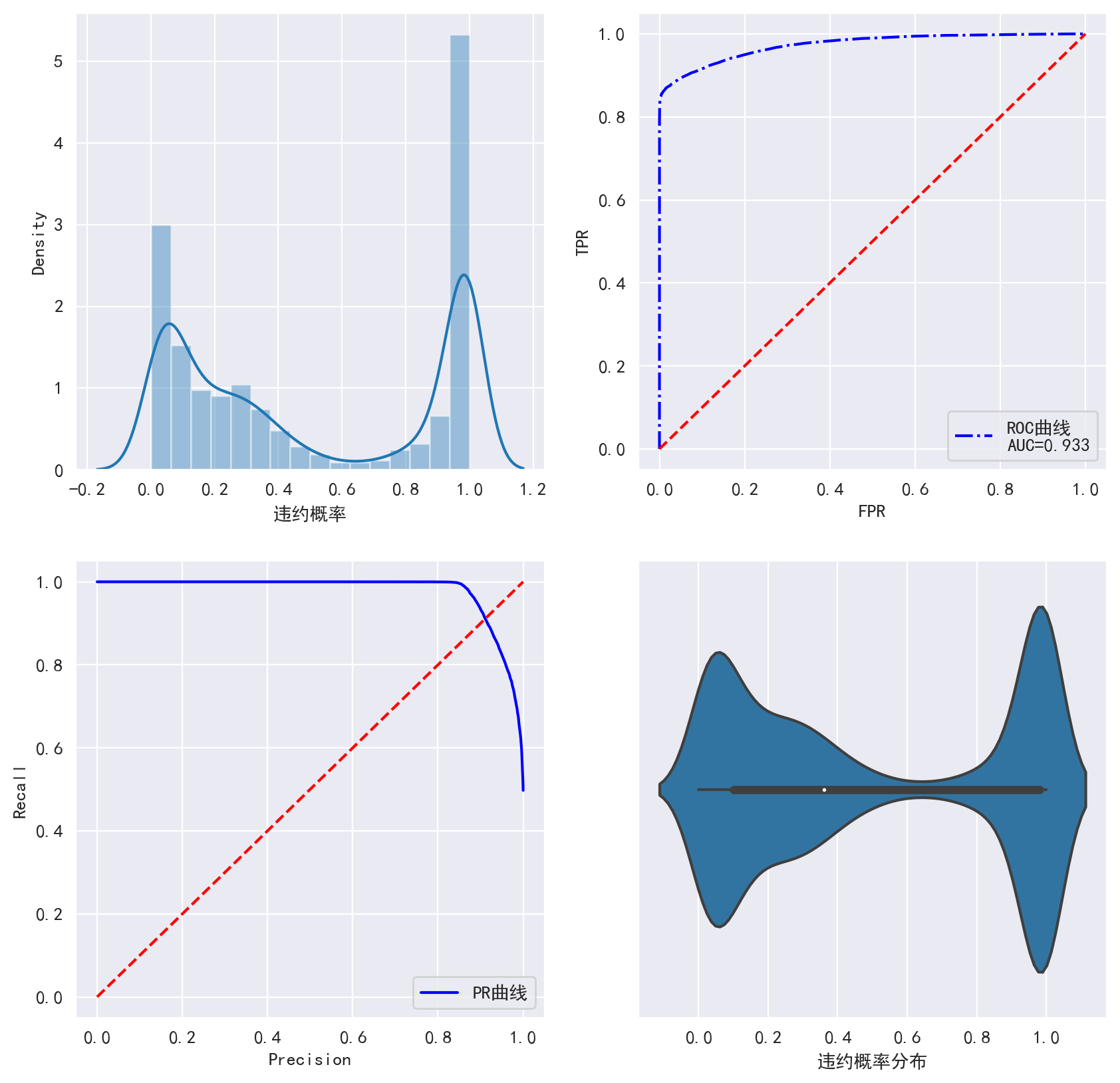


图3-13 RF单模型预测概率分布及预测结果评价

图3-13展示的是RF单模型预测结果的概率分布及评价结果。第一、四张图的柱状图和小提琴图直观地展示了预测概率的分布情况。第二、三张图分别是随机森林模型结果的ROC曲线和P-R曲线，AUC值为0.933。

RF模型结果的混淆矩阵如表3-11：

表3-11 Random Forest单模型预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.99 | 0.88 | 0.93 | 10906 |
| 1 | 0.87 | 0.99 | 0.92 | 8460 |
| accuracy |  |  | 0.93 | 19366 |
| macro avg | 0.93 | 0.93 | 0.93 | 19366 |
| wight avg | 0.93 | 0.93 | 0.93 | 19366 |

（5）基于BP神经网络建立模型

在使用Keras搭建神经网络时，调整参数至关重要。需要调整的参数有神经网络的层数、每层神经元的个数、激活函数等。在经过多次调整测试后，本文BP神经网络的参数如表3-12：

表3-12 BP神经网络单模型参数设置

|  |  |
| --- | --- |
| 参数名称 | 设置值 |
| *神经网络层数* | 3 |
| *隐层神经元个数* | 200 |
| *激活函数* | relu |
| *epochs* | 30 |
| *batch\_size* | 10 |
| *loss* | binary\_crossentropy |

BP神经网络模型预测效果如图3-14所示：

图3-14中四幅图分别是BP神经网络在训练时的损失值变化情况，违约概率核密度图，ROC曲线和PR曲线，AUC为0.914，比上述几种集成学习算法稍差。

模型结果的混淆矩阵如表3-13：

表3-13 BP神经网络单模型预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.99 | 0.83 | 0.91 | 11600 |
| 1 | 0.80 | 0.99 | 0.89 | 7766 |
| accuracy |  |  | 0.90 | 19366 |
| macro avg | 0.90 | 0.91 | 0.90 | 19366 |
| wight avg | 0.92 | 0.90 | 0.90 | 19366 |

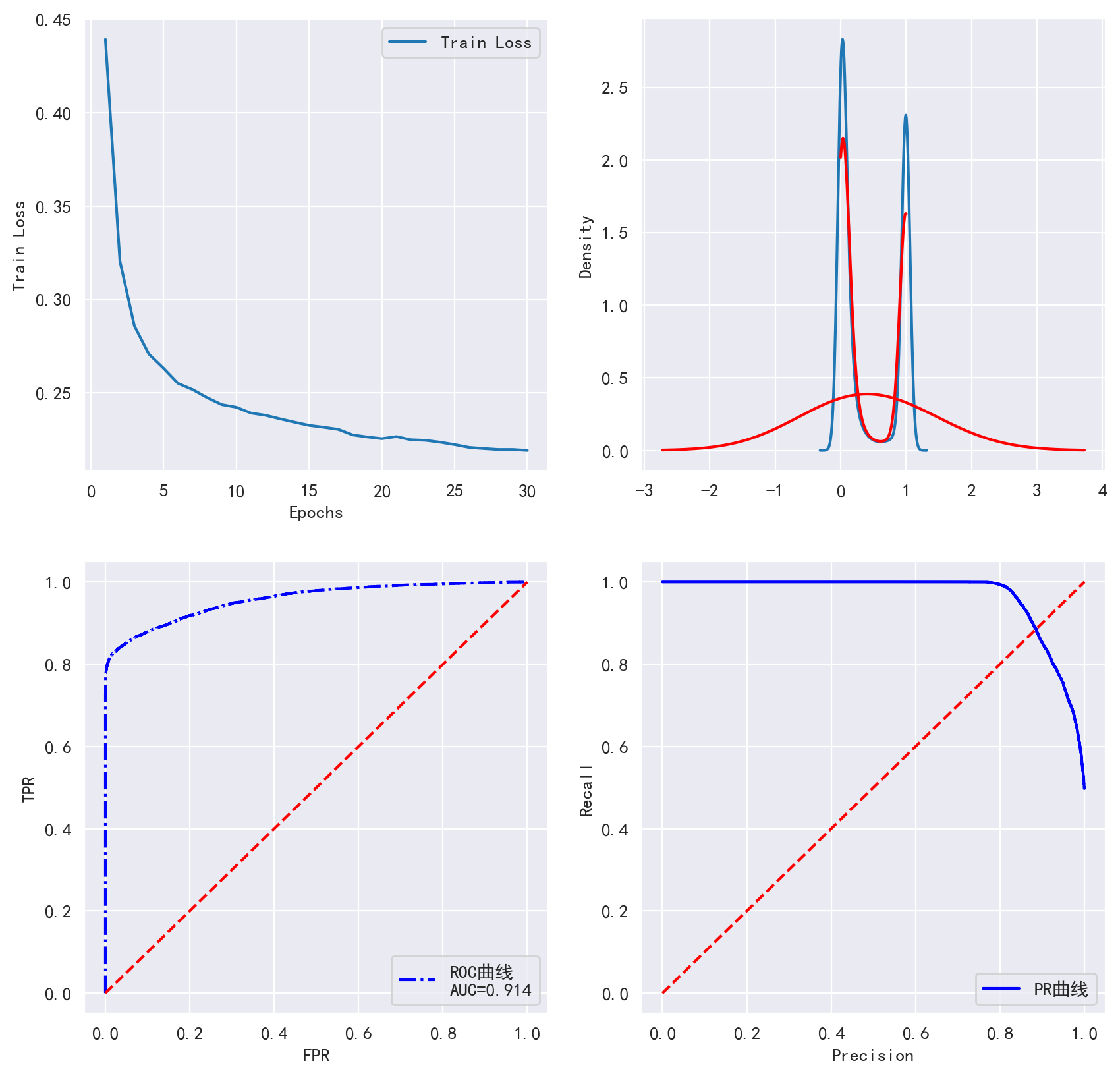


图3-14 BP神经网络训练损失变化及模型评价结果

（6）基于SVM建立模型

SVM模型有两个非常重要的参数C和gamma，其中C是惩罚系数，即对误差的容忍度，C越大，则越不能容忍出现误差，容易导致过拟合，C越小，容易欠拟合。gamma是选择RBF核函数之后自带的一个参数，隐含地决定了数据映射到新的特征空间后的分布，gamma越大，支持向量越少，而支持向量的个数影响训练与预测的速度。本文研究的SVM模型参数如表3-14：

表3-14 SVM单模型参数设置

|  |  |
| --- | --- |
| 参数名称 | 设置值 |
| *kernel* | rbf |
| *C* | 2 |
| *gamma* | 0.125 |

SVM模型的效果如图3-15：

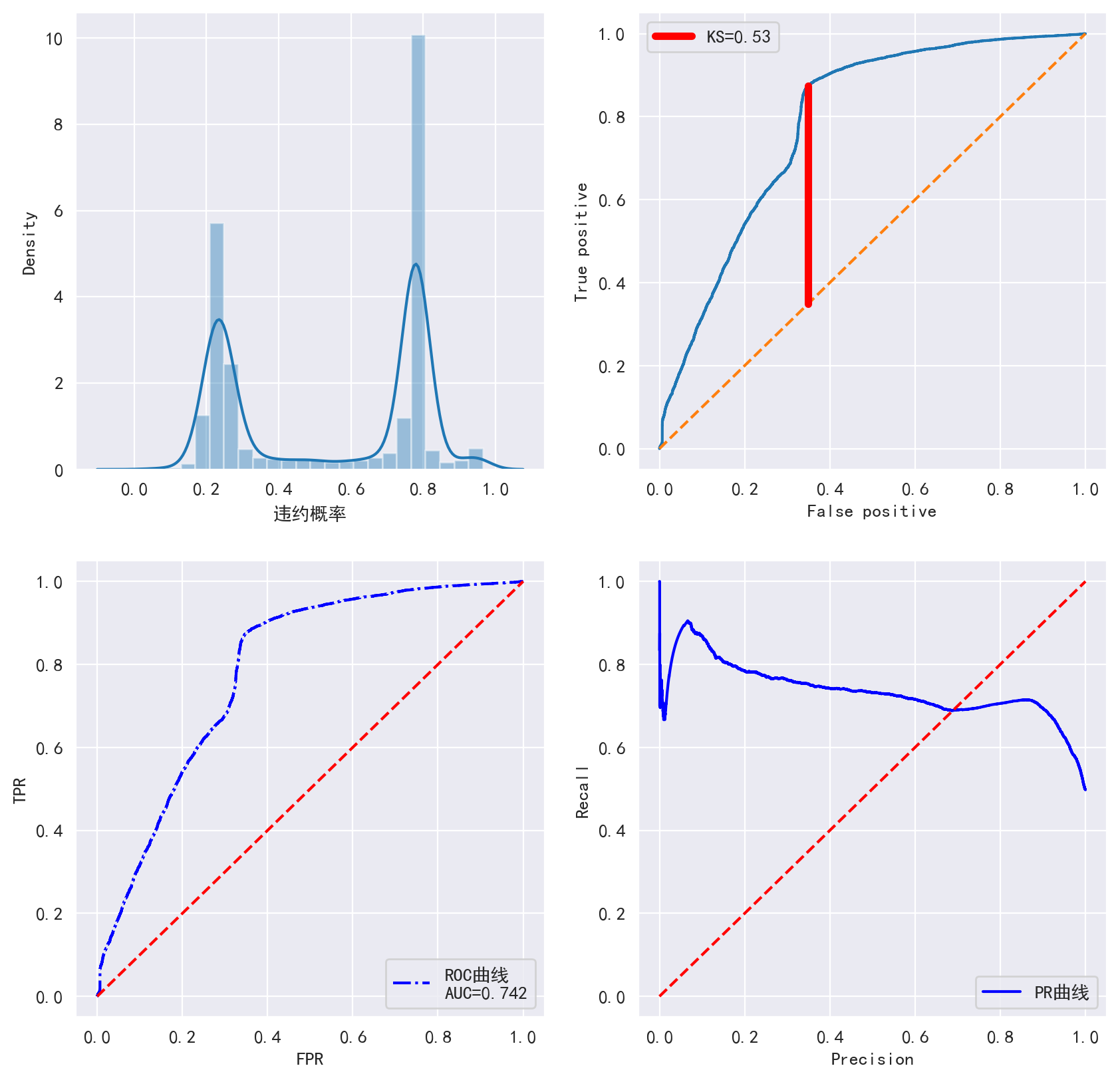


图3-15 SVM模型预测效果及指标结果

图3-15中四幅图分别是违约概率分布图，KS曲线图、ROC曲线图和PR曲线图，其中，KS值为0.53，AUC的值为0.742。

模型结果的混淆矩阵如表3-15：

表3-15 SVM单模型预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.67 | 0.78 | 0.72 | 8383 |
| 1 | 0.81 | 0.71 | 0.75 | 10983 |
| accuracy |  |  | 0.74 | 19366 |
| macro avg | 0.74 | 0.74 | 0.74 | 19366 |
| wight avg | 0.75 | 0.74 | 0.74 | 19366 |

