MathorCup
数学应用挑战赛

MathorCup数学应用挑战赛

大数据竞赛

参赛手册

（内含：参赛指南、历年赛题及优秀论文）

扫描二维码了解更多竞赛咨讯

目　录

竞赛报名通知..... 壹

历年赛题..... 1

2025 年 A 题..... 2

2025 年 B 题..... 4

2024 年 A 题..... 14

2024 年 B 题..... 18

2023 年 A 题..... 23

2023 年 B 题..... 26

2022 年 A 题..... 30

2022 年 B 题..... 34

2021 年 A 题..... 37

2021 年 B 题..... 42

优秀论文..... 48

2025 年 A 题..... 49

2025 年 B 题..... 81

2024 年 A 题..... 101

2024 年 B 题..... 139

2023 年 A 题..... 171

2023 年 B 题..... 198

2022 年 A 题..... 227

2022 年 B 题..... 254

2021 年 A 题..... 285

2021 年 B 题..... 320

MathorCup 数学应用挑战赛组委会文件

2026 年第七届 MathorCup 数学应用挑战赛—大数据竞赛报名通知

MathorCup 大数据竞赛[2026-01]号

为了展现高校学生学科基础，提高学生运用数学和大数据分析方法解决实际问题的能力，培养创新精神及合作意识，促进课外科技活动，2020 年起 MathorCup 数学应用挑战赛组委会决定在“MathorCup 数学应用挑战赛”中增加大数据独立赛项，全称“MathorCup 数学应用挑战赛—大数据竞赛”（简称“MathorCup 高校大数据竞赛”）。竞赛由中国优选法统筹法与经济数学研究会主办，MathorCup 数学应用挑战赛组委会（以下简称组委会）具体负责竞赛的组织，欢迎全国各高校的在学研究生、本科生和专科生参赛，也欢迎各高等院校组织同学报名参赛。

1. 参赛对象为高等院校在校学生。学生以队为单位报名参赛（每队最多 3 名学生，最多 1 名指导教师），允许跨专业、跨年级、跨校组队参赛。通过学校相关部门统一报名或者在官方主页（http://www.mathorcup.org）独立报名。报名截止日期为 2026 年 10 月 23 日 12:00，报名截止后不能再更改报名信息。报名时要求完整填写每位参赛者以及指导教师所在的学校全称。每所院校参赛队数量不作统一规定。

2. 竞赛初赛时间为 2026 年 10 月 23 日 18:00 至 2026 年 10 月 30 日 20:00。初赛成绩前 10%的队伍进入复赛，复赛时间为 2026 年 12 月 4 日 18:00 至 2026 年 12 月 11 日 20:00。2027 年 1 月公布获奖名单。

3. 竞赛分为研究生组、本科组、专科组进行。报名时按参赛队员中学历最高者，选择参赛组别。

4. 每个参赛队需缴纳 200 元参赛费，用于竞赛组织、赛题评阅、技术支持等。参赛费用在报名网站提交或者集体缴费。

5. 集体报名的，高校负责人须通知本校参赛队伍在报名官网：http://www.mathorcup.org 完成注册，不需要缴费，生成队伍的参赛编号后填写

集体报名表格，同集体报名支付截图一起发至 MathorCup 数学应用挑战赛组委会（以下简称组委会）邮箱（bigdata@mathorcup.org）。组委会工作人员接收集体报名表格后会在官网审核通过集体报名的队伍编号，操作完成后会发邮件给负责人核对。

6. 集体报名缴费方式：竞赛报名费由中国优选法统筹法与经济数学研究会收取，每支队伍 200 元参赛费。汇款信息如下：

银行户名：中国优选法统筹法与经济数学研究会

开 户 行：中国工商银行海淀西区支行

账　　号：0200004509089143934

注：支付时请备注 “学校+队数+支付者姓名”，集体转账后请负责人保留转账成功截图，并同集体报名表一同发至组委会邮箱（bigdata@mathorcup.org），如果需要开发票，请把发票抬头、纳税人识别号、是否开成一张发票等信息一起发送到组委会邮箱。

7. 竞赛题目共两个赛题，参赛队任选一题进行解答。题目一般来源于企业或行业经过适当简化加工的实际大数据问题，提供相关的数据，不要求参赛者预先掌握深入的行业专业知识。题目有较大的灵活性供参赛者发挥其创造能力。赛题将于竞赛开始时在竞赛官方主页以及竞赛报名网站上同时公布，不邮寄书面题目。

8. 竞赛分为初赛、复赛和金银铜牌答辩三个阶段。初赛和复赛采用网上发布题目，参赛队线下研讨、解答，并在网上提交答卷的方式进行。金银铜牌答辩采用线上 PPT 讲解并回答专家提问的方式进行。

9. 竞赛需要提交电子版论文、结果数据集和计算程序源代码。论文提交方式如下（具体请参见《MathorCup 数学应用挑战赛—大数据竞赛提交规范》）：

提交支撑材料：将与竞赛相关的其他所有文件（包括程序、数据和结果等）压缩打包上传。“支撑材料”由参赛队员在审慎考虑的基础上，选择性地添加，如竞赛题目中没有明确的要求，则不作为必须提交的材料。（如果有自己编写的程序，虽然按照论文格式要求，程序必须放入正文附录，但源程序仍然要以支撑材料的形式提供。）

论文命名为：题号+参赛编号（文件类型为 PDF）；

承诺书命名：题号+参赛编号+承诺书（文件类型为 PDF）；

支撑材料命名：题号+参赛编号+附件（文件类型为 rar 或 zip）。

例：MCB2600001 队选择的是 A 题，则论文命名应为：“AMCB2600001”，承诺书命名应为：“AMCB2600001 承诺书”，支撑材料命名应为：“AMCB2600001 附件”。

然后分别上传至报名官网的对应电子档提交处。

10. 从参赛队伍中分组别、分赛题评选出一等奖（5%）、二等奖（15%）、三等奖（30%），颁发“MathorCup 数学应用挑战赛—大数据竞赛”获奖证书。初赛中排名前 10%的队伍进入复赛。从获得一等奖的队伍中，通过答辩遴选出（不分组别、不分赛题）金奖队伍 1 支（奖金 10000 元税前）、银奖队伍 2 支（每队奖金 5000 元税前）、铜奖队伍 5 支（每队奖金 2000 元税前）。

11. 请有关参赛学校和师生在竞赛开始前认真阅读《MathorCup 数学应用挑战赛—大数据竞赛章程》，严格按照竞赛章程和参赛规则参赛。

12. 联系方式：

电话&微信：18210922591（竞赛秘书处-郭老师）

竞赛邮箱：bigdata@mathorcup.org

报名网址：http://www.mathorcup.org

竞赛章程和竞赛文件见竞赛官网：http://www.mathorcup.org，竞赛的最新消息均以该网站公布的为准。

中国优选法统筹法与经济数学研究会

MathorCup 数学应用挑战赛组委会

2026 年 8 月 10 日

历 年 赛 题

MathorCup数学应用挑战赛组委会制作

赛道 A：基于计算机视觉的集装箱智能破损检测

随着全球贸易的快速发展，集装箱作为国际货物运输的核心载体，被广泛应用于海运、陆运及多式联运环节。长期的装卸、堆叠、运输过程中，集装箱常常受到机械冲击、腐蚀、变形等外力作用，导致箱体外表面产生裂纹、凹陷、穿孔、锈蚀等不同类型的损伤。这些破损不仅会影响集装箱的结构强度与密封性能，造成货物泄漏、污染或丢失，还可能引发运输安全事故以及造成经济损失。因此，如何实现对集装箱破损的高效、准确、标准化检测，已成为港口运输管理与国际物流安全中的关键问题。

传统的集装箱破损检测主要依赖人工巡检或简单图像比对，存在检测效率低、主观性强、环境适应性差等问题，难以满足港口自动化和大规模作业的需求。近年来，随着计算机视觉与深度学习技术的迅速发展，基于图像识别的自动检测方法成为研究热点。随着港口自动化、智能化的发展，通过固定摄像头或龙门吊摄像头自动采集集装箱图像已成为可能。通过引入卷积神经网络（CNN）、目标检测模型（如 YOLO、Faster R-CNN 等），可以在复杂背景下自动识别多类型破损，实现集装箱状态的智能化评估和实时监测，为智慧港口建设和安全运输提供了技术支撑。

初赛问题

请你们队利用提供的集装箱外表面图像数据集，构建智能识别模

型。该模型需要完成以下任务：

问题 1 分类任务：结合给出的图像文件，提取图像特征，建立一个模型判断该集装箱图像是否存在任何形式的残损。

问题 2 检测与分割任务：如果存在残损，需要定位到残损所在的位置，并识别出残损的具体类别。

问题 3 模型评估：对于上述问题 1，2 中建立的模型，试从不同的维度进行评估。

利用上述建立的模型对验证集中的图片进行检测，将检测结果放在“test_result.csv”中。（注：验证集将在竞赛结束前 24 小时公布下载链接，请及时关注报名网站 https://www.saikr.com/vse/bigdata2025 的通知公告）

提示：1. 赛题求解存在多重困难，参赛队主要从以下几个方面考虑：所给数据集中图像背景复杂（港口机械、天空、地面），存在光照变化、反光、阴影、雨水、污渍等干扰；残损大小不一，大的如大面积锈蚀，小的如细微裂缝，对模型的多尺度特征提取能力要求高；无残损图片可能远多于有残损图片；某些残损类别（如“裂缝”）的样本数量可能远少于其他类别（如“凹痕”）；某些残损类别外观相似（如深凹痕与破损），难以区分。

2. 验证集中每张图片集装箱的破损按照严重程度从大到小最多取 4 个。

赛道 B：物流理赔风险识别及服务升级问题

物流企业在经历过去数年的快速扩张后，客户服务和体验提升成为核心议题。在履约过程中会不可避免存在包裹丢失、破损等问题，物流公司往往需要对其进行赔付。体验问题导致的物流理赔，成为了当前物流公司的重要环节之一。一方面，优质的理赔服务可以帮助提升品牌形象；另一方面，也需要控制理赔成本，确保企业健康发展。

理赔服务通常分为售前及售后两部分。售前环节以提前发现风险为目标，为客户推荐对应的物流服务；售后环节以理赔成本控制为目标，在保障合理理赔的前提下，减少超额赔付的出现。

我们首先关注售后部分，当发生履约问题，客户提交索赔诉求后，企业需要综合索赔商品、事故原因等信息，判断客户的索赔诉求是否合理。此时会分为如下三种情况。

1）如果客户的诉求合理，物流服务商会采用快速理赔，实现款项快速到账；

2）如果客户的诉求偏高，则需客户提供相关证据，并进入人工审核环节，随后协商赔偿方案；

3）如果客户索赔诉求严重超额，企业会采取更加严格的谈赔标准，并采用人工审核方式额外关注此部分运单。

判断客户的诉求是否合理，可以由“索赔差额”反映出来。对于已经完成谈赔的历史运单，经过人工谈赔或企业直接赔付后的实际赔

付金额，可以与客户原始要求的索赔金额进行比较。即：

索赔差额=实际赔付金额-索赔金额。

理赔服务需要首先将客户提交索赔申请的运单，标注成“合理诉求”、“诉求偏高”、“严重超额”三类，企业从而分别采用不同的策略，这一步称为“风险标注”。一般来说，“严重超额”的运单占比不会太高（通常小于3%），“合理诉求”的运单不低于85%。风险标注结果同时与索赔差额和实际赔付金额有关。实际赔付金额越高，则需要更高的索赔差额才会标注为“诉求偏高”或“严重超额”。对于相似实际赔付金额的运单标注时，分成同一类的运单，其索赔差额应尽可能接近。需要注意的是，一般来说“合理诉求”的运单索赔差额相对密集，“严重超额” 的运单索赔差额相对稀疏，且不同类型的运单索赔差额密集程度之间存在明显差别。

在进行风险标注后，需要使用标注后的结果，对未来需要索赔的运单进行分析，判断其属于哪一种类型，从而既减少人工审核介入的压力，也可以控制理赔的风险及成本。一般来说，可采用如下两种方式进行该分析。

1）预测该运单对应的“实际赔付金额”，根据预测值与客户提交的索赔金额比较，判断其属于哪一类运单；

2）直接根据运单中的信息，判断属于哪一类理赔运单，即直接预测出风险标注的分类结果。

理赔的结果，也可以反向优化物流售前的环节。在客户下单之前，物流服务商可以根据用户历史的寄件行为、历史理赔情况，对运单形

成评价，从而提供相应的服务。对于每天上千万的寄出运单，物流服务商可以根据不同的用户以及寄件情况，对其进行物流保障服务的推荐，从而对于优质客户的重要货物进行保障。

当前已有若干个历史运单，其中包含的数据信息如附件 1 所示，未来需要判断理赔服务的运单数据如附件 2 所示。在运单信息中，涉及到的相关特征及名词解释见附表 1 及附注。请根据以上数据回答以下问题。

问题 1：对附件 1 中“理赔差额”和“实际赔付金额”进行分析，并建立风险标注模型。需详细说明标注规则如何对运单进行划分，给出附件 1 的风险标注结果及分析。

问题 2：结合附件 1 中的数据，建立模型，预测附件 2 中运单的实际赔付金额，需说明预测结果的评估指标，并将预测结果填写在 result 文件中，不要改变表格中的运单号。

问题 3：使用问题 1 中给出的风险标注规则，并利用附件 1 中的数据建立风险标注的分类模型，预测附件 2 中运单的风险标注结果，并将预测结果填写在 result 文件中，与论文作品一同提交，不要改变表格中的运单号；问题 2、3 的结果放在一个 result 文件中即可。论文中需详细描述对于“严重超额”样本占比严重不均衡的处理方法，并论述使用当前方法及问题 2 中方法预测风险标注结果的优劣势。

附表 1：运单相关特征信息解释

数据类型	细分类型	特征中文名	说明
运单特征	运单属性	线路类型	线路类型 ID
		是否为 c2c	0/1 标识
		是否生鲜妥投及时	0/1 标识
		保价金额	运单保价金额
	寄收信息	始发城市	始发城市 ID
		目的城市	目的城市 ID
		寄件人账号	寄件人账号 ID
		收件人 ID	收件人 ID
		寄件是否内部	0/1 标识
	运输情况	配送超时时长	实际配送时间－预计配送时间
		异常原因	物流运输过程中员工提报异常，枚举包括 Pickup Error、Suspected Lost、Damage 等
客诉特征	客诉属性	进线渠道	枚举包括 Call、Internal Software,Mini Program 等
	进线时间	妥投到进线时长	进线时间 - 包裹妥投时间
	商品信息	商品类型	枚举：Fresh、Electronic Product、Jewelry、Appliance 等
		新旧程度	0/1/2 标识，代表未知/全新/二手
	进线人信息	进线人身份	枚举：Consiner、Receiver、Staff
		寄件 B/C	B、C、Staff、Not Known
行为特征	网点统计特征	始发网点发单量	始发网点近半年发单量
		始发网点万单理赔率	始发网点近半年发货理赔率×10000
		始发网点赔付比例	始发网点近半年赔付占索赔比例
		目的网点发单量	目的网点近半年发单量
		目的网点万单理赔率	目的网点近半年发货理赔率×10000
		目的网点赔付比例	目的网点近半年赔付占索赔比例

附注：

C2C：指个人用户之间进行商品交易时所使用的物流服务。

始发网点：指货物运输流程中，寄件人交寄包裹或货物的第一个物流操作物流站点，是整个物流链条的起点。

目的网点：指货物运输过程中最终送达收件人的末端物流操作网点，负责接收从上游转运来的货物，并完成最后的派送任务。

理赔率：指在一定统计周期内，因货物丢失、破损等原因发生理赔的运单数量占总运单数量的比例。计算公式一般为：理赔率 = （理赔单量 ÷ 总运单量）× 100%。

进线：指通过电话、小程序、APP 等渠道主动联系物流公司咨询、投诉的客户。“进线人”即发起该行为的个体，“进线渠道”即发起该行为的渠道。

妥投：指快件或货物已按照约定时间、地点和方式成功交付给收件人本人或其指定代收人，并完成签收手续，确认无误的配送状态。

保价金额：指寄件人在寄件时，向物流公司申报的货物实际价值，并支付相应保价费用后，物流公司承诺在发生损毁、丢失等情况时，按此金额上限进行赔偿的金额。

参考材料：

[1]智能供应链:数据科学理论与实战，庄晓天等，电子工业出版社，2025.8

2024 年 MathorCup 数学应用挑战赛——大数据竞赛

赛道 A：台风的分类与预测

台风属于热带气旋，是与人类生活和生产关系密切的降雨系统。然而，台风容易造成多种灾害，例如：狂风、暴雨、风暴潮、泥石流、生态破坏、疫病流行等，具有突发性强、破坏力大的特点，成为世界上最严重的自然灾害之一。

台风的成因，至今仍无法准确确定，仅已知它是由热带大气内的扰动发展而来的。在热带海洋上，海面因受太阳直射而使海水温度升高，海水容易蒸发成水汽散布在空中，造成热带海洋上的空气温度高、湿度大。这种空气因温度高而膨胀，致使密度减小，质量减轻，而赤道附近风力微弱，所以很容易上升，发生对流作用，同时周围冷空气流入补充，然后再上升，如此循环不已，使得整个气柱皆为温度较高、重量较轻、密度较小之空气，形成了“气旋”（低气压）。

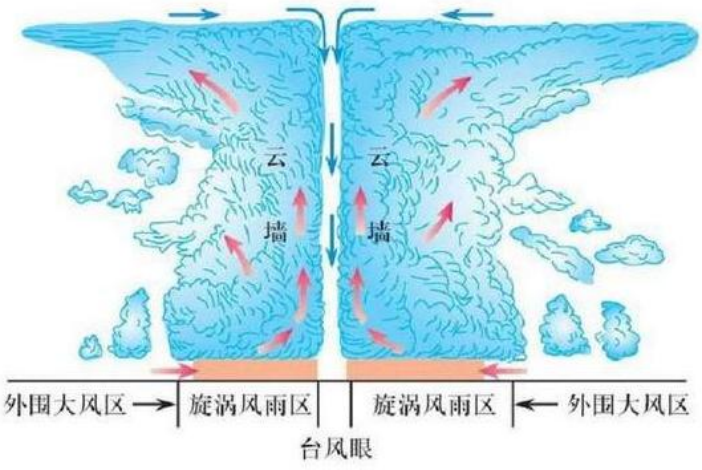


图 1 台风示意图

由于形成台风的影响因素众多，使得台风运动轨迹极其复杂，运动轨迹和影响因素之间呈现高度非线性。其中，台风的移动主要靠两种力：一种是内力，另一种是外力。内力是台风本身所产生的力。北半球台风是一团反时针方向旋转着的空气，在旋转时，空气质点的移动方向，要受到地球自转的影响而发生偏向。这种偏向往往是台风向高纬度的一侧比向赤道

的一侧来得大，促使台风向北移动，即为内力。另外，夏秋之际，太平洋上常有一个独立的高气压（副热带高气压），这个高气压四周的风向对台风的移动路径具有直接的影响。台风发生在副热带高压的南部边缘，会形成东风，使台风向西行，即为外力。如果副热带高气压西伸并加强，台风路径就在偏南地方向西行进；如果副热带高气压在台风北方东退或断裂，台风就可能在高压的西缘或裂口处转向北行，当绕到高压西北边缘，在西南风影响下，就向东北方向前进。另外，台风的路径有时还会受到其他台风的影响，出现打转、停滞的现象（例如：双台风或多台风相互影响）。

请结合大数据分析的方法完成以下问题：

初赛问题

问题 1：分析台风特征参数（强度、等级、风速等）与气温、气压、季风的关系，依据气温、气压、季风等因素的影响，建立台风的分类评价模型，明确类别划分的标准或评价算法，进行不同特征的台风类别划分。夏台风生成于 6-8 月，秋台风 9-11 月，根据所建立的分类评价模型，以 2024 年我国 7 月和 9 月为例，给出台风类别及途经省份的列表，并分析夏台风与秋台风的区别。

表 1：2024 年我国 7 月和 9 月台风类别及途经省份

月份	台风类别	途经省份
7 月 夏台风		
9 月 秋台风		

注：表格仅做示意，可根据情况修改增减。

问题 2：除温度、气压、季风外，台风的路径还受地球自转偏向力等多

重因素影响。根据气温、气压、洋流、风场等多种因素，建立台风路径预测的模型。并且，以 2024 年 9 月 13 日-17 日每日 14 点的第 13 号台风贝碧嘉为例，预测行进路线，运用 Dynamic Time Warping(DTW)动态时间规整算法与台风实际行进路线进行对比。

表 2：2024 年 9 月 13 日-17 日每日 14 点的第 13 号台风贝碧嘉的中心位置

时间	台风中心位置(经度/纬度)
13 日 14：00	
14 日 14：00	
15 日 14：00	
16 日 14：00	
17 日 14：00	

注：预测结果填到表 2 中，并放于论文正文。

问题 3：台风登陆后，风速和雨量将逐渐减弱，且某地区受台风影响而产生的风速和降水量，随着其与台风中心距离的增加而减小。请建立台风在登陆后的行进过程中降水量及风速的关系，及降水量与距台风中心距离的关系。并且，以 2024 年 9 月 16 日-18 日的第 13 号台风贝碧嘉为例，根据所建立的模型，预测贝碧嘉行进途中的中心风力及降水量，进行分析。

数据说明：

附件 1：中国近海台风路径集合(1945-2023).zip;

其他可参考数据：

数据不限于提供的数据，参赛者可根据需要自行查找其他数据，请注明数据来源。

以下参考数据已经下载见百度网盘：

链接: https://pan.baidu.com/s/1JZ_163qGoIDQkKqcDvT4Pw?pwd=8394

提取码: 8394

数据来源网址：

海洋科学大数据中心：

https://msdc.qdio.ac.cn/data/metadata-special-detail?id=1422759994058625025

北半球气压数据(1899-至今)，数据来源：https://rda.ucar.edu/datasets/d010000/dataaccess/

1929-2024 年全球站点的逐日平均风速数据（ShpExcel12000 个站点）数据来源：https://www.ncei.noaa.gov/data/global-summary-of-the-day/archive/

海洋气温数据(1991-至今)：https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html

海洋温度数据：

http://www.ocean.iap.ac.cn/pages/dataService/dataService.html?navAnchor=dataService

海洋表面温度数据、海洋温度数据(air.2m.gauss)、CHM_PRE_0.1dg_19612022（降水量场）数据压缩包，为气象数据常用的 netcdf4 形式存储的 nc 数据文件，可以利用 python 库 netcdf4、xarray 或者 pandas 进行数据处理。

降水量数据集：https://tcdata.typhoon.org.cn/rdqxfy.html

中国逐日降水数据集(1961-2022)：https://www.tpdc.ac.cn/zh-hans/data/e5c335d9-cbb9-48a6-ba35-d67dd614bb8c

其他：

欧洲哥白尼海洋中心数据：https://cds.climate.copernicus.eu/#!/home

ERA5：https://cds.climate.copernicus.eu/datasets

赛道 B：电商品类货量预测及品类分仓规划

电商企业在各区域的商品存储主要由多个仓库组成的仓群承担。其中存储的商品主要按照属性（品类、件型等）进行划分和打标，便于进行库存管理。图 1 是一个简化的示意图，商品品类各异，件数众多，必须将这些商品分散到各个仓库存储。品类分仓规划决定各商品存放在哪些仓库问题，合理的品类分仓规划对提升每个仓的管理效率、降低总体仓储成本至关重要。

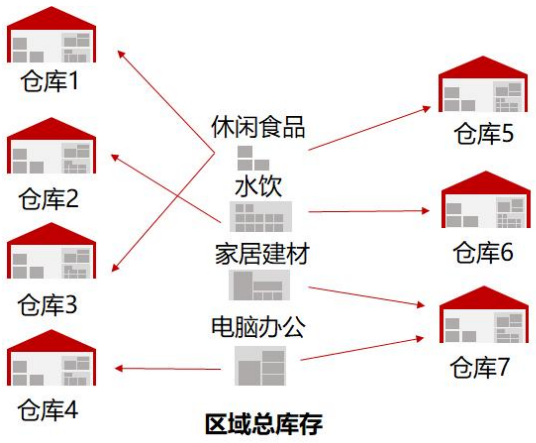


图 1 品类分仓规划示意图

准确的仓储货量预测是品类分仓规划的重要依据，对于准确的预测结果能够预见性地决定未来的仓储资源使用决策，以提前规划仓储资源，减少冗余场地的投入。一般来说，该场景需要预测两个目标，分别为库存量和销量。其中，库存量为该品类在全部仓库所需存放的总库存，分仓结果中受到仓库的仓容限制；销量为该品类在全部仓库所需打包出库的总量，分仓结果中受到产能限制。

在得到未来各品类的预测货量后，各个品类的分仓规划是供应链规划者的重要研究问题。若将品类集中存放在数量较少的仓库中，则将超过该仓的仓容及产能上限，造成履约问题；若同一品类分在多个仓库中，则会

显著增加仓库数量，增大品类库存的管理难度及总成本。此场景需考虑的上限包括两个，分别为仓容上限和产能上限，其中仓容上限为某仓库可以存放的最高库存量，产能上限为某仓库一天可以出库的最高销量。另外，若将相似的品类（使用品类关联度衡量相似性）放在同一个仓库中，同一订单中的商品更可能集中出货，可以在实际履约中减少包裹数量，从而降低履约成本。

合理的品类分仓方案，应该同时考虑仓群的复杂度及单仓仓容及产能约束，给出最优的分仓结果需综合考虑以下指标：

1) 仓容利用率：单仓总库存/仓容上限；
2) 产能利用率：单仓总出库量/产能上限；
3) 总仓租成本：使用仓库的仓租成本之和；
4) 品类分仓数：单品类存放的仓库数量；
5) 品类关联度：存放在同一仓库的所有品类之间的关联度之和。

现有一个仓储网络，包含 140 个仓库以及 350 种品类，附件 1 及附件 2 分别为各品类的历史库存量及销量，附件 3 为不同仓库相关信息（仓租日成本、仓容上限、产能上限），附件 4 为不同品类之间的关联度（表中未出现的品类组合关联度设为 0），附件 5 为不同品类的相关信息（品类编码、件型）。基于以上数据，请完成以下问题：

初赛问题

问题 1：建立货量预测模型，对该仓储网络 350 个品类未来 3 个月（7-9 月）每个月的库存量及销量进行预测，其中库存量根据历史每月数据预测月均库存量即可，填写表 1 的预测结果并放在正文中，并将完整结果填写在 result 表格文件中的“月库存预测结果”的表单中；销量需给出未来每天的预测结果，填写表 2 的预测结果并放在正文中，并将完整结果填写在 result 表格文件中的“日销量预测结果”的表单中。

表 1：月库存量预测结果

	7 月库存量	**8 月库存量**	**9 月库存量**

category1			
category31			
category61			
category91			
category121			
category151			
category181			
category211			
category241			
category271			
category301			
category331			

表 2：日销量预测结果

	7.1	**7.11**	**7.21**	**7.31**	**8.11**	**8.21**	**8.31**	**9.11**	**9.21**
category1									
category31									
category61									
category91									
category121									
category151									
category181									
category211									
category241									
category271									
category301									
category331									

问题 2：假设当前限定每个品类只能放在一个仓库中，即一品一仓，各品类之间请基于问题 1 的预测结果建立规划模型，综合考虑多个业务目标，求得品类的分仓方案，包括：应使用哪些仓库，使用的仓库需存放哪些品类的库存。填写表 3 的分仓结果并放在正文中，并将完整品类分仓结果填写在 result 表格中的“一品一仓分仓方案”表单中。

表 3：“一品一仓”分仓方案

	warehouse
category1	
category31	
category61	
category91	
category121	
category151	
category181	
category211	
category241	
category271	
category301	
category331	

问题 3：现在为每个品类按照件型及高级品类进行打标（如附件 5），并放开一品一仓假设，即允许一个品类存放于多个仓库，但同一品类存放的仓库数量不能超过 3 个，并希望同件型、同高级品类尽量放在一个仓库中。假设同一品类在不同仓库之间分布的库存量比例及出库量比例相同，当前业务的首要目标是最大品类关联度，同时兼顾其他指标。请基于问题 1 的预测结果建立规划模型，求得新的品类分仓方案，并分析不同方案中各业务指标的表现。填写表 4 的分仓结果并放在正文中，并将完整品类分仓结果填写在 result 表格中的“一品多仓分仓方案”表单中。

表 4：“一品多仓”分仓方案

	warehouse	warehouse	warehouse
category1			
category31			
category61			
category91			
category121			
category151			
category181			
category211			
category241			
category271			
category301			
category331			

注：提交论文时，请将 result 结果文件表作为计算结果提交，提交时不要改动 result 表格的格式。

参考材料：

[1] 庄晓天等. 智能供应链:预测算法理论与实战[M]. 北京: 电子工业出版社, 2023.

[2] [1] Makridakis S, Spiliotis E, Assimakopoulos V. M5 accuracy competition: Results, findings, and conclusions[J]. International Journal of Forecasting, 2022.

[3] 庄晓天等. 智能供应链:运筹优化理论与实战[M]. 北京: 电子工业出版社, 2024.

[4] Catalán, Andrés, Fisher M. Assortment Allocation to Distribution Centers to Minimize Split Customer Orders[J]. Social Science Electronic Publishing, 2013

[5] Minfang HUANG, Yuanyuan HAO, Yanxin WANG, Xiangpei HU, Ludi LI. Split-order consolidation optimization for online supermarkets: Process analysis and optimization models[J]. Frontiers of Engineering Management, 2023, 10(3): 499–516

2023 年 MathorCup 高校数学建模挑战赛——大数据竞赛

赛道 A：基于计算机视觉的坑洼道路检测和识别

坑洼道路检测和识别是一种计算机视觉任务，旨在通过数字图像（通常是地表坑洼图像）识别出存在坑洼的道路。这对于地质勘探、航天科学和自然灾害等领域的研究和应用具有重要意义。例如，它可以帮助在地球轨道上识别坑洼，以及分析和模拟地球表面的形态。

在坑洼道路检测任务中，传统的分类算法往往不能取得很好的效果，因为坑洼图像的特征往往是非常复杂和多变的。然而，近年来深度学习技术的发展，为坑洼道路检测提供了新的解决方案。

深度学习具有很强的特征提取和表示能力，可以从图像中自动提取出最重要的特征。在坑洼图像分类任务中，利用深度学习可以提取到坑洼的轮廓、纹理和形态等特征，并将其转换为更容易分类的表示形式。同时，还可以通过迁移学习和知识蒸馏等技术进一步提升分类性能。例如，一些研究者使用基于深度学习的方法对道路图像进行分类，将其分为正常、坑洼两类；另外，一些研究者还使用基于迁移学习的方法，从通用的预训练模型中学习坑洼图像的特征，并利用这些特征来分类坑洼图像。

本赛题希望通过对已标记的道路图像进行分析、特征提取与建模，从而对于一张新的道路图像能够自动识别坑洼状态。具体任务如下：

初赛问题

问题 1：结合给出的图像文件，提取图像特征，建立一个识别率高、速度快、分类准确的模型，用于识别图像中的道路是正常或者坑洼。

问题 2：对问题 1 中构建的模型进行训练，并从不同维度进行模型评估。

问题 3：利用已训练的模型识别测试集中的坑洼图像，并将识别结果放在“test_result.csv”中。（注：测试集将在竞赛结束前 48 小时公布下载链接，请及时关注报名网站 https://www.saikr.com/vse/bigdata2023 的通知公告）。

附件说明：

附件 1：data.zip；

训练数据集，文件中共包含 301 张图片。

文件名中包含“normal”字符表示正常道路，否则为坑洼道路。



图 1：正常道路示例



图 2：坑洼道路示例

附件 2：test_result.csv；

测试结果提交文件，文件中表头保持不变，数据仅做示例，提交的时候请删除后重新填写，字段描述见下表。

表 1：test_result 表字段说明

字段	说明
fnames	测试图片的文件名
label	分类标识：填写 1 和 0，1 表示正常道路；0 表示坑洼道路

附件 3：test_data.zip；

测试数据集，文件中包含几千张图片，具体数量以公布的数据为准。

测试数据集将在竞赛结束前 48 小时公布下载链接，请及时关注报名网站 https://www.saikr.com/vse/bigdata2023 的通知公告。

2023 年 MathorCup 高校数学建模挑战赛——大数据竞赛

赛道 B：电商零售商家需求预测及库存优化问题

电商平台存在着上千个商家，他们会将商品货物放在电商配套的仓库，电商平台会对这些货物进行统一管理。通过科学的管理手段和智能决策，大数据智能驱动的供应链可以显著降低库存成本，同时保证商品的按时履约。一般来说，以上供应链优化问题会包含以下方面：

需求预测

预测往往是智能供应链的决策基础，它可以让管理者提前预知各地的需求，从而将库存提前放在靠近需求的仓库中，此时的预测任务为：根据历史一段时间的需求量，预测各仓库中各商品未来需求，“预测维度”即为不同商家在各仓库中存放的各种商品每天的数量。一般来说，企业会首先根据数据的历史情况，分析出需求量序列的数理特征，对相似的需求量序列进行归类，并根据分类结果做到更加精准的预测。预测准确率有很多评价指标，较常用的指标为 1-wmape，定义如下：

$$1-\text{wmape}=1-\frac{\sum|y_i-\hat{y}_i|}{\sum y_i}$$

其中y_i为第i个序列（商家在各仓库中存放的各种商品每天的数量）的真实需求量，$\hat{y}_i$为第i个序列的预测需求量。

然而，在实际的电商供应链预测任务中，常常会出现多种不同的问题。例如，部分商品的销售时间过短、仓库存在新增或切换等情况，导致该预测维度下历史数据过少；另外，部分大型促销期间货量的陡增并由此带来的不规律性，也给需求量的精准预测带来了不小的难度。此时便需要通过算法得到历史一般规律，找出相似的历史情况（如相似的仓或商品），从而实现精准预测。

库存优化

为了有效管理库存水平，企业通常会制定各种库存策略来控制商品的补货频次和补货量。定期盘点库存策略(s，S)是一种常见的库存策略，其中

s 表示该种策略下的库存下限，S 为库存上限。在这种策略中，企业会每隔一段固定时间查看库存水平，盘点之间的间隔设为盘点周期（NRT）。如果在盘点时库存水平 T 低于最小库存水平 s，则会将库存补充至 S，此时的补货量 $Q = S - T$，如果在盘点时库存水平高于 s，则不进行补充，直到下一次库存盘点。同时需要考虑到，在每次发起补货后，补充的货物会经过一定时间后才能到达当前仓库，这个时间间隔为提前期（LT）。

使用以上的库存策略，管理者需要对两个方面进行权衡取舍。一方面，为了及时满足用户订单，公司需要保有大量库存以防止缺货；另一方面，库存水位升高带来了过高的库存成本。库存优化的目标主要为：

1）降低库存总持有成本，可由单位持有成本 h 得到（一件商品存储一天产生的成本）；

2）满足服务水平：确保有足够的产品可以及时满足客户需求，降低总缺货成本，可由缺货成本得到（每天每缺一件商品产生的成本）；

3）降低库存周转天数：库存周转天数越小，表示库存的流动性越好，企业的库存管理效率越高。

服务水平及库存周转天数计算公式为：

$$\text{服务水平}=\frac{\text{被满足的商品件数}}{\text{实际需求总件数}}\times 100\%$$

$$\text{库存周转天数} = \frac{(\text{期初库存数量}+\text{期末库存数量})}{2}\times\frac{\text{计算时间段内天数}}{\text{计算时间段内总销量}}$$

现有一张电商零售商家的历史出货量表（附件 1），给出了历史 6 个月各商家存放在电商不同仓库的商品每天的出货量。假设该出货量即为历史各商品在各仓库的需求量。同时，还可以取到各商品、商家、仓库的信息（附件 2-4），例如分类、品牌、生效日期等，这些信息的选择和引入会帮助更好的预测并管理供应链中的库存。

初赛问题：

问题一：使用附件 1-4 中的数据，预测出各商家在各仓库的商品 2023-05-16 至 2023-05-30 的需求量，请将预测结果填写在结果表 1 并上传

至竞赛平台，并对你们模型的预测性能进行评价。另外请讨论：根据数据分析及建模过程，这些由商家、仓库、商品形成的时间序列如何分类，使同一类别在需求上的特征最为相似？

问题二：现有一些新出现的商家+仓库+商品维度（附件 5），导致这种情况出现的原因可能是新上市的商品，或是改变了某些商品所存放的仓库。请讨论这些新出现的预测维度如何通过历史附件 1 中的数据进行参考，找到相似序列并完成这些维度在 2023-05-16 至 2023-05-30 的预测值。请把预测结果填写在结果表 2，并上传至竞赛平台。

问题三：每年 6 月会出现规律性的大型促销，为需求量的精准预测以及履约带来了很大的挑战。附件 6 给出了附件 1 对应的商家+仓库+商品维度在去年双十一期间的需求量数据，请参考这些数据，给出 2023-06-01 至 2023-06-20 的预测值。请把预测结果填写在结果表 3，并上传至竞赛平台。

注：结果表 1-3 打包一个压缩包，“结果表.zip”，提交 zip 文件到竞赛系统中。

附件 1-4 的字段说明：

附件 1：商家历史出货量表

字段名	字段类型	描述
seller_no	String	商家编码
product_no	String	商品编码
warehouse_no	String	仓库编码
date	String	日期
qty	Float	出货量（可看做需求量）

附件 2：商品信息表

字段名	字段类型	描述
product_no	String	商品编码
category1	String	商品一级分类

category2	String	商品二级分类
category3	String	商品三级分类

附件 3：商家信息表

字段名	字段类型	描述
seller_no	String	商家编码
seller_category	String	商家分类
inventory_category	String	库存分类
seller_level	String	商家规模

附件 4：仓库信息表

字段名	字段类型	描述
warehouse_no	String	仓库编码
warehouse_category	String	仓库类别
warehouse_region	String	仓库区域

各参赛队员可自行查找相关文献，以下参考文献仅供参考：

[1] 智能供应链:预测算法理论与实战[M]，北京：电子工业出版社，2023.

[2] Makridakis S，Spiliotis E，Assimakopoulos V．The M5 Accuracy competition: Results, findings and conclusions[J]. International Journal of Forecasting, 2020, 36(1):224-227.

[3] Makridakis S，Spiliotis E，Assimakopoulos V，et al. The M4 Competition: 100,000 time series and 61 forecasting methods[J]. International Journal of Forecasting, 2020, 36(1):54-74..

2022 年 MathorCup 高校数学建模挑战赛——大数据竞赛

赛道 A：“58 到家”家政服务订单分配问题

“58 到家”是“58 同城”旗下高品质、高效率的上门家政服务平台，平台向用户提供家政保洁、保姆、月嫂、搬家、维修等众多生活领域的服务。在家政保洁场景中，用户在平台下单购买服务后，平台会将订单分配给一个保洁阿姨，阿姨接到订单后按照用户指定的服务时间上门，进行保洁服务。平台在将订单分配给一个保洁阿姨时，一方面，为了提高对顾客的服务质量，需要尽量分配服务分较高的阿姨，其中阿姨的服务分是基于阿姨历史订单的评价情况得到，取值为[0,1]，值越大越好；另一方面，为了帮助阿姨提高接单量，需要尽量缩短阿姨相邻单之间的通行时间。

每天通过平台进行分配的订单量是巨大的，当前平台实现了一套订单分配算法，本问题研究的是如何优化系统的分配算法，提高算法的求解能力，实现提升顾客体验、节省阿姨时间。

数据说明：

数据包含一天内、一个区域的所有订单和所有保洁阿姨。

表 1：订单数据表字段说明

字段	数据类型	说明
id	int	订单唯一标识
createTime	int	订单下单时间，单位秒
serviceFirstTime	int	服务开始时间的最早时间，单位秒
serviceLastTime	int	服务开始时间的最晚时间，单位秒
serviceUnitTime	int	服务时长，单位分钟
x	int	服务地点横坐标，单位米
y	int	服务地点纵坐标，单位米

示例：

id,createTime,serviceFirstTime,serviceLastTime,serviceUnitTime,x,y
0,1661422500,1662786000,1662789600,90,-4693,-12716

1,1661474400,1662768000,1662771600,60,2176,-17371
2,1661568300,1662771600,1662775200,60,3940,-11414
3,1661575800,1662789600,1662793200,90,1791,-16026

表 2：阿姨数据表字段说明

字段	数据类型	说明
id	int	阿姨唯一标识
serviceScore	double	阿姨服务分
x	int	阿姨初始点横坐标，单位米
y	int	阿姨初始点纵坐标，单位米

示例：

id,serviceScore,x,y
0,0.9998,931,-31576
1,0.9993,7529,-7145
2,0.9992,5393,2140
3,0.999,5982,-3340

基本业务抽象

约束条件及假设：

1. 所有订单都要分配一个且只有一个阿姨；
2. 每个订单需要指定一个服务开始时间，这个时间的取值范围为[最早时间，最晚时间]，且是半点的整数倍；
3. 一个阿姨同时只能服务一个订单；
4. 阿姨需要在每个订单的服务开始时间之前到达客户位置；
5. 阿姨每天开始任务时必须从初始点位置出发；
6. 任意两点的距离为欧式距离；
7. 保洁阿姨的行驶速度为 15 千米/小时。

优化目标：

将每个订单匹配阿姨时，优化的目标是：

1. 所有订单匹配的阿姨的服务分，其平均值 A 尽可能大；
2. 最小化每单的平均通行距离 B。一个订单的通行距离指的是阿姨从上一个地点到本单地点的距离（欧式距离），其中阿姨第一个订单的通行距离等于从初始点到

第一个订单位置的距离，单位是千米；

3. 最小化阿姨服务订单的平均间隔时间 C。一个订单的间隔时间指的是，阿姨从上一个单服务结束时刻到本单服务开始时刻的时间间隔，单位是小时，其中阿姨第一个订单的间隔时间设定为 0.5 小时（阿姨首单需要做基本的准备工作，不考虑阿姨从初始点到第一个订单的通行时间）；
4. 总体目标是各个目标的加权和：αA-βB-γC，其中α=0.78、β=0.025、γ=0.195，得分四舍五入取 6 位小数。目标值越大越好。

初赛问题

问题 1：只考虑离线批量派单模式。附件 1 与附件 2 中分别给出的是一天的所有订单信息与阿姨信息。

(a) 请设计最优的订单与阿姨匹配算法，将所有订单进行分配，并将求解结果填写到 result1.txt 中。（订单必须全部分配、阿姨不需要全部匹配订单）。

(b) 基于(a)的算法，请对附件 1 中的前 50 个订单与附件 2 中前 20 个阿姨，重新运行算法，给出阿姨的执行任务列表，并画出阿姨的行动轨迹图。

表 3：result1 提交结果格式说明

字段	数据类型	说明
id	int	订单唯一标识
serviceStartTime	int	服务开始时间，单位秒
auntId	int	服务阿姨 id

示例：

0,1662786000,260
1,1662768000,61
2,1662771600,84
.....

问题 2：线上批量派单模式。在实际业务场景中，通常采用固定的频率派单，每 30 分钟将该段时间内产生的新订单统一分配；分配时允许部分订单暂时不派单，称之为压单，但是压单订单必须满足服务开始时间的最早时间比当前时间晚于 2 个小时（不包括 2 个小时）也即满足：

serviceFirstTime-currentTime>2h；请设计这种情况下的每批订单的最优分配算法。并将求解结果 1-最终决策结果填写到 result21.txt 中，结果 2-每次决策结果填写到 result22.txt 中。

表 4：问题 2 结果 1-最终决策结果格式说明

字段	数据类型	说明
id	int	订单唯一标识
serviceStartTime	int	服务开始时间，单位秒
auntId	int	服务阿姨 id

示例：

0,1662786000,260
1,1662768000,61
2,1662771600,84
.....

表 5：问题 2 结果 2-每次决策结果格式说明

字段	数据类型	说明
currentTime	int	当前决策时间，单位秒
id	int	订单唯一标识
serviceStartTime	int	服务开始时间，单位秒，如果压单，则为-1
auntId	int	服务阿姨 id，如果压单，则为-1
retainable	int	是否压单，1 为压单，0 为不压单

示例：

1661423400,0,1662786000,260,0
1661475600,1,1662768000,61,0
1661569200,2,1662771600,84,0
.....

附件

附件 1：订单数据.txt;

附件 2：阿姨数据.txt;

result: result1.txt

2022 年 MathorCup 高校数学建模挑战赛——大数据竞赛

赛道 B：北京移动用户体验影响因素研究

移动通信技术飞速发展，给人们带来了极大便利，人们也越来越离不开移动通信技术带来的各种便捷。随着网络不断的建设，网络覆盖越来越完善。各个移动运营商，越来越重视客户的网络使用体验，从而进一步提升网络服务质量。

客户满意度是客户对运营商产品服务的满意程度，反映了客户期望与实际感知的产品服务之间的差异。特别是在信息透明、产品同质化的今天，客户满意度的表现成为各大运营商市场运营状况的重要体现。数字经济时代，各大运营商需要运用数字经济的管理理念和技术手段，建立客户体验生态的全方位系统性测评体系，实现客户满意度评测的数字化转型，让客户体验赋能商业决策，让商业决策真正服务客户，共同推动移动网络高质量可持续发展。

根据客户投诉，对影响用户体验的问题逐点解决，是传统提升客户满意度的方法。但是随着用户数量的大幅增加，移动产品的种类越来越丰富，客户的需求越来越高，传统的方法已经难以有效提升客户的满意度。本研究拟通过分析影响用户满意度的各种因素，为决策提供依据，从而实现更早、更全面提升用户满意度。

中国移动通信集团北京公司，让客户根据自身在网络覆盖与信号强度、语音通话清晰度和语音通话稳定性三个方面的体验进行打分，同时还让客户根据语音通话的整体体验进行语音通话整体满意度的打分，并统计整理影响客户语音业务体验的因素，希望以此来分析客户语音业务满意度的主要影响因素，并提升客户语音业务满意度。同时，对于上网数据业务，中国移动北京公司让客户根据自身在网络覆盖与信号强度、手机上网速度、手机上网稳定性三个方面的体验进行打分，同时还让客户根据手机上网的整体体验进行手机上网整体满意度的打分，并统计整理影响客户上网体验

的因素，希望以此可以分析影响客户上网业务体验的主要因素，并提升客户的上网体验。

初赛问题

基于以上背景，请你们的团队根据附件给出的数据，通过数据分析与建模的方法帮助中国移动北京公司解决以下问题：

问题 1：根据附件 1 和附件 2，分别研究影响客户语音业务和上网业务满意度的主要因素，并给出各因素对客户打分影响程度的量化分析和结果。附件 1、2 中各字段的解释说明见附件 5。

问题 2：结合问题 1 的分析，对于客户语音业务和上网业务分别建立客户打分基于相关影响因素的数学模型，并据此对附件 3、4 中的客户打分进行预测研究，将预测结果分别填写在 result.xlsx 的 Sheet1“语音”和 Sheet2“上网”两个工作表中，并上传到竞赛平台，说明你们预测的合理性。

附件

附件 1 语音业务用户满意度数据

附件 2 上网业务用户满意度数据

附件 3 语音业务用户满意度预测数据

附件 4 上网业务用户满意度预测数据

附件 5 附件 1、2、3、4 的字段说明

result.xlsx

result21.txt

result22.txt

要求将所有问题结果写入到 reslult 文件夹下的 3 个文件中，请勿修改文件夹及文件名，并将 result 文件夹压缩为 zip 格式，将 result.zip 提交到竞赛平台。

赛道 A：二手车估价问题

随着我国的机动车数量不断增长，人均保有量也随之增加，机动车以“二手车”形式在流通环节，包括二手车收车、二手车拍卖、二手车零售、二手车置换等环节的流通需求越来越大。二手车作为一种特殊的“电商商品”，因为其“一车一况”的特性比一般电商商品的交易要复杂得多，究其原因是二手车价格难于准确估计和设定，不但受到车本身基础配置，如品牌、车系、动力等的影响，还受到车况如行驶里程、车身受损和维修情况等的影响，甚至新车价格的变化也会对二手车价格带来作用。目前国家并没有出台一个评判二手车资产价值的标准。一些二手车交易平台和二手车第三方估价平台都从自身的角度建立了一系列估价方法用于评估二手车资产的价值。

在一个典型的二手车零售场景，二手车一般通过互联网等线上渠道获取用户线索，线下实体门店对外展销和售卖，俗称 O2O 门店模式。门店通过“买手”从个人或其他渠道收购二手车，然后由门店定价师定价销售，二手车商品和其他商品一样，如果定价太高滞销也会打折促销，甚至直接以较低的价格打包批发，直至商品最终卖出。

基于以上背景，请你们团队根据附件给出的数据，通过数据分析与建模的方法帮助二手车交易平台解决下面的问题：

初赛问题

问题 1：基于给定的二手车交易样本数据（附件 1：估价训练数据），选用合适的估价方法，构建模型，预测二手车的零售交易价格，数据中会对 id 类，主要特征类等信息进行脱敏。主要数据包括车辆基础信息、交易时间信息、价格信息等，包含 36 列变量信息，其中 15 列为匿名变量。字段如下：

序号	Features	Description
1	carid	车辆 id
2	tradeTime	展销时间
3	brand	品牌 id
4	serial	车系 id
5	model	车型 id
6	mileage	里程
7	color	车辆颜色
8	cityId	车辆所在城市 id
9	carCode	国标码
10	transferCount	过户次数
11	seatings	载客人数
12	registerDate	注册日期
13	licenseDate	上牌日期
14	country	国别
15	maketype	厂商类型
16	modelyear	年款
17	displacement	排量
18	gearbox	变速箱
19	oiltype	燃油类型
20	newprice	新车价
21	anonymousFeature	15 个匿名特征
22	price	二手车交易价格（预测目标）

请采用附件 1 中的“估价训练数据”（带标签）训练模型和测试模型，自行设置测试集，使用训练完成后的模型对附件 2 中的“估价验证数据”（不带标签）进行预测，并将预测结果保存在附件 3“估价模型结果”文件中，注意不要修改格式，单独上传到竞赛平台。

其中附件 1“估价训练数据”和附件 2“估价验证数据”只相差最后 1 列数据（二手车交易价格（预测目标）），附件 3“估价模型结果”文件字段如下：

车辆 id	预测价格
id1	预测价格 1
id2	预测价格 2

附件 1、附件 2、附件 3 中各字段间采用“\t”分隔符分割，不包含表头。

模型评测标准：

$$0.2*(1-Mape)+0.8*Accuracy_5$$

Ape(相对误差)：

$$Ape=\left|\hat{y}-y\right|\Big/y$$

Mape(平均相对误差)：

$$Mape=\frac{1}{m}\sum_{i=1}^{m}Ape_i$$

其中，真实值 $y=(y_1,y_2,\cdots,y_m)$，模型预测值为 $\hat{y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_m)$；

Accuracy$_5$(5%误差准确率)：

$$Accuracy_5=count(Ape<=0.05)/count(total)$$

其中，$count(Ape<=0.05)$为相对误差 *Ape* 在 5%以内的样本数量，$count(total)$为样本总数量。

问题 2：在门店模式中，车辆在被“买手”收车以后，会进入门店进行售卖，车辆能否成功交易，除了取决于销售的谈判技巧，更重要的是车辆本身是否受消费者青睐，价格是否公道。假设你们是门店的定价师，请你们结合附件 4“门店交易训练数据”对车辆的成交周期（从车辆上架到成交的时间长度，单位：天）进行分析，挖掘影响车辆成交周期的关键因素。假如需要加快门店在库车辆的销售速度，你们可以结合这些关键因素采取哪些行之有效的手段，并进一步说明这些手段的适用条件和预期效果。

附件 4“门店交易训练数据”包括 6 个字段，如下表所示，其中所有 carid 等相关信息包含在附件 1“估价训练数据”中。各字段间采用“\t”分隔符分割，不包含表头。

序号	Features	Description
1	carid	车辆 id
2	pushDate	上架时间
3	pushPrice	上架价格
4	updatePriceTimeJson	{价格调整时间：调整后价格}
5	pullDate	下架时间(成交车辆下架时间和成交时间相同)
6	withdrawDate	成交时间

问题 3：依据给出的样本数据集，你们觉得还有哪些问题值得研究，并给出你们的思路？

将问题 1、2、3 的解决过程写成一篇论文，明确你们的思路、模型、方法和结果。

附件

附件 1：估价训练数据.txt；

附件 2：估价验证数据.txt；

附件 3：估价模型结果.txt；

附件 4：门店交易训练数据.txt。

赛道 B：信息流智能推荐算法中的序列评估问题

随着互联网信息的蓬勃发展，用户在使用互联网应用时面临着信息过载的问题。推荐算法的出现，满足了用户个性化的内容消费需求，提升了用户获取有用信息的效率，在互联网 APP 里已被广泛应用。信息流作为推荐算法的主要应用场景，是用户触达互联网信息的主要入口，已经完全融入了人们的日常生活中，成为了人们了解世界的主要方式。

图 1 为信息流产品示例。该示例中，用户在执行刷新操作后，推荐系统返回了 K 条推荐结果，构成了一个推荐序列。其中前 4 条推荐内容占满了一个手机屏幕，用户继续下滑可浏览剩余内容。一个推荐序列由多种内容类型构成，例如，内容 1 为图文内容，内容 4 为视频内容。需要说明的是，推荐系统每次返回的内容条数 K 可以是不固定的，系统可以根据用户的具体请求环境进行动态调整，以此得到最佳的用户浏览体验，如何确定 K 的大小在本题目中不做讨论。



图 1 信息流产品示例

传统推荐算法的核心思想是挖掘被推荐内容与用户兴趣的匹配关系，以及内容本身的优质程度，选择与用户最相关或者最优质的内容推荐给用户。如图 2 中（a）所示，推荐系统会对单条候选内容进行打分评估，通过内容是否匹配用户兴趣以及内容质量的高低，来预估给用户推荐这条内容后带来的综合收益大小(综合收益通常包含用户是否会点击这条内容，以及用户在这条内容上的浏览时长)。系统打出的分值则是对每条内容带来的综合收益大小的刻画。然后，系统选择出预估分值最大的 K 条内容，并按照分值从大到小的顺序推荐给用户。这种推荐方式我们称之为 point wise。

但是，研究发现，除了内容本身因素以外，内容之间的排列组合关系，也会影响用户的浏览体验，进而影响推荐收益的大小。例如，相似内容的高度集中，往往会带来较差的结果反馈，即使它们都高度匹配用户兴趣或者具有较高的内容质量。于是，越来越多的研究集中在如何选择最优的内容排列组合上，而不仅仅是最优的内容上。如图 2 中（b）所示，同样的 ABC 三条内容，按照不同的顺序（A→B→C、A→C→B、……）推荐给用户，会带来不一样的收益大小。推荐系统需要先根据候选内容生成候选推荐序列，然后对每一个候选序列进行打分评估，系统打出的分值则是对每个序列带来的整体综合收益大小的刻画。最后，系统选择出预估分值最高的一个序列，按照该序列的排列顺序将内容推荐给用户。这种推荐方式我们称之为 list wise。本题目要求参赛者设计数学模型对给定的候选序列进行序列整体收益评估。

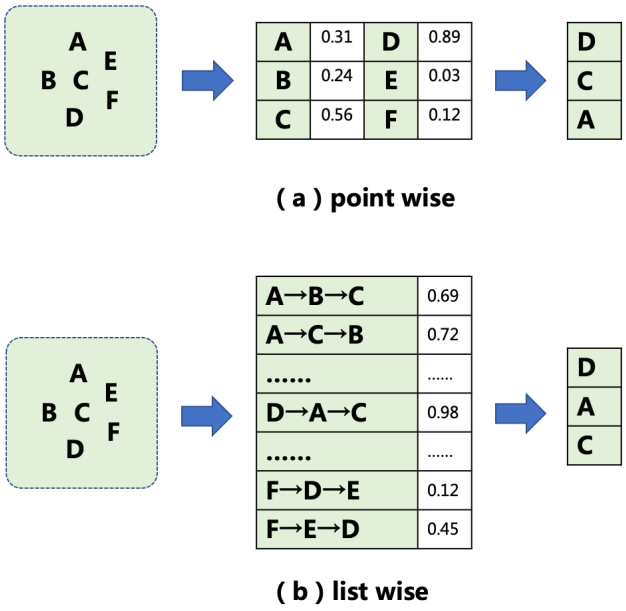


图 2 单内容评估和序列评估

题目提供近一周时间内用户在信息流产品上的曝光历史（train_data.txt）作为训练集，以供参赛者进行建模分析。附件中的 train_data.样例.txt 给出了数据格式示例，方便参赛者查看。涉及字段包括：

1. **用户 ID**：用户唯一标识，例如 1000024368；
2. **请求 ID**：用户单次请求推荐服务的唯一标识，例如 500012184_1635188998881_5305；
3. **日期**：用户单次请求推荐服务的日期，例如 20211026；
4. **时间**：用户单次请求推荐服务的时间，例如 22（代表晚上 22 点）；
5. **推荐序列**：用户单次请求推荐服务，推荐服务返回的内容列表。内容的排列顺序即为内容的真实推荐顺序，多个内容之间用“;”分隔，单个内容包括三个字段：内容 ID、用户是否点击（0 代表未点击，1 代表点击）、用户浏览时长（单位为秒），

多个字段之间用“:”分隔。例如133672454001:0:0;508896132:1:111;508969800:0:0;508870333:1:10;

同时，题目提供内容的基础属性（doc_info.txt）。附件中的doc_info.样例.txt 给出了数据格式示例。涉及字段包括：

1. **内容 ID**：内容的唯一标识，例如 133342615958；
2. **内容类型**：推荐内容分为视频（video）和图文（news）两种类型；
3. **内容类别**：内容的一二级分类，例如综艺/内地综艺；

最后，题目提供在训练集时间之后的一部分用户推荐序列作为测试集（test_data.txt），附件中的 test_data.样例.txt 给出了数据格式示例。参赛者需要根据训练集数据预测测试集序列的收益大小。涉及字段包括：

1. **请求 ID**：用户单次请求推荐服务的唯一标识，例如500012184_1635188998881_5305；
2. **用户 ID**：用户唯一标识，例如 1000024368；
3. **日期**：用户单次请求推荐服务的日期，例如 20211026；
4. **时间**：用户单次请求推荐服务的时间，例如 22（代表晚上 22 点）；
5. **推荐序列**：用户单次请求推荐服务，推荐服务返回的内容列表。内容的排列顺序即为内容的真实推荐顺序，多个内容之间用“;”分隔，单个内容只提供内容 ID，例如

508681374;133681260394;508767175;508767175;

上述完整数据集通过 https://pan.yidian-inc.com/index.php/s/QB7lhh7YPKLJWfL 进行下载获取。请参赛者对上述数据进行分析并建立模型，解决以下问题。参赛者需要将最终解决方案以论文方式进行详细阐述，包括主要模型、算法和计算结果，并以单独文件提交问题 2 的预测结果到竞赛系统中，不改变文件的格式。

问题 1：建立评估推荐序列总点击收益（序列中单条内容的点击量之和）和总时长收益（序列中单条内容的浏览时长之和）的数学模型，以及如何根据点击收益和时长收益对综合收益进行量化。不同于评估单个推荐内容收益的数学模型，在序列评估模型设计中，需要详细阐述如何考虑不同排列组合对收益大小的影响。

问题 2：基于问题 1 设计的数学模型，预测测试集（test_data.txt）中推荐序列的总点击量和总时长（单位为秒），将预测结果写入 result.csv 并提交。文件包含三列：请求 ID、总点击量、总浏览时长。请求 ID 对应测试集中的请求 ID，总点击量和总浏览时长为预估出的测试集中每个请求 ID 对应推荐序列的点击量和时长之和。附件 result.csv 中已给出的总点击量和总时长为随机生成的示例数据，参赛者需要将其替换为自己预测的总点击量和总时长值再提交。

问题 3：假设有 N 条候选内容，从中选择长度为 K（N≥K）的最优推荐序列，需要参与收益评估的序列量为A_N^K。在真实推荐场景中，由于计算性能的考虑，系统无法对所有可能序列进行收益评估，往往

需要先采用计算复杂度更低的方式对序列集合进行剪枝，圈定出少量候选序列进行精确收益评估。而剪枝策略的目标是保证候选序列集更大可能包含最优序列。请详细阐述你的建模思路，以及剪枝策略的精准度和时间复杂度。

优 秀 论 文

MathorCup数学应用挑战赛组委会制作

队伍编号	MCB2500884
题号	A

针对复杂场景下集装箱外表面损伤的智能检测与分析系统

摘 要

港口自动化与智慧物流的快速发展对集装箱外观损伤的检测精度和处理效率提出了更高要求。传统人工巡检方式效率低、误检率高，难以满足高密度、多批次的实时检测需求。为此，本文基于深度学习与计算机视觉技术，提出了一种从图像级残损判定、目标检测到结构健康评估的集装箱智能检测体系，实现了从视觉识别到安全量化的端到端建模与评估。

针对**问题一**，构建了基于残差神经网络的图像级二分类模型 **CABNet**，实现对“完好/破损”图像的快速判定。针对标签层级不一致问题，本文通过目标实例聚合策略构造图像级标签，并采用轻量卷积架构与迁移学习提升泛化性能。模型在 *test* 数据集上取得**准确率 98.0%**，能够有效完成残损样本的预筛选。

针对**问题二**，提出了改进型 **YOLOv11** 集装箱缺陷检测模型，用于识别凹陷（dent）、孔洞（hole）、锈蚀（rusty）及新增的表面划痕（scratch）四类破损。模型在主干网络中融合多尺度特征融合结构（FPN+PAN）与通道-空间注意力模块（CBAM），并采用 **COCO 预训练权重**进行迁移学习，最终模型在 *test* 数据集上取得 **Precision = 0.817**、**Recall = 0.750**、**mAP@0.5 = 0.809**、**mAP@0.5:0.95 = 0.628**。通过独立 IoU 匹配脚本验证预测框与真实框重叠率，平均 **IoU = 0.863**，检测成功率 **79.6%**，**F1-score = 0.728**，表明模型在复杂光照与背景干扰下仍具高鲁棒性。

针对**问题三**，基于检测结果构建了集装箱健康评估系统（Container Health Metrics System），实现结构级状态分析与安全等级判定。系统综合考虑缺陷类型、面积比例、位置分布及复合破损关系，建立损伤得分函数 $S_d = \sum_i \omega_{t_i} \cdot s_{a_i} \cdot \omega_{l_i} \cdot p_i$，并定义健康指数 $HI = 1 - S_d$。在此基础上，引入基于问题一当中的 **四分类检测模型**，实现对集装箱健康状态（Excellent、Good、Poor、Critical）的自动识别。批量评估结果显示，系统平均 **健康指数为 0.749**，其中 **Excellent 与 Good 样本占比 51.0%**，高风险样本（Poor + Critical）占比 **28.3%**。四分类模型在 *test* 数据集上达到 **总体准确率 80.0%**，进一步验证了健康指数在表征结构状态方面的有效性与可学习性。

关键词：深度学习；计算机视觉；CABNet；YOLOv11；集装箱健康评估系统

目录

一、 引言 1
1.1 研究背景与意义 1
1.2 国内外研究现状 1
二、 问题一求解 6
2.1 问题一分析 6
2.2 模型建立 7
2.2.1 特征提取骨干网络 8
2.2.2 注意力块 8
2.2.3 分类器模块 9
2.3 模型求解 10
2.4 结果分析 11
三、 问题二求解 14
3.1 问题二分析 14
3.2 模型建立 15
3.3 模型求解 16
3.4 结果分析 17
四、 问题三求解 19
4.1 多分类集装箱缺陷检测模型 19
4.1.1 数据处理 19
4.1.2 模型求解 20
4.1.3 结果分析 21
4.2 集装箱健康状态分析系统 22
4.2.1 系统设计 23
4.2.2 结果分析 23
五、 总结与展望 26
5.1 研究总结 26
5.2 研究展望 26
参考文献 28
附录 1： 问题一的 python 代码 31
附录 2： 问题二的 python 代码 48
附录 3： 问题三的 python 代码 127
3.1 多分类集装箱缺陷检测模型 127
3.2 集装箱健康状态分析系统 147

一、引言

1.1 研究背景与意义

在全球贸易持续扩张与现代物流体系日益复杂的背景下，集装箱作为国际货运的核心载体，其结构完整性与外观状态对保障运输安全与货物完好具有关键作用。在长期的装卸、堆叠及远途运输过程中，集装箱容易受到机械碰撞、环境侵蚀与结构疲劳等因素影响，导致表面出现裂纹、凹陷、锈蚀乃至穿孔等损伤。这些损伤若未能及时发现并处理，不仅会显著缩短集装箱的使用寿命，还可能引发货物受潮、污染等安全事故，给物流企业带来严重经济损失与运营风险。

目前，集装箱的外观检查仍主要依赖人工目视，该方法不仅效率低下，且受人员经验、疲劳程度及现场环境等多种因素影响，检测结果主观性强，难以满足现代化港口对高效率、自动化作业的需求。随着人工智能技术，尤其是计算机视觉领域的快速发展，基于图像的自动化缺陷检测逐渐成为替代人工检查的有效手段。通过在港口关键位置部署边缘算力设备，并结合先进算法，可实现集装箱外观状态的实时、高效与精准评估，显著提升检测的自动化水平与可靠性。

因此，研究具备高鲁棒性与泛化能力的智能检测技术，推动集装箱破损检测从“人眼判别”向“机器视觉”的转变，不仅有助于提升港口作业效率、降低运营成本，也对促进物流行业智能化升级、保障全球供应链安全具有重要的现实意义。

1.2 国内外研究现状

集装箱作为全球供应链的关键载体，承载了世界贸易的大部分货物流动[1]；近年在地缘冲突、航线绕行与成本波动背景下，其周转效率与安全合规的重要性进一步上升，亟需依托标准化、可量化的状态评估体系支撑运营连续性与风险控制。现场人工巡检受强反光、阴影、污渍与多尺度损伤（如凹陷、裂纹、锈蚀）干扰，存在主观性强、效率低与覆盖不足的问题[2]。基于计算机视觉的集装箱破损智能检测，能够在复杂背景下实现快速有无损筛查与关键区域精细定位，为港口与运营方提供可追溯证据链与可复制流程，因而具有突出的现实必要性与推广价值。

最早期的研究多以手工特征与规则或传统学习器为主线，围绕“增强—分割—表征—判别”的工程化流程展开。早期在海事金属结构领域，Bonnín-Pascual 与 Ortiz[3] 提出基于连通性与渗流思想的裂纹生长检测与基于颜色直方图与 GLCM 的锈蚀识别，在近实时条件下实现了裂纹连通形态的鲁棒提取与锈蚀区域的候选筛选，奠定了后续海事表面缺陷视觉化检测的算法雏形。其后，Ortiz 等人[4] 将图像采集平台与轻量前馈神经网络结合到船体巡检中，利用微型无人机拓展视角覆盖，以颜色与纹理等显式特征作为输入完成涂层破损与腐蚀的自动判别，显著缓解传统人工近距巡检的效率瓶颈。在“几何与多视”方向，Aijazi 等人[5] 基于多视位彩色三维 LiDAR 点云配准开展船体表面锈蚀斑点的自动检测与分析，通过跨视角重建减轻单幅图像光照与反光扰动的影响，提升了候选区域的一致性与可量化性。与此同时，工业表面检测的经典管线在多材料场景不断被验证。例如 Li 等人[6] 以 Gabor 滤波与 PCNN 进行织物缺陷增强与分割，体现了频域纹

理与脉冲耦合神经网络在弱对比缺陷上的互补性；针对金属与焊缝等结构化工业纹理，早期研究通常采用以梯度方向直方图等显式特征与支持向量机等分类器构成的“特征—分类器”范式来建立基线；在一般工业表面检测中，Ding 等人[7] 验证了 HOG 与支持向量机结合并配合 AdaBoost 特征选择的有效性，能够在低样本与近实时条件下保持鲁棒；而在焊缝射线图像场景，Valavanis 与 Kosmopoulos[8] 将几何与纹理描述子与多类支持向量机等学习器结合，实现对气孔、未融合与裂纹等多类型缺陷的自动识别，随后 Kasban 等人[9] 基于倒谱与高阶谱特征提出了自动检测流程，并在噪声与模糊条件下系统评估其鲁棒性。总体而言，这一时期的方法在可解释性与实现简便上具优势，但对复杂背景、强反光与阴影、多尺度与类别长尾仍敏感，跨场景泛化与极小缺陷（如细裂缝）召回率受限，从而为后续基于深度学习的端到端方法留出了改进空间。

在深度学习技术快速发展的背景下，卷积神经网络凭借其端到端学习与层次化特征表征能力，逐渐成为集装箱破损检测的主流方法，并催生了从图像分类到目标检测、再到精细分割的完整技术体系。早期研究者将深度卷积神经网络引入集装箱破损识别任务，İmamoğlu 等人[10] 对比了多种卷积神经网络架构在集装箱损伤二分类任务中的性能，通过实验验证了深度学习相比传统方法在准确率与鲁棒性上的显著优势，为后续研究奠定了基础。为实现破损区域的精确定位，目标检测技术被引入集装箱检测任务。Ren 等人[11] 提出的 Faster R-CNN 通过区域提议网络生成候选破损区域再经精细分类与边界框回归，实现了多类型破损的准确定位，其特征金字塔结构对多尺度破损具有良好适应性。Redmon 等人[12] 提出的 YOLO 系列通过单次前向传播同时完成分类与定位，显著提升了检测速度使实时监测成为可能; 其后发布的 YOLOv3[13] 采用 Darknet-53 骨干网络并引入多尺度预测策略，在保持速度优势的前提下提升了小目标检测性能。针对集装箱检测的特殊挑战，研究者们采用 Focal Loss 等损失函数缓解类别不平衡问题，设计多尺度特征融合策略增强对不同尺寸破损的检测能力。尽管目标检测方法解决了破损定位问题，但矩形边界框难以精确描述锈蚀与裂纹等不规则形状。为获得像素级精细定位结果，Ronneberger 等人[14] 提出的 U-Net 采用编码器-解码器结构与跳跃连接，通过融合浅层细节与深层语义信息实现了对破损边界的精细勾勒；He 等人[15] 提出的 Mask R-CNN 在 Faster R-CNN 基础上增加分割分支，同时完成检测、分类与像素级分割，有效处理多个破损区域重叠的复杂情况。研究者们进一步引入注意力机制突出破损特征并抑制背景干扰，为破损的定量评估提供了精确信息基础。尽管卷积神经网络在破损检测中取得显著进展，但现有方法在多尺度特征融合和跨层级信息交互方面仍存在局限，导致对不同尺寸破损目标的泛化能力不足，特别是在复杂港口环境中容易出现误检与漏检并存的现象, 为提高模型多尺度特征的判别能力提供研究方向。

随着深度学习技术在工业视觉检测中的广泛应用，基于 YOLO 系列和 ResNet 架构的改进模型逐渐成为集装箱破损检测的研究热点。这些方法在保持高推理速度的同时，通过优化网络结构、增强特征融合机制和引入注意力模块，显著提升了多尺度破损特征的提取能力与定位精度。在 YOLO 架构的改进方面，研究者针对集装箱表面破损的多尺度特性与复杂背景干扰进行了系列优化。Redmon 等人[13] 提出的 YOLOv3 首次引入 Darknet-53 骨干网络，通过残差连接缓解深层网络梯度消失问题，为后续改进奠定基础。Bochkovskiy 等人[16] 提出的 YOLOv4 整合了 Mish 激活函数、SPP 模块和尤其能增强多

尺度特征融合的 PANet 结构，显著提升了对小尺度目标的检测性能，从而为后续将其应用于锈蚀斑点等微小缺陷的检测奠定了坚实基础。

针对集装箱表面细微裂纹与锈蚀等小尺度破损难检测的问题，可借鉴自动驾驶领域的小物体检测思路。Benjumea 等人[17] 提出的 YOLO-Z 模型，通过替换骨干网络、增加检测层与优化锚框，有效增强了 YOLOv5 的微小目标检测能力，为该问题提供了直接的技术方案。Li 等人[18] 针对金属板材表面气孔、夹杂等微小点状缺陷在复杂纹理背景下难以识别的问题，将高效的 EMA 注意力机制嵌入 YOLOv7 的骨干网络末端，有效增强了模型对细微瑕疵的特征提取能力。为解决集装箱图像中破损区域尺度差异大的问题，Liu 等人[19] 提出的 YOLOv8 模型，其本身通过重新设计的特征金字塔的高效的 C2f 模块，实现了强大的多尺度特征融合能力，为检测不同尺寸的破损提供了先进的基准模型。在 YOLOv8 已为多尺度破损检测提供强大基准的基础上，后续版本在精度与效率的平衡以及特定缺陷的感知能力上持续深化。针对集装箱图像中破损形态的极端多样性与背景的复杂性，Wang 等人[20] 提出的 YOLOv9 模型引入了可编程梯度信息（PGI）和广义高效层聚合网络（GELAN），从根本上解决了深度网络中信息丢失所带来的瓶颈。该设计确保了从浅层细节特征到深层语义特征的高效传输与利用，使其在面对集装箱表面大面积油漆剥落与细微裂纹并存等复杂场景时，仍能保持卓越的特征表征能力，显著提升了复杂破损模式的综合检出率。为满足港口实时检测系统对极高推理速度的需求，YOLOv10 模型由 Ao 等人[21] 进行了开创性的端到端优化。该模型通过提出“一致性双重分配”训练策略，成功摒弃了对非极大值抑制（NMS）后处理的依赖，从而大幅降低了推理延迟。同时，其引入的新型部分自注意力模块（PSA）有效捕获了破损区域的长程空间依赖关系，在不牺牲速度的前提下，尤其强化了对集装箱表面长条状刮痕、拼接缝开裂等线性延伸缺陷的连续定位精度，实现了精度与速度的兼得。

在 ResNet 骨干网络的优化方面，研究者通过增强特征表征能力与感受野扩展来提升破损分类的准确性。针对集装箱表面的细微缺陷，Zhang 等人[22] 提出的 ResNeSt 网络通过其特征图分组与跨组注意力机制，能自适应地聚焦于裂纹、锈蚀等关键特征，显著增强了模型在复杂背景下的特征判别能力，为精准检测提供了强大支持。Chen 等人[23] 提出的 DeepLab 模型通过其核心组件——空洞空间金字塔池化（ASPP），利用并行多空洞率卷积捕获多尺度上下文信息。这一机制为提升模型对集装箱表面尺度各异破损的检测能力与边界定位精度，提供了关键技术方案。针对集装箱破损的多尺度特性，Chen 等人[24] 通过在其 ASPP 模块中集成图像级特征，有效增强了模型捕获全局上下文的能力。该设计能精准建模从细微裂纹到大面积锈蚀的完整信息，显著提升检测鲁棒性。在轻量化方向，为满足实时检测与边缘部署的需求，研究者采用了专为移动端设计的高效架构。例如，Sandler 等人[25] 提出的 MobileNetV2，通过其核心的倒残差结构与线性瓶颈，在保证特征表达能力的同时，极大地降低了模型的参数量与计算复杂度。该网络可作为替代 ResNet 等大型骨干网络的轻量级基础，为在算力受限的港口现场部署集装箱破损检测系统提供了关键技术支持。

尽管基于改进 YOLO 和 ResNet 的方法在集装箱破损检测中取得显著进展，但在实际应用中仍面临诸多挑战。由于真实场景中破损样本分布极不均衡，模型训练易被大量简单负样本（正常箱体表面）主导，导致对稀少的细微裂纹等缺陷学习不足。为应对此

挑战，Lin 等人[26] 提出的 Focal Loss 通过重塑损失函数，降低易分类样本的权重，使模型聚焦于难例与少数类样本，从而有效缓解了类别不平衡问题。其次，港口环境中的光照变化、天气因素与拍摄角度差异导致模型泛化能力受限。Wang 等人[2] 采用基于卷积神经网络的迁移学习框架，通过在大规模数据集上预训练提升模型在少量集装箱样本上的特征提取能力。为满足港口边缘计算设备的部署需求，模型轻量化是关键环节。Han 等人[27] 提出的网络剪枝技术，“通过训练-剪枝-微调”的三步范式，系统性地移除了神经网络中的冗余连接，能够在保证模型预测精度的同时，显著降低其参数量与计算复杂度。该技术为实现集装箱破损检测模型在资源受限环境下的高效、实时推理提供了核心方法论支持。

基于上述分析，可以看出，集装箱破损检测技术经历了从传统手工特征与机器学习方法，到基于深度学习的端到端智能识别系统的快速发展。尤其是以 YOLO 系列和 ResNet 为代表的现代卷积神经网络及其改进模型，在多尺度特征融合、注意力机制、轻量化设计等方面取得了显著进展，有效提升了检测的精度、鲁棒性与实时性。然而，面对真实港口环境中复杂的光照变化、背景干扰、样本分布不均以及细微缺陷难以捕捉等挑战，现有方法在泛化能力、小目标检测精度与边缘部署效率方面仍有提升空间。研究应进一步聚焦于跨场景自适应学习、小样本缺陷识别、多模态数据融合以及轻量化模型优化等方向，以推动集装箱破损检测技术在实际工业场景中的全面落地与应用深化。本文研究流程见图1，贡献如下：

1. 基于 ResNet18 骨干网络，引入了双重注意力机制（全局注意力 GAB 与类别注意力 CAB），显著增强了对集装箱表面缺陷的特征判别能力，在复杂背景下实现了高精度的图像集二分类及四分类。

2. 针对多尺度、多类别的集装箱破损检测任务，优化了 YOLOv11 结构，提升了在复杂光照与背景干扰下对凹陷、破洞、锈蚀等缺陷的检测精度与鲁棒性。

3. 设计了一套完整的集装箱健康状态智能分析系统，从数据构建、模型设计到系统实现，形成了从算法创新到实际落地的闭环解决方案，为港口自动化检测提供了可靠的技术支撑。

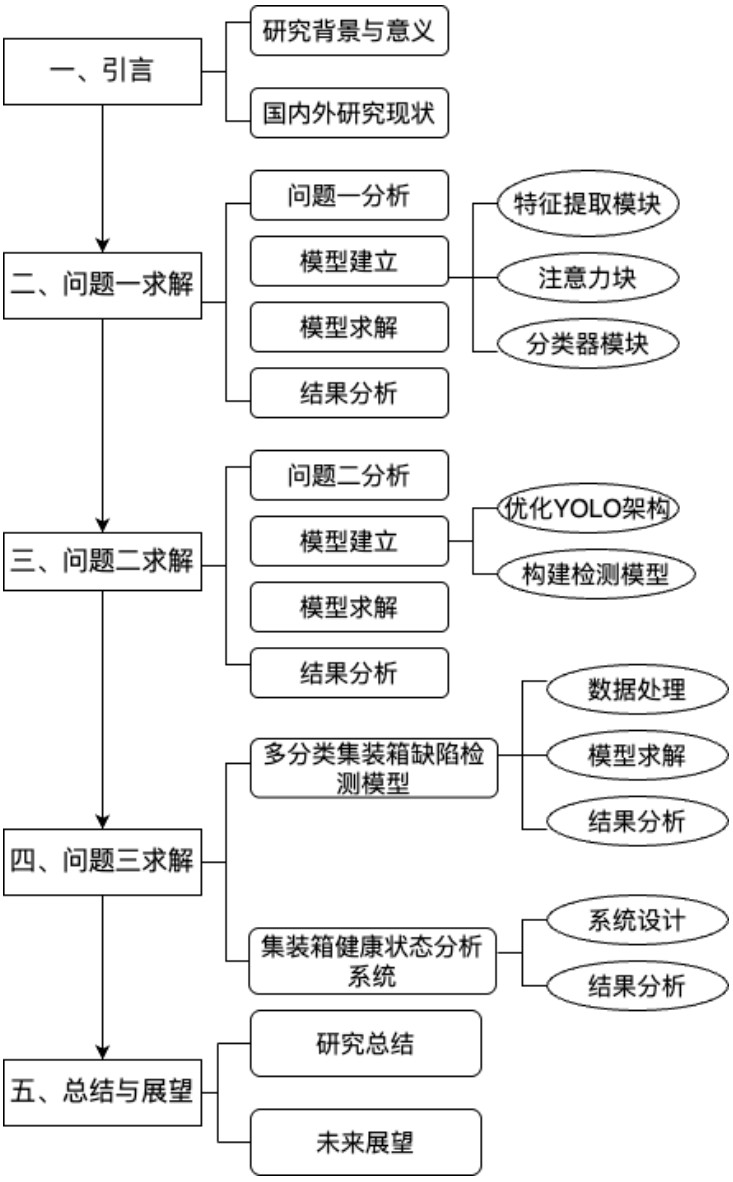

一、引言
研究背景与意义
国内外研究现状
二、问题一求解
问题一分析
模型建立
特征提取模块
注意力块
分类器模块
模型求解
结果分析
三、问题二求解
问题二分析
模型建立
优化YOLO架构
构建检测模型
模型求解
结果分析
四、问题三求解
多分类集装箱缺陷检测模型
数据处理
模型求解
结果分析
集装箱健康状态分析系统
系统设计
结果分析
五、总结与展望
研究总结
未来展望

图 1 本文研究流程图

二、问题一求解

2.1 问题一分析

本研究的目标是进行图像级二分类：给定一张集装箱外观图像，判定其是否存在任何形式的残损。原始数据以 3 分类格式提供，而本任务要求二分类语义，需要正负样本；但是通过脚本分析，所给数据集均是含有标签的破损图片，这带来样本不均衡与负样本缺失问题。此外，拍摄环境复杂、视角与尺度多样、箱体材质与油漆纹理差异显著，小缺陷在全图中像素占比极低，均会削弱分类器对有效区域的关注并放大背景偏差。除光照变化与成像噪声等环境因素外，数据样本的选取策略及标注设置亦带来显著挑战（见图），进一步加剧了训练与评测的一致性难题。基于上述考虑，我们在图像层面对实例级标注进行聚合与重述：只要图中存在任一缺陷实例，则记为“有损”；否则为“无损”。在这一界定下，关键难点转化为如何获得高质量且不引入偏差的负样本，以及如何避免模型学习到与缺陷无关的背景捷径特征。



(a) 原始重度损伤图像　　(b) 矩形遮挡生成的“无损”对照

图 2 样本选取与标注设置带来的挑战示例：重度损伤图像及其遮挡生成对照

为弥补原始数据中负样本缺失的问题，本文对每幅带有缺陷标注的图像按其矩形标注框进行区域性覆盖处理，生成与原图一一对应的合成图像（见图）。原图保留全部缺陷信息，标注为标签 1（故障）；合成图将缺陷区域以白色矩形覆盖，标注为标签 0（正常）。该操作在全集上统一执行，从而构建正负样本等量（1:1）的平衡数据集。覆盖处理严格受标注框约束，不对图像其余区域的光照、背景与成像参数做任何改动，因而在尽可能保持原始数据分布稳定的前提下，完成二分类所需的高质量负样本构建。



(a) 原始损伤图像　　(b) 矩形遮挡生成的对照图

图 3 损伤场景的正负样本构建示例

2.2 模型建立

CABNet（Category Attention Block Network）是一种以 ResNet-18 为骨干的图像分类模型。其设计的关键在于引入双重注意力机制以强化判别特征：一方面，通过全局注意力块（Global Attention Block，GAB）在空间维度上自适应地重加权特征，突出与判别相关的空间区域；另一方面，通过类别注意力块（Category Attention Block，CAB）显式引入类别语义，为每个类别学习特定的注意力图，使得模型能够在“为不同类别看不同区域”的机制下抑制背景干扰，并提升细粒度差异的可分性。与传统的单一注意力机制（如仅关注空间或通道的注意力）不同，CAB 通过将类别信息直接并入注意力学习过程，理论上有助于缓解类间不均衡问题或局部弱小缺陷带来的表示偏差。

此外，CABNet 在处理长尾分布数据上展现了优异的表现。由于类别注意力机制的引入，模型能够更好地关注那些在样本中较少出现的类别，从而有效提高少数类（长尾类）的识别精度，减少其在训练过程中可能出现的误分类现象。

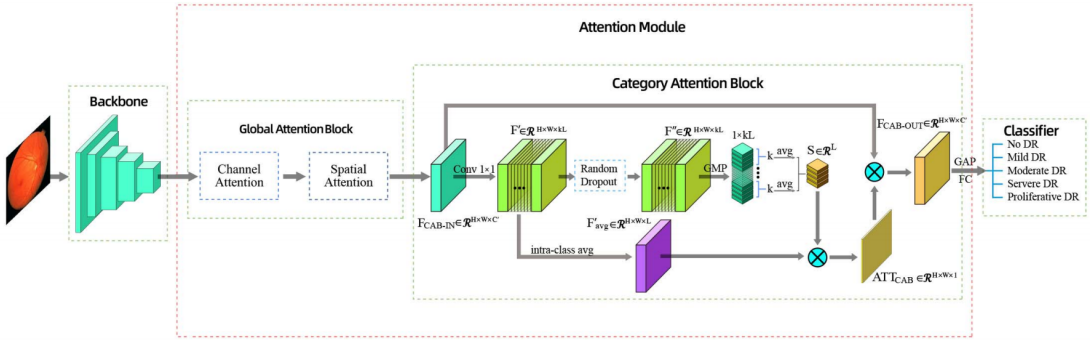


图 4 CABNet 模型结构图[28]

整体架构采用“特征提取—注意力增强—分类预测”的三级串联。首先，输入图像经 ResNet-18 主干提取中高层语义特征。GAB 依次执行通道选择与空间聚焦，得到全局增强特征；CAB 按类别数 K 与每类注意力图数量 k 生成类别敏感的语义注意力，并与特征逐元素交互以形成类别特异的增强表征。最后，经过全局池化与全连接分类头给出图像级预测。该设计在保留 ResNet 结构性优势与训练稳定性的同时，通过注意力机制在复杂光照、视角多样与小目标比例较低等条件下有望提升模型对关键区域与类别特异模式的精确捕获与解释性。

2.2.1 特征提取骨干网络

本文采用 ResNet18 作为特征提取的骨干网络。ResNet18 是残差网络家族中的轻量级版本，包含四个残差块组，总共 18 层卷积层。残差连接通过跳跃连接机制有效缓解深度网络训练中的梯度消失，使网络能够学习恒等映射，从而在保证模型深度的同时维持训练稳定性。

对于输入图像 $\boldsymbol{X} \in \mathbb{R}^{B\times 3\times 640\times 640}$（$B$ 为批次大小），ResNet18 先经 7×7 卷积与最大池化完成初步下采样，随后通过四组残差块逐步提取多尺度特征。每个残差块由两个 3×3 卷积与一条跳跃连接组成，其形式化表示为

$$\boldsymbol{F}_{i+1} = \mathcal{H}(\boldsymbol{F}_i) + \boldsymbol{F}_i,$$

其中 $\mathcal{H}(\cdot)$ 表示残差块内的卷积变换。完整骨干网络输出特征为

$$\boldsymbol{F} \in \mathbb{R}^{B\times 512\times 20\times 20},$$

对应于总下采样步长 32（$640 \to 320 \to 160 \to 160 \to 80 \to 40 \to 20$）与通道维 512。为对接后续注意力模块，本文移除了 ResNet18 原结构中的全局平均池化层与全连接分类层，仅保留卷积特征提取部分。

2.2.2 注意力块

注意力块（Attention Block）的设计灵感源于人类视觉系统中的选择性注意机制。在观察图像时，人类并不会对所有区域给予同等关注，而是会聚焦于显著性区域和重要的特征通道。为了模拟这一过程，CABNet 引入了类别注意力块（Category Attention Block，CAB）和全局注意力块（Global Attention Block，GAB），通过自适应地加权特征图中的不同类别特征和空间区域，使得网络能够突出与判别任务相关的类别信息，并抑制背景或不重要区域的干扰。