## 3.4 差异性多模型融合仿真实验及对比分析

模型融合是将多个同质或异质的学习器集成在一起做信息补充或互补，取长补短形成一个更复杂、泛化能力更强的学习器。在本文的研究中，模型融合主要采用Stacking算法。为了防止过拟合，次级学习器采用的是简单学习器逻辑斯蒂回归LR。如前面误差-分歧定理所述，个体学习器之间的差异越大，模型融合的效果越好。因此为了增强学习器之间的差异性，从算法和参数两个方面入手，分析比对，建立最佳的预测模型。

3.4.1 算法多样性的模型融合

|  |  |
| --- | --- |
| 模型名称 | 单模型组合 |
| *M1* | GBDT、XGBoost、LightGBM、RF、BP、SVM |
| *M2* | GBDT、XGBoost、LightGBM、RF、BP |
| *M3* | GBDT、LightGBM、RF、BP、SVM |
| *M4* | GBDT、XGBoost、LightGBM、RF |

通过对不同算法模型的组合融合加强差异性，提升模型的效果。算法模型的组合定义如下：

表3-16 模型融合的算法组合

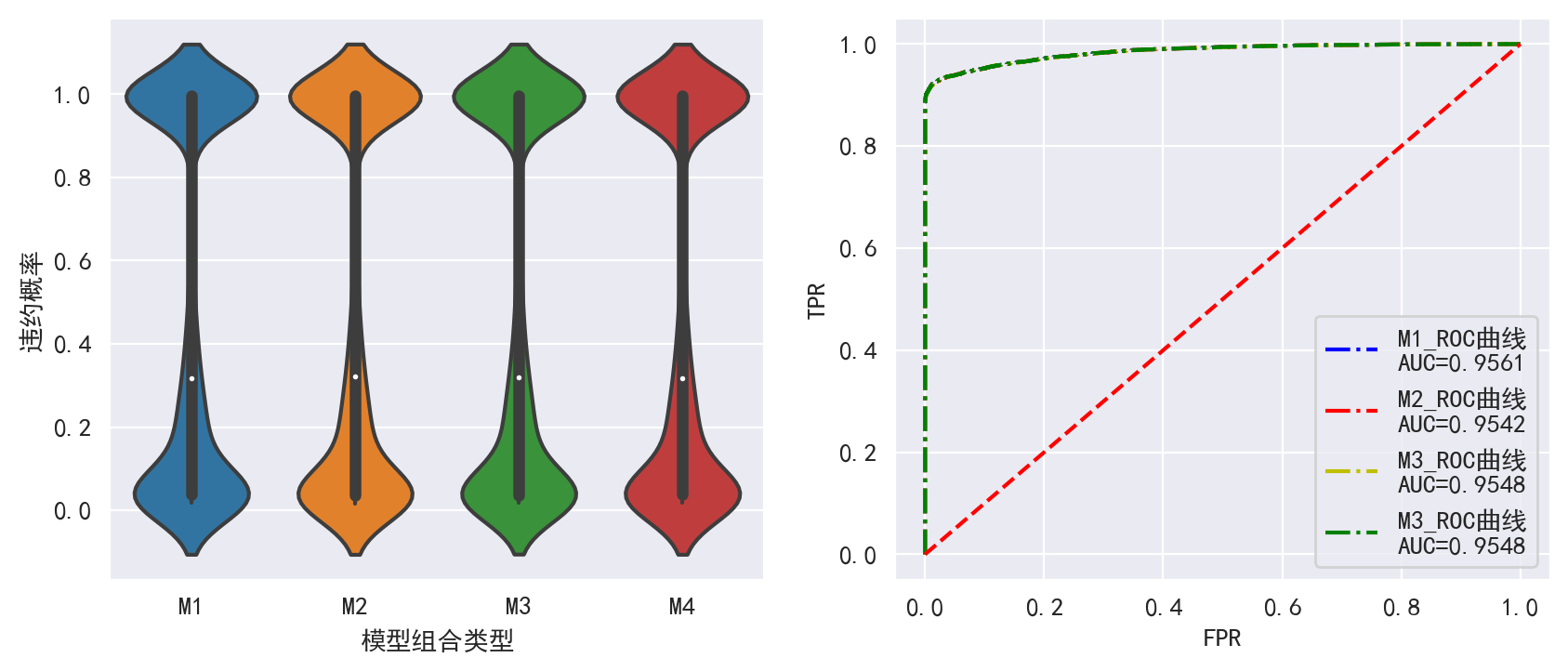
四种模型M1、M2、M3、M4不同模型组合融合的效果如图3-16：

图3-16 算法多样性的模型评价结果

如图3-16所示，四种模型融合的效果相差不大，其ROC曲线几近重合。相比于单模型，融合后的模型效果显著提升，AUC提升超过0.02。

M1、M2、M3、M4各模型的混淆矩阵如下：

表3-17 算法多样性的模型融合预测结果混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | precision | recall | F1-score |
|  | 0 | 0.98 | 0.94 | 0.96 |
| M1 | 1 | 0.93 | 0.97 | 0.95 |
| M2  M3  M4 | Accuracy  0  1  Accuracy  0  1  Accuarcy  0  1  Accuracy | 0.97  0.93  0.97  0.93  0.97  0.93 | 0.94  0.97    0.94  0.97  0.94  0.97 | 0.95  0.96  0.95  0.95  0.96  0.95  0.95  0.96  0.95  0.95 |

四个融合模型的精度、召回率等方面相较于单模型有了明显的提升。以精度为例，相比于单模型中效果最好的XGBoost，融合模型的精度提升超过3%，融合提升效果明显。模型融合效果最好的M1与单模型效果最好的XGBoost的关系以及模型融合效果最好的M1、M3和XGBoost的关系如图3-17：

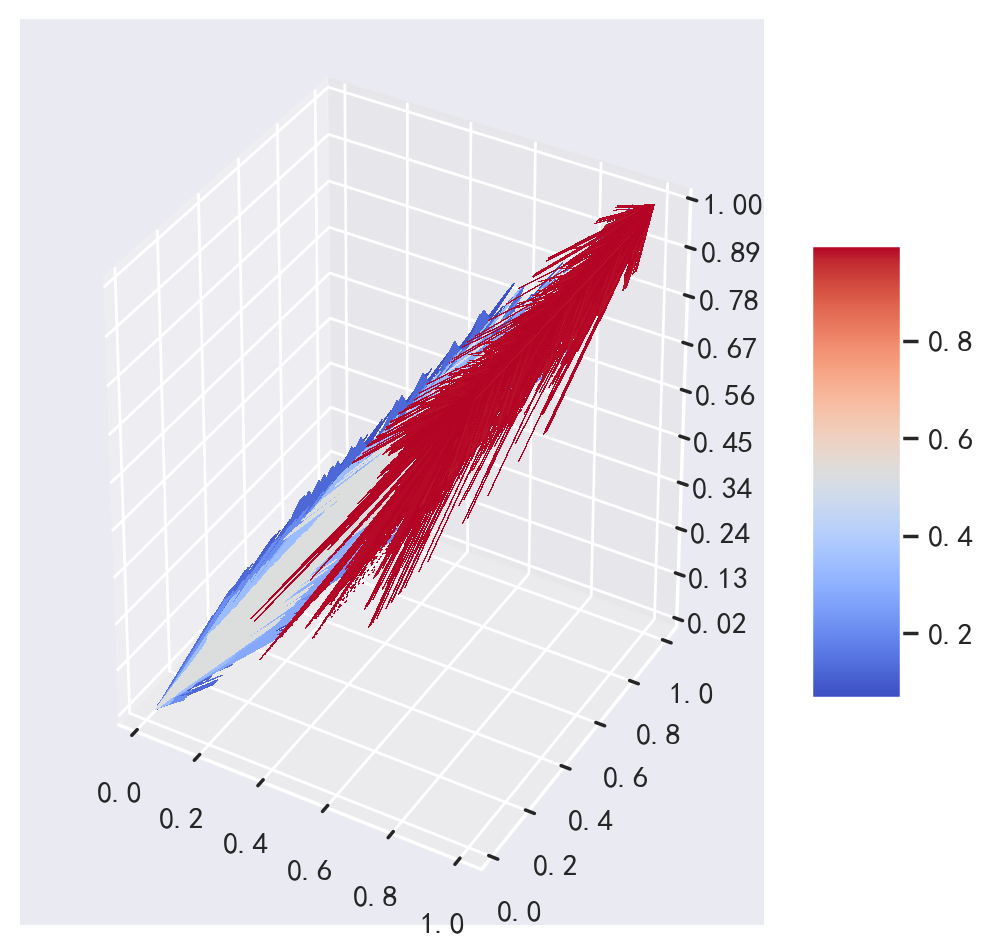
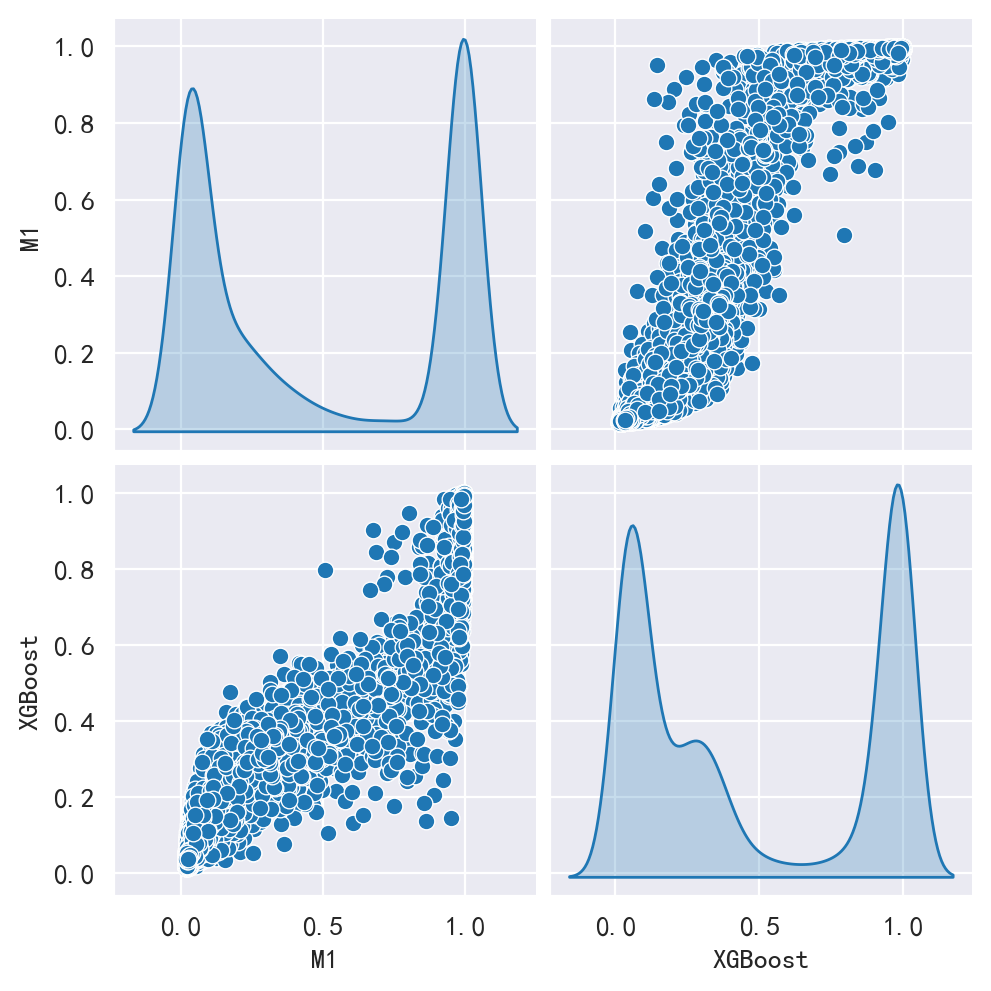


图3-17 M1与XGBoost关系图以及M1、M3与XGBoost关系图

3.4.2 参数多样性的模型融合

同种算法不同参数能训练出不同的模型，通过算法参数的扰动加强融合个体的差异性。通过算法参数的多样性提高模型融合效果的流程如图3-18：



图3-18 参数多样性的模型融合流程

从图3-18可以看出，对XGBoost、GBDT、LightGBM和RF的部分参数做小范围的扰动训练形成多个单模型，并使用Stacking再次进行模型融合。

对算法的参数做小范围的扰动，最终模型的效果图如图3-19，左图为预测概率的分布核函数图，右图展示了模型的AUC值为0.9525。

参数多样性模型效果最好的混淆矩阵如表3-18：

表3-18 参数多样性的模型融合混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.97 | 0.94 | 0.95 | 10577 |
| 1 | 0.93 | 0.96 | 0.95 | 8789 |
| accuracy |  |  | 0.95 | 19366 |
| macro avg | 0.95 | 0.95 | 0.95 | 19366 |
| wight avg | 0.95 | 0.95 | 0.95 | 19366 |

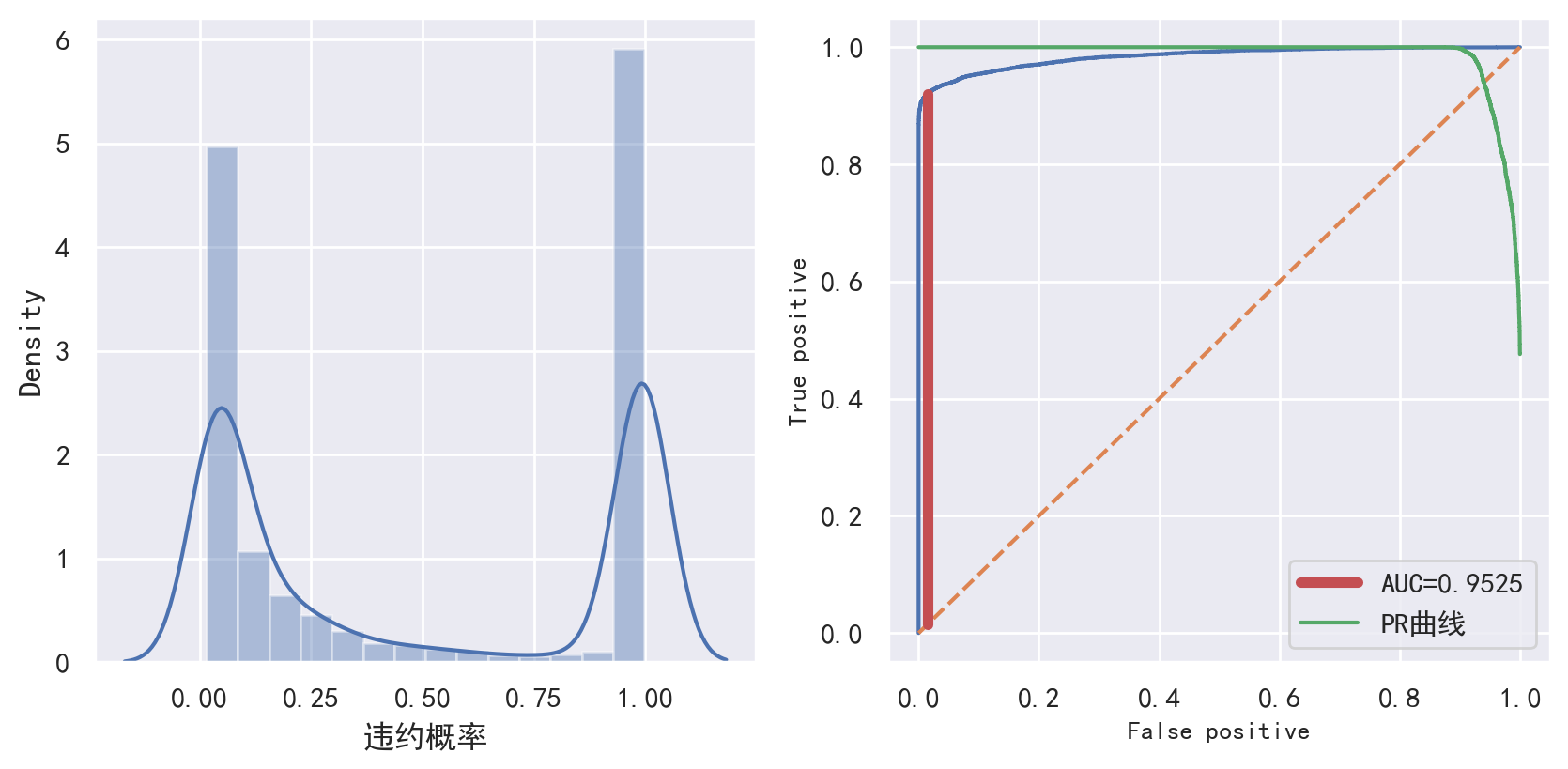


图3-19 参数多样性的模型评价结果

从表3-18模型结果的混淆矩阵来看，模型效果与算法多样性基本持平，相较于单模型仍有较大的提升效果。

各模型之间的相关关系如图3-20：

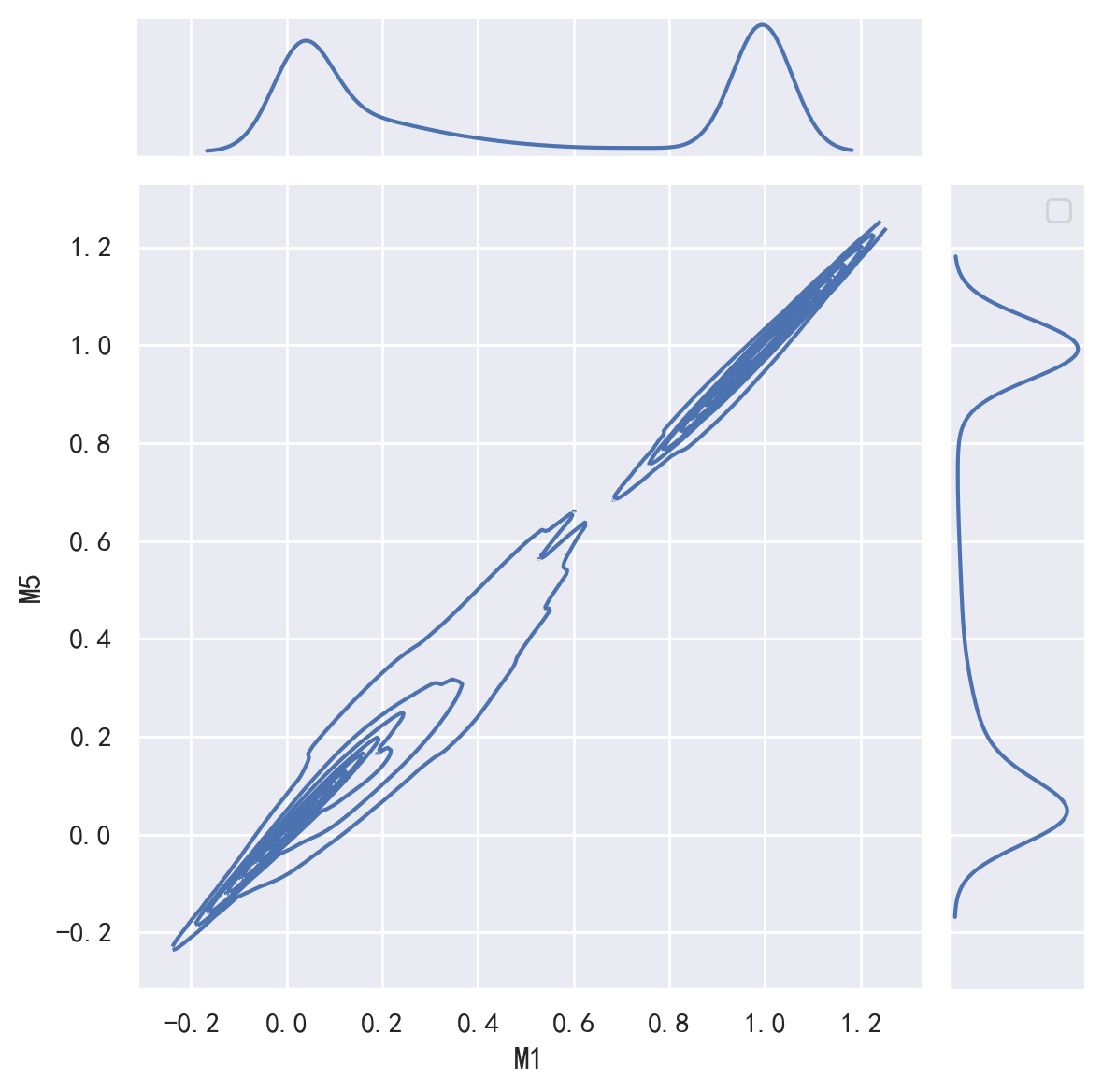
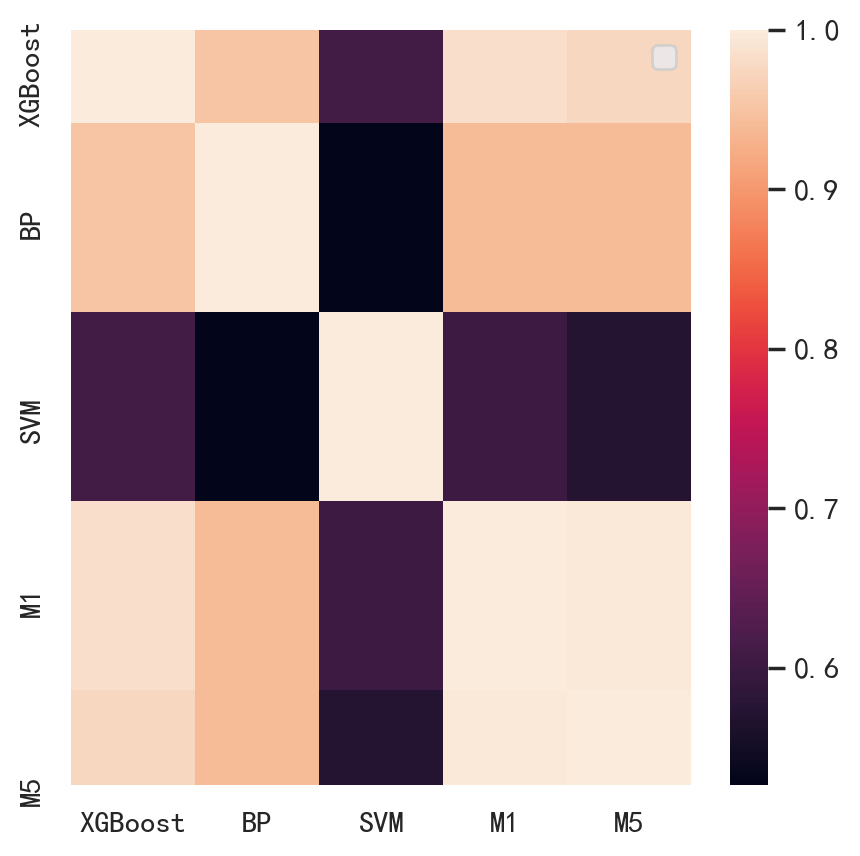


图3-20 各模型之间的相关关系

图3-20中第一张图为M1和M5，即算法多样性中效果最好的模型与参数扰动获得模型之间的相关关系比较图，可以看出，两者具有高度相关性。第二张图为XGBoost、BP、SVM、M1、M5五种模型的热力图，除了SVM与其他模型的相关度偏低，其他模型之间高度相关。

## 3.5 本章小结

本章主要内容为介绍基于多模型融合的信贷风险预测模型研究的实验部分。第一部分是对数据的探索及预处理，基于可视化对数据的部分特征进行了分析。第二部分为特征工程的构建，主要包括基础特征和显著特征。第三部分为基于机器学习的单模型仿真实验研究，并用可视化对其进行对比分析。第四部分是差异性多模型融合的仿真实验，主要包含算法多样性和参数多样性两方面。

# 第4章 基于遗传算法（GA）改进及多模型融合的信贷风险预测研究

基于XGBoost等单模型以及多模型融合在信贷风险预测研究中取得较好的预测效果，但在模型训练中若选取了不恰当的超参数可能会导致模型的欠拟合和过拟合等问题。因此，在本文研究中，引入遗传算法GA和网格搜索算法对GBDT、XGBoost、LightGBM、RF模型进行智能调参，有效选取全局最优超参数，并再次运用Stacking模型融合技术建立精准的信贷风险预测模型，提高整体模型的预测效果。

## 4.1 遗传算法和网格搜索

4.1.1 遗传算法GA思想

遗传算法起源于对生物系统所进行的计算机模拟研究。它是模拟自然界生物进化机制发展起来的随机全局搜索和优化方法，借鉴了达尔文的进化论和孟德尔的遗传学说[66]。其本质是一种高效、并行、全局搜索的方法，能在搜索过程中自动获取和积累有关搜索空间的知识，并自适应地控制搜索过程以求得最佳解[66]。

遗传算法相比于其他优化算法的主要特点是，其可以直接对所求解问题的结构对象操作，不需要求导或者连续函数的限制，具有很好的内在并行性和全局的寻优能力[67]。此外，遗传算法还采取概率化的参数寻优方法，不需要确定的规则就能自动获取优化的搜索空间，调整搜索方向[67]。

在利用遗传算法求解问题时，问题的每一个可能解都被编码成一个“染色体”，即个体，所有个体构成了群体（问题的所有可能解）。在遗传算法开始时，先随机产生一些初始解构成个体，根据预定的目标函数对每一个个体进行评估，基于给出的适应度选择下一代的个体，该操作体现了达尔文的“适者生存”的原理，大于适应度的个体被用于产生下一代，小于适应度的则淘汰[68]。再经过交叉、变异等操作组合生成新的一代，依次逐步朝着最优解的方向进化。