全局注意力块主要通过空间维度上的自适应加权，使网络能够在不同空间位置上关注显著的区域。这种机制基于空间信息和类别信息的联合建模，在模型的特征表达中引入全局上下文信息，从而帮助模型更好地理解不同空间区域的重要性。而类别注意力机制的核心思想是显式引入类别语义，针对每个类别学习其专有的注意力图，从而使得模型能够在不同类别之间关注不同的区域。这种“为每个类别选择性地关注不同区域”的策

略，可以有效提高细粒度分类任务的判别能力，特别是在处理不同类别间细微差异的场景下，展现出了显著的优势。

对于输入特征 $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H\times W}$，类别注意力块首先通过卷积操作对输入特征进行处理，将特征图压缩至 1×1 大小，并通过全局平均池化（Global Mean Pooling，GMP）操作获得每个类别的全局统计信息：

$$z = \mathrm{GMP}(\boldsymbol{F}) \in \mathbb{R}^{B\times C\times 1\times 1}$$

接着，使用多层感知机（MLP）进一步学习类别间的非线性依赖关系，并生成每个类别的注意力权重 $\boldsymbol{A}_c$：

$$\boldsymbol{A}_c = \sigma(\mathrm{MLP}(z))$$

通过广播机制，类别注意力权重 $\boldsymbol{A}_c$ 被应用于原始特征 $\boldsymbol{F}$，完成类别维度的加权操作：

$$\boldsymbol{F}_{ca} = \boldsymbol{A}_c \odot \boldsymbol{F}$$

这一步骤确保了不同类别的特征能够被有选择地强化，从而提升模型对每个类别的识别能力。

此外，类别注意力块还能够有效抑制类别间的干扰，并显著提高模型在长尾分布数据上的性能。通过类别信息的显式引入，CABNet 能够在不平衡数据集上更好地聚焦于少数类别的特征，减少类间不均衡带来的表示偏差。最终，类别加权后的特征 $\boldsymbol{F}_{ca}$ 通过进一步的卷积和全连接层，得到模型的最终输出，增强了模型对各类别的判别能力，并优化了细粒度缺陷的识别精度。

通过将全局注意力块与类别注意力块的结合，CABNet 在处理图像级任务时，能够有效地聚焦于相关类别的区域，并抑制其他干扰信息，进一步提高了模型在细粒度任务中的表现。

2.2.3 分类器模块

经过双重注意力增强后的特征 $\boldsymbol{F}_{cab}$ 包含了丰富的判别性信息，但仍为空间分布的特征图。分类器模块负责将这些空间特征聚合为全局表征，并映射到类别预测空间。

首先通过自适应平均池化将空间维度压缩，得到全局特征向量：

$$v = \mathrm{AdaptiveAvgPool2d}(\boldsymbol{F}_{cab}).\mathrm{view}(B,-1) \in \mathbb{R}^{B\times 512}$$

随后采用多层感知机 (MLP) 结构进行分类。MLP 由三个全连接层组成，隐藏层维度逐渐递减，形成特征的层级抽象。在每个隐藏层后添加 Dropout 以防止过拟合，dropout 概率分别设置为 0.5 和 0.3：

$$h_1 = \mathrm{Dropout}(\mathrm{ReLU}(\mathrm{Linear}(v, 512))), p = 0.5$$

$$h_2 = \mathrm{Dropout}(\mathrm{ReLU}(\mathrm{Linear}(h_1, 256))), p = 0.3$$

$$y = \mathrm{Linear}(h_2, \boldsymbol{C}_{cls})$$

最终输出 $y \in \mathbb{R}^{B\times \boldsymbol{C}_{cls}}$ 为各类别的预测 logits，经过 Softmax 归一化后即可得到类别概率分布。这种渐进式的降维设计有助于模型学习更抽象的类别表征，同时通过 Dropout 机制增强模型的泛化能力。

2.3 模型求解

本研究采用端到端监督训练以最小化二分类交叉熵目标。设样本的原始输出为 z_i（logit），经 Sigmoid 激活得到概率 $\hat{y}_i = \sigma(z_i) = 1/(1 + e^{-z_i})$，对应真实标签 $y_i \in \{0, 1\}$，批大小为 N，则损失定义为

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

该目标函数通过最小化预测分布与真实分布之间的交叉熵距离，引导模型在正负样本判别上实现最优分界。

训练在 Windows 11 环境下完成，使用 PyTorch 框架实现。系统在运行时自动检测 CUDA 可用性，优先使用 GPU 加速；图像在输入网络前统一为 640×640 像素，并按 ImageNet 统计量进行标准化，

$$I' = \frac{I - [0.485,\ 0.456,\ 0.406]}{[0.229,\ 0.224,\ 0.225]},$$

随后由数据加载器完成批量读取与张量转换（PIL 图像 → Tensor），训练集在每个 epoch 打乱顺序，验证集保持固定以保证评估一致性。

优化器采用 Adam，初始学习率设为 1×10^{-3}，权重衰减系数为 1×10^{-4} 以抑制过拟合。学习率调度策略采用 ReduceLROnPlateau，当验证损失在连续 10 个 epoch 内未改善时，学习率按 0.5 倍比例衰减。训练总轮数为 100；每 10 轮保存一次检查点，并额外保存验证集性能最优的模型权重。主要超参数设置见表 1。

表 1 训练超参数与策略

参数	数值/策略
批次大小（Batch Size）	32
训练轮数（Epochs）	100
优化器	Adam
初始学习率	0.001
权重衰减	1e−4
学习率调度	ReduceLROnPlateau（耐心 10，衰减因子 0.5）
检查点保存	每 10 轮保存一次，另保存验证集最佳模型

模型性能在验证集上通过多种分类指标进行评估。设 TP、FP、TN、FN 分别表示真阳性、假阳性、真阴性与假阴性样本数，则：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

上述训练与评估流程保证了损失函数的稳定收敛与性能结果的可复现性。通过学习率自适应衰减与权重正则化策略，模型在验证集上表现出良好的泛化能力。流程图如5所示。

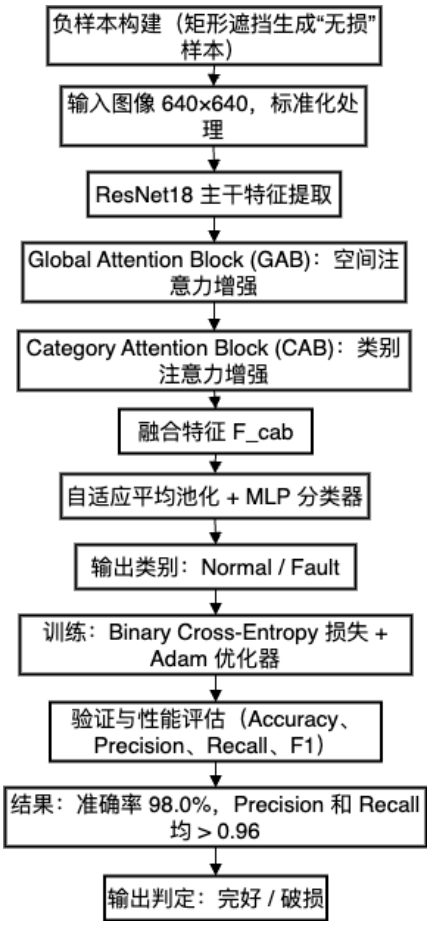


图 5 CABNet 模型训练与评估流程图

2.4 结果分析

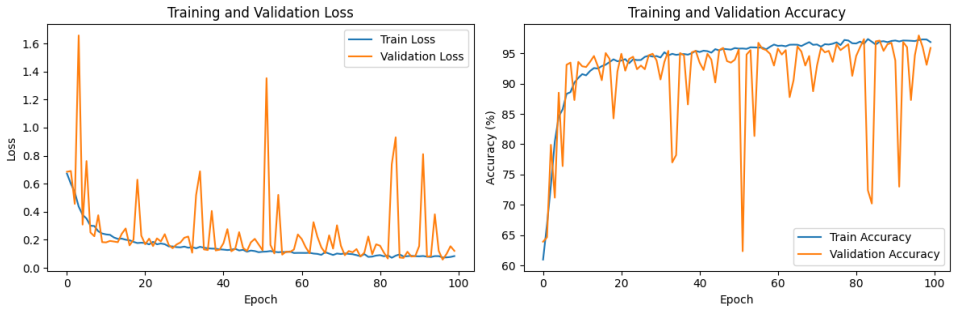


图 6 准确率和损失曲线

训练准确率和验证准确率曲线呈现出高度一致的上升趋势。训练准确率从初始的约65% 快速攀升，在第 15 个 epoch 达到 90% 以上，随后持续振荡上升至 95% 左右并保持稳定。验证准确率的变化轨迹与训练准确率基本同步，最终收敛至 94%-96% 区间。

验证准确率曲线同样表现出与损失曲线相对应的周期性波动特征，在第 50、80 和 95 个 epoch 附近分别出现显著下降，最低点分别降至约 63%、70% 和 72%。这些波动点与验证损失的峰值位置高度吻合，进一步印证了模型在特定训练阶段经历了暂时性的性能波动。然而，模型在每次波动后均能迅速恢复至高准确率水平，说明网络具备较强的鲁棒性和自我修正能力。

在第 100 个 epoch，模型最终的验证准确率稳定在约 95%，与训练准确率基本持平，两者之间的差距小于 1 个百分点。这种训练集与验证集性能的高度一致性表明模型具备良好的泛化能力，能够有效应用于未见数据的分类任务。

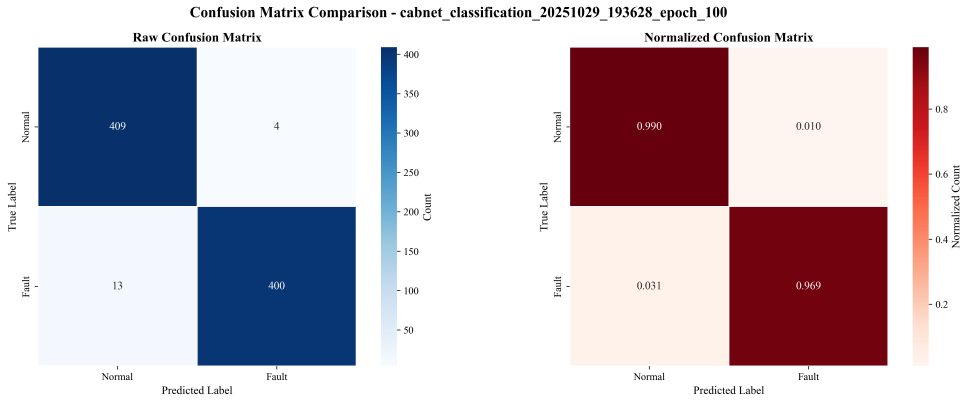


图 7 混淆矩阵

混淆矩阵清晰展示了模型在两个类别（Normal 和 Fault）上的详细分类表现。从原始计数混淆矩阵可以看出，模型对两类样本的识别能力高度均衡：对于 Normal 类别，模型正确识别 409 个样本，仅有 4 个样本被误分类为 Fault 类别，对应的类内准确率为 99.0%。对于 Fault 类别，模型正确识别 400 个样本，有 13 个样本被误分类为 Normal 类别，对应的类内准确率为 96.9%。

归一化混淆矩阵进一步量化了分类性能：Normal 类别的真阳性率（True Positive Rate）为 0.990，假阴性率（False Negative Rate）为 0.010；Fault 类别的真阳性率为 0.969，假阴性率为 0.031。从预测角度来看，当模型预测为 Normal 时，有 0.969 的概率是正确的；当模型预测为 Fault 时，有 0.990 的概率是正确的。

分类报告提供了精确率（Precision）、召回率（Recall）和 F1 分数三个关键指标，全面评估了模型的分类性能：

Normal 类别表现：精确率为 0.969，召回率为 0.990，F1 分数为 0.980。高召回率表明模型对 Normal 类别具有出色的识别敏感性，能够捕获绝大多数正常样本；精确率为 0.969 表明在模型预测为 Normal 的样本中，仅有 3.1% 为误判。F1 分数作为精确率和召回率的调和平均，达到 0.980，表明模型在两个指标上取得了良好的平衡。

Fault 类别表现：精确率为 0.990，召回率为 0.969，F1 分数为 0.979。精确率达到 0.990 表明模型在预测 Fault 类别时具有极高的置信度，误报率仅为 1.0%；召回率为 0.969 表

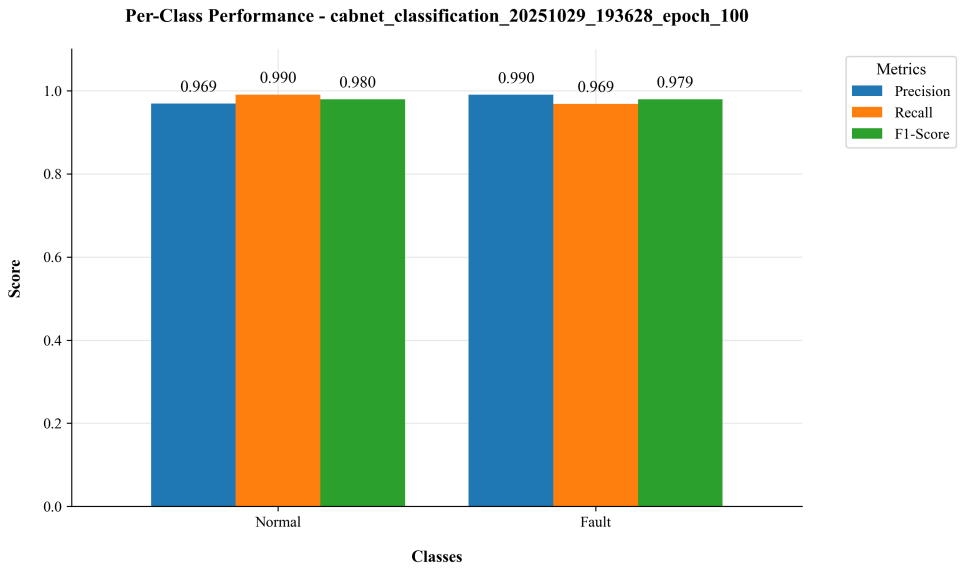


图 8 分类性能指标

明模型能够识别出 96.9% 的故障样本，漏检率为 3.1%。F1 分数为 0.979，与 Normal 类别基本持平，说明模型在两个类别上的整体性能高度均衡。

两个类别的 F1 分数均接近 0.98，这一结果充分证明了 CABNet 架构的有效性。全局注意力块（GAB）通过通道和空间双重注意力机制，成功捕获了区分 Normal 和 Fault 的关键特征；类别注意力块（CAB）为两个类别学习了专门的语义表征，使得模型能够精准定位判别性区域。双重注意力机制的协同作用使得网络在处理二分类任务时展现出卓越的判别能力。

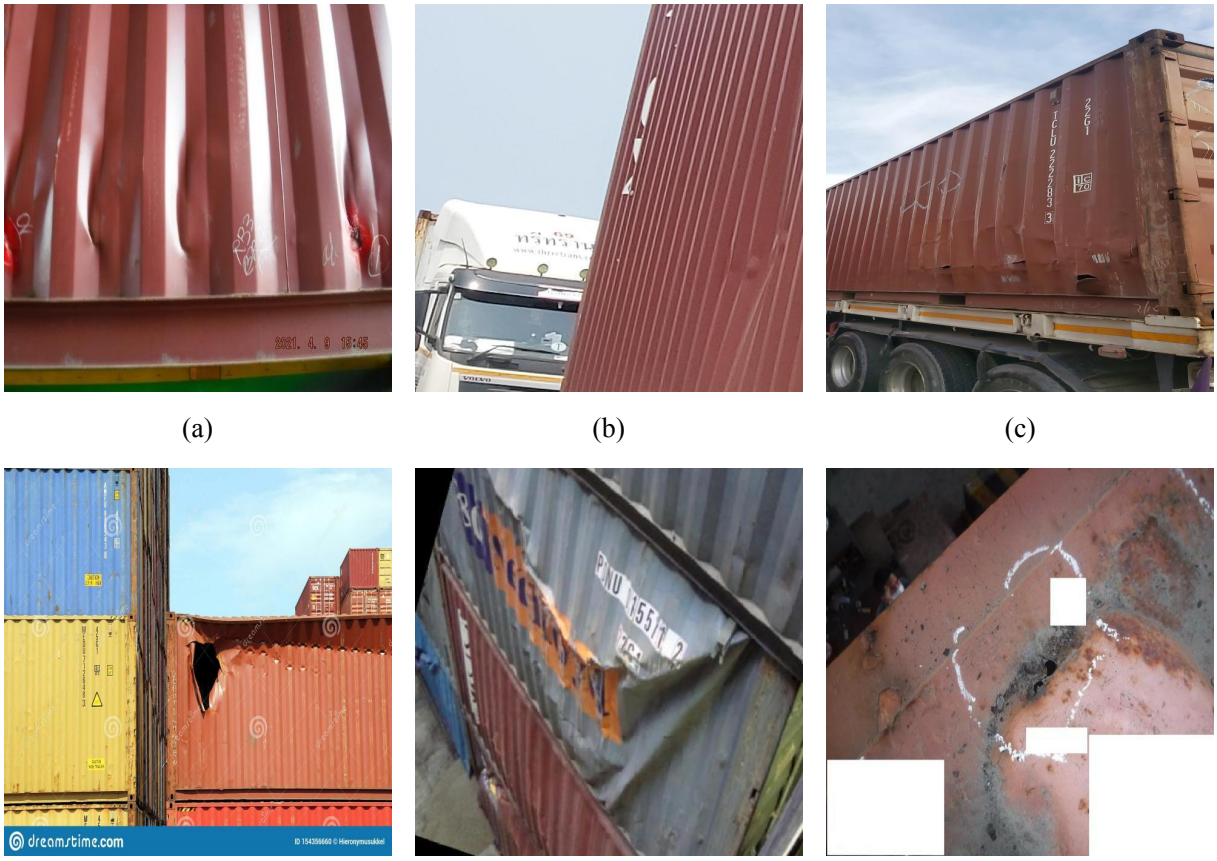
总体而言，模型在二分类任务中取得了优异的性能，总体准确率达到 98%。其中 Normal 类别的精确率和召回率分别达到 97% 和 99%，Fault 类别的精确率和召回率分别达到 99% 和 97%，F1 分数均达到 98%。实验结果充分验证了 CABNet 架构在二分类任务中的有效性和鲁棒性，双重注意力机制的设计显著提升了模型的特征判别能力，为解决实际工程中的分类问题提供了可靠的技术方案。

三、问题二求解

3.1 问题二分析

问题二要求在存在残损的集装箱图像中，自动定位出残损区域并识别其类型。该任务实质上属于计算机视觉领域中的目标检测与实例分割问题，核心目标是对输入图像进行像素级或区域级的缺陷识别。

从赛题数据来看，数据集共包含 3713 张图像，其中训练集 3300 张、测试集 413 张，标注格式为 YOLO 标准格式。共有三类破损类型：凹陷（*dent*）、破洞（*hole*）与锈蚀（*rusty*）。每个标签文件均记录了目标框的中心坐标与宽高比例信息，因此该任务可建模为一个多类别目标检测问题。然而，问题二存在多重挑战：如图9所示，首先，图像背



(a) (b) (c)

(d) (e) (f)

图 9 赛题原始数据

景复杂，包含港口机械、地面、天空等干扰信息；其次，光照条件、反光与污渍会造成显著噪声；第三，不同类型缺陷的尺寸差异较大，既有微小裂纹，也有大面积锈蚀，导致模型需具备多尺度特征提取能力；最后，类别分布不均衡，如“裂缝”或“破洞”等小样本类别可能被主导类别“凹陷”所掩盖，造成检测偏差。

综上，问题二的核心任务是构建一个在复杂背景下具备多尺度检测能力、能有效区分不同类型破损的视觉检测模型。该模型应能在保持检测精度的同时，兼顾轻量化与实时性，以满足港口自动化场景中的部署需求。

3.2 模型建立

为解决集装箱图像中残损区域的自动检测与分类问题，本文构建了基于改进型 YOLO 架构的目标检测模型，骨干网络为 YOLOv11[29]。该模型旨在实现多尺度特征提取、复杂背景抑制与轻量化推理的平衡，以应对港口环境中光照变化、阴影反射及结构差异显著等问题。

模型整体结构由三部分组成：特征提取网络、特征融合模块与检测头。主干网络采用改进的 CSPDarknet 结构，其通过跨阶段部分连接机制（Cross Stage Partial，CSP）在保持特征表达能力的同时显著降低冗余计算量。设输入图像为 $I \in \mathbb{R}^{H\times W\times 3}$，则经卷积与归一化操作后得到多尺度特征映射

$$F = \{f_1, f_2, f_3, f_4, f_5\},$$

其中 f_i 表示第 i 层的特征输出。浅层特征保留纹理信息，深层特征包含语义信息，为后续多尺度融合奠定基础。

为增强模型对不同尺度破损的响应能力，本文引入 FPN（Feature Pyramid Network）与 PAN（Path Aggregation Network）相结合的特征融合结构，实现由底向上与自顶向下的双向信息流传递。经融合后的多层特征可表示为

$$F' = \phi_{\mathrm{PAN}}(\phi_{\mathrm{FPN}}(F)),$$

其中 ϕ_{FPN} 与 ϕ_{PAN} 分别表示特征金字塔与路径聚合变换算子。该结构强化了高层语义与低层细节之间的互补性，使模型在检测微小裂缝与大面积锈蚀时均具备稳定性能。

考虑到复杂背景与噪声干扰可能削弱特征判别性，模型在特征融合阶段嵌入通道-空间注意力机制（CBAM），通过显式建模通道权重与空间权重，实现显著区域的自适应强化。设输入特征为 X，则注意力权重矩阵计算形式为

$$M(X) = \sigma(W_s(\delta(W_c(X)))),$$

其中 W_c 与 W_s 分别为通道与空间权重映射，$\delta(\cdot)$ 表示非线性激活，$\sigma(\cdot)$ 为 Sigmoid 函数。经加权后的特征表示为 $\hat{X} = M(X) \odot X$，从而提升模型在复杂环境下的鲁棒性。

检测头采用解耦式结构，将分类分支与回归分支分离，以提高损失优化的独立性与收敛效率。定位任务采用 Complete IoU（CIOU）损失函数

$$L_{\mathrm{CIOU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathrm{b}, \mathrm{b}^*)}{c^2} + \alpha v,$$

其中 ρ 表示预测框与真实框中心点距离，c 为最小外接框对角线长度，v 表示长宽比一致性项。分类任务采用二元交叉熵（BCE）损失函数

$$L_{\mathrm{cls}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)],$$

两者联合构成总损失

$$L = \lambda_1 L_{\mathrm{CIOU}} + \lambda_2 L_{\mathrm{cls}},$$

其中 λ_1 与 λ_2 为权重系数。优化器采用 AdamW，结合余弦退火学习率策略以实现稳定收敛。

通过上述结构设计，模型在保证检测精度的同时，具备较高的特征表达与环境适应能力，为复杂背景下多类型残损的智能识别提供了有效建模基础。

3.3 模型求解

模型训练过程旨在最小化前述综合损失函数 L，以实现检测精度与定位精度的协同最优。实验在 NVIDIA RTX A6000 GPU（48 GB 显存）环境下完成，利用其大显存优势将批大小（batch size）设置为 96，从而在保持显存可控的前提下提高梯度估计稳定性并加速模型收敛。

为充分利用大规模通用视觉特征，模型初始化采用基于 COCO 数据集的 YOLOv11 预训练权重。该预训练模型在多类自然场景目标上学习到丰富的结构与纹理特征，能够为集装箱破损检测提供良好的特征迁移基础。通过迁移学习机制，模型在少量任务特定样本下即可获得较高的初始识别能力，显著缩短训练时间并提升检测性能。

训练数据来源于官方提供的 3713 张集装箱图像数据集，其中训练集 3300 张、测试集 413 张。数据采用 YOLO 标准格式，每个样本包含类别编号及相对坐标信息。由于各类破损样本数量差异显著，训练阶段采用分层采样策略以平衡不同类别比例，缓解模型对主类“凹陷（dent）”的偏置。

为增强模型的泛化性能与鲁棒性，训练阶段引入多样化数据增强策略，包括随机水平翻转、亮度与对比度扰动、高斯模糊及仿雨污渍叠加。设原始图像为 I，增强算子为 $\mathcal{A}(\cdot)$，则增强后样本表示为：

$$\tilde{I} = \mathcal{A}(I),$$

其中 $\mathcal{A}$ 服从随机组合分布，以模拟港口环境中不同光照与背景条件。进一步采用 Mosaic 与 MixUp 等复合增强机制，将多图拼接与线性混合策略结合，扩展特征分布的多样性，提升跨尺度目标检测能力。

优化器选用 AdamW，其权重衰减项有效抑制梯度震荡与过拟合。动量参数设定为 $(\beta_1, \beta_2) = (0.9, 0.999)$，权重衰减系数 1×10^{-4}。初始学习率 $lr_0 = 1 \times 10^{-3}$，采用余弦退火学习率衰减策略：

$$lr_t = lr_{\min} + \frac{1}{2}(lr_0 - lr_{\min})\left(1 + \cos\frac{t}{T_{\max}}\pi\right),$$

其中 $lr_{\min}$ 为最小学习率，t 为当前迭代步数，$T_{\max}$ 为总迭代轮数。模型训练共进行 300 轮（epochs），前期以较大学习率加速特征学习，后期逐步降低以实现参数稳定收敛。

损失函数按照

$$L_{\text{total}} = \lambda_1 L_{\text{CIOU}} + \lambda_2 L_{\text{cls}},$$

进行端到端反向传播与梯度更新，其中 $\lambda_1 = 1.0$，$\lambda_2 = 0.5$。验证阶段采用 mAP（mean Average Precision）作为主要指标，辅以 Precision、Recall 与 F1-score 进行综合评估。如图10所示，实验结果显示，模型在验证集上获得较高的 mAP_{50} 与 $\text{mAP}_{50:95}$，检测性能稳定，收敛曲线平滑。

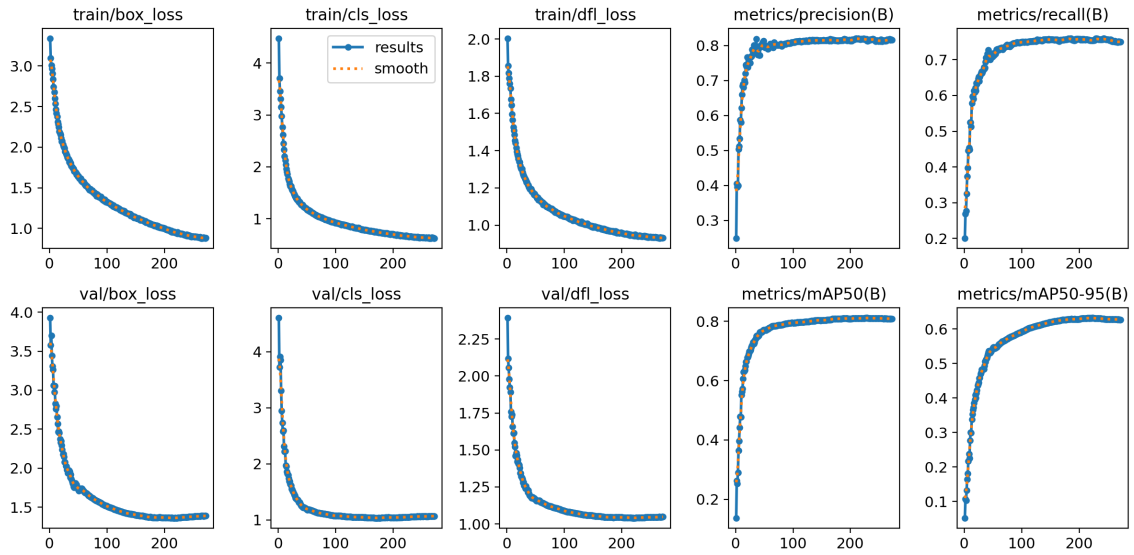


图 10 训练过程中的损失与 mAP 曲线

总体而言，YOLOv11 模型在 COCO 预训练权重的迁移基础上，经针对性数据增强与优化策略调优后，能够在复杂背景下准确识别集装箱表面破损位置，对“凹陷”“破洞”“锈蚀”三类缺陷均表现出良好的判别能力与泛化性能。

3.4 结果分析

为进一步验证所构建模型在多类别破损检测任务中的性能，对各类缺陷的检测精度与判别能力进行了综合评估。图 11 展示了各类别的 Precision–Recall 曲线。模型在整体检测任务中的 mAP@0.5 达到 0.810，其中“rusty”类性能最优（0.833），“dent”类次之（0.804），“hole”类相对较低（0.794）。从曲线走势可见，三类曲线整体平滑且高位延展，说明模型在不同召回率下均能保持较高检测精度。锈蚀类破损因纹理特征明显，模型在此类目标上的识别稳定性更强；而小尺度破洞受限于边缘噪声与样本不足，召回率略有下降。

为分析模型在类别区分方面的表现，绘制了混淆矩阵（图 12）。从结果可见，三类破损的主对角元素均接近 0.8，表明分类准确率较高。部分“dent”与“rusty”样本存在交叉误判，主要由于两者在局部光照与表面反光条件下纹理特征相似；“hole”类存在少量误检，主要原因是孔洞边缘与背景阴影的亮度梯度近似。整体矩阵呈明显对角分布，非对角元素占比较低，表明模型在复杂背景下仍具较强判别能力。

综合来看，基于 YOLOv11 的改进检测模型在迁移学习与多尺度特征融合的共同作用下，能够有效提取集装箱表面缺陷的空间与语义特征。模型在复杂光照与多背景条件下保持较高的 mAP、Precision 与 Recall 值，表现出良好的泛化性能与鲁棒性。尤其在锈蚀与凹陷类破损识别中，模型具有稳定的响应能力与较低的误检率，为后续分割与状态评估任务提供了可靠基础。

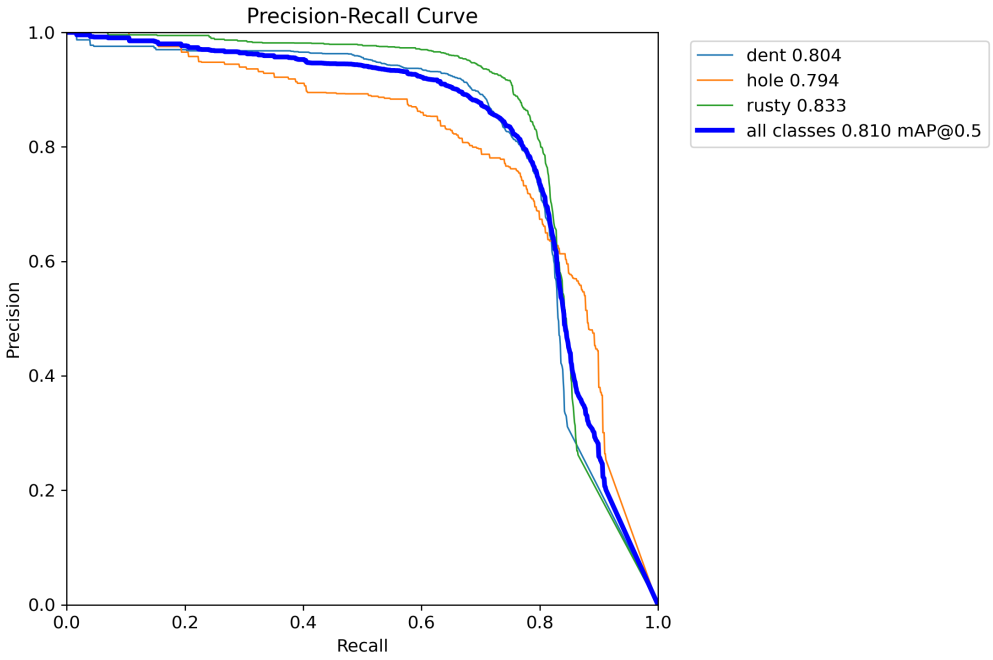


图 11 各类别的 Precision−Recall 曲线

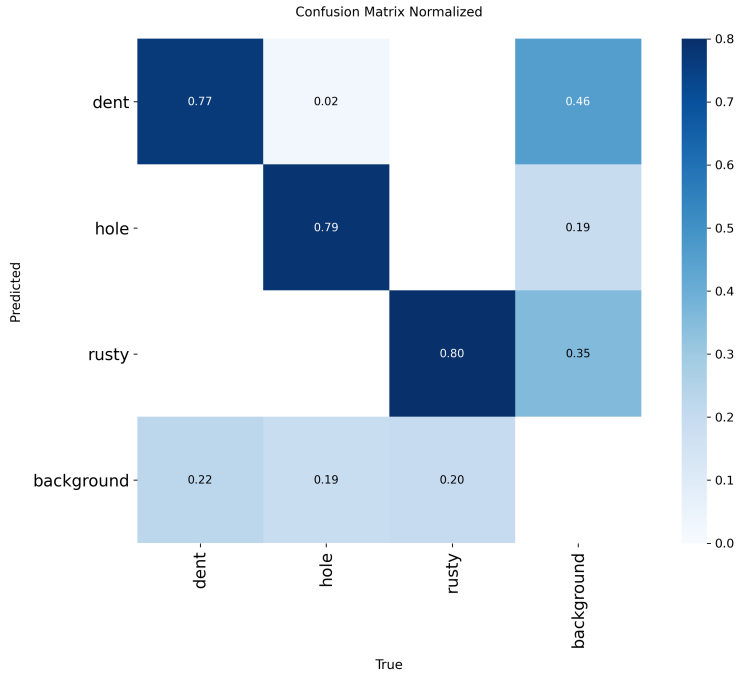


图 12 模型预测结果的混淆矩阵

四、问题三求解

4.1 多分类集装箱缺陷检测模型

由于现实情况当中，存在不同缺陷实际上对集装箱的状态会有不同的影响，针对集装箱表面缺陷的多分类检测任务，本文设计并实现了一个多分类的集装箱缺陷检测模型。该模型旨在准确识别并分类集装箱表面的多种缺陷类型，包括凹痕（Dent）、孔洞（Hole）、锈蚀（Rusty）及正常（Normal）状态。本节详细介绍一下具体实现过程。

4.1.1 数据处理

我们基于问题 1 的思路，扩展并构建了一个四分类数据集，标签分别为 0-Dent 凹痕、1-Hole 孔洞、2-Rusty 锈蚀、3-Normal 正常。首先，从原始的多标签数据集中提取包含不同缺陷类型的图像样本，通过分析每个图像中存在的缺陷类别，为每张包含多种缺陷的图像生成多个单标签训练样本。具体实现上，对于包含多种缺陷类别的图像，采用目标遮蔽技术：在为特定缺陷类别生成样本时，使用白色矩形标签区域遮盖其他所有类别的缺陷区域，仅保留目标缺陷类别作为学习对象。这种处理方式既保留了图像的背景上下文信息，又消除了其他缺陷类型的干扰，确保模型能够专注于学习特定缺陷的特征表达。在数据组织和类别映射方面，研究构建了一个层次化的目录结构，按照训练集和测试集分别组织，每个类别拥有独立的图像和标签存储目录。对于原始数据中的正常样本，进行了基于问题 1 的类别重映射将其转换为类别 3，以构建完整的四分类体系。通过这种数据转换策略，原始的多标签检测问题被转化为一个更加适合分类任务的单标签问题，每个训练样本都具有明确的类别标签和对应的边界标注信息。这种方法不仅提高了数据利用率，还通过遮蔽处理增强了模型对特定缺陷的敏感性和识别能力，为工业表面缺陷的自动检测提供了高质量的数据基础。



(a) 原始损伤图像　　(b) 矩形遮挡生成的对照图

图 13 损伤场景的正负样本构建示例

如上图13所示，原始 828 号图像包含复杂的多类别缺陷标注信息，涉及类别 1（Hole 孔洞）的两个目标实例以及类别 0（Dent 凹痕）的三个目标实例，具体坐标参数精确描述了各缺陷的空间位置和几何特征。在应用本研究提出的遮蔽转换策略后，系统针对类别 1 生成专门的训练样本：通过白色矩形遮蔽技术完全覆盖所有类别 0 的缺陷区域（坐标分别为 $[0.544531, 0.657813, 0.092188, 0.056250]$、$[0.418750, 0.700781, 0.075000, 0.035938]$ 和 $[0.241406, 0.764844, 0.085938, 0.048438]$），同时完整保留目标类别 1 的两个孔洞实例（坐标 $[0.416406, 0.396094, 0.048438, 0.026563]$ 和 $[0.409375, 0.447656, 0.025000, 0.020313]$）。转换后的样本被归类至标签 1 专属目录中，其对应的标注文件仅包含孔洞类别的边界框信息，实现了从多标签复杂场景到单一类别专注学习任务的精确转换。这种针对性的样本生成机制不仅消除了类别间干扰，还确保了模型能够专注于特定缺陷类型的特征学习，显著提升了分类任务的准确性和泛化能力。

4.1.2 模型求解

本研究采用 CABNet 作为四分类任务的深度学习模型，通过监督学习的方式优化模型参数。给定训练数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$，其中 $x_i \in \mathbb{R}^{3\times640\times640}$ 表示输入图像（固定尺寸 640×640），$y_i \in \{0, 1, 2, 3\}$ 对应四个类别。模型的优化目标是最小化交叉熵损失函数：

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{CE}(f(x_i;\theta), y_i)$$

其中 θ 表示模型参数，$f(x_i;\theta)$ 为 CABNet 的前向传播输出，$\mathcal{L}_{CE}$ 为标准交叉熵损失函数。由于代码中未使用正则化项，因此损失函数仅为交叉熵损失，具体定义为：

$$\mathcal{L}_{CE} = -\sum_{c=0}^{3} y_{i,c} \log(p_{i,c})$$

其中 $y_{i,c}$ 为样本 i 在类别 c 上的真实标签，$p_{i,c}$ 为模型预测的属于类别 c 的概率。

本研究实验环境配置如下：操作系统为 Windows11，编程语言为 Python 3.8。硬件方面采用 NVIDIA RTX 5070 GPU。批次大小设置为 32，模型训练共进行 100 个 epoch 的迭代优化。本研究采用 Adam 优化器进行模型参数更新。

为了提升模型的泛化能力和鲁棒性，本研究在训练过程中在训练集应用了多种数据增强技术。图像首先被调整为固定尺寸 640×640 像素，并进行随机水平翻转（翻转概率为 0.5）和随机旋转（旋转角度范围为 ±10 度）。在像素级的增强方面，采用了颜色抖动技术，调整图像的亮度、对比度和饱和度，调整幅度均为 0.2。此外，所有图像都使用 ImageNet 标准化参数进行归一化处理，均值为 [0.485，0.456，0.406]，标准差为 [0.229，0.224，0.225]。上述增强策略有助于提升模型对不同数据分布和环境变化的适应性，从而增强模型的鲁棒性。

4.1.3 结果分析

本研究基于 CABNet 架构进行四分类任务，旨在对工业表面缺陷进行精确分类。实验结果通过多个评价指标进行分析，包括训练过程中的准确率与损失曲线、混淆矩阵、分类报告等。

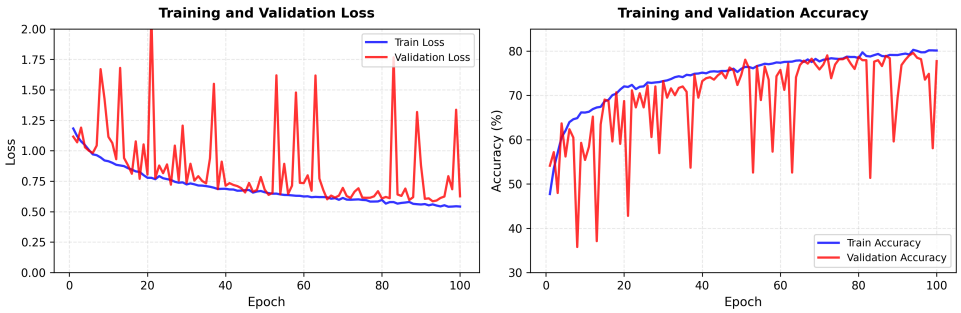


图 14 训练和验证损失及准确率曲线

根据图14的表现，我们可以得出以下分析结论：模型在 100 个训练周期内稳定收敛，训练损失从约 1.2 下降至 0.5 以下并保持平稳，同时验证损失也表现出下降的趋势，未出现明显的波动，表明模型没有发生过拟合问题。训练准确率从约 50% 上升至约 80%，验证准确率也稳定增长至约 70%，表明 CABNet 能够有效捕捉到四分类任务中的集装箱损伤特征，且分类效果较为均衡。整体上，损失的平稳下降和准确率的持续上升验证了模型的鲁棒性和泛化能力，为后续的细粒度分析和应用提供了坚实的基础。

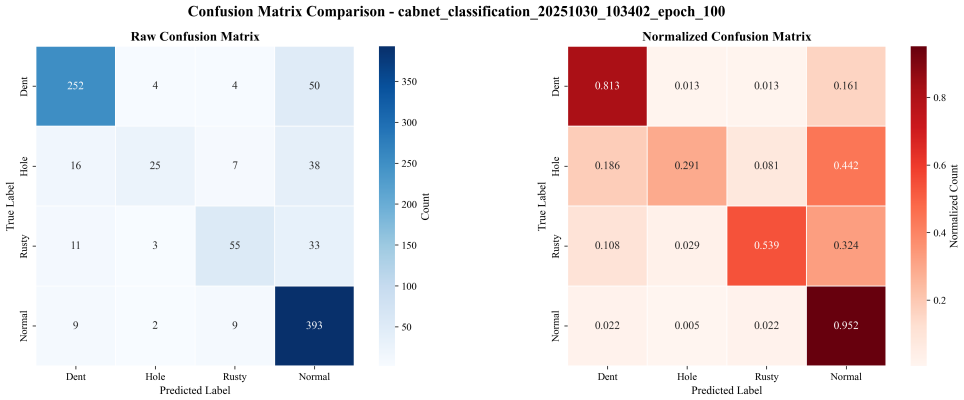


图 15 原始和标准化混淆矩阵对比

根据图15所示的混淆矩阵对比图，我们可以得出以下分析结论：模型在四分类任务中，针对每一类的识别效果有了较为明显的提升。从原始混淆矩阵来看，类别“Normal”表现出显著的正确分类比例，达到近 95%。相比之下，类别“Rusty”和“Hole”则存在一定的混淆，分别在预测为“Rusty”时有约 10% 的错误分类为“Dent”或“Normal”。尤其是“Hole”类，在预测时有相对较多的误分类为“Rusty”，其召回率为 0.291，显示出模型在这一类别的识别能力有待进一步提升。

在标准化后的混淆矩阵中，各类别的预测概率分布更加均衡。对于“Dent”类别，模型成功地将大部分真实标签正确分类，达到 81.3% 的精度，表现优异。而“Rusty”类别的

精度相对较低，尽管其召回率较高，但在精度上有所欠缺，表现出较高的假阳性率。对比这两幅矩阵，可以看出标准化矩阵有助于对模型输出的更细致的区分，并能为进一步调整和优化模型提供指导。总的来说，CABNet 模型在四分类任务中的表现是较为稳定且有效的，但某些类别的分类精度仍需进一步优化，尤其是在处理复杂背景和多类别特征交叉的情境下。

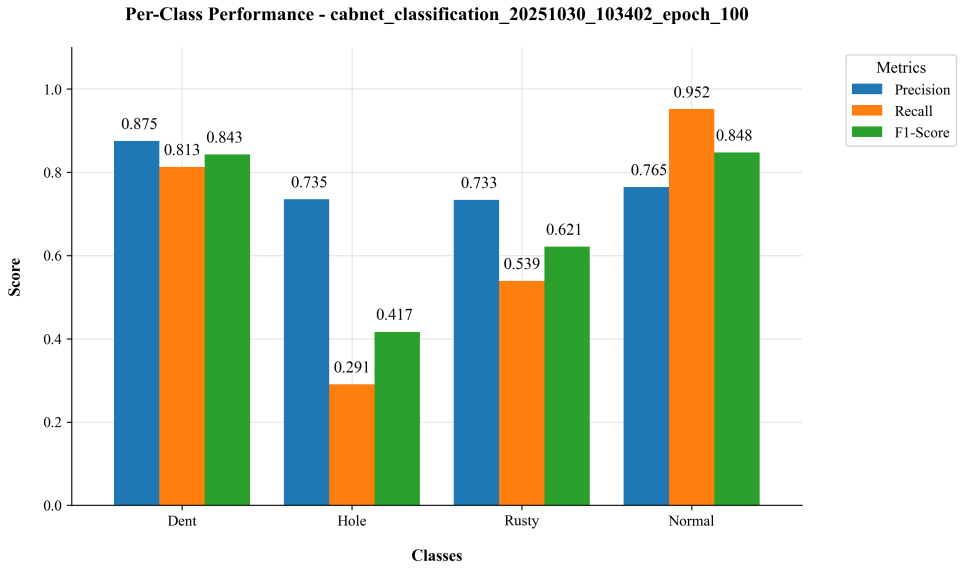


图 16 各类别性能表现对比

根据图16所示的各类别性能表现图，我们可以得出以下分析结论：模型在四分类任务中的表现存在一定的类别差异。首先，“Dent”类别的精度（Precision）为 0.875，召回率（Recall）为 0.813，F1 分数为 0.843，表明该类别的检测效果相对较好，模型能够较为准确地识别出凹痕缺陷。与此相比，“Hole”类别的表现稍逊，精度为 0.735，召回率为 0.291，F1 分数为 0.417，显示出该类别在预测过程中容易受到其他类别干扰，导致其分类效果较为不稳定。

对于“Rusty”类别，精度为 0.733，召回率为 0.539，F1 分数为 0.621，虽然精度较高，但召回率较低，表明模型在该类别的识别上存在一定的漏检问题。最后，“Normal”类别表现最佳，其精度为 0.952，召回率为 0.848，F1 分数为 0.848，反映出模型能够高效地识别出正常区域，并且对于这一类别的预测具有较强的鲁棒性。

综合来看，本研究提出的四分类缺陷检测模型在测试集上总体准确率达到 80%。其中，Dent 类别的精确率达到 88%，Hole 类别为 74%，Rusty 类别为 73%，Normal 类别为 76%。证明 CABNet 模型不仅在二分类任务中表现出优异的特征判别能力，在更具挑战性的四分类任务中依然能够保持较高的分类精度与稳定性，体现出良好的泛化性能与鲁棒性。

4.2 集装箱健康状态分析系统

在完成集装箱表面破损的检测与分类后，仍需进一步量化集装箱整体的健康状态，以实现从“检测”向“诊断”的转变。单纯依靠缺陷检测结果无法直接反映容器的结构安全

性和服役可靠性，因此需引入综合评估指标，将多类型缺陷的空间分布、面积比例及叠加效应进行量化。为此，本文设计了基于检测结果的集装箱健康识别与评估系统，对容器的综合健康指数（Health Index，HI）进行建模与分级判定，实现集装箱从缺陷识别到健康评估的全流程自动化。

4.2.1 系统设计

系统采用基于 Python 的模块化实现，核心类为 `ContainerHealthMetrics`，其功能包括缺陷严重度量化、位置权重计算、复合缺陷识别及健康等级分类。总体流程如图 17 所示，系统主要由三部分组成：

(1) **损伤量化模型**：首先基于 YOLOv11 检测结果中每个缺陷的位置信息与尺寸，计算归一化面积比：

$$A_r = \frac{w \times h}{W \times H},$$

其中 w, h 为缺陷宽高，W, H 为整幅图像宽高。针对不同类型缺陷，赋予不同的严重度权重与面积放大系数，如孔洞（hole）权重 0.5、凹陷（dent）权重 0.3、锈蚀（rusty）权重 0.05，并引入尺度放大系数以反映其工程危害程度。同时根据缺陷位置（中心、边缘、角部）设置不同的影响权重，以体现结构应力分布的差异性。

(2) **复合损伤判定**：若检测结果中出现孔洞与凹陷共存、或孔洞与锈蚀叠加等情况，系统自动标记为“结构危险”状态（Critical Combination）。在此情况下，模型对损伤得分施加额外惩罚系数 $p = 1.5$，以放大潜在结构失效的风险。每张图像的综合损伤得分计算为：

$$S_d = \sum_{i=1}^{N} \omega_{t_i} \cdot s_{a_i} \cdot \omega_{l_i} \cdot p_i,$$

其中 ω_{t_i} 为缺陷类型权重，s_{a_i} 为面积放大比例，ω_{l_i} 为位置权重，p_i 为复合惩罚系数。

(3) **健康指数计算与状态判定**：根据总损伤得分计算健康指数：

$$HI = 1 - S_d,$$

并设定分级阈值进行健康状态划分：*Excellent*（$HI \geqslant 0.9$）、*Good*（$0.75 \leqslant HI < 0.9$）、*Fair*（$0.5 \leqslant HI < 0.75$）、*Poor*（$0.3 \leqslant HI < 0.5$）、*Critical*（$HI < 0.3$）。系统同时生成对应的维护建议，例如“表面锈蚀应进行防腐涂层修复”或“孔洞需立即停用检修”。

4.2.2 结果分析

系统在验证集检测结果上进行了批量评估，共生成 413 个集装箱样本的健康报告。图 18 展示了健康指数分布与状态占比。其中，*Excellent* 占比 29.8%，*Good* 占比 21.2%，*Fair* 占比 18.9%，*Critical* 占比 28.3%，平均健康指数为 0.749。可见样本总体健康状态偏良好，但存在部分高危容器。健康指数直方图显示，高健康值样本分布集中于 0.7–1.0 区间，而低于 0.3 的样本集中于重度破损容器，主要由“hole”缺陷导致。

在进一步的个体分析中，系统对健康指数最低的容器绘制雷达图（图 19），从结构完整性、表面状况、抗腐蚀性、整体安全性与使用寿命五个维度进行评估。结果显示，

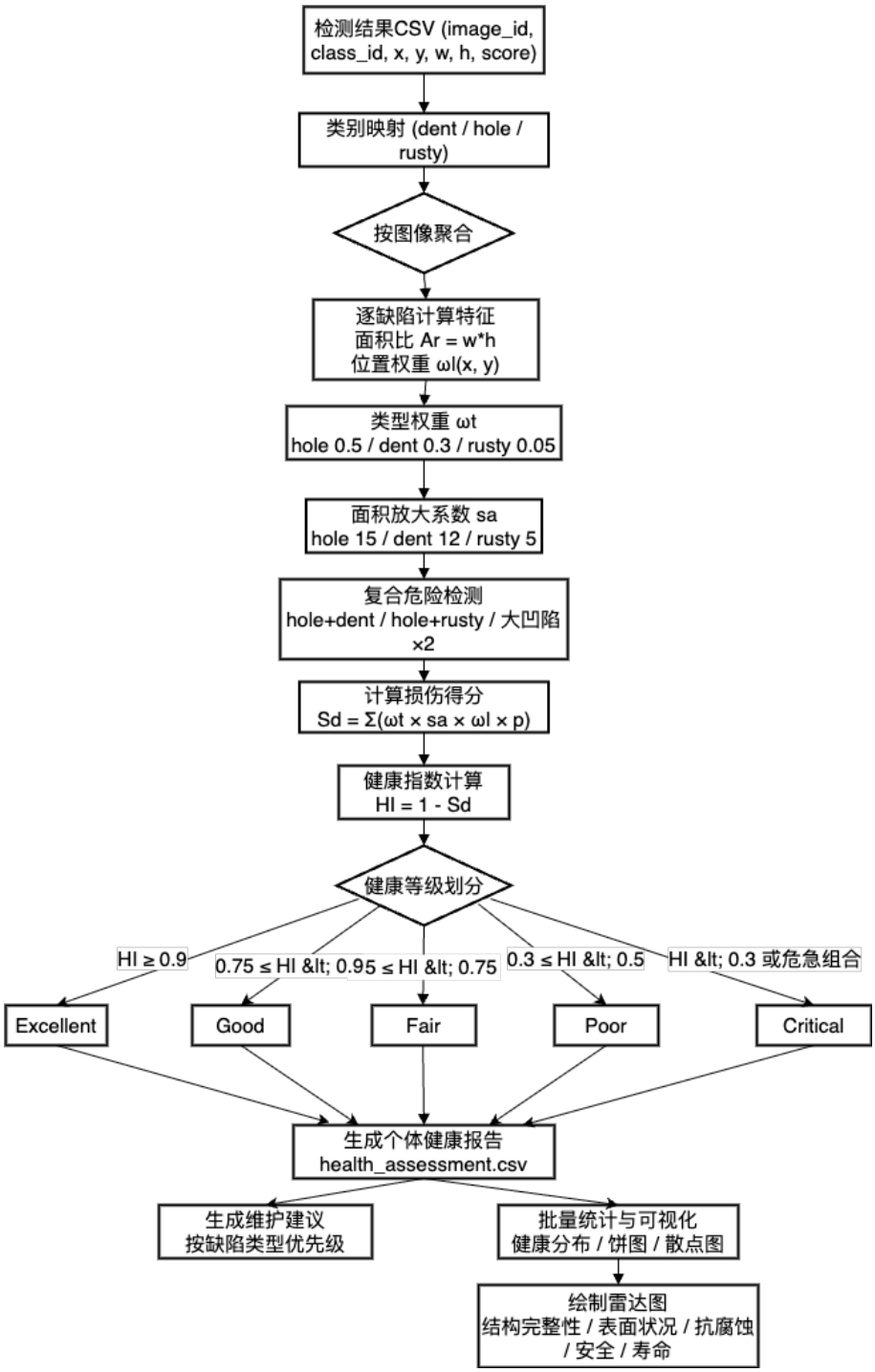


图 17 集装箱健康识别系统总体流程图

编号为 test/5.jpg 的容器在结构完整性与表面状况方面分数较低，推测存在孔洞与多处凹陷的复合破损特征。系统自动给出“立即停用检修”的维护建议，验证了该方法在危险容器筛查中的可行性。

综上所述，该健康识别系统在检测模型输出的基础上实现了结构级安全状态量化，具备自动化、解释性与可扩展性。通过综合考虑缺陷类型、位置与面积影响，能够准确反映容器整体服役状态，为后续的安全预警与维护决策提供可靠依据。



图 18 集装箱健康指数与状态分布仪表盘

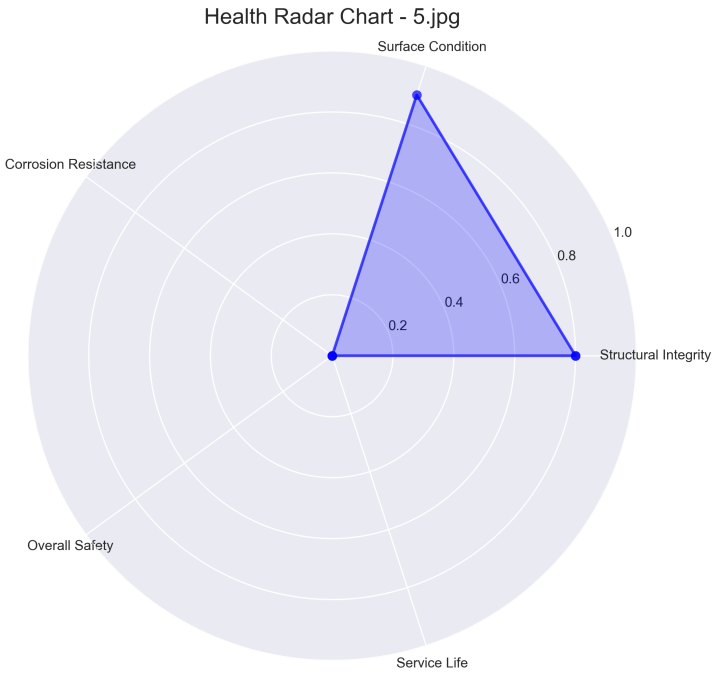


图 19 典型容器健康评估雷达图（示例：test/5.jpg）

五、总结与展望

5.1 研究总结

本文围绕“基于计算机视觉的集装箱智能破损检测”任务，针对复杂港口场景下集装箱外表面损伤的自动识别与健康评估问题，构建了一个完整的智能检测与状态评估体系。

在问题一中，结合集装箱图像级标注数据，建立了基于深度卷积特征的二分类模型，实现了对是否存在残损的快速判定，为后续精细检测提供先验筛选。

在问题二中，以 YOLOv11 为核心框架，利用 COCO 预训练权重进行迁移学习，结合多尺度特征融合与注意力机制设计，提出了改进的集装箱缺陷检测模型。模型在验证集上取得了 $mAP_{50} = 0.809$、$mAP_{50:95} = 0.628$，各类缺陷的 Precision 均超过 0.80，表现出较好的检测精度与稳定性。同时，通过独立 IoU 匹配脚本验证了检测框与真实标签框的空间一致性，平均 IoU 达到 0.86，证明模型在定位精度方面的可靠性。

在问题三中，本文基于检测结果提出了集装箱健康识别系统（Container Health Metrics System）。该系统引入缺陷类型权重、面积放大系数与位置权重，构建综合损伤得分模型；同时通过健康指数（Health Index, HI）实现容器整体状态的量化评估。系统实现了从缺陷识别到结构健康诊断的自动化流程，并通过批量评估、可视化报表与雷达图展示，实现了对集装箱群体状态的智能统计与高风险容器的精准预警。综合实验结果表明，本研究提出的检测与评估体系在复杂背景、光照变化及多尺度破损场景下均表现出良好的鲁棒性与可解释性，为港口自动化检测与安全运维提供了技术支持。

总体而言，本文的主要创新点包括：

1. 构建了端到端的集装箱破损检测与健康评估流程，实现了从视觉识别到结构量化的闭环分析；
2. 在检测阶段引入注意力机制与多尺度特征融合结构，显著提升了小目标识别能力；
3. 提出了融合缺陷类型、面积与位置的健康指数模型，为容器结构健康的定量评估提供了可扩展框架；
4. 实现了可视化评估与智能报告生成，具备工程落地价值。

5.2 研究展望

尽管本研究实现了较为完整的检测与评估体系，但仍存在若干可进一步优化与拓展的方向。首先，在数据层面，现有数据集仍以静态图像为主，缺乏时间序列或多角度观测信息。未来可通过多视角摄像机或无人机巡检系统采集连续影像，利用时空建模方法（如时序卷积网络、Transformer-based Video Detector）实现动态破损跟踪与演化分析，从而更精确地评估容器疲劳与老化过程。

其次，在模型层面，可进一步探索轻量化与硬件适配优化，以实现端侧部署与实时检测。当前的 YOLOv11 模型虽然精度较高，但计算开销较大，不利于在边缘设备上

长时运行。未来可结合网络剪枝（Pruning）、低秩分解（LoRA/SVD）、混合精度推理（FP16/INT8）等技术，构建针对国产 NPU、Jetson Orin 等硬件平台的轻量化检测模型，实现 10ms 级实时推理，从而满足港口现场自动化检测与无人巡检需求。

第三，在系统层面，健康指数模型仍以经验权重为主，未来可引入多模态学习与物理约束优化，将视觉缺陷与传感器数据（如震动、温度、应变等）融合，通过多源信息交叉验证实现结构状态的精准评估。结合数字孪生（Digital Twin）技术，可构建集装箱全生命周期的虚拟健康档案，实现“检测—诊断—预测—维护”一体化闭环。

最后，在应用层面，系统具有较强的可扩展性，可推广至铁路货厢、船体外板、桥梁面板等其他工业结构领域，特别适用于边缘计算场景下的安全巡检任务。未来结合 5G 边缘计算与云端协同推理，可构建多终端协同检测架构，实现港口、码头与仓储系统的智能互联，为智慧物流与智能制造提供支撑。

综上所述，本研究不仅为集装箱智能检测提供了可行的技术路径，也为结构健康评估领域提供了可扩展的计算框架。随着边缘 AI、传感融合与自监督学习等技术的持续发展，基于视觉的集装箱健康评估系统有望在未来实现更高的智能化、自适应化与部署灵活性。

参考文献

[1] KALUZA P, KOELZSCH A, GASTNER M, et al. The complex network of global cargo ship movements[J/OL]. Journal of the Royal Society Interface, 2010, 7(48): 1093-1103. DOI: 10.1098/rsif.2009.0495.

[2] WANG Z, GAO J, ZENG Q, et al. Multitype damage detection of container using cnn based on transfer learning[J/OL]. Mathematical Problems in Engineering, 2021, 2021: 1-12. DOI: 10.1155/2021/5395494.

[3] BONNÍN-PASCUAL F, ORTIZ A. Detection of cracks and corrosion for automated vessels visual inspection[C/OL]//ALQUÉZAR R, MORENO A, AGUILAR-MARTIN J. Frontiers in Artificial Intelligence and Applications: Vol. 210 Artificial Intelligence Research and Development - Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence, l'Espluga de Francolí, Tarragona, Spain, 20-22 October 2010. IOS Press, 2010: 111-120. https://doi.org/10.3233/978-1-60750-643-0-111.

[4] ORTIZ A, BONNIN-PASCUAL F, GARCIA-FIDALGO E, et al. Vision-based corrosion detection assisted by a micro-aerial vehicle in a vessel inspection application[J/OL]. Sensors, 2016, 16(12): 2118. https://www.mdpi.com/1424-8220/16/12/2118. DOI: 10.3390/s16122118.

[5] AIJAZI A K, MALATERRE L, TAZIR M L, et al. Detecting and analyzing corrosion spots on the hull of large marine vessels using colored 3D LiDAR point clouds[J/OL]. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., 2016, III-3: 153-160. https://doi.org/10.5194/isprs-annals-III-3-153-2016.

[6] LI Y, ZHANG C. Automated vision system for fabric defect inspection using gabor filters and pcnn[J/OL]. SpringerPlus, 2016, 5(1): 765. https://springerplus.springeropen.com/articles/10.1186/s40064-016-2452-6.

[7] DING S, LIU Z, LI C. Adaboost learning for fabric defect detection based on hog and svm[C/OL]//Proc. Int. Conf. Multimedia Technology (ICMT). 2011: 2903-2906. DOI: 10.1109/ICMT.2011.6001937.

[8] VALAVANIS I, KOSMOPOULOS D. Multiclass defect detection and classification in weld radiographic images using geometric and texture features[J/OL]. Expert Systems with Applications, 2010, 37(12): 7606-7614. https://www.sciencedirect.com/science/article/pii/S0957417410003829. DOI: https://doi.org/10.1016/j.eswa.2010.04.082.

[9] KASBAN H, ZAHRAN O, ARAFA H, et al. Welding defect detection from radiography images with a cepstral approach[J/OL]. NDT & E Int., 2011, 44(2): 226-231. https://doi.org/10.1016/j.ndteint.2010.10.005.

[10] İMAMOğLU Z, TUGLULAR T, BASTANLAR Y. Container damage detection and classification using container images[C/OL]//2020 28th Signal Processing and Communications Applications Conference (SIU). IEEE, 2020: 1-6. https://ieeexplore.ieee.org/document/9302442. DOI: 10.1109/SIU49456.2020.9302442.

[11] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149. https://arxiv.org/abs/1506.01497. DOI: 10.1109/TPAMI.2016.2577031.

[12] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 779-788. https://arxiv.org/abs/1506.02640. DOI: 10.1109/CVPR.2016.91.

[13] REDMON J, FARHADI A. Yolov3: An incremental improvement[A/OL]. 2018. https://arxiv.org/abs/1804.02767.

[14] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C/OL]//Lecture Notes in Computer Science: Vol. 9351 Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, 2015: 234-241. https://arxiv.org/abs/1505.04597. DOI: 10.1007/978-3-319-24574-4_28.

[15] HE K, GKIOXARI G, DOLLáR P, et al. Mask r-cnn[C/OL]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 2980-2988. https://arxiv.org/abs/1703.06870. DOI: 10.1109/ICCV.2017.322.

[16] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[A]. 2020.

[17] BENJUMEA A, TEETI I, CUZZOLIN F, et al. Yolo-z: Improving small object detection in yolov5 for autonomous vehicles[A]. 2021.

[18] LIU Y, WANG J, YU H, et al. Surface defect detection of steel products based on improved yolov5[C]//2022 41st Chinese Control Conference (CCC). IEEE, 2022: 5794-5799.

[19] LI C, LI L, GENG Y, et al. Yolov6 v3. 0: A full-scale reloading[A]. 2023.

[20] WANG C Y, YEH I H, MARK LIAO H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Springer, 2024: 1-21.

[21] WANG A, CHEN H, LIU L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[22] ZHANG H, WU C, ZHANG Z, et al. Resnest: Split-attention networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 2736-2746.

[23] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[24] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[A]. 2017.

[25] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[26] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[27] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network[J]. Advances in neural information processing systems, 2015, 28.

[28] HE A, LI T, LI N, et al. Cabnet: Category attention block for imbalanced diabetic retinopathy grading[J/OL]. IEEE Transactions on Medical Imaging, 2020, 40(1): 143-153. DOI: 10.1109/TMI.2020.3023463.

[29] KHANAM R, HUSSAIN M. Yolov11: An overview of the key architectural enhancements [A/OL]. 2024. arXiv: 2410.17725. https://arxiv.org/abs/2410.17725.

队伍编号	MCB2500488
赛道	B

物流理赔风险标注分类及运单赔付预测研究

摘 要

现代物流业快速发展的同时，货损理赔问题日益凸显，本研究通过构建物流理赔风险评估双阈值模型与 XGBoost 模型，深入研究物流理赔风险识别及服务升级路径。

针对问题 1，首先，进行数据预处理，剔除数据异常值，计算“理赔差额”数据列；其次，基于**分位数理论**，以第一和第三四分位数为边界划分低、中、高三层金额层，并分别将每层金额层按照 0.03 和 0.85 的阈值进行初步分类划分，得到“合理诉求”、“诉求偏高”、“严重超额”三类风险分类集；再次，对三类风险分类集分别进行**线性拟合、指数函数和幂函数拟合，**使用**最小二乘法**估计函数参数，并以决定系数 R^2 评估函数拟合效果，得出“合理诉求“阈值中线性函数的**最优参数 a=2.7224，b=234.5745**；最后，依据决定系数 R^2 评估拟合效果，得出**“合理诉求”阈值最优函数为线性函数、“严重超额”阈值最优函数为幂函数**，分别作为划分风险标注分类的阈值。

针对问题 2，构建基于梯度提升树的 XGBoost 回归预测模型。首先，进行数据预处理，对数值型特征缺失值使用中位数进行填补，对分类型特征缺失值使用众数填补，并采用**取绝对值**、变负数和**四分位距（IQR）**的方法处理异常值；其次，构造统计、比率、频率**特征工程**；再次，构建**基于梯度回归树的 XGBoost 回归模型**，使用所有特征进行训练，并通过**随机搜索法**优化模型参数；然后，基于第一阶段全量特征模型的特征重要性排序，**部分重要性排序：索赔金额>保价金额>进线人身份>进线渠道>妥投到进线时长,** 选取前 15 个较为重要的特征，并使用筛选后的特征重新训练模型；最终，得出附件 2 运单的实际赔付金额预测结果,部分结果举例：**实际赔付金额为 243.88，索赔金额为 436.69 时，标注为合理诉求**。

针对问题 3，首先，使用问题 1 的风险标注规则，在保证逻辑一致性、密集性和稀疏性的分类原则情况下，基于**四分位距（IQR）**将运单数据的实际赔付金额分为低、中、高三级金额层；其次，选用线性函数、指数函数和幂函数进行阈值拟合，依据决定系数选择最优拟合函数，并对附件 2 的运单进行风险标注预测；再次，构建 XGBoost 模型，预测附件 2 的运单风险标注分类；最后，比较双阈值函数模型和 XGBoost 模型对运单风险标注分类的优劣势，得出结论：**双阈值函数模型的优点是解释性强，计算简单**。

关键词：双阈值函数模型；XGBoost 回归模型；特征工程

目 录

一、问题重述 1

1.1 问题背景 1

1.2 问题提出 1

二、问题分析 2

2.1 针对问题 1 的分析 2

2.2 针对问题 2 的分析 2

2.3 针对问题 3 的分析 2

三、模型假设 2

四、符号说明 3

五、模型的建立与求解 3

5.1 问题 1 的模型与求解 3

5.1.1 数据预处理 3

5.1.2 基于分位数的分层方法 3

5.1.3 双阈值函数拟合模型 5

5.1.4 风险分类模型 8

5.2 问题 2 模型的建立与求解 9

5.2.1 数据预处理 10

5.2.2 特征工程 10

5.2.3 XGBoost 回归模型 11

5.2.4 参数优化 12

5.2.5 特征重要性评估 12

5.2.6 特征筛选与重训练 13

5.2.7 预测结果 14

5.3 问题 3 的模型与求解 14

5.3.1 风险标注模型 14

5.3.2 基于机器学习的风险预测模型 16

5.3.3 优劣势 18

六、模型评价 18

6.1 优点 18

6.2 缺点 18

七、参考文献 18

一、问题重述

1.1 问题背景

现代物流业是我国基础性、战略性产业。目前，我国物流市场规模已连续 9 年位居世界第一，2025 年前三季度社会物流总额达 263.2 万亿元，同比增长 5.4%[1][2]。然而，货损理赔问题日益凸显，现已成为制约行业高质量发展的重要瓶颈，2024 年货损率平均达 7.8%，远高于发达国家 3%~5%的水平[3]。传统理赔模式效率低下，从货损发现到最终赔付平均耗时 18 个工作日，45%的案件存在责任争议，实际赔付金额仅为损失金额的 50%左右[3][4]。更值得关注的是，在数字化转型背景下，仅有 35%的物流企业采用自动化理赔系统，理赔环节的数字化滞后严重制约了产业链协同发展[3]。

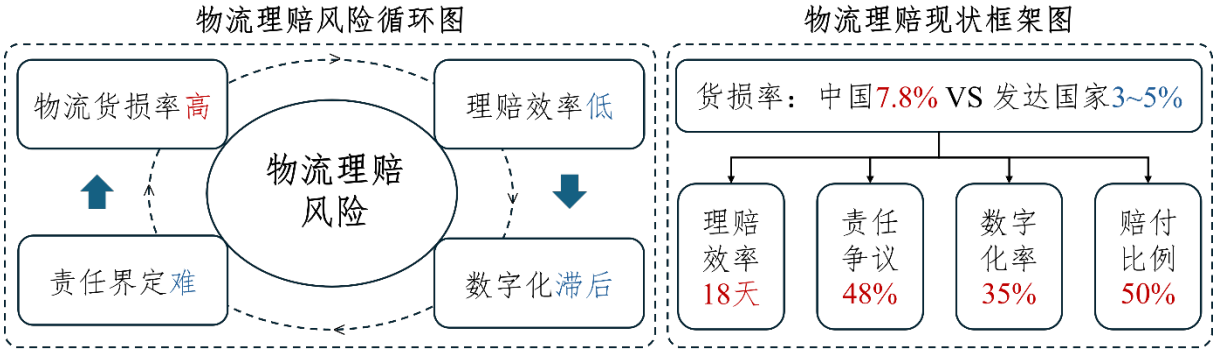


图 1 物流理赔风险研究背景图

《“十四五”现代物流发展规划》明确提出推进物流提质增效，“十五五”规划进一步强化智慧物流发展要求[5][6]。在此背景下，研究物流理赔风险识别及服务升级具有重要意义：理论上可丰富物流风险管理理论体系；实践中能有效降低企业运营成本，提高客户满意度；社会层面有助于减少资源浪费，促进绿色低碳发展，为构建现代流通体系提供有力支撑。物流理赔风险识别及服务升级已成为行业亟待解决的重要课题。

1.2 问题提出

为研究物流理赔风险识别及服务升级路径，本研究需解决以下问题：

针对问题 1：基于附件 1 中的各项数据，分析“理赔差额”与“实际赔付金额”存在的变化规律，并构建风险标注模型，阐述风险标注的原理与规则，将附件 1 中各运单的物流理赔交易行为分别标注为“合理诉求”、“诉求偏高”、“严重超额”三类风险类型。

针对问题 2：构建运单实际赔付金额的预测模型，使用附件 1 中的部分数据进行模型训练，筛选出影响实际赔付金额预测结果的评估指标，得出附件 2 中运单实际赔付金额的预测结果，并使用附件 1 中的剩余数据进行验证，得出最终优化后的预测结果。

针对问题 3：基于问题 1 设置的风险标注规则，使用附件 1 中的数据构建风险标注分类模型，并根据构建的风险标注分类模型预测附件 2 运单的理赔风险程度；处理“严重超额”样本占比严重不均衡的问题，并分析此方法与问题 2 预测方法的优劣势。

二、问题分析

2.1 针对问题 1 的分析

针对问题 1，要构建风险标注模型，对附件 1 中的运单进行风险标注。首先，进行数据预处理，剔除数据异常值，增加“理赔差额”数据列；其次，基于分位数理论，以第一和第三四分位数为边界划分低、中、高三层金额层，并分别将每层金额层按照 0.03 和 0.85 的阈值进行初步分类划分，得到“合理诉求”、“诉求偏高”、“严重超额”三类风险分类集；再次，对三类风险分类集分别进行线性拟合、指数函数和幂函数拟合，使用最小二乘法估计函数参数，并以决定系数 R^2 评估函数拟合效果；最后，依据决定系数 R^2 评估拟合效果，选择最优函数作为划分风险标注分类的阈值。

2.2 针对问题 2 的分析

针对问题 2 中物流货损的实际赔付金额预测问题，构建基于梯度提升树的 XGBoost 回归预测模型。首先，进行数据预处理，对数值型特征缺失值使用中位数进行填补，对分类型特征缺失值使用众数填补，并采用取绝对值、变负数和四分位距（IQR）的方法处理异常值；其次，构造统计、比率、频率特征工程；再次，构建基于梯度回归树的 XGBoost 回归模型，使用所有特征进行训练，并通过随机搜索优化模型参数；然后，基于第一阶段全量特征模型的特征重要性排序，选取前 N 个较为重要的特征，并使用筛选后的特征重新训练模型；最终，得出附件 2 运单的实际赔付金额预测结果。

2.3 针对问题 3 的分析

针对问题三，首先，使用问题 1 的风险标注规则，在保证逻辑一致性、密集性和稀疏性的分类原则情况下，基于四分位距（IQR）将运单数据的实际赔付金额分为低、中、高三级金额层；其次，选用线性函数、指数函数和幂函数进行阈值拟合，依据决定系数选择最优拟合函数，并对附件 2 的运单进行风险标注预测；再次，构建 XGBoost 模型，预测附件 2 的运单风险标注分类；最后，比较双阈值函数模型和 XGBoost 模型对运单风险标注分类的优劣势。

三、模型假设

假设一：附件中“始发网点万单理赔率”出现的负值、“目的网点万单理赔率”出现的正值为数据统计输入错误，可通过数据预处理进行修正。

假设二：模型迭代过程中因机器计算导致的误差可以忽略不计，且对最终的预测结果不产生实质性影响。

假设三：训练数据样本间相互独立，且服从同一概率分布。

四、符号说明

符号	说明
Q_1	第一四分位数
Q_3	第三四分位数
R^2	决定系数
f_1^*	合理诉求阈值最优拟合参数
f_2^*	严重超额阈值最优拟合参数
frequency	实体活跃度
l	损失函数
Ω	正则化项
T	树的叶子节点数
γ、λ	正则化参数
μ	理赔差额均值
σ	标准差

五、模型的建立与求解

5.1 问题 1 的模型与求解

针对问题 1，本研究构建物流理赔风险评估双阈值模型，将分层分析与函数拟合相结合，通过双阈值函数动态划分不同的风险标注等级，实现对物流理赔风险的精准分类。

5.1.1 数据预处理

首先，考虑到寄件人和收件人一般为不同人的现实情况，使用 Excel 软件的筛选功能，将附件 1 中“寄件人 id”和“收件人 id”相同的运单剔除；其次，由于保价金额均大于 0，因此剔除附件 1 中保价金额小于 0 的运单数据；最后，依据“理赔差额=索赔金额-实际赔付金额”的计算方式，增加“理赔差额”数据列。

5.1.2 基于分位数的分层方法

对预处理后的附件 1 中的运单数据按照“实际赔付金额”的数值大小升序排序，基于分位数理论，以第一和第三四分位数为边界划分为三层金额层，并分别将每层金额层按照 0.03 和 0.85 的阈值进行初步分类划分，即三层金额层分别得到“合理诉求”、“诉求偏高”、“严重超额”的风险分类集。

（1）分位数理论

由于附件 1 中的数据均为离散数据，因此，分位数的计算采用线性插值法：

$$\mathrm{Q_p} = x_{(k)} + \frac{p(n+1)-k}{k+1-k}\left(x_{(k+1)} - x_{(k)}\right) \tag{1}$$

其中，$\mathrm{k} = \mathrm{floor}\left(\mathrm{p}\left(\mathrm{n}+1\right)\right), x_{\{(k)\}}$ 表示第 k 个顺序统计量。

（2）确定分层边界

根据实际赔付金额的描述性统计分析，得出其整体集中分布在中部区域，两端差异较为显著。因此，将“实际赔付金额”使用四分位数划分为三个层次：低金额层为 $y \le Q_1$；中金额层为 $Q_1 < y \le Q_3$；高金额层为 $y > Q_3$。其中 Q_1 和 Q_3 分别为第一四分位数和第三四分位数，即 25%分位数和 75%分位数。

经过计算，得出实际赔付金额的 25%分位数为 96.86，75%分位数为 426.74。

（3）计算分层阈值

根据（2）中分得的三个层次的数据，计算每层的双阈值（X_1，X_2）。其中，第一层阈值 X_1 控制“合理诉求”占比，第二层阈值 X_2 控制“严重超额”占比。

$$\begin{cases} X_1 = F_x^{-1}(p_{合理}) \\ X_2 = F_x^{-1}(1 - p_{严重}) \end{cases} \tag{2}$$

其中，$p_{合理}$=0.85 为目标合理占比；$p_{严重}$=0.03 为目标严重超额占比；F_x^{-1} 为理赔差额的累积分布函数的反函数。

最终，三层金额层分别得到“合理诉求”、“诉求偏高”、“严重超额”的风险分类范围。

表 1 低、中、高金额层最终阈值

	合理诉求	诉求偏高	严重超额
低金额层	≤363.33	（363.33，809.85]	>809.85
中金额层	≤363.33	（843.32,1650.27]	>1650.27
高金额层	≤1950.36	（1950.36,3211.88]	>3211.88

将低、中、高金额层三级风险标注可视化处理，见图 2。

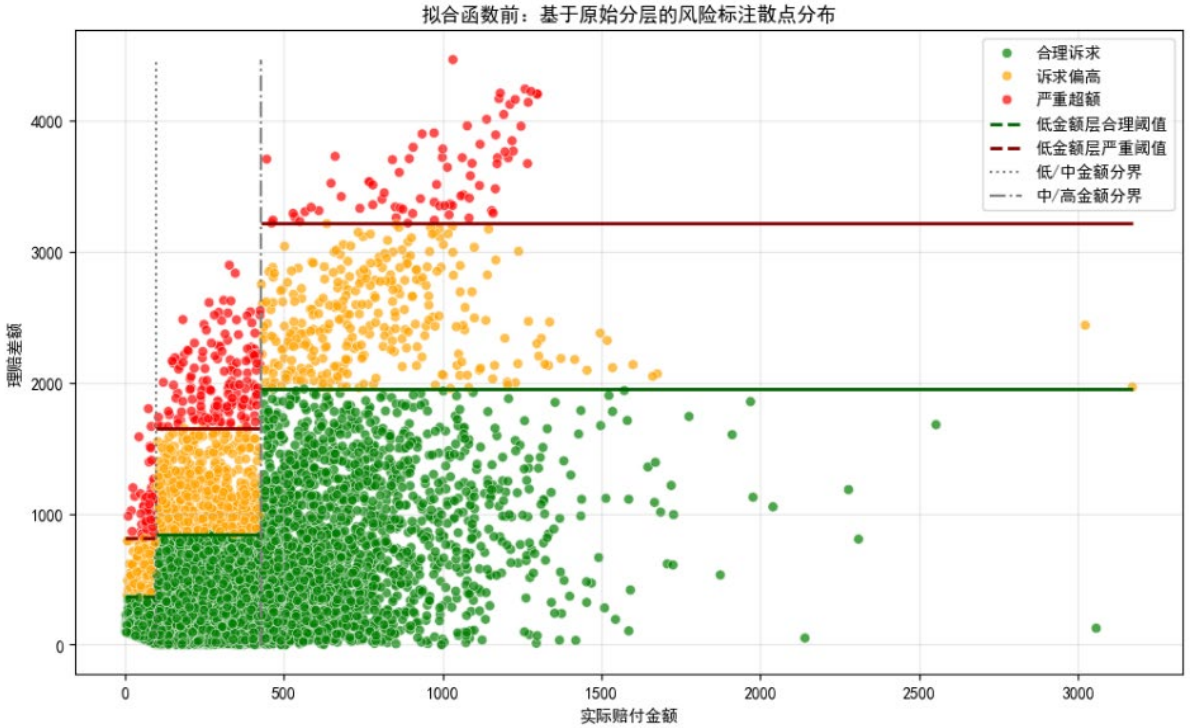


图 2 基于原始分层的风险标注

由图 2 可知，绿色区域为“合理诉求”风险标注区，黄色区域为“诉求偏高”风险

标注区，红色区域为“严重超额”风险标注区；三类风险标注区分层标注，满足“实际赔付金额越高，则需要更高的索赔差额才会标注为‘诉求偏高’或‘严重超额’”的风险标注原则。且“合理诉求”标注区占比最高，“严重超额”标注区占比最低，符合不同赔付风险运单数量的分布要求。

（4）阈值检验

为检验阈值的密集性和稀疏性是否符合“‘合理诉求’的运单索赔差额相对密集，‘严重超额’的运单索赔差额相对稀疏，且不同类型的运单索赔差额密集程度之间存在明显差别”的要求，对原始分层标注的三层索赔金额进行核密度估计（KDE），并进行可视化分析。可视化见图 3。

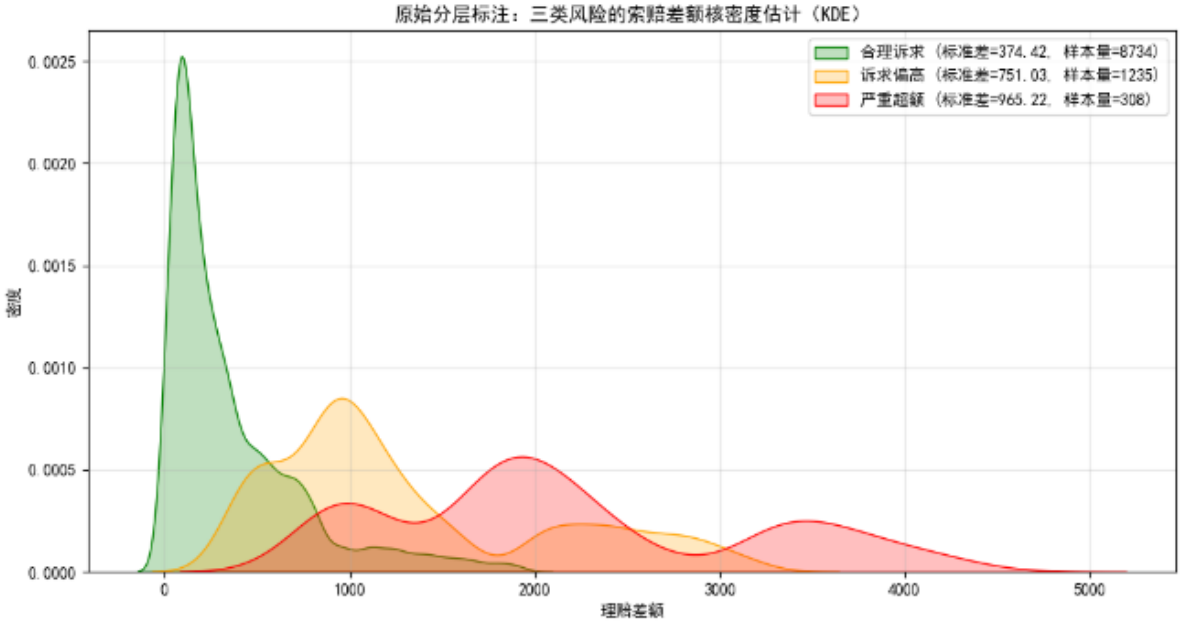


图 3 原始分层核密度估计

由图 3 可得，合理区间数据的标准差均不超过阈值、严重超额区间数据的标准差均不低于阈值，阈值的密集性和稀疏性符合要求。

5.1.3 双阈值函数拟合模型

首先，针对分位数分层后初步划分的三类风险标注分类集，分别进行线性拟合、指数函数和幂函数拟合；其次，使用最小二乘法估计函数参数，并以决定系数 R^2 评估函数拟合效果；最后，依据决定系数 R^2 评估拟合效果，选择最优函数作为划分风险标注分类的阈值。

（1）候选函数

为选取拟合效果较好的拟合函数，候选了线性拟合、指数函数、幂函数三种函数形式进行拟合：

线性函数：

$$f(y)=a\cdot y+b \tag{3}$$

指数函数：

$$f(y)=a\cdot e^{b\cdot y/y_{\max}} \tag{4}$$

其中，$y_{\max}=\max(y)$为实际赔付金额的最大值。

幂函数：

$$f(y)=a\cdot y^{b} \tag{5}$$

（2）参数估计

使用最小二乘法估计函数参数：

$$\min_{a,b}\sum_{i=1}^{m}\left(f(y_i)-x_i\right)^2 \tag{6}$$

其中，m 为分层数量，(y_i,x_i)为各层的中位数和阈值。

（3）拟合优度评估

使用决定系数 R^2 评估拟合效果：

$$\begin{cases} R^2=1-\dfrac{\sum_{i=1}^{m}\left(x_i-\hat{x}_i\right)^2}{\sum_{i=1}^{m}\left(x_i-\overline{x}\right)^2} \\ \hat{x}_i=\mathrm{f}(y_i) \\ \overline{x}=\dfrac{1}{m}\sum_{i=1}^{m}x_i \end{cases} \tag{7}$$

其中，$\hat{x}_i$ 为拟合值，$\overline{x}$ 为阈值均值。

（4）选择最优函数

经过函数拟合，最小二乘法求参数，得到合理诉求阈值 X_1 候选函数拟合结果：

表 2 合理诉求阈值 X_1 候选函数拟合结果

	函数	决定系数 R^2	参数
线性函数	$f(y)=a\cdot y+b$	0.9994	a=2.7224 b=234.5745
指数函数	$f(y)=a\cdot e^{b\cdot y/y_{\max}}$	0.9769	a=431.8817 b=1.5139
幂函数	$f(y)=a\cdot y^{b}$	0.9971	a=17.2508 b=0.7322

同理，可得严重超额阈值 X_2 候选函数拟合结果：

表 3 严重超额阈值 X_2 候选函数拟合结果

	函数	决定系数 R^2	参数
线性函数	$f(y)=a\cdot y+b$	0.9942	a=4.0665 b=669.1786
指数函数	$f(y)=a\cdot e^{b\cdot y/y_{\max}}$	0.9657	a=931.1027 b=1.2469
幂函数	$f(y)=a\cdot y^{b}$	0.9978	a=74.6561 b=0.5824

由表 2、表 3 可知，合理诉求阈值 X_1 候选函数中，线性函数的决定系数 R^2=0.9994 最大，严重超额阈值 X_2 候选函数中，幂函数的决定系数 R^2=0.9978 最大。