4.1.2 遗传算法的设计与实现

遗传算法求解最优值是一个由可行解组成的群体初步进化过程。在求解过程中，涉及到几个重要的步骤。分别是编码和遗传的操作。其中，遗传操作又包含选择、交叉、变异。选择和交叉基本上完成了遗传算法的大部分搜索功能，变异操作增加了遗传算法找到最优解的能力。

（1）编码操作

在利用遗传算法求解问题时，首先需要确定问题的目标函数和变量，之后对这些变量进行编码。在遗传算法求解过程中，遗传算子直接对串操作，且问题的解以数字串形式存在[68]。编码可以分为二进制编码和实数编码。若用二进制编码表示个体，则通过公式4-1将二进制数转化为十进制数的解码公式：

|  |  |  |
| --- | --- | --- |
|  |  | （4-1） |

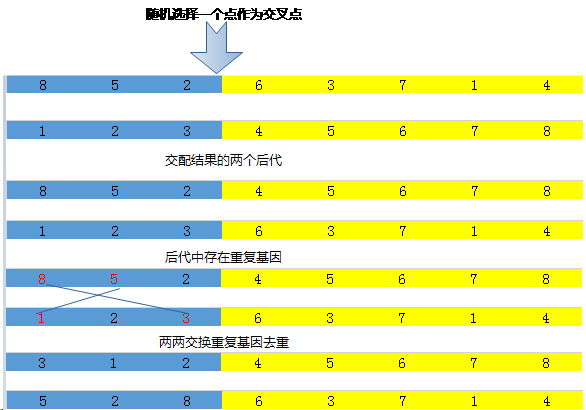
其中，为某个个体的第i段，每段段长为l，和是分量的定义域的两个端点。

二进制编码、解码操作简单易行，交叉和变异等遗传操作便于实现，且符合最小字符集编码的原则。除了二进制编码，常用的还包括浮点编码法、符号编码法、袋鼠染色体编码等。

（2）选择操作

选择操作是指从群体中选择符合条件的优良个体，并淘汰劣质个体的操作。符合条件在这边是建立在适应度评估的基础上。由适应度函数计算出适应度，适应度越大的个体，被选中的可能性越大，在生成下一代的个体就越多[68]。

目前常用的选择方法有轮赌盘方法、最佳个体保留法、期望值法、排序选择法、竞争法、线性标准化法等。以轮赌盘方法为例，创建赌盘时，将种群中所有个体的适应度求和，称为总和适应度，然后将每个个体的适应度除以总和适应度，将得到的商逐个累加得到累积总和为1，如图4-1所示：



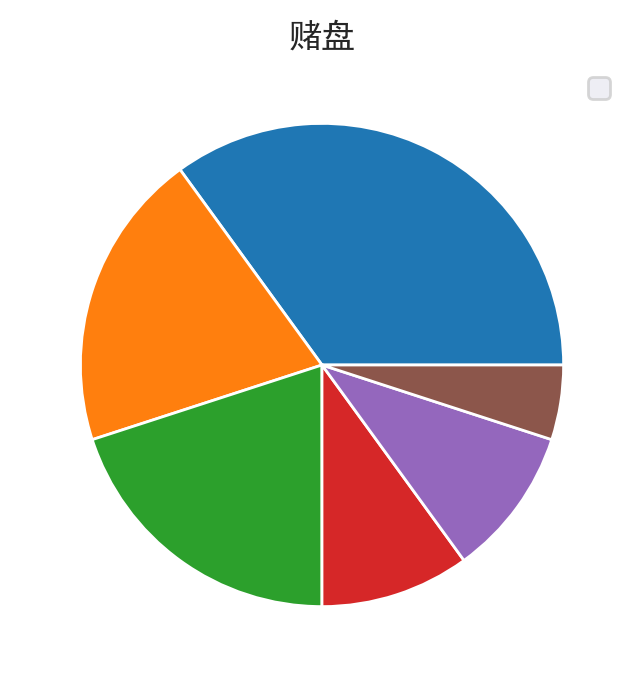




图4-2 单点交叉示意图 图4-1 轮赌盘方法

（3）交叉操作

交叉操作是指将两个父代个体的部分结构加以替换重组生成新的个体的操作。交叉可以在下一代产生新的个体，通过交叉操作，遗传算法的搜索能力有了极大地提升，是获得优良个体和最优解的重要手段[68]。常见的交叉算子有单点交叉、两点交叉、多点交叉、均匀交叉以及算数交叉等。单点交叉示意图如图4-2所示。

（4）变异操作

变异操作是指以很小的变异概率p随机改变种群中个体的某些基因的值。变异操作本身是一种局部随机搜索过程，与选择、交叉算子结合在一起，能够避免由于选择和交叉算子引起的信息丢失，保证遗传算法的有效性[67]。变异操作主要包括基本位变异、均匀变异、边界变异、非均匀变异、高斯近似变异等。

综上，遗传算法的流程如图4-3：



图4-3 遗传算法操作流程

从图4-3可以看出，遗传算法首先需要确定实际问题的参数集，并对参数集进行编码。编码完成后初始化问题的解，即初始化遗传算法中群体，采用不同的适应度计算方式评价群体，并完成选择、交叉、变异这三个基本算子的遗传操作，群体升级，并重新评价群体的适应度，观察是否满足停止准则，若满足，则输出可行解，若不满足则如此循环迭代，直到满足停止的条件。

4.1.3 遗传算法实例

遗传算法是用于解决最优化问题的一种搜索算法。以二元函数求解最大值为例，待求解的二元函数如下：

|  |  |  |
| --- | --- | --- |
|  |  | （4-1） |

利用遗传算法的思想，求得，在x约等于0，y约等于1.5时取得最大值约7。原函数以及寻找到的最大值点图像如图4-4与4-5所示：

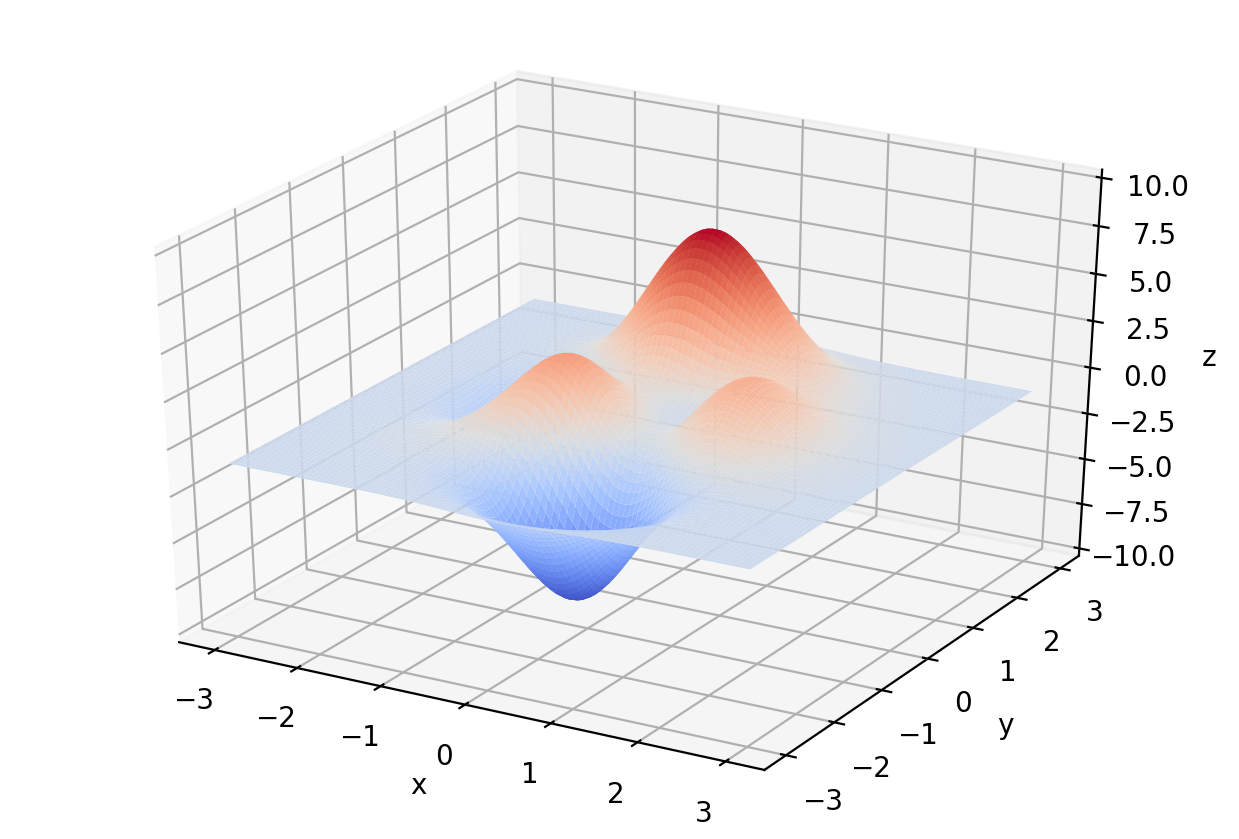
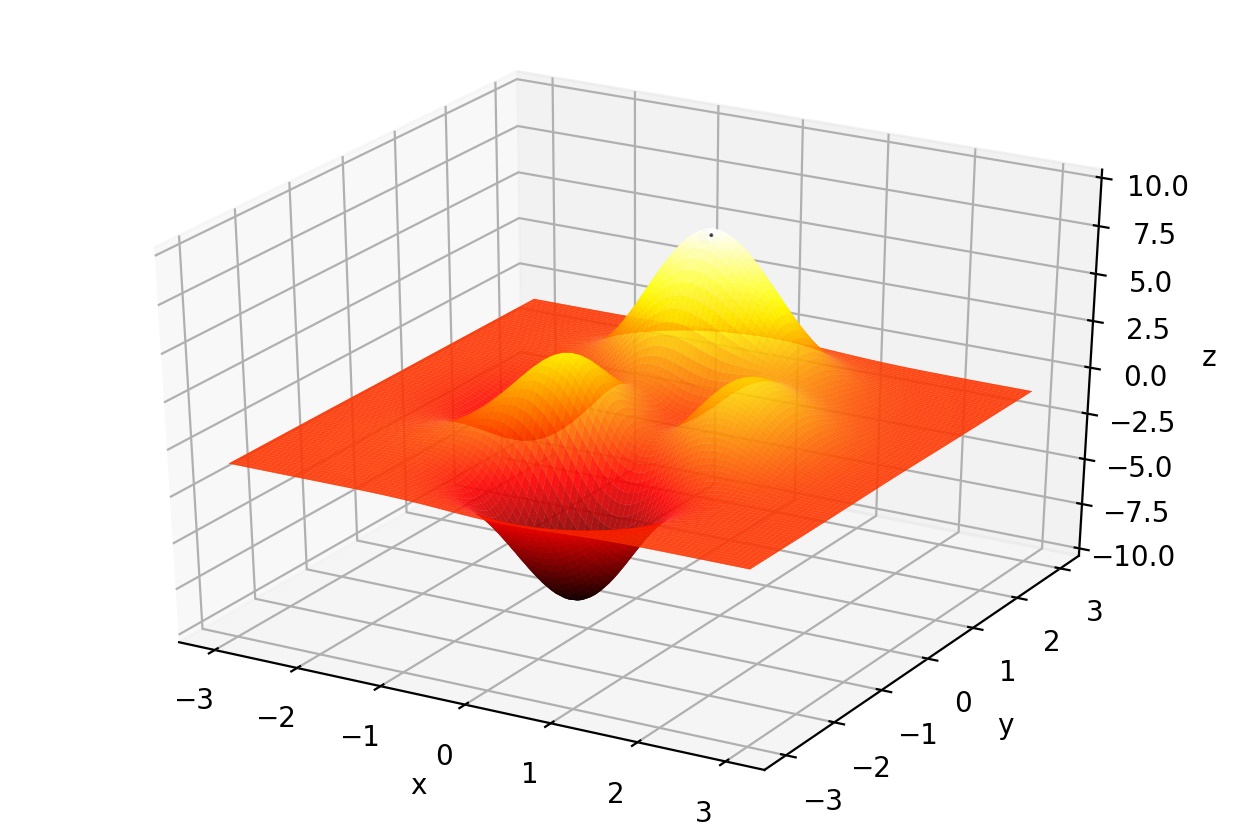


图4-4 原函数图像 4-5 包含最大值图像

4.1.4 网格搜索

网格搜索法是一种对参数值的穷举搜索方法。网格搜索方法通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。具体地，将算法每个参数的可能取值进行排列组合，列出所有可能的组合结果生成网格，将这些网格用于模型算法的训练，使用交叉验证对每次训练进行评估[69]。在算法对所有参数组合训练完成后，选择并返回一组泛化性能最好的参数作为最佳参数组合。为了确定搜索参数，每次训练后需要对训练效果进行评分。评估的机制是根据实际情况来确定，常规的包括Accuracy、F1-score、 F-beta、Precision、Recall等。

网格搜索是最简单、应用最广泛的超参数搜索寻优方法，通过查找搜索范围内所有的点来确定最优值。在使用网格搜索时，若采取较大的搜索空间和较小的步长，则网格搜索具有较大的概率可以找到全局最优值[69]。但网格搜索最大的问题是其十分消耗计算资源和时间，尤其对于需要调优的参数比较多的时候，计算资源和时间呈指数级增长。因此，在实际应用中，网格搜索一般先使用较广的搜索范围和较大的步长，之后逐步缩小搜索范围和步长，寻找更精确的最优值。

## 4.2 单模型算法改进仿真实验及对比分析

4.2.1 单模型超参数优化设计

同特征工程一样，模型算法的参数调节是非常重要的一项工作。模型的超参数优化问题本质上是黑盒函数优化问题。根据模型复杂程度的不同，需要调节的参数数量也不尽相同，如简单的逻辑回归，需要调节的参数只有正则项系数，复杂的如XGBoost需要调节的变量会增加许多。