（4）阈值检验

对双函数分层标注的三层索赔金额进行核密度估计（KDE），并进行可视化分析。可视化见图 4。

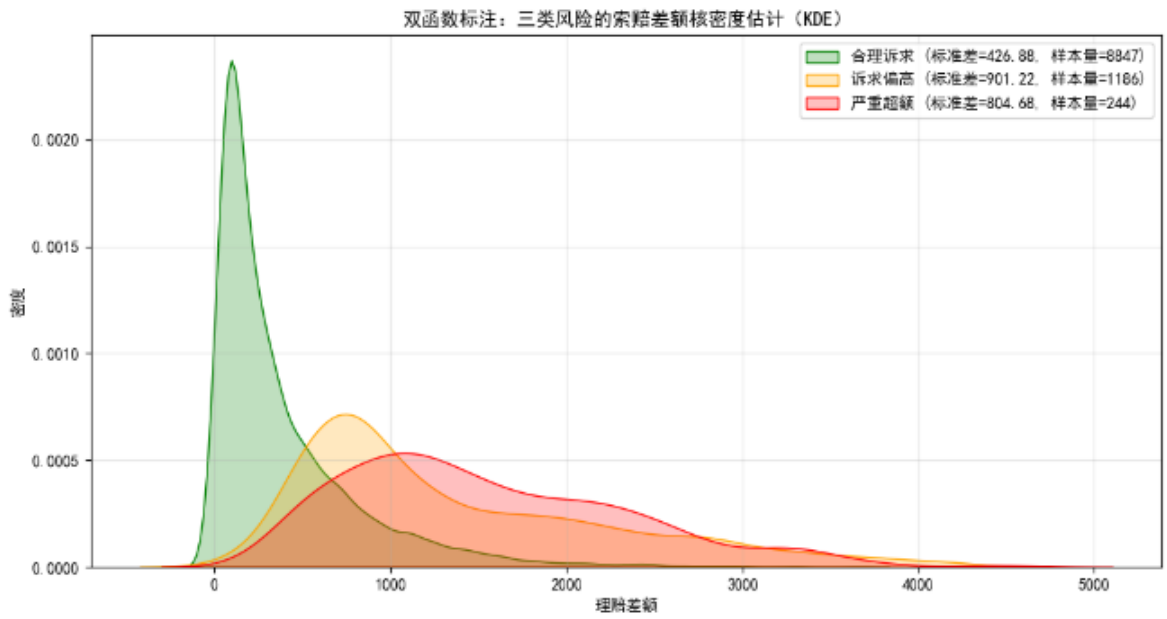


图 4 双函数标注核密度估计

由图 4 可得，合理区间数据的标准差均不超过阈值、严重超额区间数据的标准差均不低于阈值，阈值的密集性和稀疏性符合要求。

将候选函数进行可视化处理，详见图 5。

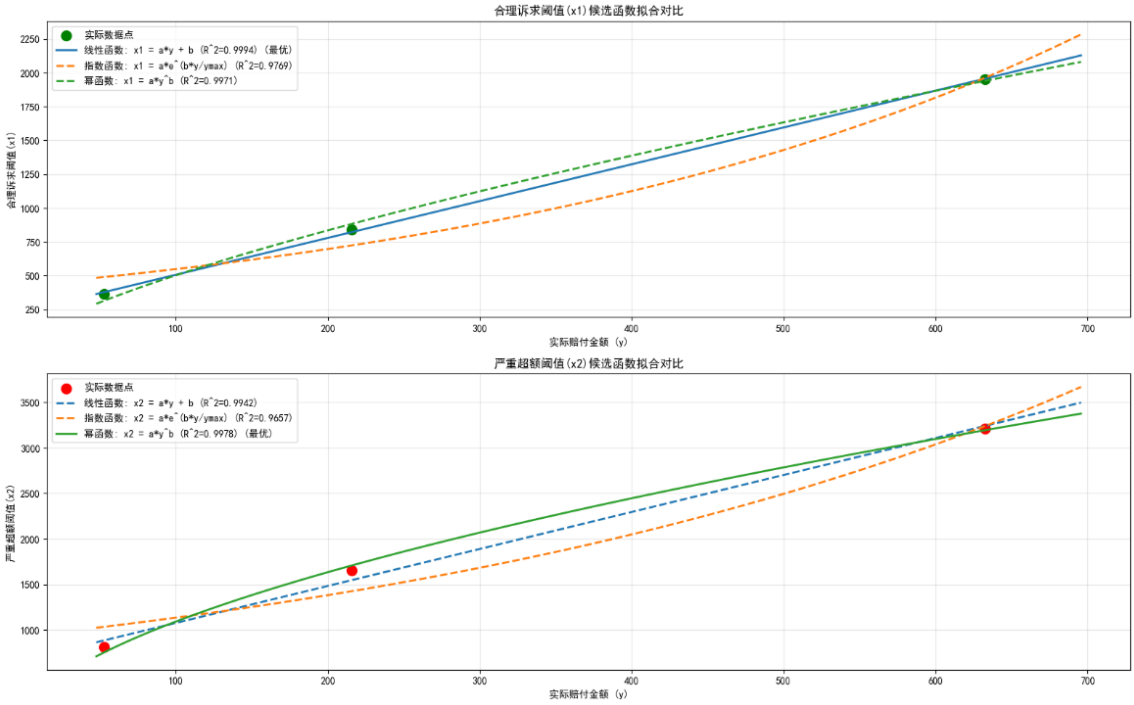


图 5 候选函数拟合对比图

由图 5 可知，“合理诉求”与“诉求偏高”边界阈值的三条候选函数中，线性函数的决定系数 R^2=0.9994，高于其他候选函数，说明该函数对 X_1 与实际赔付金额的关系解释能力最强，拟合效果最好；呈现单调递增趋势，符合“实际赔付金额越高，合理诉求阈值应适度提高”的分类规则；且相比其他函数，该函数趋势更平稳，避免对极端值过度敏感。“严重超额”与“诉求偏高”边界阈值的三条候选函数中，幂函数的决定系数

R^2=0.9978，高于其他候选函数，对 X_2 与实际赔付金额的关系刻画更准确，拟合效果最好；函数增长速率较快，符合“高金额案件需更高超额才判定为严重”的风险控制需求。因此，分别选用线性函数、幂函数作为三级风险标注分类的边界拟合曲线：

“合理诉求”与“诉求偏高”风险标注分界函数：

$$f(y)=a_1\cdot y+b_1 \tag{8}$$

其中，a_1=2.7224，b_1=234.5745。

“诉求偏高”与“严重超额”风险标注分界函数：

$$f(y)=a_2\cdot y^{b_2} \tag{9}$$

其中，a_1=74.6561，b_1=0.5824。

最终，得到实际赔付金额与双阈值的函数关系图：

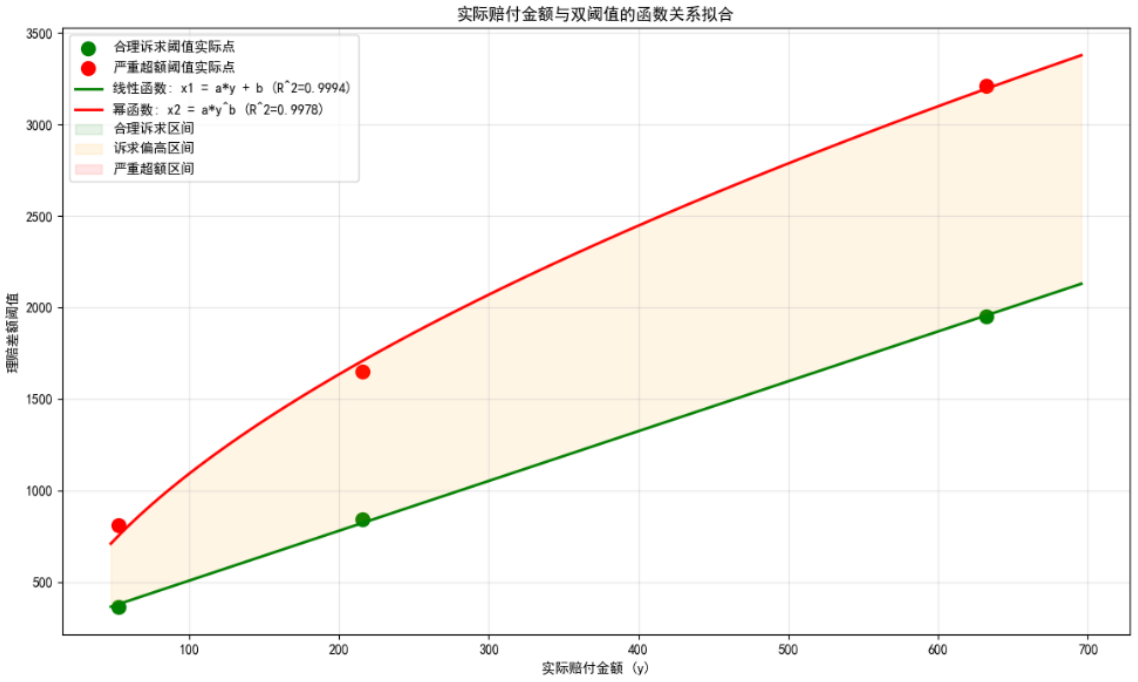


图 6 实际赔付金额与双阈值得函数关系拟合

5.1.4 风险分类模型

（1）动态阈值计算

对于任意实际赔付金额 y，计算动态双阈值：

$$\begin{cases} X_1(y)=f_1^*(y) \\ X_2(y)=f_2^*(y) \end{cases} \tag{10}$$

其中，f_1^*和f_2^*分别为“合理诉求”阈值和“严重超额”阈值的最优拟合函数。

（2）风险等级划分

根据理赔差额 x 与动态阈值的关系进行风险分类：$x\le X_1(y)$划分为“合理诉求”风险分类；$X_1(y)<x\le X_2(y)$划分为“诉求偏高”风险分类；$x\le X_1(y)$划分为“严重超额”风险分类。可视化见图 7。

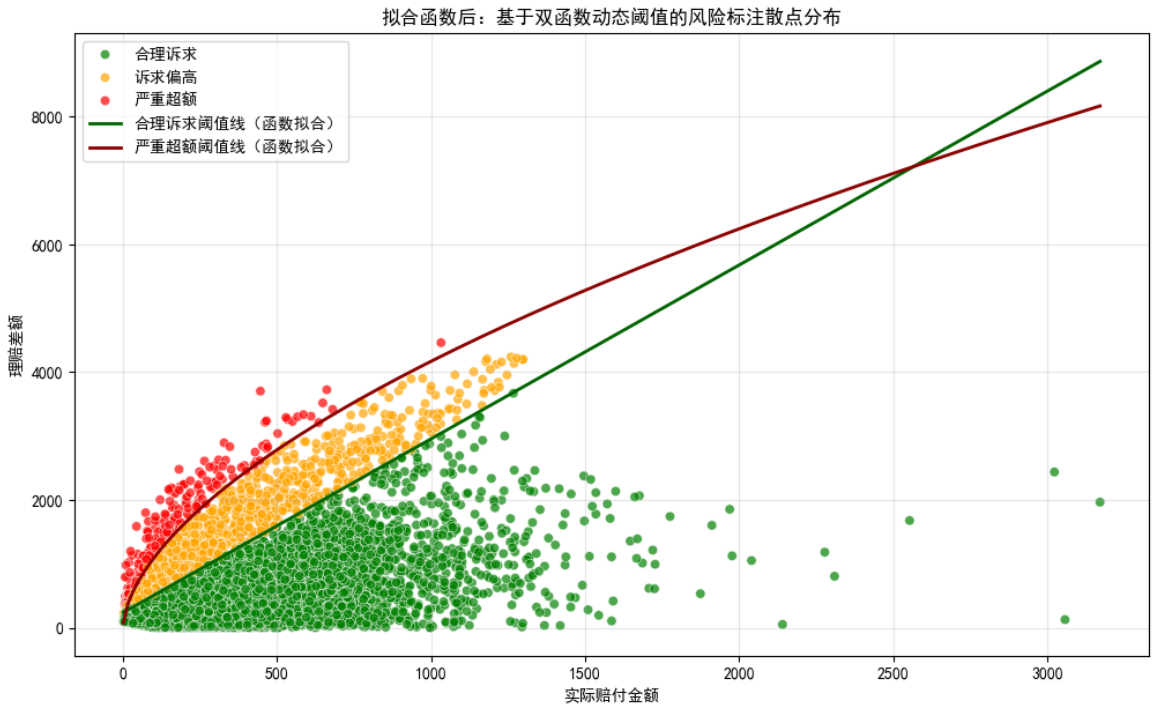


图 7 基于双函数动态阈值的风险标注

由图 7 可知，绿色区域为“合理诉求”风险标注区，黄色区域为“诉求偏高”风险标注区，红色区域为“严重超额”风险标注区；“合理诉求”风险标注区与“诉求偏高”风险标注区。

部分风险标注结果展示如下：

表 4 部分风险标注结果

实际赔付金额	理赔差额	风险标注	风险标注_双函数拟合
336.47	454.27	合理诉求	合理诉求
55.7	93.47	合理诉求	合理诉求
136.18	127.56	合理诉求	合理诉求
97.29	291.27	合理诉求	合理诉求
457.81	15.96	诉求偏高	严重超额
1094.63	396.53	诉求偏高	合理诉求
590.86	126.16	合理诉求	诉求偏高
128.2	1.79	合理诉求	诉求偏高
1600.35	472.61	合理诉求	诉求偏高
1440.21	432.28	合理诉求	诉求偏高
571.18	101.81	合理诉求	诉求偏高
351.43	8.06	合理诉求	严重超额
3715.57	1172.64	严重超额	诉求偏高

5.2 问题 2 模型的建立与求解

针对问题 2 中物流货损的实际赔付金额预测问题，构建了基于梯度提升树的回归预

测模型。研究基于附件 2 所给运单数据，即包含物流订单特征的数据集 $D=\{(x_i,y_i)\}_{i=1}^{n}$，其中，$x_i\in R^d$ 为 d 维特征向量，$y_i\in R^+$ 为对应的实际赔付金额，以学习映射函数 $f:R^d\to R^+$ 为目标，使得对于新的样本 x，能够准确预测其赔付金额。

5.2.1 数据预处理

首先，进行缺失值处理，对于数值型特征缺失值，使用中位数进行填补；对于分类型特征，使用众数填补。

其次，采用取绝对值、变负数和四分位距（IQR）的方法处理异常值。对“始发网点万单理赔率”出现的负值取绝对值；对“目的网点万单理赔率”出现的正值“变正为负”；对“配送超时时长”、“妥投到线时长”、“始发网点发单量”、“目的网点发单量”等特征中数值过大的特征值进行截断：

$$x_{ij}^{clipped}=\begin{cases}Q_1-1.5\times IQR, & x_{ij}<Q_1-1.5\times IQR\\ x_{ij}, & Q_1-1.5\times IQR\le x_{ij}\le Q_3+1.5\times IQR\\ Q_3+1.5\times IQR, & x_{ij}>Q_3+1.5\times IQR\end{cases}\tag{11}$$

其中，Q_1 和 Q_3 分别为第一和第三四分位数，$IQR=Q_3-Q_1$。

再次，对“始发网点发单量”和“目的网点发单量”等特征进行可视化处理，将离群的异常值剔除，以“始发网点发单量”和“目的网点发单量”为例，详见图 8、图 9。

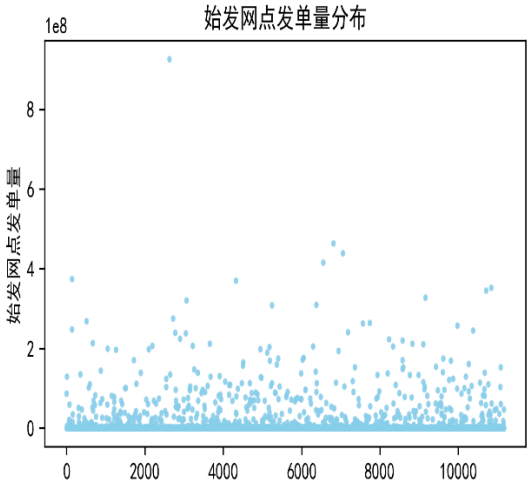


图 8 始发网点发单量分布图

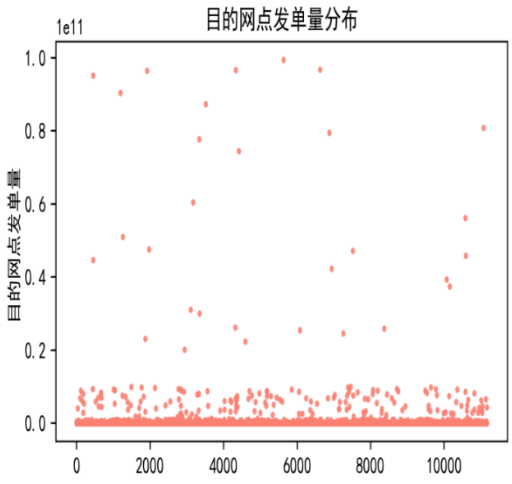


图 9 目的网点发单量分布

最后，对数值型特征进行标准化处理：

$$x_{ij}^{norm}=\frac{x_{ij}-\mu_j}{\sigma_j}\tag{12}$$

其中，μ_j 为特征 j 的均值，σ_j 为特征 j 的标准差。

5.2.2 特征工程

针对“商品类型”、“寄件 B/C”、“进线人身份”等非数值型数据，通过构造特征工程进行特征编码处理。

（1）统计特征构造

构造网点指标均值特征：

$$\begin{cases} avg\_claim\_rate = \dfrac{origin\_claim\_rate + dest\_claim\_rate}{2} \\ avg\_payout\_ratio = \dfrac{origin\_payout\_ratio + dest\_payout\_ratio}{2} \end{cases} \tag{13}$$

（2）比率特征构造

构造时长比率特征，反映进线效率：

$$efficiency\_ratio = \frac{claim\_response\_time}{delivery\_delay\_time + \varepsilon} \tag{14}$$

其中，ε 为极小值，防止分母为零。

（3）频率特征构造

构造频率特征，反映实体的活跃度：

$$frequency(v) = \frac{count\left(x_j = v\right)}{n} \tag{15}$$

其中，v 为特征 j 的某个取值，n 为总样本数。

5.2.3 XGBoost 回归模型

首先，构建基于梯度回归树的 XGBoost 回归模型，使用所有特征进行训练，并通过随机搜索优化模型参数；其次，基于第一阶段全量特征模型的特征重要性排序，选取前 N 个较为重要的特征，并使用筛选后的特征重新训练模型；最终，得出预测结果。

（1）模型原理

XGBoost 模型为基于梯度提升的集成学习算法，通过迭代训练多颗决策树并组合成强学习器，可以进行精准预测。其预测函数为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{16}$$

其中，$f_k\left(x_i\right)$ 表示第 k 颗决策树对样本 x_i 的预测值，K 为决策树的数量。

（2）目标函数

模型使用加法训练策略，第 t 轮的目标函数为：

$$\tau^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathrm{x}_i\right)\right) + \Omega\left(f_t\right) \tag{17}$$

其中，$\hat{y}_i^{(t-1)}$ 为第 t-1 轮的预测值，l 为损失函数，Ω 为正则化项。

对于回归问题，采用平方误差损失函数：

$$l\left(y_i, \hat{y}_i\right) = \left(y_i - \hat{y}_i\right)^2 \tag{18}$$

正则化项定义为：

$$\Omega\left(f_t\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{19}$$

其中，T 为树的叶子节点数，ω_j 为第 j 个叶子节点的权重，γ 和 λ 为正则化参数。

（3）梯度提升算法

对目标函数进行二阶泰勒展开：

$$\begin{cases} \tau^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right) \\ g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial\left(\hat{y}_i^{(t-1)}\right)^2} \end{cases} \tag{20}$$

（4）树结构学习

对于给定的树结构，叶子节点的最优权重为：

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{21}$$

其中，I_j 表示属于第 j 个叶子节点的样本集合。

对应的目标函数值为：

$$\tau\left(\omega^*\right) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{22}$$

5.2.4 参数优化

使用随机搜索法进行参数优化：

$$\theta^* = \arg\ \max_{\theta} CV\_score\left(X_{all}, y, \theta\right) \tag{23}$$

首先，定义搜索空间：

表 5 搜索空间定义

搜索空间	定义
树深度	max_depth~Uniform(3,8)
子节点最小权重	min_child_weight~Uniform(1,4)
样本采样率	subsample~Uniform(0.7,1.0)
特征采样率	colsample_bytree~Uniform(0.7,1.0)
剪枝阈值	γ~Uniform(0,0.2)
学习率	learning_rate~Uniform(0.03,0.1)
树数量	n_estimators~Uniform(150,500)

其次，采用 5 折交叉验证评估模型性能：

$$CV_{score} = \frac{1}{5}\sum_{k=1}^{5} score\left(D_k^{val}, \hat{y}_k\right) \tag{24}$$

其中，D_k^{val} 为第 k 折的验证集，$\hat{y}_k$ 为对应的预测值。

5.2.5 特征重要性评估

XGBoost 模型评估特征重要性主要有权重重要值、增益重要性、覆盖重要性三种计算方法，本研究选用增益重要性作为特征选择的依据，即特征在分裂时带来的增益：

$$gain\left(j\right) = \sum_{k=1}^{K}\sum_{splits\ on\ j\ in\ f_k} \Delta\tau \tag{25}$$

经过计算，得出 30 个特征值的重要性排序，前 15 项重要性指标见表 6。

表 6 部分特征重要值排序

	feature	importance
1	索赔金额	0.396410

2	保价金额	0.060308
3	进线人身份	0.037980
4	是否 c2c	0.037533
5	进线渠道	0.036139
6	妥投到进线时长	0.035418
7	配送超时时长	0.035349
8	始发网点发单量	0.030370
9	商品类型	0.026864
10	目的网点发单量	0.026061
11	收件人 id_频率	0.026021
12	异常原因	0.022917
13	始发城市	0.022587
14	目的网点赔付比例	0.021485
15	进线时长/配送时长	0.021088

将特征重要值可视化处理，见图 10。

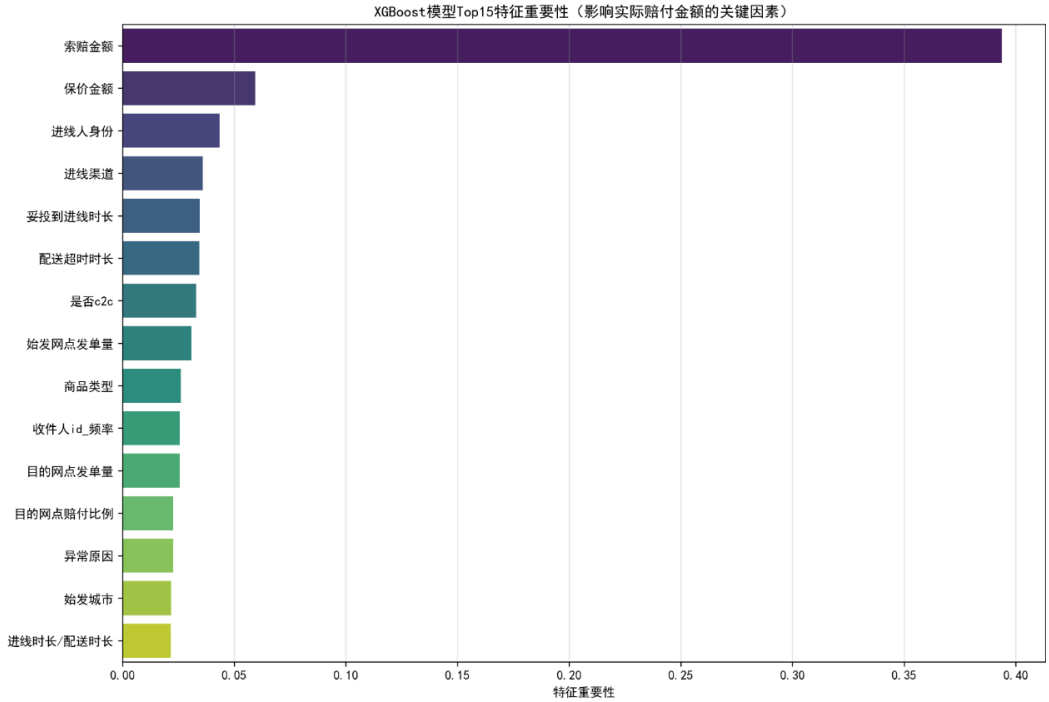


图 10 XGBoost 模型特征重要性前 15 项特征

由图 10 可得，索赔金额对实际赔付金额的影响最大，其次是保价金额和进线人身份。最终筛选出“索赔金额”、“保价金额”、“进线人身份”、“进线渠道”、“妥投到进线时长”、“配送超时时长”、“是否 c2c”、“始发网点发单量”、“商品类型”、“收件人 id_频率”、“目的网点发单量”、“目的网点赔付比例”、“异常原因”、“始发城市”、“进线时长/配送时长”等 15 项特征，用于后续模型的再训练。

5.2.6 特征筛选与重训练

基于第一阶段模型的特征重要性，选择前 15 个最重要的特征：

$$S=\left\{j \mid rank\left(gain\left(j\right)\right)\le N\right\} \tag{26}$$

使用筛选后的特征重新训练模型：

$$f^{*}=\arg\ \min_{f}\tau\left(f\left(X_{S}\right),y\right)+\Omega\left(f\right) \tag{27}$$

其中，X_S 为筛选后的特征矩阵。

5.2.7 预测结果

（1）实际值与预测值对比

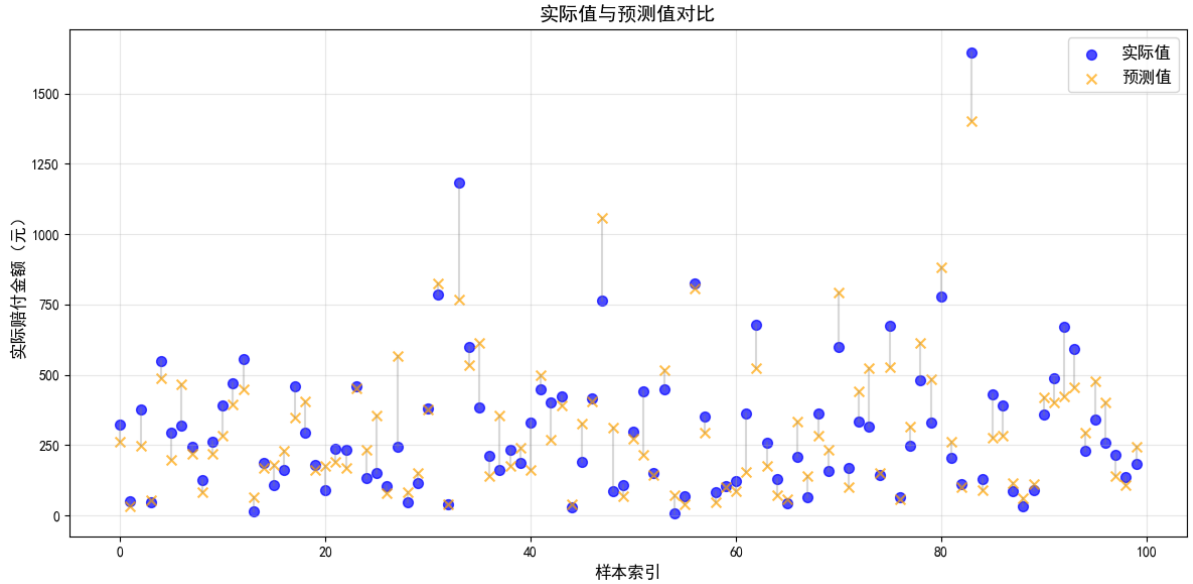


图 11 实际值与预测值对比图

由图 11 可知，蓝色实点为实际值、黄色叉点位预测值，预测值与实际值变化趋势基本一致，由此可得，模型预测效果良好。

（2）部分预测结果展示

表 7 部分实际赔付金额预测结果

运单号	实际赔付金额
1	243.8800049
2	265.3099976
3	697.5499878
4	261.5400085
5	50.95999908
6	171.8399963
7	247.8399963
8	272
9	212.9600067
10	304.7200012
11	314.7099915
12	229.3600006
13	58.06000137

5.3 问题 3 的模型与求解

5.3.1 风险标注模型

（1）数据分层策略

基于实际赔付金额的分位数将数据分为三层：低金额层：$y_i \le Q_1$；中金额层：$Q_1 < y_i \le Q_3$；高金额层：$y_i > Q_3$。

（2）分层阈值计算

首先，设定初始阈值。对于每个分层，基于目标占比计算初始阈值：合理诉求占比85%，严重超额占比 3%。

$$X_1^{(0)} = x_{(\lfloor n\alpha \rfloor)} X_2^{(0)} = x_{(\lceil n(1-\beta) \rceil)} \quad (28)$$

其中，$x_{(k)}$表示排序后的第 k 个观测值，n 为样本数量。

其次，设置阈值调整机制：

①逻辑一致性调整： $X_1 < X_2$。若 $X_1^{(0)} \geq X_2^{(0)}$，则 $X_1^{(1)} = \mu - 0.5\sigma$，$X_2^{(1)} = \mu + 1.5\sigma$。其中，$\mu$ 为理赔差额均值，σ 为标准差。

②密集性调整： 优化合理区间的密集程度。计算合理区间数据的标准差：$\sigma_1 = std\left(x_i \mid x_i \leq X_1\right)$。若 $\sigma_1 > \max_{\text{合理标准差}}$，则迭代调整 X_1。

③稀疏性调整： 优化严重超额区间的稀疏程度。计算严重超额区间数据的标准差：$\sigma_2 = std\left(x_i \mid x_i > X_2\right)$。若 $\sigma_2 < \min_{\text{严重标准差}}$，则迭代调整 X_2。

（3）模型拟合

基于问题 1 的研究方法，这里仍采用线性函数、指数函数和幂函数作为候选函数，进行阈值函数拟合；使用最小二乘法估计参数值；选用决定系数 R^2 作为模型选择的评价指标，选择决定系数 R^2 最大的函数作为最优拟合函数。

（4）新数据预测

对于附件 2 中新的理赔运单，应用拟合的双阈值函数进行风险等级预测：

$$\hat{C}_i = f\left(x_i, y_i, X_1^*, X_2^*\right) \quad (29)$$

其中，X_1^* 和 X_2^* 为最优拟合函数。

（5）预测风险标注结果

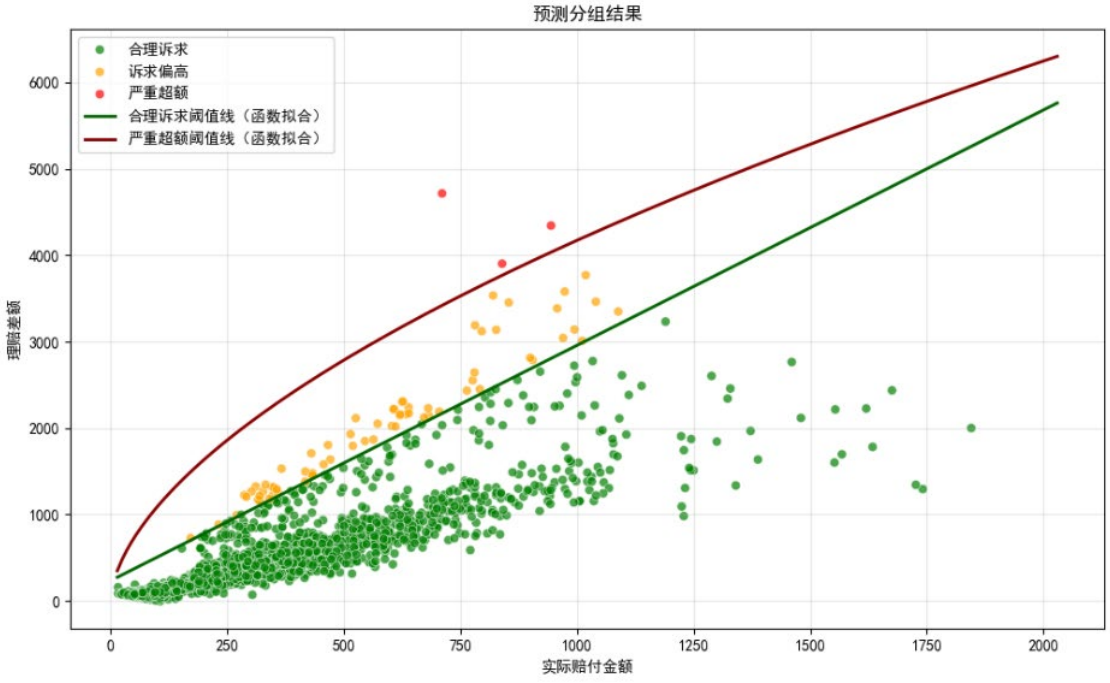


图 12 预测分组结果拟合图

由图 12 可知，使用问题 1 风险标注规则预测附件 2 运单风险标注结果，红色部分标注为“严重超额”，黄色部分标注为“诉求偏高”，绿色部分标注为“合理诉求”；“严重超额”部分占比最少、“诉求偏高”部分占比适中、“合理诉求”部分占比最多。

（6）阈值检验

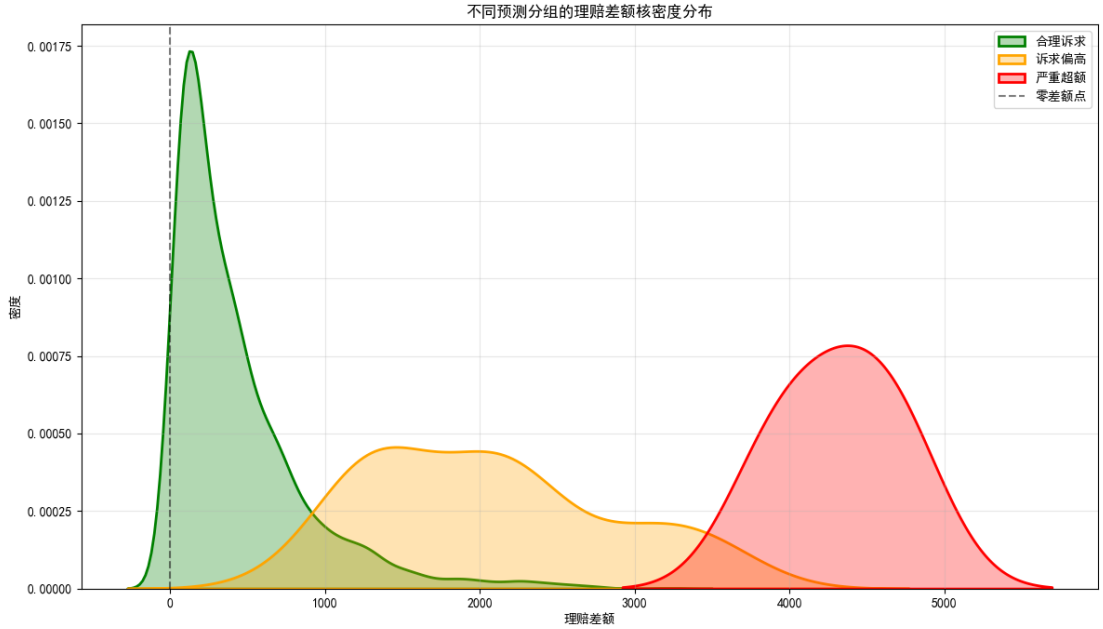


图 13 不同预测分组的理赔差额核密度分布图

表 8 预测分组结果表

运单号	实际赔付金额	理赔差额	预测分组
1	243.8800049	436.6899951	合理诉求
2	265.3099976	457.9400024	合理诉求
3	697.5499878	853.9500122	合理诉求
4	261.5400085	465.7699915	合理诉求
5	50.95999908	49.17000092	合理诉求
6	243.8800049	436.6899951	合理诉求
7	265.3099976	457.9400024	合理诉求
40	791.3400269	2449.029973	诉求偏高
119	525.3599854	2114.110015	诉求偏高
121	638.5499878	2239.510012	诉求偏高
178	171.25	727.32	诉求偏高
283	270.2099915	992.9200085	诉求偏高
301	710.460022	4714.329978	严重超额
481	310.8099976	1323.990002	诉求偏高

5.3.2 基于机器学习的风险预测模型

（1）特征工程

首先，对“实际赔付金额”、“保价金额”、“配送超时时长”等数值特征进行标准化处理，对“线路类型”、“是否生鲜妥投及时”、“商品类型”等分类特征进行编码处理。

其次，进行特征值变换。对“始发网点万单理赔率”出现的负值取绝对值；对“目

的网点万单理赔率”出现的正值“变正为负”；构造比例特征和频率特征。

（2）XGBoost 分类模型

使用问题 2 使用的模型对附件 2 运单的风险等级进行预测。对于三分类问题，使用 softmax 目标函数：

$$p_{ik}=\frac{e^{f_k(x_i)}}{\sum_{c=1}^{3}e^{f_c(x_i)}}\ \hat{y}_i=\arg\ \max_k p_{ik} \tag{30}$$

采用随机搜索法优化超参数:树深度 3-7；学习率 0.03-0.1；子样本比例 0.7-1.0；列样本比例 0.7-1.0。优化目标为宏平均 F1 分数：

$$F_1^{macro}=\frac{1}{3}\sum_{k=1}^{3}F_1(k)F_1(k)=2\cdot\frac{precision(k)\cdot recall(k)}{precision(k)+recall(k)} \tag{31}$$

（3）特征重要性分析

首先，计算基于树模型的特征重要性：

$$\text{importance}(j)=\sum_{k=1}^{t}gain(j,k) \tag{32}$$

其中，*gain(j,k)*表示第 *k* 颗树中特征的信息增益。

其次，筛选出重要性排名前 15 的特征：$S^*=\{j\mid rank(\text{importance(j)})\le N\}$。

（4）模型的评估与验证

首先，采用 5 折交叉验证评估模型性能：

$$CV(k)=\frac{1}{5}\sum_{i=1}^{5}F_1^{(i)}(k) \tag{33}$$

其次，通过混淆矩阵分析模型在不同类别上的表现：

$$CM=\begin{bmatrix}TP_1 & FN_{12} & FN_{13}\\ FN_{21} & TP_2 & FN_{23}\\ FN_{31} & FN_{32} & TP_3\end{bmatrix} \tag{34}$$

（5）预测结果

表 9 问题 3 预测结果表

运单号	实际赔付金额	理赔差额	预测分组
1	243.8800049	436.6899951	合理诉求
2	265.3099976	457.9400024	合理诉求
3	697.5499878	853.9500122	诉求偏高
4	261.5400085	465.7699915	合理诉求
5	50.95999908	49.17000092	合理诉求
6	243.8800049	436.6899951	合理诉求
7	265.3099976	457.9400024	合理诉求
40	791.3400269	2449.029973	合理诉求
119	525.3599854	2114.110015	严重超额
121	638.5499878	2239.510012	严重超额
178	171.25	727.32	诉求偏高
283	270.2099915	992.9200085	诉求偏高
301	710.460022	4714.329978	严重超额

481	310.8099976	1323.990002	诉求偏高

5.3.3 优劣势

双阈值函数模型的优点是解释性强，计算简单；缺点是对数据分布假设较强；XGBoost 模型的优点是预测精度高，适应性强；缺点是解释性较差，计算复杂。

六、模型评价

6.1 优点

①双阈值函数模型对于特征分布清晰、类别边界明确的数据，双阈值模型能够提供较好的分类效果。

②双阈值函数模型相比单阈值模型，双阈值模型能够更好地处理类别之间的过渡区域，减少误分类。

③XGBoost 模型能够输出每个特征的重要性得分，有助于理解模型决策过程和特征选择。

6.2 缺点

①双阈值模型分类精度通常低于更复杂的模型。

②XGBoost 模型对极端异常值较为敏感，可能影响模型性能。

七、参考文献

[1]中国物流与采购联合会.前三季度中国社会物流总额同比增长 5.4%[EB/OL].2025-10-29.

[2]中国物流与采购联合会.“物畅其流”!透过亮眼数据看“十四五”期间我国现代物流发展重要成就[EB/OL].2025-08-27.

[3]人人文库.货损理赔通道 2025 年物流行业风险管理案例分析报告[R].2025-09-10.

[4]安壹保险律师.2024 年物流保险诉讼事由分析[EB/OL].2024-12-24.

[5]国务院办公厅.关于印发“十四五”现代物流发展规划的通知[Z].2024-12-10.

[6]政策快讯."十五五"规划物流专项解读[EB/OL].2025-10-26.

[7]张李明,洪蕾.基于双阈值函数的音符检测旋律时间序列化分析[J].电子技术,2022,51(09):22-25.

[8]杨晓南,刘佳明,艾恒雨,等.XGBoost 模型驱动的 MOFs 吸附重金属性能预测[J/OL].哈尔滨理工大学学报,1-10[2025-10-31].

[9]李鸿霄,张怀念,杨洁,等.基于 XGBoost 对大数据驱动地形气候交互模型的构建与暴雨灾害预测[J/OL].数学的实践与认识,1-11[2025-10-31].

[10]Haoyuan F ,Xuebin Z ,Peiji S , et al.Environmental interpretation of spatial heterogeneity in the trade-offs and synergies of land use functions: A study based on the XGBoost-SHAP model[J].Journal of Arid Land,2025,17(10):1378-1401.

队伍编号	MCB2400837
赛道	A

台风的分类与预测

摘　要

台风作为一种极端气象现象，对沿海地区的安全和经济活动构成了严重威胁。随着全球气候变化的加剧，台风的频率和强度可能会发生变化，这使得准确预测其路径、强度及其对环境影响的能力显得尤为重要。台风的形成和发展受多种气象因素的影响，包括气温、气压、洋流和风场等，这些因素的复杂相互作用使得台风的研究充满挑战。近年来，借助大数据和机器学习技术，研究人员在台风路径预测和影响因素分析方面取得了显著进展。然而，现有的研究仍然面临数据维度不一致、特征提取不足以及模型鲁棒性等问题，限制了其实际应用效果。因此，系统性地研究台风的分类、路径预测及其登陆后衰减特征，不仅可以提升对台风行为模式的理解，还能为灾害预警和应对策略的制定提供有力支持。

本次赛题围绕台风分类、路径预测和登陆后风速与降水量衰减的需求展开。研究团队首先针对台风分类评价，**利用多个权威数据集及中国气象局获取的气象数据，提取台风的时序观测数据，并转化为一致性的 17 个特征向量，以确保后续建模时特征维度统一。**团队采用层次聚类算法构建分类树状图，**识别具有相似特征的台风类别，共计分类 21 类，并对每个类别进行了命名，具体结果见表 4.6-4.7**。同时，团队使用循环神经网络（RNN）捕捉台风路径的短期和长期特征，预测 2024 年 7 月及 9 月的分类情况，结果也见表 4.8。通过分析夏季与秋季台风的特征差异发现，夏季台风受低气压和西南季风影响，路径复杂，易深入内陆。秋季气压较高、气压梯度增大，受东北季风影响，台风多在沿海转向偏东北，路径较稳定，影响范围较小。此外，**研究发现随着全球气候变暖，秋季台风的行径路径呈现出与夏季台风行径路径相似的趋势。**这一发现为进一步理解台风行为模式及其气候变化影响提供了重要依据。

其次，在台风路径预测方面，研究团队探索气温、气压、洋流和风场等多重因素对路径的影响，以预测 2024 年 9 月 13 日至 17 日第 13 号台风贝碧嘉的中心位置。为提高预测精度，**团队选择了发生时期相近、强度相似和中心位置相近的历史台风数据作为训练样本。**由于台风生命周期和数据记录频率差异，团队采用步长为 2 天的滑动窗口法将每条台风数据划分为多个训练样本，从而详细刻画不同发展阶段。**最终提取了 147 个时序特征**，包括经度、纬度、风速、气压和气温等，并利用长短期记忆网络（LSTM）算法进行逐步预测，结果见表 5.2。为评估模型准确性，团队采用动态时间规整（DTW）算法计算预测路径与实际路径的差异，DTW 距离为 8.8774，表明预测路径与实际路径相似性较高；地理位移误差为 0.4035°，**显示预测点与真实点的空间偏差小，模型的路径预测效果良好**。这一研究为理解台风路径的动态变化提供了重要依据。

最后，针对台风登陆后风速和降水量的衰减预测，研究团队分析台风中心风速和降水量与登陆距离的关系，在问题二的特征基础上，**继续采用步长为 4 天的滑动窗口法，以每日 14 时的台风数据进行特征扩充，共扩充 54 个时序特征。**并结合 LSTM 与多重注意力机制，逐步预测风速和降水量的动态变化。研究还分析影响预测结果的关键气象指标，为有效应对台风灾害提供数据支持。模型通过 LSTM 与多重注意力机制，成功捕捉了台风贝碧嘉登陆后的风速与降水变化趋势。误差分析和 DTW 对比结果表明，模型在台风衰减趋势的预测上表现良好

关键词：**分类评价；系统聚类；RNN；路径预测；LSTM；DTW；多重注意力机制**

目　录

一、问题重述 1
1.1 问题背景 1
1.2 目标任务 1
二、模型假设及符号 2
2.1 模型假设 2
2.2 主要的符号说明 2
三、问题分析 3
3.1 针对问题一 3
3.2 针对问题二 4
3.3 针对问题三 5
四、问题一模型的建立与求解 6
4.1 数据预处理 6
4.1.1 数据来源 6
4.1.2 数据过滤 7
4.1.3 数据清洗与缺失值处理 8
4.1.4 数据整合 9
4.2 特征提取 11
4.2.1 原始数据局限与特征提取的必要性 11
4.2.2 搭建特征空间 11
4.2.3 关键特征的提取 12
4.3 基于层次聚类的分类模型 13
4.3.1 特征选择与数据标准化处理 13
4.3.2 层次聚类模型的构建 13
4.3.3 模型参数与阈值设置 14
4.3.4 台风路径类型划分 14
4.3.5 模型评估 15
4.3.6 结果展示与分析 16
4.4 基于循环神经网络的台风数据分类评价 16
4.4.1 模型选择与架构 16
4.4.2 模型训练与验证 17
4.4.3 分类规则与评估 17
4.5 夏台风与秋台风的对比分析 17
4.5.1 结果展示 17
4.5.2 差异分析 18
4.5.3 结果应用与分析 18
五、 问题二模型的建立与求解 19
5.1 台风的数据融合 19
5.1.1 台风数据筛选 19
5.1.2 台风数据融合 19
5.2 数据预处理 20

5.3 搭建特征空间 20
5.3.1 滑动窗口的建立 20
5.3.2 指标的选取 22
5.4 基于 LSTM 的路径预测模型 24
5.4.1 训练集与测试集的构建 24
5.4.2 LSTM 模型的实施 24
5.4.3 预测结果与分析 25
5.4.4 问题二小结 26
六、 问题三模型的建立与求解 28
6.1 台风数据融合 28
6.1.1 数据筛选 28
6.1.2 数据融合 29
6.2 搭建特征空间 30
6.3 基于 LSTM+多重注意力机制的预测模型 31
6.3.1 滑动窗口处理数据 31
6.3.2 模型的实施 32
6.3.3 预测结果分析 32
6.4.4 问题三小结 33
七、模型的评价与推广 34
7.1 模型评价 34
7.2 模型推广 34
八、参考文献 35
九、附录 36

一、问题重述

1.1 问题背景

台风是全球范围内破坏力最强的**自然灾害**之一，尤其在**夏秋季节**，常伴随**强风、暴雨、风暴潮**等极端天气，对沿海地区的**经济和社会发展**造成严重影响。台风的形成通常依赖于**海洋温度、气压系统和地球自转偏向力**等气象因素。当**海洋表面温度上升**时，水汽蒸发形成低压气旋，并在地转偏向力的作用下旋转并发展成台风，特别是**西北太平洋地区**，台风活动频繁，**路径复杂**，常呈现出**西进型、转向型、迷走型**等多种模式。

然而，近年来**全球气候变暖**，海洋表面温度上升，**台风的频率和强度**发生了显著变化，**路径预测的难度**也随之增加。2024 年第 13 号台风**贝碧嘉**便是这一现象的**典型案例**。贝碧嘉的**初期路径预测**显示它将影响浙江沿海，然而在逼近浙江时，台风发生**转向**，未登陆台州，最终路径偏向了其他区域。这次**预测偏差**引发了防灾资源的调度浪费，给多个地区带来**误判的风险**。

这次台风的路径变化强调了**台风分类与预测中大数据技术的必要性**。通过**历史气象数据、海洋温度、风速、气压**等多源数据，结合**动态时间规整算法（DTW）等技术**，可以**更加精确地预测台风的路径和强度**，从而提高预警的准确性。**大数据技术**在这一背景下显得尤为关键，能够通过**机器学习**和**数据挖掘**对台风特征进行**多维度分析**，提升预测模型的**科学性**和**时效性**。

大数据驱动的台风研究和预测对现代社会有着**重要现实意义**：

①提高台风预测的准确性：基于**历史数据**和**先进算法**，如**动态时间规整（DTW）**和深度学习，可以更准确地**预测台风路径和强度**，降低误差，帮助决策者做出**及时反应**。

②防灾减灾的提前部署：准确的台风路径预测能帮助**政府和救援机构**提前部署防灾措施，减少因预测不准确带来的**资源浪费**及**潜在损失**。

③促进经济和基础设施的规划：台风预测对**物流、能源和建筑**等行业有着重要的**决策支撑**，帮助优化**资源调配**，减少**经济损失**。

基于**大数据分析的台风分类与预测**能够更好地应对**气候变化**背景下的**复杂气象系统**，有效提高预测的**科学性和社会响应能力**，助力**防灾减灾与可持续发展**。

1.2 目标任务

本次研究主要针对**台风的分类与路径预测**问题，结合大数据分析技术，通过**数据建模与算法优化**，解决赛题提出的三个问题：

问题 1：分析**台风特征参数**与**气象因素**的关系，构建**台风分类评价模型**，明确**台风类别划分标准**。预测 2024 年 **7 月和 9 月的台风类别**及其途经省份，并**区别夏台风与秋台风**。

问题 2：利用**动态时间规整算法（DTW）**，结合**气温、气压、季风**等因素，建立**台风路径预测模型**，并以 **2024 年第 13 号台风贝碧嘉**为例进行**路径对比分析**。

问题 3：根据**台风登陆后的强度衰减**，建立**降水量和风速变化模型**，验证并分析 **2024 年第 13 号台风**的数据。

二、模型假设及符号

2.1 模型假设

为了简化台风路径预测与分类模型的建立，在不影响模型精度的前提下，团队做出以下假设：

（1）假设台风路径中的观测点与台风中心的距离在**500 公里**内，并且观测点的日期记录与台风中心的日期为**同一天**，则可以认为该观测点的气象数据能够较好地反映台风在该位置的气象特征。

（2）假设台风主要影响的范围为**季风区**。

（3）假设**海表温度**是反映海洋整体温度最有效的变量。

（4）假设海表温度的**测量深度为 2 米**，足以反映海洋表层的热状态。

（5）假设**月度海表温度的变化幅度较小**，用**当月** 2 米深度的海表温度值来**代表该月所有**台风期间的海表温度情况。

（6）假设台风的实际路径为曲线，但由于观测数据仅包含两个经纬度坐标点，且时间间隔较短（每 6 小时记录一次），为简化计算，**假设两点间的台风路径为直线**。

2.2 主要的符号说明

符号	意义
T	台风强度
M	季风影响参数
P	气压(pa)
SST	海表温度(℃)
V	最大风速(m/s)
C_t	台风类别
$\boldsymbol{R}_t$	台风路径
RH	相对湿度
lat	纬度
lon	经度

三、问题分析

3.1 针对问题一

问题一以台风分类评价为基础，重点探讨台风特征参数（如强度、等级、风速等）与气温、气压、季风等气象因素的关系。**研究依托赛题组提供的多个权威数据集，以及从中国气象局爬取的气象数据**，获取台风路径、气温、降水、风速和海洋环境等完整信息，用于涵盖台风生成和移动过程中的重要特征。由于台风生命周期和持续时间各异，刻画台风的记录数据条数不一致，这会导致后续建模时维度不统一。因此，**为确保特征维度一致，需要对海量数据进行特征提取**，将台风的时序观测数据转化为一致性的特征向量，以便后续的分类评价模型能够有效处理这些数据。

在构建台风分类评价模型中，研究团队将**考虑多个分类视角，包括台风季节、方向和具体路径等**，搭建可以刻画这些分类视角的关键气象指标，如使用哈弗塞恩公式计算台风路径点之间的空间地理距离、气温最大变化率、气压最大变化率和路径转向角等。在此基础上，**团队选择层次聚类算法。这种方法可以通过树状图直观地展示不同类别的分层关系，**有助于识别具有相似特征的台风类别。算法采用完全链接法逐步合并特征相似的台风样本。这一过程最终形成分类树状图，揭示出台风类别的自然分布，从而为后续研究和应对策略提供数据支持。

为了深入构建台风的分类评价模型，在特征提取和层次聚类结果的基础上，**研究团队选择使用循环神经网络（RNN）来捕捉台风路径的短期和长期特征。**基于历史台风特征数据，训练这些模型学习不同类别台风的特征模式，从而在 2024 年 7 月和 9 月的台风特征基础上，对其分类预测，并列出各类别台风的途径省份。最终，基于模型的分类结果，研究团队需要分析夏季（6-8 月）与秋季（9-11 月）台风的特征差异，包括气温、台风强度、路径模式和降水等方面，以全面理解不同气象背景和海洋环境对台风行为模式的影响。**具体流程图见图 3.1。**

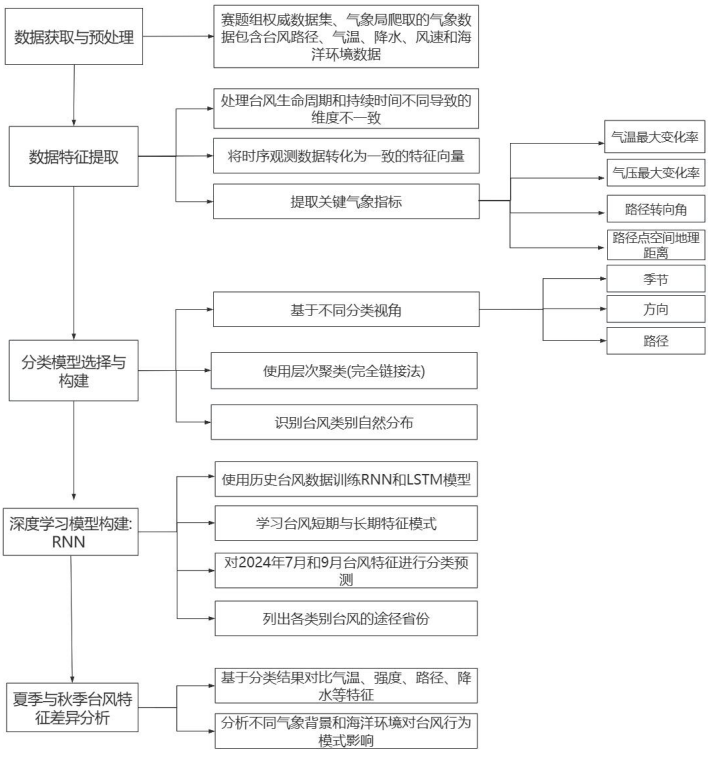


图 3.1 问题一模型流程图

3.2 针对问题二

针对台风路径预测需求，研究团队将探索气温、气压、洋流和风场等多重因素对台风路径的影响，以预测 2024 年 9 月 13 日至 17 日第 13 号台风贝碧嘉每天 14 时的中心位置。为了更精准地预测台风趋势，**团队选择了发生时期相近、强度相似和中心位置相近的历史台风数据作为训练样本。**同时，由于台风的生命周期和数据记录频率差异，导致数据记录数不一致，传统模型难以处理。为提高模型的鲁棒性，团队采用滑动窗口法，通过固定台风时长，将每条台风数据划分为多个训练样品，从而详细刻画台风的不同发展阶段。

团队基于已知的贝碧嘉在 9 月 10 日 14 时至 9 月 12 日 14 时的数据，**将滑动窗口长度设置为两天，以 6 小时为时间间隔，整合 8 组台风记录数据形成一组台风数据，从而有效捕捉台风路径的短期变化特征。**此外，选择经度、纬度、风速、气压和气温等核心气象指标，结合差值、对数变换、平方变换和比率等方法进一步挖掘数据中的时序变化特征，增强模型的预测能力。

由于台风路径的变化复杂且受多种因素影响，**长短期记忆网络（Long Short-Term Memory，LSTM）算法能够处理长时间依赖问题，从而适合处理台风路径的时序预测**。研究团队利用滑动窗口法截取的时序特征作为输入层，将后一天的台风特征设为输出层，以此训练 LSTM 算法。通过学习台风路径的动态变化，模型能够逐步预测下一时间段的经纬度、气温、气压和风速等关键特征，构建了一个完整的路径预测模型。

为了评估预测路径的准确性，研究团队将模型预测的路径与第 13 号台风贝碧嘉的实际路径进行对比，**并使用动态时间规整（Dynamic Time Warping ,DTW）算法来衡量两条时间序列之间的差异。**这种比较有助于分析模型在路径预测中的误差分布，并揭示台风路径预测中的关键影响因素。**具体流程图见图 3.2。**

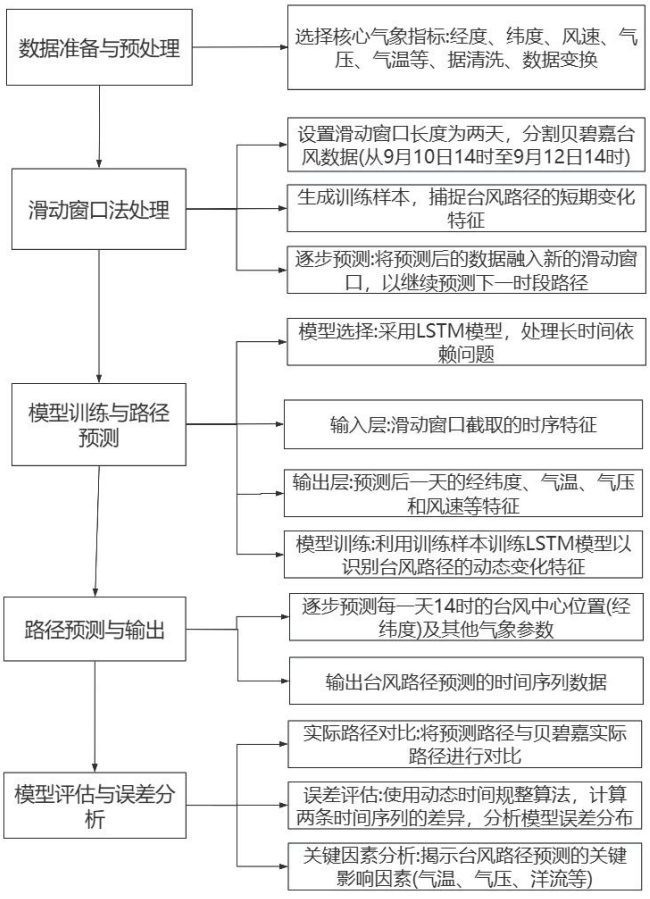


图 3.2 问题二模型流程图

3.3 针对问题三

针对台风登陆后风速和降水量衰减的预测需求，研究团队重点分析台风中心风速和降水量如何随着距离登陆点的增加而逐渐减弱。以 2024 年 9 月 16 日至 18 日第 13 号台风贝碧嘉为例，团队将建立风速与降水量的衰减关系模型，并通过实际数据进行分析。为准确刻画台风的降水量和中心风速，**团队挑选了发生时期相近、生命周期相似、登陆点相近的历史台风数据作为模型训练样本。**与问题二的预测方式类似，团队继续采用滑动窗口的特征提取方法，逐步预测接下来几天的降水量和风速数据。

与问题二不同的是，由于台风的降水量目前仅有按天统计的数据，团队将**提取台风登陆后四天每天 14 时的观测数据特征输入训练模型，以学习台风随着登陆距离增加而衰减的动态变化规律。**团队在问题二的特征基础上，丰富了台风中心与登陆点的距离、与台风降水量和风速关系密切的核心气象指标，包括台风中心风速、气温、距离登陆的距离，以及台风中心以 100 公里、200 公里、300 公里为半径的降水量等。这些特征全面反映了台风的降水量和风速变化，以及降水量与距台风中心距离的关系。

考虑到 LSTM 在处理长序列时可能面临的困难，尤其是早期信息的遗忘，**研究将 LSTM 与多重注意力机制相结合，使模型既能学习台风衰减的时序特征，又能聚焦于影响风速和降水量变化的关键特征。**最终，模型将逐步预测下一时间段的风速和降水量，从而模拟台风在行进过程中的衰减情况。

通过分析预测结果的时间序列趋势，包括风速和降水量的变化模式，团队还将识别和分析影响预测结果的关键气象指标，例如气温、气压、洋流和风场的变化对风速和降水量的影响及降水量与距台风中心距离的关系。**具体流程图见图 3.3。**

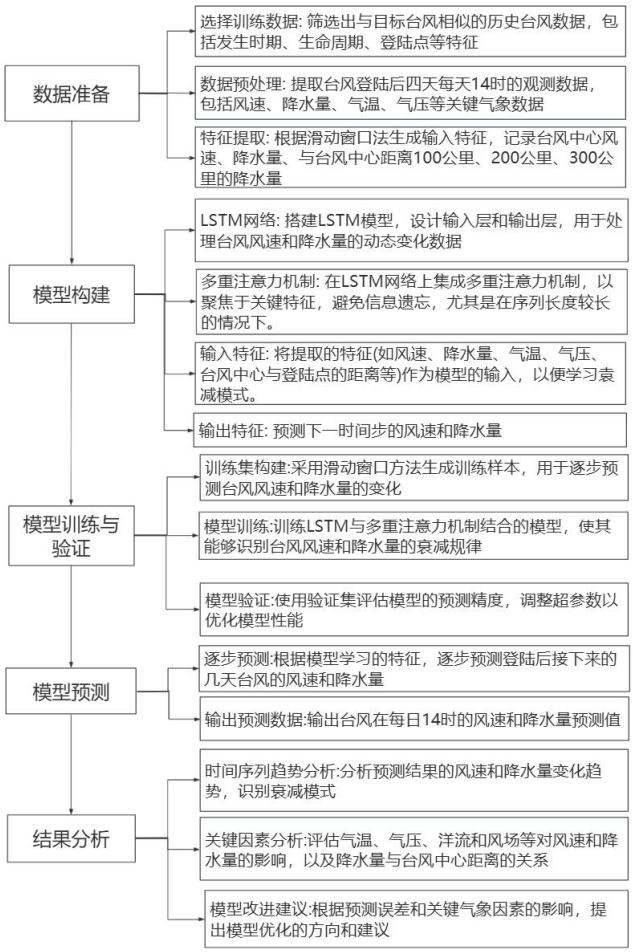


图 3.3 问题三模型流程图

四、问题一模型的建立与求解

4.1 数据预处理

4.1.1 数据来源

本研究**依托多个权威数据集**，以获取**台风路径、气象参数和海洋环境**等方面的完整信息。不同数据集包含**多样化的变量类型**，如气温、降水、风速等，涵盖了**台风生成和移动过程中的重要特征**。具体数据来源如下：

（1）中国近海台风路径集合(1945-2023 年)

该数据集由**海洋科学大数据中心**提供，大部分以 **6 小时为时间步长**记录，包含了**台风的起始时间、结束时间、观测时间、经纬度、台风等级、风速、气压、移动方向和速度**等关键指标。利用这一数据集可以支持台风路径预测和强度分类的历史数据分析，确保路径模型的准确性和科学性。

表 4.1 中国近海台风路径集合数据

台风编号	中文名称	英文名称	起始时间	结束时间	当前时间	经度	纬度	台风强度	台风等级	风速	气压	移动方向	移送速率
201820	西马仑	Cimaron	2018-08-18T20:00:00	2018-08-24T14:00:00	2018-08-23T08:00:00	135.2	30.1	强台风(STY)	14	45	950	北西北	40
201820	西马仑	Cimaron	2018-08-18T20:00:00	2018-08-24T14:00:00	2018-08-23T14:00:00	134.3	31.7	强台风(STY)	14	42	955	北	45
201820	西马仑	Cimaron	2018-08-18T20:00:00	2018-08-24T14:00:00	2018-08-23T17:00:00	134.5	32.6	强台风(STY)	14	42	955	北	51

（2）海洋温度数据 (air.2m.gauss)

由 Ocean and Climate 平台提供，记录了**全球海洋表面 2 米深度的气温变化，**是台风**生成和强度变化的重要影响因素**。海表温度能够**直接影响台风的能量获取和强度变化，**尤其在台风途经海域时至关重要。通过这一数据集，我们能够更好地**捕捉台风的热量来源，**为**路径预测和强度评估**提供关键气象支持。

表 4.2 海洋温度数据

时间	经度	纬度	海表 2m 温度
2018-08-18	88.542	65.625	273.6
2018-08-18	82.8508	101.25	273.425
2018-08-18	77.1394	127.5	274.15

（3）降水量数据集

来源于**中国气象局热带气旋资料中心，**记录了**多个观测站点在台风路径中的逐日降水量**。降水量是**台风强度和路径预测的关键变量，**对**台风降雨影响、路径复杂性和强度变化的精确分析**至关重要。

表 4.3 降水量数据集

热带气旋序列号	热带气旋数	观测点编号	最高降雨强度的次数	最大小时降水量(单位:毫米)	首次最大降雨强度开始和结束时间
201824	1800	57993	1	0.3	201808251800-201808251900
201824	1800	58362	1	0.8	201808250700-201808250800
201824	1800	58457	1	6.6	201808251200-201808251300

（4）2012-2024 年我国观测站点的逐日气象指标

为保证研究中台风路径与强度预测的准确性，团队通过 python **爬虫技术**，从**美国国家海洋和大气管理局（NOAA）**的数据库中，收集了 2012 年至 2024 年间**中国各地级市气象观测站点的逐日气象数据**。该数据集涵盖了每天各站点的**降水量、气温和露点温度**等关键气象指标。

表 4.4 2012-2024 年我国观测站点的逐日气象指标

站点编号	站点名称	日期	纬度	经度	气温	露点温度	降水量
45032099999	TA KWU LING, CH	2018/8/22	22.5333333	114.15	81.7	77.5	0.02
45032099999	TA KWU LING, CH	2018/8/23	22.5333333	114.15	81.6	78.1	2.74
45032099999	TA KWU LING, CH	2018/8/24	22.5333333	114.15	84.7	79.1	0.51

4.1.2 数据过滤

为了提高数据的相关性和模型预测的精度，我们对部分无关或无效的数据进行了剔除。以下是数据剔除的具体规则：

（1）远海台风数据过滤

远海台风通常不会对大陆区域产生直接影响，因此在数据预处理时设定距离中国大陆沿海线超过 **500 公里**的台风为“远海台风”，并将其过滤掉。

图 4.1 展示了这一筛选的效果：初始数据包含 **6327 条**记录（左侧），经过筛选后，数据条目减少至 **5635 条**（右侧），从而提高了数据对大陆影响的相关性。

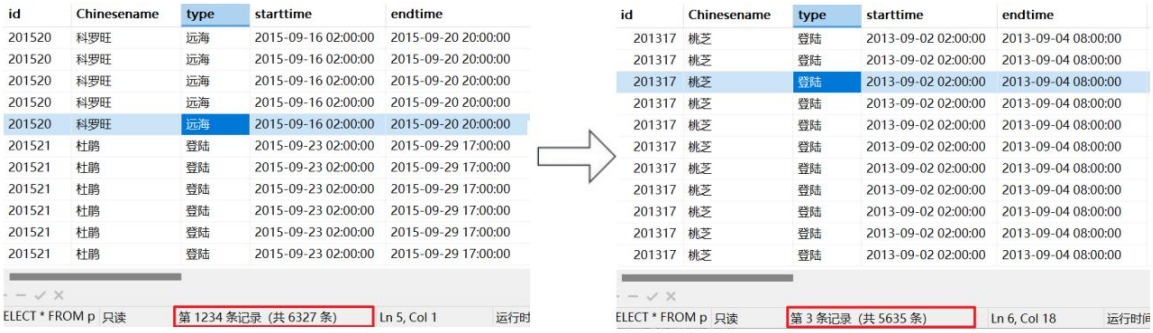


图 4.1 远海台风数据过滤前后对比

（2）降水量数据过滤与最大降雨量替换

在降水量数据处理过程中，我们执行了以下两个主要步骤：

①降水量过滤：对于距离台风中心超过 **400 公里**的降水观测点，将其降水量**设置为 0**，以排除非台风引起的降水干扰，减少噪音，提高模型精度。见图 4.2。

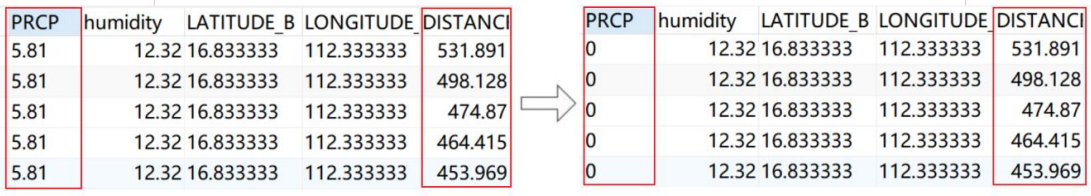


图 4.2 降水量数据处理前后对比

②最大降雨量替换：在完成“0 值”处理后，从“最大小时降雨量数据”中提取每个台风的最大降雨量时段，并与总表中的观测时间匹配。匹配到的记录将其降雨量替换为相应的**最大降雨量**数据，确保数据的代表性。

（3）时间无效数据剔除

为确保数据准确性，我们删除了中国近海台风路径集合中**观测时间早于台风起始时间**的记录。此类记录通常反映背景气象条件，与台风活动无关。清除这些数据后，数据集仅保留与台风实际生成和发展的相关信息。

starttime	endtime	nowtime	longitude
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 05:00:00	123.3
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 08:00:00	122.3
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 11:00:00	121.4
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 14:00:00	120.7
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 17:00:00	119.9
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 20:00:00	119
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-02 23:00:00	118.3
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-03 02:00:00	117.2
2018-01-03 08:00:00	2018-01-04 11:00:00	2018-01-03 05:00:00	116
2018-06-05 08:00:00	2018-06-09 05:00:00	2018-06-02 14:00:00	111.3

图 4.3 台风观测时间早于起始时间的数据

（4）数据时效性筛选

近年来，**全球气候变化**尤其明显，特别是**海表温度的持续上升**对台风生成与强度的影响显著。图 4.3 反映了近 80 年来特定海域 7、8 月平均海表温度的变化趋势，显示出温度整体呈上升趋势，特别是在近十年内多次达到峰值。这种变化为台风的生成和增强提供了更多能量支持[1]。因此，我们团队选择**近十年（2013-2022 年）**的数据不仅能够更好地反映当前气候变化对台风的影响，还能提升模型的时效性，使预测结果更加符合实际。

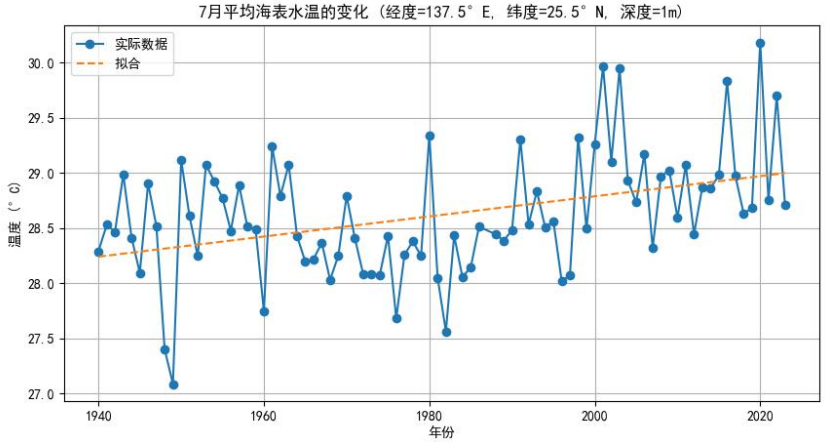


图 4.4 1940-2023 年 7 月平均海表温度变化趋势图（经度=137.5° E，纬度=25.5° N，深度=1m）

4.1.3 数据清洗与缺失值处理

（1）数据清洗

在数据筛选完成后，我们首先对数据中的异常值进行了清洗，确保模型输入数据的准确性。具体的清洗步骤包括：

①**降水量（PRCP）异常处理**：在降水量数据中，部分记录出现了无效的高值（如 99.99）。我们使用线性插值法对这些异常值进行了修正，以保持数据的连续性和平滑性。见图 4.5。

TEMP	PRCP	humidity	LATITUDE_B
27	99.99	8.53	32.5
27	99.99	8.53	32.5
27	99.99	8.53	32.5
27	99.99	8.53	32.5
27	99.99	8.53	32.5
27	99.99	8.53	32.5
27	99.99	8.53	32.5
27	99.99	8.53	32.5

temp	PRCP	humidity	latitude_B
27	10.24	8.53	32.5
27	10.24	8.53	32.5
27	10.24	8.53	32.5
27	10.24	8.53	32.5
27	4.746667	8.53	32.5
27	4.103333	8.53	32.5
27	0.4575	8.53	32.5
27	0.395	8.53	32.5

图 4.5 降水量异常值处理前后对比

②**统一温度单位**：为了确保数据的一致性，我们统一将不同数据源中的所有温度数据单位转换为摄氏度：

开尔文转摄氏度公式：$T_{Celsius} = T_{Kelvin} - 273.15$

华氏度转摄氏度公式：$T_{Celsius} = \frac{5}{9}(T_{Fah\ renh\ eit} - 32)$

通过这些转换，如图 4.6，我们确保了温度数据的标准化处理，从而避免了不同温度单位对模型训练的影响。

id	temp	humidity	prcp	air
201317	53.4	7.47	0.08	300.33
201317	53.4	7.47	0.08	300.33
201317	53.4	7.47	0.08	299.65
201317	53.4	7.47	0.08	299.65
201317	53.4	7.47	0.08	299.65
201317	53.4	7.47	0.08	299.65

id	temp	humidity	prcp	air
201317	25.22	9.27	0.04	27.33
201317	25.22	9.27	0.04	27.33
201317	25.22	9.27	0.04	26.65
201317	25.22	9.27	0.04	26.65
201317	25.22	9.27	0.04	26.65
201317	25.06	7.67	0	26.65

图 4.6 温度单位统一前后对比

（2）缺失值处理

在数据清洗完成后，我们对数据中的缺失值进行了处理，以确保数据的完整性和一致性。这一步对于后续模型的稳定性至关重要。具体步骤如下：

①**均值填充法**：对于少量随机缺失的气象数据，我们采用均值填充法。即使用同一区域、相邻时间段内的平均值填补缺失值。这种方法能够保持数据的完整性，避免因缺失值造成的数据不连续问题。

②**线性插值法**：对于时间序列数据中连续缺失的部分，采用线性插值法进行填补。通过计算缺失点前后时间点的数据，推算出合理的缺失值，确保数据的平滑性与一致性。

4.1.4 数据整合

数据整合是本研究关键步骤之一，旨在通过台风编号和时间戳等将多个数据源中的不同气象变量合并为统一的数据集。具体步骤如下：

（1）温度数据的整合

温度数据来自两个主要来源：**2012-2024 年我国观测站点的逐日气象指标**（用于陆地气温）和**海洋温度数据**（air.2m.gauss）（用于海洋气温）。在数据整合过程中，根据台风中心与观测站的距离，动态选择最适合的温度数据：

为**判断台风中心点相对于陆地和海洋的接近性**，我们使用**哈弗赛恩公式**（Haversine Formula）计算台风中心点到最近陆地与海洋观测点的最短距离，分别记录为 DISTANCE_B（距离陆地）和 DISTANCE_C（距离海洋）。

哈弗赛恩公式如下：

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\mu}{2}\right)}\right)$$

其中，ϕ 和 μ 分别为台风中心和观测点的纬度和经度，r 为地球半径。

当台风中心位置距离海洋观测点更近（即 DISTANCE_C 较小）时，选用海表 2 米深度的温度数据 air 作为该位置的温度值；当台风中心位置距离陆地观测站更近（即 DISTANCE_B 较小）时，则使用陆地观测站的温度数据 TEMP 作为该位置的温度值。图 4.7 展示了**原始数据**中的陆地与海洋温度，以及台风中心点与最近陆地和海洋观测点之间的距离。

TEMP	PRCP	humidity	LATITUDE_B	LONGITUDE_B	DISTANCE_B	air	LATITUDE_C	LONGITUDE_C	DISTANCE_C
91.4	0	2573353.64	35.66667	-117.83333	5883.177035	299.65	33.3328	176.25	52.153276
91.4	0	2573353.64	35.66667	-117.83333	5874.597929	299.65	33.3328	176.25	58.373913
94.1	0	2366802.38	35.66667	-117.83333	5857.433946	299.55	33.3328	176.25	72.789219
94.1	0	2366802.38	35.66667	-117.83333	5836.13448	299.55	33.3328	176.25	94.34528
53.402	0.08	7.47	52.783333	103.683333	3424.387129	300.33	25.7139	125.625	113.103245
53.402	0.08	7.47	52.783333	103.683333	3424.387129	300.33	25.7139	125.625	113.103245
53.402	0.08	7.47	52.783333	103.683333	3429.944643	300.33	25.7139	125.625	93.239608
53.402	0.08	7.47	52.783333	103.683333	3428.625386	300.33	25.7139	125.625	99.316513
53.402	0.08	7.47	52.783333	103.683333	3417.430627	299.65	27.6186	125.625	91.325822

图 4.7 温度数据整合前的记录状态

经过筛选和整合处理，最终将选择的温度值**统一存入 TEMP 列**中，如图 4.8 所示。这一步骤确保了所有位置的温度数据都经过合理选择，为后续分析提供了可靠的数据支撑。

longitude	latitude	province	strength	level	windvelocity	pressure	direction	movevelocity	TEMP	PRCP
129	17.1	Bicol	热带低压(TD)	7	15	1001	北东北	5	29.65	0
129	17.5	Bicol	热带低压(TD)	7	15	1001	北东北	5	29.7	0
129.6	17.3	Bicol	热带风暴(TS)	8	18	1000	北东北	14	29.7	0
129.7	17.5	Bicol	热带风暴(TS)	8	18	1000	北东北	15	29.7	0
130.1	17.9	Bicol	热带风暴(TS)	8	18	998	北东北	15	29.3	0
130.2	18.6	Bicol	热带风暴(TS)	8	18	998	北东北	15	29.3	0
131	19.3	Okinawa	热带风暴(TS)	8	20	995	东北	10	28.9	0

图 4.8 温度数据整合后的记录状态

（2）降水量数据的整合

降水量数据选取了两个来源：**2012-2024 年我国观测站点的逐日气象指标**和**降水量数据集**。优先选择降水量数据集中的最大降水值。如果该数据集中没有所需的记录，则使用逐日气象指标中的降水量数据。通过这种方式，确保了台风期间降水量的准确反映。

（3）湿度的计算

湿度数据并未直接提供，为了计算台风中的相对湿度（Relative Humidity, RH），我们使用了基于露点温度（dewp）和实际气温（temp）的计算公式。该公式如下：

$$RH = \frac{6.11 \times 10^{7.5 \times \frac{dewp}{237.3+dewp}}}{6.11 \times 10^{7.5 \times \frac{dewp}{237.3+temp}}}$$

通过这种方式，我们能够准确计算出各个时间点的湿度值。

（4）台风路径与气象数据的融合

以**台风编号、时间戳**作为主键，将中国近海台风路径集合数据与其他数据集进行匹配和合并，生成了一个包含完整气象信息的数据集。见图 4.9，表格展示了**台风编号、名称、时间、经纬度、强度、风速、气压、移动速度和方向**等核心指标的整合结果。

id	Chinesename	Englishname	starttime	endtime	nowtime	longitude	latitude
201317	桃芝	Toraji	2013-09-02 02:00:00	2013-09-04 08:00:00	2013-09-02 02:00:00	124.700000	26.300000
201317	桃芝	Toraji	2013-09-02 02:00:00	2013-09-04 08:00:00	2013-09-02 05:00:00	124.700000	26.300000
201317	桃芝	Toraji	2013-09-02 02:00:00	2013-09-04 08:00:00	2013-09-02 08:00:00	125.300000	26.500000
201317	桃芝	Toraji	2013-09-02 02:00:00	2013-09-04 08:00:00	2013-09-02 09:00:00	125.500000	26.600000
201317	桃芝	Toraji	2013-09-02 02:00:00	2013-09-04 08:00:00	2013-09-02 10:00:00	125.700000	26.800000

province	strength	level	windvelocity	pressure	direction	movevelocity	TEMP	PRCP	humidity
Okinawa	热带风暴(TS)	8	18.00	1000.00	东北	9.00	11.89	0.08	7.47
Okinawa	热带风暴(TS)	8	18.00	998.00	东北	9.00	11.89	0.08	7.47
Okinawa	热带风暴(TS)	8	18.00	998.00	东北	10.00	11.89	0.08	7.47
Okinawa	热带风暴(TS)	8	20.00	995.00	东北	10.00	11.89	0.08	7.47
Okinawa	热带风暴(TS)	8	20.00	995.00	东北	10.00	11.89	0.08	7.47

图 4.9 数据整合后的数据集展示

4.2 特征提取

4.2.1 原始数据局限与特征提取的必要性

（1）原始数据局限性

在台风路径预测中，直接使用**原始数据存在诸多挑战**，主要原因是台风的**持续时间不一致**，导致每个台风的**观测记录数量不同**。台风路径数据通常每隔六小时记录一次，但由于台风的**生命周期差异大**，记录的时间点数量会有所不同。例如，某个台风可能有 20 条记录，而另一个台风可能只有 10 条。这种数据不一致性在后续分析中产生了以下几方面的问题：

①**输入数据维度不一致**：机器学习模型通常要求输入数据维度相同。台风生命周期不同，观测记录的数量不同，导致无法统一输入模型，影响训练效果。

②**路径比较复杂化**：台风持续时间不同，直接比较其路径存在困难。例如，一个台风持续 48 小时，另一个持续 72 小时，路径时序数据无法对齐，难以准确地进行分析和预测。

因此，为了**应对原始数据在时序长度和数据维度上的不一致问题**，我们采用**特征工程方法**，将台风的时序数据转换为具有一致性特征的向量，以保证输入数据的统一性和模型的预测能力。

（2）特征提取的必要性

特征工程的核心任务在于从原始数据中**提取出对台风分类和路径预测起关键作用的变量**，通过规范化处理，将台风的时序观测数据**转化为同维度的特征向量**，确保模型能够处理格式一致的数据。特征提取在此过程中的**作用不可或缺**。

首先，通过**统一时间步的特征提取**，可以生成维度一致的特征向量，使模型训练变得可行。其次，**特征提取能够有效捕捉台风路径中的关键变化模式**，例如通过差值、对数和平方等变换来**突出变量的短期变化和趋势**，增强模型在复杂气象条件下的预测能力。

此外，特征提取还具备**降维和减少计算复杂度的优势**。在保留路径预测关键特征的前提下，特征提取**有效降低了数据冗余**，缩短了模型训练时间，减小了过拟合风险。这一过程使模型在不同台风样本上获得**一致性更好的预测表现**，提高了台风路径预测的**准确性。**

4.2.2 搭建特征空间

在**台风路径预测的特征空间构建**过程中，**气温、气压和季风**等关键气象因素对台风路径的变化和强度有着直接影响。因此，针对这三方面的**特征提取**，可以为模型提供全面的预测输入[2]。具体包括以下几个方面：

（1）从气温角度

气温，尤其是**海表温度**，对台风的形成和路径有重要影响。**热带气旋**需要**高温的海水**来提供能量，海温的变化也直接影响台风的**强度和行进方向**。因此，团队提取了**气温的最大变化率**，以便衡量在台风生命期内**温度的波动幅度**。此特征不仅捕捉到了台风所处环境的热力变化，也反映了台风可能经历的**强度调整过程**。

（2）从气压角度

气压是衡量**台风强度的核心指标**之一，**气压越低**通常意味着**台风越强**。为了捕捉气压对台风路径的影响，团队提取了**气压最大变化率**，这一特征体现了气压在台风生命周期中的**剧烈变化速率**。气压的大幅波动往往伴随着台风强度的变化，这种变化不仅影响台风的路径方向，还会影响路径的曲折程度。

路径曲度与气压关系：通过计算**台风实际路径长度与直线距离的比率**，我们能够更好地量化气压变化对路径曲度的影响。**气压的急剧变化**通常意味着台风可能经历了**强烈的路径转向和风速变化**。此特征不仅有助于模型理解**气压与路径偏移的关系**，还能揭示台风在**高强度阶段的路径行为模式**。

（3）从季风角度

季风是影响台风路径的主要**大气动力**之一，**强季风气流**会将台风引导至特定方向。因此，团队通过**多个方向角的特征提取**（如 **4.1 方向角至 4.4 方向角**）来捕捉季风对台风路径的引导作用。这些**方向角特征**测量了台风路径的**转向趋势和变化率**，反映了季风对台风路径的**分阶段影响**。

一阶纬度和经度方差：台风在受到季风引导时，其路径的**纬度和经度变化**会表现出特定的波动模式。通过提取一阶纬度和经度方差，我们能够**反映路径波动的幅度**，揭示季风在不同阶段对台风路径的**调节作用**。例如，季风引导下的路径偏移会使台风朝特定方向转向，而经纬度方差能够反映这种变化的**累积效应**。

通过以上特征的提取，我们能够在**气温、气压和季风**的多重影响下**构建具有代表性的特征空间**。这种特征空间为**台风路径的聚类分析和分类模型**提供了**标准化的输入数据**，增强了**模型在不同台风样本上的泛化能力**。

4.2.3 关键特征的提取

基于**台风路径、强度及相关气象因素**的分析，本研究提取了具有代表性的**关键特征**，如表 4.1 所示，以便更好地支持路径预测和分类模型的构建：

表 4.5 台风特征参数、定义及其计算公式

特征	定义	公式
最大转向角	台风每段路径相对起始角度的最大偏转角度，衡量台风路径的突然转折情况。	$\theta = \arccos\left(\frac{\overrightarrow{AA'}\cdot\overrightarrow{BB'}}{\lvert\overrightarrow{AA'}\rvert\cdot\lvert\overrightarrow{BB'}\rvert}\right)$
路径转角波动	台风路径方向变化的波动性，即转向角的方差，反映台风路径的曲折程度。	$S^2 = \frac{\sum_{i=1}^{n}(\theta_i-\theta)^2}{n}$
气温最大变化率	（陆地的近地面气温，海洋的近海面气温）最大变化率	$\mathrm{Max}\left(\frac{\Delta T}{\Delta t}\right)$
气压最大变化率	台风生命周期内气压的最大变化速率，反映台风强度变化。	$\mathrm{Max}\left(\frac{\Delta P}{\Delta t}\right)$
路径长度与直线距离的比率	台风实际路径长度与起点到终点的直线距离之比，反映台风路径的曲折程度。	$\frac{s}{S}$
最大降雨量	台风生命期内观测到的最大降雨量，反映了台风携带的水汽含量。	$\mathrm{Max}(R_i)$

最大湿度	台风生命周期内的最大湿度，衡量水汽含量对台风强度的影响。	$\mathrm{Max}(\mathrm{R}H_i)$
最大风速	台风过程中观测到的最大风速，反映台风强度	$\mathrm{Max}(v_i)$
生命周期时长	台风从生成到消失的总持续时间，衡量台风的生命周期长度。	$T_{末} - T_{初}$
一阶纬度平方和	台风在每个观测时间点的纬度变化量的平方和，反映纬度的变化幅度对台风路径的影响。	$\sum_{i=1}^{n} (\Delta LAT_i)^2$
一阶经度平方和	台风在每个观测时间点的经度变化量的平方和，反映经度的变化幅度对台风路径的影响。	$\sum_{i=1}^{n} (\Delta LON_i)^2$
平均航向角	台风路径上每段轨迹的航向角度的平均值，衡量台风路径的整体方向趋势。	$\theta_{avg} = \frac{1}{n}\sum_{i=1}^{n} \theta_i$
4.1 方向角	台风路径的第一段方向角，计算台风路径起始至第一四分段的方向。	$(\arctan2(\sin(\mathrm{lon}_2 - \mathrm{lon}_1) \cdot \cos(\mathrm{lat}_2), \cos(\mathrm{lat}_1) \cdot \sin(\mathrm{lat}_2) - \sin(\mathrm{lat}_1) \cdot \cos(\mathrm{lat}_2) \cdot \cos(\mathrm{lon}_2 - \mathrm{lon}_1)) \cdot \frac{180}{\pi} + 360) \bmod 360$
4.2 方向角	台风路径的第二段方向角，计算台风路径第一四分段至第二四分段的方向。	
4.3 方向角	台风路径的第三段方向角，计算台风路径第二四分段至第三四分段的方向。	
4.4 方向角	台风路径的第四段方向角，计算台风路径第三四分段至路径终点的方向。	

通过这些特征提取，能够有效概括台风在不同时间点的路径和强度变化特征，为后续的聚类分析和分类模型提供标准化输入数据。

4.3 基于层次聚类的分类模型

在特征提取完成后，我们应用聚类算法对台风数据进行分析，以识别台风的不同类别特征。由于类别标签和数量不清晰，因此选择**层次聚类算法**[3]，它能够通过对特征空间的分析自动生成类簇。

4.3.1 特征选择与数据标准化处理

在**台风特征提取**过程中，我们选取了 **16 项特征**，这些特征能够较为全面地反映台风的**生命周期、气象条件和运动特性**。数据集包含多个时间步长的台风记录，因此需要对每个台风实例按 ID 进行分组，并计算每个特征的**平均值**，得到一个最终数据集。完成分组后，对提取的特征数据进行标准化，去除特征间**尺度差异**，以适合聚类分析。

4.3.2 层次聚类模型的构建

凝聚层次聚类（Agglomerative Hierarchical Clustering）是一种自底向上的方法，最初将每个数据点视为独立的簇，然后逐步合并两个最相似的簇，迭代此过程，直至所有数据点合并为一个簇。该方法能够生成**树状图**（dendrogram），直观展示数据的**分层关系**，与**分裂型层次聚类**相比，凝聚层次聚类在**计算复杂度**上更具优势[4]。

①构建距离矩阵：通过**哈弗塞恩**(Haversine)公式计算样本点之间的**实际经纬度距离矩阵**，作为层次聚类的基础。

②聚类链接方式：使用**完全链接**（Complete Linkage）的方式，定义两个簇之间的**距离为两个簇中任意两点之间的最长距离**。这种方法倾向于生成**紧凑的球形集群**，对异常值和噪声不太敏感，更注重簇的**整体分离度**。在针对于**台风分类**中，完全链接可能更适合，因为它可以减少异常路径对聚类结果的影响，从而提高分类的**稳定性和准确性**。

③树状图构建：依据层次合并结果，分别绘制**夏、秋台风的分类树状图**

（dendrogram），用于直观展示聚类过程及各类间的**距离关系**。

4.3.3 模型参数与阈值设置

根据**夏秋台风分类树状图**，如图 4.10-4.11。我们分别将**阈值**分别定为 **9.1 和 8.0**，在该阈值下的树状图分枝较为清晰，分别形成了 **10 个、11 个聚类类别**。阈值的确定基于**分层距离**的观察，9.1，8.0 的阈值选择保证了**类间的差异性**，同时将**台风数据**分为适量的类别以便于**后续分析**。

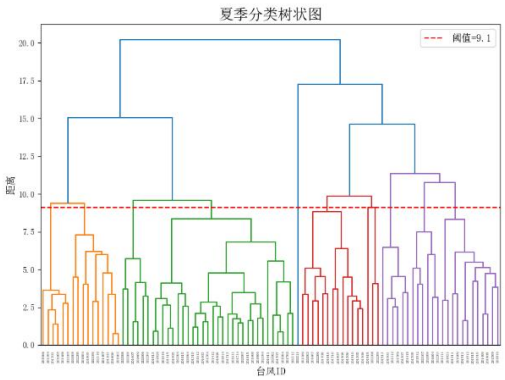
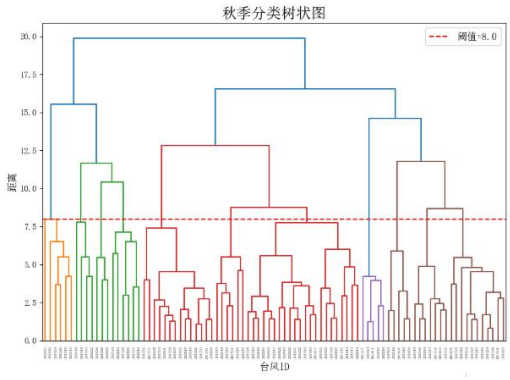


图 4.10 夏季台风分类树状图



4.11 秋季台风分类树状图

4.3.4 台风路径类型划分

在夏秋台风的层次聚类分析之后，本研究依据台风的**路径形态和移动方向**，将台风路径划分为 **21 种**典型类型。这些分类不仅能够展示台风在路径上的多样性，同时也为路径预测模型提供了清晰的分类标准。以下是台风路径类型的详细分类说明。

（1）夏季台风路径类型

夏季台风的活动较为频繁，路径特征复杂多变。本研究通过对夏季台风数据的聚类分析，总结出表 4.6 所示典型的路径类型：

表 4.6 夏季台风路径类型及定义

台风类型	定义
近海西北-东北转向型	这类台风在近海区域首先向西北方向移动，随后转向东北，呈现出明显的路径转向特征。
近海东北直线型	路径呈直线形态，从近海向东北方向移动，基本保持稳定的移动方向。
近海北直线型	路径从近海向北方移动，直线型的路径显示出较为稳定的北向趋势。
登陆西-西北转向型	这类台风在登陆后从西向西北转向，通常在陆地上呈现出一定的偏移。
近海西-东北回旋型	路径在近海区域先向西移动，随后回旋至东北方向，呈现出明显的回旋特征。
登陆西北-东北转向型	在登陆后，台风从西北方向转向东北方向，显示出转向路径。
登陆西北波动型	登陆后，主要沿西北方向移动，但路径上伴有明显的波动性，呈现出多次小幅度转向或曲折。
登陆西北-西南回旋型	这类台风在登陆后从西北方向回旋至西南方向，路径呈现出回旋的特点。
近海“L”型转向型	路径在近海区域形成“L”型，通常表示先向一个方向移动，随后转向形成直角形路径。

（2）秋季台风路径类型

秋季台风相较夏季台风，活动路径更为多样，常受季风、海洋洋流和地理环境的复杂作用影响。聚类分析得出如表 4.7 所示秋季台风的典型路径类型：

表 4.7 秋季台风路径类型及定义

台风类型	定义
“鱼钩型”路径	路径形似鱼钩，在区域呈现出明显的曲折。
登陆西北“蛇形”路径	路径曲折变化，形似蛇形，通常在西北方向上呈现出较大的波动。
近海“S”型路径	路径形成“S”形状，表示台风在移动过程中出现多次方向变化。
近海“上凹”路径	路径呈现出上凹形，路径整体向上凸起，然后下弯，形成独特的曲线。
近海西直线型	秋季近海台风路径直线向西移动，路径变化不大。
登陆西北波动型	台风在登陆后沿西北方向呈波动移动，路径具有一定的曲折性。
登陆西北-西转向型	在登陆后最初沿西北方向移动，随后路径逐渐转向西方。
近海东北直线型	路径从近海向东北方向移动，保持直线形态，移动方向未发生显著变化。
近海弓形路径	路径呈弓形，表示台风移动过程中受到外力影响，形成弧形曲线。
近海西北-东北转向型	路径在近海区域从西北方向向东北方向转向。
近海撬棍型路径	路径呈现撬棍状，多次发生方向变动，路径复杂且不规则。

（3）典型路径展示

考虑到论文篇幅限制，本文将选择几类具有代表性的台风路径类型进行展示，以说明不同类型台风在路径特征上的差异。如图 4.12、图 4.13 和图 4.14 所示。

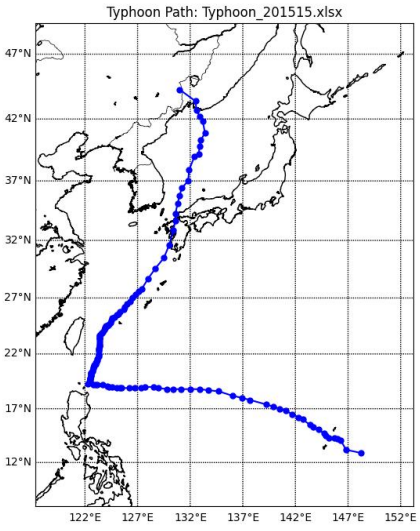


图 4.12 夏季近海“L”型转向台风

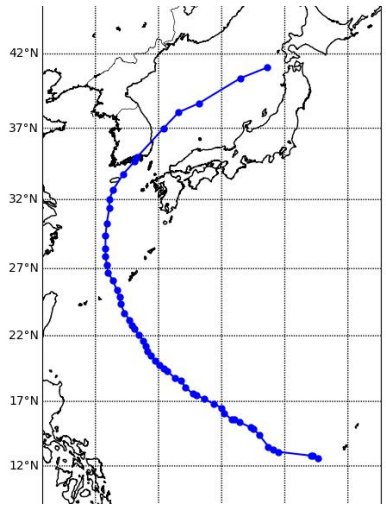
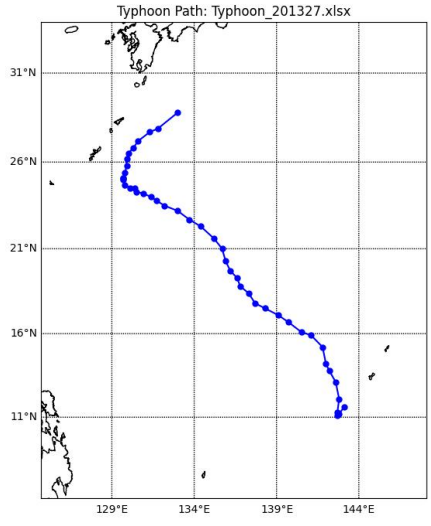


图 4.13 秋季近海弓型台风图



4.14 秋季近海撬棍型台风

4.3.5 模型评估

（1）针对夏台风聚类模型

Calinski-Harabaz 指数（CH）：通过计算**簇内方差和与簇间方差的比值**来评估聚类效果，值越大表示簇内越紧凑且簇间越分离，聚类分析得到的指数为 **10.377785518789183。**

轮廓系数（Silhouette Coefficient）：度量**聚类结果的紧密性和分离性**，经过分析得到轮廓系数为 **0.9701509843108841**，展现较好的聚类效果。

Davies-Bouldin Index（DB）：综合考虑类内样本的**相似度**以及类间样本的**差异度**，通过计算类内样本到其**聚类中心的平均距离**与不同聚类中心之间的**距离之比**来评估聚类效果，聚类分析得到的指数为 **0.1384741965735143。**

Dunn 指数（Dunn’s indices，简称 DVI）：通过比较**聚类内部的紧密度**和**聚类之间的间距**来评估聚类效果，模型的 Dunn 指数为 **0.9772564418327485**，体现聚类内部的**紧密度高**，聚类之间的**间距较大**，聚类质量较好。如图 4.15。

```
Calinski-Harabasz Index (CH): 10.377785518789183
Silhouette Coefficient: 0.9701509843108841
Davies-Bouldin Index (DB): 0.1384741965735143
Dunn Index: 0.9772564418327485
```

图 4.15 夏台风聚类模型评价指标

（2）针对秋台风聚类模型

CH 指数：11.84662976321819　　轮廓系数：0.9658314569172365

DB 指数：0.1195347762187334　　Dunn 指数：0.9725734126784659

```
Calinski-Harabasz Index (CH): 11.84662976321819
Silhouette Coefficient: 0.9658314569172365
Davies-Bouldin Index (DB): 0.1195347762187334
Dunn Index: 0.9725734126784659
```

图 4.16 秋台风聚类模型评价指标

4.3.6 结果展示与分析

为了更好地理解**台风聚类结果**，我们对聚类结果进行了**可视化分析**，如图 4.14-4.15。**聚类特征分布—热力图(Heatmap)**：对每个**类别的特征均值**进行分析，以观察**不同类别**在生命周期、最大风速、降雨量等**特征上的差异**。结果表明，不同类别的台风在生命周期长度、最大风速、最大湿度等各方面**存在显著差异**，有助于针对不同台风类型进行灾害预防和应对策略的定制。

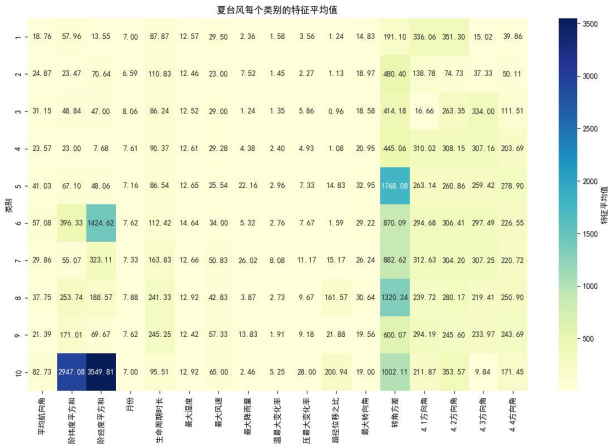


图 4.14 夏台风每个类别的特征均值

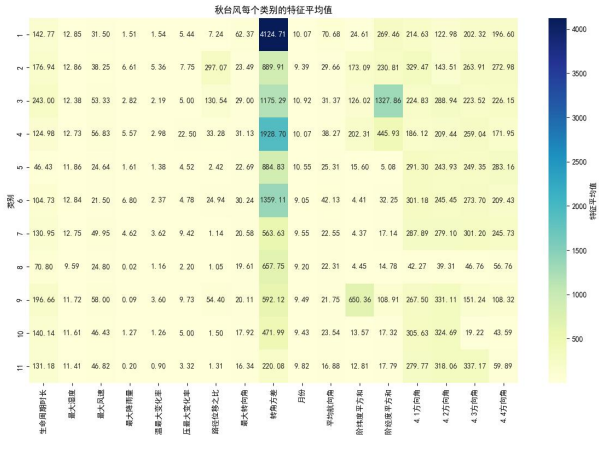


图 4.15 秋台风每个类别的特征均值

4.4 基于循环神经网络的台风数据分类评价

4.4.1 模型选择与架构

基于**台风特征参数的时序变化特征**以及**气温、气压、季风**等多重因素的影响，本研究选择了**循环神经网络（RNN）**作为分类模型[5]。

RNN：RNN 通过内部循环机制在隐藏层中建立了前后时间步的数据依赖，能够有效地捕捉台风路径在**前后**的动态变化，保持长期依赖性。

4.4.2 模型训练与验证

为确保分类模型的精确性和稳定性，研究将数据集划分为 **80%训练集**和 **20%测试集**，并采用**五折交叉验证**进行模型验证和优化，以**减少过拟合风险**并**提高模型的泛化能力**[6]。训练通过**误差反向传播**（Backpropagation Through Time, BPTT）算法完成，以最小化分类损失。训练与验证过程如下：

①**数据预处理**：根据表 4.1 中的 16 项特征，将**台风的时序观测数据转化为标准化的特征向量**，确保所有**数据维度的一致性**。

②**模型训练**：在训练集上，通过输入时序特征数据，**让 RNN 分别学习台风路径的动态特征**。模型逐步建立对**不同气象条件下台风路径变化的识别能力**。

前向传播（Forward Propagation）：

输入数据通过网络的每一层，直到最后一层（输出层）。

反向传播（Backpropagation）：

利用**链式法则**，从输出层开始反向通过网络的每一层，计算损失函数对每个参数的梯度，从而更新网络中的权重和偏置。

③**模型验证**：在测试集上验证**模型性能**，衡量在不同**类别台风上的预测精度**，以评估模型在未见台风数据上的表现。

④**模型优化**：通过**五折交叉验证**，调整**超参数**（如神经元数量、学习率等）以优化模型表现，确保在**不同台风类别上的分类准确性**。

4.4.3 分类规则与评估

本研究基于**层次聚类**方法，将台风分为多个**类别**（依据**季节**、**路径模式**、**移动方向**等维度）。**RNN 模型**通过学习这些类别的**特征模式**，能够对新的台风进行**分类预测**。**五折交叉验证**的引入，有助于确保模型在**季节变化（夏台风和秋台风）以及路径模式和强度条件下的稳健性，**该模型的构建可为气象灾害监测和应急响应提供支持。

4.5 夏台风与秋台风的对比分析

4.5.1 结果展示

通过模型分析，2024 年 7 月的夏台风主要影响了**台湾、福建、江西和湖北**等东南沿海和南部内陆省份。该月的夏台风表现出较高的强度和明显的降水特征，尤其是在途经台湾和福建时，降水强度尤为显著。相比之下，2024 年 9 月的秋台风主要影响了**海南、上海、江苏、安徽和河南**等省市，路径相对稳定，降水量较为温和但持续时间较长。这种差异可能与当年气温的异常偏高有关。由于气温升高，秋季台风在路径和强度上表现出更接近夏季台风的特征，其影响范围也更为广泛。这种夏季台风的“北移”趋势和秋季台风的“强化”现象，为灾害防范提供了重要的预测依据。

表 4.8 2024 年我国 7 月和 9 月台风类别及途经省份

月份	台风类别	途径省份
7 月 夏台风	**夏季登陆西北波动型台风**	**台湾省、福建省、江西省、湖北省**
	夏季登陆西-西北转向台风	**海南省**
9 月 秋台风	**秋季登陆西北-西转向台风**	**海南省**
	秋季登陆西北“蛇形”台风	**上海市、江苏省、安徽省、河南省**

	秋季“鱼钩型”台风	上海市、江苏省
	秋季近海“上凹”型台风	台湾省

4.5.2 差异分析

（1）台风强度方面

分析模型结果表明，夏季台风的强度普遍较高，尤其是近海回旋和波动型台风，表现出更强的风速和能量。这与夏季海表温度较高密切相关，海洋热能为台风提供了额外的能量支持。相比之下，秋季台风的强度较弱，多数表现为转向和直线路径。然而，由于2024年气温异常偏高，一些秋季台风表现出更强的强度，部分秋台风接近夏台风的强度水平。

（2）路径与方向方面

夏季台风多向北移动，路径复杂，回旋和转向台风较为常见；秋季台风则多沿西北方向移动，路径相对平稳，以直线型为主。这种路径差异与季风的强弱变化息息相关，而今年气温的异常偏高，使得秋季台风的路径更接近夏季模式，路径复杂度有所增加。

（3）降水与湿度方面

夏台风的降水量丰富、集中，尤其是在近海回旋和转向路径中，湿度也相对较高。秋台风降水量较小，但持续时间长，且湿度相对较低。今年因气温偏高，秋台风的湿度和降水量也有所增加，进一步表明气温对台风降水特征的影响。

4.5.3 结果应用与分析

基于分类评价模型的分析结果，可以总结出夏季台风与秋季台风在路径、强度和降水方面的显著差异。2024年气温偏高的情况下，秋季台风逐渐表现出类似夏季台风的特征。这提示我们在未来的防灾减灾中，应更加关注气温升高对台风路径和强度的潜在影响，为灾害预测和防范提供科学支持，从而提升防灾减灾的有效性。

五、 问题二模型的建立与求解

5.1 台风的数据融合

5.1.1 台风数据筛选

（1）研究范围的确定

鉴于近年来全球变暖加剧，尤其是**2024 年温度异常偏高**，我们将传统的**秋季台风数据范围从 9 月-11 月延伸至 8 月-10 月**，包含夏末和初秋的台风数据。通常情况下，8 月的数据被划分为夏台风数据集。然而，为更全面地捕捉秋季台风的生成特征及其路径变化，本研究特别将 **8 月纳入考量**，以便更好地反映在**全球气候变暖背景下台风活动的潜在变化趋势**，尽管我们的目标台风——台风贝碧嘉发生于 9 月中旬。

筛选出的数据涵盖 **1945 年至 2022 年间 8 月-10 月**的台风记录。

（2）台风等级筛选

为聚焦于对大陆具有显著影响的**高强度台风**，我们筛选了**台风等级在 10 至 18 级的数据**。较低等级的台风通常影响范围较小且对路径预测的参考价值有限，而**贝碧嘉台风的强度较高**，因此筛选出同样强度较大的台风能够增强模型对高等级台风路径的预测效果，并确保数据的代表性。

（3）登陆台风筛选

本研究专注于可能对大陆造成实际影响的台风路径预测，因此**仅保留登陆台风的数据**。在数据筛选过程中，我们应用了与问题一中一致的标准，对**远海和近海台风**进行了预处理，以确保数据筛选的一致性和严谨性。

（4）生命周期筛选

我们筛选出**生命周期大于或等于 8 天的台风**，以确保数据的连续性和充足的观测时间点。该筛选条件的设定主要基于**贝碧嘉台风的生命周期较长**，且我们需要进行 **9 月 13 日至 17 日的路径预测**。**长生命周期的台风数据**能够为模型提供足够的时序信息，捕捉台风在不同时间段的路径变化，提升模型的预测准确性。

（5）方向角筛选

台风的移动方向是路径预测的重要因素之一。为使数据集的台风路径特征与贝碧嘉相似，我们选择了**方向角在 260° 至 350°** 之间的台风数据。这一区间的方向角更贴近贝碧嘉的实际走向，有助于模型学习和预测在**类似移动方向上的台风路径**，提升路径预测的精度。

基于以上筛选条件，我们构建了一个符合研究需求的高质量台风数据集，如图 5.1 所示。这些数据为后续的台风路径预测模型奠定了坚实基础，并能够更准确地应用于台风贝碧嘉的路径预测分析。

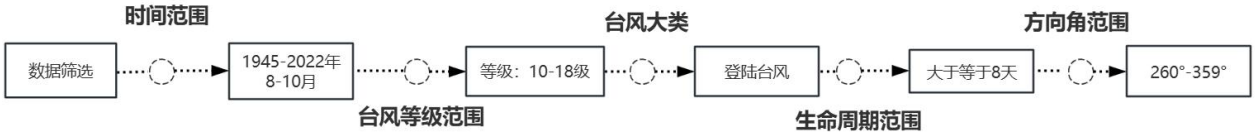


图 5.1 台风数据筛选条件

5.1.2 台风数据融合

在问题二的台风数据融合过程中，**整体方法和问题一中所用的步骤基本一致**。整合后的数据集如图 5.2 所示，包含**台风编号、名称、观测时间、经纬度、强度、风速、气压、移动速度及方向**等核心指标，为**路径预测提供了全面且一致的数据支持**。

id	starttime	duringtime	longitude	latitude	windvelocity	pressure	temp
195025	1950/9/7 14:00:00	1950-09-07 14:00:00	143.6	17.5	30	983	27.97
195025	1950/9/7 14:00:00	1950-09-07 20:00:00	143.5	17.9	35	980	27.97
195025	1950/9/7 14:00:00	1950-09-08 02:00:00	143.4	18.4	35	980	28.17
195025	1950/9/7 14:00:00	1950-09-08 08:00:00	143.1	18.9	40	975	28.17
195025	1950/9/8 14:00:00	1950-09-08 14:00:00	142.9	19.4	40	970	28.05
195025	1950/9/8 14:00:00	1950-09-08 20:00:00	142.7	20.0	45	965	28.05
195025	1950/9/8 14:00:00	1950-09-09 02:00:00	142.3	20.7	50	960	28.52
195025	1950/9/8 14:00:00	1950-09-09 08:00:00	141.9	21.5	50	955	28.27
195025	1950/9/9 14:00:00	1950-09-09 14:00:00	141.4	22.2	55	950	27.95
195025	1950/9/9 14:00:00	1950-09-09 20:00:00	140.9	22.8	55	950	27.95
195025	1950/9/9 14:00:00	1950-09-10 02:00:00	140.1	23.7	55	950	28.82
195025	1950/9/9 14:00:00	1950-09-10 08:00:00	139.5	24.6	60	945	28.27
195025	1950/9/10 14:00:00	1950-09-10 14:00:00	138.9	25.2	60	945	27.92

elect * from answeritems　只读　查询时间: 0.016s　第 2 条记录 (共 1460 条)

图 5.2 台风数据融合

5.2 数据预处理

在数据预处理阶段，我们进行了以下几项操作，以确保**输入数据的质量和一致性**，为台风路径预测模型提供可靠的基础：

（1）删除风速为 0 的记录

由于**风速为 0 通常不符合台风活动的特征**，可能是**数据异常或记录误差**，因此我们剔除了所有风速为 0 的观测记录，以确保**数据的准确性和代表性**。见图 5.3。

longitude	latitude	strength	level	windve	pressure
139.9	5.7		0	0	1006
139.3	5.9		0	0	1006
138.7	6.3		0	0	1006
138	6.7		0	0	1006
137.3	7.2		0	0	1005
136.7	7.7		0	0	1005
136	8.2		0	0	1005
135.3	8.8		0	0	1005

图 5.3 风速为 0 数据显示

（2）筛选统一的观测时间点

台风路径数据通常以 **6 小时为步长**记录，为确保**模型输入的统一性和时序一致性**，我们选择了**每日的 2 时、8 时、14 时和 20 时数据**。这些时间节点均匀分布，能够涵盖台风路径中的不同时段特征，如**温度、风速、气压**等，从而为路径预测模型提供**稳定的输入序列**。

（3）删除无效观测时间记录

为避免数据偏差，我们**删除了所有观测时间早于台风起始时间的记录**，我们应用了与问题一中一致的处理方法。

5.3 搭建特征空间

5.3.1 滑动窗口的建立

在台风路径预测模型的构建中，**滑动窗口法被用于处理台风生命周期长度不一带来的数据维度差异问题**，确保**模型输入的统一性**。由于台风的生成和消亡时间差异较大，不同台风的观测数据包含的时间步数量不一，造成输入数据的维度不一致，而**机器学习模型通**

常要求输入数据维度统一。因此，为了实现**有效的时序建模和规范化特征提取**，我们引入了滑动窗口法。

（1）滑动窗口的目的

滑动窗口法的核心在于**构建统一的输入特征维度**，以满足模型训练需求，同时使得**模型能够捕捉到台风路径的时序动态特征**。通过设置一个固定大小的滑动窗口，我们可以将台风路径数据按时间段**分割为连续的小样本**。具体而言，**每个滑动窗口包含两天的数据**（每 6 小时取一次数据，共 8 个时间点），这些样本作为模型输入。滑动窗口的长度设计为两天，主要考虑到**台风路径和气象特征的短期变化**有助于捕捉台风路径的连续性，而过长的窗口可能增加输入特征的冗余性，从而**降低模型的学习效率**。

（2）滑动窗口的构建与应用

在滑动窗口法应用中，每次滑动窗口以**过去两天的数据作为输入层的特征，用于预测下一天的目标变量。**每个输入样本包括过去两天的**经度、纬度、气温、气压和风速**等关键气象特征的时间序列信息，使得**模型能够学习台风路径随时间变化的动态趋势**。具体步骤如下：

①**构建输入样本**：每次滑动窗口从当前时间点开始，**向前取两天的数据**，这些数据点包括**经度、纬度、气温、气压、风速**等关键特征。对于每个时间步，还**计算前一时间步的变化特征**，如**差值、对数变化和平方变化**，以捕捉特征变化的复杂性。

②**生成输出样本**：在每次滑动后，**窗口的输出为第三天的经度、纬度、气温、气压和风速**。这个目标数据提供了**模型预测下一时间步路径的依据**，使得滑动窗口法能够**逐步建立台风路径的动态预测能力。**

③**滑动窗口的逐日预测**：在滑动窗口法的应用过程中，模型**每次使用前两天的数据生成新的样本并预测下一个时间点的特征**。通过这种**逐日滑动**的方式，模型能够**生成一条连续的路径预测。**

（3）滑动窗口法的设定依据

选择**“前两天预测第三天”**的滑动窗口设定，是为了**符合题目要求的预测需求**。具体来说，题目要求预测**台风贝碧嘉在 2024 年 9 月 13 日-17 日的路径情况**，而贝碧嘉台风的观测数据从 **9 月 11 日**开始。因此，为准确模拟 9 月 13 日及之后的台风路径，**滑动窗口法利用 9 月 11 日和 12 日的已知观测数据推算 9 月 13 日的情况**，以**逐日预测后续路径**。这种设定不仅**确保了预测的时间连续性**，也为模型提供了足够的**上下文信息**，以增强预测的**可靠性和准确性。**

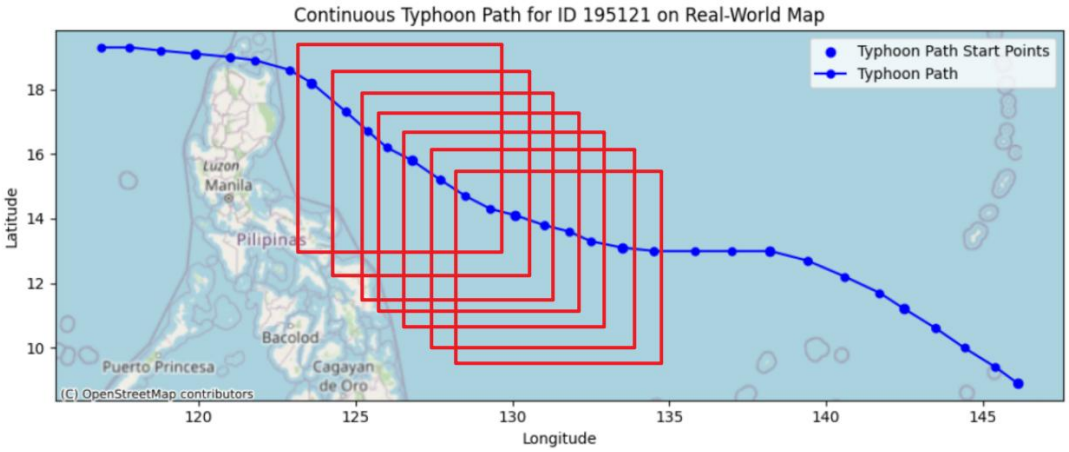


图 5.4 台风路径滑动窗口示意图

（4）滑动窗口法的优点

通过滑动窗口法，可以**有效解决台风路径预测中的数据维度不一致性问题**，为模型提

供了**标准化的输入特征**。同时，滑动窗口法在**时间序列预测中确保了每一时间步输入数据的连续性**，能够**捕捉到台风路径随时间变化的潜在规律**，从而**提升模型在台风路径预测上的精度**。

5.3.2 指标的选取

在构建台风路径预测模型的特征空间时，我们**综合考虑了台风路径的时序动态特征**以及**各类气象变量对台风路径变化的潜在影响**。为此，选取了**经度、纬度、风速、气压、气温**等核心指标。这些指标不仅**直接反映台风的强度和位置**，还能够通过**差值、对数、平方、比率**等变换**捕捉到不同时间段内特征变化的复杂性**，提升模型的**特征表达能力**。

在表 5.1 中，我们使用**下标形式“天数,小时”来表示台风的具体观测时间点**。下标的第一个数字代表观测天数，第二个数字表示具体的小时数。**举例解释如下：**

下标(1,14)表示观测数据来源于**预测台风的前 48 小时到前 24 小时之间的第 1 天 14 点时的观测值**。

下标(2,20)表示观测数据来源于**预测台风的前 24 小时到 0 小时之间的第 2 天 20 点时的观测值**。

这种**时间下标形式在特征选择中确保了数据的时序一致性**，使得模型能够根据**连续时间点的观测数据进行动态分析**。

表 5.1 台风路径预测特征指标及其时间跨度

序号	特征	备注
1-8	$LAT_{1,14}$、$LAT_{1,20}$、$LAT_{1,2}$、$LAT_{1,8}$、$LAT_{2,14}$、$LAT_{2,20}$、$LAT_{2,2}$、$LAT_{2,8}$	前 48h-前 24h，前 24h-0h 纬度
9-16	$LONG_{1,14}$、$LONG_{1,20}$、$LONG_{1,2}$、$LONG_{1,8}$、$LONG_{2,14}$、$LONG_{2,20}$、$LONG_{2,2}$、$LONG_{2,8}$	前 48h-前 24h，前 24h-0h 经度
17-24	$WND_{1,14}$、$WND_{1,20}$、$WND_{1,2}$、$WND_{1,8}$、$WND_{2,14}$、$WND_{2,20}$、$WND_{2,2}$、$WND_{2,8}$	前 48h-前 24h，前 24h-0h 风速
25-32	$PRES_{1,14}$、$PRES_{1,20}$、$PRES_{1,2}$、$PRES_{1,8}$、$PRES_{2,14}$、$PRES_{2,20}$、$PRES_{2,2}$、$PRES_{2,8}$	前 48h-前 24h，前 24h-0h 气压
33-40	$TEM_{1,14}$、$TEM_{1,20}$、$TEM_{1,2}$、$TEM_{1,8}$、$TEM_{2,14}$、$TEM_{2,20}$、$TEM_{2,2}$、$TEM_{2,8}$	前 48h-前 24h，前 24h-0h 气温
41-47	$LAT_{1,14}-LAT_{1,20}$、$LAT_{1,20}-LAT_{1,2}$、$LAT_{1,2}-LAT_{1,8}$、$LAT_{1,8}-LAT_{2,14}$、$LAT_{2,14}-LAT_{2,20}$、$LAT_{2,20}-LAT_{2,2}$、$LAT_{2,2}-LAT_{1,8}$	前 48h-前 24h，前 24h-0h 一阶纬度差
48-54	$LONG_{1,14}-LONG_{1,20}$、$LONG_{1,20}-LONG_{1,2}$、$LONG_{1,2}-LONG_{1,8}$、$LONG_{1,8}-LONG_{2,14}$、$LONG_{2,14}-LONG_{2,20}$、$LONG_{2,20}-LONG_{2,2}$、$LONG_{2,2}-LONG_{1,8}$	前 48h-前 24h，前 24h-0h 一阶经度差
55-61	$WND_{1,14}-WND_{1,20}$、$WND_{1,20}-WND_{1,2}$、$WND_{1,2}-WND_{1,8}$、$WND_{1,8}-WND_{2,14}$、$WND_{2,14}-WND_{2,20}$、$WND_{2,20}-WND_{2,2}$、$WND_{2,2}-WND_{1,8}$	前 48h-前 24h，前 24h-0h 一阶风速差
62-68	$PRES_{1,14}-PRES_{1,20}$、$PRES_{1,20}-PRES_{1,2}$、$PRES_{1,2}-PRES_{1,8}$、$PRES_{1,8}-PRES_{2,14}$、$PRES_{2,14}-PRES_{2,20}$、$PRES_{2,20}-PRES_{2,2}$、$PRES_{2,2}-PRES_{2,8}$	前 48h-前 24h，前 24h-0h 一阶气压差
69-75	$TEM_{1,14}-TEM_{1,20}$、$TEM_{1,20}-TEM_{1,2}$、$TEM_{1,2}-TEM_{1,8}$、$TEM_{1,8}-TEM_{2,14}$、$TEM_{2,14}-TEM_{2,20}$、$TEM_{2,20}-TEM_{2,2}$、$TEM_{2,2}-TEM_{2,8}$	前 48h-前 24h，前 24h-0h 一阶气温差
76-83	$\cos(LAT_{1,14})$、$\cos(LAT_{1,20})$、$\cos(LAT_{1,2})$、$\cos(LAT_{1,8})$、	前 48h-前 24h，前

84–91	$\cos(LAT_{2,14})$、$\cos(LAT_{2,20})$、$\cos(LAT_{2,2})$、$\cos(LAT_{2,8})$ $(WND_{1,14})^2$、$(WND_{1,20})^2$、$(WND_{1,2})^2$、$(WND_{1,8})^2$、 $(WND_{2,14})^2$、$(WND_{2,20})^2$、$(WND_{2,2})^2$、$(WND_{2,8})^2$	24h-0h 纬度的 cos 值 前 48h-前 24h，前 24h-0h 风速的平方
92–99	$\log(WND_{1,14})$、$\log(WND_{1,20})$、$\log(WND_{1,2})$、$\log(WND_{1,8})$、 $\log(WND_{2,14})$、$\log(WND_{2,20})$、$\log(WND_{2,2})$、$\log(WND_{2,8})$	前 48h-前 24h，前 24h-0h 风速的对数值
100–107	$\log(PRES_{1,14})$、$\log(PRES_{1,20})$、$\log(PRES_{1,2})$、$\log(PRES_{1,8})$、 $\log(PRES_{2,14})$、$\log(PRES_{2,20})$、$\log(PRES_{2,2})$、$\log(PRES_{2,8})$	前 48h-前 24h，前 24h-0h 气压的对数值
108–115	$(PRES_{1,14})^2$、$(PRES_{1,20})^2$、$(PRES_{1,2})^2$、$(PRES_{1,8})^2$、 $(PRES_{2,14})^2$、$(PRES_{2,20})^2$、$(PRES_{2,2})^2$、$(PRES_{2,8})^2$	前 48h-前 24h，前 24h-0h 气压的平方
116–123	$PRES_{1,14}/WND_{1,14}$、$PRES_{1,20}/WND_{1,20}$、$PRES_{1,2}/WND_{1,2}$、 $PRES_{1,8}/WND_{1,8}$、$PRES_{2,14}/WND_{2,14}$、$PRES_{2,20}/WND_{2,20}$、 $PRES_{2,2}/WND_{2,2}$、$PRES_{2,8}/WND_{2,8}$	前 48h-前 24h，前 24h-0h 气压与风速比
124–131	$WND_{1,14}/PRES_{1,14}$、$WND_{1,20}/PRES_{1,20}$、$WND_{1,2}/PRES_{1,2}$、 $WND_{1,8}/PRES_{1,8}$、$WND_{2,14}/PRES_{2,14}$、$WND_{2,20}/PRES_{2,20}$、 $WND_{2,2}/PRES_{2,2}$、$WND_{2,8}/PRES_{2,8}$	前 48h-前 24h，前 24h-0h 风速与气压比
132–139	$PRES_{1,14}/WND_{1,14}^2$、$PRES_{1,20}/WND_{1,20}^2$、$PRES_{1,2}/WND_{1,2}^2$、 $PRES_{1,8}/WND_{1,8}^2$、$PRES_{2,14}/WND_{2,14}^2$、$PRES_{2,20}/WND_{2,20}^2$、 $PRES_{2,2}/WND_{2,2}^2$、$PRES_{2,8}/WND_{2,8}^2$	前 48h-前 24h，前 24h-0h 气压与风速平方比
140–147	$WND_{1,14}/PRES_{1,14}^2$、$WND_{1,20}/PRES_{1,20}^2$、$WND_{1,2}/PRES_{1,2}^2$、 $WND_{1,8}/PRES_{1,8}^2$、$WND_{2,14}/PRES_{2,14}^2$、$WND_{2,20}/PRES_{2,20}^2$、 $WND_{2,2}/PRES_{2,2}^2$、$WND_{2,8}/PRES_{2,8}^2$	前 48h-前 24h，前 24h-0h 风速与气压平方比

表格中的特征包括原始指标及其多种变换形式。以下是每类特征及其选择的理由：

（1）差值与对数变换

在时间序列数据中，前后时间步的**差值与对数变换**可以有效反映数据的**变化趋势**。

①差值：差值特征能够直接捕捉**台风路径及气象特征的短期变化率**。例如，风速和气压的差值特征能够揭示台风**强度变化的趋势**，有助于**预测台风路径变化**。

②对数变换：对数变换通过缩小数据的量级差异，**减少极端值的影响**，使模型在训练过程中更稳定。对于像风速这样的变量，对数值能够**更好地反映相对变化**，适应不同强度的台风路径变化。

（2）平方和平方根变换

对气压和风速等变量进行了**平方变换**，以强调这些指标**在极端天气条件下的特性**。

平方值：平方变换可以**放大数据的波动性**，突出极端变化，适用于描述台风路径中的剧

烈变化。同时，这种处理方式确保了**模型输入的非负性**，避免了负值对训练过程的不利影响。例如，**风速的平方值更能突出风力变化的显著性**，从而帮助模型在极端天气事件中捕捉到风速变化的影响。

（3）比率特征

比率特征用于分析**两个变量之间的相对关系**，从而揭示复杂的相互作用。

①气压与风速比：通过气压和风速的比率，模型可以识别**气压变化对台风风力的调节作用**。例如，当风速与气压的比值升高时，通常意味着**台风中心气压降低**，增强了风场的影响。因此，引入**气压与风速比**能够让模型**更准确地预测台风路径**。

②对数比：对比率进行**对数变换**进一步平滑了数据波动，避免由于数值范围不同而导致模型的训练偏差，提升模型的泛化能力。

（4）角度特征与三角函数转换

台风的经纬度变化反映了路径方向的变化，因此使用了 cos **变换**捕捉路径的**空间特征**。

cos 变换：使用纬度和经度的 cos 值能够消除路径方向的角度影响，使模型可以**专注于路径的波动性而非地理方位**。这种处理方式能**更好地捕捉台风路径在空间维度上的连续性**。

（5）组合与交互特征

为了增强模型的**特征表达能力**，我们构建了一些**组合特征**。

①风速与气压平方的比率：通过引入**气压平方与风速比**，模型可以进一步捕捉气压和风速在台风路径变化中的**相互作用**，这种复杂的交互关系有助于**提高预测精度**。

②多维组合特征：通过多种特征组合（例如经度和纬度、风速和气压）可以揭示出不同变量在台风路径变化中的**联合作用**，提高模型的预测准确性。

综上所述，本研究通过**差值、对数、平方、比率**等多种变换以及**组合特征**，为模型提供了**多维度、结构化的输入数据**。这些特征的变换不仅**提高了模型对时序特征的感知能力**，也为台风路径的**精准预测**提供了充分的特征信息，从而使得模型在**捕捉台风路径变化趋势**上更加准确。

5.4 基于 LSTM 的路径预测模型

为了实现台风路径的高效预测，本研究采用了**长短期记忆网络（LSTM）模型**。LSTM 是一种特殊的**递归神经网络（RNN）**，具有捕捉长时间依赖关系的能力，能够有效建模时间序列数据中特征的动态变化，非常适用于台风路径预测。台风路径的变化受**气温、气压、风场等多重因素**的影响，而 LSTM 的记忆单元能够保留和选择性遗忘信息，从而更准确地捕捉台风路径的复杂动态变化。

5.4.1 训练集与测试集的构建

为确保模型的**预测精度和泛化能力**，我们将数据集划分为 **80%的训练集和 20%的测试集**。训练集用于模型学习，测试集则用于验证模型在未见数据上的表现。

训练集构建：通过**滑动窗口法**生成训练样本，每次窗口包含**连续两天的观测数据（8 组特征）**，预测下一天的路径和气象特征。逐日推进的滑动窗口确保输入序列的**时间维度一致**，并捕捉路径的短期变化。模型学习这些样本以建立对台风路径的动态预测能力。

测试集构建：测试集使用与训练集相同的滑动窗口法生成，用于验证模型的**泛化性能**。测试集包含在训练中未曾出现的样本，以确保模型在新数据上也能有效预测台风路径。

5.4.2 LSTM 模型的实施

在 LSTM 模型的设计中，我们团队特别关注**输入层和输出层的特征选择**，以充分利用台

time	longitude	latitude	strength	power	wind_speed	pressure	move_speed	move_direction	radius7	radius10	radius12
2024-09-10 20:00:00	145.40	12.40	热带风暴	8	18	998	26	北西	230	0	0
2024-09-10 23:00:00	144.90	12.90	热带风暴	8	18	998	24	北西	230	0	0
2024-09-11 02:00:00	144.50	13.30	热带风暴	8	18	998	22	北西	230	0	0
2024-09-11 05:00:00	143.90	13.40	热带风暴	9	23	990	20	北西	260	0	0
2024-09-11 08:00:00	143.30	13.60	热带风暴	9	23	990	22	北西	260	0	0
2024-09-11 11:00:00	143.20	14.00	强热带风暴	10	25	985	24	北西	260	0	0
2024-09-11 14:00:00	142.50	14.70	强热带风暴	10	25	985	21	北西	260	0	0
2024-09-11 17:00:00	141.80	15.00	强热带风暴	10	25	985	20	北西	260	0	0
2024-09-11 20:00:00	141.60	15.10	强热带风暴	10	25	985	22	北西	260	0	0
2024-09-11 23:00:00	141.40	15.20	强热带风暴	10	28	982	26	北北西	260	65	0
2024-09-12 02:00:00	140.80	16.00	强热带风暴	10	28	982	22	北北西	260	80	0
2024-09-12 05:00:00	140.40	17.10	强热带风暴	10	28	982	23	北北西	260	80	0
2024-09-12 08:00:00	140.00	17.30	强热带风暴	10	28	982	25	北北西	260	80	0
2024-09-12 11:00:00	139.90	18.10	强热带风暴	10	28	982	28	北北西	260	80	0
2024-09-12 14:00:00	139.80	19.10	强热带风暴	11	30	980	25	北西	200	60	0

风路径数据中的时序特征。贝碧嘉台风的分析数据由我们团队**自行从浙江省台风网爬取**，数据包含了**时间、经度、纬度、强度、等级、风速、气压、移动速度、移动方向、7 级风圈半径、10 级风圈半径、12 级风圈半径**等多个关键气象指标。

图 5.5 贝碧嘉数据集

输入层：如表 5.1 中所示，输入特征包括了 **147 个关键特征**，这些特征是通过**滑动窗口法生成的**，涵盖了台风的**经度、纬度、气温、气压、风速**等观测数据，分别以 **24 小时的前 48 小时、前 24 小时和当前时间点（0 小时）为基础**进行采样，每天选择 **2 时、8 时、14 时、20 时**的观测数据。这种特征选择能够捕捉台风路径中的**短期变化趋势**，确保数据的多维度输入，同时提供完整的**时序信息**供模型分析。

输出层：模型的输出层用于**预测下一天的路径目标**，包括**经度、纬度、气温、气压和风速**。同样在 **24 小时的时段内，每隔 6 小时输出 2 时、8 时、14 时、20 时**四个时间点的预测值。这样不仅保证了路径预测的连续性，还能够更详细地刻画台风在不同时段的**变化趋势**，以确保模型在每个时间步都能提供**精准的预测数据**。

这种设计流程确保了模型在**台风路径预测任务中的精准性和泛化能力**。通过滑动窗口法生成的**训练集和测试集**，LSTM 模型能够逐步学习并捕捉台风路径的时序特征变化，而输入层与输出层的精心设计进一步增强了模型在不同时间步上的**预测精度**。

模型不仅能够**识别和预测台风路径的连续变化**，还能**细致地刻画台风在每日各时段的动态趋势**。最终，以 **2024 年 9 月 13 日至 17 日台风贝碧嘉的路径预测为例**，LSTM 模型展示了在复杂路径预测任务中的卓越表现，为台风路径预测的实际应用提供了有力支持。

5.4.3 预测结果与分析

在这一部分，我们对 **2024 年 9 月 13 日至 17 日第 13 号台风贝碧嘉**的路径进行了每日 **14 点的精准预测**，预测路径如下表 5.2 所示。表格中记录了每日 **14 点的台风中心位置**，包括经度和纬度信息。该表格便于将**预测路径与实际路径**进行对比，为后续的**误差分析**和**动态时间规整（DTW）**提供数据基础。

表 5.2 2024 年 9 月 13 日-17 日每日 14 点的第 13 号台风贝碧嘉的中心位置

时间	**台风中心位置**（经度,纬度）
2024-09-13 14:00:00	135. 38,23.74
2024-09-15 14:00:00	134.22,24,42
2024-09-16 14:00:00	133.04,26.1
2024-09-17 14:00:00	130.38,27.52

（1）**预测结果的准确性**

LSTM 模型在**台风路径的整体趋势与具体位置**的预测上表现优异，尤其是在**路径平稳阶段，预测路径与实际路径的吻合度**较高，表明模型对**稳定路径的跟踪能力**较强。针对贝碧嘉路径的每日变化，LSTM 预测结果显示出较好的**精确度**，能够较为细致地反映路径的连续变化。

图 5.6 展示了**台风贝碧嘉在 2024 年 9 月 13 日至 17 日的实际路径（蓝色）与 LSTM 模型预测路径（红色）的对比。实际路径和预测路径在路径平稳阶段吻合较好，而在路径转向区域出现一定偏差**，为后续的**误差分析和动态时间规整（DTW）评估**提供了可视化支持。

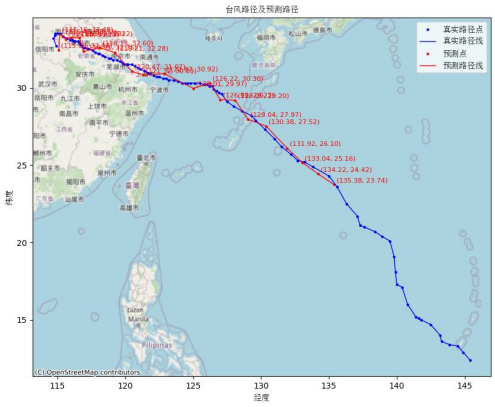


图 5.6 实际路径与预测路径对比图

（2）误差分析与动态时间规整（DTW）对比

为了进一步分析预测结果的**准确性**，我们运用 **DTW 算法**将预测路径与实际路径进行了比对。**DTW 评估结果**表明，LSTM 模型在**台风路径平稳期的预测误差较低**，显示出对路径稳定时的良好适应性。然而，**台风路径中的急剧转向和复杂变化区域**出现了一定误差，这可能是受**路径突变、地理条件及多因素耦合作用**的影响。**DTW 分析**进一步验证了 LSTM 模型在**路径平稳阶段的准确性**，同时也揭示了**路径复杂阶段的优化空间**。

为评估模型准确性，团队采用动态时间规整（DTW）算法计算预测路径与实际路径的差异，DTW 距离为 8.8774，表明预测路径与实际路径相似性较高；地理位移误差为 0.4035°，**显示预测点与真实点的空间偏差小，模型的路径预测效果良好**。

（3）特征因素影响分析

模型预测的结果表明，**气温和气压**等直接气象参数对**台风路径的影响较为显著**，尤其在**稳定的路径**中作用更为突出。而在**路径变化较大的时段，洋流、风场等环境因素**的作用尤为明显，模型在这些阶段的误差有所增加。通过 **LSTM 和 DTW 的联合分析**，模型能够有效捕捉**贝碧嘉路径随时间的动态变化趋势**。

（4）路径转向原因分析

通过 **LSTM 模型的预测结果**以及与**实际路径**的对比，我们发现 **2024 年 9 月 14 日台风贝碧嘉**发生了明显的**路径转向**。这一转向现象主要受以下几个因素的影响：

气压变化：路径转向前，**气压出现了显著波动**。低气压区域可能对台风路径产生了引导作用，促使台风向新的方向移动。

风场变化：9 月 14 日前后，**周边风场发生了变化**，尤其是受高空风切变和低层引导气流的影响，导致台风路径发生调整。风场的变化会对**台风中心的移动方向**产生重要影响。

洋流因素：贝碧嘉路径经过的区域受到**洋流的影响**，尤其是**暖流的分布**改变了台风的热源供应，增强了其强度，并影响了其路径的变化。

季风作用：9 月的季风活动通常较强，贝碧嘉台风在靠近大陆的过程中，可能受到了**季风系统的引导**，使其在 **9 月 14 日出现明显的路径转向**。

通过上述分析，我们可以得出**贝碧嘉台风在 9 月 14 日的路径转向是多因素耦合作用的结果**，这些**气象和环境因素共同作用**，引导了台风路径的调整。这种路径转向的分析为未来**台风路径预测提供了重要参考依据**。

5.4.4 问题二小结

通过对 **LSTM 模型**的构建与优化，**问题二**的研究探索了**气温、气压、洋流和风场等多因素**对台风路径的影响。LSTM 模型结合**滑动窗口法及 DTW 对比分析**，展示了在**台风路径预测**上的高效性与可靠性。模型在**路径平稳期**的表现尤其突出，说明了**气象特征对台风路径变化的预测能力**；同时，通过 **DTW 分析**进一步揭示了**路径急剧变化区域的误差原因及改进**

方向。本研究不仅**验证了 LSTM 模型在台风路径预测中的适用性，**还为未来的**气象路径建模**提供了潜在的改进思路。

六、 问题三模型的建立与求解

6.1 台风数据融合

6.1.1 数据筛选

（1）台风月份筛选

为了解决贝碧嘉登陆后的风速和降水量预测问题，我们最初筛选了以往数据中 9、10、11 月份的台风进行训练。然而，由于与贝碧嘉路径相似的台风在这几个月份中非常少，加之全球变暖等因素的影响，贝碧嘉产生月份的气温与往年 678 月份更加相似，1949-2018 年间 6、7、8 月的台风路径更贴近贝碧嘉的特征，因此最终选择了这几个月份的数据作为训练集。

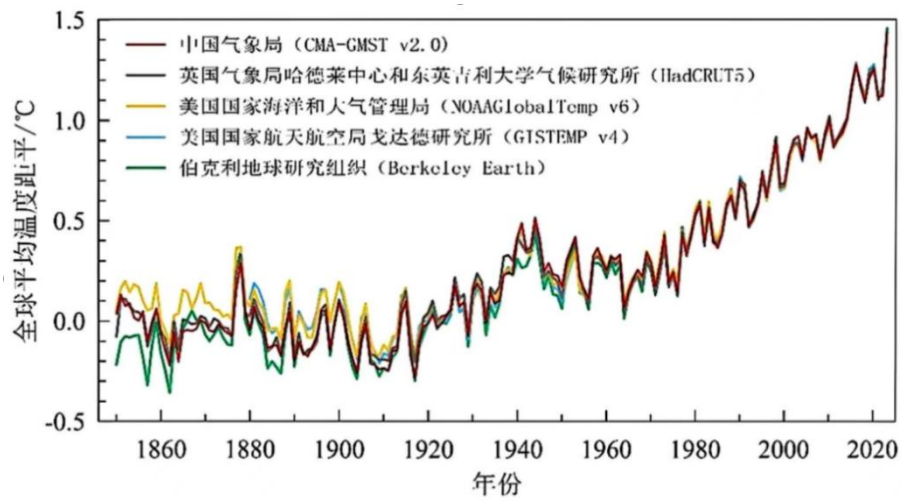


图 6.1 全球平均温度距平 1895~2023（相对 1850~1900 年平均值）

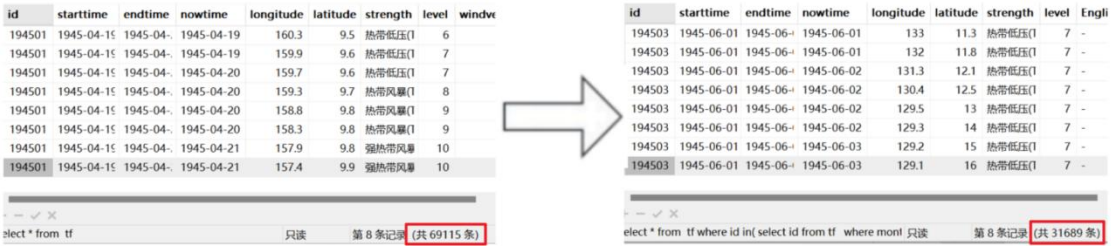


图 6.2 月份筛选变化图

（1）登陆点筛选

结合贝碧嘉从上海登陆的特点，选取了历史数据中登陆点位于我国浙江省沿海一带城市的台风记录。台风的路径特征在不同地区会有明显的差异，尤其是在登陆后的衰减和影响范围上。贝碧嘉从上海登陆，位置接近浙江沿海，因此选择浙江省沿海城市的历史台风记录能够更贴合其路径特征。

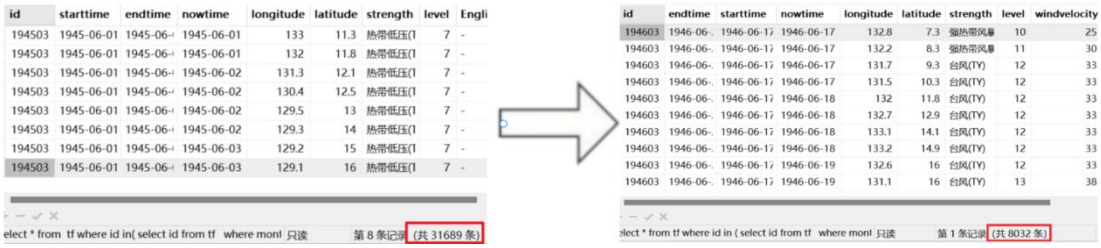


图 6.3 登陆点筛选变化图

（2）观测数据时间点筛选

由问题二预测出的 9 月 13 日-17 日每日 14 点贝碧嘉的位置经纬度，通过选择固定时间点（即每日 14 时）的数据输入模型，可以提供一致的观测视角，减少短期波动对模型的干扰，让模型更专注于较长期的趋势变化。这种方式有助于模型捕捉台风的自然衰减规律，提高预测的稳定性和准确性。

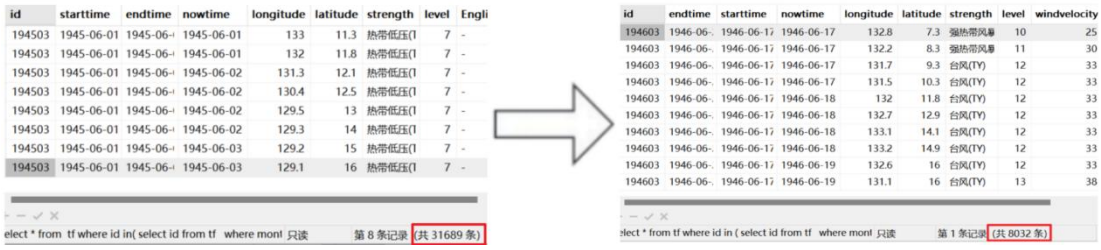


图 6.4 观测数据时间点变化图

（3）生命周期的筛选

本次模型训练中，我们需要通过前四天的台风数据来**预测未来三天**的风速和降水量，因此选择**生命周期大于等于 7 天**的台风数据尤为重要。生命周期较短的台风在观测数据中可能仅停留在生成初期或登陆初期，无法提供充足的衰减过程数据。而生命周期较长的台风则能够完整覆盖前四天的观察期和后续三天的预测期，使模型能够从完整的时序数据中学习到台风在生成、发展、登陆及衰减各阶段的特征变化。

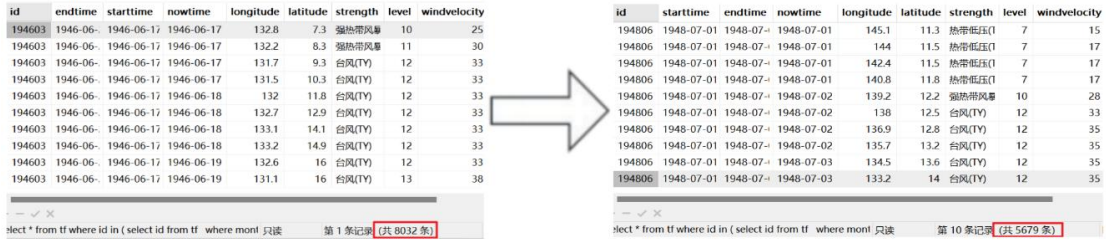


图 6.5 生命周期筛选变化图

6.1.2 数据融合

（1）降水量数据的融合

降水量的数据通过题目文件中给出的中国观测点信息表和日降水量数据表，通过计算出最近的监测点，得出台风每天 14：00 的降水量数据。

id	Chinesename	Englishname	starttime	endtime	nowtime	longitude	latitude	strength	level	windvelocity	pressure	province	precipitation
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-25 14:00	133.4	13.6	热带低压(TD	6	12	1004	Kayangel	14
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-26 14:00	131	16.5	热带低压(TD	7	15	1002	stern Visay	11
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-27 14:00	126.7	17.1	热带风暴(ST	10	25	990	Bicol	42
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-28 14:00	124.2	19.1	台风(TY)	13	40	965	agayan Vall	5
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-29 14:00	122.4	22.2	台风(TY)	13	40	965	台湾省	2
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-30 14:00	120.6	25.3	热带风暴(ST	11	30	980	台湾省	25
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-31 14:00	119.6	27.7	热带低压(TD	7	15	998	浙江省	6
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	8-01 14:00	120.3	34.5	热带低压(TD	7	15	995	江苏省	7
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	8-02 14:00	120.2	40.2	热带低压(TD	6	12	1004	辽宁省	3
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-01 14:00	132	14.6	热带风暴(TS	9	23	990	stern Visay	8
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-02 14:00	128.9	18	热带风暴(ST	11	30	980	Bicol	4
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-03 14:00	126	21.1	台风(TY)	13	40	960	Okinawa	10
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-04 14:00	124.7	23.9	台风(TY)	13	40	955	Okinawa	43

图 6.6 降水量数据融合

（2）气温数据的融合

我们通过海上两米处的气温数据来获取台风的气温，同时团队还编写代码爬取了全球监测点的气温数据，在数据筛选时，根据台风中心与两个观测点之间的距离，选择距离最近的观测点气温数据，以确保数据的代表性和准确性。（数据预处理方式与第一、二问处理方式相同）

id	Chinesename	Englishname	starttime	endtime	nowtime	longitude	latitude	strength	level	windvelocity	pressure	province	precipitation	TEMP
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-25 14:00	133.4	13.6	热带低压(TD	6	12	1004	Kayangel	14	24.72
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-26 14:00	131	16.5	热带低压(TD	7	15	1002	stern Visay	11	16.72
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-27 14:00	126.7	17.1	热带风暴(ST	10	25	990	Bicol	42	26.33
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-28 14:00	124.2	19.1	台风(TY)	13	40	965	agayan Vall	5	25.22
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-29 14:00	122.4	22.2	台风(TY)	13	40	965	台湾省	2	23.61
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-30 14:00	120.6	25.3	热带风暴(ST	11	30	980	台湾省	25	23.61
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	7-31 14:00	119.6	27.7	热带低压(TD	7	15	998	浙江省	6	23.61
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	8-01 14:00	120.3	34.5	热带低压(TD	7	15	995	江苏省	7	24.33
200108	桃芝	TORAJI	7-25 14:00	8-02 14:00	8-02 14:00	120.2	40.2	热带低压(TD	6	12	1004	辽宁省	3	25.06
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-01 14:00	132	14.6	热带风暴(TS	9	23	990	stern Visay	8	28.22
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-02 14:00	128.9	18	热带风暴(ST	11	30	980	Bicol	4	28.8
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-03 14:00	126	21.1	台风(TY)	13	40	960	Okinawa	10	28.8
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-04 14:00	124.7	23.9	台风(TY)	13	40	955	Okinawa	43	28.8
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-05 14:00	123.2	26.6	强台风(STY	14	45	950	台湾省	103	28.8
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-06 14:00	120.4	28.9	台风(TY)	12	33	975	浙江省	20	28.22
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-07 14:00	118.3	31.7	热带风暴(TS	9	23	994	安徽省	5	28.33
200509	麦莎	Matsa	7-31 20:00	8-09 05:00	8-08 14:00	119.1	37.3	热带风暴(TS	8	18	995	山东省	10	28.33

图 6.7 气温数据融合

6.2 搭建特征空间

在特征空间的构建过程中，我们考虑了台风在**登陆后的降水量和风速的变化趋势**。通常情况下，台风的降水量和风速会逐渐减弱，但这种变化不仅受到**历史降水量和风速的影响**，还受台风路径、气温、湿度、台风与沿海观测点的相对位置等多重因素的影响。因此，为了更准确地刻画台风的减弱过程，我们选择了以下关键特征：前四天每天下午两点的风速、降水量、方向角、与登陆点的距离、台风中心经度纬度值。这些特征能够全面反映台风的强度变化、行进路径、环境因素等，从而提高模型对降水量和风速的预测准确性。

以下是每类特征及其选择的理由：

（1）登陆点经纬度：登陆点的地理位置为台风预测提供了一个重要的参考点。台风行进距离和登陆点之间的关系能够帮助模型了解台风登陆后的变化趋势，特别是在风速和降水的减弱阶段。

（2）台风中心与登陆点的距离：是反映台风强度变化的重要特征。通常随着台风逐渐远离登陆点，风速和降水量逐渐减弱，因此这项特征能够帮助模型更好地预测台风的衰减过程。

（3）风速：台风在不同时间步的风速变化能够揭示其强度的动态调整，对未来时刻的风速预测具有直接的参考作用。

（4）降水量：模型能够更好地捕捉台风带来的降雨趋势，从而预测台风在接下来的时间中降水量的变化。

（5）方向角：方向角表示台风的行进方向，通过观察台风方向角的变化，模型可以识别出台风可能的转向趋势，从而更准确地预测其路径变化对风速和降水量的影响。

（6）气温：气温是影响台风强度变化的重要环境因素。较高的气温通常会提供更多的能量，使台风增强，而较低的气温则可能导致台风减弱。因此，气温特征能够帮助模型更好地理解台风在行进过程中的能量变化及其对风速和降水量的影响。

综上所述，我们所选取的指标如下表所示（以下特征是在第二题基础上增加的特征）：

在表格中，我们使用下标形式“天数,小时”和“天数，距离，小时”来表示台风的具体观测时间点。下标的第一个数字代表观测天数，第二个数字表示具体的小时数。举例解释如下：

1,14,100 表示观测数据来源于台风的第 1 天 14 点，即前 48 小时到 24 小时之间的第 1 天 14 点时距离台风中心点 100 公里行径路线点上的观测值。

2,14,200 表示观测数据来源于台风的第 2 天 14 点，即前 24 小时到 0 小时之间的第 2 天 14 点时距离台风中心点 200 公里行径路线点上的观测值。

序号	特征	备注

1-4	$LAT_{1,14}$、$LAT_{1,14}$、$LAT_{3,14}$、$LAT_{4,14}$	前 4 天每天 14 时纬度
5-8	$LONG_{1,14}$、$LONG_{2,14}$、$LONG_{3,14}$、$LONG_{4,14}$	前 4 天每天 14 时经度
9-10	LAT_2、$LONG_2$	登陆点经度纬度
11-14	$DISTANCE_{1,14}$、$DISTANCE_{2,14}$、$DISTANCE_{3,14}$、$DISTANCE_{4,14}$	前 4 天每天 14 时登陆点到台风中心距离
15-18	$WND_{1,14}$、$WND_{2,14}$、$WND_{3,14}$、$WND_{4,14}$	前 4 天每天 14 时风速
19-22	$PRCP_{1,14}$、$PRCP_{2,14}$、$PRCP_{3,14}$、$PRCP_{4,14}$	前 4 天每天 14 时降水量
23-26	$AZI_{1,14}$、$AZI_{2,14}$、$AZI_{3,14}$、$AZI_{4,14}$	前 4 天每天 14 时方向角
27-30	$TEMP_{1,14}$、$TEMP_{2,14}$、$TEMP_{3,14}$、$TEMP_{4,14}$	前 4 天每天 14 时温度
31-54	$PRCP_{1,14,100}$、$PRCP_{1,14,200}$、$PRCP_{1,14,300}$、$WND_{1,14,100}$、$WND_{1,14,200}$、$WND_{1,14,300}$ $PRCP_{2,14,100}$、$PRCP_{2,14,200}$、$PRCP_{2,14,300}$、$WND_{2,14,100}$、$WND_{2,14,200}$、$WND_{2,14,300}$ $PRCP_{3,14,100}$、$PRCP_{3,14,200}$、$PRCP_{3,14,300}$、$WND_{3,14,100}$、$WND_{3,14,200}$、$WND_{3,14,300}$ $PRCP_{4,14,100}$、$PRCP_{4,14,200}$、$PRCP_{4,14,300}$、$WND_{4,14,100}$、$WND_{4,14,200}$、$WND_{4,14,300}$	前四天每天 14 时距离台风中心 100、200、300 公里行进路线上的降雨量、风速
55-		

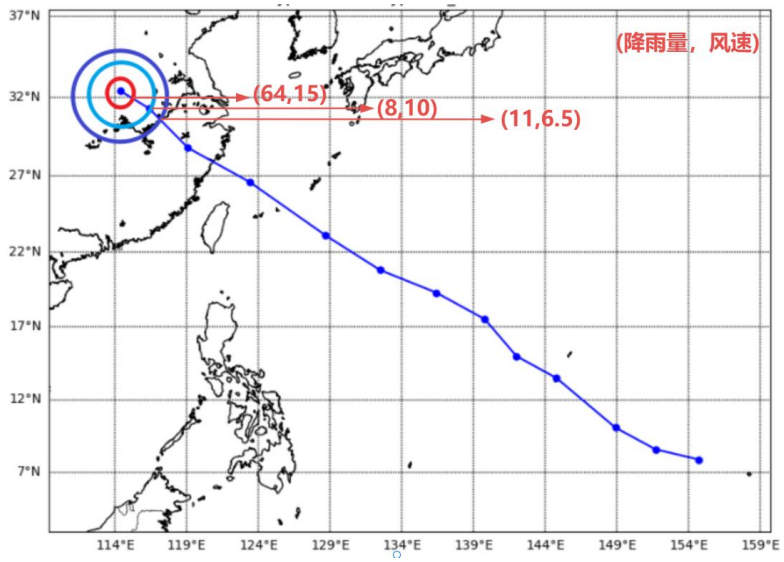


图 6.8 距离中心点 300 公里的降雨量风速

6.3 基于 LSTM+多重注意力机制的预测模型

6.3.1 滑动窗口处理数据

在本次模型中，我们采用了与**第二问相似的滑动窗口方法**进行预测。然而，与第二问

中选择的**两天中八个时间点数据**不同，我们专注于预测 **14 号 14 时的降水量、风速**等数据，因此选取了**更多天的 14 时数据**来进行模型训练。这一选择的原因在于，我们希望利用**更适配的历史数据**来提高预测的准确性。

具体而言，我们选择的窗口包括了**前 4 天中每天 14 时的降水量、风速**等数据，这样可以充分捕捉到台风在**同一时间点的影响变化**。此外，选择**一天的降水量数据**也是基于前期多表融合的结果，因为在数据整合过程中，降水量数据的提供只有一整天的数据。通过这种方式，我们的模型能够在更加准确的数据集上进行训练，同时利用较长时间范围内的观测数据，从而更好地反映台风的动态特征和变化规律。

滑动窗口通过创建时间序列的**重叠子集**就是**前四天每天 14 时的台风经纬度、风速、降雨量、方向角和与登陆点的距离**，使得每个时间步都可以作为一个样本输入到模型中。同时我们通过**重叠窗口增加训练样本的数量**，帮助提高模型的**泛化能力**。

6.3.2 模型的实施

LSTM 模型的结构设计包含以下几个主要层次：

（1）输入层：输入的数据为滑动窗口法构造的前四天每天 14 时的台风特征（如经度、纬度、方向角、气压、风速、气温和台风中心与登陆点的距离等）。

（2）LSTM 层：该层包含若干个 LSTM 单元，每个单元通过记忆和遗忘门，捕捉台风特征随时间的动态变化。

（3）全连接层：经过 LSTM 层处理后，连接至全连接层，用于进一步处理特征数据。

（4）输出层：最终输出预测的目标变量，即台风后一天的降水量和风速。

第二步：更新数据，输入第二天到第五天每天 14 时的数据来预测第六天的数据

第三步：接着滑动窗口更新，输入第三天到第六天的数据来预测第七天的数据

由此预测出了贝碧嘉 16-18 号的全部数据，我们可以得到台风行进过程中降水量与风速的关系。（通过 python 代码我们团队从中央气象台台风网爬取了贝碧嘉 12-15 号的数据）

6.3.3 预测结果分析

（1）预测结果的准确性

基于 **LSTM 和多重注意力机制**的模型在预测**台风贝碧嘉的降水量和风速**方面表现出**较高的准确性**。模型能够准确捕捉**降水量和风速的主要变化趋势**，特别是在降水峰值和风速增减的关键时段表现良好。通过引入**多重注意力机制**，模型能够**聚焦**于影响台风特征的关键因素，从而进一步提升了预测精度。

（2）误差分析

我们建立了**台风登陆后降水量和风速的关系模型**，以及**降水量与距台风中心距离的关系模型**，并基于该模型预测了 2024 年 9 月 16 日至 18 日第 13 号台风贝碧嘉行进过程中的**中心风力和降水量**。通过误差分析发现，模型在台风**风速和降水量的逐渐减弱阶段**表现较为准确，但在剧烈变化的时段仍存在一定偏差。

为进一步验证模型预测的准确性，我们将模型预测的数据与实际数据进行了**逐步对比分析**。对比结果表明，模型能够捕捉台风**登陆后风速和降水量随距离增大的减弱趋势**，但在个别时段中，预测结果与实际数据的时间步上存在一定差异。尤其在台风快速减弱或移动过程中，模型在降水量和风速的衰减速率上略有提前或滞后。

这一**误差分析表明**，尽管模型能够较好地拟合台风登陆后的整体衰减趋势，但在应对台风**行进速度加快或风速、降水量剧烈波动**的情境下，模型的动态适应性仍有提升空间。未来的优化可以着重于提升模型在这些突变场景下的**响应能力**，以进一步提高预测的准确性和时间一致性。

日期	经纬度	经过省份	降雨量	风速
16号14时	（131，26.1）	江苏苏州	98	33
17号14时	（130.38，27.52）	安徽淮北	231.7	18
18号14时	（114.8，33.4）	河南周口	79.4	12

图 6.9 贝碧嘉 16-18 号 14 时降水量、风速数据

（3）结果分析

在这次结果中，基于 **LSTM 与多重注意力机制**的模型成功实现了对**台风贝碧嘉降水量和风速**的预测。通过对 16 日至 18 日每日 14 时的预测数据与实际观测数据进行对比分析，我们发现模型在捕捉降水量峰值和风速变化趋势方面表现良好，**注意力机制**的引入进一步帮助模型聚焦于关键特征（如台风中心与登陆点的距离），从而有效提高了**预测精度**。这些结果表明，该模型在实际台风监测与预警中具有应用潜力。

通过分析预测结果的时间序列趋势，包括风速和降水量的变化模式，可以识别和分析影响预测结果的关键气象指标。观察风速和降水量的变化模式发现，气温、气压、洋流和风场的波动对风速和降水量的预测有重要影响：

（1）气温的变化直接影响台风的能量供给。在气温升高的区域，台风从海洋吸收更多能量，导致风速和降水量的增加；而在气温较低区域，风速和降水量往往呈减弱趋势。因此，气温的波动对台风路径及强度有明显的调节作用。

（2）气压的梯度对台风的移动和发展有重要影响。当气压较低且变化剧烈时，台风会受到更强的推动力，导致风速加快；反之，气压稳定或气压差较小时，台风的能量供给会减少，从而降低风速。气压对降水量的影响也体现在当气压较低时，台风吸收水汽的能力更强，导致更强的降水。

（3）洋流的变化对台风路径和风速的调整作用尤为显著。温暖洋流可以为台风提供热量和水汽，从而增强风速和降水；而冷洋流则会减少台风的能量供给，导致风速和降水减弱。在洋流交替的区域，台风往往会发生路径偏转，影响风速和降水量的变化。

（4）大尺度风场对台风路径有明显的引导作用。风场的方向和强度会对台风的行进速度和方向产生直接影响，进而影响风速的大小。在风场增强的区域，台风的移动速度更快，导致风速变化更为显著；而在风场减弱的区域，台风可能停滞或路径偏转，影响降水量的分布。

6.4.4 问题三小结

在问题三中，我们建立了**台风登陆后风速与降水量的关系模型**，以及降水量随台风中心距离变化的关系模型，并应用于 2024 年 9 月 16 日至 18 日期间第 13 号台风贝碧嘉的预测。通过 **LSTM 与多重注意力机制的滑动窗口模型**，我们有效捕捉了贝碧嘉的风速和降水变化趋势。**误差分析**及动态时间规整（DTW）对比结果表明，**模型在台风整体衰减趋势上的拟合效果良好。**总体而言，该模型为台风登陆后降水和风速的预测提供了可靠的参考。

七、模型的评价与推广

7.1 模型评价

本文针对台风路径、强度和降水量的预测进行了建模分析。对于台风路径预测问题，由于台风数据集的记录时间间隔不一致且数据量有限，本文采用滑动窗口法将台风路径数据按固定时间长度分割，构建了时间序列样本，增强了数据的时序特征表达。尽管本文采用了基于 LSTM 和多重注意力机制的时序预测模型，该模型在预测路径变化趋势时能够有效捕捉风速、气温、气压等关键影响因素，但在台风剧烈变向或强度突变的时段仍存在一定的误差。模型的误差分析表明，该模型在台风路径平稳阶段的预测效果较好，但在风速和降水量剧烈波动的时段适应性稍显不足。此外，本文使用的动态时间规整（DTW）算法与实际路径对比分析也揭示了模型在应对突变情况时的改进空间。总体而言，基于多层模型集成方法的台风预测模型在精度和时间一致性方面具有较好的表现。

7.2 模型推广

本文建立的台风预测模型在台风路径、风速和降水量预测方面具有较高的准确性，能够推广到灾害预警、气象服务、交通管理、农业生产和保险行业等多个领域：

①灾害预警领域：可以用来预测台风路径和强度，帮助防灾部门提前做好应急预案，制定撤离和防护措施，降低人员和财产损失。

②气象服务领域：可以将该模型集成到气象预报系统中，为公众和政府提供实时的台风路径和强度预报，提升气象预报的准确性和及时性，提高台风预警能力。

③交通管理领域：可以用来预测台风路径和降水量，帮助交通管理部门提前规划交通路线，做好防护措施，确保公路、铁路和航空交通的安全性，减少因台风造成的交通中断和拥堵。

④农业生产领域：台风预测模型可用于监测台风对农业区域的影响，帮助农户和农业管理部门及时采取保护措施，减少台风对农作物的破坏，保障农业生产安全。

保险行业：可以用来辅助保险公司进行风险评估，根据台风预测结果对受影响地区的客户提供有针对性的保险方案，提高保险服务的精准性，同时帮助保险公司合理设置理赔范围和额度，降低损失。

⑤城市提前预警准备：城市可以有序地组织危险区域的人员疏散，提前设置和开放避难所，为可能的撤离人群提供充足的食物、饮水和医药等物资，保障市民的安全。

八、参考文献

[1]Bhatia, K. T., Vecchi, G. A., Knutson, T. R., Murakami, H., Kossin, J. P., Dixon, K. W., & Whitlock, C. E. (2019). *Recent increases in tropical cyclone intensification rates*. Nature Communications, 10(1), 1-9.

[2]Zhang, W., Zhang, Y., & Zhang, L. (2021). A deep learning-based typhoon intensity and trajectory prediction model using environmental parameters. *Atmospheric Research*, 250, 105413Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Advances in neural information processing systems, pages 3320–3328, 2014.

[3]Chen, J., Huang, W., Chen, K., & Xu, H. (2020). Clustering analysis of tropical cyclone trajectories over the western North Pacific based on hierarchical clustering and self-organizing maps. *Journal of Climate*, 33(8), 3199-3217.

[4]Li, Q., Wang, Y., Zhang, J., & Chen, X. (2022). Improved typhoon path prediction using ensemble clustering algorithms with meteorological factors. *Advances in Meteorology*, 2022, 5643210.Cortes, C.、Vapnik, V. 支持向量网络。马赫学习 20 , 273–297 (1995)。

[5]Xu, H., et al. (2019). *Applications of clustering algorithms in meteorology*. *Meteorological Applications*, 26(2), 287-303.

[6]Zhang, W., Zhang, Y., & Zhang, L. (2021). *Deep learning applications in atmospheric sciences: A review*. *Environmental Modelling & Software*, 137, 104946.