在本文的研究中，主要利用遗传算法对XGBoost、GBDT、LightGBM以及Random Forest进行改进，为模型寻找最优的超参数组合配置，使模型具有更好的泛化能力。

通过第二章对GBDT等集成学习算法的介绍以及第三章单模型调参研究发现，learning\_rate，n\_estimators，max\_depth等参数对模型的准确率影响较大。根据经验，学习率的参数范围一般设置在[0.01, 0.8]之间，学习率过小，模型收敛速度较慢，学习率过大，收敛速度快但容易导致欠拟合。树模型的最大深度过小或过大都会极大影响模型最后的效果。因此，确定本文研究各单模型的超参数搜索空间如表4-1所示：

表4-1 遗传算法改进的单模型超参数搜索空间

|  |  |  |  |
| --- | --- | --- | --- |
| 模型名称 | 超参数 | 参数说明 | 参数范围 |
| XGBoost | learning\_rate | 学习率 | [0.01，0.8] |
|  | n\_estimators | 个体学习器个数 | [10，2000] |
| GBDT  LightGBM | max\_depth  min\_child\_weight  learning\_rate  n\_estimators  max\_depth  subsample  learning\_rate  n\_estimators  num\_leave | 最大深度  子节点最小权重和  学习率  个体学习器个数  最大深度  子样本比例  学习率  个体学习器个数  树的叶子数量 | [1，15]  [0，10]  [0.01，0.85]  [10,1500]  [1,15]  [0.01,0.9]  [0.01,0.8]  [10,1500]  [10，1200] |
| RF | n\_eatimators  max\_depth | 个体学习器个数  最大深度 | [8，2000]  [1， 20] |

4.2.2 遗传算法改进单模型

通过使用遗传算法对GBDT等单模型的参数搜索空间寻找全局最优的超参数组合，达到优化单模型的效果，进而改善模型融合最终的效果。

利用在训练集中得到的最优超参数融合到LightGBM等单模型建立信贷风险预测模型，单模型以LightGBM为例，GA-LightGBM预测模型的建立结构如图4-6：



图4-6 GA-LightGBM预测模型的建立结构图

由上图可知，由遗传算法得到各单模型全局最优超参数组合，再用此参数做模型训练。经过仿真实验，最终得到的四组最优参数组合如表4-2：

表4-2 遗传算法改进的单模型超参数组合

|  |  |  |  |
| --- | --- | --- | --- |
| 模型名称 | 超参数 | 参数说明 | 参数值 |
| XGBoost | learning\_rate | 学习率 | 0.19 |
|  | n\_estimators | 个体学习器个数 | 136 |
| GBDT  LightGBM | max\_depth  min\_child\_weight  learning\_rate  n\_estimators  max\_depth  subsample  learning\_rate  n\_estimators  num\_leave | 最大深度  子节点最小权重和  学习率  个体学习器个数  最大深度  子样本比例  学习率  个体学习器个数  树的叶子数量 | 9  5.83  0.25  330  14  0.8  0.23  129  96 |
| RF | n\_estimators  max\_depth | 个体学习器个数  最大深度 | 92  18 |

遗传算法改进前和改进后以及利用网格搜索参数寻优的集成学习算法预测对比结果如表4-3所示：

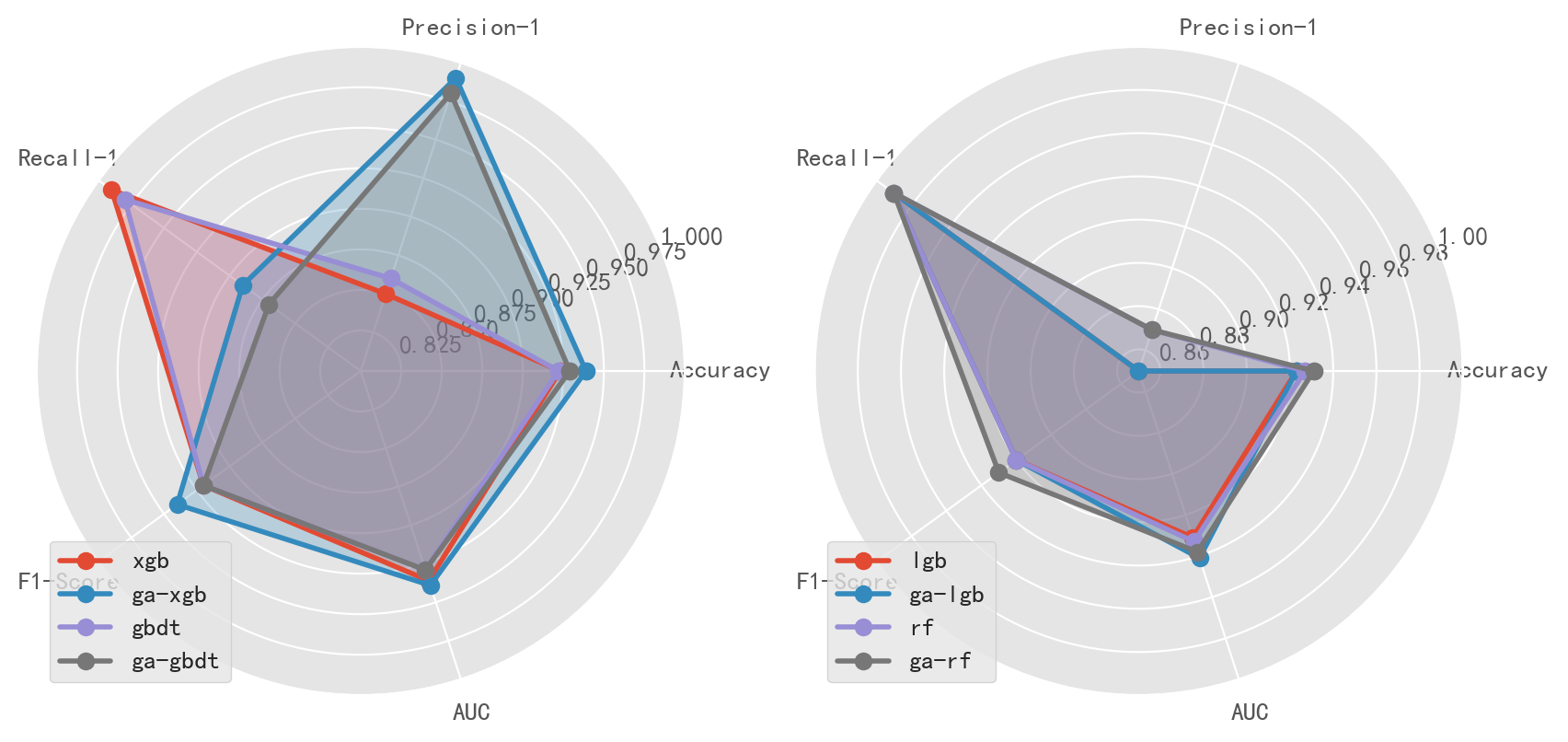
表4-3 遗传算法和网格搜索改进前后模型效果对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型名称 | Accuracy | Precision-0 | Recall-0 | Precision-1 | Recall-1 | F1-score-1 |
| XGBoost  GA-XGBoost  GS-XGBoost  GBDT  GA-GBDT  GS-GBDT  LightGBM  GA-LightGBM  GS-LightGBM  RF  GA-RF  GS-RF | 0.923  0.939  0.931  0.922  0.929  0.925  0.923  0.936  0.932  0.927  0.931  0.929 | 0.99  0.90  0.89  0.99  0.89  0.91  0.99  0.99  0.99  0.99  0.99  0.99 | 0.87  0.99  0.98  0.87  0.98  0.97  0.87  0.89  0.88  0.88  0.88  0.87 | 0.85  0.99  0.99  0.86  0.98  0.99  0.85  0.88  0.89  0.87  0.87  0.86 | 0.99  0.89  0.88  0.98  0.87  0.86  0.99  0.99  0.99  0.99  0.99  0.99 | 0.92  0.94  0.95  0.92  0.92  0.91  0.92  0.93  0.93  0.92  0.93  0.92 |

表4-3中GA-XGBoost为用遗传算法改进XGBoost后的模型，GS-XGBoost为用网格搜索最优参数组合后的模型，其他的以此类推。Precision-0是该模型下0（未违约）的准确率，Recall-0是该模型下0的召回率。F1-score-1为该模型下1的F1值。从上表可以看出，改进后的模型识别违约客户的准确率上有较为明显的提高，且总体精度也有一定程度的提升。

上述指标以及AUC值对比雷达图可视化如图4-7所示，从雷达图可以明显看出，利用遗传算法改进前后的差别，改进后识别信贷高风险的准确率有了显著提升。

四种集成学习算法的遗传算法改进前后ROC曲线对比如图4-8所示，由ROC曲线可以看出，在评价指标AUC维度遗传算法也有了一定程度的提升。



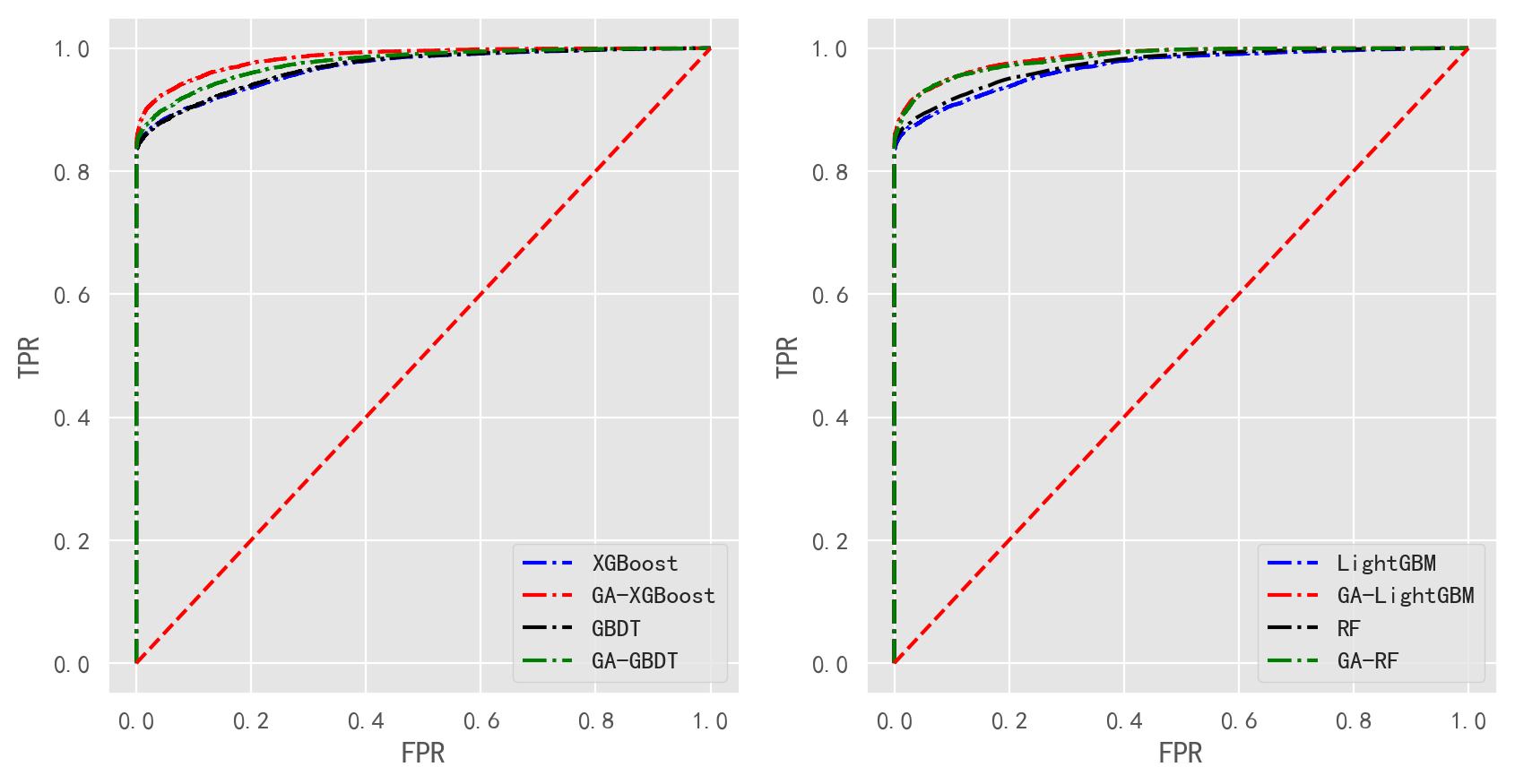
图4-7 遗传算法改进前后各指标结果对比雷达图

图4-8 四种集成学习算法的遗传算法改进前后的ROC曲线对比

从各个模型预测的概率分布情况来看，四个集成学习算法模型的预测逾期概率往0和1两极聚集，中间段较少，证明模型对于是否逾期具有较好的可分性。各个模型预测概率的分布小提琴图如图4-9所示。

再次对比分析模型之间的相关性关系，以GBDT和GA-GBDT为例，如图4-10所示，GBDT和GA-GBDT之间的相关性较强，但又各有差异，尤其在于0.2-0.4之间的分布。