队伍编号	MCB2403007
赛道	B

基于大数据分析的电商货量预测与仓库分配优化研究

摘 要

随着电商行业的发展，货量预测与品类分仓规划问题日益成为供应链管理中的重要研究课题。本文针对电商中的货量预测问题，基于大数据分析和机器学习思想，通过确定销售趋势、市场需求和季节性变化等指标，以提高预测精度和优化库存管理为目标，建立了相应的预测模型，并使用先进的算法对模型进行求解。

针对问题一，本研究针对库存量和销量预测问题，采用 **SARIMA 模型**和**随机森林回归模型**进行分析和建模。在库存量预测中，利用 SARIMA 模型通过数据预处理和参数优化，有效捕捉了**数据的季节性和趋势**。ADF 检验结果显示中 p-value 远高于 0.05，表明数据可能具有趋势或季节性波动，适配该模型的预测模式。销量预测则使用**随机森林回归模型**，通过构建特征和处理历史数据，展示出良好的鲁棒性和准确性。随机森林回归模型预测的日销量结果的 MAE，MSE 和 RMSE 绝大部分均为 0，，R^2为 1。这表明模型在这些品类上的拟合基本完全正确。两种模型的应用为库存管理和分仓规划提供了可靠的数据支持。

针对问题二，在第二问的仓库分配模型中，我们构建了一个**线性规划模型**，旨在通过最小化仓租成本来优化各品类的仓库分配，特别强调“一品一仓”原则，以确保每个品类仅分配至一个仓库，从而简化管理流程并提高效率。同时，该模型也满足仓库的容量和产能限制。采用 **PULP_CBC_CMD 求解器**，通过**分支割平面法**求解，使得模型能够有效探索可行的分仓方案，最终实现了整体仓租成本的降低。最终发现求解时间设定为 400 秒时，此时为最优解，总成本为 6977267.16125076 元。

针对问题三，在多目标优化的背景下，本研究旨在利用**模拟退火算法**解决电商品类的多仓分配问题，以最大化仓储布局的整体收益。我们首先定义了一个综合收益函数，涵盖高级品类收益、件型一致性收益、品类关联度收益和总仓储成本，并为各收益项赋予不同权重，以合理平衡多目标收益，确保优化过程的全面考量。通过逐步降低温度的**模拟退火算法**，探索解空间以优化收益并控制成本，逐渐减少非优解的接受概率，引导解的收敛，最终保留当前最优解。总成本为 27,502,455.91 元。关联度收益为 390,167.564。高级品类收益达到了 7,860,000。件型一致性收益为 5,670,000，虽然仓租总成本上升但是各项业务指标兼顾考虑，能够更好的适应现实情况。

该方案在处理**非线性关系**和**动态变化**时表现出色，能够适应市场波动，提升预测的准确性。此外，模型的灵活性和系统化决策支持功能，使得资源配置更加合理，运营效率得到显著提升，能够在不同业务场景中推广，为企业提供科学依据和数据支持。

关键词： SARIMA 模型　随机森林回归模型　线性规划模型　模拟退火算法

目录

一、 问题重述 1
1.1 问题背景 1
1.2 问题提出 1
二、 问题分析 1
2.2 问题一的分析 2
2.3 问题二的分析 3
2.4 问题三的分析 3
三、 模型假设与符号说明 4
3.1 模型基本假设 4
3.2 符号说明 4
四、 问题一模型建立与求解 5
4.1 数据预处理 5
4.1.1 处理历史销量中的缺失值 5
4.2 库存量预测模型的建立与求解 6
4.2.1 数据与模型的适配性分析与处理 6
4.2.2 模型构建与求解逻辑 9
4.2.3 SARIMA 模型描述 9
4.2.4 基于网格搜索的自动参数优化 10
4.2.5 模型评估结果与分析 10
4.3 销量预测模型的建立与求解 11
4.3.1 数据与模型的适配性分析 11
4.3.2 模型构建与求解逻辑 11
4.3.3 随机森林回归模型描述 11
4.3.4 模型评估与参数检测 12
4.3.5 模型评估结果与分析 14
五、 问题二模型建立与求解 15
5.1 仓库分配模型的建立与求解 15
5.2 数据与模型的适配性分析 16
5.3 模型构建与求解逻辑 16
5.4 模型求解与结果分析 17
5.5 模型评估与应用 19
六、 问题三模型建立与求解 19
6.1 数据准备与处理 19
6.2 模型建立与求解过程 20

6.2.1 决策变量与目标函数设定 20
6.2.2 目标函数 20
6.2.3 收益计算逻辑 20
6.2.4 约束条件 22
6.2.5 决策变量约束 22
6.2.6 求解过程 22
6.2.7 最优解输出 23
6.3 模型评估与分析 24
6.4 分析不同方案中各业务指标的表现 24
七、 模型分析检验 26
7.1 灵敏度分析 26
7.2 误差分析 27
7.2.1 问题一的误差分析 27
7.2.2 问题二的误差分析 27
7.2.3 问题三的误差分析 27
7.3 整体误差分析 27
八、 模型评价与推广 27
8.1 模型的优点 27
8.2 模型的不足 28
8.3 模型的推广 28
参考文献 29
附录 30

一、 问题重述

1.1 问题背景



电商企业的商品存储系统由多个仓库组成，商品的存放主要依赖于其属性（如品类和件型）的分类和标记，以实现更高效的库存管理。在分仓规划过程中，必须预测两个关键指标：库存量和销量。为了有效应对仓容和产能的限制，同时考虑品类库存管理的复杂性和整体成本，准确的预测结果对于决策者至关重要。这些结果将帮助他们合理分配未来的仓储资源，减少不必要的空间浪费。因此，在确定各品类的预测货量之后，如何进行品类的分仓规划成为供应链管理者的重要课题。

在制定最优的分仓方案过程中，需要全面考虑仓群的复杂性以及各个仓库的仓容和产能限制。关键指标包括仓容利用率、产能利用率和总仓租成本等。当前的仓储网络由140 个仓库和 350 种品类组成，并提供了各品类的历史库存量和销量、不同仓库的相关信息（包括仓租日成本、仓容上限和产能上限）、品类之间的关联度（未出现的组合设为 0）以及不同品类的详细信息（如品类编码和件型）。因此，构建一个高效的模型来解决商品存放问题显得至关重要。对此我们建立模型解决下列问题：

1.2 问题提出

基于前述的研究背景，本题目给出了 5 个具有相关数据的信息附件和 1 个上传答案的参考附件。本题所要解决的问题如下：

问题一：建立货量预测模型，对该仓储网络 350 个品类未来 3 个月（7-9 月）每个月的库存量及销量进行预测，其中库存量根据历史每月数据预测月均库存量即可，填写表 1 的预测结果并放在正文中，并将完整结果填写在 result 表格文件中的“月库存预测结果”的表单中；销量需给出未来每天的预测结果，填写表 2 的预测结果并放在正文中，并将完整结果填写在 result 表格文件中的“日销量预测结果”的表单中。

问题二：假设当前限定每个品类只能放在一个仓库中，即一品一仓，各品类之间请基于问题 1 的预测结果建立规划模型，综合考虑多个业务目标，求得品类的分仓方案，包括：应使用哪些仓库，使用的仓库需存放哪些品类的库存。填写表 3 的分仓结果并放在正文中，并将完整品类分仓结果填写在 result 表格中的“一品一仓分仓方案”表单中。

问题三：现在为每个品类按照件型及高级品类进行打标（如附件 5），并放开一品一仓假设，即允许一个品类存放于多个仓库，但同一品类存放的仓库数量不能超过 3 个，并希望同件型、同高级品类尽量放在一个仓库 中。假设同一品类在不同仓库之间分布的库存量比例及出库量比例相同，当前业务的首要目标是最大品类关联度，同时兼顾其他指标。请基于问题 1 的预测结果建立规划模型，求得新的品类分仓方案，并分析不同方案中各业务指标的表现。填写表 4 的分仓结果并放在正文中，并将完整品类分仓结果填写在 result 表格中的“一品多仓分仓方案”表单中。

二、 问题分析

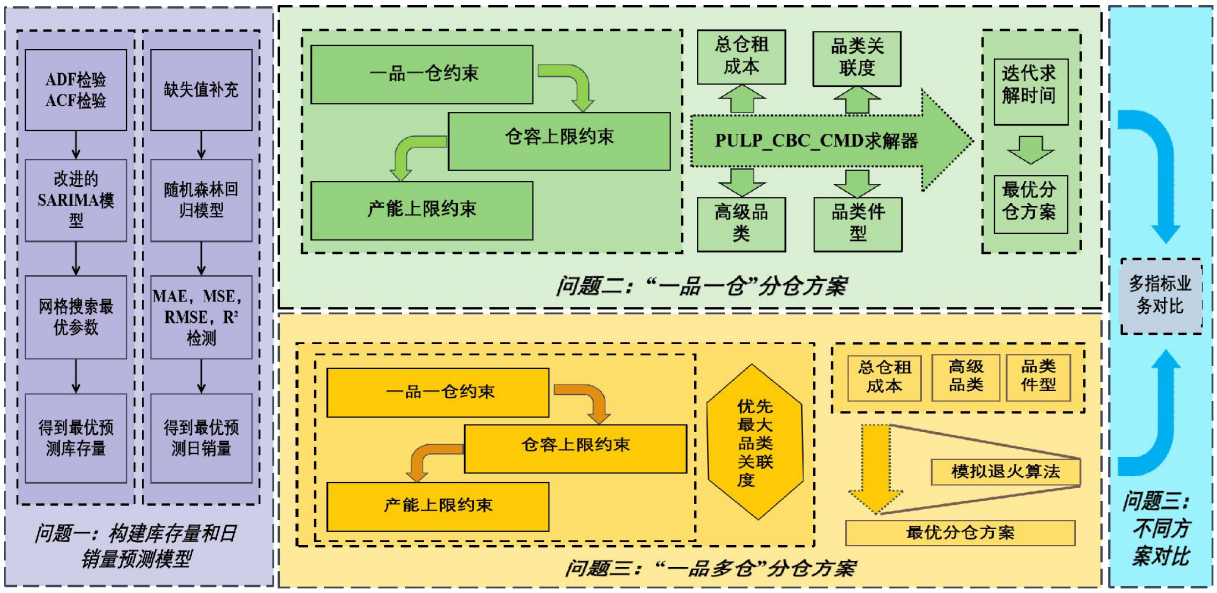


图 1 问题分析流程图

2.2 问题一的分析

在本次研究中，我们针对销量预测问题进行了深入的数据预处理和模型构建，以确保预测结果的准确性和可靠性。初步分析发现，历史销量数据中存在缺失值，这可能会影响后续模型的性能。因此，首先进行了数据的清洗和缺失值填补。

数据预处理:

在对历史销量数据的分析中，发现某些品类的销量记录不足 183 天，可能导致数据的完整性受到影响。为了提高模型的预测效果，我们采用**随机森林回归模型**对缺失值进行了填补。随机森林的优势在于其强大的非线性拟合能力和鲁棒性，能够有效应对时间序列数据中的波动性和异常值。通过提取日期特征（如日、月和星期），我们成功捕捉到了销量的时间规律，为后续的模型训练提供了高质量的输入数据。

经过多轮的交叉验证，随机森林模型在填补缺失值的过程中成功模拟了品类销量的历史趋势，形成了一个覆盖 183 天的完整数据集。这一过程不仅保留了原始数据的重要趋势，还揭示了更多的细节波动，显著提升了数据的分析价值。

库存量预测模型的建立:

在构建库存量预测模型时，我们对附件数据进行了详细的适配性分析。通过 ADF 检验和自相关性分析，发现多个品类的数据表现出明显的非平稳性和季节性特征。针对这些特征，我们采用线性插值法处理缺失数据，并通过差分操作将非平稳序列转化为平稳序列，为建立 SARIMA 模型打下了基础。

SARIMA 模型因其能够有效处理时间序列中的季节性波动，成为我们库存量预测的理想选择。通过网格搜索优化模型参数，我们确保了每个品类的 SARIMA 模型都能够准确捕捉库存量的趋势与季节性特征，为后续的库存管理提供了可靠的依据。

销量预测模型的建立:

在销量预测部分，我们使用**随机森林回归模型**对未来 3 个月的销量进行预测。通过对历史销量数据的深入分析，我们发现两段时间的数据存在明显的季节性和周期性波动，随机森林模型在处理这些复杂数据时展现出良好的鲁棒性和预测能力。

模型评估过程中，我们采用多种指标如 MAE、MSE、RMSE 和 R^2，全面验证了模型的预测性能。绝大部分品类的预测结果表明，模型能够有效捕捉销量变化趋势，且预测误差保持在合理范围内，进一步为后续的库存管理和分仓规划提供了坚实的支持。

通过本次问题一的分析与求解，我们不仅成功填补了历史销量数据的缺失值，还构建了适应性强的库存量和销量预测模型。整个过程强调了数据预处理的重要性以及模型选择的合理性，为未来的预测任务奠定了良好的基础。

2.3 问题二的分析

在电子商务的仓储管理中，合理的品类分仓决策是提升仓库管理效率和降低仓储成本的关键。参考《电商物流分仓备货优化研究》[3]中的思路，问题二旨在解决“一品一仓”的品类分仓问题，即每个品类需唯一分配到某个仓库。该问题具有以下特点：

业务背景与问题性质：在分仓场景中，电子商务企业通常面对复杂的品类和仓库布局。合理分仓可以防止仓库资源浪费，并提升管理便捷性。但因仓库容量和产能的限制，以及仓租成本的存在，每个品类的仓库选择需考虑多个因素，包括仓库的存储能力、出库能力和运营成本，增加了决策难度。由于需要满足仓库的容量和产能约束，且目标是最小化仓租成本，本问题可建模为一个具有约束条件的资源分配优化问题。

核心优化目标：问题二的核心目标在于通过最小化总仓租成本，为每个品类分配一个仓库，同时确保在满足仓库仓容和产能约束的前提下实现仓库资源的合理利用。将总仓租成本降至最低有助于降低运营开支，并确保资源配置符合公司业务需求。

关键约束条件如下：

1. 一品一仓约束：确保每个品类仅分配到一个仓库，以简化仓储管理并优化仓库资源利用。

2. 仓容上限约束：每个仓库的品类库存总量不得超过仓库容量上限，以避免仓库超载、影响存储安全和管理效率。

3. 产能上限约束：仓库每日的出库量不得超过仓库的产能上限，以确保在订单量高峰期满足履约需求，防止出库瓶颈。

4. 模型的适用性与求解方法：本问题的解答过程采用了线性规划模型，适合此类资源分配问题。线性规划通过简单的约束定义和目标函数构建，能够快速求解大规模的资源分配问题，提供高效、接近最优的解。特别地，在求解器的选择上，**分支割平面法**被用于处理整数解要求，有助于在求解过程中快速收敛，找到最优分仓方案。此外，通过增加计算时间和迭代次数，可以更接近理论最优解，使得解的质量和求解效率得到良好平衡。

5. 业务决策的参考指标：为进一步评估分仓方案的合理性，模型解答过程还综合考虑了仓容利用率、产能利用率、总仓租成本、品类分仓数和品类关联度等指标。合理的仓容和产能利用率确保仓库资源有效分配；较低的仓租成本降低运营支出；适当的品类分仓数控制管理复杂度。这些指标共同构成了对模型性能和分仓效果的多维评价基准，有助于决策者在多目标之间进行权衡。

综上所述，问题二的品类分仓问题是一个典型的**线性规划**资源分配优化问题，具有约束条件多、成本优化复杂的特点。通过构建线性规划模型、利用分支割平面法求解、在较短时间内找到接近最优解，该模型能够满足仓储分配的业务需求，同时提供科学合理的分仓决策支持。

2.4 问题三的分析

在问题三中，我们需要解决的核心问题是**电商品类的多仓分配优化，以最大化仓储布局的整体收益**。针对电商行业多品类、多仓储的特点，传统单仓存储模式在满足复杂市场需求和快速响应客户需求方面存在一定的局限性。因此，如何在满足各项业务需求的基础上合理分配仓储资源，成为了一个具有挑战性的多目标优化问题。

为实现这一目标，我们提出了一个**综合收益函数**，涵盖了影响仓储布局的关键因素：高级品类收益、件型一致性收益、品类关联度收益和总仓储成本。高级品类收益旨在提升高价值商品的存储与管理效率；件型一致性收益则注重于提升相似件型商品的运营效率；品类关联度收益通过优化相关商品的集中存放以提高交叉销售的机会；而总仓储成本则直接影响仓储方案的经济性。通过赋予各收益项和成本项不同的权重，我们希望在整体收益最大化和成本控制之间实现合理的平衡。

本题目中，选择**模拟退火算法**来求解多仓分配优化问题，主要是因为其在复杂解空间中的强大寻优能力。模拟退火算法通过逐步降低温度的方式减少非优解的接受概率，从而引导解的收敛。该算法具有探索解空间和避免局部最优的特性，特别适合于多目标优化问题。在求解过程中，通过多次迭代和温度递减控制，模拟退火算法能够在庞大的解空间中逐步接近全局最优解，并通过持续的优化改善方案的整体收益。

为了确保模型的实际应用性，我们在模型构建中注重数据的完整性和合理性。具体而言，通过数据准备与处理步骤加载仓库信息、品类关联度、日销量预测、月库存量预测以及件型和高级品类信息，保证输入数据的准确性。同时，将仓储成本、仓容上限、产能上限等从日单位转为月单位，简化了计算维度，确保所有数据在相同的时间单位内进行比较和运算。通过数据结构设计，我们也提高了收益计算的效率和模型的可维护性。

综合来看，本问题分析强调了多仓分配在优化电商仓储资源方面的重要性。通过构建多目标综合收益函数并使用模拟退火算法求解，我们的模型能够提供灵活、平衡的仓储布局方案，以提高整体的运营效率和市场适应性。这一方案不仅适用于不同品类的电商品仓储规划，同时为电商企业在实际运营中的决策提供了有价值的参考。

三、 模型假设与符号说明

3.1 模型基本假设

(1) **数据真实性假设**：通过假设所提供的数据是准确和可靠的，没有严重的错误或缺失。

(2) **季节性规律假设**： 假设库存量的季节性变化规律是可预测的，并且在未来会保持一致。

(3) **线性关系假设**：假设仓租成本、仓容和产能与分配方案之间存在线性关系。

(4) **静态资源假设**：假设仓库的容量和产能是固定的，并且在预测期间不会发生变化。

(5) **数据稳定性假设**：假设品类之间的关联度在预测期间保持稳定。每个品类的件型在预测期间保持不变。每个品类的高级品类分类在预测期间保持不变。

(6) **月份固定假设**：假设每个月固定为 30 天，从而方便仓容上限，产能上限，仓租日成本转换为以月为单位进行计算。

3.2 符号说明

表 2 符号说明

符号	含义
x_{ij}	品类i是否分配到仓库j（1 表示分配，0 表示不分配）
R_j	仓库j的月租成本
I_i	品类i的库存量
S_i	品类i的销量

C_j	仓库j的月仓容上限
P_j	仓库j月产能上限

四、 问题一模型建立与求解

4.1 数据预处理

4.1.1 处理历史销量中的缺失值

在进行销量预测之前，我们需要对历史销量数据进行预处理，以确保数据的完整性和准确性。通过对所提供的历史销量数据进行分析发现，部分品类的销量记录存在缺失，即其记录天数少于 183 天。这种数据缺失可能会对预测模型的效果造成不利影响，因此我们决定采用合理的方法进行数据补全，为模型预测提供高质量的输入。图 1 为历史销量的天数统计。

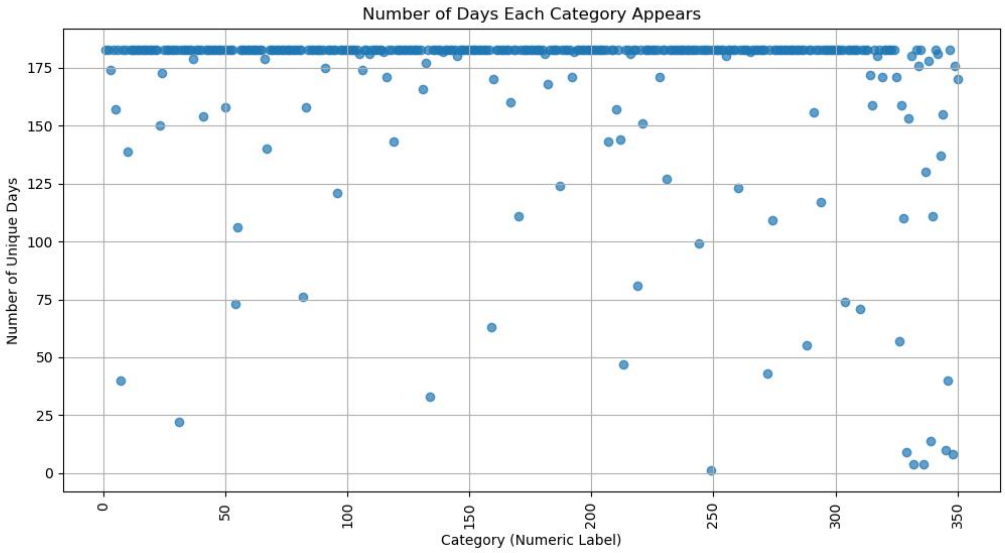


图 2 历史销量天数统计

针对这些缺失值的补全，我们选择了随机森林回归模型作为填补方法。随机森林回归是一种基于多棵决策树的集成学习算法，能够捕捉数据中的非线性关系，并在面对不规则缺失数据时表现出色。与传统的均值填补或线性插值方法相比，随机森林模型更适合处理时间序列数据中的波动性问题。此外，该模型的鲁棒性较强，能够有效应对噪声数据和异常值对填补结果的干扰。

随机森林的回归公式为：

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad (1)$$

其中，$\hat{y}$是最终的预测值，N是随机森林中的决策树数量，$T_i(x)$是第 i 棵树对输入 x 的预测。

在数据处理的具体操作中，我们对所有品类的数据进行了清洗和重构，并与完整的日期范围进行对齐。对于缺失值的部分，我们提取了日期特征，如日、月和星期，以捕捉销量的时间规律。这些特征与已知的销量数据一起作为模型的输入，用于训练随机森林回归模型，并通过该模型预测缺失日期的销量。

在模型训练过程中，我们采用了交叉验证的方法评估模型的性能，确保其在不同数据集上的预测准确性。经过多次验证，模型成功模拟了品类销量的历史趋势，并将预测结果作为缺失数据的补充。通过这一过程，我们生成了覆盖 183 天的完整历史销量数据集。

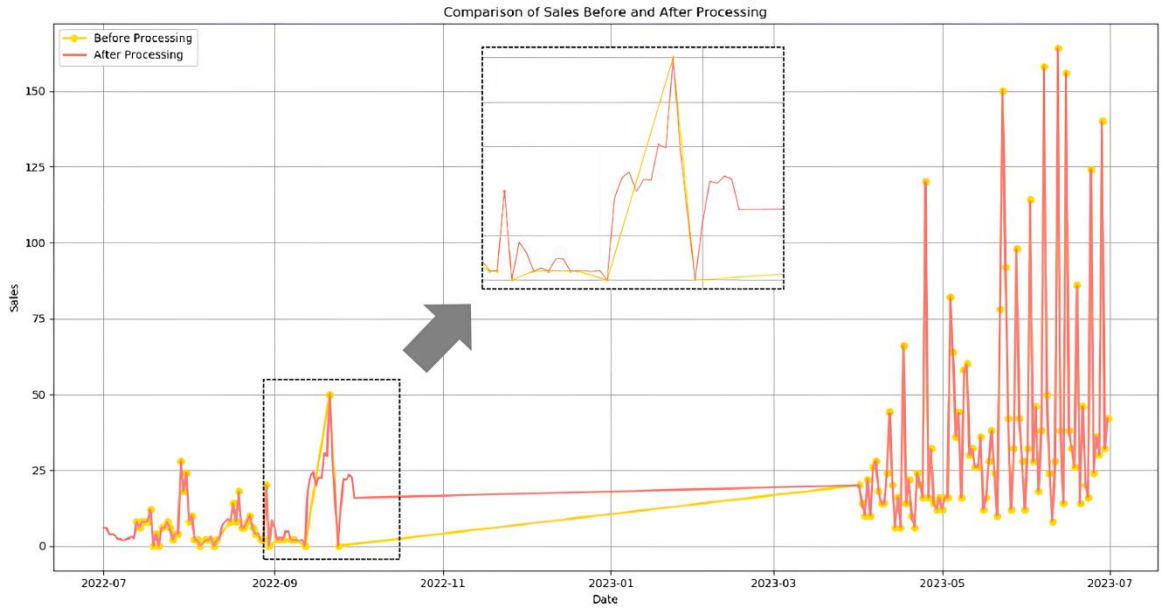


图 3 品类 category10 数据预处理前后对比图

图 2 是展示品类 category10 数据预处理前后对比，在 2022 年 9 月份的销售数据中，可以明显观察到，经过数据预处理后（淡红色曲线），出现了多个拐点和波动，而处理前（黄色曲线）相对平稳。这些拐点的增多说明预处理不仅保留了原始数据的重要趋势，还揭示了更多的细节波动，使得数据的分析价值进一步提升。从局部放大图的展示可以看到，数据处理前的销量曲线变化较为单一，而处理后的曲线在同一时间段内呈现出更为复杂的变化。这些变化使得销售数据的关键点得以更加清晰地展现。整体来看，数据预处理的过程增强了数据的使用性和可解释性。通过揭示更多的销售拐点和趋势变化，预处理后的数据能够更好地支持未来销售趋势的预测与异常检测，提高了数据的分析价值。这表明预处理不仅是数据清洗的步骤，更是发掘数据潜力的重要手段。

这一数据预处理过程为后续的预测和优化模型提供了可靠的数据支持。完整且准确的历史数据不仅提高了预测模型的精度，也为品类的合理分仓规划奠定了坚实的基础。

4.2 库存量预测模型的建立与求解

4.2.1 数据与模型的适配性分析与处理

在参考《电商销量预测及其分仓备货策略研究》[2]这篇文献后对问题一有了较为清楚的思路，先分析附件 1 的数据可以发现，缺失了 2022 年 9 月，10 月，11 月的数据，本文选择先对所有数据进行特征分析，即对数据进行特征分析，观察各品类是否具有季节性或周期性变化，为后续建模提供更多信息。

（1）时间序列稳定性（平稳性）分析：

我们首先采用了 ADF 检验，该检验是一种用于检验时间序列数据是否为平稳序列的统计方法。时间序列的平稳性很重要，因为许多时间序列模型要求输入数据是平稳的。若数据不平稳，需要通过**差分**或其他方法将其转换为平稳序列。

ADF 检验的回归模型表示为：

$$\Delta X_t = \alpha + \beta t + \emptyset X_{t-1} + \sum_{i=1}^{p} \theta_i \Delta X_{t-i} + \epsilon_t \quad 2)$$

Δ 是差分运算符，即 $\Delta X_t = X_t - X_{t-1}$ ，α 是常数项，β_t是趋势项，允许检验序列中是否存在趋势。∅是待检验的参数，若ϕ=0，则表示存在单位根，序列非平稳，θ_i是滞后项的系数，ϵ_t是白噪声误差项。

（2）设置原假设与备择假设：

- **原假设（H0 ）**：设置序列存在单位根，即序列是非平稳的。
- **备择假设（H1 ）**：设置序列没有单位根，即序列是平稳的。

p-value 是在原假设 H0 为真时，观察到的统计量等于或更极端的概率。可以用以下公式表示：

$$p - value = P(T \leq t | H_0) \quad 3)$$

其中，T 是检验统计量的随机变量，t 是在样本中计算出的统计量。

判断标准：

p-value < 0.05：拒绝原假设 (H_0)，认为序列是平稳的。

p-value ≥ 0.05：无法拒绝原假设 (H_0)，认为序列是非平稳的。随机抽取的 60 种品类的检验结果如图 1 所示。其余结果见附件。

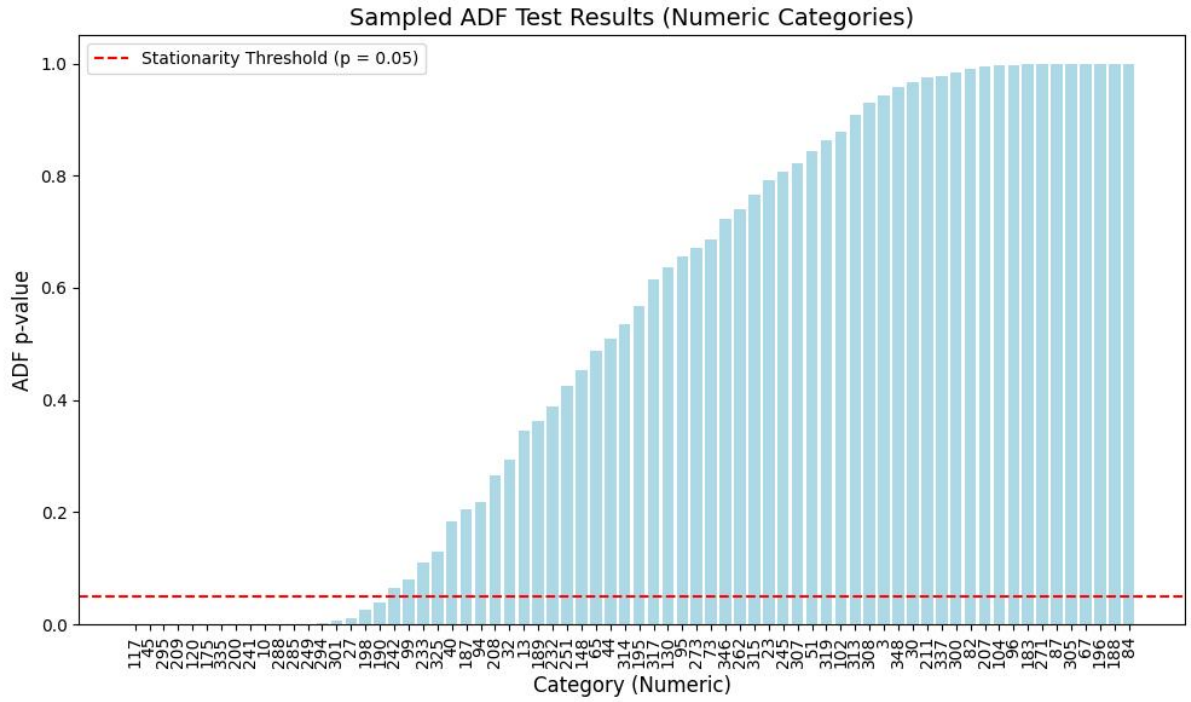


图 4 部分数据的 ADF 检验图

这张 ADF 检验图表明库存量的时间序列数据中存在明显的**非平稳性**，其中超过一半的品类表现出显著的趋势或波动（右侧的高 p-value 类别）。这些类别的 p-value 远高于 0.05，说明这些数据的时间序列非平稳。这些非平稳序列可能代表了存在趋势性或季节性波动的库存品类。

（3）周期性与自相关性检测：

我们下一步采用了自回归条件异方差（Autoregressive Conditional Heteroskedasticity, ACF）分析，该分析是一种时间序列分析方法，主要用于检测序列中是否存在条件异方

差现象，也就是通过观察时间序列的自相关性来判断数据在不同时间段是否存在显著的波动差异。

本文对各品类进行 ACF 分析进而判断数据的自相关性，部分分析结果如图 2 所示。其余结果见附件。

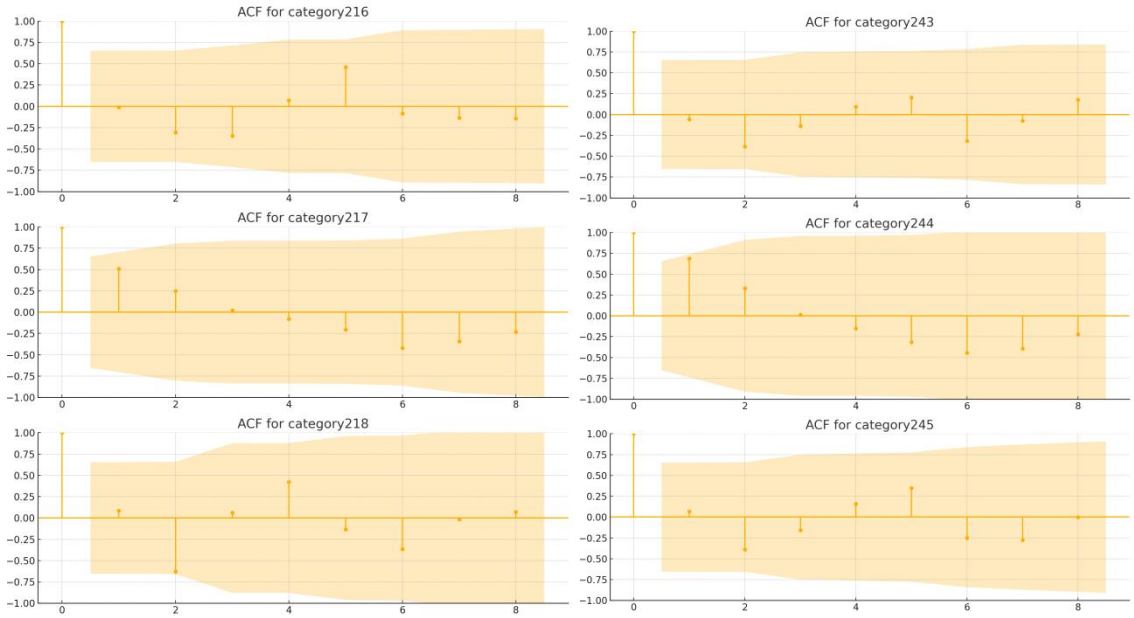


图 5 部分品类的 ACF 图

多个品类（如 216, 217, 219）显示出强烈的滞后 1 自相关，这意味着它们之间可能存在**短期的强趋势或季节性依赖**。例如 218，245 这样的品类，滞后 2 和滞后 4 的自相关值显著，可能存在**短周期的波动性**。例如 243 品类别，似乎呈现出平稳的特征，自相关值较低且快速下降。

（4）缺失值处理：

在附件 1 的数据中，2022 年 9 月、10 月和 11 月的数据缺失。由于时间序列数据通常具有连续性，本文选择线性法插值可以有效地在缺失值与已知值之间建立一种线性关系，以尽可能保持数据的连续性和一致性。

通过对时间序列的特征分析，如平稳性检验和自相关性分析，我们发现多个品类表现出明显的非平稳性及季节性特征。线性法插值可以在这种趋势或季节性波动的基础上进行推断，使得填补的缺失数据能够反映出数据的整体趋势，而不是简单地插入平均值或中位数。

ADF 检验结果显示，许多品类的数据存在显著的非平稳性，p-value 远高于 0.05，表明这些序列可能具有趋势或季节性波动。线性插值在这种情况下，能够在已知数据点的上下文中平滑过渡，反映数据的潜在趋势。同样，ACF 分析结果表明，部分品类存在明显的自相关性。这意味着前期的数据对后期的值有一定的影响，而线性插值可以利用这些信息，确保插值的合理性和有效性。

与其他插值方法相比，线性法插值不需要假设数据的具体分布形式或其他复杂结构。它仅依赖于已有数据点的直线连接，这减少了对模型假设的依赖，使得插值结果更具普适性。

（5）差分处理：

在时间序列分析中，确保数据的平稳性是建立有效模型的前提条件。正如前文所述，经过 **ADF 检验**的分析结果显示，多个品类的库存量数据存在显著的非平稳性，这使得

我们需要采取措施将这些非平稳序列转化为平稳序列。为此，**差分处理**是一种常用且有效的技术。

差分的基本公式为：

$$Y_t^{'} = Y_t - Y_{t-1} \quad (4)$$

其中，$Y_t^{'}$是差分后的时间序列，Y_t是当前时刻的值，Y_{t-1}是前一时刻的值。通过差分，可以有效消除时间序列中的趋势性，使其趋于平稳。

部分差分结果如下图所示。

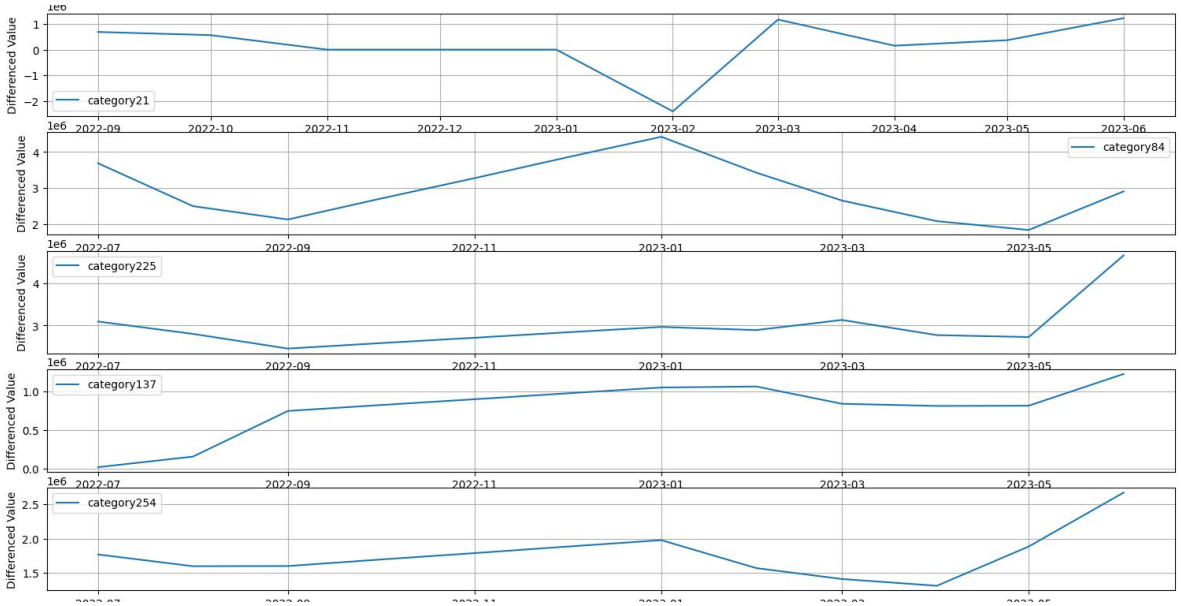


图 6 部分品类差分处理后的 ADF 检验

4.2.2 模型构建与求解逻辑

基于上述特征分析，SARIMA 模型是一个合理且有效的选择，主要原因包括：能够处理时间序列中的季节性波动（适合周期性明显的品类）。通过差分操作将非平稳序列转换为平稳序列，从而满足建模需求。能够捕捉短期自相关性，通过 AR 和 MA 机制来处理这些特征。季节性部分使得 SARIMA 模型在面对缺失数据或波动性强的数据时有更强的鲁棒性。对于分析中提到的这些库存品类的时间序列数据，SARIMA 模型能够综合处理非平稳性、周期性、趋势性以及自相关特征。因此，本文基于 SARIMA 季节性自回归积分滑动平均模型建立库存量预测模型。

4.2.3 SARIMA 模型描述

SARIMA 模型通常用记号表示为 SARIMA(p, d, q)(P, D, Q)[m]，其中：p, d, q 分别是 ARIMA 模型中自回归（AR）、移动平均（MA）阶数和差分次数。P, D, Q 是季节性部分的自回归（SAR）、差分（SD）、和移动平均（SMA）阶数[1]。m 是季节的长度（周期），例如对于每年重复的月度数据，m = 12。

SARIMA 模型是季节性部分与非季节性部分的结合，定义为：

$$\boldsymbol{\Phi_P(B^m)\phi_p(B)(1-B)^d\ (1-B^m)^D y_t = \Theta_Q(B^m)\theta_q(B)\epsilon_t} \quad (5)$$

其中：

y_t表示时间 t 的观测值。

ϵ_t表示白噪声或随机误差项（通常假设为正态分布）。

B 是滞后算子（lag operator），即$B_{y_t} = y_{t-1}$。

$\phi_p(B)$和$\Phi_{\mathrm{p}}(B^m)$分别是非季节性和季节性自回归部分的多项式。
$\theta_p(B)$和$\Theta_{\mathrm{Q}}(B^m)$分别是非季节性和季节性移动平均部分的多项式。
$(1-B)^d$是非季节性差分，用来使数据平稳。
$(1-B^m)^D$是季节性差分，用来消除季节性成分。

4.2.4 基于网格搜索的自动参数优化

在本研究中，为了提升库存预测的准确性，我们使用了**SARIMA（季节性自回归积分滑动平均）模型**。SARIMA 模型的效果在很大程度上依赖于参数的合理设置，因此我们采用了**网格搜索（Grid Search）算法**，对模型的参数进行自动优化。

由于不同品类的库存数据具有各自的趋势和季节性特点，因此每个品类的 SARIMA 模型参数可能不同。为了确保模型在每个品类上的适用性，我们采用网格搜索（Grid Search）对参数进行自动优化。网格搜索遍历所有可能的参数组合，并通过比较每个组合对应的 AIC（Akaike 信息准则）值，选择使 AIC 最小的参数作为最优参数。

AIC 是衡量模型复杂度与拟合效果的标准，计算公式为：

$$\boldsymbol{AIC = 2k - 2ln(L)}$$

其中，k 为模型参数个数，L 为最大似然估计。AIC 值越小，模型的拟合效果越好，同时避免了模型过拟合 。

网格搜索实现步骤:

1.定义参数搜索空间：p, d, q：在[0,1,2]之间搜索。P, D, Q：在[0,1]之间搜索。s：设为 12（月度数据的季节周期）。

2.遍历参数组合：使用 sklearn 库中的 ParameterGrid 函数生成所有可能的参数组合，并逐一尝试拟合 SARIMA 模型。

3.选择最优参数：对每个参数组合，计算对应的 AIC 值。最终选取 AIC 最小的参数组合作为该品类的最优参数。

4.拟合最优模型并预测：使用最优参数拟合模型后，对未来 3 个月进行滚动预测，并将预测结果用于更新库存数据。

4.2.5 模型评估结果与分析

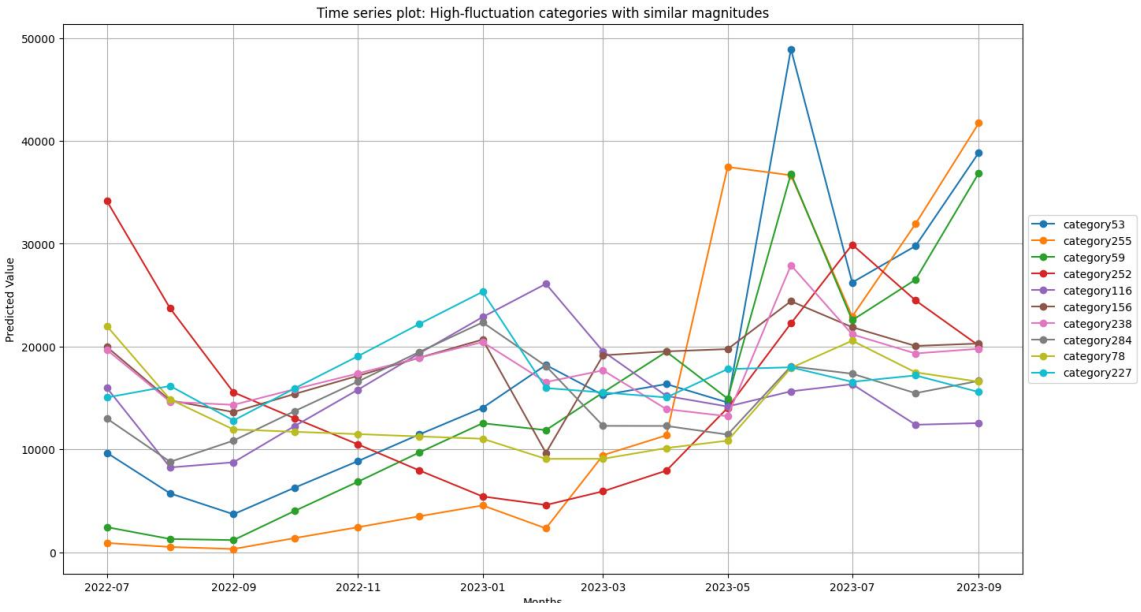


图 7 部分品类的时间序列预测图

基于参数优化的 SARIMA 模型在此预测中表现出色，成功捕捉到大部分商品类别的趋势与季节性特征。对于需求波动剧烈的类别（如 category53 和 category255），模型在捕捉其整体趋势上表现良好：在年中达到需求高峰，之后逐渐回落。这表明，模型成功识别了年中的季节性需求高峰。未来三个月的预测结果表明，不同类别的需求变化趋势各异，为库存管理提供了有效的参考。对于波动大的类别，建议灵活调整库存策略，平稳类别则可采用稳定库存策略以降低成本。

4.3 销量预测模型的建立与求解

在本次销量预测任务中，我们使用了**随机森林回归模型**对未来 3 个月（2023 年 7 月 1 日至 9 月 30 日）每天的销量进行预测。该模型基于**历史销量数据**和**日期相关特征**进行训练，以捕捉销量随时间变化的趋势和周期性波动特性。我们使用的数据涵盖了**2022 年 7 月 1 日至 9 月 30 日**和**2023 年 4 月 1 日至 6 月 30 日**两个时间段的销量记录。

4.3.1 数据与模型的适配性分析

分析历史销量数据后发现，这两个时间段的数据存在季节性和周期性波动。例如，7 月至 9 月可能是消费高峰期，而 4 月至 6 月可能为淡季，销量变化规律较为复杂。此外，由于数据来自两个不连续的时间段，存在一定的不规则性和缺失值。因此，我们选择了随机森林回归模型，它适用于处理非线性数据，并能够在应对数据缺失和噪声方面表现出良好的鲁棒性。

4.3.2 模型构建与求解逻辑

为了准确预测未来的销量，我们对各品类构建了日期特征和滞后特征，并使用随机森林回归模型进行训练和预测。

日期特征的构建：

日（day）：捕捉每月不同日期的销量规律。

月（month）：反映不同月份之间的季节性波动。

星期几（weekday）：识别工作日和周末的销量差异。

滞后特征的构建：

滞后 1 天、7 天、14 天和 30 天的销量值：这些特征帮助模型捕捉销量的短期和长期周期性趋势，如每周的销售高峰或每月的变化。

训练与测试集划分：

我们将数据按时间顺序划分为训练集和测试集，以避免打乱时间序列的结构。训练集用于模型的学习，测试集用于验证模型的预测能力。

逐步预测未来销量：

在训练完成后，使用最后 30 天的历史数据作为基准，通过逐步生成滞后特征，对未来 3 个月内每天的销量进行连续预测。

4.3.3 随机森林回归模型描述

设在某日t的销量为y_t,日期的日、月和星期几分别为d_t, m_t, w_t。滞后 1 天、7 天、14 天和 30 天的销量分别表示为$\text{lag}_1, \text{lag}_7, \text{lag}_{14}, \text{lag}_{30}$。对于未来某日$t+1$，模型的输入特征向量为：

$$[\boldsymbol{X_{t+1}} = [\boldsymbol{d_{t+1}}, \boldsymbol{m_{t+1}}, \boldsymbol{w_{t+1}}, \mathbf{lag_1}, \mathbf{lag_7}, \mathbf{lag_{14}}, \mathbf{lag_{30}}]] \quad 7)$$

基于这些输入特征，随机森林回归模型生成的预测值为：

$$\widehat{y_{t+1}} = f(X_{t+1}) \quad (8)$$

其中，f 是随机森林回归模型构建的回归函数。

4.3.4 模型评估与参数检测

为了确保模型的准确性和稳健性，我们在模型训练完成后，对每个品类在测试集上的预测结果进行了评估，采用了多种评估指标来验证模型的性能：

MAE（平均绝对误差）：

测量预测值与实际值之间的平均绝对误差。计算公式为：

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (9)$$

MAE 越小，说明预测值与实际值的偏差越小。

MSE（均方误差）：

反映预测值与实际值之间误差的平方平均值，对大误差更加敏感。计算公式为：

$$[MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2] \quad (10)$$

MSE 较小时，表明模型对大幅偏差的预测较少。

RMSE（均方根误差）：

通过对 MSE 开平方得到 RMSE，直接衡量误差的大小。计算公式为：

$$[RMSE = \sqrt{MSE}] \quad (11)$$

RMSE 比 MSE 更容易直观地反映误差的大小。

R^2（决定系数）：

评估模型对数据的拟合程度。其取值范围为 0 到 1，值越接近 1，表示模型对数据的解释能力越强。计算公式为：

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (12)$$

其中，$\bar{y}$是实际值的平均值。

表 3 部分品类预测检验参数

品类	MAE	MSE	RMSE	R^2
Category304	1.3507911	3.748737	1.9361656	0.9009371
Category336	0.0136564	0.000541	0.0232735	0.9981120
Category339	0	0	0	1
Category345	0	0	0	1
Category31	0	0	0	1
Category348	0	0	0	1
Category332	0	0	0	1

对于 category304，模型表现良好，R^2 接近于 0.9，表明模型能够较好地拟合该品类的销量变化趋势，预测误差处于合理范围内。这表明该品类的数据具有较好的规律性，模型能够有效捕捉其销量的变化。

category336 的预测效果非常优异，R^2 达到 0.998，MAE 和 RMSE 几乎为 0，说明模型对该品类的销量变化进行了高度准确的拟合。模型的表现表明该品类的历史数据高度稳定，容易预测。

对于 category339、category345、category31、category348 和 category332，所有误差指标均为 0，R^2 为 1。这表明模型在这些品类上的拟合完全正确，预测值与实际值完全一致。这可能是由于这些品类的数据非常稳定，历史销量没有明显波动，或者测试集中销量保持不变，因此模型能够完美预测。

总体来看，模型在这些品类上的表现非常优异，尤其是对于稳定性较强的品类，预测结果完美。这表明随机森林回归模型在处理数据波动较小、具有较强规律性的品类时表现非常出色。对于 category304，虽然预测误差比其他品类稍大，但依然处于较高的预测水平，模型能够为后续的库存管理和分仓规划提供可靠的数据支持。

4.3.5 模型评估结果与分析

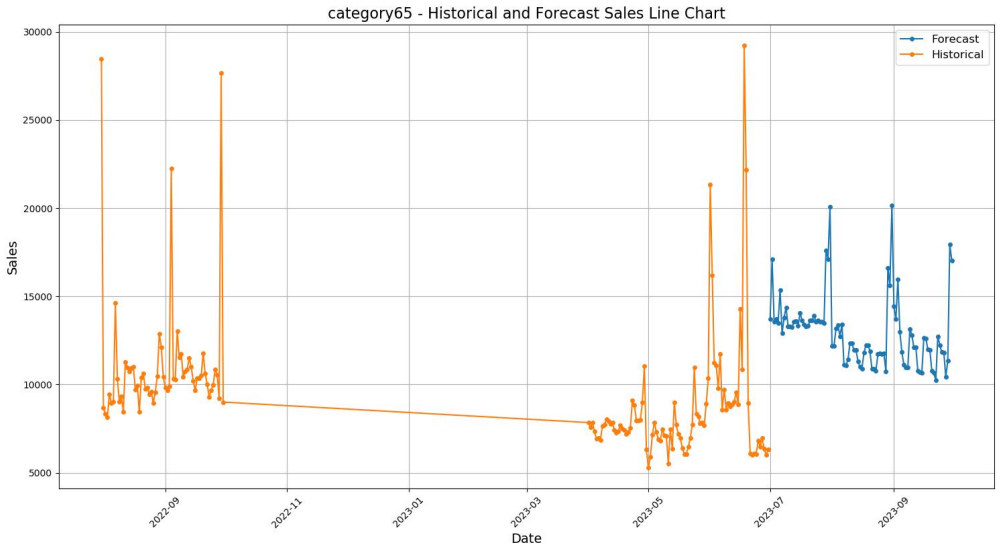


图 8 品类 category65 销量预测情况图

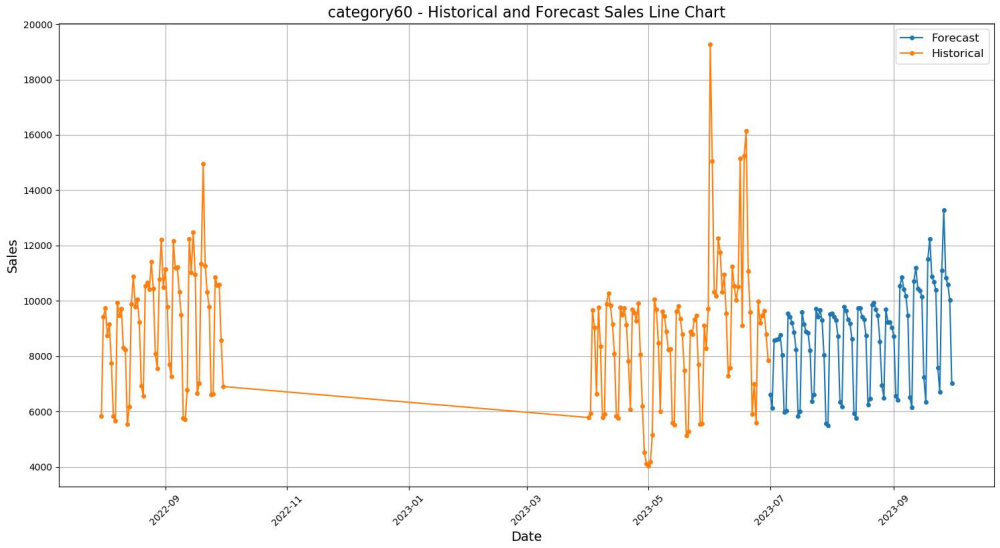


图 9 品类 category60 销量预测情况图

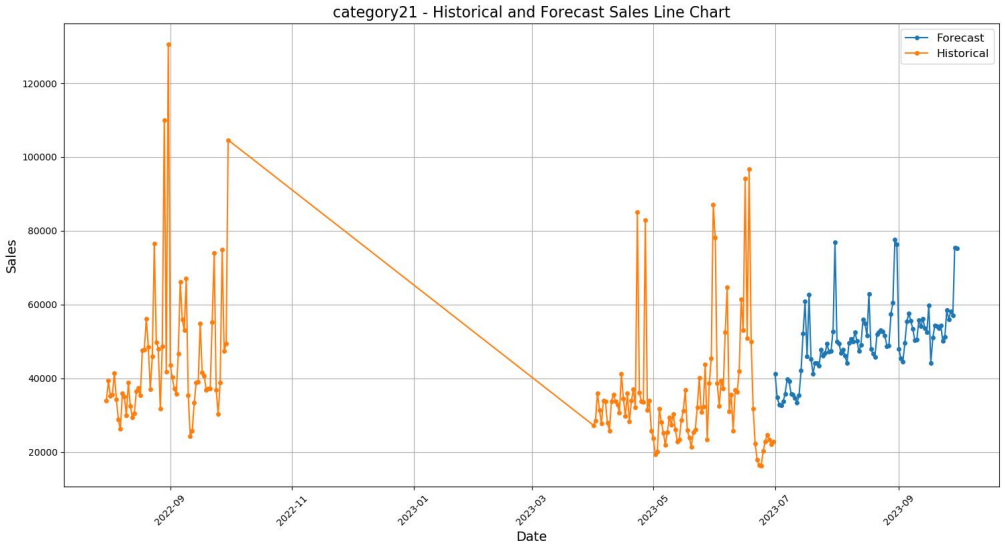


图 10 品类 category21 销量预测情况图

从三张预测结果图可以看出，模型在不同品类的销量预测中表现良好。对于 category65 和 category60，尽管历史销量中存在多次剧烈波动和极端峰值，模型的预测结果能够成功捕捉整体的趋势，并在预测区间内保持合理的波动。这表明模型在处理复杂的销量数据时具备一定的稳健性，能够有效避免因少数异常值导致的预测偏差。在 category21 中，模型不仅延续了销量的上升趋势，还在一定程度上反映了历史数据中的波动性，说明模型对销量变化的捕捉较为准确。尽管个别极端值未完全反映在预测结果中，但这种平滑处理使得预测结果更加稳健，为业务决策提供了可靠的数据支持。总体而言，模型在面对复杂数据时表现出色，既能把握长期趋势，又能控制波动范围，充分展现了其预测能力的优越性。通过进一步优化特征输入，模型在未来的实际应用中有望继续提升其准确性和适应性。

经过测试，我们发现随机森林回归模型在不同品类上的预测效果较为稳定。大部分品类在测试集上的 MAE 和 RMSE 保持在合理范围内，表明模型的预测误差较小。同时，R^2决定系数大部分接近于 1，表明模型能够较好地拟合历史销量数据并捕捉其变化趋势。

多种评估指标的使用确保了模型的稳健性和准确性。在面对不连续的时间段和复杂的销量波动时，模型依然能够提供稳定的预测结果。这为未来 3 个月的库存管理和分仓规划提供了可靠的数据支持。

通过构建基于随机森林回归的销量预测模型，我们成功捕捉了销量的时间规律和滞后关系，并对模型进行了全面的评估。随机森林模型的非线性拟合能力、鲁棒性和逐步预测机制，使其在应对复杂时间序列和不规则数据时表现出色。多维度的评估指标验证了模型的准确性和稳健性，为后续的分仓规划和库存管理提供了强有力的支持。

五、 问题二模型建立与求解

5.1 仓库分配模型的建立与求解

在本次仓库分配任务中，我们构建了一个线性规划模型，以优化每个品类的仓库分配方案。该模型旨在通过最小化仓租成本，同时满足仓库的容量和产能限制，为每个品类分配一个仓库，从而提高仓库的管理效率。

5.2 数据与模型的适配性分析

通过分析现有的仓库和品类数据，我们发现仓库的仓容和产能限制是进行合理分配的关键因素。每个仓库的租金、容量和产能都影响着分仓决策。为了确保模型能够有效反映实际情况，我们采用线性规划方法，适合于解决这类资源分配问题。线性规划模型具备高效求解的特性，能够快速处理大规模的决策问题。

线性规划模型的适用性还体现在其对约束条件的灵活性。通过定义一品一仓、仓容上限和产能上限等约束，模型能够在确保可行解的前提下，实现目标函数的优化。这种特性使得模型在电商环境中具有较强的适应性

5.3 模型构建与求解逻辑

为了准确实现仓库分配，我们建立了一个包含目标函数和约束条件的线性规划模型。

目标函数： 我们的目标是最小化总的仓租成本，目标函数形式为：

$$\textbf{Minimize } Z = \sum_{i \in I} \sum_{j \in J} R_j \cdot x_{ij} \tag{13}$$

其中，R_j表示仓库j的月租成本，x_{ij}是决策变量，表示品类c是否分配到仓库j（1 表示分配，0 表示不分配）。

约束条件： 模型中包含以下关键约束：

一品一仓约束：确保每个品类只能分配到一个仓库，避免资源的过度分散，简化管理流程。

$$\sum_{i \in I} x_{ij} = 1, \ \forall i \in I \tag{14}$$

仓容上限约束：限制每个仓库的库存量不超过其仓容上限，确保仓库的安全性和高效性。

$$\sum_{i \in I} x_{ij} \cdot I_i \leq C_j, \ \forall j \in J \tag{15}$$

产能上限约束：限制每个仓库的出库量不超过其产能上限，以保障在高峰期能够满足客户需求。

$$\sum_{i \in I} x_{ij} \cdot S_i \leq P_j, \ \forall j \in J \tag{16}$$

其中，I_i为品类i的库存量，S_i为品类i的销量，C_j和P_j分别为仓库j的月仓容和月产能上限。

5.4 模型求解与结果分析

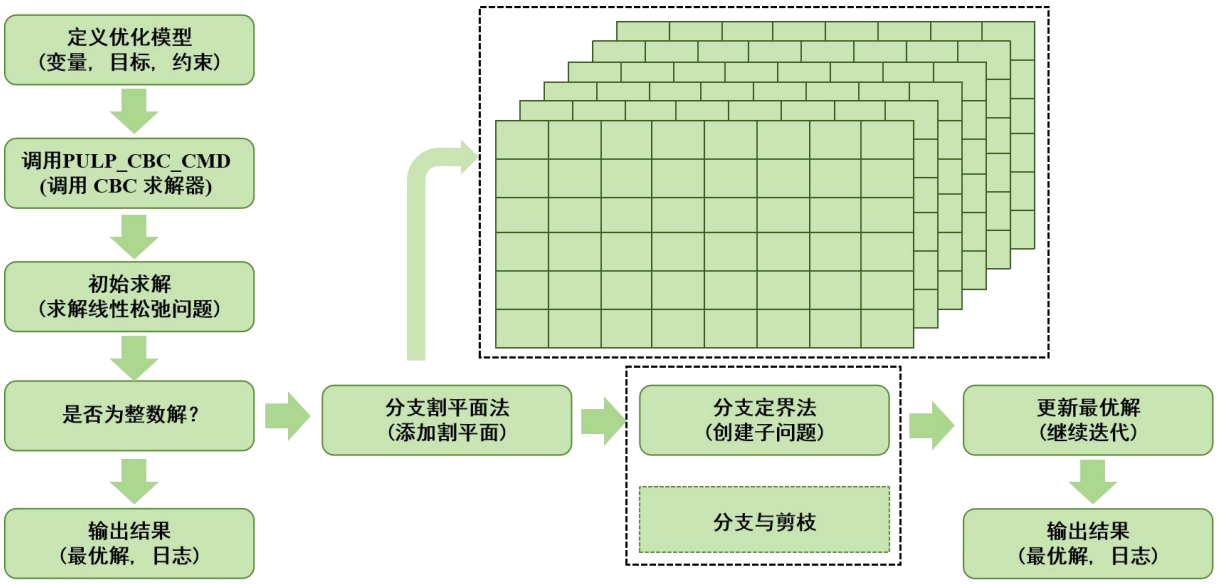


图 11 PULP_CBC_CMD 求解器原理图

在分仓模型中，使用分支割平面法[4]能够有效地优化品类的仓库分配。该方法通过添加割平面来限制非整数解，同时保持所有整数解的可行性。例如，在进行初步求解后，如果发现某个品类的仓库分配并未能有效利用仓库的容量或者导致了产能超限，割平面将帮助排除不合适的分配方案。

分支过程的实施：

在模型实施过程中，求解器首先会对线性松弛问题进行求解，得到初步的最优解。接下来，根据当前的解进行分支，创建子问题以进一步探索可能的分仓方案。每次分支后，求解器会计算新的下界，通过比较当前最优解与新解的下界，判断是否需要继续深入搜索。通过这种方式，可以有效地缩小解空间，从而提高求解效率。

综合考虑的业务目标：

在进行品类分仓规划时，需综合考虑多个业务目标，包括仓容利用率、产能利用率、总仓租成本、品类分仓数及品类关联度等。这些指标共同决定了分仓方案的合理性。例如，合理的仓容利用率意味着仓库空间得到了充分利用，而产能利用率的提高则可以确保出库效率最大化。同时，降低总仓租成本和控制品类分仓数可以减少管理复杂度，提升整体运营效率。

迭代与优化结果：

在每次迭代中，求解器会根据新的割平面更新最优解，并在限制条件的框架下不断优化分仓方案。最终，求解器将输出最佳的分仓结果，包括使用的仓库和存放的品类，确保每个仓库的容量和产能限制得以满足。该结果将反映在结果文件中，并用于进一步的业务分析和决策支持。

在模型构建完成后，我们使用线性规划求解器对其进行求解。求解的过程确保了在满足所有约束条件的前提下，能够找到一个全局最优解。通过对结果的分析，我们可以明确每个品类应分配至哪个仓库，并了解各仓库的使用情况。

表 4 300 秒求解时间的规划结果参数表

指标	参数
Objective value	6977267.16125076
Lower bound	6059643.349

Gap	0.15
Enumerated nodes	14106
Total iterations	218943
Time (CPU seconds)	300.08
Time (Wallclock seconds)	300.08

这组结果展示了优化模型求解的性能与输出情况，目标值（Objective value）为6,977,267.16，表示当前解对应的总成本，这是优化模型所要最小化的关键指标。也就是说，该解代表了一个可行的分仓方案，其总成本达到了接近最优的水平。下界（Lower bound）为 6,059,643.35，这是算法推断出的理论上可能达到的最低总成本，意味着在理想情况下，总成本不会低于该值。通过这些值的对比，我们可以评估当前解的质量。差距（Gap）为 0.15%，反映了当前解与理论最优解之间的相对差异。这么小的差距表明，当前解已十分接近最优解，进一步优化的潜力有限，表明算法已经在较高程度上实现了成本的优化。

求解过程中，共枚举了 14,106 个节点（Enumerated nodes），这说明算法在解空间中进行了大量的探索，以找到最优解。这种枚举操作通常是分支定界（Branch-and-Bound）或类似算法的一部分，用于逐步缩小解的范围。与此同时，算法共迭代了 218,943 次（Total iterations），这反映了优化过程中的计算强度和复杂性——每次迭代都涉及更新解或寻找更优的解。如此多的迭代次数表明问题的规模较大，且求解器在逐步收敛的过程中付出了大量计算努力。

此外，整个求解过程的 CPU 时间和实际时间（Wallclock seconds）均为 300.08 秒，说明求解过程持续了大约 5 分钟，并且计算过程与实际耗时一致。这表明算法在求解过程中没有遇到显著的等待时间或资源竞争问题，从而保证了计算的高效性。这些指标综合表明，优化模型在有限的时间内成功地找到了一个高质量的解，并且该解与理论最优解之间的差异非常小。总体而言，这一分仓方案在成本控制上表现良好，表明算法在成本优化和求解效率之间取得了良好的平衡。

模型求解后的分仓方案能够有效降低整体仓租成本，同时合理利用仓库资源。我们对各品类的分仓结果进行了详细评估，并计算了总的仓容利用率和产能利用率，以进一步分析仓库分配的合理性。

表 5 三次不同求解时间的规划参数对比表

求解时间	目标成本	迭代次数	差距
300.08	6977267.16125076	218943	0.15
400.25	6823616.24231076	264415	0.13
500.17	6823616.24231076	379526	0.13

从表格数据可以看出，不同的求解时间对目标成本、迭代次数和差距（Gap）都有明显的影响。首先，目标成本随着求解时间的增加呈现下降趋势。在求解时间为 300.08 秒时，目标成本为 6977267.16125076；当时间延长至 400.25 秒时，目标成本下降至 6823616.24231076，表明求解器在更长的时间内找到了更优的解。然而，进一步增加求解时间到 500.17 秒后，目标成本保持在 6823616.24231076 不变，这说明求解器在延长的时间内未能找到更低的成本解，可能已经接近最优解。

此外，求解时间的增加伴随着迭代次数的显著增长。在 300.08 秒内，求解器共进行了 218,943 次迭代；延长到 400.25 秒时，迭代次数增加到 264,415 次；而在 500.17 秒时，

迭代次数进一步增加到 379,526 次。尽管迭代次数显著增多，但目标成本在 400.25 秒之后不再下降，这表明随着计算时间和迭代次数的增加，模型的优化效果趋于平缓，即解的质量提升空间有限。

差距（Gap）方面也展现出类似的规律。最初在 300.08 秒时，差距为 0.15%，表明当前解与理论最优解之间有 0.15%的差异；当求解时间增加到 400.25 秒和 500.17 秒时，差距缩小至 0.13%，进一步接近理论最优解。这种差距的减少说明求解时间增加确实使解更接近最优解，但在 400 秒左右达到稳定，延长求解时间对差距的改善有限。

综上所述，这些数据表明，延长求解时间到 400 秒左右时，模型的求解效果已经达到较高的精度，目标成本和差距均趋于稳定，进一步增加求解时间并未带来明显的优化效果。这意味着 400 秒的求解时间在成本与精度上达到了较为理想的平衡，适合在实际应用中作为求解时间的参考值。

5.5 模型评估与应用

为确保模型的准确性与有效性，我们对分仓结果进行了综合评估。通过比较不同品类的分仓效率，我们能够识别出仓库资源利用的潜在问题，并对模型进行进一步的优化调整。此外，通过与实际运营数据的对比，验证模型的预测能力和适应性。

本模型的优势在于其结构简单、求解效率高且易于解释，适合在快速变化的电商环境中进行实时决策。未来，我们将继续探索更复杂的模型，以处理品类之间的关联性和变化性，从而提高分仓方案的优化程度。

六、 问题三模型建立与求解

在第三问中，我们的目标是在多目标优化的背景下，通过模拟退火算法求解电商品类的多仓分配问题，以最大化仓储布局的整体收益。为实现这一目标，我们首先定义了一个综合收益函数，涵盖高级品类收益、件型一致性收益、品类关联度收益及总仓储成本等关键因素，并为这些收益项和成本项赋予不同权重，以合理平衡各收益目标。

基于模拟退火算法的原理，我们逐步降低温度、探索解空间，从而优化收益并控制成本。这一过程通过减少非优解的接受概率，引导解的收敛，并在多次迭代中保留当前最优解，持续改善方案的整体收益。模拟退火算法在复杂解空间中寻优的能力，尤其适合多目标优化问题，确保有效平衡各收益目标。最终，模型输出的最佳分仓方案实现了多品类、多仓库的高效资源配置，为实际电商分仓布局提供了可操作和优化的参考方案。

6.1 数据准备与处理

在解决电商品类的多仓分配问题时，数据准备与处理是至关重要的第一步。我们的目标是最大化仓储布局的整体收益，因此必须确保输入数据的准确性和完整性，以便为模型提供高质量的基础。

首先，我们加载多个数据源，包括仓库信息、品类的关联度、日销量预测、月库存量预测以及品类的件型和高级信息。这些数据涵盖了仓储布局的核心要素，有助于全面分析和理解当前的资源配置。

为了便于后续计算，我们将仓租成本、仓容上限和产能上限从日单位转换为月单位。这一转换不仅能够简化后续的收益和成本计算，还能使得所有计算在同一时间维度上进行，提高模型的可比性和一致性。

在分析销量数据时，我们计算每个品类的最高月销量，确保在分仓规划中考虑到潜在的需求峰值。此外，通过获取每个品类的预测库存量最大值，可以为实际分仓决策提供更为合理的依据。

对关联度数据的处理则通过创建字典来实现，便于在计算收益时快速检索相关信息。这种数据结构的设计不仅提高了计算效率，还减少了代码的复杂度，使得整个算法更具可读性和可维护性。

综上所述，精心的数据准备和处理过程为后续的决策制定打下了坚实的基础，确保模型在多目标优化的求解中具备高效性和可靠性。

6.2 模型建立与求解过程

6.2.1 决策变量与目标函数设定

在本模型中，决策变量被定义为每个品类是否被分配到特定的仓库。这样的设计是为了灵活地应对每个品类在不同仓库中的需求特性，同时允许多个品类共享同一仓库。这一策略不仅能够提高仓储资源的利用效率，还能使得仓储布局更加灵活，适应市场需求的变化。

目标函数的构建是模型设计的关键部分。我们通过最大化综合收益函数来实现仓储布局的优化，该函数包括高级品类收益、件型一致性收益、品类关联度收益及总仓储成本。这一综合收益函数的设计有几个优点：

多目标平衡：通过为各项收益赋予不同的权重，我们能够在多个优化目标之间进行合理的平衡。例如，高级品类收益和件型一致性收益的权重设置使得品类在仓库中的分布既考虑到集中存储的优势，又保持了相同件型的管理便利性。

收益最大化：目标函数的设计直接与电商的盈利模式相关，确保我们的求解方向与实际业务目标一致。通过优化这些收益项，我们能够帮助企业实现更高的市场竞争力。

成本控制：在综合收益函数中，加入总仓储成本的考虑，使得模型不仅关注收益的增加，也重视成本的控制。这一设计能够有效避免因过度追求收益而导致的成本失控问题。

定义决策变量$x_{i,j}$表示品类i是否分配到仓库j：

$$x_{cw} = \begin{cases} \mathbf{1}，\text{品类}\boldsymbol{i}\text{分配到仓库}\boldsymbol{j} \\ \mathbf{0}，\text{品类}\boldsymbol{i}\text{不分配到仓库}\boldsymbol{j} \end{cases} \tag{17}$$

6.2.2 目标函数

我们希望通过最大化以下综合收益函数 F，使仓储分配达到最优效果，主要考虑品类关联度、件型一致性、高级品类收益和仓储成本的影响。目标函数可表示为：

$$\textbf{Maximize } F = w_1 \cdot H + w_2 \cdot U + w_3 \cdot A - w_4 \cdot C \tag{18}$$

其中：

H表示高级品类收益；

U表示件型一致性收益；

A表示品类关联度收益；

C表示总仓储成本；

w_1，w_2，w_3，w_4为各项的权重系数。

6.2.3 收益计算逻辑

收益计算逻辑的设计旨在全面反映各类收益对整体仓储布局的贡献，确保模型在评估分仓方案时具备准确性和可靠性。

在计算高级品类收益时，我们关注品类集中存储在同一仓库的情况。这种设计有助于避免高级品类的分散，提高资源的使用效率。通过增加重复高级品类的收益，我们鼓励在仓储布局中实现高价值品类的集中存放，从而增强整体盈利能力。

高级品类收益计算：

高级品类收益H定义为将相同高级品类集中存储在同一仓库中的收益，以避免高级品类的分散：

$$\boldsymbol{H = \sum_{j} \sum_{i \in I_j} h_i \cdot x_{ij}} \tag{19}$$

其中：

h_i为品类i的高级品类收益值；

I_i表示分配到仓库j的品类集合。

件型一致性收益则衡量相同件型品类的集中存储效果，促使管理过程更加便捷，提高物流效率。通过奖励相同件型的集中存放，我们不仅提高了库存管理的便利性，也降低了跨仓调度的需求。

件型一致性收益U衡量将相同件型的品类集中存储在同一仓库的效果，提高一致性：

$$\boldsymbol{U = \sum_{j} \sum_{i \in T_j} t_i \cdot x_{ij}} \tag{20}$$

其中：

t_i为品类i的件型一致性收益；

T_j 为分配到仓库j的同件型品类集合。

品类关联度收益计算的设计，通过考虑品类之间的关联性，帮助减少物流成本。关联度较高的品类放置在同一仓库中，可以降低配送成本，提高客户服务质量。这一设计思路的好处在于，它强化了仓储布局的灵活性，使得业务能够更好地应对市场变化。

品类关联度收益：

品类关联度收益A衡量关联度较高的品类放置在同一仓库中的情况，以降低履约成本：

$$\boldsymbol{A = \sum_{j} \sum_{i,k \in A_j} r_{ik} \cdot x_{ij} \cdot x_{kj}} \tag{21}$$

其中：

r_{ik}表示品类i和品类k的关联度；

A_j为分配到仓库j的品类集合对。

总成本的计算则通过惩罚逻辑确保每个仓库的资源分配均衡，避免因某一仓库资源过剩或不足而导致的效率低下。这一设计能够有效提升仓储布局的稳定性，增强模型的实用性。

总仓储成本：

仓储成本C表示使用各个仓库带来的日成本：

$$\boldsymbol{C = \sum_{j} c_j \cdot y_j} \tag{22}$$

其中：

c_j为仓库 j 的日租成本；

y_j为仓库 j 是否使用的指示变量，即 $y_j = 1$ 表示仓库 j 被使用，否则 $y_j = 0$

通过上述收益计算逻辑的合理设计，我们能够全面评估不同分仓方案的有效性，从而为最终的决策提供有力的数据支持。

件型一致性收益：

6.2.4 约束条件

(1) 仓容上限约束

仓容上限约束确保每个仓库的总库存不超过其存储容量C_j ：

$$\sum_{i} \boldsymbol{p_i \cdot x_{ij} \leq C_j \cdot y_j} \ \forall \boldsymbol{j} \tag{23}$$

其中：

p_i表示品类i的预测库存量。

(2) 产能上限约束

产能上限约束确保每个仓库的出库量不超过其出库产能P_j ：

$$\sum_{i} \boldsymbol{s_i \cdot x_{ij} \leq P_j \cdot y_j} \ \forall \boldsymbol{j} \tag{24}$$

其中：

s_i表示品类i的日销量预测值。

(3) 品类分仓限制

每个品类最多可存储在 3 个仓库内：

$$\sum_{j} \boldsymbol{x_{ij} \leq 3} \ \forall \boldsymbol{i} \tag{25}$$

6.2.5 决策变量约束

定义各决策变量的取值范围，确保每个变量的有效性：

$$\boldsymbol{x_{ij} \in \{0,1\}, y_j \in \{0,1\}} \ \forall \boldsymbol{i,j} \tag{26}$$

6.2.6 求解过程

首先，为每一个品类 i 初始化一个随机的仓库分配方案。根据题目要求，每个品类最多可以分配到三个仓库，这样能够在满足库存和产能约束的情况下，为仓储布局提供一定的灵活性。具体的初始化步骤为：

对每个品类 i，随机选择 1 到 3 个仓库进行初始分配。

生成一个“当前解”，该解包含了所有品类的仓库分配情况。

在模拟退火算法的迭代求解过程中，我们使用如下参数与方法进行优化：

温度参数：初始温度 T_0 设置为 1000，随着每次迭代降低温度，逐渐减少对不优解的接受概率，使解不断收敛[5]。

迭代更新。

生成新解：在当前解的基础上，随机选择一个品类并对其分配仓库进行小幅调整，如增加或减少一个仓库，确保调整后仓库总数不超过三个。

计算新解的综合收益：利用综合收益函数计算新的分仓方案的收益。

接受准则：若新解的收益更高，则接受该解；若收益更低，则根据以下公式以一定概率接受：

$$\mathbf{P(accept\ new\ solution) = exp(\frac{current\_profit - new\_profit}{T})} \quad \textbf{(27)}$$

其中：

current_profit 是当前解的综合收益。

new_profit 是新解的综合收益。

T 是当前温度（控制接受较差解的概率）。

这样可以在初始阶段探索更多解空间，避免局部最优。

温度递减：每次迭代后按冷却速率 α=0.99 更新温度 T，直到温度降到接近零或达到最大迭代次数。

算法如下：

Algorithm

1. $\boldsymbol{T_n=T_0}$
2. V_n = calculate_value(I)[0]
3. For _ in range(M)
4. I[id] = 1 –I[id]
5. V_{new} = calculate_value(I)[0]
6. Δ = V_{new} - V_n
7. if Δ < 0 or random() < $e^{-\frac{\Delta}{T_n}}$:
8. V_n = V_{new}
9. else: I[id] = I[id]
10. T_n *= Co_R

其中，T_n表示当前温度，T_0表示初始温度，V_n表示当前个体适应度，I 表示待优化的个体，M 表示最大迭代次数，id 表示每一个品类，V_{new}表示变异后个体的适应度值，Δ表示新个体与当前个体的适应度差异，Co_R表示降温过程。

6.2.7 最优解输出

在所有迭代完成后，最终输出最佳分仓方案和对应的综合收益值。此时的解是基于多目标收益优化和仓储约束下的最优方案，可以提高整体的分仓效益。

表 6 一品多仓的分仓结果

分仓方案	总成本	关联度收益	高级品类收益	件型一致性收益
一品多仓	27502455.91	390167.564	7860000	5670000

在本次模型求解中，得到了多个关键指标的结果，包括总成本、关联度收益、高级品类收益和件型一致性收益。以下是对每个指标的详细分析。

首先，总成本为 27,502,455.91 元。这一数值反映了电商在仓储管理中需要承担的费用，包括仓租、运营成本等。合理的总成本是保证企业盈利的基础，过高的成本会直接影响整体利润。因此，本模型在设定分仓方案时，致力于在优化收益的同时，控制总成本，以实现更为合理的资源配置。

接下来，关联度收益为 390,167.564。该收益指标体现了关联品类在同一仓库中的存放效果，能够减少配送成本，提高物流效率。通过优化品类的存放布局，企业可以降低跨仓调度的需求，从而提升整体运营效率。尽管这一数值相较于其他收益项较低，但它在整体仓储管理中的重要性不容忽视，尤其是在提高客户满意度和优化服务响应时间方面。

第三，高级品类收益达到了 7,860,000。这一指标突出了高价值品类集中存放所带来的经济效益，集中管理高级品类可以提高资源利用率和管理效率。通过优先配置这些品类到合适的仓库，企业能够在满足市场需求的同时，最大限度地发挥其经济效益，进而提升整体盈利能力。

最后，件型一致性收益为 5,670,000。这一收益指标反映了将相同件型的商品集中存放的效果，能够提高管理的便捷性，降低库存管理成本。件型一致性使得仓库运营更加高效，有助于实现快速周转和准确的库存管理，从而提升整体供应链的灵活性。

综上所述，本次结果分析显示，模型不仅成功优化了总成本，同时也通过多项收益的合理配置实现了整体效益的提升。高级品类收益和件型一致性收益的贡献显著，显示出本模型在仓储布局优化中的有效性。通过关注各个收益指标，企业可以在未来的仓储管理中更加精准地调整资源配置，以实现更高的经济效益和运营效率。

6.3 模型评估与分析

在模型评估与分析阶段，我们重点验证了模型的有效性、解的合理性和应用可行性。通过引入高级品类、件型一致性和品类关联度等收益项，模型量化了分仓方案的核心目标，并为这些收益赋予不同权重，确保优化方向符合实际业务需求。这种综合收益函数设计实现了收益的最大化，同时优化了整体分仓效益。

在成本控制与仓储约束方面，我们明确量化了总仓租成本，并限制每个品类的仓储数量，确保合理分配资源，避免仓储瓶颈。模拟退火算法的应用使得模型在多目标优化中具备灵活性，权重参数的调整增强了模型的适应性，使其能够更好地应对不同的业务需求。

此外，算法的收敛性与解的稳定性得到保障，模拟退火过程通过温度控制实现了解的稳定与合理，适用于大规模的分仓优化任务。模型输出的最佳分仓方案直观展示了仓库分配结果及综合收益值，为实际决策提供了有力支持，提升了仓储资源的利用效率。

尽管模型表现优异，但仍存在局限性，如对参数设置的敏感性和权重选择的依赖性。未来可以引入其他优化算法，或增加自适应权重调整机制，以更好适应动态的仓储需求。

6.4 分析不同方案中各业务指标的表现

不同方案主要指问题二中的一品一仓模式和问题三中的一品多仓模式。我们将根据题目描述的四类重要数据，即总成本、关联度收益、高级品类收益、件型一致性收益，作为业务指标进行衡量，以评估各自的表现和优势。

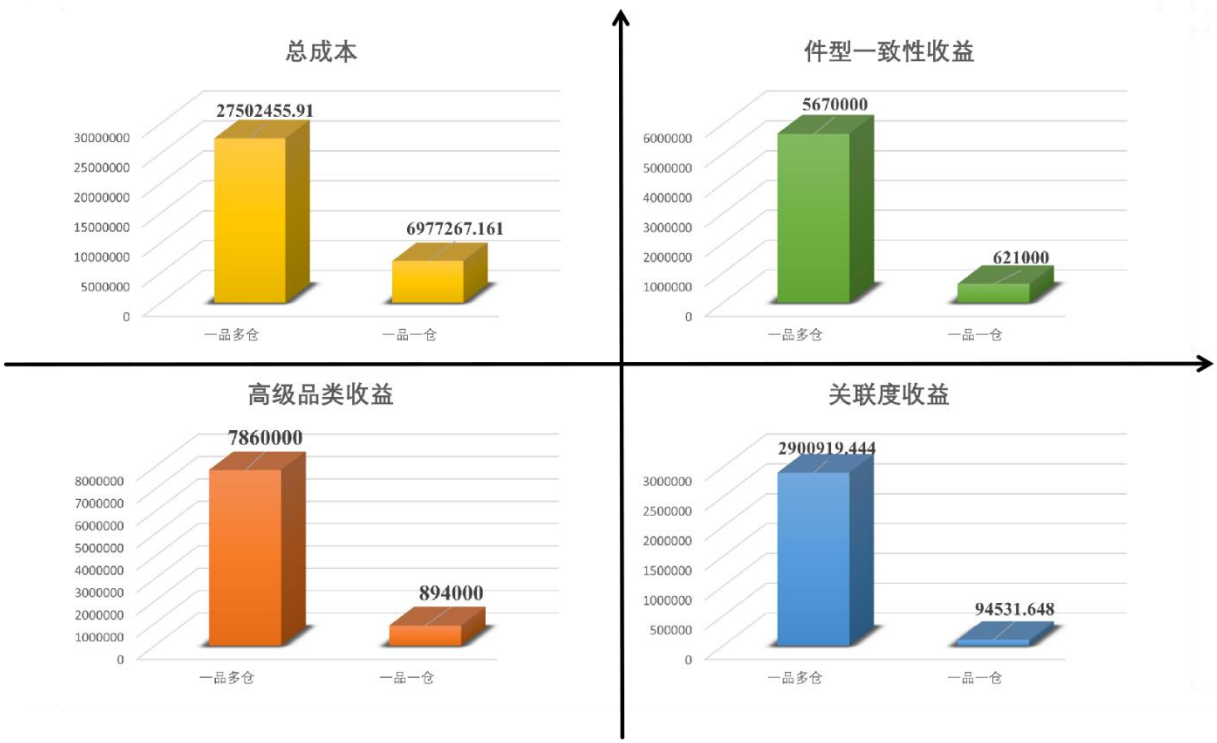


图 12 “一品一仓”和“一品多仓”的各项业务指标对比图

在本次模型求解中，对“一品多仓”和“一品一仓”两种模式下的四个关键指标进行了对比分析，包括总成本、关联度收益、高级品类收益和件型一致性收益。这些指标不仅反映了不同仓储模式的成本和收益差异，也为选择适合的仓储方案提供了数据支持。以下是对每个指标的详细分析。

总成本：一品多仓模式的总成本为 27,502,455.91 元，而一品一仓模式的总成本为 6,977,267.161 元，多仓模式的成本几乎是单仓模式的四倍。总成本是衡量仓储方案经济性的重要指标，它包括仓租成本、运营成本、物流成本等。多仓模式的高成本主要源于多个仓库的租赁和维护费用。虽然分仓可以降低部分运输距离，但在实际运营中，由于各个仓库之间的协调、运输和管理需求增加，导致总成本显著上升。管理多个仓库不仅需要更多的人力，还增加了操作复杂性，特别是在库存和订单管理方面，这进一步提升了运营成本。

在一品一仓模式中，集中管理一个仓库的成本较低，仓租和维护费用大幅减少，且由于无需多仓之间的调配，整体运输成本也更易控制。这种模式简化了管理流程，有助于降低成本，是适合较小市场或需求相对稳定的选择。然而，单仓模式在满足快速市场需求变化方面存在劣势，可能难以实现高效的区域配送。因此，从总成本的角度看，一品一仓具有显著的成本优势，但在需求分布广泛的市场中，其成本节省可能会受到配送时效的限制。

关联度收益：一品多仓模式的关联度收益为 390,167.564 元，而一品一仓模式为 94,531.648 元，前者的收益大约是后者的四倍。关联度收益反映了商品之间的交叉销售潜力，即将相关商品存放在同一仓库中，可以在履约过程中集中发货，减少包裹数量，降低物流成本，同时增加客户在一次购买中选购多种相关商品的可能性。多仓模式的高关联度收益表明，它能够更好地利用商品间的关系，优化仓储布局，实现更高效的履约。通过集中存放关联商品，一品多仓模式在交叉销售和降低配送成本方面更具优势。

相比之下，一品一仓模式的关联度收益较低，虽然可以通过集中管理简化仓储操作，但无法充分发挥商品关联度带来的价值。在单仓模式下，所有商品存放在同一地点，缺

少地理上的分散优势，导致无法利用区域市场需求的细微差异，影响了交叉销售机会。因此，虽然单仓模式能够减少管理复杂性，但在关联度收益方面处于劣势，难以充分挖掘商品间的协同效益。

高级品类收益：在高级品类收益上，一品多仓模式的收益为 7,860,000 元，而一品一仓模式为 894,000 元，多仓模式的收益是单仓模式的近十倍。高级品类收益是通过集中特定高价值商品而实现的收益提升。在一品多仓模式下，多个仓库能够覆盖更广泛的地理区域，使得高端商品更易接触到潜在高价值客户群体。分仓策略可以针对不同区域的消费特点实施差异化库存管理，提高高端商品的可获得性。此外，分仓策略还可以防止高端商品缺货，确保市场需求波动时仍能提供及时供应。

相较之下，一品一仓模式的高级品类收益较低，因其集中存储限制了高端商品的市场接触面。单仓模式在地理分布上缺乏灵活性，可能导致高端商品的物流时间延长，难以满足高端客户对服务效率的期望。这不仅影响客户满意度，也可能降低品牌形象。因此，一品一仓模式的高级品类收益远低于多仓模式，不利于企业在高端市场上的拓展和品牌建设。

件型一致性收益：一品多仓模式的件型一致性收益为 5,670,000 元，而一品一仓模式仅为 621,000 元，多仓模式在该收益项上高出单仓模式约 9 倍。件型一致性收益反映了相似件型商品在同一仓库中的存放效率。将同一件型的商品集中存放，能够提高包装和发货的便捷性，有助于减少订单错误率，提升发货效率。在一品多仓模式下，不同仓库可以根据件型特征进行商品分布，从而提升包装和配送效率。这种策略不仅减少了运输和包装过程中的错误，还可以显著降低运输成本。

而在一品一仓模式下，所有商品都集中在一个仓库，件型一致性的分布难以灵活调整，可能导致发货时的操作效率降低。特别是在订单量较大或商品种类繁多的情况下，单仓模式的作业压力增大，可能影响整体订单处理效率。因此，虽然一品一仓模式在管理复杂性上较低，但在实现件型一致性收益方面效果不如一品多仓。

综合分析这四个指标，一品一仓的优势在于成本低、管理简单，适合较小或需求稳定的市场。单仓模式可以集中资源、降低管理费用，并简化运输协调。然而，这一模式的灵活性不足，无法快速响应市场需求的变化，特别是在高端商品销售和关联商品的集中发货方面表现欠佳。

一品多仓模式则具有显著的市场覆盖和灵活性优势。通过多仓布局，企业能够更好地接触不同区域的消费者，尤其是高端客户群体，同时优化关联商品的集中存放，提升交叉销售的可能性。在件型一致性方面，多仓模式通过灵活的商品分布提升了包装和发货效率。然而，这一模式的缺点在于较高的运营成本和管理复杂性。多个仓库带来的成本压力较大，增加了协调难度，需要更成熟的管理机制和资源投入。

因此，企业在选择仓储模式时，需要综合考虑市场需求、产品特性、区域覆盖和管理能力。一品一仓适合成本敏感、管理资源有限的企业，而一品多仓更适合追求市场扩展、高效供应链和服务优化的企业。这种权衡有助于在降低成本和提升收益之间找到平衡点，确保仓储策略符合企业的长远发展目标。

七、 模型分析检验

7.1 灵敏度分析

灵敏度分析有助于理解输入特征对各模型结果的影响。在库存和销量预测模型中，通过评估特征重要性和调整 SARIMA 模型参数，可以识别对预测结果影响较大的因素，帮助企业优化关键特征，从而提高预测的准确性。在仓库分配的线性规划模型中，灵敏度分析通过调整仓库容量、成本等参数，观察对最优分仓方案的影响，识别关键资源和

约束。模拟退火算法的灵敏度分析则关注温度衰减率和初始温度的设定，评估这些参数对收敛速度和解质量的影响。这些分析有助于优化模型参数，增强解的稳健性，使企业在面对市场波动时具备更好的应对能力。

7.2 误差分析

7.2.1 问题一的误差分析

问题一的误差主要来源于数据补全和模型预测。数据补全过程中，销量数据的缺失可能导致补全误差，尤其是在销售波动较大的时间点，如促销期间，补全后的数据可能未能真实反映实际情况。模型预测误差也可能出现，尤其对于波动频繁的品类，补全模型可能无法准确捕捉这些波动趋势。因此，问题一的预测结果虽然较为准确，但在高波动品类上可能仍存在一定误差。

7.2.2 问题二的误差分析

问题二中的线性规划模型假设仓库容量和需求为静态，未考虑实际变化带来的误差。此外，线性模型的简化可能忽略了非线性特征，使得模型与实际情况间存在一定的差异。同时，约束条件的松弛和求解时间限制可能影响解的精确性。数据的准确性对线性规划解的可靠性也有较大影响，如果问题一的数据预测存在误差，将进一步影响分仓方案的合理性。

7.2.3 问题三的误差分析

问题三使用模拟退火算法优化多仓分配方案。模拟退火引入的随机性会导致解的不稳定性，每次运行结果可能不同。此外，目标函数的简化在合并多个收益和成本项时可能忽略一些市场波动因素，影响实际解的效果。参数设定和数据准确性也是影响解质量的重要因素，若权重设置不准确，可能导致收益优化方向偏差。

7.3 整体误差分析

在整个问题求解中，误差来源主要来自数据补全、模型假设、算法随机性和参数设定。问题一的预测误差会影响后续问题的输入数据，从而在问题二和问题三中引发连锁误差。这表明模型的各部分紧密关联，强调了在模型构建和求解过程中保持数据准确性和模型灵活性的重要性，以减少误差对整体决策的影响。

八、 模型评价与推广

8.1 模型的优点

随机森林回归模型具有强大的非线性关系捕捉能力，能够处理复杂的特征交互，使得对多维数据的分析更加灵活和精准。它能有效处理非线性关系，减少过拟合的风险，并对异常值和缺失值具有较强的鲁棒性。这种模型通过集成多棵决策树，减少了过拟合的风险，同时提高了对新数据的预测准确性。

SARIMA 算法在时间序列数据建模方面表现出色，能够有效捕捉历史数据中的趋势和季节性变化。其自回归和移动平均的组合，使得 SARIMA 能够为时间序列数据提供稳健的预测，特别是在存在显著时间依赖性的情况下。因此，随机森林与 SARIMA 的结合，使得预测结果更为准确，并增强了对未来需求波动的适应能力。

采用线性规划模型优化仓库分配方案，这种方法的优势在于能够精确满足资源约束条件，如仓库的容量和产能限制。线性规划提供了一种系统化的决策支持工具，使得优化目标（如最小化仓租成本）能够在数学上严谨地表达。这一特性确保了分仓方案的合

理性和可操作性，为企业在资源配置上的决策提供了科学依据。通过线性规划，企业能够实现成本最小化或利润最大化，从而提升运营效率和竞争力。

模拟退火算法是一种基于随机搜索的优化技术，灵感来源于物理退火过程。其优点在于能够在复杂的解空间中寻找近似最优解，特别适合处理组合优化问题。通过在搜索过程中接受一定的劣解，模拟退火能够有效避免局部最优，从而提高全局搜索能力。在电商品类的多仓分配问题中，模拟退火算法可以灵活适应多种约束条件，快速找到有效的仓库分配方案，帮助企业在动态市场环境中优化资源配置，提升客户满意度和运营效率。

整体而言，将随机森林回归、SARIMA 算法、线性规划和模拟退火结合起来，可以形成一个综合性解决方案。这样的组合模型能够发挥各自的优势，处理复杂的预测与优化问题。例如，利用随机森林和 SARIMA 算法进行准确的销量和库存预测，再通过线性规划模型优化仓库分配，最后使用模拟退火算法进一步精细化解决方案。这种多层次、多维度的策略能够为企业提供全面的数据支持和决策依据，帮助其在竞争激烈的市场中保持领先地位。

8.2 模型的不足

在本次预测任务中，数据补全误差是需要特别关注的问题。实际应用中，商品库存量和销量的波动性和季节性可能也是重要的因素，由于部分品类的历史销量记录存在缺失，在补全这些缺失值的过程中可能引入误差，尤其是在销售趋势波动较大或数据不稳定的品类中。例如，在重大促销日（如“双十一”、“双十二”）或节假日，销量常出现短期急剧上升，这类突发波动通常难以通过常规的补全方法（如随机森林回归）精确还原。因此，缺失数据的补全往往只能基于整体趋势，而较难完全反映销售的突发变化，这使得在特定日期的销量可能被低估或高估。但本文未能考虑到这些因素的影响，一定程度上影响了模型的准确性。

由于第二问模型使用线性规划方法进行求解，模型中的目标函数和约束条件均为线性关系，这意味着模型假定仓租成本、容量、产能等指标与分配方案间呈线性关系。然而，实际仓租成本和产能往往具有非线性特征，例如产能的利用率可能呈现边际递减效果。线性近似导致的误差会影响模型输出的精确性，使得最终的分仓方案在实际应用中可能需要进一步调整。

8.3 模型的推广

本文提出的综合性解决方案在特定业务目标的实现上具备广泛的应用潜力。我们可以推广的应用场景有：

(1)电商需求预测与库存管理：

针对电商行业，结合随机森林回归和 SARIMA 算法的预测能力，可以优化库存管理策略。企业可以根据历史销量数据及市场趋势，灵活调整库存水平，降低缺货和过剩的风险。在此基础上，利用线性规划优化仓库分配，确保资源高效利用。模拟退火算法可用于在动态市场环境中快速响应需求变化，进一步提升客户满意度。

(2)制造业生产调度优化：

在制造业中，结合上述模型可以有效优化生产调度。通过随机森林和 SARIMA 算法分析生产需求，预测未来的原材料需求和产出量，随后应用线性规划模型进行生产资源的合理分配。模拟退火算法可以在多种生产约束条件下探索最优调度方案，确保生产效率的同时降低成本，从而提高整体运营效率。

参考文献

[1] 赵广富,寇玮华,刘俊,等.基于 Prophet-SARIMA 模型的唐包线曹妃甸西站货运量预测研究[J].综合运输,2024,46(09):157-163.DOI:10.20164/j.cnki.cn11-1197/u.2024.09.030.

[2] 李景.电商销量预测及其分仓备货策略研究[D].南京大学,2019.DOI:10.27235/d.cnki.gnjiu.2019.001925.

[3] 王之威.电商物流分仓备货优化研究[D].北京交通大学,2016

[4] 苏欣欣, 伊廷刚, 秦虎. 分支定价割平面法求解带时间窗和人力分配的车辆路径问题[J]. 交通运输工程与信息学报, 2021.

[5] 董灵波.基于模拟退火算法的森林多目标经营规划模拟[D].东北林业大学,2016.DOI:10.27009/d.cnki.gdblu.2016.000042.

队伍编号	MCB2302202
赛道	A

基于深度学习的坑洼道路检测和识别研究

摘 要

道路的状况对于交通安全至关重要。坑洼道路是指路面出现了凹凸不平、有大大小小的裂缝和洞坑等问题，不仅会给车辆和行人带来不便，还可能导致交通事故的发生。随着近年来科技的进步，深度学习技术迅速发展，为坑洼道路检测任务提供了新的解决方案。本文结合所给的图像文件，提取了图像特征，建立了识别图像道路坑洼与否的模型，并从不同维度进行模型评估，并利用该模型对测试集中的图像进行分类识别。

针对问题一，需要建立一个识别道路坑洼的模型，由于数据集中 normal 和 potholes 类的样本数量不平衡，所以对 potholes 类数据进行**翻转**和**四等分裁剪**操作以达到**平衡数据集**的效果。在图像特征提取方面，本文使用 **VGG16 获取图片特征向量**并使用 **PCA 降维**，同时分析图像颜色分布直方图和使用 **Canny 算法检测轮廓特征**。为了提高模型准确率，本文引入两个深度学习模型 **Yolov5** 和 **Yolov8**，同时加入一个基于集成学习的机器学习模型。并采用**加权投票**的方法将**机器学习、Yolov5** 和 **Yolov8** 的结果进行集成，权重分别为 0.2、0.4、0.4。这样可以有效解决两个深度学习模型分类结果歧义问题。

针对问题二，对模型进行了完整的训练和评估，从**指标评价、鲁棒性**和**生成对抗样本**三个方面进行了分析。分别考察了**损失函数变化、精确度**和**召回率、高斯噪声鲁棒性、生成对抗样本**及识别等，得出了模型具有一定的鲁棒性和抗对抗样本能力的结论。

对于 Yolov5 模型，其各项**损失函数随着迭代的增加而逐渐降低、精确度和召回率整体呈上升趋势**，具体见图 22；在噪声数据上，有 4 个样本前后识别结果不一致，**分类准确率为 97.010%，鲁棒性较强**；不考虑对抗攻击强度因素，在对抗样本上，**仅有 1 个对抗样本前后识别结果不同，抗对抗样本能力较强**。

对于 Yolov8 模型，其**损失函数同样随着迭代周期增加而降低**，具体见图 24，**F1** 在置信度为 0.453 时达到**最大值 0.59**，**精确度**随置信度增加呈上升趋势，**在 0.914 置信度达到 1.0**；在噪声数据上，9 个样本前后识别不一致，分类准确率下降了 2.325%，**为 96.678%，具有一定鲁棒性**；不考虑对抗攻击强度因素，在对抗样本上，**仅有 2 个对抗样本前后识别结果不同，抗对抗样本能力较强**。

综合上述评估指标，**本文所构建的模型识别率较高、分类准确、鲁棒性较好，且具有一定的抗对抗样本的能力**。

针对问题三，将训练好的模型用于识别所给训练集中，对于正常道路需要标记为 1，对于坑洼道路需要标记为 0，并将结果存入到“test_resulu.csv”文件中。

关键词：Yolov5　Yolov8　集成学习　VGG16　Canny 算法　生成对抗样本

目录

一、 问题重述 1
1.1 问题背景 1
1.2 问题提出 1
二、 问题分析 1
2.1 问题一的分析 1
2.2 问题二的分析 2
2.3 问题三的分析 2
三、 符号说明 2
四、 模型建立与求解 2
4.1 问题一的模型建立与求解 2
4.1.1 数据预处理 3
4.1.2 图像特征提取 4
4.1.3 模型建立 7
4.1.4 基于集成学习的机器学习模型 8
4.1.5 基于 Yolov5 的深度学习模型 9
4.1.6 基于 Yolov8 的深度学习模型 13
4.2 问题二的模型建立与求解 16
4.2.1 模型训练 16
4.2.2 模型的指标评价 16
4.2.3 模型的鲁棒性 19
4.2.4 添加高斯噪声 19
4.2.5 生成对抗样本 20
4.3 问题三的模型建立与求解 23
五、 模型评价与推广 24
5.1 模型优点 24
5.2 模型缺点 24
5.3 模型推广 24
六、 参考文献 24

一、 问题重述

1.1 问题背景

基于计算机视觉的坑洼道路检测和识别是一个具有实际应用意义的问题。及时检测和识别坑洼道路可以帮助城市和交通管理部门快速定位和修复道路问题，从而提升道路的安全性和可行性。通过减少坑洼道路的存在，可以减少车辆事故和损坏，并提高驾驶者和乘客的舒适度。对于自动驾驶车辆而言，准确检测和识别坑洼道路至关重要。自动驾驶车辆需要实时感知道路状况并作出相应的决策，以确保安全行驶。通过坑洼道路的检测和识别，自动驾驶车辆可以调整车速、避免碰撞或选择更平整的道路。除了日常生活，坑洼道路检测和识别在地质勘探、航天科学和自然灾害等领域的研究和应用具有重要意义。

随着近年来科技的进步，深度学习技术迅速发展，为坑洼道路检测任务提供了新的解决方案。在识别坑洼的过程中，深度学习可以提取坑洼的轮廓、纹理和形态等特征，并将其转化为更易于分类的表示形式。例如，一些研究者采用基于深度学习的方法对道路图像进行分类，将其分为正常和坑洼两类。他们利用深度学习模型从大量带有标签的训练图像中学习坑洼图像的特征，并通过训练模型进行分类预测。此外，一些研究者利用迁移学习的思想，利用在大规模图像数据上预训练的通用模型（如 ImageNet 上的预训练模型），学习坑洼图像的特征表示。他们通过微调或仅训练分类层，将预训练模型的知识迁移到坑洼图像分类任务中，以提高分类性能。

总而言之，深度学习技术在坑洼道路检测任务中具有重要作用。它能够自动提取关键特征，并利用迁移学习等方法提升分类准确性。这些方法为解决坑洼道路检测问题提供了更可靠和有效的解决方案。

1.2 问题提出

本次赛题提供了道路图像数据，对此进行分析、特征提取和建模，根据建立的模型自动识别一张新的道路图像的坑洼状态。基于如上要求，完成下列任务：

问题一：利用所提供道路图像数据，提取图像特征，建立一个识别图像道路是正常还是坑洼的模型，要求识别率高、速度快并且分类准确。

问题二：训练问题一中所构建的用于识别道路坑洼与否的模型，并从不同维度评估该模型。

问题三：利用上述问题中构建并训练的模型判断测试集中的道路图像是正常或者坑洼，并将识别结果保存至“test_result_csv”文件中。

二、 问题分析

2.1 问题一的分析

问题一要求结合所给的图像文件，提取图像特征，在此基础上建立一个识别道路坑洼与否且性能好的模型。通过分析发现，所给的图像文件中 normal 类和 potholes 类的样本数量相差较大，因此需要在原来 potholes 类数据集上进行翻转和四等分裁剪操作来扩充 potholes 类数据集，以到达平衡数据集的效果。

在图像特征提取方面，1)使用深度学习 VGG16 获取图片的特征向量并使用主成分分析法进行降维；2)分析图像的颜色分布直方图；3)使用 Canny 算法检测图像轮廓特征。

在模型建立方面，为了提高模型的识别率和分类准确度我们引入两个深度学习模型 Yolov5 和 Yolov8，但由于这两个模型的分类会存在歧义问题，所以我们使用集成学习的方法并引入机器学习。通过加权投票的方法，将机器学习、Yolov5、Yolov8 的权重分别设置为 0.2、0.4、0.4，这样就可以有效解决两个深度学习模型分类结果歧义问题，即当深度学习模型分类结果相同时，采用深度学习结果；当深度学习模分类结果不同时，采用机器学习结果。

2.2 问题二的分析

问题二中首先要完整对模型的训练，接着从不同的维度对模型进行评估。在评估上我们从指标评价、鲁棒性、生成对抗样本三个维度进行了分析。在指标评价上，我们展示了各损失函数随着训练周期增长的变化规律，以及模型的精确度和召回率；在模型鲁棒性上，在原始数据集中添加高斯噪声，并将原分类结果和加入噪声后的分类结果进行对比，从而评估模型的鲁棒行；在生成对抗样本上，采用遗传算法生成 35 张有坑洼图片在 Yolov5 和 Yolovv8 上的对抗样本，并将对抗样本进行识别，找到前后变化的对抗样本，如果样本数量不多，则认为具有抗对抗样本的能力，有一定的安全性。

2.3 问题三的分析

在问题三中，需要我们利用已经训练好的模型去识别训练集中的坑洼图像，并将结果存入到“test_resulu.csv”文件中，对于正常道路需要标记为 1，对于坑洼道路需要标记为 0。

三、 符号说明

名称	解释
X	图像特征向量
$f(\cdot)$	优化目标函数
$F(\cdot)$	图像识别模型
c	权衡攻击成和最终图像的权重系数
l_{model}	与识别模型相关的损失函数
$l_{metric}(x+\delta)$	模型对样本$x+\delta$的置信度

四、 模型建立与求解

4.1 问题一的模型建立与求解

在问题一中面临两个关键任务：首先，需进行图像特征提取以区分正常类和坑洼类两个类别；其次，需要构建一个专用模型，以便对图像进行分类，以确定道路状态是否正常或存在坑洼。

4.1.1 数据预处理

(1) 平衡数据集

当目标检测模型面临类别不平衡的数据集时，存在过度拟合样本更多类别、忽视小样本类别的问题。这种情况下，模型往往倾向于偏向于预测数量较多的类别，而对于数量较少的类别则表现出相对较低的关注度和准确性。数据扩增成为了该领域中一种常用的解决方案。在本次赛题中，我们发现其提供的道路图像数据集存在 normal 类 266 张、potholes 类 35 张的不平衡情况，为此，我们对 potholes 类数据集进行了扩增。具体而言，我们通过将 35 张 potholes 类道路图像进行翻转，并进行四等分裁剪的方式，得到了更多的数据集。这一方法能够有效地缓解模型对数量较多类别的偏向，提高对数量较少类别的关注度和准确性。



图 1 翻转操作图



图 2 裁剪操作图

对图片进行预处理的目的是增加 potholes 类的数据集，因此裁剪的过程中只保留有坑洼的图片，**如图 2 中排除了无坑洼的图片**。

(2) 标识坑洼

为了应对本次赛题中的坑洼图像识别任务，我们采用了 labelimg 工具，通过给 potholes 类添加 potholes 标记，以便用于 Yolo 模型的训练。

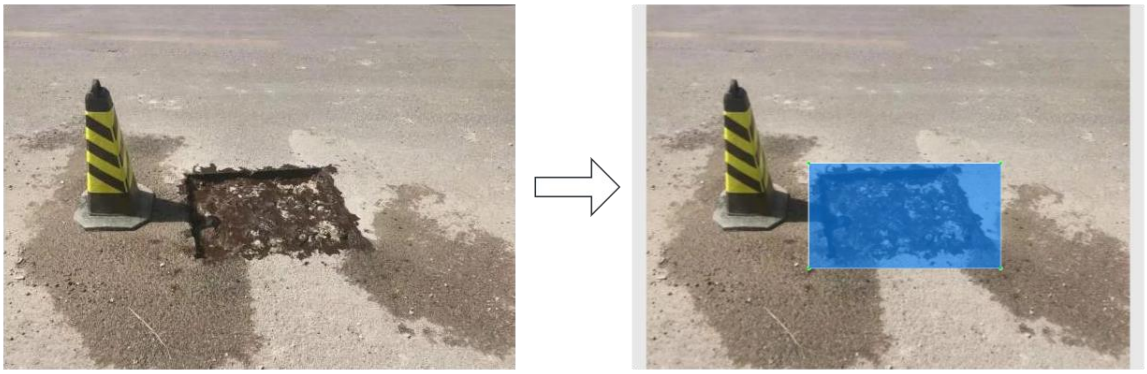


图 3 标识坑洼图

4.1.2 图像特征提取

在提取图像特征方面我们进行了三个方面的分析：

1、使用深度学习 VGG16 得到每张图片的特征向量，接着使用主成分分析法将其降维到一个长度为 100 的特征向量$X_1=(x_{1_1},x_{1_2},\cdots,x_{1_{100}})$；

2、分析图像的颜色分布直方图得到图像 RGB 上三个维度上长度为 256 的特征向量，即$X_i=(x_{i_1},x_{i_2},\cdots,x_{i_{256}}),i=2,3,4$；

3、首先使用 Canny 边缘检测算法找到有用的边缘，接着在 Canny 检测到的边缘中查找封闭的轮廓，最后计算出封闭轮廓面积占该图片面积的比例，得到一个一维特征向量$X_5=(x_5)$。

最后利用这三个方面得到的特征向量，接着便可以构建出每张图片对应的特征向量$X=(X_1,X_2,X_3,X_4,X_5)=(X_{1_1},\cdots,X_{1_{100}},X_{2_1},\cdots,X_{2_{256}},X_{3_1},\cdots,X_{3_{256}},X_{4_1},\cdots,X_{4_{256}},X_{5_1})$。并利用该特征向量来区分道路是正常还是坑洼。

（一）深度学习模型提取图像特征

VGG16 模型在 2014 年的 ImageNet 图像分类与定位挑战赛中确实取得了显著的成绩，展现了其出色的性能。其主要优点之一是其结构的简洁性和深度。这种深度使得模型能够捕捉到图像中更复杂和更高级的特征。

VGG16 特意使用小的3×3卷积核，这与许多其他模型使用的较大卷积核不同。小的卷积核有几个优点。首先，它们减少了参数数量，这有助于防止过拟合，同时保持了模型的计算效率。其次，使用多个小卷积核的连续卷积层可以模拟较大卷积核的效果，但是带有更少的参数和更高的非线性，这使得模型可以学习更复杂的特征。

此外，VGG16 的设计也确保了每一个卷积层的输出与前一个卷积层具有相同的空间维度（宽和高），这意味着仅当达到池化层时，特征图的尺寸才会减半。使用2×2大小、步幅为 2 的 max 池化方式可以确保每一层的输出尺寸是前一层的一半。

VGG16 由若干卷积层和池化层按顺序堆叠而成，通过这种方式，VGG16 能够形成一个深度网络结构，这对于捕捉图像中的细微和高级特征至关重要。

当使用 VGG16 进行特征提取时，并不需要得到类别的概率分布，所以此处我们移除了最后的 Softmax 激活函数。修改后 VGG16 进行特征提取的具体结构图如下：

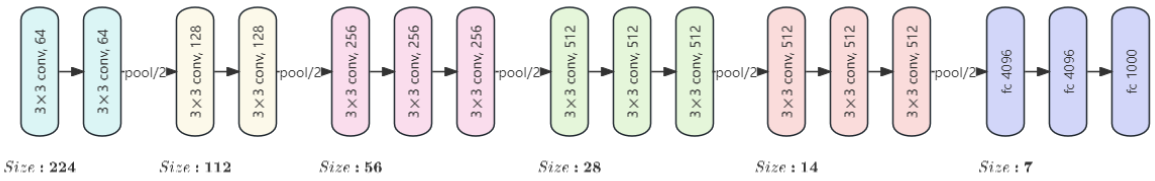


图 4 用于特征提取的 VGG16 结构图

使用 VGG16 进行完特征提取后得到了长度为 1000 的特征向量，然而高维数据会导致更高的计算成本，因此选择使用主成分分析法对该特征向量进行降维，最后得到长度为 100 的特征向量 $X_1=(x_{1_1},x_{1_2},\cdots,x_{1_{100}})$。

（二）颜色分布直方图

颜色分布直方图是一种用于描述图像中颜色分布的重要特征。通过研究直方图的形状和峰值可以反映出主要色调分布和对比度情况。

在绘制颜色分布直方图时，首先，计算出每张图片在 RGB 三个通道上的像素分布情况；接着，按 normal 类中的 R 通道、normal 类中的 G 通道、normal 类中的 B 通道、potholes 类中的 R 通道、potholes 类中的 G 通道、potholes 类中的 B 通道进行分类；最后，将每个类中的分布进行累加并绘制出直方图。得到的颜色分布直方图如下：

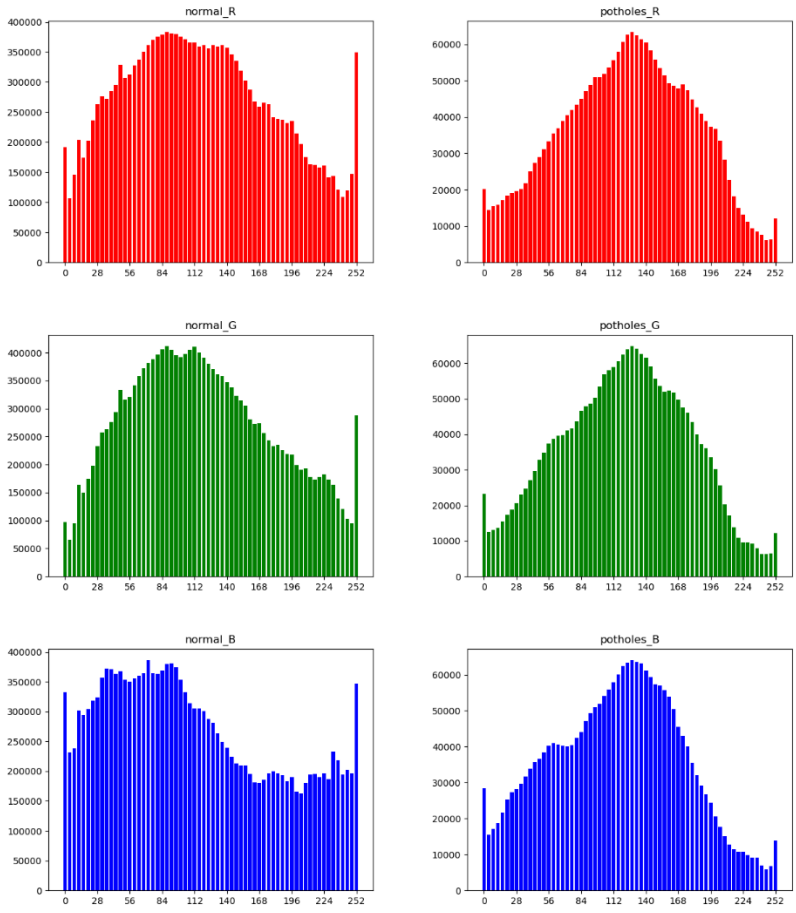


图 5 颜色分布直方图

图 5 中左侧的直方图为 normal 类颜色分布直方图，右侧的直方图为 potholes 类颜色分布直方图。

由此可以得到：

potholes 类直方图的整体分布比 narmal 类直方图的整体分布更靠右，其原因是 potholes 类中坑洼处更多，这些区域由于收到反光的影响，像素值普遍偏高；

通过计算各颜色分布直方图的标准差，得到 normal 类颜色分布直方图的标准差分别为 23391.56、26302.91、20491.49，potholes 类颜色分布直方图的标准差为 4349.04、4627.55、4516.47，对比发现 normal 类的标准差明显比 potholes 类的标准差大，说明 normal 类像素分布更加均匀、对比度小，potholes 类像素分布不均匀、对比度大。

最后将每张图片 RGB 通道像素分布作为特征向量$X_i=(x_{i_1},x_{i_2},\cdots,x_{i_{256}}),i=2,3,4$。

（三）Canny 检测图像轮廓特征

normal 类和 potholes 类最明显的区别为没坑和有坑，直觉上认为道路有坑处会有更明显的轮廓形状，因此使用 Canny 算法对图片进行了轮廓检测，以提取出图像的轮廓特征。其基本流程如下：

1、噪声去除：由于边缘检测对噪声非常敏感，因此首先需要使用高斯滤波器对图像进行平滑处理。高斯滤波基于高斯函数，其形式为$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}$。其中，$\sigma$是标准差，决定了滤波器的宽度。这个函数的离散近似被用作卷积核。

2、计算梯度强度和方向：使用 Sobel 算子计算图像的梯度。Sobel 算子是两个3×3的矩阵，分别是S_x和S_y。分别用于计算图像x方向像素梯度矩阵G_x和y方向像素梯度矩阵G_y。最后梯度大小通过$G=\sqrt{G_x^2+G_y^2}$计算，方向通过$\theta=\arctan(G_x,G_y)$计算。

3、非极大值抑制：为了得到细致的边缘，需要在梯度方向上进行非极大值一致。在这一步中会检查像素是否是其邻域中沿着梯度方向爱的局部最大值。如果不是，像素被设为 0。

4、设置高阈值和低阈值来确定边缘：基于梯度强度，像素被分类为强、弱或非边缘。强边缘像素的强度高于高阈值，弱边缘像素的强度在两个阈值之间，而非边缘像素的强度低于低阈值。

5、边缘跟踪：弱边缘像素被标记为真实边缘或非边缘像素。基于连通性，如果边缘像素周围有一个强边缘像素，则该像素被标记为真实边缘像素；否则，它被抑制。

调用 Canny 算法对文件“potholes16.jpg”进行识别，结果如下：



图 6 原图（左）和轮廓图（右）

从图 6 中可以看出，坑处的轮廓信息相比于周围的环境更加明显，这与我们的猜测一致。

最后在每张图的轮廓图中寻找封闭的轮廓，计算出所有封闭轮廓总面积与图片面积的比值，并作为一个特征向量$X_5=(x_5)$。得到了 normal 类的平均封闭轮廓面积占比为 0.04899，potholes 类的平均封闭轮廓面积占比为 0.09545，有着显著的区别。

4.1.3 模型建立

为了提高模型的识别率和分类准确度，在建立模型时采用了集成学习中的 bagging 算法来搭建模型。尽管集成学习会对多个个体学习器进行学习，但是由于可以并行学习，所以在训练时间上并不会损失过多。

在个体学习器上选择异质的学习器，即个体学习器的种类并不完全相同。为了避免多个学习器集成后结果反而变得更糟，对于个体学习器需要其具有一定的准确性，同时需要个体学习器间具有差异。因此我们选择了一个 Yolov5 模型、一个 Yolov8 模型、一个基于集成学习的机器学习模型作为个体学习模型。

结合策略上采用加权投票法。虽然可以利用提取的特征使用机器学习算法进行分类，但效果仍然有限，所以在权重上我们将机器学习模型权重设为$\alpha_1=0.2$、Yolov5 模型权重设置为$\alpha_2=0.4$、Yolov8 模型权重设置为$\alpha_3=0.4$。当$\alpha_1X_1+\alpha_2X_2+\alpha_3X_3\geq 0.5$则分为 normal 类、$\alpha_1X_1+\alpha_2X_2+\alpha_3X_3<0.5$则分为 potholes 类。

这样的权重选择就可以解决两个深度学习模型造成分类不同时难以决策的问题：当两个深度学习模型分类结果相同时，结果即为深度学习分类结果；当两个深度学习模型分类结果不同相同时，结果即为机器学习分类结果。

基于集成学习的机器学习模型的结构与上述相同，选择了一个支持向量机 SVM 模型、逻辑回归模型、CART 决策树模型作为个体学习器。结合策略上同样采用了相对多数投票法。

问题一的模型结构如下：

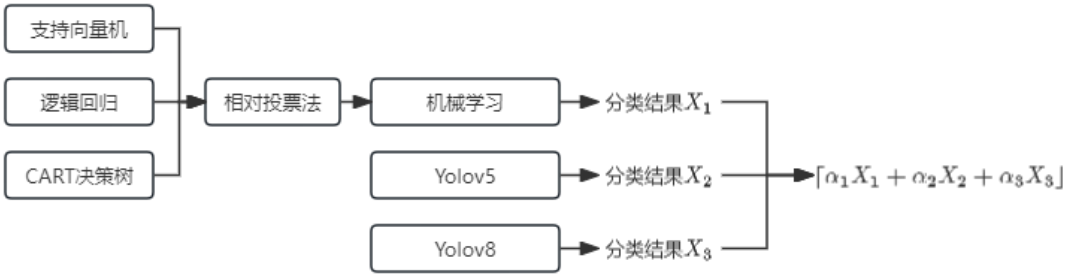


图 7 模型结构图

4.1.4 基于集成学习的机器学习模型

在 Bagging 中，每个个体学习器都是在不同的训练集上独立地进行训练。这些训练集是通过自助采用方法从原始训练集中有放回地抽取得到，因此每个训练集都是从原始训练集中有重复地采样得到的，以增强模型的多样性。最后通过相对投票法得到最终的分类结果。

在训练集上，事先选出的 80%图像作为训练集，每张图片经过特征提取得到的特征向量用于训练个体学习器。剩下 20%图片则作为测试集，用于评估模型的分类效果。

在个体学习器上我们选择了支持向量机 SVM、逻辑回归、CART 决策树这三个个体学习器。

支持向量机是一种用于二分类和多分类问题的有监督学习算法，其目标是通过找到具有最大间隔的决策超平面来实现最大间隔分类。找到最大间隔决策边界可以通过求解对偶问题得到支持向量机的最优解，所以支持向量机算法实际上是通过求解一组二次优化问题来寻找最优的超平面和分类决策函数。其数学原理如下：

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ &\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \alpha_i \geqslant 0, i = 1, 2, \cdots, N \end{aligned} \tag{1}$$

求解后即可得到决策超平面：

$$f(x) \quad = w^T \phi(x) + b = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)^T \phi(x) + b = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \tag{2}$$

其中核函数$k(x_i, x_j)$采样高斯核函数，即$k(x_i, x_j) = e^{(-\frac{\| x_i - x_j \|^2}{2\sigma^2})}$。

逻辑回归是用于二分类问题的有监督学习算法。其假设函数使用了线性回归的形式，并通过 Sigmod 函数进行映射：

$$h_\theta(x) = sigmoid(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{3}$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} (y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))) \tag{4}$$

使用最大似然估计，通过最小化损失函数$J(\theta)$来估计参数θ。此处选择的优化算法

为梯度下降法。通过最大化似然函数来调整模型参数，使得模型对观测样本的预测概率尽可能接近真实标签的概率。

CART 决策树与一般决策树的区别在于 CART 树是一颗二叉树，在进行分类任务时其选择特征采用基尼系数，基尼系数代表了模型的不纯度，基尼系数越小，不纯度越低，特征越好。

$$Gini(D)=\sum_{i=1}^{n}p(x_i)*(1-p(x_i))=1-\sum_{i=1}^{n}p(x_i)^2 \tag{5}$$

$Gini(D)$ 反映了从数据集 D 中随机抽取两个样本，其类别标记不一致的概率。因此 $Gini(D)$ 越小，则数据集 D 的纯度越高。

4.1.5 基于 Yolov5 的深度学习模型

Yolov5 的结构如图 8 所示。

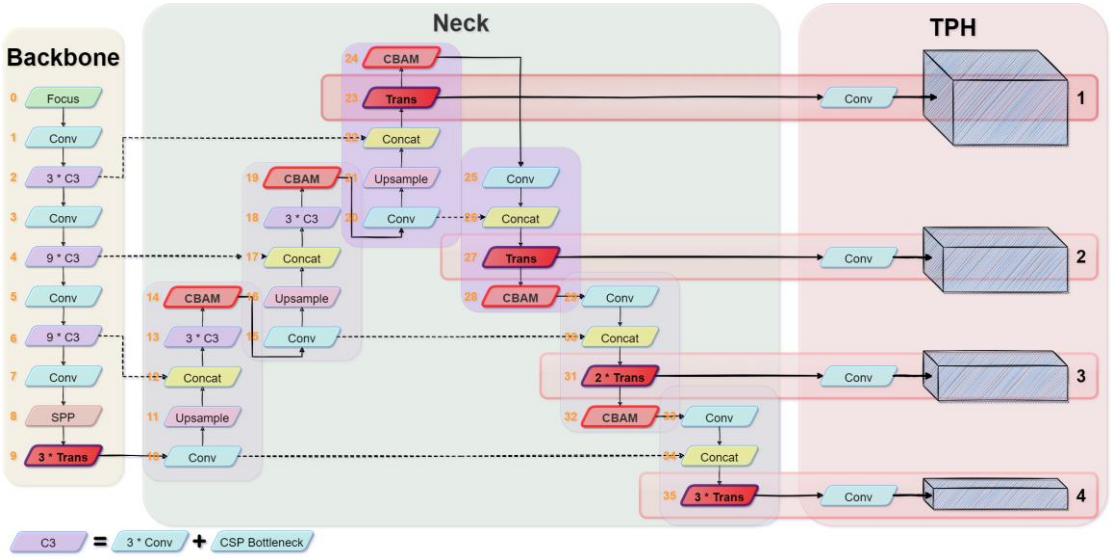


图 8 Yolov5 结构图

(1) 数据输入

Yolov5s 模型在输入端采用了先进的技术，有效地提升了图像数据的信息丰富度，并在目标检测框的调整方面具有更高的灵活性。本文将详细介绍 Yolov5 模型在输入数据上的预处理方法，主要包括 Mosaic 数据增强技巧和自适应锚框计算。

- Mosaic 数据增强

Mosaic 数据增强主要思想是将四张图片进行随机裁剪，再拼接到一张图上作为训练数据。Mosaic 数据增强有助于提高模型的鲁棒性和泛化能力，因为它引入了更多的场景变化和复杂性，避免因部分类型数据过少而导致过拟合，且提高了对小目标物体的检测性能。

在拼接前，程序会生成一个大的合成图像背景，并在背景上随机生成一个点，以该点在图中的横纵坐标作十字分割，对 4 张裁剪后图片所在区域进行划分。在图 10 中，4 张不同样本图的分配位置以不同颜色呈现。其中蓝色点是随机分割点；黑色框是 mosaic 画布；蓝色框是 4 张样本图实际所占空间。由于样本图的裁剪是随机的，难免发生某样

本图大小超过背景的情况，图 11 所示。此时需要对超出部分进行裁剪，将超出部分切除，保留在背景图像上的部分，左上、左下和右下的部分同理。



图 9 Mosaic 样例

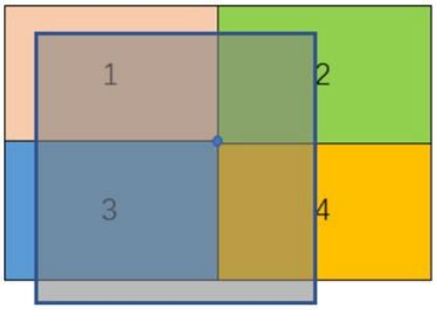


图 10 Mosaic 分割思路

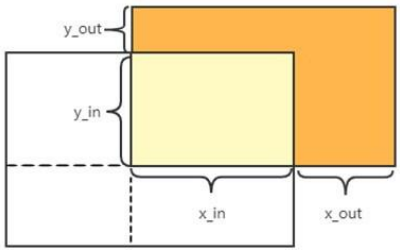


图 11 Mosaic 分割实例

- 自适应锚框计算

大部分传统的图像检测模型往往只能对固定尺寸图像进行单目标检测。而在数据集中，训练集样本的每一目标框的尺寸、数量、位置不一，难以具体确定锚框大小，目标框的尺寸没有规律，加上图片尺寸较大，不能通过简单的数值约定约束锚框大小。因此，自适应锚框计算的核心思想是根据数据集中的目标尺寸和分布情况来自动计算和调整锚框的大小和长宽比。

YOLOv5s 网络在每次训练迭代中，按照事先定义的输入长宽尺寸设置锚框，然后根据新数据集的标签信息，自动学习适应锚框的尺寸。如果召回率达到了 0.98 以上，就不再进行优化，而是直接返回结果；但如果召回率低于 0.98，那么模型将采用 K 均值聚类方法，结合自定义数据集的分析，以获得适合数据集目标边界框的预设锚定边框。具体来说，在可能的位置根据预设锚定框来选取目标区域，然后在这些预设锚定框的基础上进行微调。其次，利用真实标签位置相对于预设锚定框的偏移来构建预定义的锚框，模型在这些锚框的基础上输出目标检测的预测框。接下来，将这些预测框与真实标签进

行比对，计算它们之间的差异，并通过反向传播来更新模型参数，迭代使损失值最小化，以使预测框尽可能与初始的输入锚框相匹配。

这一过程实现了自适应计算不同训练数据集中的最佳锚框值。在目标检测任务中，首先生成多个锚框，然后预测每个锚框的类别信息和位置偏移量。根据这些预测的偏移量，调整锚框的位置，从而得到最终的预测边界框。最后，通过筛选，确定需要输出的预测边界框。

(2) Focus 结构分析

Focus 结构是 Yolov5s 的创新点，在历史 Yolo 系列的模型中未出现此种结构。其切片操作如图所示。具体操作是在一张图片中每跨步一个像素取出一个值，类似于邻近下采样，得到四张互补的图片，并未产生信息丢失。这样一来，将 W、H 信息就集中到了通道空间，输入通道扩充了 4 倍，即拼接起来的图片相对于原先的 RGB 三通道模式变成了 12 个通道，最后将得到的新图片进行卷积操作，最终得到了没有信息丢失情况下的二倍下采样特征图（Focus 下采样的参数量略多于 Conv 下采样，但是有一定提升效果）。

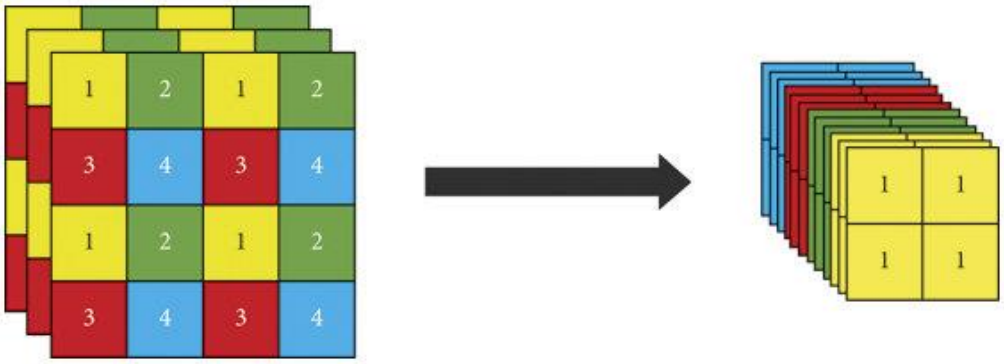


图 12 Focus 切片操作

在模型中，原始 608*608*3 的图像输入 Focus 结构，采用切片操作，先变成 304*304*12 的特征图，再经过一次 80 个卷积核的卷积操作，最终变成 352*352*80 的特征图。其流程如图 13 所示

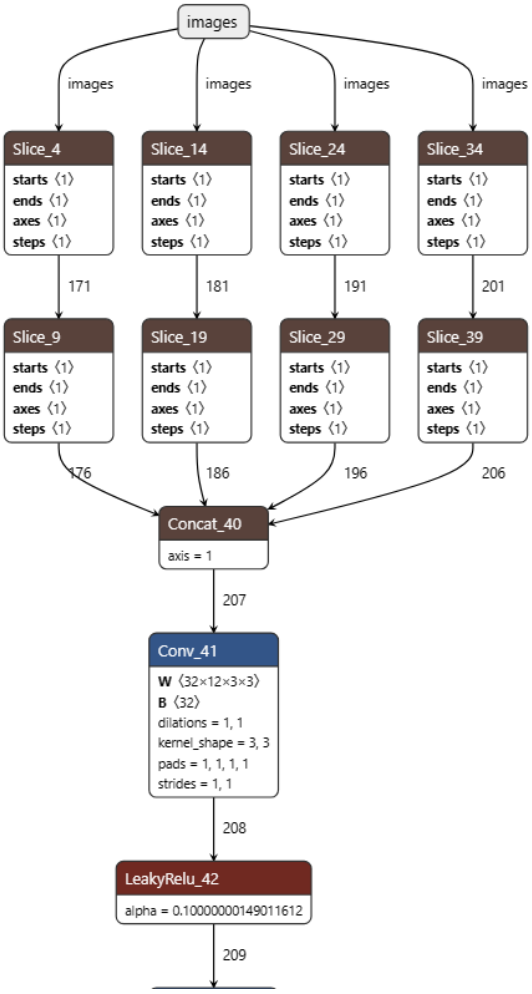


图 13 Focus 结构示意图

(3) Backbone 骨干网络

Backbone 骨干网络是 Yolov5s 中一个预先训练的网络，其主要负责提取图像数据的丰富特征表示，得到多个大小不同的特征图，这有助于降低图像的空间分辨率并提高其特征（通道）分辨率。增加残差结构，可以增加层与层之间反向传播的梯度值，避免因为加深而带来的梯度消失，从而可以提取到更细粒度的特征且不必担心网络退化。Backbone 的网络结构其大致流程如图 14 所示（输入图片尺寸为 608x608）：

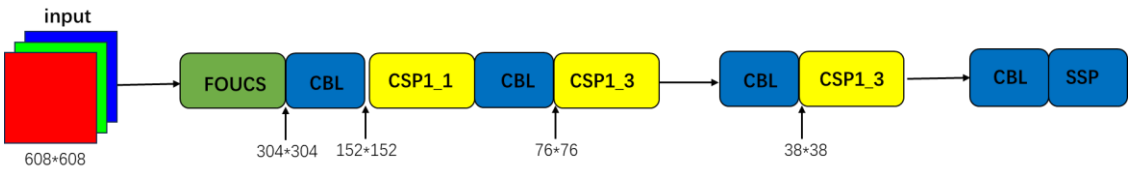


图 14 Backbone 骨干网络结构示意图

(4) Neck 网络

Yolov5s 的 Neck 最新和 Yolov4 中一样，都采用 FPN+PAN 的结构。Yolov4 在 FPN 层的后面还添加了一个自底向上的特征金字塔。FPN 的功能在于有效实现了从高维度向低维度的语义信息传递，加强了不同层次特征的融合，从而实现了多尺度的目标预测，

从而使模型在大目标的识别方面表现更为明确；而 PAN 则在低维度向高维度再次传递语义信息，从而增强了模型对小目标的识别准确性。这两种结构的联合应用使得不同特性互相补充，从而进一步提升了模型的性能。

由图 15 所示，FPN 是自顶向下的，将高层的特征信息通过上采样的方式进行传递融合，得到预测的特征图。每个 CSP 模块前面的卷积核大小都是 3*3，步长为 2，相当于下采样操作。因此得到的特征图的大小依次为 76*76、38*38、19*19。

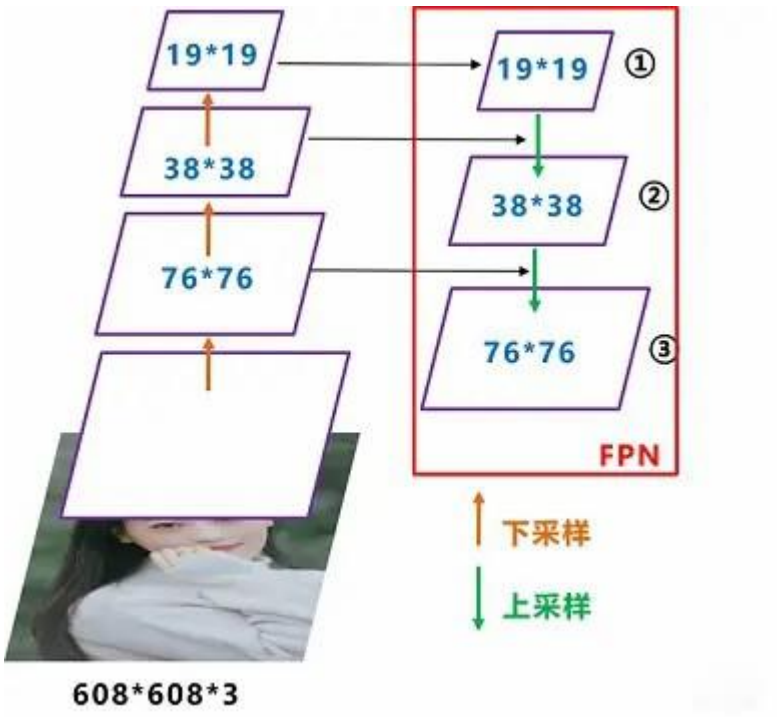


图 15 FPN+PAN 结构示意图

和 Yolov3 的 FPN 层不同，Yolov5s 在 FPN 层的后面还添加了一个自底向上的特征金字塔。其中包含两个 PAN 结构。这样结合操作，FPN 层自顶向下传达强语义特征，而特征金字塔则自底向上传达强定位特征，二者协同，从不同的主干层对不同的检测层进行参数聚合。

Yolov5 的 Neck 结构中，采用借鉴了 CSPnet 设计的 CSP2_X 结构。与 CSP1_X 相比，CSP2_X 将 Bottleneck 部分换成了 2x 次 CBL 操作。具体见图 16。

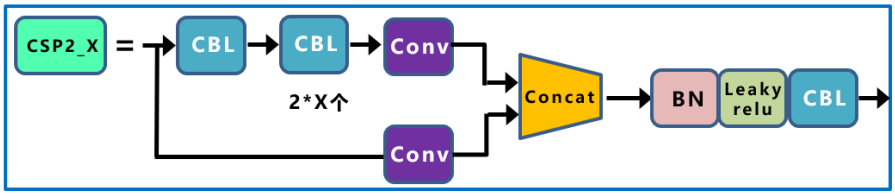


图 16 CSP2_X 结构示意图

4.1.6 基于 Yolov8 的深度学习模型

Yolov8 是一种基于深度学习的目标检测算法，它是 Yolo 系列算法的最新版本。Yolov8 引入了新的功能和改进，进一步提升性能和灵活性，使其成为实现目标检测、图像分割、姿态估计等任务的最佳选择。其结构图如图 17 所示[1]。

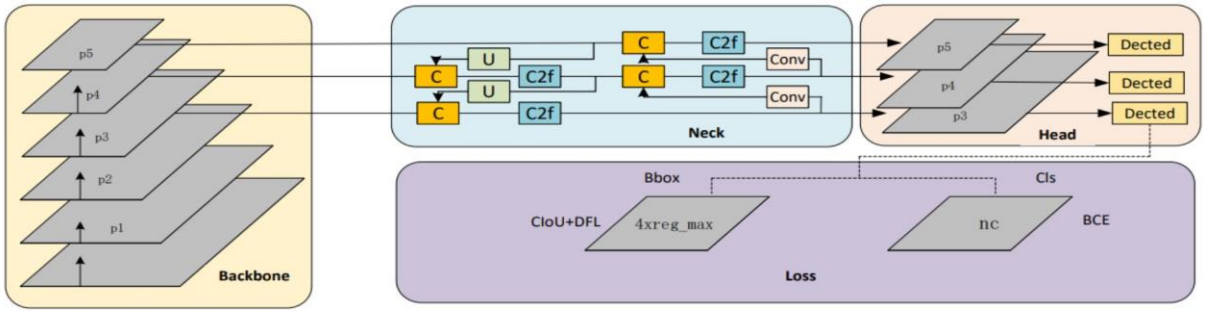


图 17 Yolov8 网络结构图

相较于 Yolov5 模型，Yolov8 模型的改进主要在于 Backbone、Neck、Head 和 Loss 等方面。

(1) Backbone

Yolov8 的 Backbone 部分与 Yolov5 基本相同，基于 CSP 思想，将 C3 模块替换为 C2f 模块。C2f 模块借鉴了 Yolov7 中的 ELAN 思路，将 C3 和 ELAN 结合在一起组成了 C2f 模块，使 Yolov8 在保证自身重量轻的同时可以获得更丰富的梯度流信息。

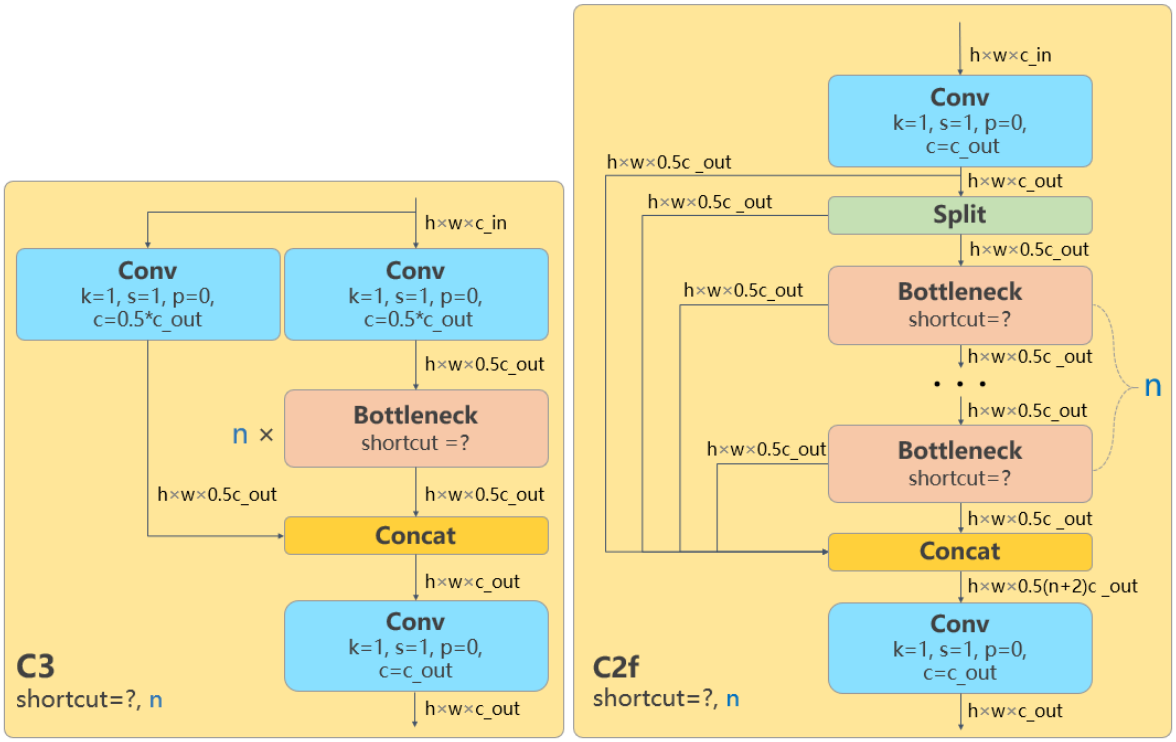


图 18 C3 与 C2f 对比图

在 Backbone 末端仍然使用最流行的 SPPF 模块，依次传递 3 个大小为 5*5 的 Maxpools，然后将每一层进行串联，这样既保证了不同尺度下物体的精度，同时有保证了物体的轻量化。

(2)Neck

在 Neck 部分，Yolov8 使用的特征融合方法仍然是 PAN-FPN，加强了不同尺度下特征层信息的融合和利用。Yolov8 使用了两个上采样和多个 C2f 模块以及最终解耦的头部结构来组成 Neck 模块。

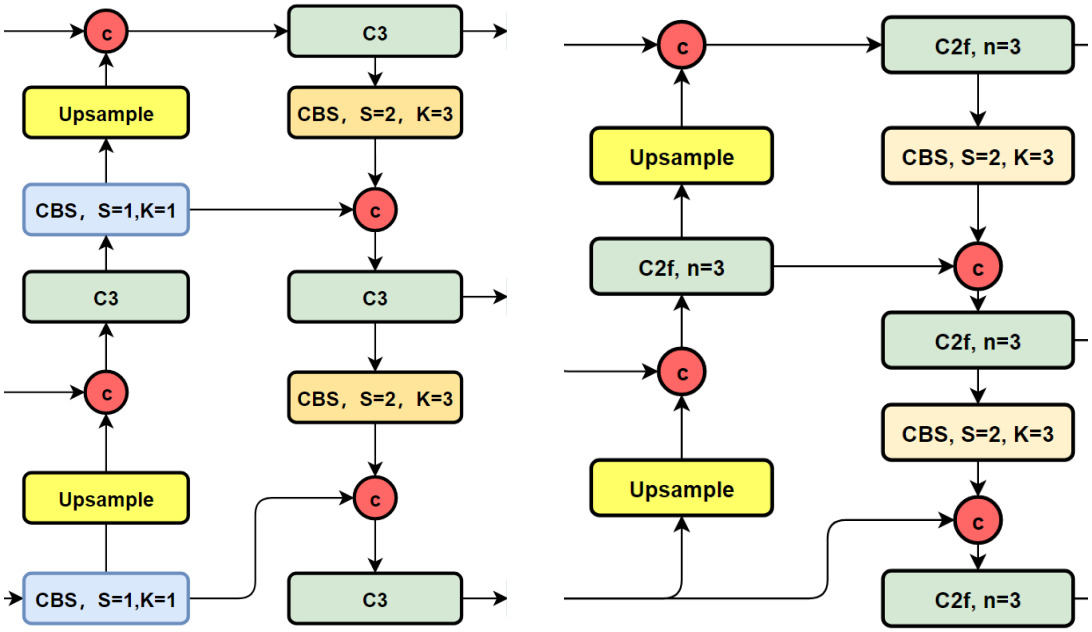


图 19 Yolov5(左)和 Yolov8(右)Neck 结构对比图

(3)Head

Yolov8 与其前身 Yolov5 相比，最重要的改进在于 Head 部分。具体而言，通过引入一种解耦头结构，类似于 Yolox，Yolov8 成功将回归分支和预测分支进行了有效分离。此外，针对回归分支的处理中，Yolov8 采用了 Distribution Focal Loss（DFL）策略中提出的积分形式表示法，将原本以确定性单值作为预测标准的回归坐标转化为一个分布，从而进一步提高了模型的精度和鲁棒性。

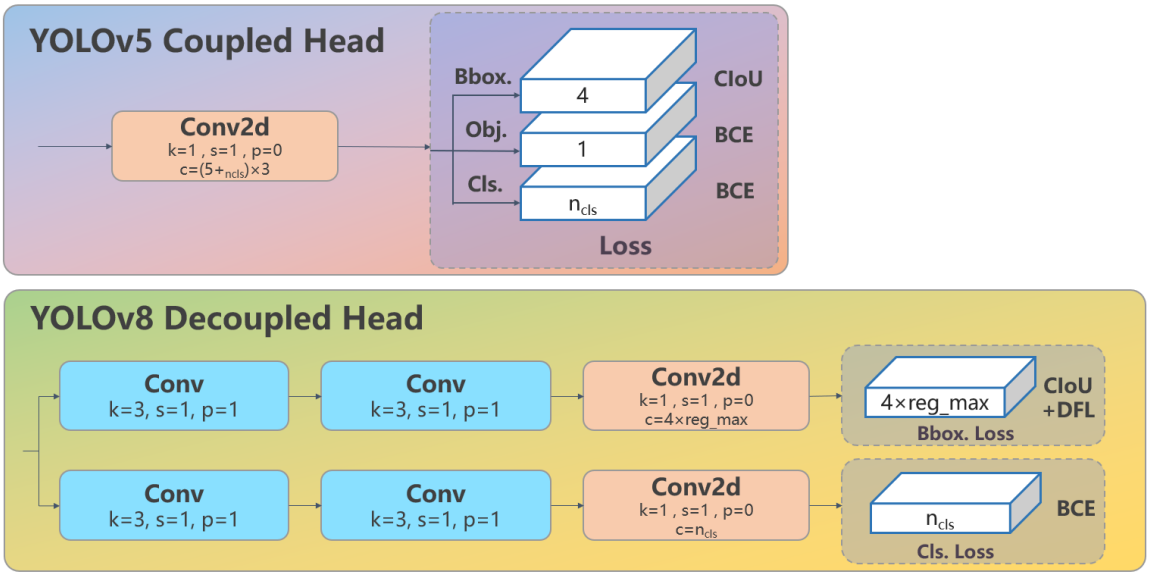


图 20 Yolov5 与 Yolov8 的 Head 结构对比图

4.2 问题二的模型建立与求解

4.2.1 模型训练

我们将经过翻转和裁剪操作扩充的数据集按 8:2 的比例划分训练集和测试集，并将问题一中构建的基于集成学习的机器学习、Yolov5 和 Yolov8 三种模型进行训练。

4.2.2 模型的指标评价

(1) 基于集成学习的机器学习模型

表 1 机器学习模型的评价指标表

准确率	精确率	召回率	F1 值
0.6179	0.6646	0.7004	0.6821

准确率（Accuracy）为 0.6179，表示模型在整体样本中预测正确的比例较低。

精确率（Precision）为 0.6646，表示在所有预测为正类的样本中，有 66.46%是真正的正类。较高的精确率意味着模型在预测为正类的样本中的误判较少。

召回率（Recall）为 0.7004，表示在所有真实正类样本中，有 70.04%被正确预测为正类。较高的召回率意味着模型能够较好地捕捉到正类样本。

F1 分数（F1 Score）为 0.6821，综合考虑了精确率和召回率。F1 分数是精确率和召回率的调和平均值，用于综合评估模型的性能。较高的 F1 分数表示模型在同时考虑精确率和召回率时取得了较好的平衡。

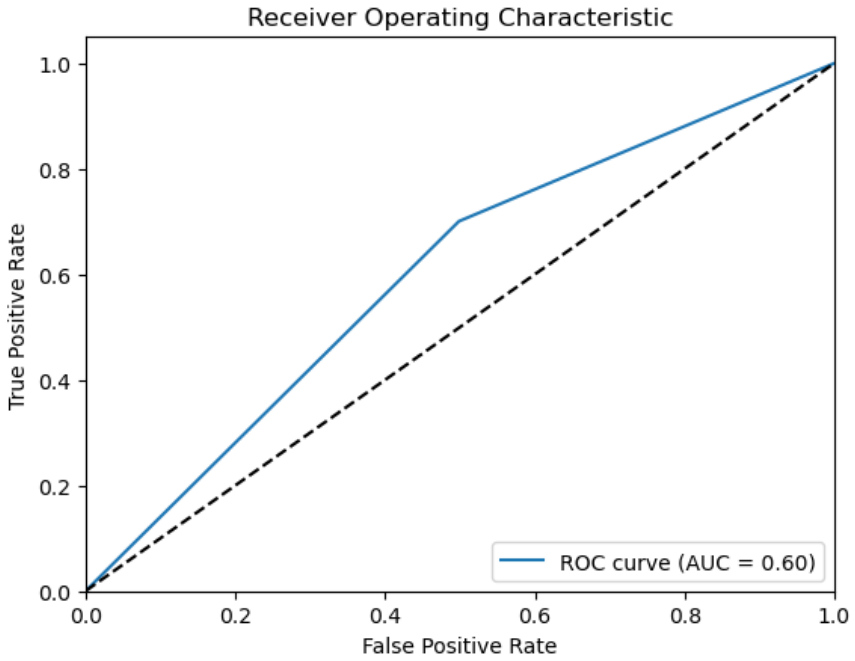


图 21 ROC 曲线图

该 ROC 曲线大部分时间都位于对角线之上，这表示模型的预测性能超过了随机预测。AUC (Area Under Curve) 值是 0.60。AUC 的范围是 0 到 1。一个完美的模型的 AUC 是 1，而一个完全随机的模型的 AUC 是 0.5。AUC 值为 0.60 表示该模型的性能比随机预测要好，但仍有提升的空间。

(2) 基于 Yolov5 的深度学习模型

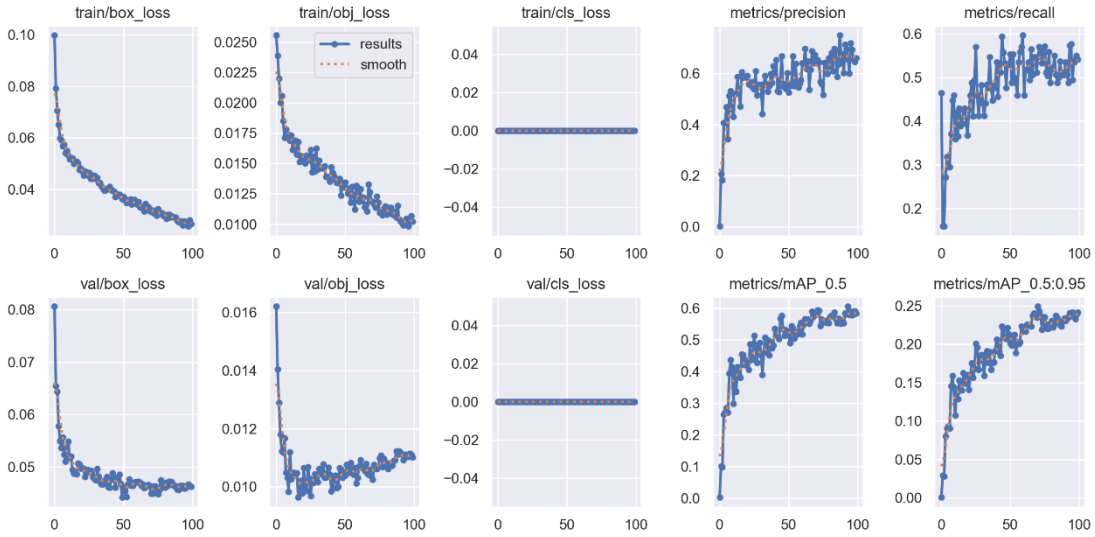


图 22 Yolov5 性能模型曲线

train/box_loss 和 val/box_loss：这两张图展示了模型的 bounding box 预测损失。从图中可以看出，训练损失随着训练周期数的增加呈下降趋势，验证损失也有相似的趋势。

train/obj_loss 和 val/obj_loss：这些图展示了模型预测物体存在的损失。同样，训练和验证的损失都在下降，但是在验证过程中下降得更为显著。

train/cls_loss 和 val/cls_loss：这是模型的分类损失。从图中可以看出，分类损失在训练和验证过程中基本保持不变。可能的原因是类别非常均衡，或者模型在分类任务上已经达到了很好的效果。

metrics/precision 和 metrics/recall：这两张图表示模型的精确度和召回率。精确度和召回率是评价模型性能的关键指标，特别是在物体检测任务中。从图中可以看出，随着训练的进行，精确度和召回率都有所波动，但总体上呈上升趋势。

metrics/mAP_0.5 和 metrics/mAP_0.5:0.95：这两张图展示了模型的 mean Average Precision (mAP)。mAP 是评价物体检测模型性能的一个重要指标。第一张图（mAP_0.5）只考虑 IoU>0.5 的预测，而第二张图（mAP_0.5:0.95）则考虑了从 0.5 到 0.95 的所有 IoU 值。后者提供了一个更全面的评估。两者都显示了随着训练周期的增加，mAP 也在逐渐上升，这意味着模型的预测效果在改善。

这组图表显示，随着训练周期数的增加，Yolov5 模型的性能在不断提高。所有的损失都在下降，而精确度、召回率和 mAP 等性能指标都在上升。

(3) 基于 Yolov8 的深度学习模型

Yolov8 模型的 F1-Confidence Curve、Precision-Confidence Curve、Precision-Recall Curve 和 Recall-Confidence Curve 性能曲线图如下：

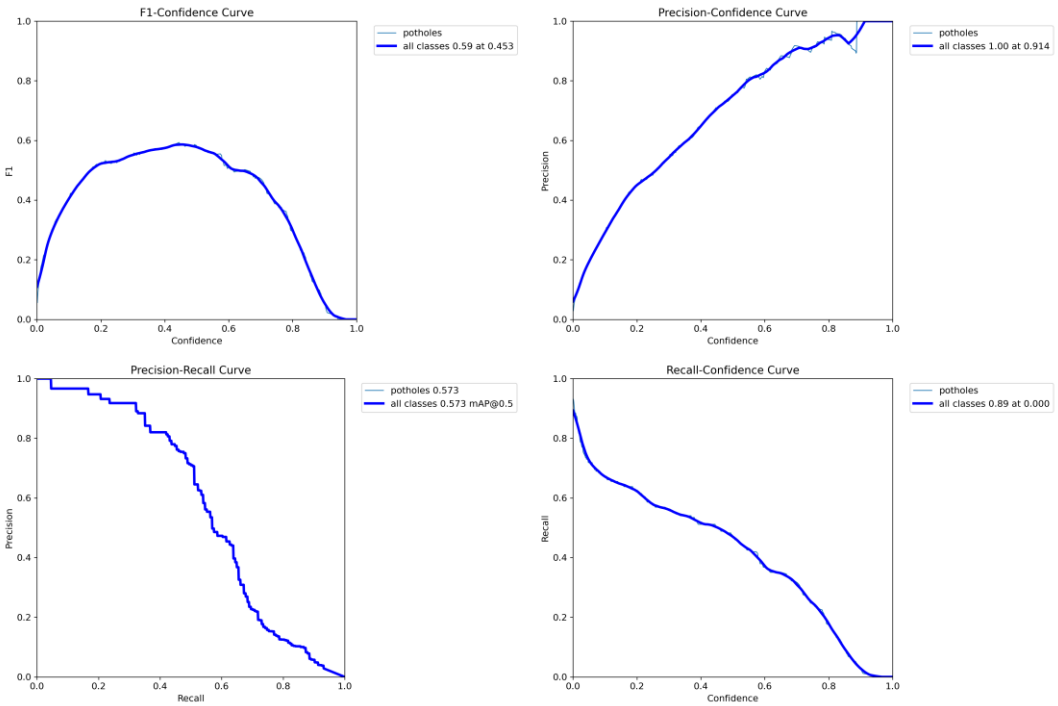


图 23 Yolov8 模型性能曲线 1

F1-Confidence Curve：对于“potholes”类别，当置信度在约 0.453 时，F1 分数达到了其最大值，约为 0.59。这意味着模型对“potholes”分类的综合性能（精确度和召回率的结合）在这个置信度下最好。F1 曲线的形状呈现了一个明显的山峰形状，随着置信度的增加，F1 分数先增加然后下降。

Precision-Confidence Curve：对于“potholes”类别，随着置信度的增加，精确度呈现出一个上升的趋势，并在约 0.914 的置信度处达到 1.0，这意味着在这个置信度水平上，模型的精确度达到了最大值。

Precision-Recall Curve：这条曲线展示了精确度和召回率之间的权衡关系。对于“potholes”类别，其最大的平均精确度（mAP）为 0.573，这是一个中等的性能指标。曲线从左上角向右下角移动，表示在提高召回率的过程中，精确度有所下降。

Recall-Confidence Curve：当置信度为 0 时，对于“potholes”类别的召回率为 0.89。随着置信度的增加，召回率逐渐下降。这意味着在高置信度下，模型可能错过了一些真正的“potholes”实例。

Yolov8 模型在“potholes”分类上的表现呈现出一定的权衡。虽然在某些置信度水平上，它的精确度可以达到很高，但其召回率可能较低。

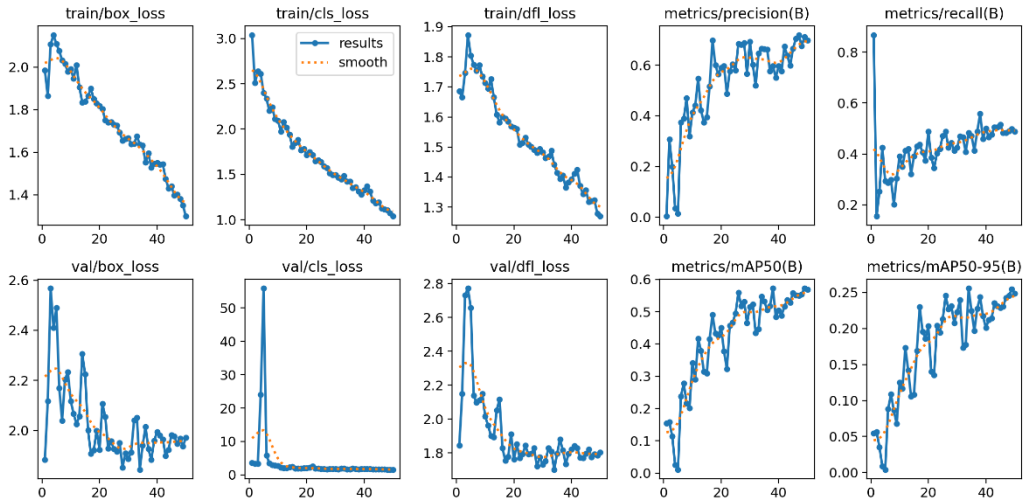


图 24 Yolov8 模型性能曲线 2

train/box_loss：训练过程中的框损失 (box loss) 显示了一个下降的趋势，这表示模型在学习过程中在预测边界框位置上逐渐变得更准确。

train/cls_loss：分类损失 (class loss) 也呈现下降趋势，这意味着模型在区分不同的类别上变得更好。

train/dfl_loss：这个损失也呈现了下降趋势。

验证损失 (Validation Loss)：val/box_loss, val/cls_loss,和 val/dfl_loss：验证集上的损失在初始阶段有较大的波动，但随着时间的推移，这些波动逐渐减小，而且呈下降趋势。这意味着模型在验证集上的性能逐渐稳定并且提高。

metrics/precision(B)和 metrics/recall(B)：在训练过程中，精确度 (precision) 和召回率 (recall) 呈现了波动的上升趋势。这表示模型在正确识别目标的能力逐渐增强。

metrics/mAP50(B)：这个指标代表在 IoU 阈值为 0.5 时的平均精确度 (mean average precision)。它也显示了一个上升的趋势，说明模型的整体性能在提高。

metrics/mAP50-95(B)：这是在多个 IoU 阈值（从 0.5 到 0.95）上的平均精确度。这个指标也在逐渐上升，表示模型在多个阈值上的性能都在提高。

Yolov8 模型在训练过程中逐渐获得了更好的性能。损失值在下降，而性能指标在上升。但是，还需要注意验证损失和性能指标的波动，这可能表示模型对于训练数据可能存在一定的过拟合。

4.2.3 模型的鲁棒性

鲁棒性是指系统或模型在面对各种不确定性、异常情况或攻击时的稳定性和健壮性。在深度学习和人工智能领域，鲁棒性是指模型对于输入数据的变化、扰动、噪声以及对抗性攻击等的处理能力。验证模型的鲁棒性是一个重要的任务，本文通过添加高斯噪声和生成对抗样本两种方法来验证模型的鲁棒性。

4.2.4 添加高斯噪声

高斯噪声（Gaussian noise）是一种常见的随机噪声模型，也被称为正态分布噪声。它是指服从高斯分布的随机变量所引入的噪声。



图 25 高斯噪声

我们使用训练后的模型分别对提供的 301 个原始图像文件进行分类预测并计算其准确率，接着对该数据集添加高斯噪声重新进行分类预测并计算准确率，对比准确率的变化而评价模型的鲁棒性。

- 基于 Yolov5 的深度学习模型

原始图像分类准确率：97.010%，添加噪声图像分类准确率：97.010%。

表 2 Yolov5 模型有无噪声识别对比表（识别结果不同部分）

图像文件名称	无噪声	有噪声
Normal9.jpg	1	0
normal227.jpg	0	1
normal268.jpg	0	1
normal309.jpg	1	0

Yolov5 模型在面对噪声时表现出了一定的鲁棒性。尽管无噪声和有噪声的识别准确率相同，但仍 4 个样本在两种情况下的识别结果不一致。这意味着模型在面对噪声时可能会受到一些干扰，导致分类结果与无噪声情况下有所不同。

- 基于 Yolov8 的深度学习模型

原始图像分类准确率：99.003%，添加噪声图像分类准确率：96.678%。

表 3 Yolov8 模型有无噪声识别对比表（识别结果不同部分）

图像文件名称	无噪声	有噪声
normal106.jpg	1	0
normal14.jpg	1	0
normal151.jpg	1	0
normal177.jpg	1	0
normal190.jpg	1	0
normal208.jpg	1	0
normal297.jpg	1	0
potholes00000.jpg	1	0
potholes00024.jpg	0	1

通过比较原始图像分类准确率和添加噪声图像分类准确率，我们可以评估模型的鲁棒性。添加噪声导致图像分类准确率下降了约 2.325%。在这种情况下，96.678% 的分类准确率表明 Yolov8 模型在存在噪声的图像上仍然能够较好地进行分类。虽然添加噪声导致准确率下降，但仍然保持了相对较高的分类性能。

4.2.5 生成对抗样本

生成对抗样本是深度学习邻域的热门研究话题。对抗样本是通过在原始输入数据上添加微小的扰动来生成的，这些扰动足以欺骗模型，使其产生不正确的预测，但对人类

来说这种变化是几乎不可察觉的。深度神经网路在许多任务上表现出色，但对抗样本揭示了它们的脆弱性，对抗攻击可能会影响现实世界的应用，例如自动驾驶、面部识别。其中坑洼道路就被广泛应用于自动驾驶领域，考虑到现实世界中安全性的要求，此处我们对模型进行了对抗样本攻击，以评估模型的安全性。

现有的对抗样本大多是通过在正常图像上叠加扰动生成的，在保证人类无法察觉的同时，使模型尽可能识别错误。在生成对抗样本时实际上是解决一个优化问题，因此我们需要对一个目标函数最小化以达到目标效果。通用的优化目标函数$f(\cdot)$定义公式如下：

$$f(x,\delta,t)=l_{model}(F(x+\delta),t)+c\cdot l_{metric}(x+\delta) \tag{6}$$

其中$F(x+\delta)$是图像识别模型；c是权衡攻击成功和最终图像的权重系数；l_{model}是与识别模型相关的损失函数，用来衡量期望的模型输出t和实际输出$F(x+\delta)$之间的差异；$l_{mrtric}(x+\delta)$是模型对样本$x+\delta$的置信度。

基于遗传算法的生成对抗样本的优点是它不需要直接访问模型的参数或内部结构。这使其特别适合于黑盒攻击情境。此外，遗传算法提供了一种灵活的方法来调整和探索不同的对抗扰动，这有助于找到有效的对抗样本。算法流程图如下：

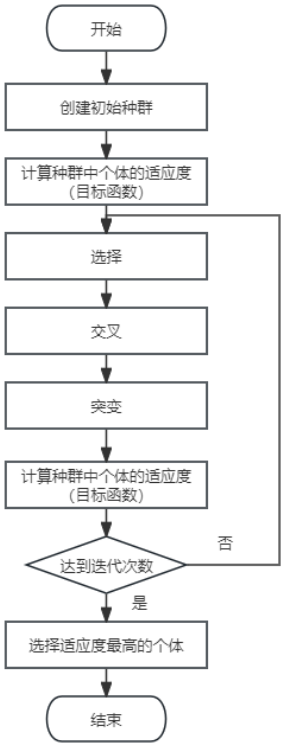


图 26 遗传算法流程图

我们对 Yolov5、Yolov8 这两个深度学习模型分别进行了生成对抗样本攻击，将初始种群数量设置为 100、迭代次数为设置 500、扰动系数设置为 1，在原始 pothels 类中 35 张图片生成了对抗样本。结果如下：

表 4 生成对抗样本结果

文件名	v5	v5 对抗样本	v8	v8 对抗样本
potholes00000.jpg	1	1	1	0
potholes00001.jpg	1	1	1	1
potholes00002.jpg	1	1	1	1
potholes00003.jpg	1	1	4	1
potholes00004.jpg	0	0	2	2
potholes00007.jpg	0	0	0	0
potholes00008.jpg	1	1	1	1
potholes00009.jpg	2	2	2	2
potholes00010.jpg	2	2	1	1
potholes00012.jpg	1	1	1	1
potholes00013.jpg	1	1	1	1
potholes00018.jpg	1	1	1	1
potholes00019.jpg	1	1	1	1
potholes00020.jpg	1	1	1	1
potholes00024.jpg	2	2	1	1
potholes00028.jpg	2	2	2	2
potholes1.jpg	2	2	2	2
potholes12.jpg	4	4	2	2
potholes16.jpg	1	1	1	1
potholes19.jpg	6	6	2	2
potholes2.jpg	8	8	2	2
potholes22.jpg	3	3	4	4
potholes24.jpg	9	9	10	10
potholes26.jpg	1	1	2	2
potholes29.jpg	5	5	8	8
potholes3.jpg	27	7	18	18
potholes31.jpg	3	3	2	2
potholes32.jpg	7	7	9	9
potholes34.jpg	1	1	1	1
potholes4.jpg	5	5	3	3
potholes5.jpg	1	1	1	1
potholes6.jpg	2	2	2	2
potholes7.jpg	1	1	1	1
potholes8.jpg	9	9	7	7
potholes9.jpg	1	1	1	1

表中的数字表示模型识别出的坑数。从结果上看，Yolov5 模型在受到对抗样本攻击后有 1 个文件识别结果出现了差异，为文件“potholes3.jpg”；Yolov8 模型在受到对抗样本攻击后有 2 个文件识别后出现了差异，为文件“potholes00000.jpg”、“potholes00003.jpg”。

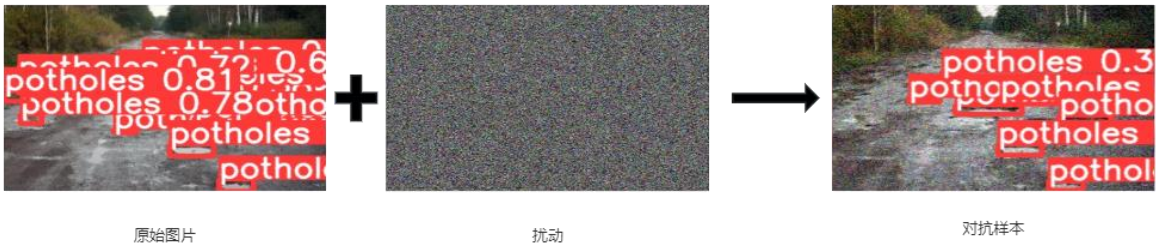


图 27 Yolov5 生成对抗样本结果

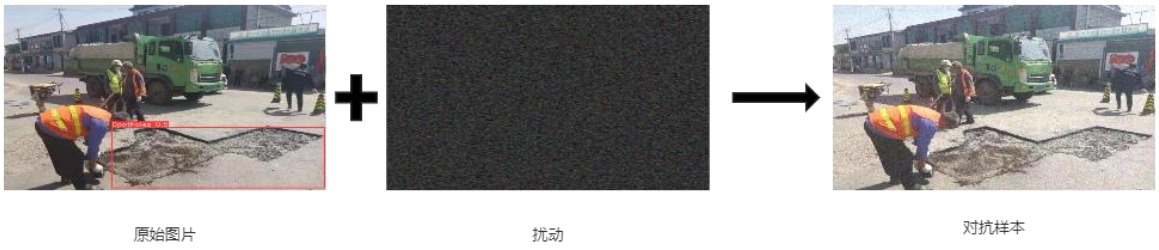


图 28 Yolov8 生成对抗样本结果

文件“potholes3.jpg”在生成对抗样本后，Yolov5 的识别结果从 27 个坑洼变为了 7 个坑洼；文件“potholes00000.jpg”在生成对抗样本后，Yolov8 的识别结果从 1 个坑洼变为了 0 个坑洼。

不考虑对抗攻击强度因素，从总体结果来看，生成的对抗样本对原始识别结果影响较小，该模型具有一定的抗对抗样本的能力，鲁棒性较强。

4.3 问题三的模型建立与求解

本题我们利用上述构建的机器学习、Yolov5 和 Yolov8 三个模型对测试数据集进行分类，其中机器学习模型权重为 0.2，Yolov5 和 Yolov8 模型的权重均为 0.4。若这三个模型的加权投票大于或等于 0.5，则分为正常道路，反之则为坑洼道路。在该模型下，若 Yolov5 模型和 Yolov8 模型分类结果一致，则选用该分类结果，否则选用机器学习的分类结果。部分分类结果如表 5，全部分类结果见“test_result.csv”文件。

表 5 模型分类结果

图像名称	机器学习	Yolov5	Yolov8	加权投票	分类
00aqq56m.jpg	1	1	1	1	1
00c9sdfq.jpg	1	0	0	0.2	0
00jjy9i3.jpg	1	1	1	1	1
00q3q8py.jpg	1	0	1	0.6	1
00yf8rxj.jpg	1	0	0	0.2	0
01jyy8yk.jpg	1	1	1	1	1
01lx7dej.jpg	0	0	1	0.4	0
01paoyy0.jpg	0	1	0	0.4	0
02f4ee5p.jpg	1	1	1	1	1
02ld1b80.jpg	1	1	1	1	1
		……			
zx9exkdw.jpg	1	0	0	0.2	0

zxfhwbki.jpg	1	0	0	0.2	0
zxgm3ar3.jpg	1	1	1	1	1
zy7m9c8x.jpg	0	1	0	0.4	0
zynixpux.jpg	1	0	0	0.2	0
zyylqfs0.jpg	0	0	0	0	0
zzhe6767.jpg	0	0	0	0	0
zzipq9zn.jpg	0	1	1	0.8	1
zzvom0ax.jpg	0	0	0	0	0
zzvth22q.jpg	1	1	1	1	1

五、 模型评价与推广

5.1 模型优点

1. 本文主要实现了基于 YOLO 目标检测深度学习模型的道路坑洼检测项目，并结合机器学习进行集成学习，所成的融合模型识别率高、分类准确，能够快速准确地识别图像中的指定物体。

2. 在前期数据预处理中，由于所给训练数据集中正常和坑洼道路类别不平衡。我们通过翻转、裁剪操作对 potholes 类数据集进行扩增，显著提高模型预测的准确性。

3. 我们通过对图片增加噪声和生成对抗样本来评估模型的鲁棒性。通过数据对比可得其在处理复杂背景、部分遮挡和小目标等情况下表现出较强的鲁棒性和抗干扰能力，能够适应不同的场景。

5.2 模型缺点

1. YOLOv5 的性能高度依赖于大规模、多样化的训练数据。如果训练数据不足，模型的性能可能会下降。

2. YOLOv5 深度学习模型在处理大规模数据时需要较强的计算资源（GPU）支持，这限制了模型训练的最终性能，使得模型特征提取能力和识别准确度受到进一步限制。

3. 由最终得取的测试结果说明模型的泛化预测能力还需进一步提升。

5.3 模型推广

深度学习在大数据中具有较强的特征提取能力，极大地提高了对于复杂数据集的处理能力，以较快的速度和准确率平衡而广受推崇。由于 YOLO 模型泛化能力、特征提取能力较强，其对于大多数目标检测准确率都较高，这使其应用广泛，其可应用于各种物体检测项目，比如道路人流检测，人脸识别技术等领域。

六、 参考文献

[1]倪云峰,霍洁,侯颖等.基于 YOLOv8—OCR 的井下人员检测算法[J/OL].无线电工程:1-8[2023-11-02].

[2]闫佳,聂楚江,苏璞睿.基于遗传算法的恶意代码对抗样本生成方法[J].电子与信息学报,2020,42(9):2126-2133.

[3] 孟东宇. 黑盒威胁模型下深度学习对抗样本的生成[J]. 电子设计工程, 2018, 26(24): 164-167.

[4]任安虎,姜子渊,马晨浩.基于改进 YOLOv5s 的道路裂缝检测算法[J/OL].激光杂志:1-7[2023-11-03].

[5]刘彦清. 基于 YOLO 系列的目标检测改进算法[D].吉林大学,2022

[6]谈世磊,别雄波,卢功林等.基于 YOLOv5 网络模型的人员口罩佩戴实时检测[J].激光杂志,2021,42(02)

[7]Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (pp. 2778-2788).