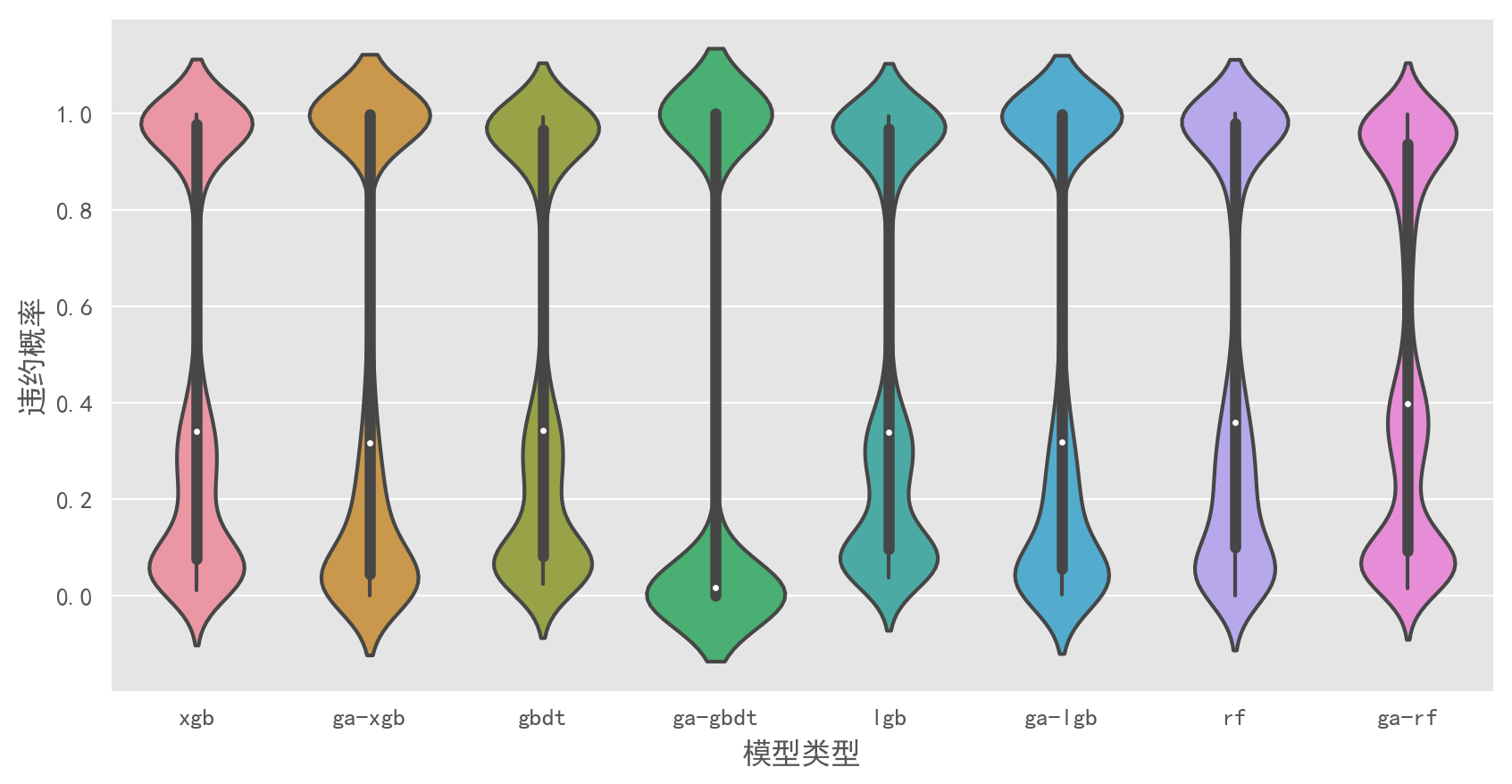


图4-9 遗传算法改进前后概率分布图

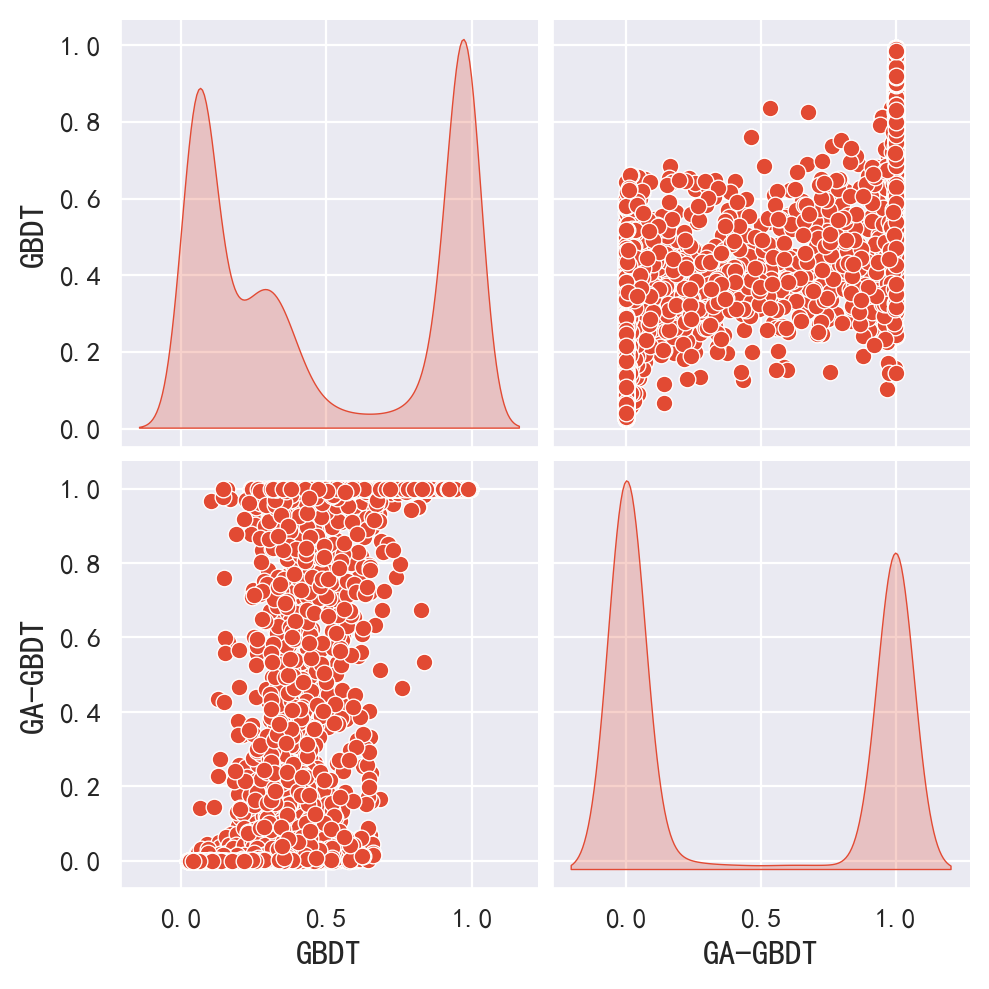


图4-10 GBDT与GA-GBDT的相关关系图

## 4.3 单模型算法改进后的模型融合

上一小节使用遗传算法对GBDT、XGBoost、LightGBM以及Random Forest做全局的参数组合寻优，建立了GA-GBDT，GA-XGBoost，GA-LightGBM和GA-RF模型，小幅度提升了单模型的泛化能力。为了进一步提升整体模型预测信贷风险的能力，再次使用Stacking算法对GA-GBDT、BP等算法等做进一步融合。从算法多样性和参数多样性两个维度出发，建立最佳的信贷风险预测模型。

4.3.1 算法多样性的模型融合

|  |  |
| --- | --- |
| 模型名称 | 单模型组合 |
| *M1* | GA-GBDT、GA-XGBoost、GA-LightGBM、GA-RF、BP、SVM |
| *M2* | GA-GBDT、GA-XGBoost、GA-LightGBM、GA-RF、BP |
| *M3* | GA-GBDT、GA-LightGBM、GA-RF、BP、SVM |
| *M4* | GA-GBDT、GA-XGBoost、GA-LightGBM、GA-RF |

通过对不同算法模型的组合融合加强差异性，提升模型的效果。算法模型的组合定义如下：

表4-4 模型融合的算法组合

模型的AUC值和ROC曲线以及概率分布效果如图4-11：

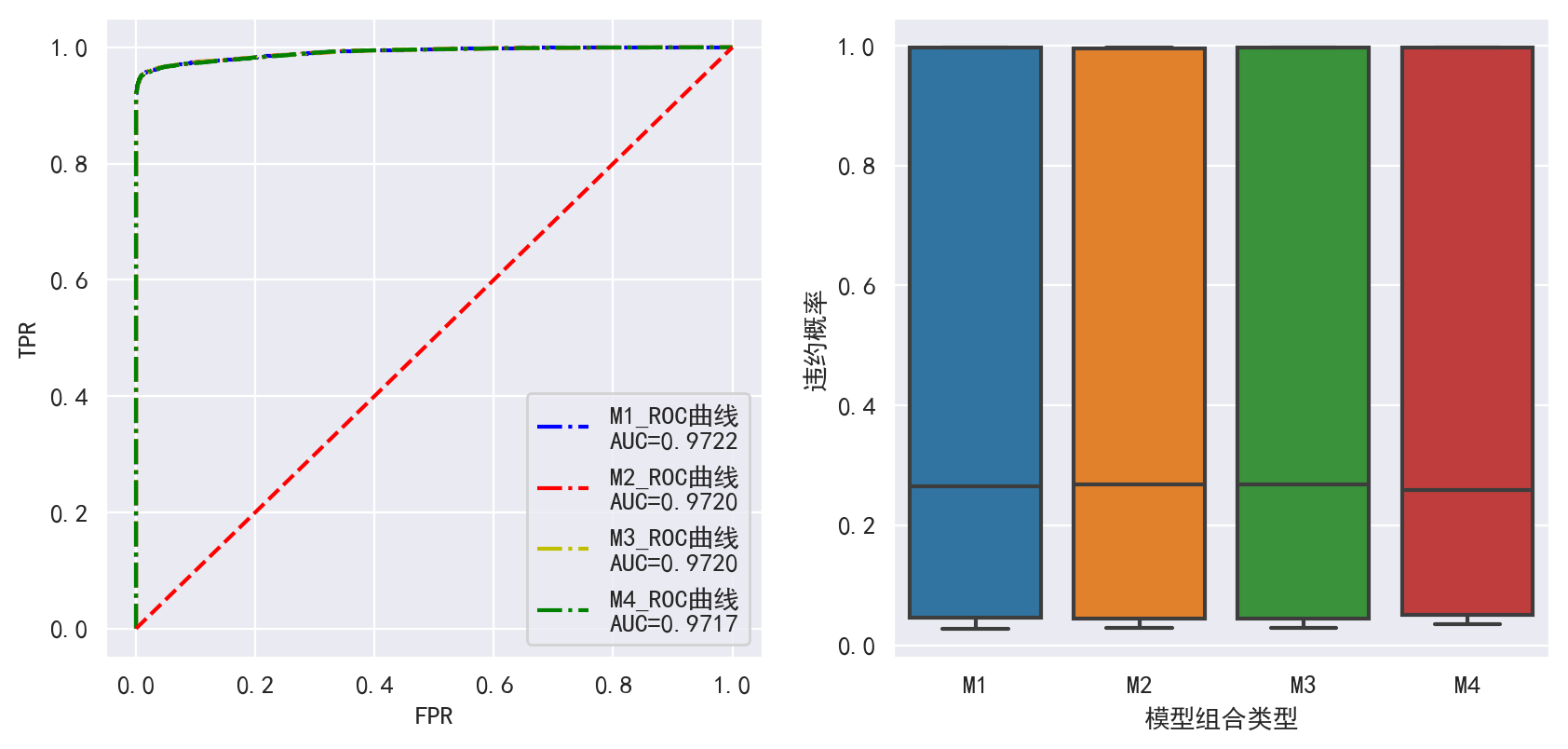


图4-11 M1、M2、M3、M4模型的ROC曲线和违约概率分布

从图4-11的ROC曲线可知，AUC的值均超过0.97，相比于未算法改进前，AUC的值提升了超过0.02，提升效果明显。同时，从4-11中的箱图看出，算法多样性生成的模型之间相似度较高。

M1、M2、M3、M4各模型的混淆矩阵如下：

表4-5 算法多样性各模型的混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | precision | recall | F1-score |
|  | 0 | 0.98 | 0.96 | 0.97 |
| M1 | 1 | 0.96 | 0.98 | 0.97 |
| M2  M3  M4 | Accuracy  0  1  Accuracy  0  1  Accuarcy  0  1  Accuracy | 0.98  0.96  0.98  0.96  0.98  0.96 | 0.96  0.98    0.96  0.98  0.96  0.98 | 0.97  0.97  0.97  0.97  0.97  0.97  0.97  0.97  0.97  0.97 |

经过算法多样性的模型融合建立的M1到M4在精度、准确率、召回率等方面均有明显的提升，平均提升相较于改进前提升2%以上，且总体模型的精度超过97%。

4.3.2 参数多样性的模型融合

同种算法不同参数能训练出不同的模型，通过算法参数的小范围扰动加强融合个体学习器之间的差异性。参数多样性生成模型的混淆矩阵如下：

表4-6 参数多样性生成模型的混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.98 | 0.96 | 0.97 | 10665 |
| 1 | 0.96 | 0.97 | 0.96 | 8701 |
| accuracy |  |  | 0.97 | 19366 |
| macro avg | 0.97 | 0.97 | 0.97 | 19366 |
| wight avg | 0.97 | 0.97 | 0.97 | 19366 |

从混淆矩阵来看，算法多样性和参数多样性生成的模型效果基本持平，算法多样性的泛化能力略好于参数多样性生成的模型。

## 4.4 信贷风险预测模型的最终建立

通过上述实验对比分析，最终确定信贷风险的预测模型，通过Stacking对GA-GBDT、GA-XGBoost、GA-LightGBM、GA-RF、BP、SVM进行模型融合建立最终模型。