队伍编号	MCB2302397
赛道	B

基于特征融合与集成学习的电商供应链分层预测方法

摘 要

需求预测往往是电商平台智能供应链的决策基础，它可以让管理者提前预知各地的需求，从而将库存提前放在靠近需求的仓库中，显著地降低库存成本，同时保证商品的按时履约。本文对各商品的历史出货量数据进行挖掘与分析，构建了基于数理特征与专家经验的多模型集成预测框架，旨在协助电商平台优化库存管理，更好地规划和定制促销活动，提高库存周转率。

针对问题一，需要对商品序列进行分类并预测未来需求量。本文首先根据附件数据构建全体需求量序列集。考虑部分字段信息覆盖了功能场景、仓库区域等组成的先验分类知识，先基于 8 个字段特征进行第一层级分类。然后，采用 K-means 算法对第一层级类别进行细化聚类，建立 9 个不同尺度的分类标准。针对预测问题，本文提出了一种基于自顶向下和分层聚类的集成预测方案。首先，基于 9 个分类标准获得各类别下总的需求量序列。然后，基于特征工程构建各类别汇总序列的特征集，分别采用指数平滑、Auto ARIMA、XGBoost、LightGBM、LSTM 等 5 个算法对各分类标准各类别下的总需求量序列构建预测模型，并设计合适的指标评估各算法的优劣。最终，采用 9 个分类标准下的模型预测值的集成量作为最后的预测输出，结果表明融合后的模型获得更高的 1-WMAPE 指标且预测更为稳定。

针对问题二，需要预测新出现序列组合的未来需求量。首先，将附件 1 与附件 5 的数据整合（2206 条），对所有组合截取等维长度的序列并提取数理特征，并将构建的数理特征与独热编码后的 8 个解释变量特征汇整，采用 K-means 算法对所有实例进行聚类（2206 条序列聚为 14 类）。然后，设置滑动窗口对聚类后的各类别序列进行特征提取和样本集构建，采用 SVM、XGBoost、LightGBM、LSTM、KNN 算法分别训练预测模型，基于 5 折交叉验证法调整模型参数并评估算法性能。在预测阶段，首先对待预测实例进行分类，采用类别下的 5 个模型预测值的集成量作为最后的预测输出，结果表明集成后预测模型可以挖掘出更为显著的相似序列特征，实现对新组合的精准预测。

针对问题三，需要预测规律性大型促销活动期间各商品的未来需求量。首先，提取附件 1 中各商品每月 1 日-11 日的出货量数据（共 5 个月）组成日常序列，并将各商品以月为基准取均值得到日常均值序列。提取附件 6 中各商品 1 日-11 日的出货量数据组成促销序列。然后，根据日常均值数据与促销数据计算各商品的促销力度值，基于问题一提出的集成预测框架对各商品在 6 月 1 日-6 月 20 日的日常值进行预测。最后，将预测结果与对应的促销力度值求积，得到最终的促销预测值，结果表明构建的事件影响模型可以分析出促销活动期间商品序列的波动规律，使预测值更符合促销期间的实际情况。

关键词：特征融合；集成学习；分层时序预测；LightGBM；K-means；LSTM

目 录

1 问题重述 2
1.1 问题背景 2
1.2 问题提出 2
2 问题分析 2
2.1 问题一的分析 2
2.2 问题二的分析 4
2.3 问题三的分析 5
3 模型假设 6
4 符号说明 6
5 问题一模型建立与求解 6
5.1 基于数理特征与专家经验的分层时序聚类算法 7
5.1.1 聚类特征提取 7
5.1.2 K-means 聚类算法构建 9
5.1.3 聚类结果评估 9
5.2 基于自顶向下和分层聚类的集成预测算法 10
5.2.1 指数平滑预测模型构建 11
5.2.2 Auto ARIMA 预测模型构建 12
5.2.3 XGBoost 预测模型构建 13
5.2.4 LightGBM 预测模型构建 13
5.2.5 LSTM 预测模型构建 14
5.3 任务预测与性能评价 14
6 问题二模型建立与求解 19
6.1 基于融合特征的单层聚类算法 19
6.1.1 特征提取 19
6.1.2 K-means 聚类算法构建 20
6.1.3 聚类结果评估 21
6.2 多模型集成预测算法 22
6.2.1 预测特征提取 23
6.2.2 预测算法构建 23
6.3 任务预测与性能评价 24
7 问题三模型建立与求解 25
7.1 问题分析与模型建立 25
7.2 预测结果分析 27
8 模型评价与推广 27
8.1 模型的优点 27
8.2 模型的缺点 28
8.3 模型的推广 28
参考文献 28
附录 29

1 问题重述

1.1 问题背景

电商平台存在着上千个商家，他们会将商品货物放在电商配套的仓库，电商平台会对这些货物进行统一管理。通过科学的管理手段和智能决策，大数据智能驱动的供应链可以显著降低库存成本，同时保证商品的按时履约。需求预测往往是智能供应链的决策基础，它可以让管理者提前预知各地的需求，从而将库存提前放在靠近需求的仓库中，此时的预测任务为：根据历史一段时间的需求量，预测各仓库中各商品未来需求，“预测维度”即为不同商家在各仓库中存放的各种商品每天的数量。一般来说，企业会首先根据数据的历史情况，分析出需求量序列的数理特征，对相似的需求量序列进行归类，并根据分类结果做到更加精准的预测。

然而，在实际的电商供应链预测任务中，常常会出现多种不同的问题。例如，部分商品的销售时间过短、仓库存在新增或切换等情况，导致该预测维度下历史数据过少；另外，部分大型促销期间货量的陡增并由此带来的不规律性，也给需求量的精准预测带来了不小的难度。此时便需要通过算法得到历史一般规律，找出相似的历史情况（如相似的仓或商品），从而实现精准预测。

1.2 问题提出

基于前述研究背景，该题目给出了 6 个信息附件和 3 个答案附件。基于 6 个信息附件内容，拟要解决的研究问题如下：

问题 1：使用附件 1-4 中的数据，预测出各商家在各仓库的商品 2023-05-16 至 2023-05-30 的需求量，并将预测结果填写在结果表 1 并上传至竞赛平台，并对模型的预测性能进行评价。另外讨论：根据数据分析及建模过程，这些由商家、仓库、商品形成的时间序列如何分类，使同一类别在需求上的特征最为相似。

问题 2：当出现新的商家+仓库+商品维度（附件 5），讨论这些新出现的预测维度如何通过历史附件 1 中的数据进行参考，找到相似序列并完成这些维度在 2023-05-16 至 2023-05-30 的预测值。把预测结果填写在结果表 2，并上传至竞赛平台。

问题 3：每年 6 月会出现规律性的大型促销，为需求量的精准预测以及履约带来了很大的挑战。附件 6 给出了附件 1 对应的商家+仓库+商品维度在去年双十一期间的需求量数据，参考这些数据，给出 2023-06-01 至 2023-06-20 的预测值。把预测结果填写在结果表 3，并上传至竞赛平台。

2 问题分析

2.1 问题一的分析

问题一主要是基于附件 1-4 中的数据，对各商家在各仓库的商品未来 15 天的需求量进行预测，并分析由商家、仓库和商品形成的需求量序列的数理特征，对相似的需求量序列进行归类，根据分类结果提高需求量预测的精度。因此，问题一主要涉及聚类和

预测两个任务。针对上述任务，对附件 1-4 进行分析，梳理的经验特征有 8 个，包括字段名：商品一级分类、二级分类等。这些字段信息既包含了各需求量序列本身的特征属性，也覆盖了一些由行业标准、功能场景、仓库区域等组成的先验分类知识。如字段“qty_seq”为附件中所有商家-仓库-商品的组合，由这些组合构成的数据集无显式的内在关联，只能依赖聚类算法抽象处理；而字段“category1”-“warehouse_region”可以利用先验知识进行数据的初步划分，再借助聚类算法进行二次规整，使得划分的合理性和解释性更强。比如，往往规模大的商家出货量较多，对应的需求量序列变化趋势相似。因此，在进行序列分类时，可以先依据该字段名和特征类别对所有实例进行初步划分，而后采用聚类算法对相应类别进行细化。

具体地，针对分类问题，本文提出了一种基于数理特征与专家经验的分层时序聚类算法，从需求量序列的时域分布、稳定性、所属的商家、商品、仓库类别划分等尺度提取特征构建分类标准，算法架构如图 2.1 所示：

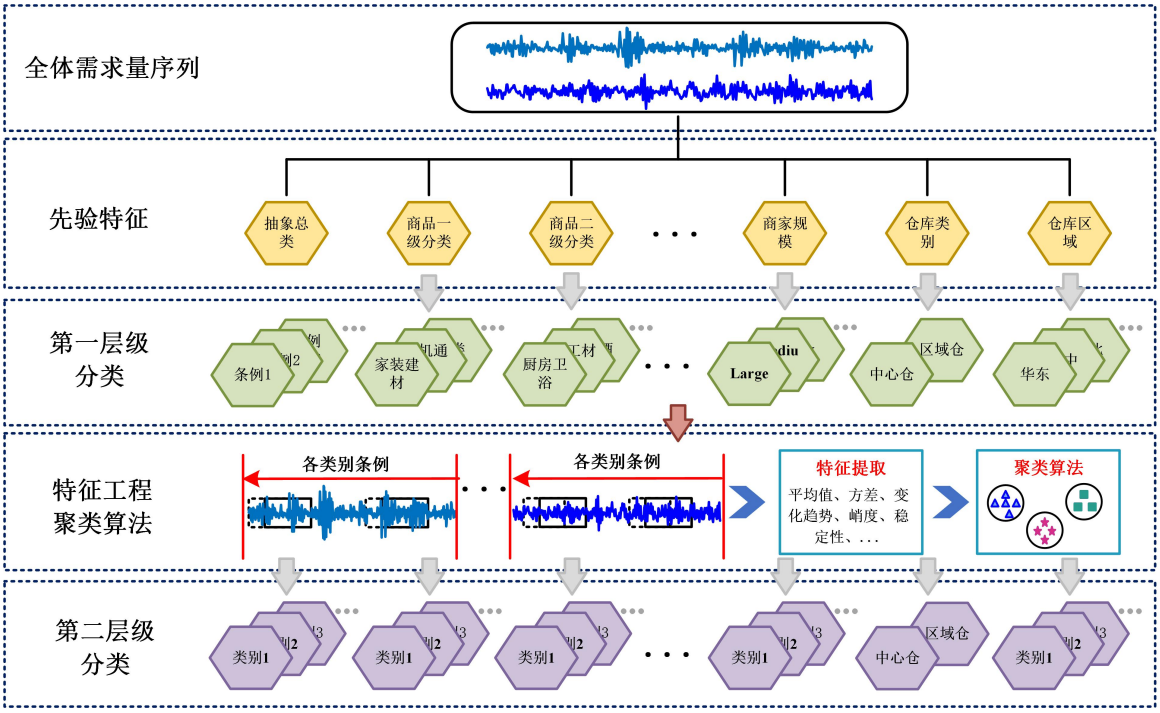


图 2.1 基于数理特征与专家经验的分层时序聚类算法

首先，基于附件 1-4 的数据构建全体需求量序列集。然后，根据表 2.1 中的经验特征“category1”-“warehouse_region”分别进行第一层级分类，获得 9 个分类标准，各分类标准下的类别数不等，如“category2”标准下有 79 类，“category3”标准下有 291 类，各类别所包含的实例数也不同。为了缩小各分类标准所含类别的个数，对第一层级分类结果中各类别数据进行数理特征提取[1-2]。最后，采用 Kmeans 聚类法[3]对每个分类标准下的类别数进行聚类，其目的在于将时序变化趋势差异较大的类别分离开，进而得到各分类标准下的最终类别数。比如，“category2”分类标准在第一层级下将数据按照市场经验划分为了 79 类，通过聚类算法在第二层级下被抽象分为 4 类。聚类后各标准对应的类别数如表 2.1 所示。

表 2.1 聚类后的类别数

分类标准	第一层级类别数	第二层级类别数
qty_seq	1996	5
category1	23	3
category2	79	4
category3	291	4
seller_category	17	4
inventory_category	4	4
seller_level	4	4
warehouse_categoty	2	2
warehouse_region	7	3

针对预测问题，本文提出了一种基于自顶向下和分层聚类的集成预测算法[4-5]。首先，对 9 个分类标准各类别下包含的实例进行求和处理，获得各类别下总的需求量序列，共计 33 条。然后，对各类别下总的需求量序列进行特征提取，获得表征各类别需求量序列变化趋势的显著特征。最后，分别采用极致梯度提升（XGBoost）[6]、轻量级梯度提升机器学习（LightGBM）[7]、长短时记忆网络（LSTM）[8]、随机森林[9]等预测算法对各类别下的总需求量序列构建时序预测模型（各算法下均构建 33 个模型），并设计合适的指标评估各算法的优劣，进而选择最优的预测算法。在预测阶段，针对待预测的需求量序列，先按照 9 个分类标准将其分别划分到第二层级的各类别中，并基于各类别构建的模型采用自顶向下的方案得到 9 个预测值。最终，采用 9 个模型预测值的集成量作为最后的预测输出，具体的实现流程如图 2.2 所示：

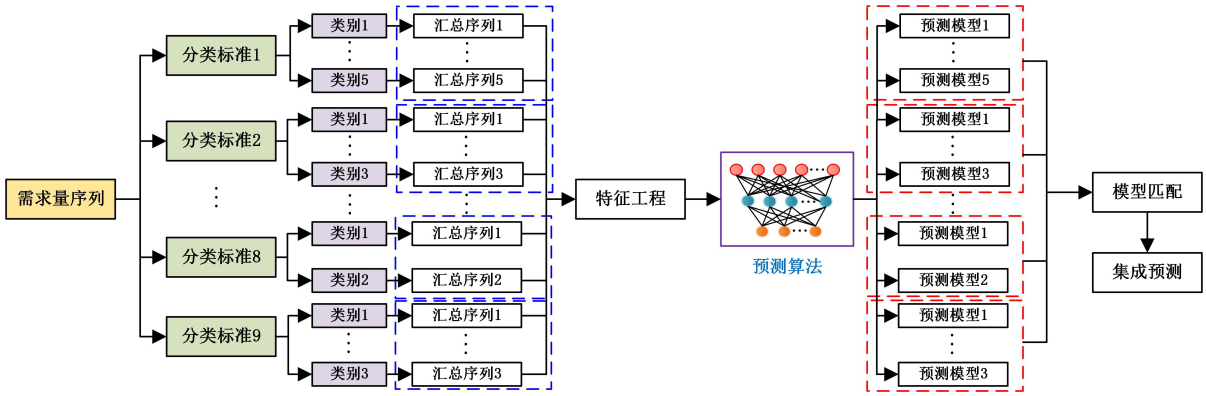


图 2.2 基于分层聚类的集成预测算法实施架构

2.2 问题二的分析

问题二主要是考虑出现了新的商家+仓库+商品维度，对这些实例未来 15 天的需求量进行预测。这些实例主要有以下特征：1）出货量序列信息不一致，部分实例由于是新上市或改变了原先存放的仓库，导致不同序列长度不一致；2）出现了新的第一层级类别，使得难以直接借助问题一的架构进行分类和预测。

针对上述问题，首先，将附件 1 与附件 5 的数据整合（1996 条+210 条），发现整合后的商家-仓库-商品序列在 2023/4/23-2023/5/15 期间需求量序列维度一致。对

2023/4/23-2023/5/15 期间所有数据提取数理特征，并与独热编码后的 8 个解释变量特征（商品一级类别等）汇整，构建特征集合，采用 K-means 算法对所有实例进行聚类，将 2206 条序列分为 14 类。然后，采用滑动窗口对聚类后各类别序列进行特征提取，采用支持向量机（SVM）、XGBoost、LightGBM、LSTM 等算法分别训练预测模型。在预测阶段，首先对待预测实例进行分类，基于该类别中 n 个算法模型得到 n 个不同的预测值，并将各模型预测值进行集成得到最终的预测输出，整体流程如图 2.3 所示。

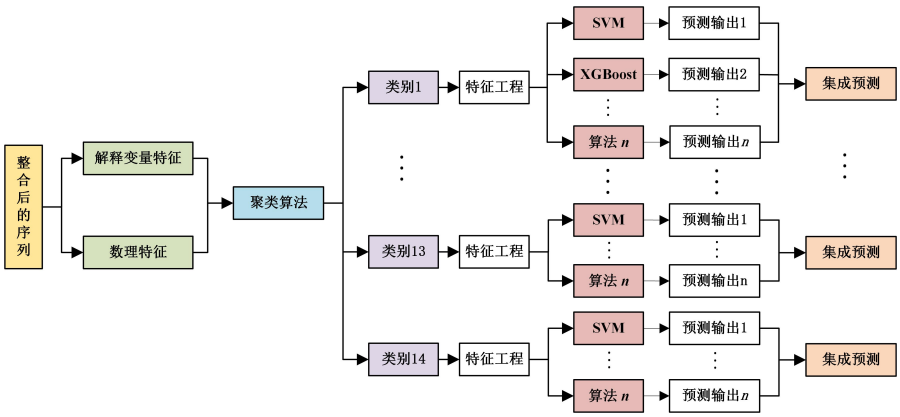


图 2.3 问题二求解流程

2.3 问题三的分析

问题三主要是考虑在出现规律性大型促销活动时，对实例在 2023/06/01-2023/06/20 期间的需求量进行预测。根据实际分析，将 2023/06/01-2023/06/20 期间活动峰值确定在 2023/06/18（京东 618）。大型促销活动期间，商品的需求量将出现以下特点[10]：1）序列无序性增强：在促销活动期间，商品出货量通常会经历销售峰值，出货量显著增加，商品库存快速减少，导致序列的规律性大幅降低；2）样本稀少：规律性的促销活动只在一段时间内定期出现，相应的历史数据较为稀少，样本的数理特征难以挖掘和关联。

针对上述问题，首先，提取附件 1 中各商品每月 1 日-11 日的出货量数据（共 5 个月）组成日常序列，并将各商品以月为基准取均值得到日常均值序列。提取附件 6 中各商品 1 日-11 日的出货量数据组成促销序列。然后，根据日常均值数据与促销数据计算各商品 1 日-11 日的促销力度值。基于问题一提出的集成预测框架对各商品 6 月 1 日-20 日的日常值进行预测。最后，将预测结果与对应的促销力度值求积，得到最终的促销预测值，整体流程如图 2.4 所示。

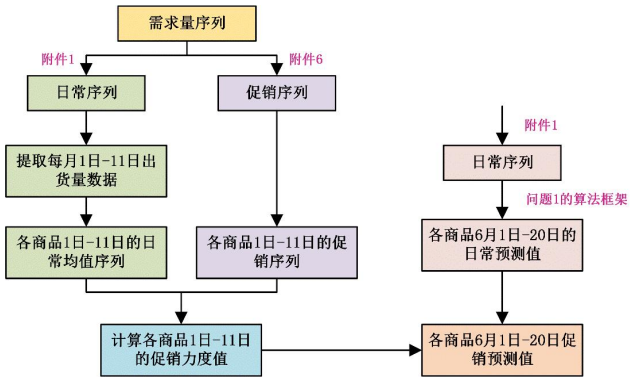


图 2.4 问题三求解流程

3 模型假设

1、假设所调查的商品需求量序列的数据都比较准确；
2、假设附件中的数据在一定时间内是稳定的，除季节性、趋势性和周期性特征外，统计特性在相近时间段内基本保持一致；
3、假设附件中数据点之间存在一定程度的自相关性和依赖性，即过去的数据点对未来的数据点有影响；
4、同一由商家-仓库-商品组成的序列在规律性促销期间出货流量无显著性差异；

4 符号说明

符号	符号说明
C_i	第 i 个分类标准
$\bar{x}$	序列 x 的均值
σ	标准差
$Kurtosis$	峰度
S	偏度
$\hat{y}$	序列的预测值
α	平滑系数
d	滑动窗口长度
W	权重系数
δ	K-means 收敛阈值
f	傅里叶频率
$I(\bullet)$	指示函数

5 问题一模型建立与求解

问题一主要涉及分类和预测两个任务。针对聚类任务，本文提出了一种基于数理特征与专家经验的分层时序聚类算法。首先，基于附件 1-4 的数据构建全体需求量序列集。然后，根据表 1 中的经验特征分别进行第一层级分类。为了缩小各分类标准所含类别的个数，对第一层级分类结果中各类别数据进行数理特征提取。最后，采用 K-means 对每个分类标准下的类别数进行聚类，获得各标准下的第二层级细化类别。

针对预测问题，本文提出了一种基于自顶向下和分层聚类的集成预测算法。首先，对 9 个分类标准各类别下的实例进行汇总，获得各类别下总的需求量序列。然后，对各类别下总的需求量序列进行特征提取，获得表征各类别需求量序列变化趋势的显著特征。最后，分别采用极致梯度提升（XGBoost）、轻量级梯度提升机器学习（LightGBM）、长短时记忆网络（LSTM）等算法对各分类标准类别下的总需求量序列构建预测模型，并设计合适的指标评估各算法的优劣。在预测阶段，基于各分类标准构建的模型采用自顶向下的方案得到 9 个预测值并将其集成作为最终的预测量。关键流程如图 5.1 所示。

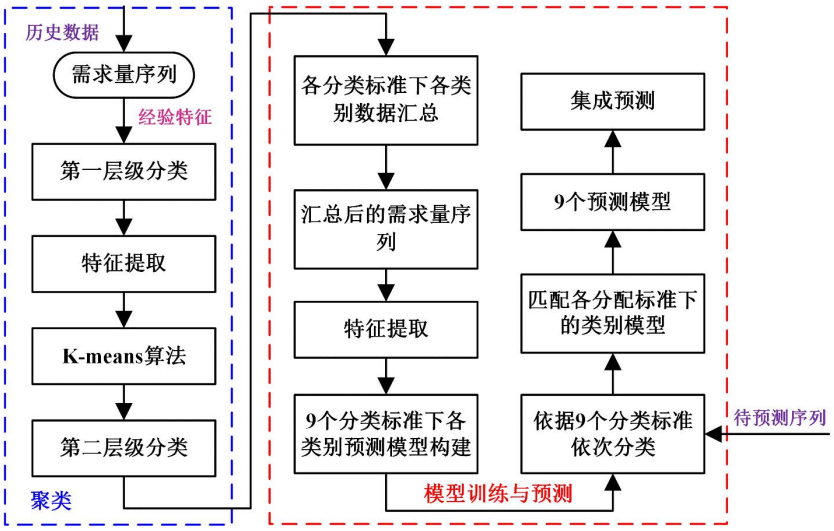


图 5.1 问题一模型建立与求解流程

5.1 基于数理特征与专家经验的分层时序聚类算法

5.1.1 聚类特征提取

根据附件 1 中的数据，首先构建由商家-商品-仓库形成的需求量序列，共计 1996 条时间序列实例，其中每个实例是由 2022/12/01-2023/5/15 的出货量组成，样本长度为 166。附件 2-4 中包含了 8 个经验特征：商品一级分类、商品二级分类、商品三级分类、商家分类、库存分类、商家规模、仓库类别、仓库区域。分别统计 8 个经验特征的类别数，结果如表 2.1 所示。这些经验特征覆盖了一些由功能场景、仓库区域等组成的先验分类知识，可以为需求量序列的划分提供经验知识。然而，若直接以经验特征划分类别存在两个问题：1）部分经验特征包含的类别数较多，如指标“商品二级分类”包含厨房卫浴、电工电料、洋酒等 79 个类别，类别繁杂且难以直接合并；2）若仅依靠 8 个专家指标将忽视需求量序列的数理特征和本身属性，难以保证各类别下序列的时序趋势相似。

针对上述问题，构建分层聚类算法，融合抽象特征与领域知识，为不同序列的划分提供多个尺度的分类依据。其中，第一层级为各经验特征形成的分类标准，对于第一层级分类标准下类别数较多的特征采用聚类算法进行二次分类，待聚类特征如表 5.1 所示。

表 5.1 经验特征与聚类标准

特征字段名	编号	第一层级类别数	是否聚类
qty_seq	分类标准 1（C_1）	1996	✓
category1	分类标准 2（C_2）	23	✓
category2	分类标准 3（C_3）	79	✓
category3	分类标准 4（C_4）	291	✓
seller_category	分类标准 5（C_5）	17	✓
inventory_category	分类标准 6（C_6）	4	×
seller_level	分类标准 7（C_7）	4	×
warehouse_categoty	分类标准 8（C_8）	2	×
warehouse_region	分类标准 9（C_9）	7	✓

其中，分类标准 1 表示每个需求量序列为一个类别，为最细化的情形。从表 5.1 可以看出，共有 6 个分类标准需要进行二次聚类。其中，每个由商家-商品-仓库组成的需求量序列长度为 166，包含了 5 个半月的出货量数据。为了捕获需求量序列的周期性、季节性等特征，同时考虑实际商品销售时，相近日期出货量的强相关性，对序列进行划分，以半个月为周期将原始的需求量序列分为 11 段长度为 15 的新序列。针对每段新序列，构建如下的聚类特征：**平均值、标准差、最大值、最小值、趋势性强度、峰度和偏度**，共计 7 个特征。其中，趋势性强度、峰度和偏度的定义与计算方式如下：

1）趋势性强度：衡量一段时间序列数据中趋势的强度或趋势的持续性。具体地，采用线性回归拟合时间序列数据，得到如下方程：

$$y = k_a x + b \tag{5-1}$$

其中，y 是需求量序列，x 是时间，k_a 为斜率反映序列趋势的方向和强度。正斜率表示上升趋势，负斜率表示下降趋势，斜率的绝对值越大，趋势的强度越大。

2）峰度：峰度用于度量随机变量概率分布的陡峭程度，表征了数据分布的尖锐程度，或者说是数据集中在均值附近的程度。计算公式如下：

$$Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{\sigma}\right)^4\right] \tag{5-2}$$

其中，n 为序列数据点的长度，x_i 为每个数据点，$\bar{x}$ 为均值，σ 为标准差。

3）偏度：偏度用来度量随机变量概率分布的不对称性。用来衡量数据分布的偏斜程度，即数据分布的尾部在均值的哪一侧更加重或更加轻。计算公式如下：

$$S = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{x_i - \bar{x}}{\sigma}\right)^3\right] \tag{5-3}$$

各分段序列长度为 15，由上述 7 个特征表示，则每条序列可以由 77 个特征点进行刻画。对于分类标准 1，需要对 1996 条序列直接进行抽象聚类，构造的聚类特征集为 1996*77。对于分类标准 2-5 和 9，由于在第一层级已经借助先验知识进行了数据的初步划分，则聚类的目标是将第一层级的类别进行二次细化聚类，使得类别划分的合理性更强，具体的聚类流程如图 5.2 所示。

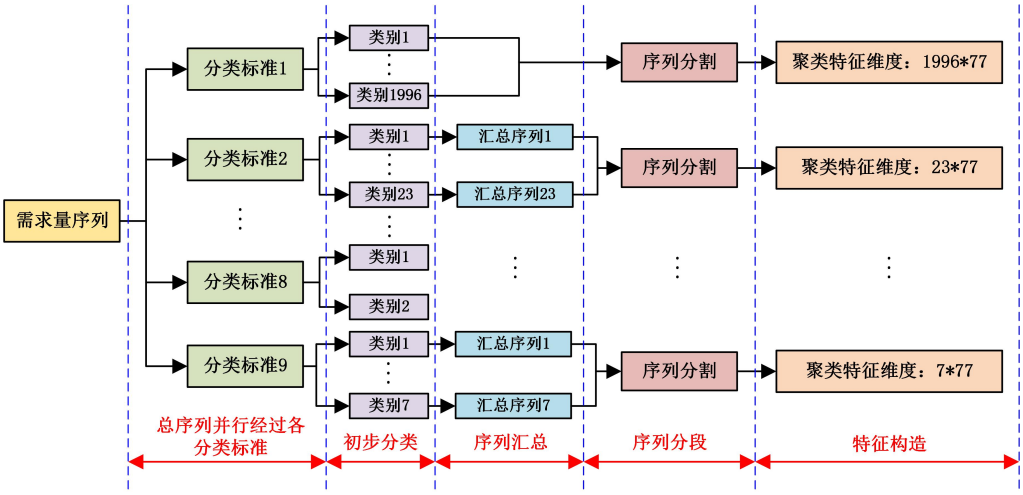


图 5.2 各分类标准特征提取流程

5.1.2 K-means 聚类算法构建

基于上述构造的特征集对各分类标准的第一层级类别进行规整，采用 K-means 算法进行聚类细化，算法执行流程如下：

1）设定 k 值范围 $k \in [1, k_{\max}]$，计算每个数据到 k 个初始化聚类中心的距离，将数据分到距离聚类中心最近的数据集中，当所有数据对象都划分以后，就形成了 k 个数据集。

2）重新计算每个簇的数据对象的均值，将均值作为新的聚类中心。

3）计算每个数据对象到新的 k 个初始化聚类中心的距离，重新划分，直到所有的数据对象无法更新到其他的数据集中。

4）采用肘部法则和轮廓系数评估聚类结果，进而选择合适的 k 值。

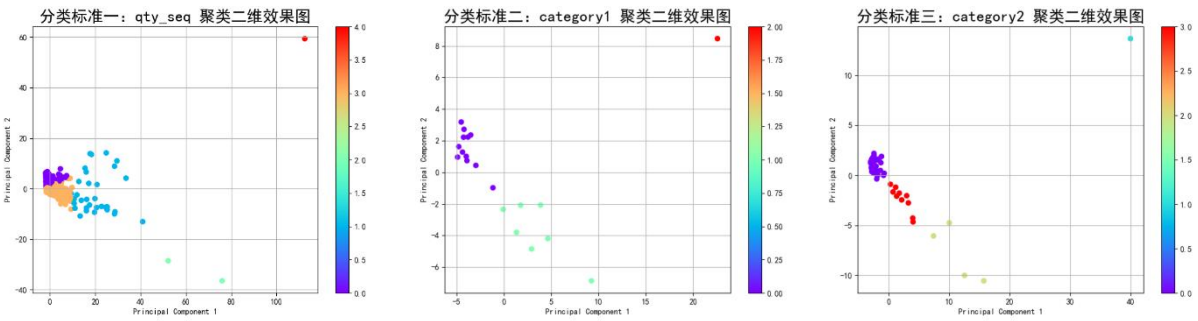
具体地，针对每个分类标准的特征集进行 K-means 聚类，得到的聚类结果如表 5.2 所示。其中，第二层级类别数相对于第一层级显著减少，使得类别划分更为详细且同一类别序列的特征分布更为相似。本文提供了 9 个分类标准，当出现新的序列时，可以选择任意分类标准对其进行分类。

表 5.2 K-means 聚类后的类别数

特征字段名	第一层级类别数	第二层级类别数
qty_seq	1996	5
category1	23	3
category2	79	4
category3	291	4
seller_category	17	4
inventory_category	4	4
seller_level	4	4
warehouse_categoty	2	2
warehouse_region	7	3

5.1.3 聚类结果评估

基于 K-means 聚类方法，采用 5.1.1 中所述的几类特征指标对第一层级类别数较多的特征进行聚合，其中分类标准 2-5 和 9 需要进行 K-means 聚类细化。将聚类后的结果使用主成分分析（PCA）进行降维并可视化聚类结果，结果如图 5.3 所示。



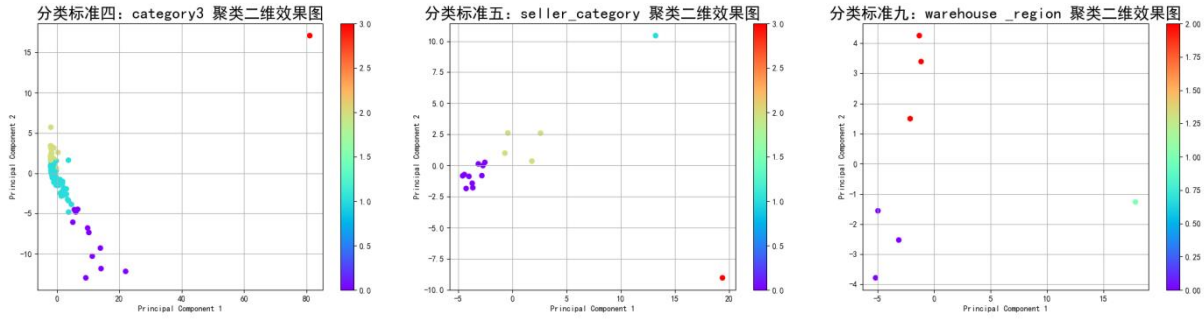


图 5.3 K-means 聚类后的可视化分类图

为了更直观的分析聚类效果，将聚类之后各分类标准下属于同一类别的需求量序列绘制为折线图，结果如图 5.4 所示。从图中可以看出，本文构建的聚类方案可以较好地挖掘出序列中类分离的显著特征。基于第一层级的字段名分类利用到了领域的先验知识，使得类别划分的可解释性增强。提出的 9 个分类标准都能较好地时间序列进行分类，使同一类别在需求上的特征最为相似。

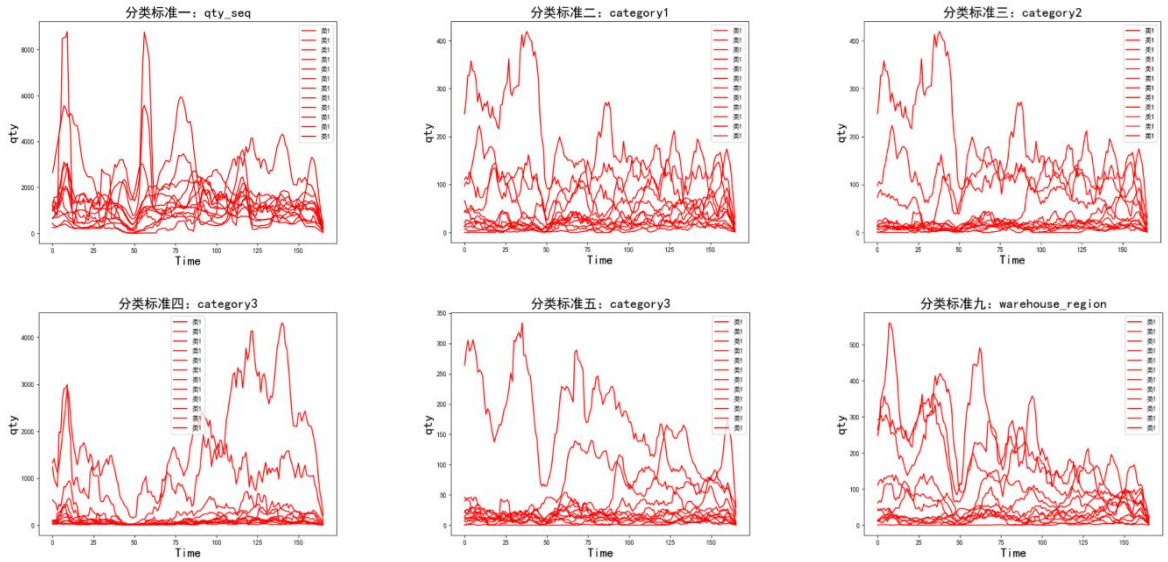


图 5.4 各分类标准下类别 1 中序列的分布情况

5.2 基于自顶向下和分层聚类的集成预测算法

针对预测任务，本文提出了一种基于自顶向下和分层聚类的集成预测算法。首先，对 9 个分类方案各类别下包含的实例进行求和处理，获得各类别下汇总的需求量序列，共计 33 条。然后，对各分类方案的各类别下汇总的需求量序列进行特征提取，获得表征各类别需求量序列变化趋势的显著特征。最后，分别采用指数平滑（ES）、自调节的差分整合移动平均自回归模型（AutoARIMA）、XGBoost、LightGBM、LSTM 预测算法对各类别下的总需求量序列构建时序预测模型（各算法下均构建 33 个模型），并设计合适的指标评估各算法的优劣，进而选择最优的预测算法。

在预测阶段，针对待预测的需求量序列，先按照 9 个分类标准将其分别划分到第二层级的各类别中，获得各类别汇总的需求量序列。然后，将单序列的预测任务转化为其所属类别汇总需求量的预测任务，并基于各类别构建的时序预测模型获得汇总序列的预

测值。其次，在各分类标准下，计算待预测序列占其归属类别汇总的序列的比例，采用自顶向下的方案获得单序列的预测值。最后，将 9 个分类标准下的预测值进行集成作为最后的预测输出，框架执行流程如图 5.5 所示。

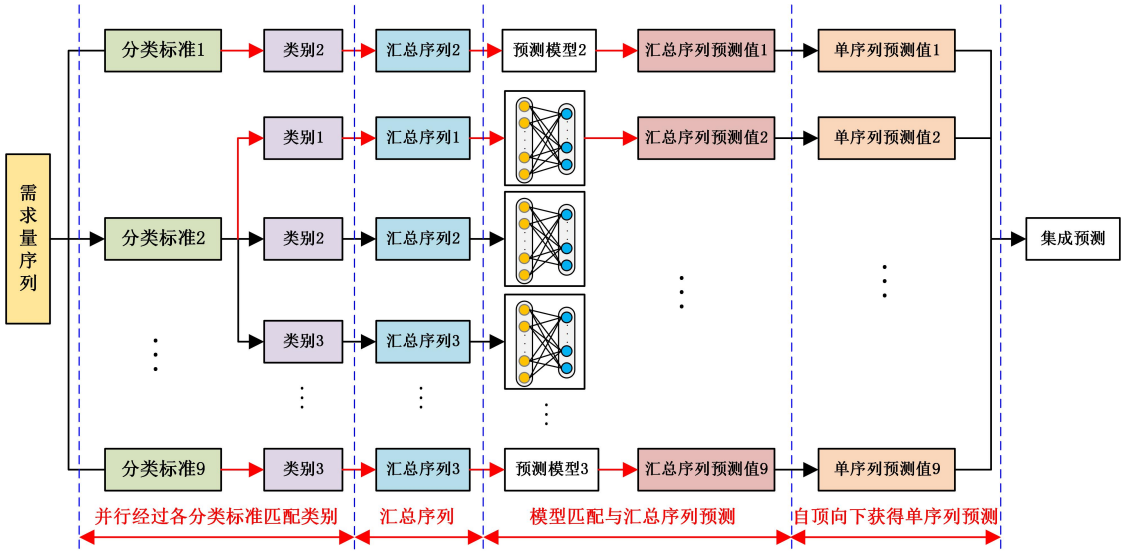


图 5.5 基于自顶向下和分层聚类的预测任务流程

注意到，在训练阶段，9 个分类标准下的 33 个预测模型均基于各类别汇总后的需求量序列进行构建，因此在对单需求量序列进行预测时，需要先匹配其在 9 个分类标准下的类别归属，将任务转换为汇总序列的预测，然后计算单需求量序列在其所属类别汇总需求量序列的比例因子，采用自顶向下的方案求得单序列的预测值：对层次结构的顶层进行预测，然后通过预定义的比例将其分解到所属类别的节点。

假设 y_A 为某条需求量序列，长度为 166，现需对其未来 15 天的需求量进行预测。在分类标准 1 中其被划分到了类别 2，类别 2 中汇总后的需求量序列为 $y_{C_{12}}$，为类别中所有实例的加和，长度也为 166。采用预测模型对汇总后的序列 $y_{C_{12}}$ 进行预测，得到未来 15 天汇总序列的预测值 $\hat{y}_h$，则单需求量序列的预测值可以通过下式计算：

$$\hat{y}_{A,h} = p \bullet \hat{y}_h, \quad p = \frac{1}{T}\sum_{t=1}^{T}\frac{y_{A,t}}{y_{C_{12},t}} \tag{5-4}$$

式中，p 为比例因子，T 表示天数。比例因子通过计算 T 天内单序列每天的出货量占汇总序列每天出货量的均值求得。在本文中，采用预测任务之前的 30 天每天的出货量占加和序列每天出货量比重的均值进行表征。依次计算待预测序列在各分类标准下的预测值，采用 9 个预测量的均值作为最后的预测量：

$$\hat{y}_A = \frac{1}{9}\sum_{h=1}^{9}\hat{y}_{A,h} \tag{5-5}$$

基于图 5.5 的框架，分别采用 ES、Auto ARIMA、XGBoost、LightGBM、LSTM 算法训练预测模型，并设计合适的指标评估各算法的优劣，进而选择最优的预测算法。

5.2.1 指数平滑预测模型构建

指数平滑又称为指数修匀，其原理是将历史数据进行加权平均作为未来时刻的预测

结果，并且随着过去观测值离预测值距离的增大，权重呈指数型衰减。具体的算法流程如下：

1）对原始数据进行预处理，消除时间序列中偶然性的变化，提高收集的数据中近期数据在预测中的重要程度。

2）考虑趋势与季节性的影响，构建三次指数平滑模型：

$$S_{t+1}^{(1)}=\alpha X_t+(1-\alpha)S_t^{(1)},S_{t+1}^{(2)}=\alpha S_t^{(1)}+(1-\alpha)S_t^{(2)},S_{t+1}^{(3)}=\alpha S_t^{(2)}+(1-\alpha)S_t^{(3)} \quad (5\text{-}6)$$

式中，S_t 为第 t 天的预测值，X_t 为第 t 天的实际值，α 为平滑系数。预测未来 T 天：

$$y_{t+T}=A_t+B_tT+C_tT^2 \quad (5\text{-}7)$$

式中，

$$\begin{cases} A_t=3S_t^{(1)}-3S_t^{(2)}+S_t^{(3)},\ C_t=\left(\dfrac{\alpha}{2(1-\alpha)^2}\right)\left(S_t^{(1)}-2S_t^{(2)}+S_t^{(3)}\right) \\ B_t=\left(\dfrac{\alpha}{2(1-\alpha)^2}\right)\left((6-5\alpha)S_t^{(1)}-2(5-4\alpha)S_t^{(2)}+(4-3\alpha)S_t^{(3)}\right) \end{cases} \quad (5\text{-}8)$$

3）使用历史时间序列数据训练指数平滑模型，通过趋势调整、添加趋势修正值改进指数平滑法的预测结果。最后，使用训练好的模型生成未来时间点的预测值。

5.2.2 Auto ARIMA 预测模型构建

ARIMA 模型是一种广泛应用的时间序列预测模型，它结合了自回归（AR）和移动平均（MA）的概念，具有较好的灵活性和准确性。Auto ARIMA 提供了模型阶次自适应寻优方案，指定一个阶数上限和信息准则计算方法，从而找到信息准则最小的阶数组合，避免了传统 ARIMA 复杂的参数调整过程。其中，ARIMA 模型的公式为：

$$Y_t=c+\varphi_1Y_{t-1}+\varphi_2Y_{t-2}+\ldots+\varphi_pY_{t-p}+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}+\varepsilon_t \quad (5\text{-}9)$$

式中，Y_t 为考虑的时间序列数据，φ_1 到 φ_p 为 AR 模型的参数，θ_1 到 θ_q 为 MA 模型的参数。ε_t 为 t 时间点上的误差项，c 为常数项。对于第二层级各类别汇总的需求量序列分别训练 Auto ARIMA 模型（共计 33 个），算法构建过程如图 5.6 所示。

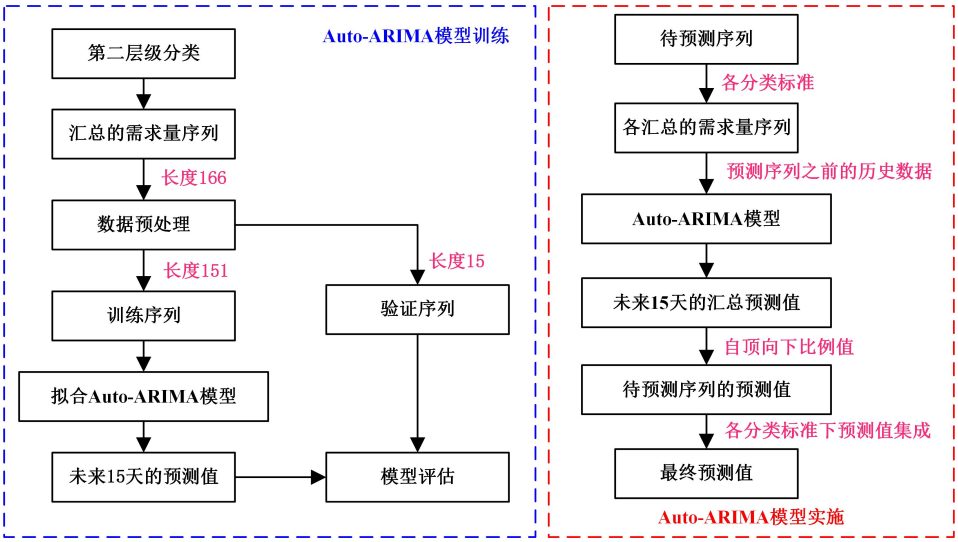


图 5.6 Auto ARIMA 模型的构建流程

其中，训练样本为 2022/12/01-2023/4/30 期间出货量组成的长度为 151 的汇总序列。验证样本为汇总实例未来 15 天的出货量，即 2023/5/01-2023/5/15 期间的出货量。在模型实施阶段，用各需求量序列 2023/5/15 之前的历史数据经过构建好的 Auto ARIMA 模型预测 2023/05/16-2023/05/30 的需求量。

5.2.3 XGBoost 预测模型构建

XGBoost 模型是梯度分类和回归问题的有效实现，它是决策树算法的集合，其中新树修复了那些已经属于模型的树的错误，将添加树直到无法对模型进一步改进为止。XGBoost 提供了随机梯度提升算法的高效实现，并提供了一组模型超参数，这些参数旨在提供对模型训练过程的控制。已知训练数据集 $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$，损失函数 $l(y_i,\hat{y}_i)$，正则化项 $\Omega(f_k)$，则整体目标函数为：

$$L(\phi)=\sum_i l(y_i,\hat{y}_i)+\sum_k \Omega(f_k) \tag{5-10}$$

式中，$L(\phi)$ 是线性空间的表达，i 为样本编号，k 表示数的编号，$\Omega(f_k)$ 反映了 k 棵树的复杂度。当 XGBoost 模型构建完成，第 i 个样本的预测值为：

$$\hat{y}_i=\sum_{k=1}^{K} f_k(x_i) \tag{5-11}$$

XGBoost 为有监督学习，针对问题一中的任务，采用 XGBoost 构建预测模型的核心是数据集的设计。取 2022/12/01-2023/4/30 期间出货量组成的长度为 151 的汇总序列为训练集，2022/5/01-2023/5/15 期间的出货量为验证集，设置滑动窗口为 15，截取长度为 15 的切割序列作为输入样本 x_i，切割序列后一天的需求量为输出标签 y_i，通过滑动窗口构建最终的训练集合，如图 5.7 所示。当模型训练结束后，按照图 5.5 的流程进行预测。

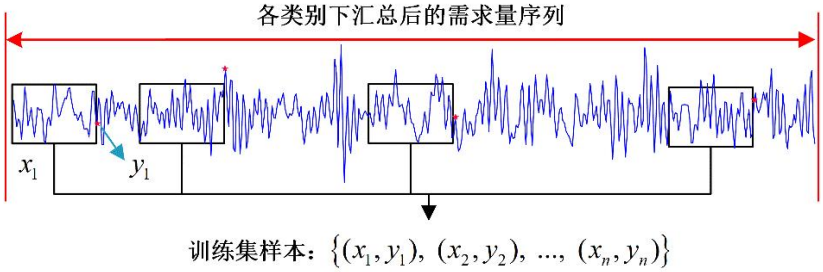


图 5.7 XGBoost 训练集构造

5.2.4 LightGBM 预测模型构建

LightGBM 是个快速、分布式、高性能的基于决策树算法的梯度提升框架。可以克服 XGBoost 空间消耗和计算量大的缺陷，并且能够在不损害准确率的条件下加快梯度提升决策树（GBDT）模型的训练速度，算法的基本流程如图 5.8 所示。

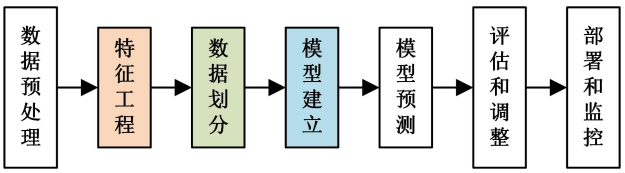


图 5.8 LightGBM 模型工作流程

与 XGBoost 相似，LightGBM 也为有监督学习，合理的特征工程会大幅提高模型的效果。因为树模型是基于分裂点的，它对于特征与特征之间的关系不敏感。为了捕获需求量序列的周期性、季节性、滞后性等特征，同时考虑在实际商品销售时，相邻日期出货量的强相关性，对于 2022/12/01-2023/4/30 期间出货量组成的长度为 151 的汇总序列进行分段，将其作为训练数据。设置滑动窗口为 30，截取长度为 30 的切割序列作为新数据，切割序列后一天的需求量为输出标签，获得的切割序列个数为：

$$N_s = l_s - d - 1 \tag{5-12}$$

式中，l_s 为原始序列的长度，d 为滑动窗口的长度。按照上述参数设定，切割后的序列个数为 120 个，长度为 30。针对每段新序列，构建如下的特征：**1）滞后特征**：包括各切割序列倒序 1 天到倒序 7 天的需求量之和；倒序 7 天、14 天、21 天、28 天的需求量均值和标准差；**2）数理特征**：切割序列的最大值和最小值；切割序列的趋势性强度；**3）周期性/季节性特征**：周期性/趋势性强度。共计 7 个特征，通过特征工程后，获得 120*7 的特征集，取 2022/5/01-2023/5/15 期间的出货量作为验证集训练 LightGBM 模型。当模型训练完成后，按照图 5.5 的流程预测 2023/05/16-2023/05/30 的需求量。

5.2.5 LSTM 预测模型构建

LSTM 是循环神经网络（RNN）的变体，它通过使用记忆单元和门控机制来解决传统 RNN 中的梯度消失和梯度爆炸问题。核心是由输入门、遗忘门、输出门和记忆单元组成。其中，输入门用于控制输入信息的更新，遗忘门用于控制记忆的遗忘，输出门用于控制输出信息的选择，记忆单元用于存储和更新信息。基于上述机制，能够挖掘出数据中的时序依赖关系，算法的基本流程如图 5.9 所示。

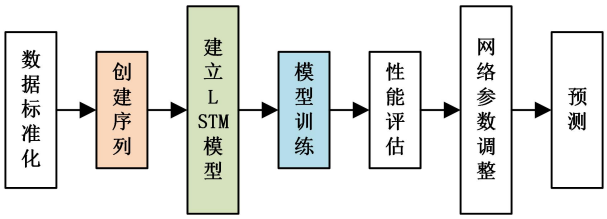


图 5.9 LightGBM 模型构建流程

LSTM 数据集的构造 XGBoost 相似。其中，训练样本为 2022/12/01-2023/4/30 期间各类别所包含实例出货量组成的长度为 151 的汇总序列，设置滑动窗口为 15，截取长度为 15 的切割序列作为输入样本，切割序列后一天的需求量为输出标签。验证样本为汇总实例未来 15 天的出货量，即 2023/5/01-2023/5/15 期间的出货量。

5.3 任务预测与性能评价

基于附件 1-4 中商品-商家-仓库在 2022/12/01-2023/5/15 期间形成的需求量序列和图 5.5 的框架，分别采用 ES、Auto ARIMA、XGBoost、LightGBM、LSTM 算法训练预测模型，各算法的拟合能力和在验证集上的表现如图 5.10 和 5.11 所示。从图中可以看出，提出的各算法均可以挖掘出时序数据中的抽象特征，实现对出货量的预测。其中，ES

和 Auto ARIMA 主要是基于统计学原理进行模型训练，预测输出较为平缓。XGBoost 和 LightGBM 基于决策树的架构进行模型训练，可以捕获序列中周期性、季节性、滞后性等特征，具有适中的计算复杂度。LSTM 主要是基于神经网络强大的学习能力来抽取序列中的抽象特征，在训练集上展现了优异的能力，较好地预测出未来需求量的走势。

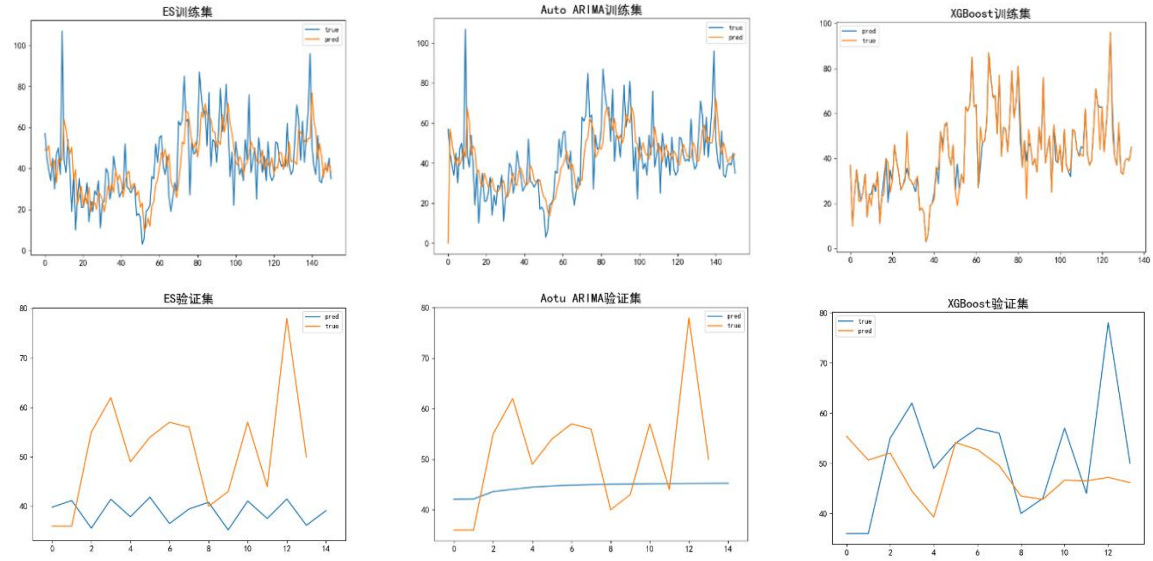


图 5.10 ES、Auto ARIMA 和 XGBoost 对部分类别的拟合与预测结果

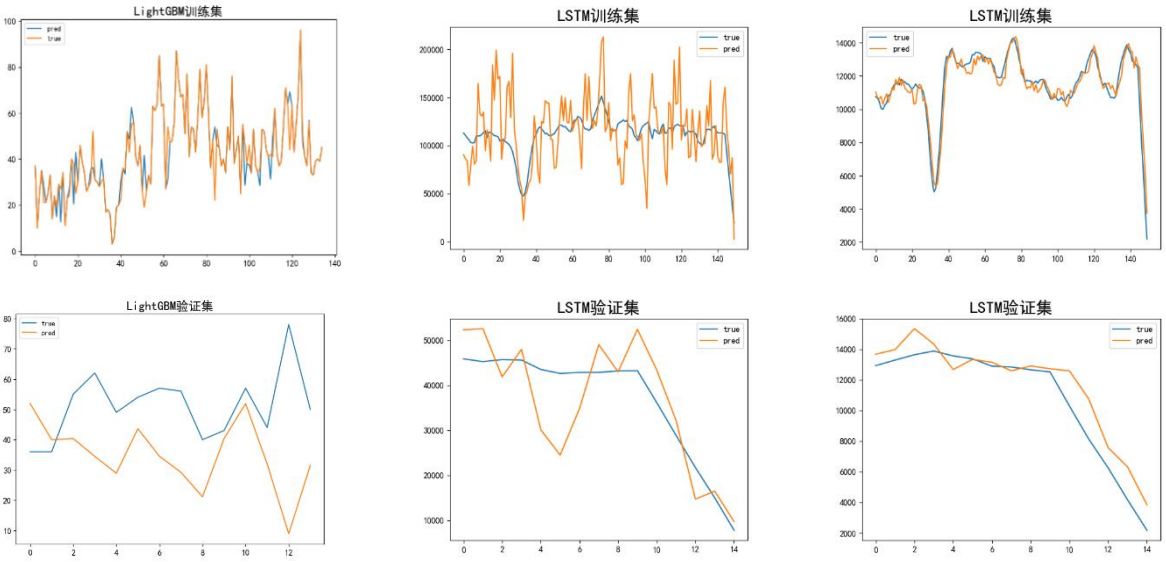


图 5.11 LightGBM 和 LSTM 对部分条例的拟合与预测结果

针对各类算法，均取 2023/05/01-2023/05/15 期间 15 天的出货量作为验证集。选取以下指标对各算法性能进行评价：

1）1-wmape：表示预测值与实际值之间的平均绝对百分比误差。考虑了不同预测值的权重以更好地反映不同数据点的重要性。较低的 WMAPE 值拥有更准确的预测，计算公式如下：

$$1-\text{wmape}=1-\frac{\sum|y_i-\hat{y}_i|}{\sum y_i} \tag{5-13}$$

式中，y_i 为验证集的真实值，$\hat{y}_i$ 为模型的预测值。由于部分序列在 2023/05/01-2023/05/15

期间真实值均为 0，为了避免分母奇异，采用修订后的下式进行计算：

$$1-\text{wmape}=1-\frac{\sum \text{len}(\hat{y}_i \neq 0)}{\sum \text{len}(y_i)} \tag{5-13}$$

式中，$\text{len}(\hat{y}_i \neq 0)$ 表示预测序列中值不为 0 的天数。修订后的指标反映了预测序列中不为 0 的个数占总预测长度的比例。

2）RMSSE：表示模型的均方根比例误差。它量化了模型的预测误差相对于一个简单基准模型（通常是历史平均值）的性能。它不受数据规模的影响，因此可以用于不同问题和数据集的比较。较低的 RMSSE 拥有更好的预测性能，其计算公式为：

$$\text{RMSSE}=\sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(y_t-\hat{y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(y_t-y_{t-1})^2}} \tag{5-14}$$

式中，n 为训练样本序列的长度，h 为验证集序列的长度。

3）MAPE：平均绝对百分比误差，用于度量预测值与实际观测值之间的百分比误差的平均值。有助于反映预测值相对于实际值的偏离程度，其不受数据的绝对值或单位的影响。较低的 RMSSE 拥有更好的预测性能，计算公式为：

$$\text{MAPE}=\frac{1}{h}\sum_{i=1}^{h}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{5-15}$$

4）RMSE：均方根误差，用于量化模型的预测值与实际观测值之间的差异。该指标较为稳定，不受异常值的干扰，不依赖于残差的分布，可解释性较强。较低的 RMSE 拥有更好的预测性能，计算公式为：

$$\text{RMSE}=\sqrt{\frac{1}{h}\sum_{i=1}^{h}(y_i-\hat{y}_i)^2} \tag{5-16}$$

基于上述指标，各算法预测性能的评价值如表 5.3 所示。从表中看出，在指标 1-WAMAPE 下，LightGBM 算法在各分类标准中均取得了较好的预测性能。虽然，Auto ARIMA 在 RMSSE、MAPE、RMSE 指标上与 LightGBM 具有相媲美的性能，但是它在验证集上的预测往往表现为输出一段平缓的均值，难以捕获序列中的滞后性和波动性特征。同时，针对实际预测任务，当待预测样本按照 9 个分类标准划分结束后，电商平台也可以依据 4 个指标选取各类别下最优的算法执行集成预测任务。

表 5.3 不同算法下预测性能评价值

平均评价指标		1-WAMAPE					RMSSE				
		LightGBM	ARIMA	EM	XGBoost	LSTM	LightGBM	ARIMA	EM	XGBoost	LSTM
数理特征	类别 1	0.7929	**0.8058**	0.7887	0.7451	0.7537	2.9965	2.3264	2.3310	3.5776	10.3783
	类别 2	**0.8967**	0.8733	0.8799	0.8568	0.7887	1.7860	1.3154	1.4094	2.5044	9.5255
	类别 3	0.7212	0.4428	0.5179	0.4178	**0.7225**	4.5650	7.3000	6.2816	9.5181	9.2683
	类别 4	**0.9576**	0.9416	0.8665	0.7022	0.9254	2.8811	2.7681	5.0897	12.8474	11.9762

	类别 5	0.6298	**0.6351**	0.5847	0.5732	0.3040	1.9538	0.7305	0.8943	2.1995	9.7517
商品一级分类	类别 1	**0.9166**	0.8412	0.8450	0.8142	0.8727	4.8969	8.1522	8.0219	9.1089	12.4504
	类别 2	0.9541	**0.9549**	0.7798	0.7189	0.8219	2.0891	1.7577	7.2852	10.5636	103.0375
	类别 3	0.7922	**0.8088**	0.7070	0.702	0.5322	1.9268	1.3005	1.7243	2.8260	25.6239
商品二级分类	类别 1	**0.9384**	0.8667	0.8163	0.8039	0.8851	3.8935	6.3905	8.3513	9.7088	10.6806
	类别 2	0.9667	**0.9681**	0.8317	0.7709	0.8141	1.4344	1.3744	5.3125	7.9500	12.9854
	类别 3	0.7619	**0.7829**	0.6446	0.6570	0.4993	1.8878	1.2696	1.8020	2.7629	28.3575
	类别 4	**0.9245**	0.8984	0.5880	0.8009	0.8371	3.0316	2.9967	10.2036	6.0462	9.9205
商品三级分类	类别 1	0.7807	0.6252	0.6298	0.7438	**0.8170**	7.6614	12.4993	12.1666	9.9546	9.7974
	类别 2	0.7592	**0.7748**	0.6218	0.7067	0.4924	1.8512	1.1853	1.8102	2.2477	28.2107
	类别 3	**0.9613**	0.9519	0.8565	0.7525	0.8946	2.3173	2.5553	5.7816	11.0620	9.2232
	类别 4	**0.9401**	0.9193	0.7203	0.7584	0.9136	2.0580	2.4157	7.5645	7.2030	6.0436
商家分类	类别 1	**0.9479**	0.9120	0.8540	0.8159	0.8138	3.1818	4.7265	6.5790	8.3305	13.3033
	类别 2	**0.9334**	0.8968	0.7442	0.8222	0.8901	1.9781	2.7678	5.9227	4.4713	8.3497
	类别 3	0.7922	0.8088	0.7070	0.7020	**0.8350**	1.9268	1.3005	1.7243	2.8260	11.9548
	类别 4	**0.9606**	0.9560	0.7738	0.7779	0.8969	2.1822	1.8856	7.9513	9.2250	8.1652
库存分类	类别 1	0.8762	0.8710	0.8321	0.7656	**0.8886**	2.2821	1.3868	1.5677	3.7032	12.2229
	类别 2	**0.9584**	0.9512	0.8942	0.7305	0.8610	1.7716	1.4814	2.8839	8.4323	7.6353
	类别 3	0.8944	0.9040	**0.9123**	0.6705	0.5268	2.8595	2.4452	2.4742	9.8862	27.0617
	类别 4	**0.8968**	0.8834	0.6222	0.8529	0.8882	3.7815	3.5827	9.6849	5.1024	9.6006
商家规模	Large	**0.9183**	0.9181	0.8911	0.6802	0.7928	2.3097	1.4875	1.7610	7.1659	36.6812
	Medium	**0.9543**	0.9159	0.8458	0.8875	0.9112	2.1395	2.6708	4.4599	4.0573	7.1379
	Small	**0.9345**	0.9286	0.7998	0.9005	0.8554	2.2313	2.5315	5.2862	3.3347	9.1847
	Special	0.8192	0.6198	0.5230	0.8280	**0.8282**	4.4013	8.8819	10.8039	4.2843	10.4215
仓库类别	中心仓	**0.9261**	0.9190	0.8862	0.7248	0.7828	2.2625	1.4305	1.8714	6.5381	40.1806
	区域仓	**0.9388**	0.9278	0.8966	0.7654	0.8209	2.1593	1.7124	2.0898	6.0806	10.7697
仓库区域	类别 1	**0.9329**	0.9272	0.8985	0.8308	0.8345	1.8766	1.3808	1.7912	4.0763	10.7157
	类别 2	**0.9306**	0.9259	0.8879	0.7157	0.8131	2.1638	1.3579	1.8335	6.7081	11.6326
	类别 3	**0.9289**	0.8865	0.7888	0.8501	0.9128	3.8721	5.0272	8.4665	6.8619	8.5619

评价指标		MAPE					RMSE				
		LightGBM	ARIMA	EM	XGBoost	LSTM	LightGBM	ARIMA	EM	XGBoost	LSTM
数理特征	类别 1	17.87	17.07	21.37	22.66	32.92	969.57	902.37	904.14	1125.16	5488.83
	类别 2	10.83	13.32	12.11	13.34	20.40	708.90	805.97	863.61	1016.83	8148.83
	类别 3	28.16	56.45	48.23	58.42	32.12	166.09	358.27	308.29	359.53	9255.03
	类别 4	4.03	5.55	13.09	29.54	12.41	1164.74	1470.36	2703.55	5487.15	9381.22

	类别 5	41.05	40.49	53.49	47.67	139.78	447.64	421.97	516.64	477.66	6550.42
商品一级分类	类别 1	7.81	15.13	14.76	18.10	20.66	413.72	695.07	683.97	756.35	3160.76
	类别 2	4.61	4.48	21.71	28.04	24.34	922.84	955.10	3958.66	4862.54	17378.10
	类别 3	20.01	16.77	32.35	28.79	63.30	2032.82	2163.20	2868.02	2887.99	22699.74
商品二级分类	类别 1	5.87	12.82	17.85	19.35	17.81	571.40	1068.17	1395.92	1438.68	4893.00
	类别 2	3.32	3.12	16.63	22.77	25.32	433.53	483.93	1870.47	2469.21	10954.27
	类别 3	23.23	19.27	40.90	34.06	61.71	1975.65	2093.97	2972.03	2793.62	22678.32
	类别 4	7.54	9.99	40.60	19.72	20.54	459.15	578.66	1970.29	1000.26	4127.42
商品三级分类	类别 1	19.25	32.88	32.68	20.96	12.84	337.77	510.55	496.96	416.54	921.70
	类别 2	23.20	20.48	44.19	30.27	66.07	1849.54	1895.18	2894.42	2170.53	17222.74
	类别 3	3.76	4.59	14.12	24.54	14.30	660.25	886.82	2006.53	3270.88	8202.16
	类别 4	5.98	7.99	27.57	24.06	12.17	582.74	808.22	2530.87	2134.82	4607.71
商家分类	类别 1	4.91	8.26	14.13	18.10	23.45	401.30	639.11	889.60	1062.58	10239.67
	类别 2	6.56	10.42	25.28	17.69	12.83	471.18	742.79	1589.49	1081.31	9064.97
	类别 3	20.01	16.77	32.35	28.79	16.15	2032.82	2163.20	2868.02	2887.99	304.65
	类别 4	3.93	4.29	22.18	22.29	11.82	479.29	545.83	2301.66	2155.17	3177.66
库存分类	类别 1	11.04	11.53	16.79	21.76	17.20	1836.02	1899.52	2147.33	2879.92	4669.14
	类别 2	4.11	4.78	10.45	26.83	18.94	734.77	832.70	1621.01	3734.82	4384.15
	类别 3	10.88	10.06	8.90	33.45	67.02	36.54	31.42	31.79	124.94	23973.51
	类别 4	10.53	11.19	36.85	14.48	17.96	561.22	662.68	1791.40	840.86	6574.14
商家规模	Large	7.83	7.77	10.58	31.23	26.70	2671.31	2736.20	3239.18	8313.93	30451.76
	Medim	4.52	8.11	15.10	10.95	14.70	77.45	113.79	190.01	147.85	623.50
	Small	6.29	6.58	19.45	9.44	18.96	126.83	159.47	333.00	189.27	1287.88
	Special	16.36	33.52	43.02	15.83	19.51	336.37	678.49	825.03	321.01	1639.46
仓库类别	中心仓	7.24	7.96	11.11	26.91	27.19	1755.00	1797.14	2351.02	5137.40	20918.57
	区域仓	5.85	6.81	10.15	23.07	20.19	960.25	1095.87	1337.40	2701.69	12372.48
仓库区域	类别 1	6.52	7.00	9.93	16.30	21.71	668.02	701.90	910.49	1446.37	8757.07
	类别 2	6.68	7.21	10.97	27.87	19.58	1821.48	1839.51	2483.74	5705.42	22442.78
	类别 3	6.79	10.70	20.38	14.59	16.00	207.21	299.32	504.10	368.83	1263.04

LightGBM 在 9 个分类标准下和集成后的预测输出与真实结果如图 5.12 所示。从图中可以看出，在不同的分类标准下，9 个预测输出序列表现了相似的走势特征。但是，由于类别划分和训练模型时所用特征集的不同，9 个预测输出序列的变化区间存在差异。为了克服单模型泛化能力弱与不稳定等问题，对 9 个分类标准下的预测值进行集成作为最终的预测输出。集成后的预测值有效地平滑了各模型的缺陷，获得了更小的 WAMAPE、RMSSE 指标。

基于上述分析，采用图 5.5 中的流程框架和 LightGBM 算法对附件中各商品在 2023/05/16-2023/05/30 期间的需求量进行预测，相应结果已填写至附件中的表 1。

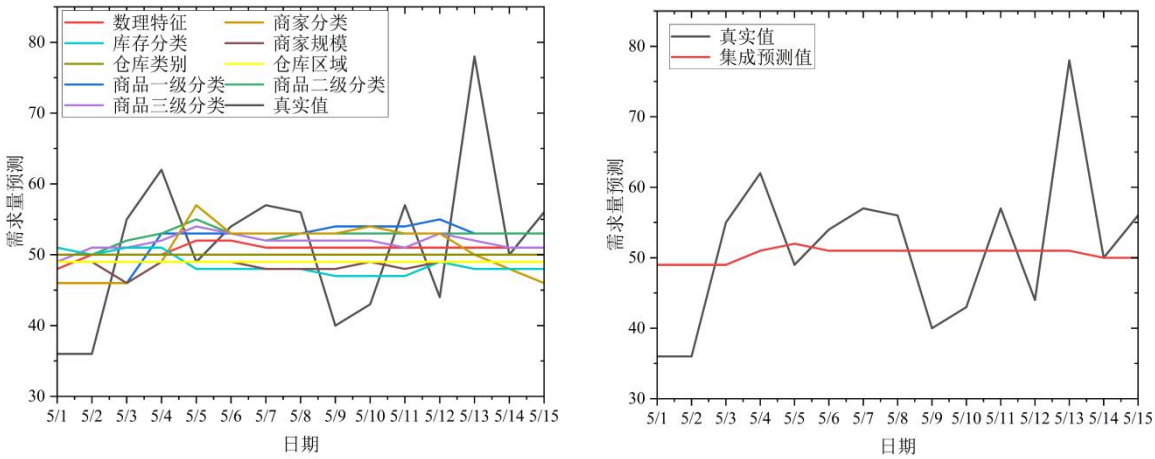


图 5.12 各分类标准下以及集成后的预测输出与真实结果

6 问题二模型建立与求解

问题二主要涉及预测任务。相对于问题一，主要是出现了新的商家+仓库+商品维度，这些实例主要有以下特征：1）出货量序列信息不一致，部分实例由于是新上市或改变了原先存放的仓库，导致不同序列长度不一致；2）出现了新的第一层级类别，使得难以直接借助问题一的架构进行分类和预测。

针对上述问题，本文提出了一种融合 K-means 聚类的多模型集成预测算法。首先，将附件 1 与附件 5 的数据整合（1996 条+210 条），对 2023/4/23-2023/5/15 期间所有数据提取数理特征，并与独热编码后的 8 个解释变量特征（商品一级类别等）汇整，构建分类特征集合，采用 K-means 算法对所有实例进行聚类，将 2206 条序列分为 14 类。然后，采用滑动窗口对聚类后各类别序列进行特征提取和样本集构建，采用支持向量机（SVM）、XGBoost、LightGBM、LSTM、KNN 算法分别训练预测模型，基于 5 折交叉验证法调整模型参数并评估算法性能。在预测阶段，首先对待预测实例进行分类，基于该类别中 5 个算法模型得到 5 个不同的预测值，并将各模型预测值进行集成得到最终的预测输出，整体流程如图 6.1 所示。

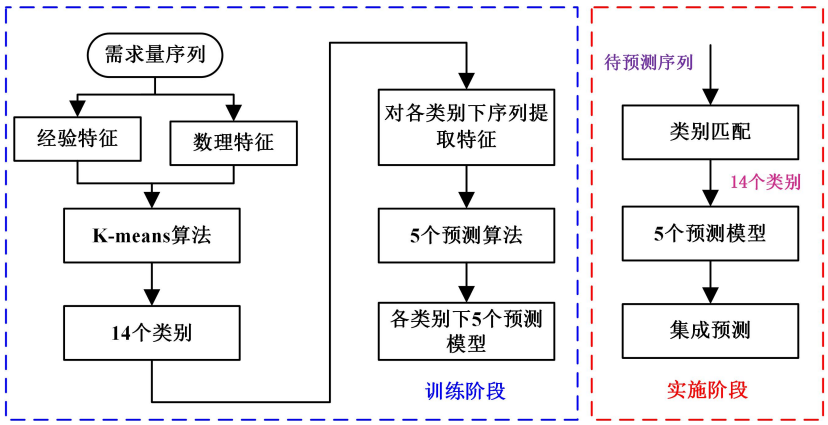


图 6.1 问题 2 框架的训练与实施流程

6.1 基于融合特征的单层聚类算法

6.1.1 特征提取

根据附件 1 和附件 5 中的数据，首先构建由商家-商品-仓库形成的需求量序列，共计 2206 条时间序列实例。由于出现了新上市的商品或改变了某些商品所放仓库未知，导致不同实例的需求量序列长度不一致。 针对该问题，对附件信息进行分析整理，截取商家-仓库-商品序列在 2023/4/23-2023/5/15 期间的数据，保证 2206 条序列长度一致。与问题一不同，问题二出现了新的第一层级类别，无法使用先验字符段特征进行初步分类。为克服该问题，采用独热编码对字符段特征进行量化表示，将这些特征与数理特征堆积共同用于抽象聚类。具体地，独热编码适用于特征取值离散且无实际意义的场景，采用 0 和 1 进行处理，以商家规模为例，编码结果如图 6.2 所示。

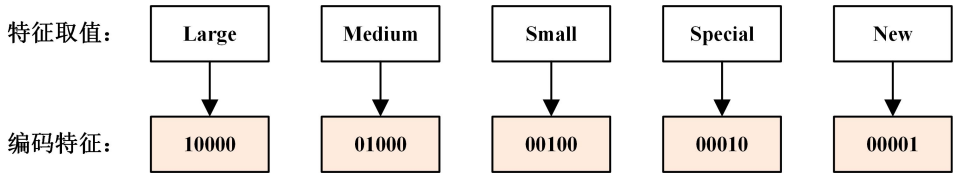


图 6.2 独热编码示意图

为了捕获需求量序列的周期性、季节性等特征，同时考虑在实际商品销售时，相邻日期出货量的强相关性，对序列进行划分。首先，以一周为周期对每条需求量序列进行划分得到 3 段长度为 7 的新序列。针对每段新序列，提取如下的聚类特征：平均值、标准差、最大值、最小值、趋势性强度、峰度和偏度，共 21 个特征。然后对原始序列提取傅里叶变换后的主频分量作为附加特征，与片段特征堆积得到维度为 2206*22 的特征集。时域信号的傅里叶变换可以表示为：

$$X(f)=\int_{-\infty}^{\infty}x(t)\bullet e^{-j2\pi ft}dt,\ \ A(f)=\left|X(f)\right| \qquad (6\text{-}1)$$

式中，$X(f)$ 为频域表示，$x(t)$ 是时域信号，f 是频率。$A(f)$ 为频率 f 处的振幅。在幅度谱中，寻找最大的振幅或能量对应的频率分量。

6.1.2 K-means 聚类算法构建

基于上述构造的特征集，采用 K-means 算法对 2206 条序列进行聚类，基于肘部法则和轮廓系数评估聚类结果，最优 k 值的寻优过程如图 6.3 所示。可以看出，当 $k=14$ 左右时，轮廓系数开始收敛，簇内平方和达到了适中的量级。算法执行流程如表 6.1 所示。

表 6.1 K-means 聚类算法流程

算法 1：K-means 聚类算法

输入：类别数量 k，最大迭代次数 $t_{\max}$，质心变化阈值 δ

选择 k 个随机样本作为初始质心

while $i<i_{\max}$ or 质心变化量$<\delta$

　　分配样本到最近的质心，并计算每个聚类的平均值

　　更新质心，并计算当前质心与更新后质心的差距

　　$i=i+1$

end while

输出：聚类结果

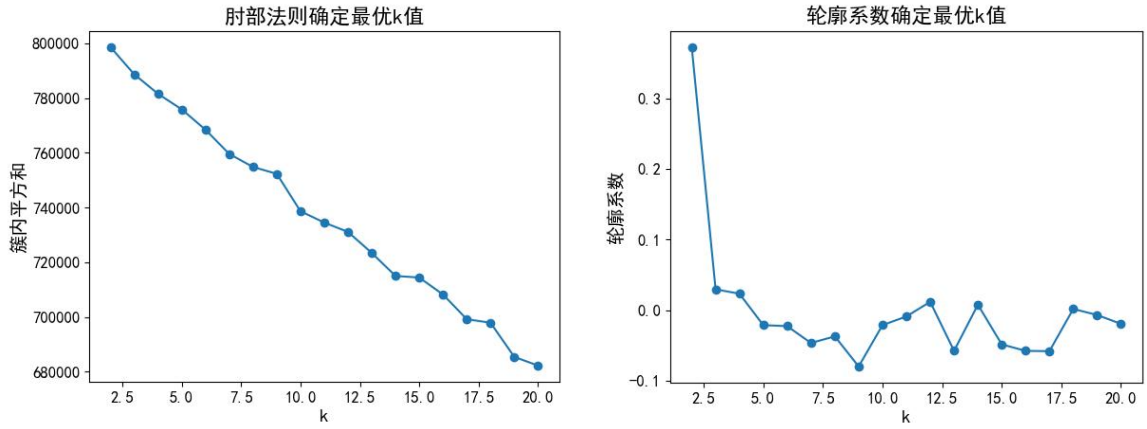


图 6.3 肘部法则和轮廓系数选择最优的聚类数量

6.1.3 聚类结果评估

基于 K-mans 算法，采用章节 6.1.1 中所述的几类特征指标对附件 1 和附件 5 中的 2206 条实例进行聚合。将聚类后的结果使用主成分分析（PCA）降维并可视化聚类结果，结果如图 6.4 和 6.5 所示。从图中可以看出，本文构建的聚类方案可以较好地挖掘出序列中类分离的显著特征。各簇边界明显且簇类中心点相互远离，使同一类别在需求上的特征最为相似。

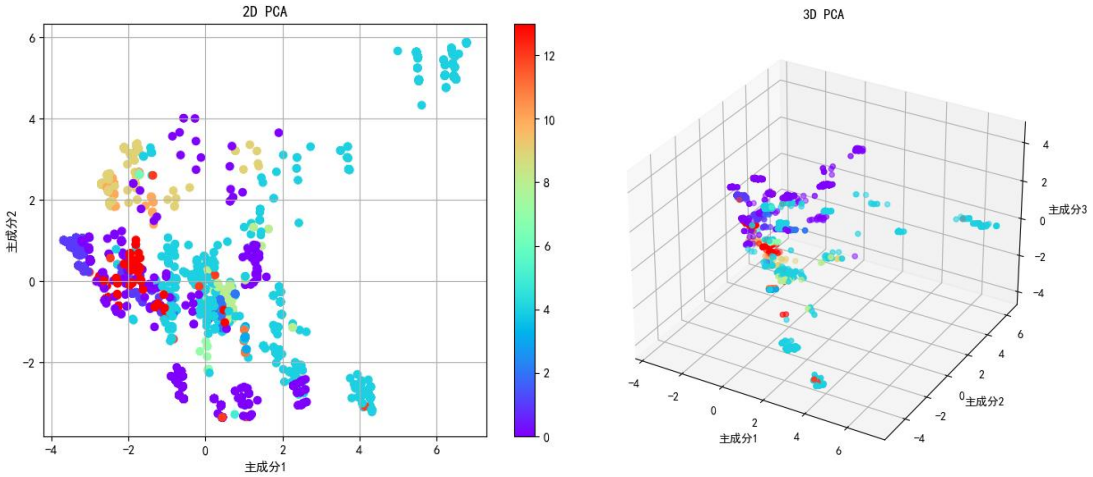
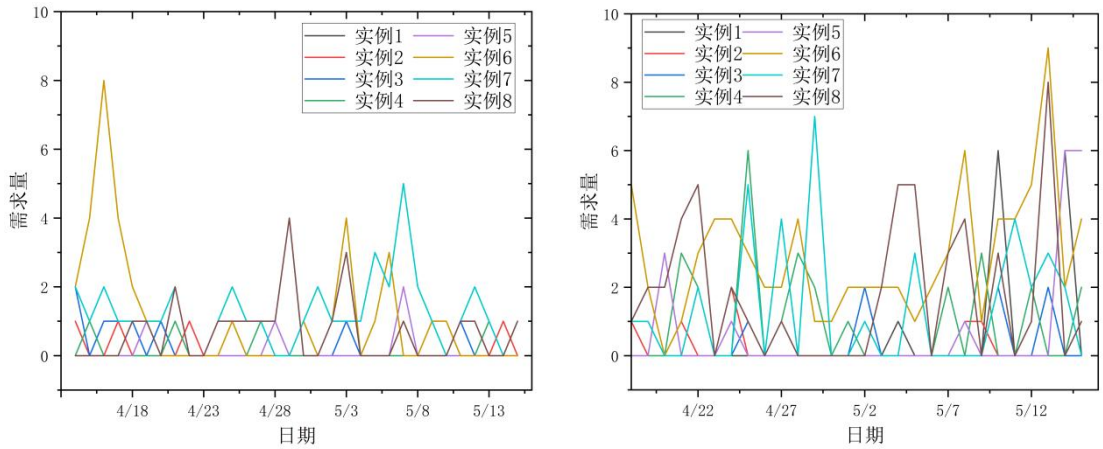


图 6.4 K-means 聚类后的可视化分类图



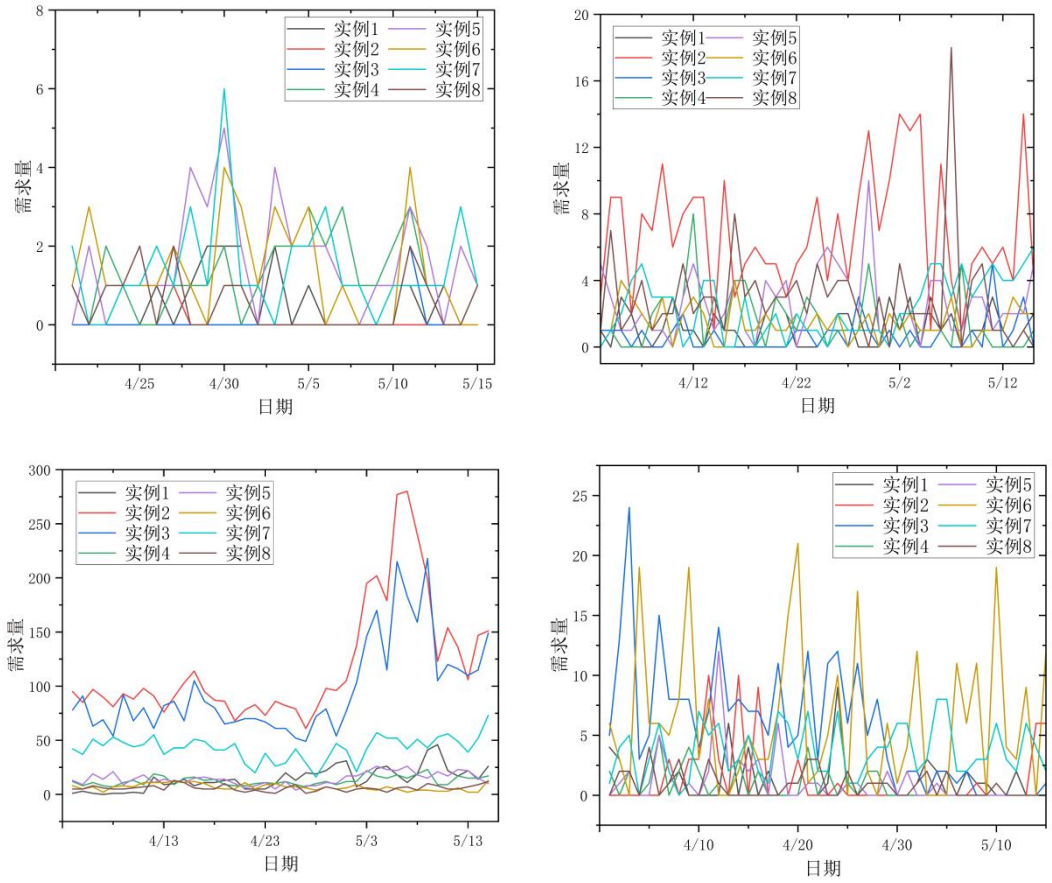


图 6.5 分类标准 1-6 下序列的分布情况

6.2 多模型集成预测算法

在获得聚类标准后，对 2206 条序列进行分类。不同于问题一，问题二不对各类别中的序列数据进行加和，避免各样例的个性化特征被淹没。采用滑动窗口对聚类后各类别中的序列依次进行特征提取和样本集构建。比如，若类别 1 中具有样本量 112 条，每个序列提取 20 个特征，则构造的数据集维度为 112*20。然后，对于各类别分别采用 SVM、XGBoost、LightGBM、LSTM、KNN 算法训练预测模型（各算法下均训练 14 个模型），基于 5 折交叉验证法调整模型参数并评估算法性能。在预测阶段，首先对待预测实例进行分类，基于该类别中 5 个算法模型得到 5 个不同的预测值，并将各模型预测值进行集成得到最终的预测输出，以类别 1 中的序列预测任务为例，实施流程如图 6.6 所示。

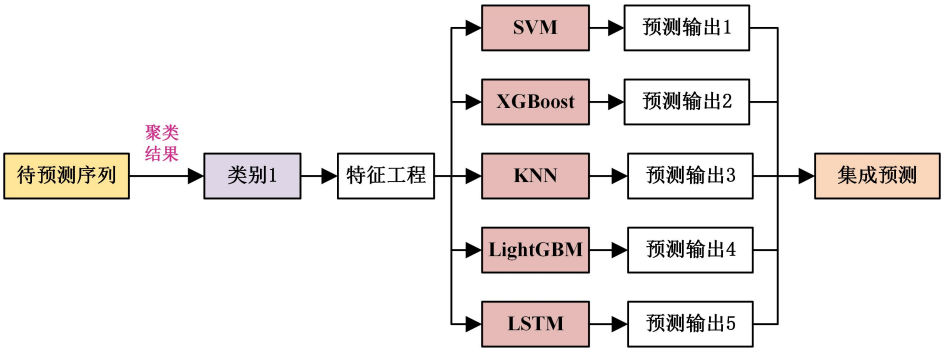


图 6.6 类别 1 中需求量序列预测任务流程

6.2.1 预测特征提取

合理的特征工程会大幅提高模型的效果。根据聚类结果匹配附件 1 和附件 5 中 2022/12/01-2023/5/15 期间的需求量序列，设置滑动窗口长度为 15，截取长度为 15 的序列作为输入样本，该序列后一天的需求量为输出标签，通过滑动窗口构建最终的训练集合。在聚类阶段，由于附件 1 和附件 5 中不同实例的需求量序列长度不一致，我们采取了等维分割取 2023/4/23-2023/5/15 期间的数据进行分类。而在预测模型构建阶段，为了充分地利用数据量，采用滑动窗口构建切割序列集合，克服了不同序列长度不一致的问题。考虑到预测模型强大的挖掘能力，采用基于滑动窗口构造的数据集直接训练 SVM、XGBoost、LSTM 和 KNN 预测模型。

针对基于决策树的 LightGBM 算法，需要进一步挖掘数理特征以保证预测任务的实施。为了捕获需求量序列的周期性、季节性、滞后性等特征，同时考虑在实际商品销售时，相邻日期出货量的强相关性，对于每个切割序列，构建如下的特征：各切割序列倒序 1 天到倒序 7 天的需求量之和；倒序 7 天、14 天的需求量均值和标准差；切割序列的最大值和最小值；切割序列的趋势性强度；周期性/趋势性强度，共计 6 个特征。

6.2.2 预测算法构建

针对聚类后的 14 个类别，分别采用 SVM、XGBoost、LightGBM、LSTM、KNN 算法训练预测模型。其中，XGBoost、LightGBM、LSTM 算法原理与构建流程与问题一相似，具体见章节 5.2。SVM 和 KNN 算法的原理和执行流程如下：

1）基于 SVM 的预测模型构建

支持向量机（SVM）的基本思想是将数据映射到高维空间中，并在该空间中找到一个超平面，使得各类数据点到该超平面的距离最大。具体来说，对于给定的训练数据集，SVM 会通过计算每个样本点与超平面之间的距离，进而确定最佳的决策边界。为了避免过拟合和提高泛化性能，SVM 引入了核函数，可以将线性不可分的数据映射到高维空间，从而实现非线性分类。训练任务可以表示为：

$$\min \ \frac{1}{2}\|W\|^2, \quad \text{s.t.} \ \ y_i(W^T x_i + b) \geq 1, \ i = 1, \ 2, \ ..., \ N \tag{6-2}$$

式中，$y_i(\bullet)$ 为分类超平面。采用构建的切割序列集优化上述问题得到最终的预测模型。

2）基于 KNN 的预测模型构建

K 近邻算法是一种常用的回归方法。基本思想是对于给定的样本，在训练数据集中寻找与它最近的 K 个邻居，通过 K 个邻居的信息来预测这个样本的数值。训练任务可以表示为：

$$\begin{gathered} y = \arg\max_{c_j} \sum_{x_i \in N_k(x)=1} I(y_i = c_j) \\ i = 1, \ 2, \ ..., \ N; \ j = 1, \ 2, \ ..., \ K \end{gathered} \tag{6-3}$$

式中，x_i 为实例的特征向量，y_i 为实例的标签，$I(\bullet)$ 为指示函数。采用构建的切割序列集来训练 KNN 模型，计算样本与训练数据的相似度，筛选最近邻的 K 个样本，采用 K 个邻居标签的平均值作为序列的预测值，进而实现预期的预测任务。

6.3 任务预测与性能评价

基于附件 1 和 5 中商品-商家-仓库在 2022/12/01-2023/5/15 期间形成的需求量序列和图 6.1 的框架，分别采用 SVM、XGBoost、LightGBM、LSTM、KNN 算法训练预测模型，各算法的拟合能力和在验证集上的表现如图 6.7 所示。从图中可以看出，提出的各算法均可以挖掘出时序数据中的抽象特征，实现对出货量的预测。集成后的预测值克服了单模型泛化能力弱的问题，有效地平滑了各模型的缺陷，可以更好地逼近真实值。

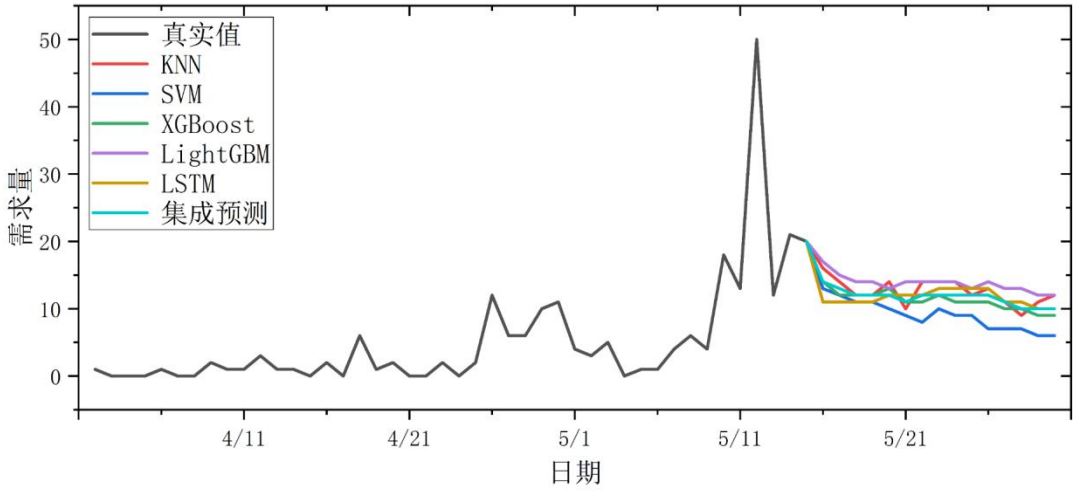


图 6.7 某实例在各算法下以及集成后的预测输出与真实结果

针对各类算法，采用 5 折交叉验证法进行模型训练与性能评估。选取 1-wmape、MAPE 和 RMSE 指标评价各算法优劣，结果如表 6.2 所示。从表中看出，由于各类别下序列分布的数理特征存在差异，同种算法在不同类别下展现了互异的预测性能。比如，在指标 1-WAMAPE 下，KNN 模型在类别 3、7、8、13 中获得了最好评价，SVM 模型在类别 1、4、6 中获得了最好评价，XGBoost 在类别 12 上获得最好评价，LightGBM 在类别 2、5 上获得了最好评价，LSTM 在类别 9、10、11、14 上获得了最好评价。因此，在实际预测任务中，电商平台可以在给定指标下对于每个类别选取最优的模型进行预测。基于上述分析，采用图 6.1 中的算法架构对附件 5 中各商品在 2023/05/16-2023/05/30 期间的需求量进行预测，相应结果已填写至附件中的表 2。

表 6.2 不同算法下预测性能评价

五折交叉验证平均指标	1-WMAPE					MAPE				
	KNN	SVM	XGBoost	LightGBM	LSTM	KNN	SVM	XGBoost	LightGBM	LSTM
类别 1	0.3691	**0.3801**	0.3618	0.3744	0.3686	743.0695	**603.0286**	833.8921	762.4272	844.0827
类别 2	0.5762	0.5845	0.5895	**0.5914**	0.5887	349.2767	**306.6765**	424.9912	367.4352	429.6076
类别 3	**0.5306**	0.4916	0.5532	0.5040	0.5028	334.8504	394.5991	**324.5619**	312.2125	470.2088
类别 4	0.5318	**0.5414**	0.5278	0.4778	0.5562	249.0347	251.4863	268.2551	272.1805	**241.0364**
类别 5	0.6240	0.5945	0.6365	**0.6441**	0.6431	293.1223	**263.2717**	387.4426	303.1533	331.3930

类别 6	0.7026	**0.7240**	0.7184	0.6865	0.6587	39.7623	38.0004	**37.7473**	42.3758	35.4796
类别 7	**0.5556**	0.2051	0.4583	0.5231	0.3566	**1516.7200**	3029.2272	2089.5253	1738.4156	1981.6913
类别 8	**0.3643**	0.3522	0.3693	0.2311	0.3648	284.9975	**219.1434**	335.5463	365.6069	342.4850
类别 9	0.5138	0.4997	0.5225	0.5419	**0.5625**	296.7115	**277.3866**	342.2849	308.6810	359.6722
10	0.4858	0.4757	0.4958	0.5019	**0.5316**	555.7322	**460.9982**	619.9490	547.2010	649.9891
11	0.3735	0.3687	0.3763	0.3693	**0.3992**	487.2293	**398.3599**	578.7195	520.9615	526.2898
12	0.4245	0.3960	**0.4556**	0.4370	0.4297	388.0059	**341.6599**	395.3457	399.0321	360.2626
13	**0.5435**	0.5229	0.5285	0.5212	0.4779	1536.7208	1971.2279	1934.4179	**1513.0256**	2403.5921
14	0.5213	0.3865	0.5551	0.5346	**0.5986**	322.8405	**290.1643**	360.4213	322.9552	305.4850

评价指标		类别 1	类别 2	类别 3	类别 4	类别 5	6	7	8	9	10	11	12	13	14
RMSE	KNN	11.2279	8.0715	**7.7664**	6.2921	8.1224	13.1281	325.2324	4.6855	14.8003	27.3384	5.8995	5.1284	**83.0691**	11.1947
	SVM	11.9593	8.3930	8.9299	6.4535	11.1154	**12.5144**	419.3711	5.0289	16.0401	28.5039	6.6965	5.7457	86.6652	18.3486
	XGBoost	11.3804	7.6395	7.3462	6.1802	7.7060	12.6448	**319.2013**	4.5505	14.9167	26.2668	5.8070	4.8282	86.3768	10.0020
	LightGBM	11.5449	7.8339	8.2468	6.7877	7.6340	13.8704	326.2910	5.1463	13.5937	25.8895	5.9375	4.9541	88.0488	10.6524
	LSTM	**11.0880**	**7.6021**	7.8458	**5.9545**	**7.3311**	15.8262	433.8531	**4.5657**	**10.0356**	**17.4091**	**5.5168**	**4.7861**	89.7018	**7.7281**

7 问题三模型建立与求解

7.1 问题分析与模型建立

问题三主要是考虑在出现规律性大型促销活动时，对实例在 2023/06/01-2023/06/20 期间的需求量进行预测。根据实际分析，将 2023/06/01-2023/06/20 期间活动峰值确定在 2023/06/18（京东 618）。大型促销活动期间，商品的需求量将出现以下特点：1）序列无序性增强：在促销活动期间，商品出货量通常会经历销售峰值，出货量显著增加，商品的库存快速减少，导致序列的规律性较常规期间大幅降低；2）样本稀少：规律性的促销活动只在一段时间内定期出现，相应的历史数据较为稀少，样本的数理特征难以挖掘和关联。

针对上述问题，首先根据附件 1 中 1996 件商品在 2022/12/01-2022/12/11、2023/01/01-2023/01/11 、 2023/02/01-2023/02/11 、 2023/03/01-2023/03/11 、2023/04/01-2023/04/11 这 5 个月 1 日至 11 日的需求量，计算各商品每日均值，得到附件 1 中每件商品在 1 日至 11 日的需求量均值，作为该商品在非促销期间的每日销售量。则第 i 个商品在第 j 日的非促销期间每日平均销售量 $y_{\mathrm{cm}\_(i,j)}$ 的计算公式为：

$$y_{\mathrm{cm}\_(i,j)}=\frac{1}{5}\sum_{m=1}^{5}y_{\mathrm{cm}\_(m,i,j)} \tag{7-1}$$

式中，$i=1,\ 2,\ ...,\ 1996$，$j=1,\ 2,\ ...,\ 11$。$y_{\mathrm{cm}\_(m,i,j)}$ 为第 i 个商品在第 m 月第 j 日的销售

量。

然后将附件 6 中 1996 件商品每件商品在 2022/11/01-2022/11/11 这 11 天的销售量作为该商品在促销期间的每日销售量 $y_{\text{act}\_(i,j)}$。随后计算每个商品在 1 日至 11 日的促销期间每日销售量与非促销期间每日销售量的比值，得到每一件商品在 1 日至 11 日的促销力度值 $P_{(i,j)}$，计算公式为：

$$P_{(i,j)}=\frac{y_{\text{act}\_(i,j)}}{y_{\text{cm}\_(i,j)}},\quad i=1,\ 2,\ ...,1996;\ j=1,\ 2,\ ...,\ 11 \tag{7-2}$$

问题三需要预测的促销期间每日销售量等于模型基于非促销期间每日销售量的预测值与促销力度值的乘积，需要计算出各商品每日对应的促销力度值。由于问题三需要预测出附件 1 中 1996 件商品每个商品在 2023/06/01-2023/06/20 期间 20 天的商品需求量，不能与根据附件 6 计算得到的每一件商品在 1 日至 11 日的 11 个促销力度值一一对应，因此将根据附件 6 计算得到的每一件商品在 1 日至 11 日的促销力度值划分为 3 组，分别为 1 日-5 日、6 日-10 日、11 日。计算每组的促销力度值均值，计算方法如下：

$$P_{(i)}^{1}=\frac{1}{5}\sum_{j=1}^{5}P_{(i,j)},\ P_{(i)}^{2}=\frac{1}{5}\sum_{j=6}^{10}P_{(i,j)},\ P_{(i)}^{3}=P_{(i,10)} \tag{7-3}$$

同时将问题三需要预测的 1996 件商品在 2023/06/01-2023/06/20 期间 20 天的商品需求量分为 4 组，分别为 2023/06/01-2023/06/09、2023/06/10-2023/06/17、2023/06/18、2023/06/19-2023/06/20。根据附件 6 计算得到的每一件商品在 1 日至 11 日的 3 组促销力度值，与问题三需要预测的 1996 件商品在 2023/06/01-2023/06/20 期间 20 天的 4 组商品需求量，其对应关系为：将这 3 组促销力度值的均值分别作为 2023/06/01-2023/06/09、2023/06/10-2023/06/17、2023/06/18 这 3 个时间段的促销力度值，第 4 组（2023/06/19-2023/06/20）的促销力度值根据节后的商业经验判断其应该为 1，即认为其与非促销期间该日每日销售量一致，整体流程如图 7.1 所示。

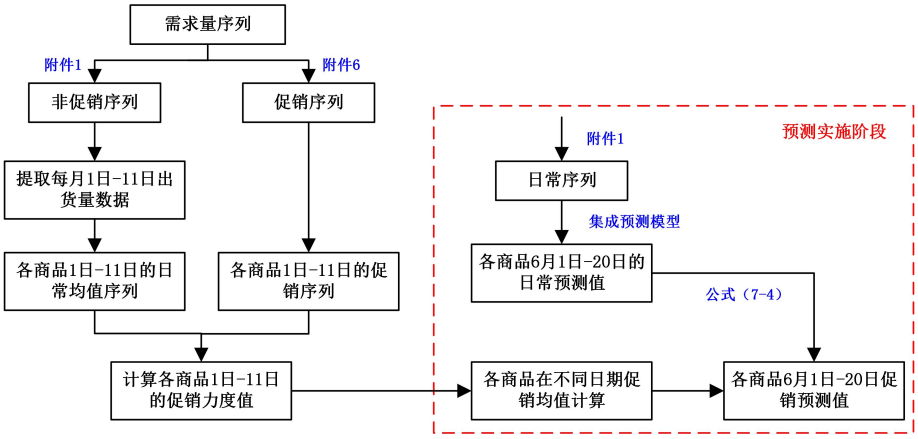


图 7.1 问题 3 框架的训练与实施流程

问题三需要预测的促销期间每日销售量等于模型基于非促销期间每日销售量的预测值与促销力度值的乘积。$y_{\text{pre\_cm}\_(i,j)}$ 表示问题一中模型预测的第 i 个商品在第 j 日的非促销期间每日销售量，则第 i 个商品在第 j 日的促销期间每日销售量 $y_{\text{pre\_act}\_(i,j)}$ 可以通过

下式求解：

$$\begin{cases} y_{\text{pre\_act}\_(i,j)} = y_{\text{pre\_cm}\_(i,j)} \bullet P_{(i)}^{1}, j = 1,\ 2,\ ...,\ 9 \\ y_{\text{pre\_act}\_(i,j)} = y_{\text{pre\_cm}\_(i,j)} \bullet P_{(i)}^{2}, j = 10,\ 11,\ ...,\ 17 \\ y_{\text{pre\_act}\_(i,j)} = y_{\text{pre\_cm}\_(i,j)} \bullet P_{(i)}^{3}, j = 18 \\ y_{\text{pre\_act}\_(i,j)} = y_{\text{pre\_cm}\_(i,j)}, j = 19,\ 20 \end{cases} \tag{7-4}$$

7.2 预测结果分析

以某一件商品为例，图 7.2 展示了促销力度值作用前后的效果。其中蓝色曲线代表未考虑促销力度值时模型预测出的常规期间商品销量，黄色曲线代表对模型基础预测施加促销因子之后的销量。根据该图可以看出，本文提出的促销事件影响模型对商品的促销期间销量进行了优化，使其更符合促销期间的实际情况。

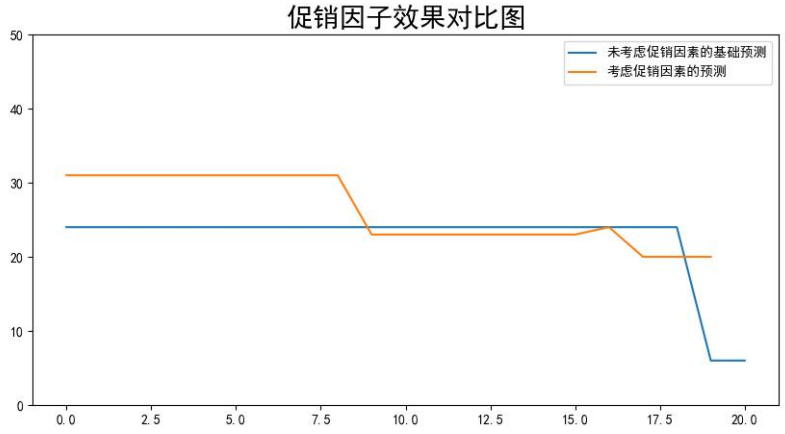


图 7.2 促销因子效果对比图

基于上述分析，采用图 7.1 中的算法架构对附件 6 中各商品在 2023/6/01-2023/06/20 期间的需求量进行预测，相应结果已填写至附件中的表 3。

8 模型评价与推广

8.1 模型的优点

1、针对电商供应链预测任务的特点，设计预测任务的层次结构来实现更好的预测，充分挖掘时序数据特征实现精准预测，**模型具有一定的可解释性**。

2、对数据集进行充分挖掘，建立特征工程，同时考虑了数据的数理特征和解释变量，**对数据集的处理比较完整**。

3、引入**集成学习思想**，综合考虑多模型预测效果，减少极端数据带来的误差，降低个例的随机性对总体预测的影响，多方面学习数据中内在联系，结合多个模型优良方面，避免陷入局部最优。

4、对机器学习模型采用了**交叉验证**的方式，更好地利用数据集，减少数据浪费，提高模型的泛化能力，验证模型的稳健性，**防止过拟合**情况的发生。

5、**模型具有可拓展性**，可以引入更多性能优良的时序预测模型或考虑更多解释变

量，来支撑电商供应链更精准的预测任务。

8.2 模型的缺点

1、特征工程的数理特征提取仍有改进空间，后续可以对特征工程更好的处理。

2、受限于时间，各时序预测模型参数可以进一步寻优。

8.3 模型的推广

精准的需求预测对于电商商家降低成本、保证履约等具有重要意义。本文采用的层次时序预测思想对于稀缺数据、新维度数据等的预测能够起到良好的效果，综合考虑数理特征与外生变量的融合，有利于充分挖掘相似数据模式，指导同一类别数据的预测。另外，本文引入了多种机器学习时序预测模型，基于大数据驱动的方法能够极大方便当今对复杂数据的处理。本文的模型不仅仅能够在供应链优化当中提供整体和局部的需求预测，也有潜力应用于能耗预测、金融市场风险管理、公共健康领域等，利用不同层级的时间序列间的依赖关系和模式进行精准预测，实现良好的效益。

参考文献

[1] 杨博, 罗嘉琛, 宋艳涛, 等. 基于对比学习的时间序列聚类方法[J/OL]. 计算机科学: 1-13 [2023-11-03].