最终建立的模型的预测概率分布如图4-12：

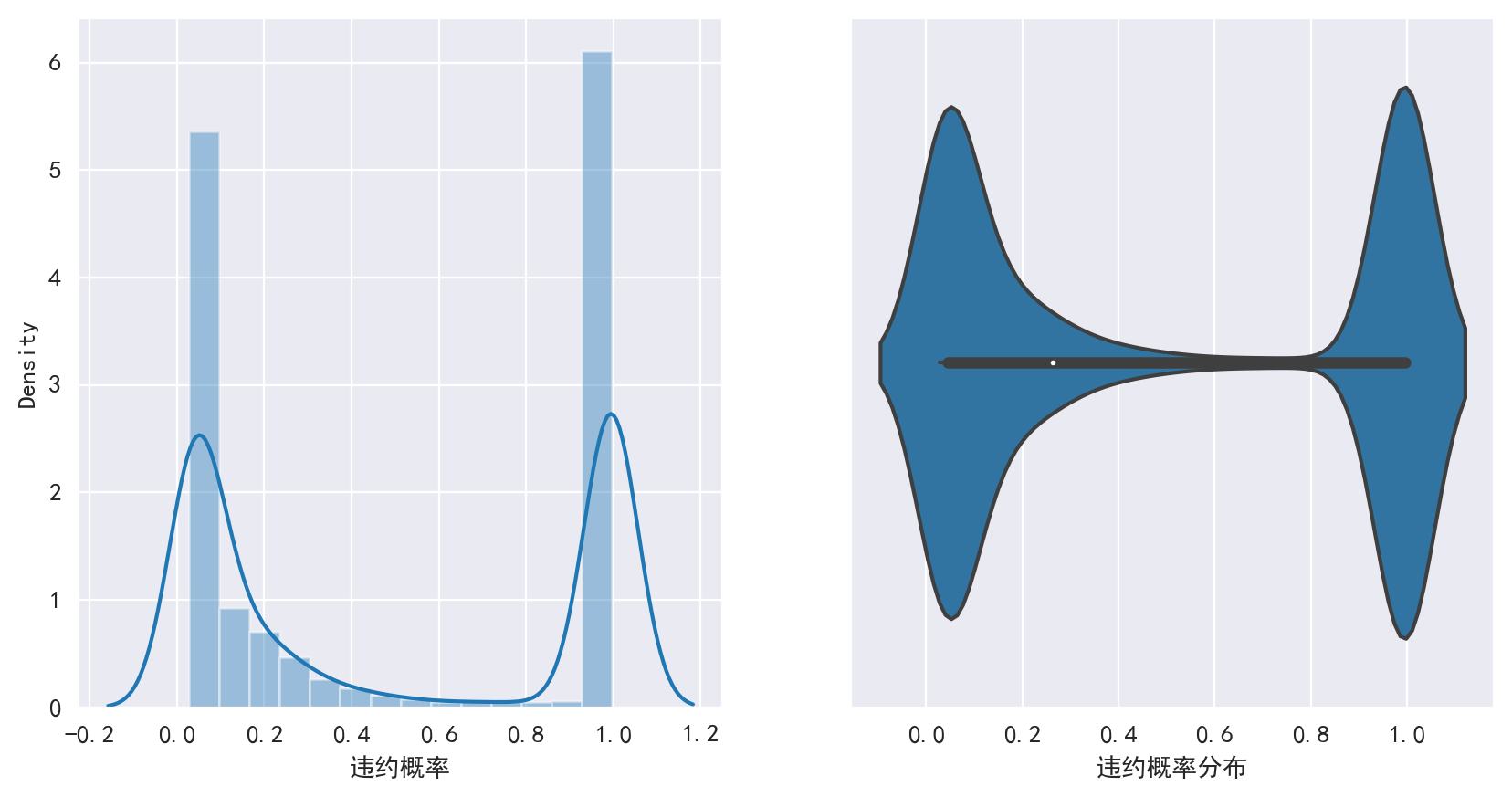


图4-12 最终模型预测结果的概率分布

从图4-12中的柱状图和小提琴图可以看出模型预测的概率集中在两端，中间较少，尤其违约概率在[0.4，0.8]之间。同时也可以说明模型对是否违约逾期有较为准确的判定，可以将正负样本较为“安全”的识别出来。

最终模型的混淆矩阵如下：

表4-7 最终模型预测结果的混淆矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.98 | 0.96 | 0.97 | 10700 |
| 1 | 0.96 | 0.98 | 0.97 | 8666 |
| accuracy |  |  | 0.9713 | 19366 |
| macro avg | 0.97 | 0.97 | 0.97 | 19366 |
| wight avg | 0.97 | 0.97 | 0.97 | 19366 |

从表4-7的混淆矩阵看，模型整体的精度、准确率、召回率以及F1-Score均达到96%以上，提升效果显著。

最终模型预测结果的ROC曲线和PR曲线如图4-13，最终模型预测结果的AUC达到0.972，预测效果较好，且相较于改进前，提升超过3%。

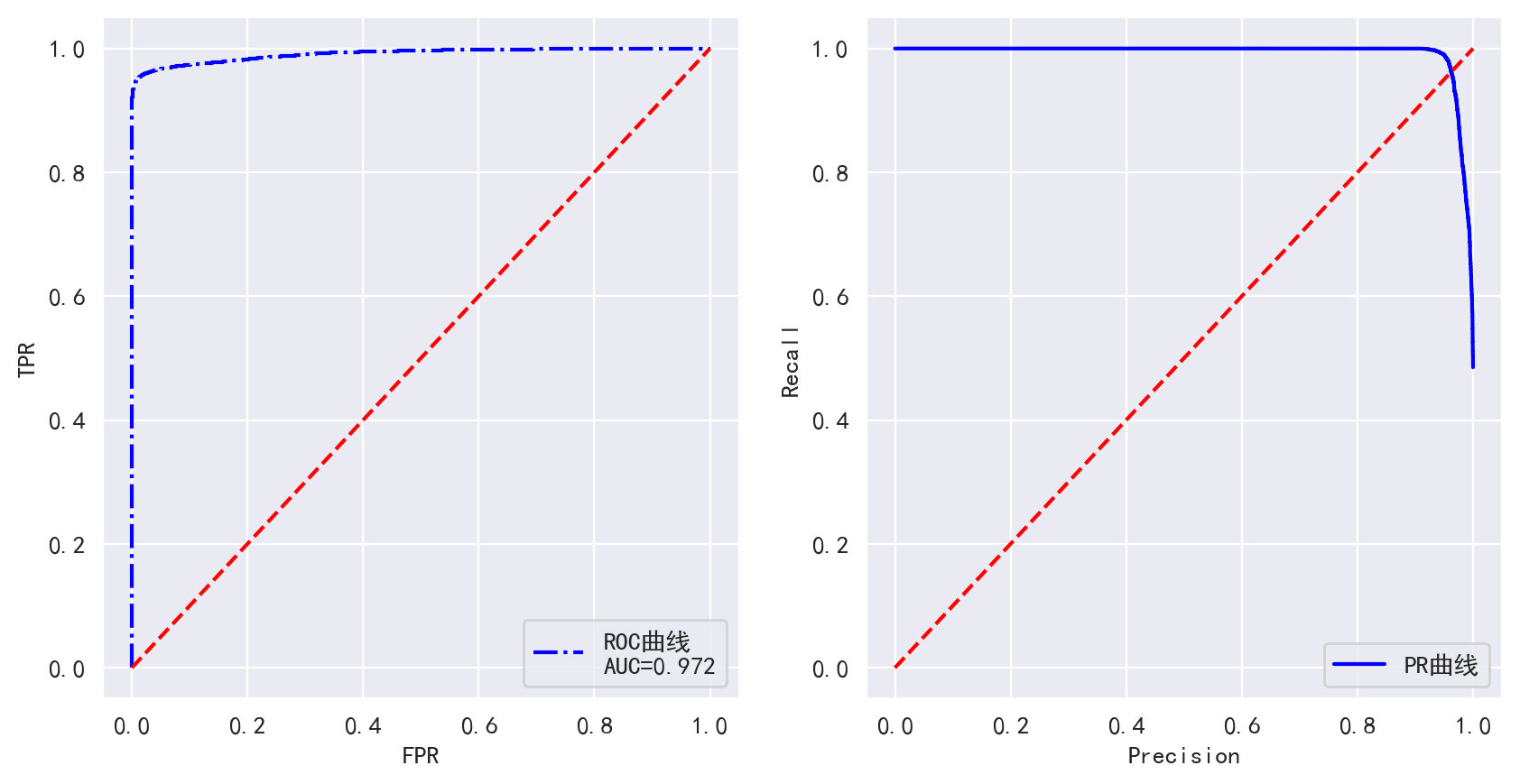


图4-13 最终模型预测结果的ROC曲线和P-R曲线

## 4.5 本章小结

本章重点介绍了使用遗传算法和网格搜索算法优化GBDT、XGBoost、LightGBM、Random Forest这四个单模型超参数的过程。遗传算法针对四个模型确定不同的搜索空间，优化不同的超参数组合，有效解决了该类模型容易陷入局部最优等问题。完成并提升了单模型的信贷风险预测的准确性和泛化能力，并对其相关性进行了比较。紧接着从算法多样性以及参数多样性增加模型之间的差异性，提升模型融合的效果。最终经过多个维度的比较，选定通过Stacking对GA-GBDT、GA-XGBoost、GA-LightGBM、GA-RF、BP、SVM进行模型融合建立最终的模型，并对模型预测概率的分布进行了可视化研究。

# 第5章 基于多模型融合的信贷风险预测模型的应用

本章是对基于多模型融合的信贷风险预测模型的应用。本文研究中的应用是基于Django框架建立系统模块，对基于多模型融合的信贷风险预测模型进行可视化的展示及应用。

## 5.1 系统开发环境

系统的硬件开发环境同第三章保持一致。信贷风险预测系统的软件开发环境如表5-1所示。

表5-1 系统软件环境

|  |  |
| --- | --- |
| 软件 | 名称及版本号 |
| 操作系统 | Win10 64位 |
| 开发环境 | Pycharm 2019 |
| Python | 3.6.1 |
| 服务器  数据库 | Django-1.9.8  Mysql-8.0.19 |
| 浏览器 | Google Chrome |

## 5.2 系统设计

Django是一款强大的、基于Python的开源Web应用框架，该框架采用了MTV的框架模式，MTV即模型M，视图V以及模板T，分层达到解耦的目的[70]。

在Django的项目目录中，主要包含以下几个文件。urls.py，views.py，models.py，forms.py，templates文件夹，admin.py以及settings.py。在Django框架中，控制器能够接受用户输入的部分并由框架自动处理。因此，在框架的整体结构中，Django更关注的是模型M、模板T以及视图V，其含义分别如下：

（1）M代表模型（Model）：即数据存取层，处理与数据相关的所有事物。

（2）T代表模板（Template）：即表现层，处理与表现相关的决定。

（3）V代表视图（View）：即业务逻辑层，用于存取模型及调取恰当模板的相关逻辑。

图5-1是Django框架的MVT访问流程示意图。



图5-1 MVT模式访问流程示意图

其中利用Django的ORM技术可以方便的通过model层对数据库进行操作，其本质是将关系数据库映射为程序语言中的对象，通过操作对象完成对数据库的操作。



图5-2 ORM与数据表对应关系图

信贷风险预测模型系统的可视化主要展示的是各个模型的预测结果查询对比。通过前台获取查询用户的ID，并选取相应的模型，后台模型预测出用户是否违约以及违约的概率，并将结果可视化。

## 5.3 预测查询模块及预测设计实现

对信贷风险模型预测结果的可视化主要包括用户是否会违约和会违约的概率。

（1）查询页面

图5-3是预测结果查询页面展示，通过输入用户的ID以及选择对应的算法模型来获取相关的预测结果。



图5-3 查询页面图

（2）预测结果展示

图5-4是信贷风险预测模型的预测结果，包括选择的模型，是否会违约以及违约的概率，图中所示的是用户ID为19366，选择的模型是GA-XGBoost，经后台计算，该用户不会违约，违约的概率为0.02163。



图5-4 预测结果展示

## 5.4 本章小结

本章主要介绍了基于多模型融合的信贷风险预测模型的应用实现过程。首先介绍了开发预测系统的硬件和软件环境，并对Django的框架原理做了重点阐述。紧接着结合Django框架和数据库相关技术完成了信贷风险预测模型查询模块的可视化工作，展示了预测模块具体实现的界面，同时给出了模型预测的结果，包含是否违约以及违约的概率等相关信息。

# 总结与展望

本文总结

本文探讨建立了一种基于遗传算法与多模型融合技术的信贷风险预测模型。信贷风险的管理不仅是金融服务行业的重要组成部分，尤其对于互联网金融服务来说，也是金融机构的核心竞争力之一。当前个人信用风险评估主要依赖于央行的征信系统，但央行的征信系统存在结构性缺陷，会大大限制评估客户的范围。基于此问题，本文通过分析国内外的研究动态，从用户的基本信息、银行卡消费记录、信用卡账单以及用户的浏览行为出发，挖掘信用风险相关特征，并利用集成学习方法和多模型融合技术建立信贷风险预测模型。同时，为了进一步提高模型的泛化能力，使用遗传算法寻找最优超参数组合，进一步提高了模型预测的准确性。本文的研究工作可以概括为以下几个方面：

1、针对目前信贷风险预测模型存在的问题，围绕国内外研究动态，定量分析了各自存在的缺陷。结合本文研究中使用的用户相关行为数据集，确定了使用集成学习方法和多模型融合的技术来进行信贷风险预测模型的方案。

2、本文的研究主要使用GBDT、XGBoost、LightGBM、RF四种集成学习算法和BP神经网络以及SVM预测信贷风险。围绕当前预测模式存在的问题，根据用户行为挖掘信贷风险相关影响因素，量化具象为相关特征，并使用上述学习算法建立模型。同时，为了提高模型的准确性，引入多模型融合技术，取长补短，确定各模型之间最佳权重系数。通过单模型以及多模型融合之后的模型之间对比，结果表明多模型融合的预测结果明显优于其他几种模型。

3、为了进一步提升模型的泛化能力，本文引入模拟达尔文进化论的遗传算法对四种集成学习方法进行改进，寻找最优的模型超参数组合，减小陷入局部最优的缺陷。结果表明使用遗传算法改进后的集成学习相较于改进前有了一定程度的提升。同时，从算法多样性和参数多样性两方面再次使用多模型融合技术构建加权融合模型，取得了较好的预测效果，证明了该模型的有效性。

4、设计并实现了信贷风险预测模型的应用。通过Django框架和数据的可视化展示了系统的查询模块，使模型结果更加直观地展示。

进一步的研究方向

1、本论文研究中使用的训练集数据量较小，且存在大量的数据缺失和类别不均衡的影响，可能存在泛用性问题。若需要对模型推广使用，需要对模型做进一步的研究。

2、本文引入的遗传算法只能以一定的概率找到全局最优的超参数组合，使用的是概率机制进行迭代，具有随机性，后续研究可结合一些启发式算法进行改进。同时，遗传算法较难解决大规模计算量问题，容易陷入“早熟”。

3、本文信贷风险预测模型的应用仅包含预测模型查询结果的模块，功能较为单一，后续可根据需求方面对系统功能做进一步的扩充。

# 致 谢

论文至此接近尾声，也意味着我在交大三年的研究生学习生涯即将结束。三年时光白驹过隙，依稀记得第一次走进西南交大的校园，被路旁美丽的银杏大道所吸引，被巍峨高大的图书馆所震撼。三年时间，我享受着实验室学习的快乐时光，不仅收获了专业知识，更收获了友谊和感动。在此，我对所有曾经帮助过我的人表示由衷的感谢。

首先非常感谢我的导师楼新远副教授。读研期间，楼老师在学习、生活上给予了我悉心的指导与关怀，从学术研讨会、到参与科技竞赛，再到论文的发表都给予了我极大地启发与帮助。楼老师治学严谨、学识渊博、平易近人，对我三年的学习生涯有着深远的影响。本篇论文的撰写，从选题、论文的研究内容、技术路线、关键实验到最后论文的完成，都给出了详细深入的指导意见，再次感谢楼老师。

我还要感谢我的女朋友赫倚和我的父母在生活等各方面给予了大力支持，我才能专心地完成学业，让我对未来充满了期待，让我更有动力、更从容的走向社会。同时我还要感谢实验室的同门、师弟师妹，在科研方面，我们共同探讨，一起分享科研知识，使我的研究生生涯更加丰富，感谢你们陪我度过这三年的校园时光。

再次由衷感谢所有关心和帮助我的人，祝愿大家前程似锦，生活美满！

# 参考文献

1. 肖义. 互联网金融业务增长模式及其对传统金融的借鉴[D]. 2015.
2. 常敏. 互联网＋形势下大力推进智能制造[J]. 中国石油和化工经济分析, 2015, 000(008):P.10-11.
3. 张鑫, 曹帅. 基于BP神经网络的P2P网贷信用风险甄别研究——以拍拍贷为实例[J]. 现代商业, 2020, 000(011):105-106.
4. 无[1]. 我国征信业的改革与发展[J]. 中国银行业, 2019, 000(008):P.19-21.
5. 张莹. 商业银行互联网信贷业务的风险管理研究[J]. 财经界(学术版), 2019(20).
6. 张万军. 基于大数据的个人信用风险评估模型研究[D]. 2016.
7. 沙希杜·伊斯兰姆, 和红梅, 基肖尔·库梅尔·班萨克,等. 孟加拉国视角下的"一带一路"及孟中印缅经济走廊建设[J]. 南亚东南亚研究, 2018(3).
8. 闫真宇. 关于当前互联网金融风险的若干思考[J]. 浙江金融, 2013(12):40-42.2.
9. 王裕粟. 基于数据挖掘的客户信用评级模型的设计与实现[D]. 2012.
10. 洪娟, 曹彬, 李鑫. 互联网金融风险的特殊性及其监管策略研究[J]. 中央财经大学学报, 2014.
11. 周贤. 新形势下银行信贷风险管理问题研究[J]. 中国市场, 2020, No.1033(06):20+71.
12. 刘佳蒙. 商业银行信贷风险管理存在的问题与对策研究[J]. 农村经济与科技, 2020, v.31;No.480(04):166-167.
13. 徐溪蔓. 基于时间序列模型的商行信贷规模与风险管理分析[J]. 现代商业, 2019(24)..
14. 谢昌文, 李黎. 基于层次分析法的商业银行小企业信贷风险预警模型初探[J]. 金融经济:下半月, 2015.
15. Carter M M . ENGINE, SYSTEM AND METHOD OF PROVIDING CLOUD-BASED BUSINESS VALUATION AND ASSOCIATED RULES-BASED SERVICES[J]. 2012.
16. Sandoval, Leonidas. Structure of a Global Network of Financial Companies Based on Transfer Entropy[J]. Entropy, 2014, 16(8):4443-4482.
17. 姜迪. 互联网金融风险与风险防范[J]. 经贸实践, 2019, 243(01):138.
18. 李长征. 完善我国互联网金融风险管理机制的建议[J]. 商场现代化, 2019, 000(004):92-94.
19. 章豪. 借款人未按协议履行分期还款义务应视为预期违约[J]. 商, 2014, 000(018):136-137.
20. 胡佳. 我国商业银行信贷风险评估模型比较研究[D]. 2007..
21. 陈晓兰, 任萍. 基于Logistic混合模型的企业信用风险评价研究[J]. 山东财政学院学报, 2011(02):90-93.
22. 光丽.基于Logistic回归模型的P2P网络借贷平台借款人信用风险评估研究——以人人贷为例[D].甘肃:兰州大学,2017. DOI:10.7666/d.D01298439.
23. 呼振凯. P2P网络借贷中借款人的信用风险评估研究[D]..
24. 徐晓霞,李金林.基于决策树法的我国商业银行信用风险评估模型研究[J].北京理工大学学报（社会科学版）,2006,8(3):71-74. DOI:10.3969/j.issn.1009-3370.2006.03.018.
25. 张红梅. 基于决策树算法的房贷信用风险评估研究[D]. 哈尔滨理工大学, 2012.
26. 马文勤. 基于BP神经网络的农户小额信贷信用风险评估研究[D]. 西北农林科技大学, 2010.
27. Chen H , Joslin S . Generalized Transform Analysis of Affine Processes and Applications in Finance[J]. Review of Financial Studies, 2012, 25.
28. 方先明, 熊鹏, 张谊浩. 基于Hopfield神经网络的信用风险评价模型及其应用[J]. 中央财经大学学报, 2007, 000(008):34-40.
29. 张佳维. 基于模糊神经网络的个人信用风险评估[D]. 2014..
30. EDUARDOFRIEDMAN. AN EMBEDDING THEOREM FOR QUATERNION ALGEBRAS[J]. Journal of the London Mathematical Society, 1999, 60(1):33-44.
31. Venter A , Laurie D P . A Doubly Adaptive Integration Algorithm Using Stratified Rules[J]. BIT, 2002, 42(1):183-193.
32. 柯国霖. 梯度提升决策树(GBDT)并行学习算法研究[D]. 2016.
33. 汪伟. 特征选择在几种决策树集成方法中的表现研究[D]. 2017.
34. 曹礼园, 李深洛. 基于多目标最优化的最小代价决策树构建与实现[J]. 计算机与数字工程, 2019, 47(12):3020-3024.
35. 孙慧然. 一种基于决策树的机器学习方法和推理机设计[D]. 长春工业大学, 2014.
36. 李超, 张文辉, 李然,等. 基于Stacking集成学习的恒星/星系分类研究[J]. 天文学报, 2020, 061(002):102-111.
37. 杨立洪, 白肇强. 基于二次组合的特征工程与XGBoost模型的用户行为预测[J]. 科学技术与工程, 2018, 018(014):186-189.
38. 谢勇, 项薇, 季孟忠,等. 基于Xgboost和LightGBM算法预测住房月租金的应用分析[J]. 计算机应用与软件, 2019(9).
39. 熊苏生. 基于改进LightGBM的交通模式识别算法[J]. 计算机与现代化, 2018, 278(10):72-77+130..
40. 郭步豪. 基于梯度提升机器学习算法的ECG身份识别[D]. 2020.
41. 王喆, 李冬冬, 范奇,等. 一种基于单边动态下采样的不平衡数据分类系统:, 2016.
42. 祁海洋, 史源源, 王培勇,等. 一种识别虚开发票企业的方法,装置及设备:, 2019.
43. 简艺恒, 余啸. 基于数据过采样和集成学习的软件缺陷数目预测方法[J]. 计算机应用, 2018.
44. Kandhoul N , Dhurandher S K , Woungang I . Random forest classifier‐based safe and reliable routing for opportunistic IoT networks[J]. International Journal of Communication Systems, 2021, 34(1).
45. 何世建. 基于梯度提升决策树与深度信念网络融合的推荐算法研究[D]. 2017.
46. 王玉晶, 莫建麟. 基于TS特征选择的生理情感状态分类[J]. 齐齐哈尔大学学报:自然科学版, 2013, 29(3):19-19.
47. 凌强, 费习宏, 李峰. 一种基于CatBoost模型的车辆尾气浓度超标判别方法:, 2020..
48. Wang X , Lou X Y , Hu S Y , et al. Evaluation of Safe Driving Behavior of Transport Vehicles Based on K-SVM-XGBoost[C]// 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). IEEE, 2020.
49. Hussain M , Wajid S K , Elzaart A , et al. A Comparison of SVM Kernel Functions for Breast Cancer Detection[C]// Eighth International Conference Computer Graphics. IEEE Computer Society, 2011.
50. 松井,正一. NFL定理(No Free Lunch Theorem)[J]. Journal of Japan Society for Fuzzy Theory and Systems, 2002, 14.
51. 洪铭, 汪鸿翔, 刘晓芳,等. 采用负相关学习的SVM集成算法[J]. 华侨大学学报(自然科学版), 2018, v.39;No.164(06):162-166.
52. Jeong H K , Lee Y P , Lahaye R J W E , et al. Evidence of Graphitic AB Stacking Order of Graphite Oxides[J]. Journal of the American Chemical Society, 2008, 130(4):1362-1366.
53. Baurle R A , Tam C J , Edwards J R , et al. Hybrid simulation approach for cavity flows: Blending, algorithm, and boundary treatment issues[J]. Aiaa Journal, 2003, 41(8):1463-1480.
54. 王永吉. 支持向量机泛化性能的研究及其应用[D]. 江南大学, 2009.
55. 汪云云, 陈松灿. 基于AUC的分类器评价和设计综述[J]. 模式识别与人工智能, 2011.
56. Grant T , Kluge A G . Data exploration in phylogenetic inference: scientific, heuristic, or neither[J]. Cladistics-the International Journal of the Willi Hennig Society, 2003, 19(5):379-418.
57. Saita C A . Declarative Data Cleaning: Language, Model, and Algorithms[J]. 2001.
58. Lekadir K , Merrifield R , Yang G Z . Outlier Detection and Handling for Robust 3-D Active Shape Models Search[J]. IEEE Transactions on Medical Imaging, 2007, 26:212-222.
59. Birant D , Kut A . ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60(1):p.208-221.
60. Yu H , Lo H , Hsieh H , et al. Feature engineering and classifier ensemble for KDD Cup 2010[C]// Jmlr Workshop & Conference. 2010.
61. Guyon I M , Andr, Elisseeff. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003.
62. Guyon I , Elisseeff, André. An introduction to variable and feature selection.[C]// Joint International Conference on Artificial Neural Networks & Neural Information Processing. Springer-Verlag, 2003.
63. 陈岩, 来海锋, 王清,等. 基于filter-wrapper的两步特征变量提取方法[J]. 机电工程, 2010, 27(4):67-71.
64. 廖伟智, 严伟军, 阴艳超,等. 基于Filter和Wrapper选择算法的特征选择方法:, 2018.
65. Wang J , Dang Y , Wang Q , et al. Embedded method for embedded interaction code array[J]. 2006.
66. Goldberg D E . Genetic Algorithm in Search Optimization and Machine Learning[J]. Addison Wesley, 1989, xiii(7):2104–2116.
67. Yang J , Honavar V . Feature subset selection using a genetic algorithm[J]. IEEE Intelligent Systems & Their Applications, 2002, 13(2):44-49.
68. 刘成锴. 基于遗传算法的文本特征降维方法研究[D]. 2020.
69. 王兴玲, 李占斌. 基于网格搜索的支持向量机核函数参数的确定[J]. 中国海洋大学学报(自然科学版), 2005(05):169-172.
70. 张翠丽, 孟小艳, 杨抒. 基于Django框架的管理系统的设计与开发[J]. 计算机技术与发展, 2019.

# 攻读硕士学位期间发表的论文及科研成果

1. 已发表的论文：

1. X Wang, X Y Lou, S Y Hu and S C He. Evaluation of safe driving behavior of transport vehicles based on K-SVM-XGBoost[J]. Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2020:84-92.

2. 参与科研项目：