[2] 韩立, 李芳. 基于动态特征选择和长短期记忆模型的气温预报方法[J]. 计算机时代, 2023(10): 148-156.

[3] Adnan R M, Parmar K S, Heddam S, et al. Suspended Sediment Modeling Using a Heuristic Regression Method Hybridized with Kmeans Clustering[J]. Sustainability, 2021, 13. DOI:10.3390/su13094648.

[4] Makridakis S, Spiliotis E, Assimakopoulos V. The M5 Accuracy competition: Results, findings and conclusions[J]. International Journal of Forecasting, 2020, 36(1): 224-227.

[5] Makridakis S, Spiliotis E, Assimakopoulos V, et al. The M4 Competition: 100,000 time series and 61 forecasting methods[J]. International Journal of Forecasting, 2020, 36(1): 54-74.

[6] Cai R, Xie S, Wang B, et al. Wind speed forecasting based on extreme gradient boosting[J]. IEEE Access, 2020, 8: 175063-175069.

[7] 王鹏新, 王颖, 田惠仁, 等. 基于 LightGBM 的冬小麦产量估测及可解释性研究[J/OL]. 农业机械学报: 1-16 [2023-11-03].

[8] 韩克宁, 王伟, 樊冬艳, 等. 基于产量递减与 LSTM 耦合的常压页岩气井产量预测[J]. 油气藏评价与开发, 2023, 13(05): 647-656.

[9] 荣新, 覃卫坚, 韦文山. 基于 3 种机器学习算法的台风频数预测[J]. 海洋预报, 2023, 40(05): 1-9.

[10] 李明宇. 大型商超白酒产品促销效能评价方法研究[J]. 中国酿造, 2023, 42(09): 246-252.

队伍编号	MCB2200703
赛道	（B）

基于决策树模型对北京移动用户体验影响因素研究

摘要

随着移动通信技术的飞速发展，网络覆盖越来越完善，各个移动运营商越来越重视用户的网络使用与体验，从而进一步提升网络服务质量。

本文则是对北京移动用户体验影响因素的探究，通过分析影响用户体验的主要影响因素，并以此来为决策提供依据，进而提升客户的上网体验。本文基于附件中提供的打分项以及影响打分的各个因素建立了对于客户语音业务和上网业务的数学模型，最终得出量化分析结果，并对附件 3，附件 4 中的客户打分进行了预测。具体本文解决了以下问题。

针对问题一：首先进行数据清洗，根据附件五的描述字段对附件一二数据进行转换，针对特征缺失值的类型分别采用 K 近邻算法（KNN）以及均值填补将特征缺失值补全。之后进行特征工程，对部分类别特征数据进行标签化和独热编码，将附件一“用户描述”字段和附件二“用户备注”字段中的文本进行合并分词，并通过 tf(词频)绘制词云图量化分析，对影响用户打分的主要因素类型进行初步判断。为了对数据集有一个整体把握，我们对处理好的数据做了描述性统计进行分析，并给出了分析结果。之后，引入了 CART 决策树模型针对每项打分提取所有特征的重要性程度，并根据特征的重要性得出了影响客户语音业务满意度的主要因素为 GPRS 总流量（KB）、前 3 月 MOU、当月 ARPU、是否遇到过网络问题，影响客户上网业务满意度的主要因素为重定向次数、年龄、当月 GPRS 资源使用量（GB）、网络信号差/没有信号。然后给出了各因素对客户打分影响程度的量化分析和结果。

针对问题二：为了结合相关因素对用户打分进行一个更好的预测，需要对问题一处理好的数据再次进行特征工程，采用 PCA 主成分降维和特征重构的方式重新构造数据集，基于 CART 决策树模型设定阈值针对每项打分数据进行特征选取，并结合附件三和附件四中已有的特征筛选出主要特征，同时对附件 3 和附件 4 部分缺失特征进行众数填补和特征转换使其格式调整为与训练集相同的格式。之后划分数据集并进行模型训练，首先针对决策树的泛化性能进行后剪枝优化。随后引入了集成学习模型随机森林和 XGBoost，通过网格搜索和五折交叉验证的方式对 CART 决策树模型、随机森林、XGBoost 模型进行超参数优化，最后采取软投票融合机制对上述三个模型进行加权融合，得到最终预测模型，数据带入模型得到了最终用户打分预测结果，并对结果的合理性做出解释说明。

关键词：特征工程，CART 决策树，PCA 主成分降维，随机森林，XGBoost 模型，模型融合

目录

一、问题重述..... 1
1.1 问题背景..... 1
1.2 问题提出..... 1
二、问题分析..... 1
2.1 针对问题一的分析..... 1
2.2 针对问题二的分析..... 1
三、模型假设..... 2
四、符号说明..... 2
五、问题一模型的建立与求解..... 3
5.1 数据清洗..... 3
5.2 特征工程..... 6
5.3 基于决策树模型的特征重要性提取..... 7
5.3.1 特征选择..... 7
5.3.2 构建 CART 分类树模型..... 7
5.3.3 特征重要性..... 8
5.3.4 整体满意度打分主要因素分析..... 9
5.3.5 其余打分影响程度的量化分析..... 10
六、问题而模型的建立与求解..... 12
6.1 模型的选取和数据的优化..... 12
6.1.1 选取目标值和特征值..... 12
6.1.2 代价复杂度 CPP 决策树后剪枝..... 15
6.2 随机森林模型..... 18
6.3 XGBoost 模型..... 19
6.4 随机森林和 XGBoost 模型的超参数优化..... 20
6.5 模型融合..... 23
6.6 基于决策树模型的预测及结果合理性..... 24
七、模型评价..... 24
八、模型推广..... 25
参考文献..... 25
附录..... 26

一、问题重述

1.1.问题背景

随着移动通信技术的飞速发展，运营商们都非常重视客户的网络使用体验，从而进一步提升网络服务质量。在信息透明，产品同质化的今天，客户满意度的表现成为各大运营商市场运营状况的重要体现，客户满意才能推动商业决策的实行，从而进一步推动移动网络的高质量可持续发展。由于创痛的方法已经难以有效的提升客户的满意度，这次拟通过分析影响用户满意度的各种因素，从而更快、更全面的提升用户满意度。

1.2.问题提出

基于以上背景，要求我们通过数据分析与建模的方法帮助中国移动北京公司解决以下问题：

1.分别给出影响客户语音业务和上网业务满意度的主要因素，并且给出各个因素对于客户的满意度影响程度的量化分析和结果。

2.通过问题一的分析，创建客户语音业务和上网业务客户打分的相关影响因素的数学模型，并对附件 3，附件 4 中的客户打分进行预测。

二、问题分析

2.1 针对问题一的分析

首先对附件 1 和附件 2 的数据进行数据清洗，根据附件 5 的字段描述，将原始数据进行转换，针对特征缺失值的类型分别采用 K 近邻算法（KNN）以及均值填补将特征缺失值补全，然后进行特征工程，利用分类特征编码和 tf(词频)对数据集特征进行了量化。我们还对处理后的数据进行了描述性统计，以便方便后续的研究。之后建立 CART 决策树模型，对数据集进行分类并计算每个特征的重要性，最终得出影响客户打分项目的主要因素及量化分析结果。

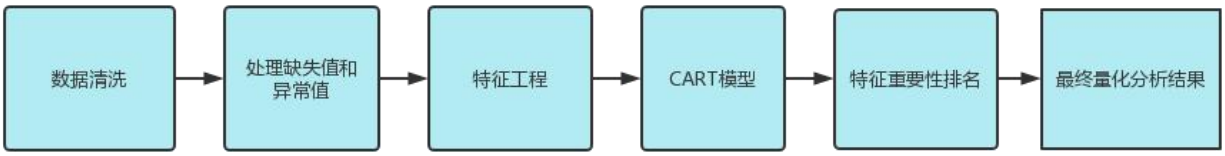


图 1.问题一思路流程图

2.2 针对问题二的分析

首先对数据集再进行特征工程，利用 PCA 主成分降维和特征重构重新选取，选取出重要性较大的特征。之后利用选取的特征建立 CART 决策树，随机森林，XGBoost 模型，并以 7：3 的比例划分训练集与测试集，然后利用网格搜索和五折交叉验证对模型调参寻取最优参数，之后对模型进行软投票加权法进行融合，从而达到进一步稳定预测的目的。随后，我们对附件 3，附件 4 中的数据进行了预处理，使之成为与训练集格式一样的数据，最终得到了预测结果，并解释了预测结果的合理性。

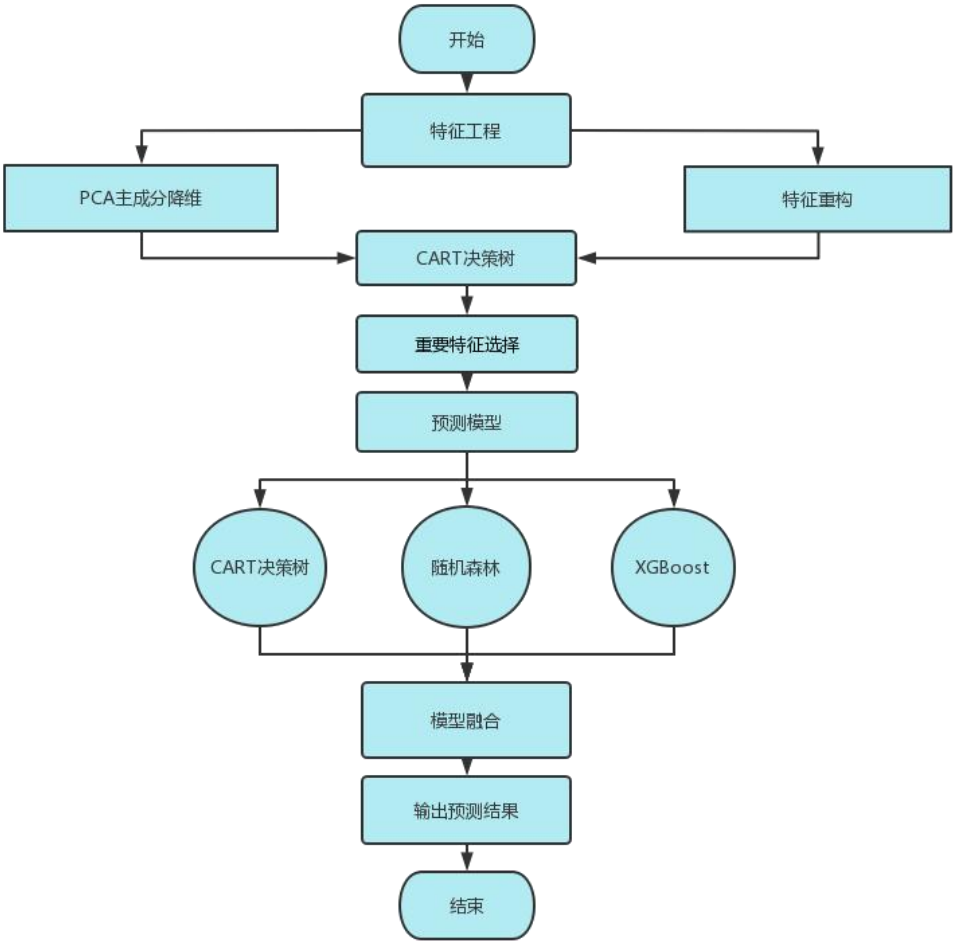


图 2 问题二思路流程图

三、模型假设

1. 假设题目中所给出的数据真实有效。
2. 假设客户对语音业务和上网业务的满意度不受除附件给出的其它因素影响。
3. 假设调查群体基本具有相同的特征。

四、符号说明

下表列出了一部分符号，文中出现的其它符号在相应位置给予解释

符号	符号说明
I	特征重要性
$Gini$	基尼系数
$Gini-index$	基尼指数
result	表示在分类时属于某类的概率

五、问题一　模型的建立与求解

5.1 数据清洗

附件 1、附件 2 给出的原始数据集中，介于给出的数据量非常庞大，信息复杂，涉及到的影响因素众多等，可能会导致原始数据集出现缺失值、异常值等现象，影响本文数据挖掘及分析的结果，因此本文先对原始数据集进行初步数据清洗。

1.对数据集进行初步转换

根据附件 5 给出的附件 1、附件 2 的字段说明，对数据进行一系列处理，根据描述字段的分析，对数据集进行以下转换，转换方式见**图 3**：

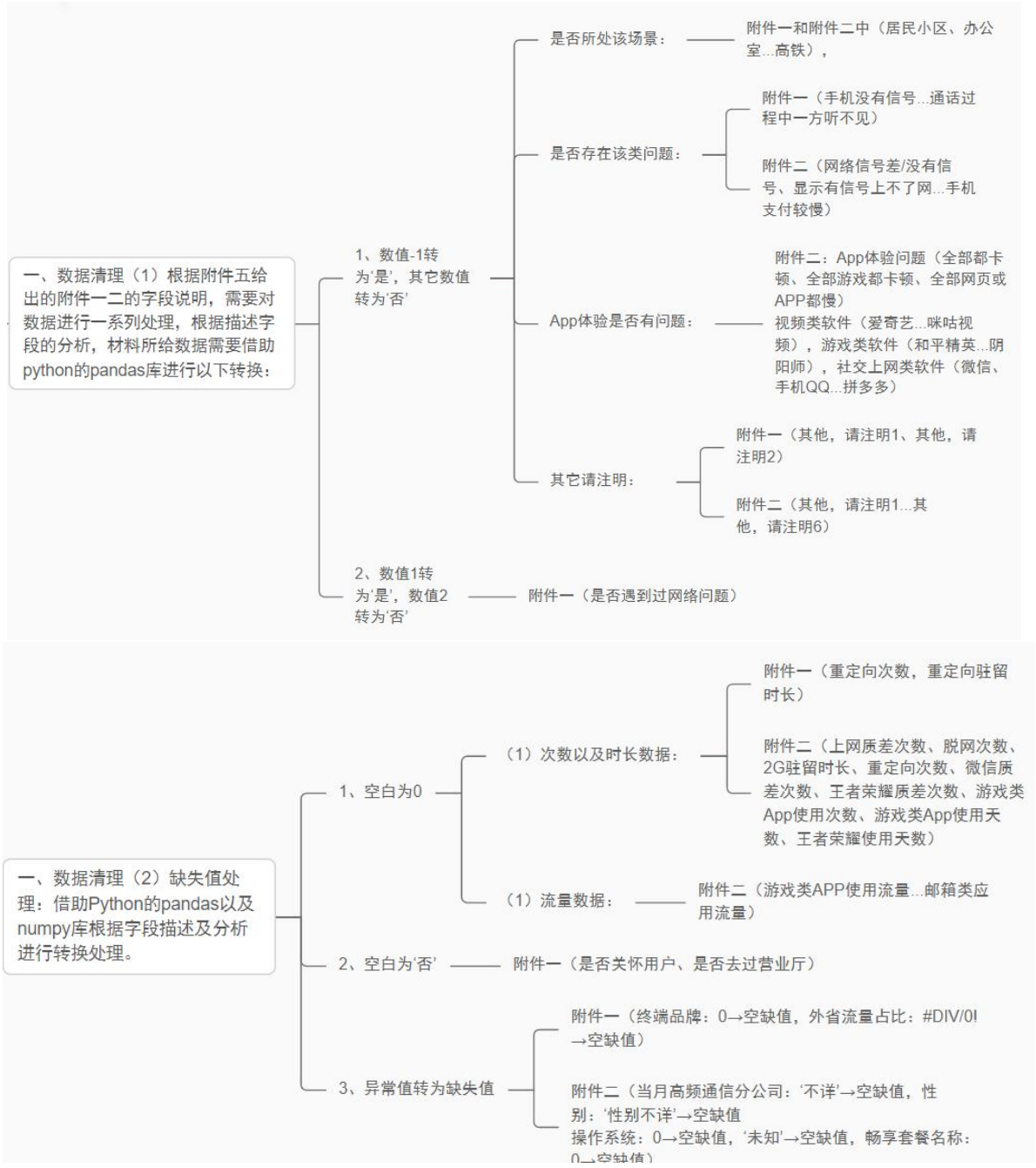


图 3 数据集转换详情图

经过以上处理之后，对附件 1 和附件 2 的数据列缺失值进行统计分析，结果见**图 4** 和**图 5**：

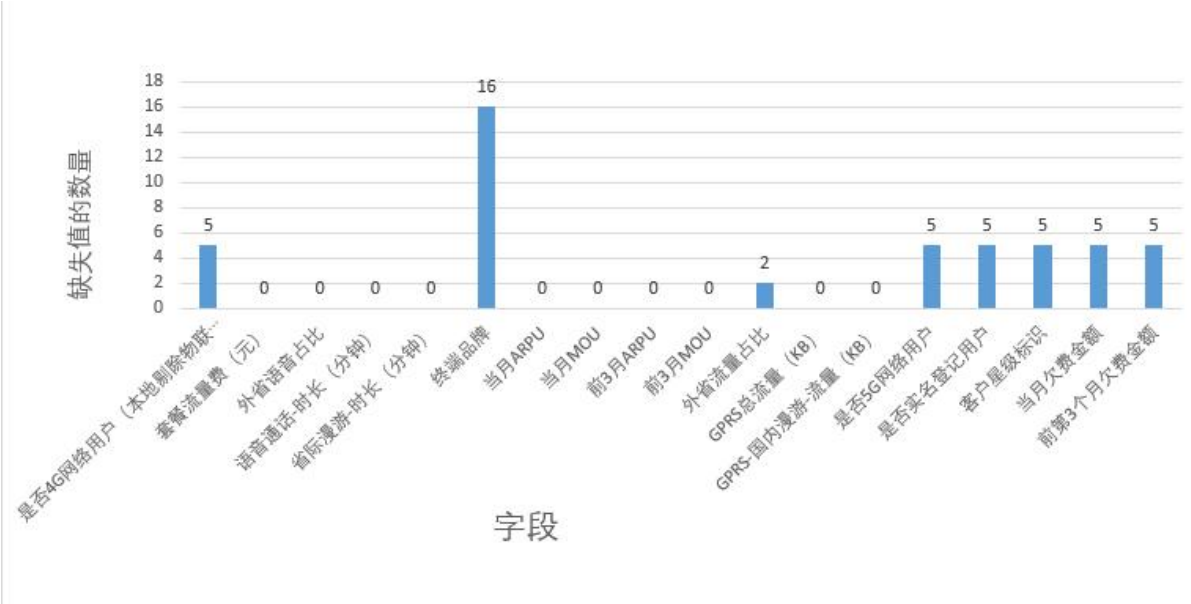


图 4 附件 1 缺失值统计

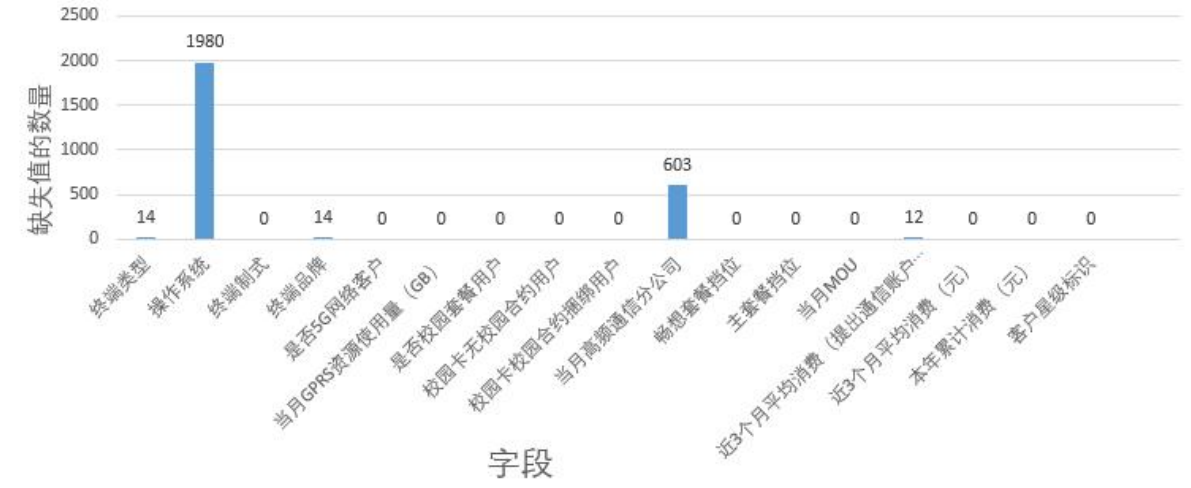


图 5 附件 2 缺失值统计

2. 删除缺失值过量的特征

数据集中“畅享套餐名称”、“码号资源-激活时间”、“码号资源-发卡时间”特征缺失值较多，超过 50%，缺失值占比见**图 6**：

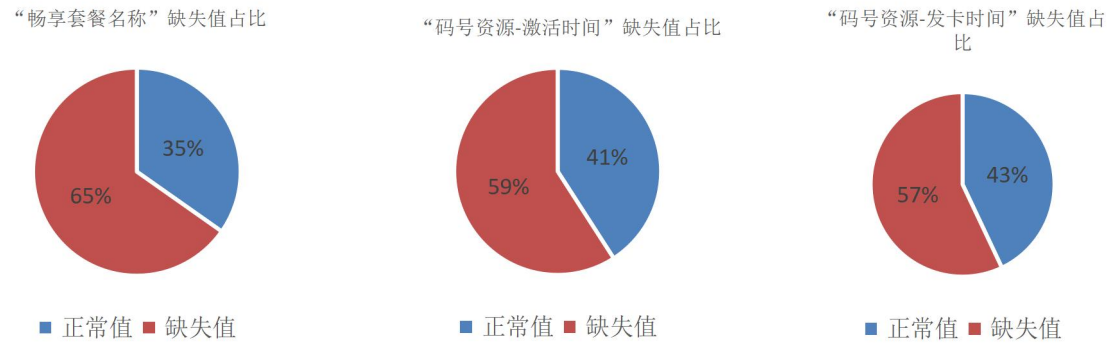


图 6　畅享套餐名称、码号资源-激活时间、码号资源-发卡时间特征缺失值占比

上述特征由于缺失值占比过高，强行填补会造成较大的偏差，因此将该类特征删除。

3. 缺失值填补

（1）连续型数据

附件 1 中“外省流量占比”、“当月欠费金额”、“前三个月欠费金额”

附件 2 中“近三个月平均消费（剔除通信账户支付）”均为连续型数据，采用平均值填补其缺失值。

（1）类别型缺失值

通过观察数据集，附件 1 中“是否 4G 网络客户（本地剔除物联网）”、“终端品牌”、“是否 5G 网络客户”、“是否实名登记用户”、“客户星级标识”和附件 2 中“终端类型”、“终端品牌类型”、“操作系统”、“终端品牌”、“当月高频通信分公司”、“性别”均为类别变量，因此本文采用 K 近邻（KNN）分类算法预测填补。

KNN 算法的主要思想就是待预测样本点与其它样本点距离进行计算，然后保存并排序，选出距离最近的 K 个样本点，若这 K 个样本大多数属于类别 M，则将该待预测样本归为 M。具体步骤如下：

1）选取特征

附件 1 中选取“语音通话整体满意度”、“语音通话清晰度”、“语音通话稳定性”、“网络覆盖与信号强度”这 4 个字段作为特征变量。

附件 2 中选取“手机上网整体满意度”、“网络覆盖与信号强度”、“手机上网速度”、“手机上网稳定性”这 4 个字段作为特征变量。

2）计算欧式距离

本文利用欧式距离公式计算待预测样本点和其它样本点的距离，即预测样本点与其他样本点的相似程度。本文假设待测样本点 A 的坐标为(A_1, A_2, A_3, A_4)，其它样本点 B 的坐标为(B_1, B_2, B_3, B_4)，那么 A、B 两点的欧式距离为：

$$|AB| = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + (A_3 - B_3)^2 + (A_4 - B_4)^2}$$

3）根据 K 值判定类别

计算得到待预测样本与所有样本的欧式距离后，根据距离的远近对其它样本进行升序排序。最终选取前 K 个其它样本并得到其中最多的类别（本文选取的 K 值为 9），实现待预测样本的归类，完成类别型缺失值的填补。

4.数据的整体分析

在完成上述数据预处理工作后，为了对数据有一个整体的把握，对处理好的数据和附件 3，附件 4 中的数据进行描述性统计，分别计算了样本量、最大值、最小值等统计量来研究定量数

据的基本情况。根据统计结果发现绝大部分数据的变异系数（CV）都大于 0.15，在这些数据基本上都存在异常值，即题目中给出的真实数据中每一项都存在异常值，这是由于打分数据太过于主观而导致的。在相同的因素下，不同的人的满意状况不相同，因此会造成很大的打分差距。**（总体描述结果见附录 1）**

5.2. 特征工程

完成数据清洗工作后，接下来是对数据集的特征工程。

1. tf（词频）

tf 是用以评估一字词对于一个文件的重要程度的一种统计方法。通过观察数据集，我们发现附件 1 中的“用户描述”和附件 2 中的场景备注数据、现象备注数据、APP 大类备注、APP 小类视频备注、APP 小类游戏备注、APP 小类上网备注等文本数据是主观评判数据，因此决定通 tf 来对这些数据进行处理。本文通过 Python 对每个用户的描述数据进行了合并，将所有词语合成在一个文件中，并通过 jieba 库进行分词，然后计算每一个词所对应的 tf 值。

tf 公式如下：

$$tf = \frac{n_i}{\sum_k n_k}$$

其中 n_i 表示所有用户评语中某词出现的次数，$\sum_k n_k$ 表示所有评语的总词数。

右图是根据**附件一和附件二**用户描述和备注字段词频绘制的词云图，可以看到，影响用户语音业务满意度和手机上网满意度的主要因素类型是网络信号和所处的地点。



图 7　词云图

2. 分类特征编码

（1）LabelEncoder 编码

本文中针对二分类特征对数据进行 LabelEncoder 编码，转换方式见**图 8**：

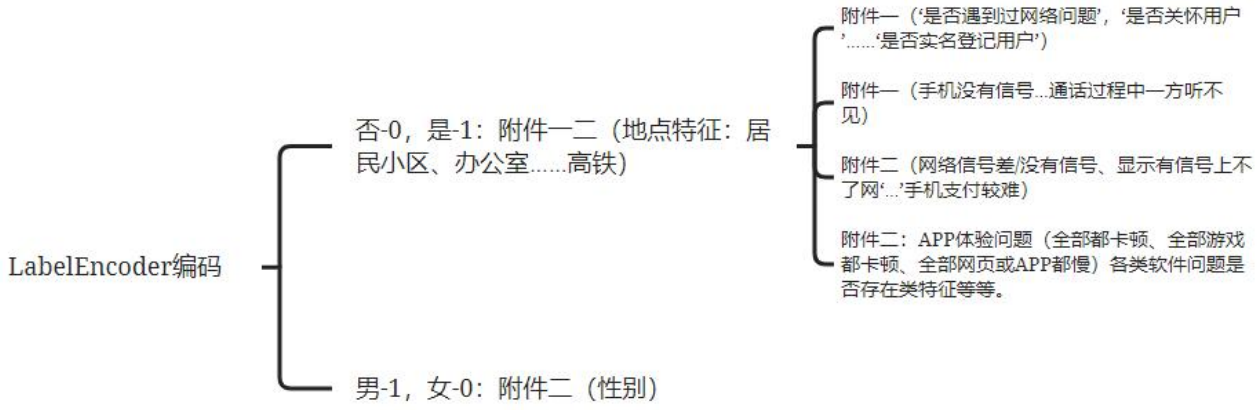


图 8　LabelEncoder 编码详情图

（2）独热编码（One-Hot Encoding）：

针对多分类特征（大于三个类别），并且类别之间不存在大小比较意义，因此，本文对多分类特征（大于三个类别）进行独热编码。

以附件 2“终端制式”样本为例：

终端制式	终端制式=4G终端	终端制式=5G终端	终端制式=2/3G终端
4G终端	1	0	0
5G终端	0	1	0
2/3G终端	0	0	1

图 9　“终端制式”样本独热编码

由于附件 1、附件 2 中的“终端品牌类型”种类过多（共 631 项），用独热编码会造成维度暴涨，所以对类别特征编码时排除该特征。

5.3 基于决策树模型的特征重要性提取

决策树（Decision Tree）是一种自上而下，对样本数据进行树形分类的过程。决策树由结点和有向边组成。每个内部结点表示一个特征或属性，叶节点表示类别。样本从从顶部的根节点开始划分，样本被分到不同的子节点中。再根据子节的特征进行进一步划分，直至所有样本都被归到某一叶结点中。

本文采用的是一种常见的决策树构造算法 CART 分类树。

5.3.1. 特征选择

CART 分类树中是利用基尼指数（Gini index）作为节点选取标准点的特征进一步划分，直至所有样本都被归到某一叶节点中。

设样本集合 D 中有 K 类样本，样本集合 D 中第 K 类样本所占的比例为 $p_k=(k=1,2,...,|K|)$，那么数据集 D 的基尼值定义为：

$$\mathrm{Gini}(D)=1-\sum_{k=1}^{|K|}p_k^2$$

基尼值反映的是两个样本类别标记不一致的概率，数据集的基尼值越小，纯度越高。

设离散属性 a 有 V 个可能的取值 $\{a_1,a_2,...,a_v\}$，用 a 来对样本集合进行划分会产生 V 个分支节点，其中，第 v 个分支节点包含了 D 中所有在属性 a 上取值为 V 的样本，记为 D_v。属性 D 的基尼指数为：

$$Gini-index(D,a)=\sum_{v=1}^{V}\frac{|D_v|}{|D|}Gini(D_v)$$

在构建决策树过程中，选取使划分后基尼指数最小的属性作为最优的划分属性，即 $a'=\arg\min_{a\in A} Gini-index(D,a)$

5.3.2. 构建 CART 分类树模型

在构建树的过程中，选择最优属性作为测试属性，将测试属性取值相同的样本划分到同一个样本子集中；如果该样本集类别中只有一个属性，则分支为叶子节点，否则对该样本集采用递归调用算法。模型具体建立过程见图 10：

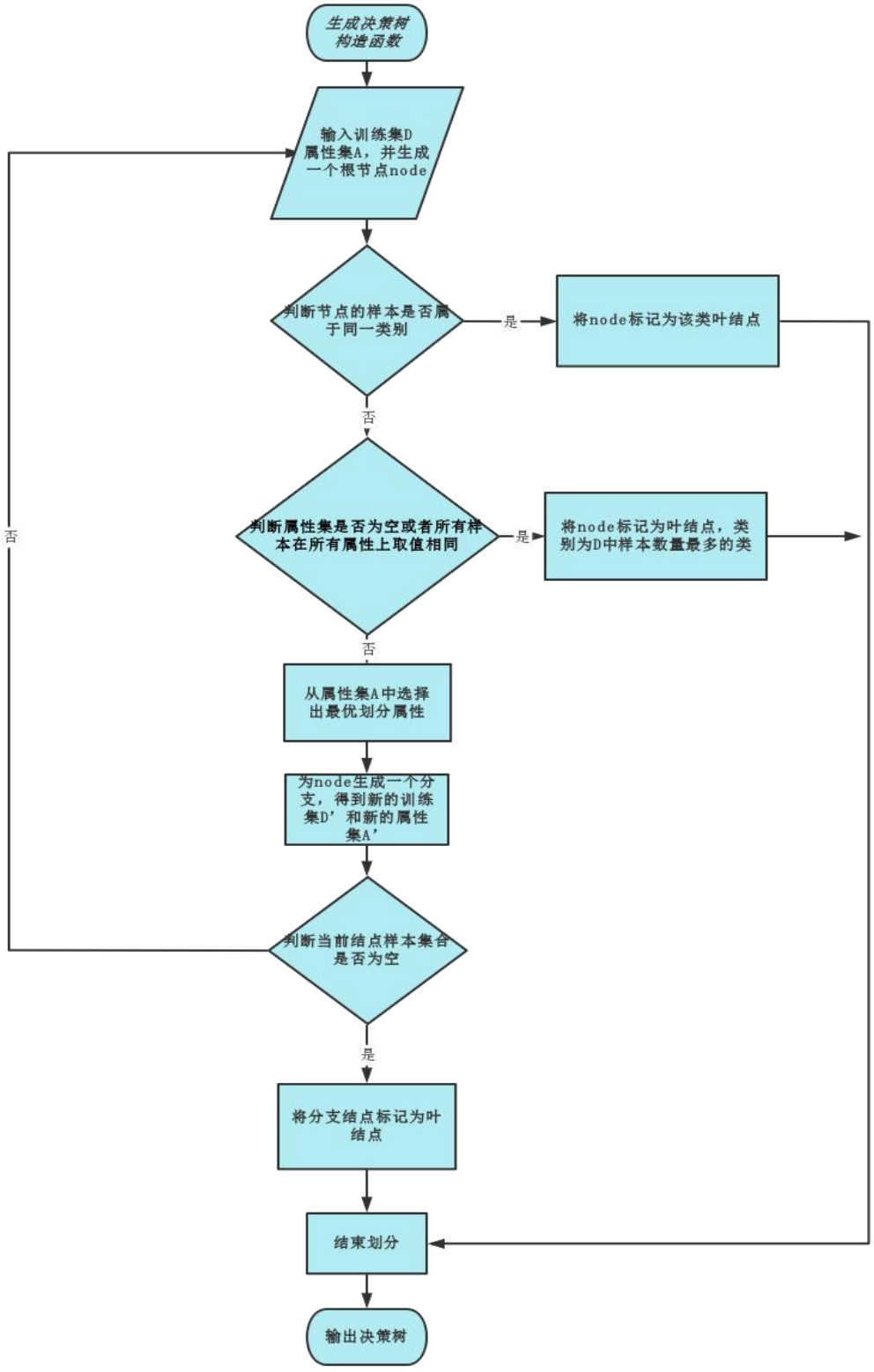


图 10　CART 模型构建具体流程

5.3.3.特征重要性

为了研究影响客户语音业务和上网业务满意度的主要因素，需要对决策树中的特征计算出特征的重要性。特征重要性越高，表示该特征对其满意度的影响越大。本文决定用特征重要性来评估对影响客户语音业务和上网业务满意度的主要因素。

CART 模型构造好后，对每一个特征计算其重要性，对每个特征计算它的基尼系数，特征重要性为基尼系数减少量的归一化值。

设在该数据集中的样本总数为 N， N_t 是决策树中某一结点的样本数目， L_t 是在这一结点处左孩子的样本数目，R_t 是在这一结点处右孩子的样本数目，该处结点的基尼系数为 $Gini(D)$，左孩子的基尼系数为 $Gini(L)$，右孩子的基尼系数为 $Gini(R)$；那么该结点处所对应的特征重要性为：

$$I = \frac{N_t}{N}\left(Gini(D) - \frac{R_t}{N_t}Gini(R) - \frac{L_t}{N_t}Gini(L) \right)$$

最后根据每个特征的特征重要性进行降序排序，并将数据汇总在 csv 文件，存放在支撑材料中。

5.3.4. 整体满意度打分主要因素分析

对于语音通话整体满意度和手机上网整体满意度的两个打分项目采用 5.3.1 中建立的模型，采用 5.3.3 中的公式计算出每个特征的重要性。由于篇幅有限，仅展示出特征重要性排名前十名结果，结果见**表 1** 和**表 2**：

表 1　语音通话整体满意度特征重要性前十名

GPRS 总流量（KB）	0.125644
前 3 月 MOU	0.105648
当月 ARPU	0.102283
是否遇到过网络问题	0.090247
当月 MOU	0.06138
语音通话时长（分钟）	0.058187
Mos 质差次数	0.049293
脱网次数	0.0352
未接通掉话次数	0.028615
重定向驻留时长	0.021474

从上表可以看到影响语音通话整体满意度的主要因素为:“GPRS 总流量(KB)”“前 3 月 MOU”、“当月 ARPU”、“是否遇到过网络问题”。

表 2　手机上网整体满意度特征重要性前十名

重定向次数	0.08170442
年龄	0.051158538
当月 GPRS 资源使用量（GB）	0.046827913
网络信号差/没有信号	0.039974137
通信类应用流量	0.033789474
当月 MOU	0.032047617
网页类应用流量	0.029131373
优酷视频使用流量	0.028134721
小视频系 APP 流量	0.026828634
阿里系 APP 流量	0.026439612

从上表可以看到影响手机上网整体满意度的主要因素为：“重定向次数”、“年龄”、“当月 GPRS 资源使用量（GB）”、“网络信号差/没有信号”。

5.3.5 其余打分影响程度的量化分析

由于附件 1 和附件 2 中两个满意度各存在三个打分项目，依然采用 5.3.1 中建立的模型，并采用 5.3.3 中的公式计算特征重要性，由于篇幅限制，仅展示特征重要性排名前十结果，结果见**表 3-表 8。（剩余特征的量化分析结果见附录 2）**

表 3　附件 1 网络覆盖与信号强度特征重要性前十名

GPRS 总流量（KB）	0.12798
当月 ARPU	0.10791
前 3 月 MOU	0.09952
是否遇到过网络问题	0.09703
Mos 质差次数	0.06102
当月 MOU	0.05995
语音通话-时长（分钟）	0.05946
未接通掉话次数	0.03023
脱网次数	0.02915
重定向驻留时长	0.02000

表 4　附件 1 语音通话清晰度特征重要性前十名

GPRS 总流量（KB）	0.11892
当月 ARPU	0.11012
前 3 月 MOU	0.09129
是否遇到过网络问题	0.09222
Mos 质差次数	0.06423
当月 MOU	0.05820
语音通话-时长（分钟）	0.05752
终端品牌-苹果	0.01882
脱网次数	0.03945
重定向驻留时长	0.02247

表 5　附件 1 语音通话稳定性特征重要性前十名

GPRS 总流量（KB）	0.11073
是否遇到过网络问题	0.10525
前 3 月 MOU	0.10456
当月 ARPU	0.086006
语音通话-时长（分钟）	0.06882
当月 MOU	0.06078
Mos 质差次数	0.05986
脱网次数	0.03810
未接通掉话次数	0.02198
重定向驻留时长	0.01767

表 6　附件 2 网络覆盖与信号强度特征重要性前十名

GPRS 总流量（KB）	0.12798
当月 ARPU	0.1079
前 3 月 MOU	0.0995
是否遇到过网络问题	0.0973
Mos 质差次数	0.06102
当月 MOU	0.0599
语音通话-时长（分钟）	0.05946
未接通掉话次数	0.0302
脱网次数	0.0291
重定向驻留时长	0.02000

表 7　附件 2 手机上网速度特征重要性前十名

重定向次数	0.07989
年龄	0.04123
通信类应用流量	0.03499
今日头条使用流量	0.0349
腾讯视频使用流量	0.03368
当月 MOU	0.03353
手机上网速度慢	0.03130
当月 GPRS 资源使用量	0.03062
上网过程中网络时断时续或时快时慢	0.02940
视频类应用流量	0.02817

表 8　附件 2 手机上网稳定性特征重要性前十名

重定向次数	0.07784
年龄	0.04307
当月 MOU	0.04271
通信类应用流量	0.0397
当月 GPRS 资源使用量（GB）	0.0373
上网过程中网络时断时续或时快时慢	0.03725
网络信号差/没有信号	0.03694
网页类应用流量	0.02841
今日头条使用流量	0.02712
本年累计消费（元）	0.02581

六、问题二　模型的建立与求解

在本问题中将要建立基于决策树模型的预测模型，并且引入集成学习模型，为了提高模型的预测效果，首先对问题一中已处理好的数据进一步优化处理，对降维后的数据进行特征重构和特征选择，同时对决策树模型进行剪枝以提高其泛化能力，随后建立了随机森林模型及 XGBoost 分类模型，对选取的数据集进行拟合，同时通过网格搜索和五折交叉验证对模型超参数进行优化，最后将三种模型通过软投票加权法进行融合，进一步提高模型预测的准确率。

6.1 模型的选取和数据的优化

在上述决策树模型中，由于对多分类特征进行了独热编码，使得原数据集维度扩张，在这种情况下，建立的树模型会出现过拟合的风险，降低模型泛化性能。因此，为了提高模型的预测效果，对整体数据集进行了特征降维重构，接着基于重构的特征带入 CART 决策分类树模型设定阈值对特征进行选择选取用来预测的特征值。

6.1.1 选取目标值和特征值

下文中模型的目标值分别为 8 个打分项目，即“语音通话整体满意度”、“网络覆盖与信号强度”、“语音通话清晰度”、“语音通话稳定性”、“手机上网整体满意度”、“网络覆盖与信号强度”、“手机上网速度”、“手机上网稳定性”。

接下来是对特征值的选取，在选取之前，需要对数据集进一步加工处理，本文采用 PCA 主成分降维及特征合并两种方式进行特征重构。

（1）PCA 主成分降维

主成分分析(Principal Component Analysis，PCA）是一种无监督学习的多元统计分析方法，主要思想是将 n 维特征映射到 k 维上，这 k 维特征被称为主成分，是在原有 n 维特征的基础上重新构造出来的 k 维特征，本文将降维的数据原有的 n 维特征全部降至一维。过程如下：

假设该类别的数据集中有 n 条 k 维数据，将数据按列组成 k 行 n 列矩阵 k

1）对数据样本中心化，即将矩阵 k 中样本数据均值为零，方差为一；

$$\sum_i x_i = 0$$

2）计算出协方差矩阵 $C = \frac{1}{m} XX^T$ 即：

$$C = \begin{pmatrix} cov(x_1,x_1) & cov(x_1,x_2) & \dots & cov(x_1,x_m) \\ cov(x_2,x_1) & cov(x_2,x_2) & \cdots & cov(x_2,x_m) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_m,x_1) & cov(x_m,x_2) & \cdots & cov(x_m,x_m) \end{pmatrix}$$

其中 $cov(x,y)$ 为两个样本变量之间的协方差，$\bar{x}$ 、$\bar{y}$ 为两个样本变量的均值，即：

$$cov(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

4）求出协方差矩阵的特征值及对应的特征向量，并将求出的特征向量按对应特征值大小排序，取第一行组成矩阵 P，最终求得 $Y=PX$ 即为数据降到一维后的数据。

对数据集中一部分数据进行 PCA 主成分降维，详情见图 11。

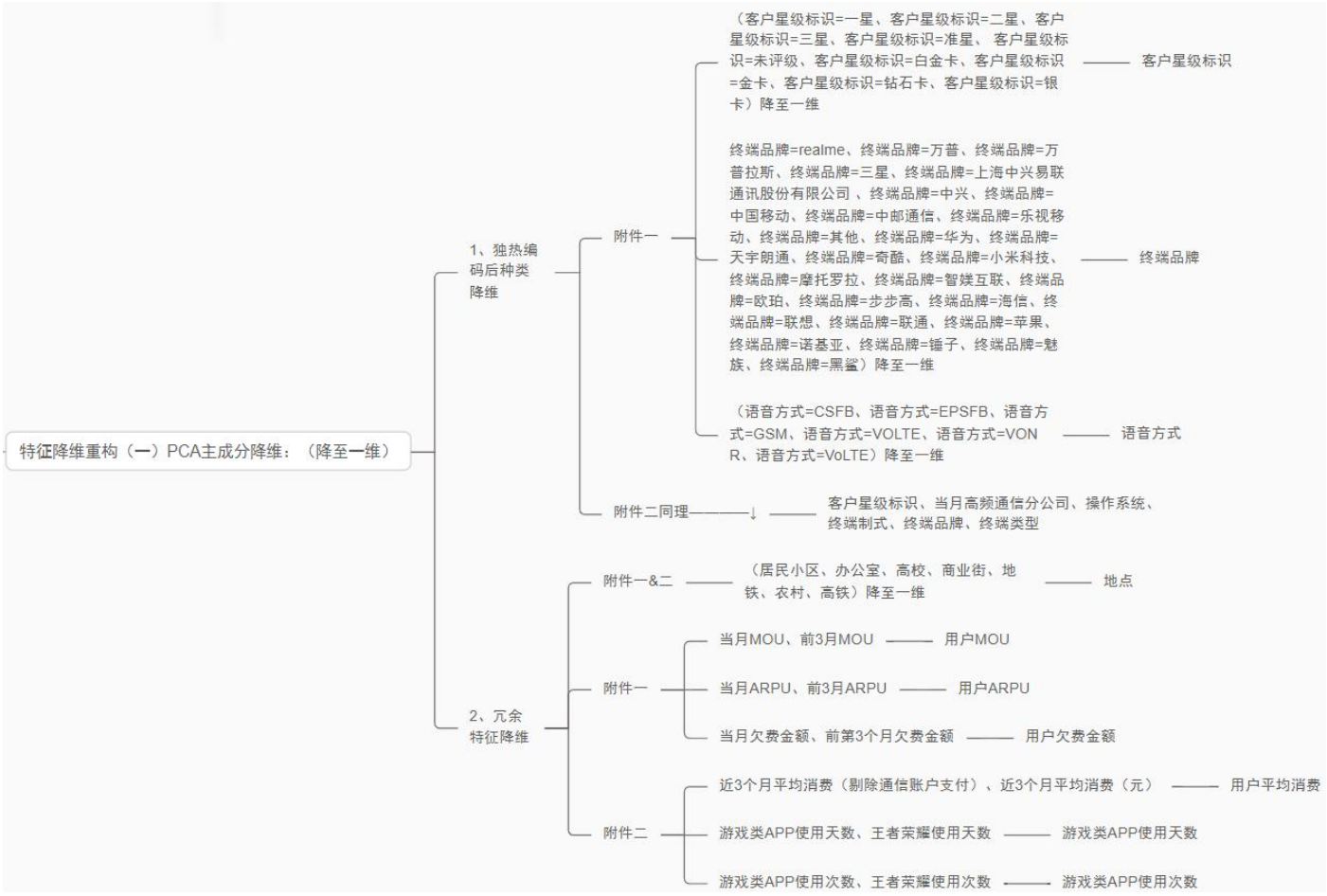


图 11　PCA 主成分降维详细情况

（2）特征合并

根据描述字段将一些相似特征进行相加合并归为一类特征。例如对附件 1 中（手机没有信号、有信号无法拨通、通话过程中突然中断、通话中有杂音、听不清、断断续续、串线、通话过程中一方听不见）这些特征进行相加变为新的特征“网络问题出现总个数”。

对数据集特征合并的详细处理方式见图 12：

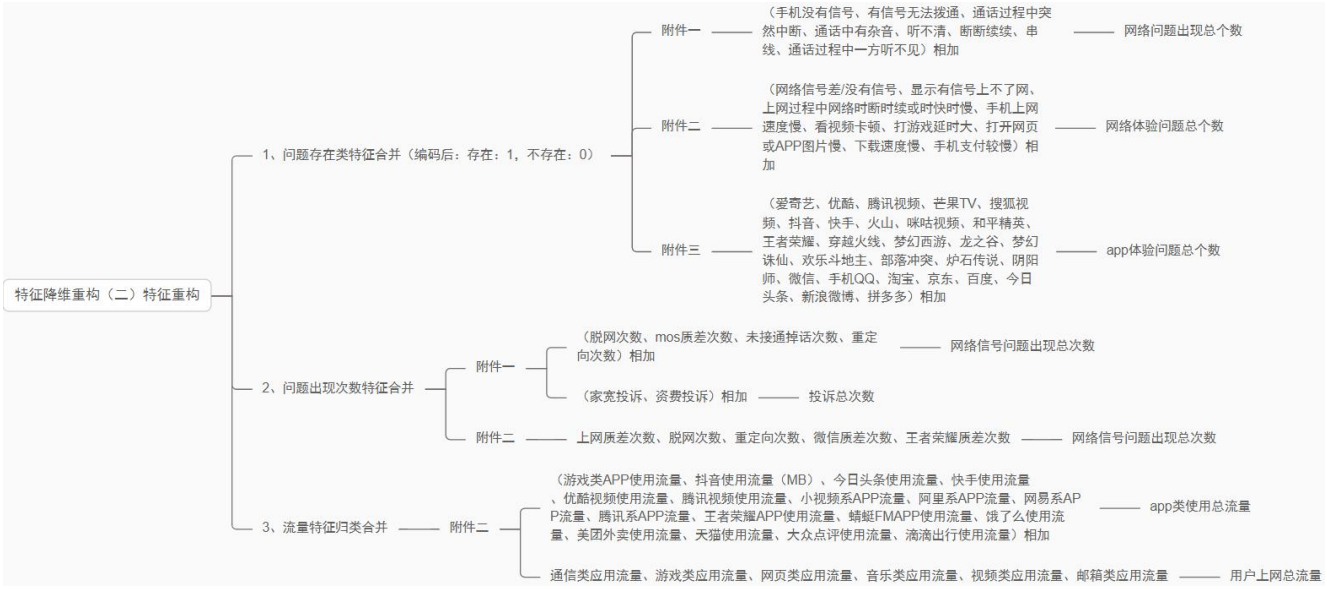


图 12　特征合并详情

（3）基于决策树模型的重要特征选择

1）附件 1 重要特征选择

将重构后的特征数据引入 CART 分类树模型中，得到了新的特征重要性结果（**附件 1 和附件 2 中特征重要性结果见附录 3）。**通过 sklearn 库中 SelectFromModel 方法设定特征选择的阈值为特征重要性均值进行特征选取，最终选择出特征重要性大于均值的特征作为用于模型训练的特征值，最终选取的特征见**表 9：**

表 9　附件 1 语音业务用户满意度特征选择结果

GPRS 总流量（KB）
用户 MOU
用户 ARPU
语音通话-时长（分钟）
网络信号问题出现总次数
是否遇到过网络问题
地点
网络问题出现总个数

2）附件 2 重要特征选择

采用与附件 1 重要特征选择一样的方法，先选取特征值大于均值的特征，考虑到附件 2 给出的数据中的部分特征在附件 4 中并未出现，附件 4 中给出的数据集与附件 2 中给出的数据集相差过大，为了使附件 2 与附件 4 数据集统一为相同的格式，排除了进行特征融合后在附件 2 中存在，但在附件 4 中缺失过多的部分特征，之后再选取特征重要性排名前十的特征，选取结果见表 10.

表 10　附件 2 上网业务用户满意度特征选择结果

网络体验问题总个数
当月 MOU
当月 GPRS 资源使用量（GB）
地点
网络信号问题出现总次数
性别
app 体验问题总个数
全部网页或 APP 都慢
全部游戏都卡顿
全部都卡顿

3）对附件 3 数据的处理

在预测之前，将附件 3 的数据处理成与附件 1 数据集相同的格式：将“重定向次数”采用众数进行填补，剩余数据采用与 5.1，5.2 中所描述的方式处理。然后选取和**表 9** 一样的特征作为预测时的特征。

4）对附件 4 数据的处理

在预测之前，将附件 4 的数据处理成与附件 2 中训练集相同的格式：将缺失特征“王者荣耀质差次数”、“重定向次数”根据附件二数据进行众数填补，附件四的‘GPRS 总流量（KB）’需要进行单位转换转转换为附件二中的‘当月 GPRS 资源使用量（GB）’剩余数据采用与 5.1，5.2 中所描述的方式处理。然后选取和**表 10** 一样的特征作为预测时的特征。

（4）划分数据集

借助 sklearn 库中 model_selection.train_test_split 方法设定随机种子和比例，可以很好地将原数据集中的目标值和特征值按 7：3 的比例划分为训练集和测试集用于后续模型的训练。

6.1.2 代价复杂度 CPP 决策树后剪枝（Cost Complexity Pruning）

为了进一步提高决策树的泛化能力，降低决策树的过度生长造成过拟合的风险，对决策树进行后剪枝处理。

后剪枝是先从训练集生成一棵完整决策树后，对树进行剪枝简化处理。CCP 代价复杂度则是通过下式损失函数的最小化为目标对决策树进行剪枝：

$$L=\sum_{i=1}^{T}\frac{N_i}{N}L_i+\alpha T$$

其中 T 是树中叶子结点的个数，N 是样本总个数，N_i 是第 i 个叶子结点上的样本数量，L_i 是第 i 个叶子结点的损失函数。

L_i 是待定系数，L_i 越大，表示叶节点的个数对损失函数的影响越大，剪枝后的决策树就更容易选择复杂度较小的树，即 T 较小。L_i 越小，表示叶节点的个数对损失函数的影响越小，此时剪枝后的决策树就更容易选择负责都较大的树，即 T 较大。因此，L_i 控制了预测误差与树的复杂度对剪枝的影响。

为了确定最优参数 L_i 的取值，首先通过绘制学习曲线选出较优的取值范围，采用 Python 编程，将测试样本与训练样本代入模型，绘制不同 L_i 取值下决策树在训练样本和测试样本上的精确度。以附件 1“语音通话整体满意度”和附件 2“手机上网整体满意度”两项打分项目为例，参数与模型训练集和测试集精度关系见**图 13** 和**图 14**。

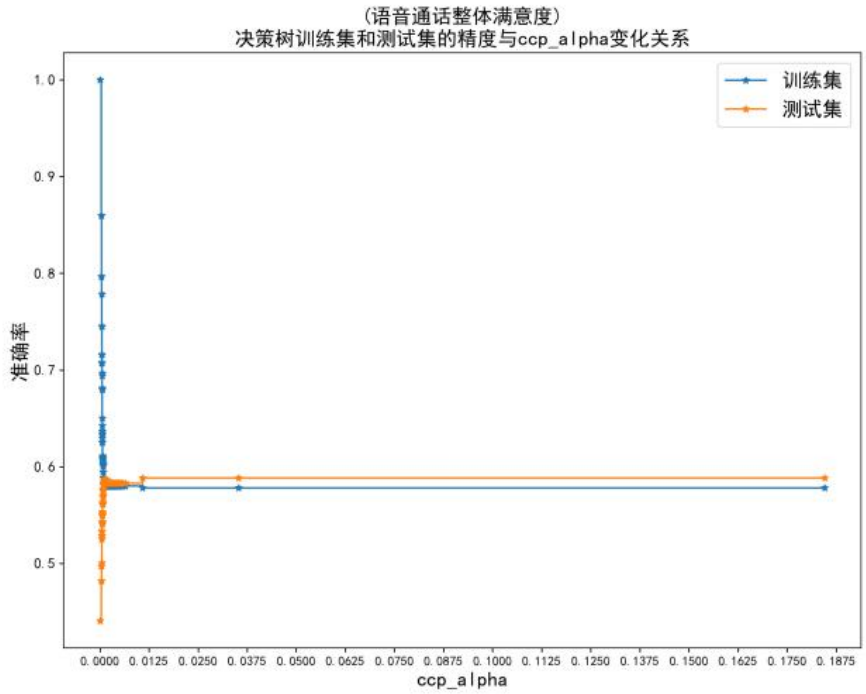


图 13 模型参数α与模型测试集、训练集精度变化关系（语音通话整体满意度）

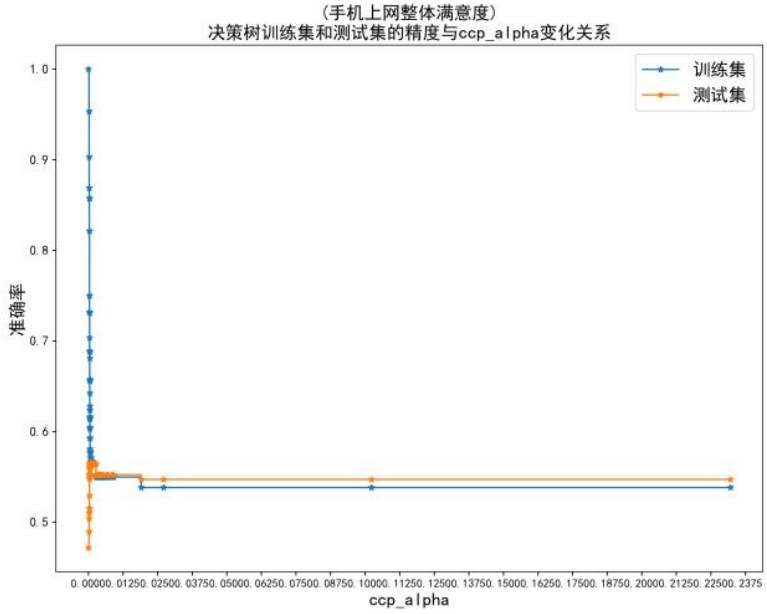


图 14 模型参数α与模型测试集、训练集精度变化关系（手机上网整体满意度）

可以看出最优精度区间都在[0.0000～0.0125]之间。为了提高模型的准确率，在后续超参数优化过程中会在最优区间内进行选择，本文采取网格搜索与交叉验证的方式进一步调参。

（1）网格搜索（Grid Search）

网格搜索是一种穷举方法，即在上述得出的最优精度区间内，将该区间内的参数依次代入模型中进行调试，最终选出较好的参数。随后进行交叉验证的处最优参数组合的得分。

（2）交叉验证（Cross Validation）

交叉验证是将数据集 D 划分为 k 个大小相似的互斥子集，然后用 $k-1$ 个子集的并集作为训练集，余下的子集作为测试集，总共可以获得 k 组测试集和训练集，即进行 k 次训练及测试，最终返回 k 次测试结果的均值。这种方法被称为“ k 折交叉验证”。本文取 $k-1$，即 5 折交叉验证。示意图见**图 15**：

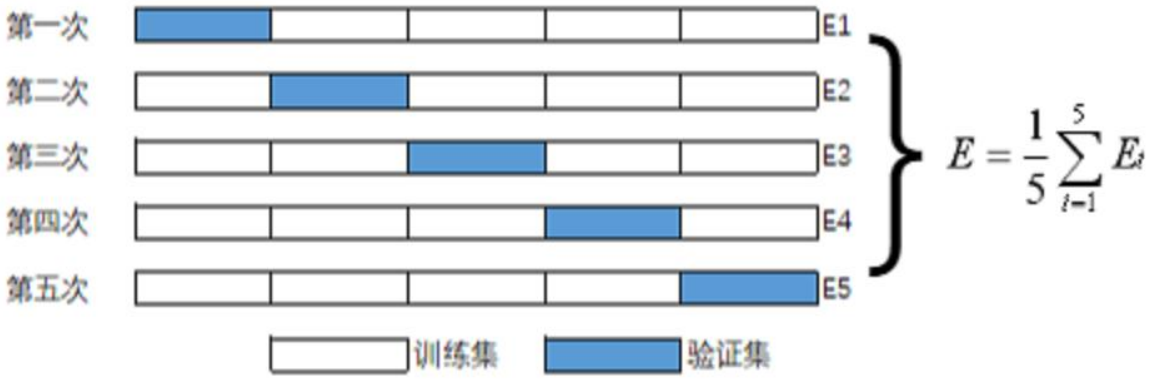


图 15　五折交叉验证示意图

经过网格搜索和 5 折交叉验证后选取了最优参数，并对 CART 决策树完成 CCP 剪枝操作，绘制出后剪枝可视化图，由于篇幅有限，此处展示针对语音通话整体满意度的后剪枝部分可视化，见**图 16**。

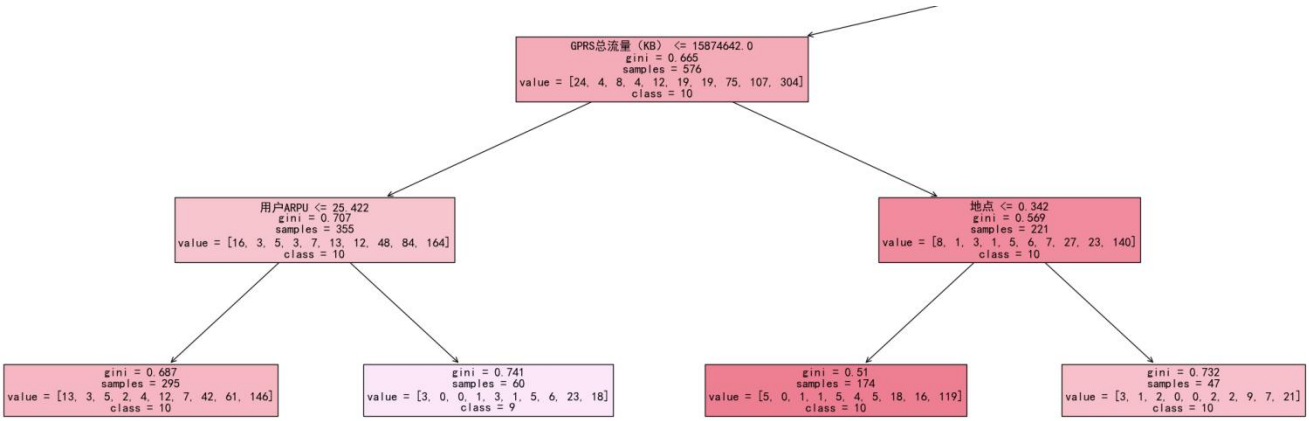


图 16　　CCP 后剪枝后 CART 决策树部分可视化图

本文对剪枝前与剪枝后的 CART 模型的精度做了对比，并绘制出剪枝前后模型的精度对比图，见**图 17** 和**图 18**。

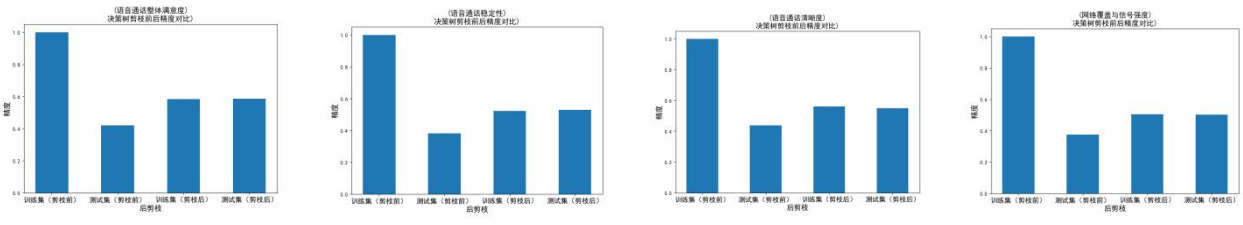


图 17　附件 1 各打分项目 CART 模型剪枝前后精度对比

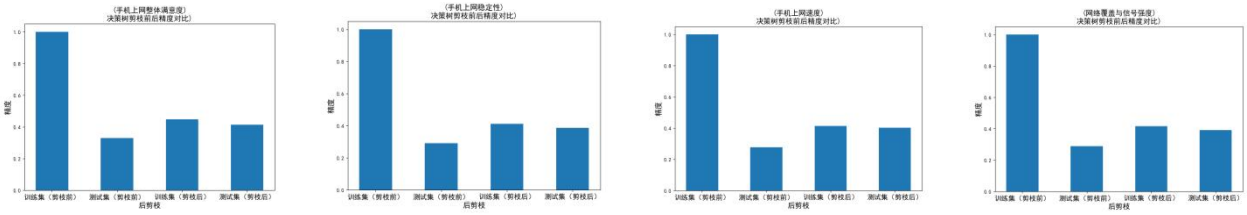


图 18　附件 2 各打分项目 CART 模型剪枝前后精度对比

从图中可以看出，剪枝后的测试集精度都得到了有效的提升，过拟合现象得到大幅缓解。

6.2 随机森林模型

随机森林（Random Forest）是一种集成学习模型，先产生一组基学习器，然后用某种策略将这些学习器组合起来，通常决策树被用做基学习器。决策树本身属于弱学习器，但经过集成之后会有很强的预测效果，成为强学习器。随机森林本质上是多棵决策树组合在一起的多数投票结果。

随机森林的算法步骤如下：

（1）使用训练数据训练模型，从数据样本集中抽取训练集，共进行 k 轮抽取，共得到 k 个相互独立的训练集。

（2）特征选取：根据上文所述的重要特征选择选取特征。

（3）每次使用一个训练集训练模型，共 k 个模型。

（4）输出结果：采用投票的方式得到分类结果

随机森林模型建立大致流程图见图 19：

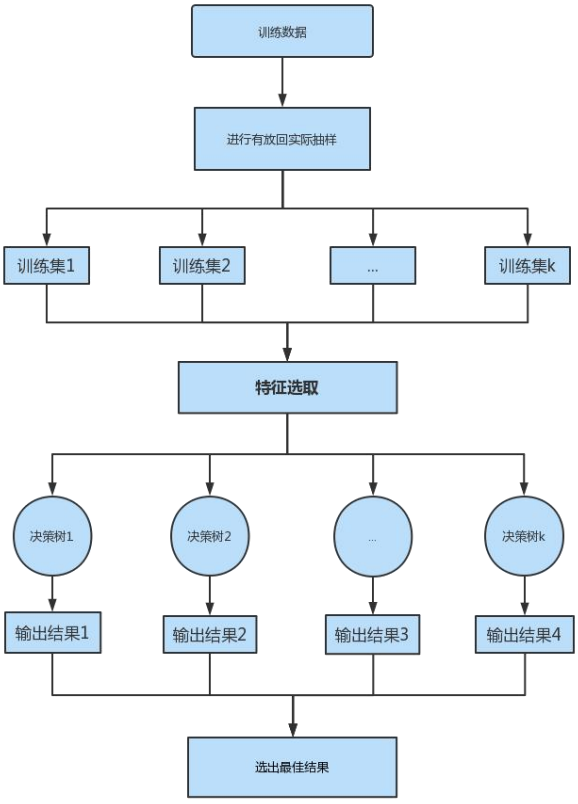


图 19　随机森林模型流程图

6.3 XGBoost 模型

XGBoost 模型是一种以 CART 决策树模型为基础的集成学习算法，可将弱学习器提升为强学习器，是基于梯度提升树 GBDT 的思想，对 GBDT 模型的优化。XGBoost 模型是一个加法模型，它串行的训练一组 CART 决策树，对预测结果进行加权求和，得到一个强学习器。模型常用的目标函数为：

$$Obj(\theta)=L(\theta)+\Omega(\theta)$$

其中 $Obj(\theta)$ 为目标函数；$L(\theta)$ 为模型训练损失，用以评估模型的优劣；$\Omega(\theta)$ 为正则化项；θ 为样本空间。

XGBoost 基于单一树模型进行线性加权集成，训练过程采用加法训练。其预测原理是：某时刻的预测值是前一时刻的预测值与此刻的函数之和，目标是选取最优的函数值。设训练数据集 $\{(x_1,y_1),(x_2,y_2)...(x_n,y_n)\}$，其目标函数的表达式为：

$$Obj^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t)}\right)+\sum_{i=1}^{t}\Omega(f_i)$$

其中，$\hat{y}_i$ 表示第 i 个样本 x_i 的预测值；$\sum_{i=1}^{n}l\left(y,\hat{y}_i^{(t)}\right)$ 为损失函数，本文所使用的为 softmax 的损失函数，即 $L_i=-\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$；$\sum_{i=1}^{t}\Omega(f_i)$ 为正则化项，表示 i 棵树的复杂度。由于 GBDT 梯度提升树有 $y_i=\hat{y}_i^{(t-1)}+f_t(x_i)$，将其代入目标函数并对正则化项进行拆分，其目标函数变为：

$$\sum_{i=1}^{n}l\left(y_i,y_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t)$$

f_t 表示 t 时刻的函数。该原理就是通过寻找最优的 f_t 来使目标函数最小化。

运用泰勒公式对目标值进行展开，泰勒公式为：

$$f(x+\Delta x)\approx f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2$$

利用泰勒公式可将目标函数表示为：

$$\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{t-1}+f_t(x_i)\right)=\sum_{i=1}^{n}\left[l\left(y_i,\hat{y}_i^{t-1}\right)+l'\left(y_i,\hat{y}_i^{t-1}\right)f_t(x_i)+\frac{1}{2}l''\left(y_i,\hat{y}_i^{t-1}\right)f_t^2(x_i)\right]$$

用 g_i 记为第 i 个样本损失函数的一阶导数，h_i 记为第 i 个样本损失函数的二阶导数，即：

$$g_i=l'\left(y_i,\hat{y}_i^{t-1}\right)$$

$$h_i=l''\left(y_i,\hat{y}^{t-1}{}_i\right)$$

如此可以将目标函数表示成如下形式：

$$Obj^{(t)}=\sum_{i=1}^{n}\left[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t)$$

其中，树模型的复杂度定义为$\Omega(f)$，可以利用该复杂度的参数来更新目标函数。

6.4 随机森林和 XGBoost 模型的超参数优化

在构建随机森林和 XGBoost 模型时，首先以各参数的默认值构建模型，得到初始模型的评价标准值，然后对随机森林重要参数:
max_depth,n_estimators,min_samples_leaf,min_samples_split 以及 xgboost 重要参数：max_depth,n_estimators,learning_rate 等超参数进行调节。其中 n_estimators 为迭代次数，即弱分类器的个数；learning_rate 为学习率，其作用是用来控制迭代速率；max_depth 为树的最大深度，用来避免过拟合；min_samples_leaf 和 min_samples_split 分别为叶子节点需要的最小样本数和拆分内部节点所需的最少样本数，该参数可以限制模型的复杂程度，产生平滑模型的效果，减轻过拟合现象。

由于对不同参数组合进行网格搜索以及交叉验证需要耗费大量时间，为了提高参数优化效率，通过绘制学习曲线确定单个参数的最优精度区间，之后针对每个参数的最优区段通过网格搜索配合五折交叉验证得到最优参数组合。

以语音通话整体满意度和手机上网整体满意度两个打分数据为例：

1）首先需要控制基学习器单个树的深度 max_depth，由于分类树的拟合能力与其复杂程度成正比，一般而言，树的深度越大则表示其分支越多，更加具特性而失去通用性，所以限制最大深度可以有效避免过拟合现象，随机森林与 xgboost 训练集与测试集随深度误差变化见**图 20**。

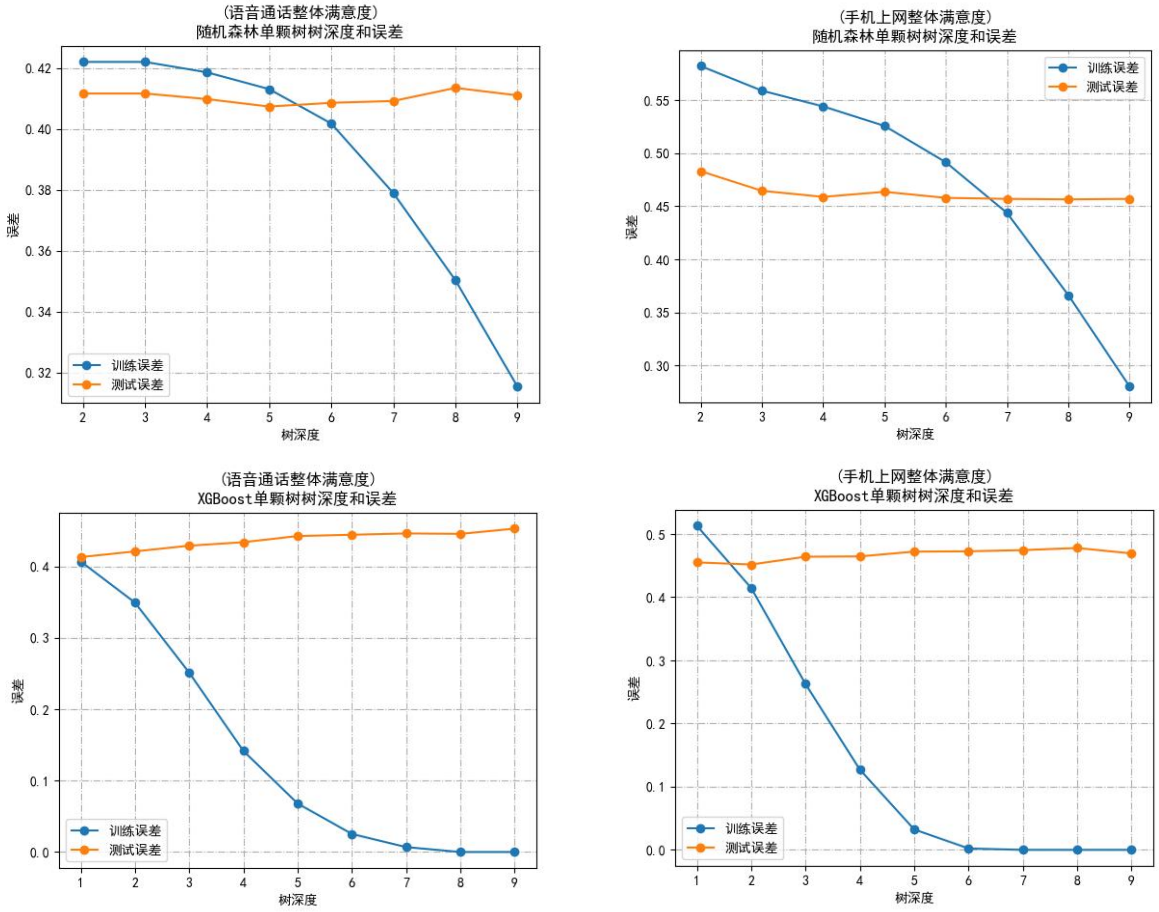


图 20　随机森林与 xgboost 训练集与测试集随深度误差变化

可以看出两类模型随着树深度的增加出现了严重的过拟合现象，针对语音通话整体满意度和手机上网整体满意度的预测误差分析得出：随机森林模型单个树的最大深度区段位于 5~7 之间时测试误差和训练误差会逐渐达到中和，之后随着深度的增加，过拟合程度会大幅加深；而

xgboost 模型针对两项打分数据测试误差和训练误差的中和深度区段则为 1~2，所以综合上述分析，针对随机森林和 xgboost 分别选定[5,7]、[1,2]为基于两项打分预测模型单个树深度的最优区段，接下来需要控制随机森林模型单个树叶子节点需要的最小样本数和拆分内部节点所需的最少样本数来进一步控制过拟合现象，最后通过网格搜索和五折交叉验证找到随机森林这三项参数的最佳组合。

2）同样，为了更好地确定 xgboost 模型的最佳迭代次数区段以及提高该模型的鲁棒性，我们需要对 xgboost 的学习率 learning_rate 进行调节，以便控制每次迭代更新权重的步长。xgboost 根据学习率误差变化曲线见**图 21：**

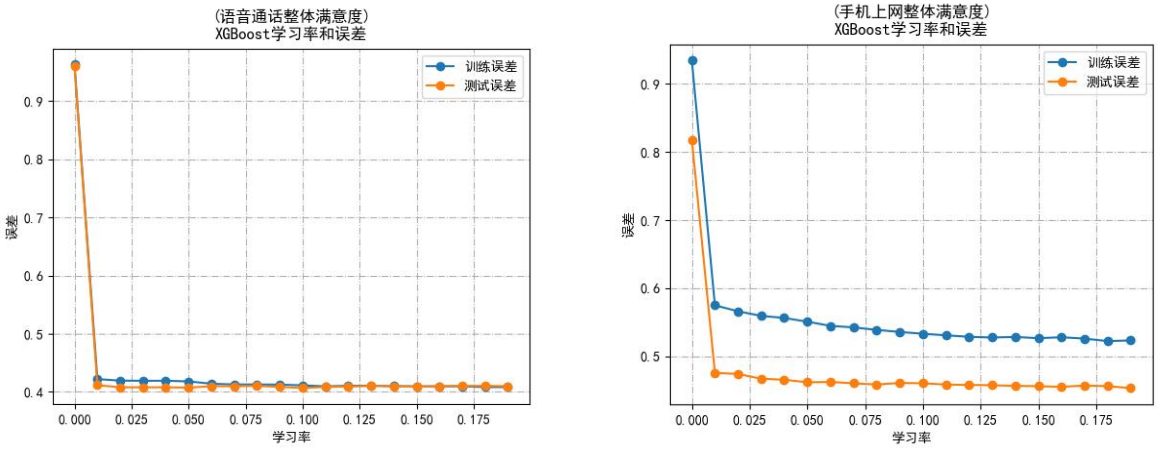


图 21　XGBoost 模型学习率误差变化曲线

由图可以看到初始学习率的增加对于测试集和训练集的误差带来极大的改善，泛化能力不断提升，针对两项打分数据最低误差区段集中在 0.1~0.175 之间，接下来需要控制每棵树随机采样比例参数 subsample，决定最小叶子节点样本权重和 min_child_weight 以及限制每棵树权重改变的最大步长 min_delta_step，同时引入权重的 L2 正则化项参数 reg_lambda 来进一步改善过拟合以及欠拟合现象，同样通过网格搜索找到最优参数组合。

3）接下来，我们需要控制两类集成模型中的基学习器的数量（迭代次数），找到每个模型在低误差下的收敛区间来达到一个较为理想的预测效果。误差变化曲线图见**图 22：**

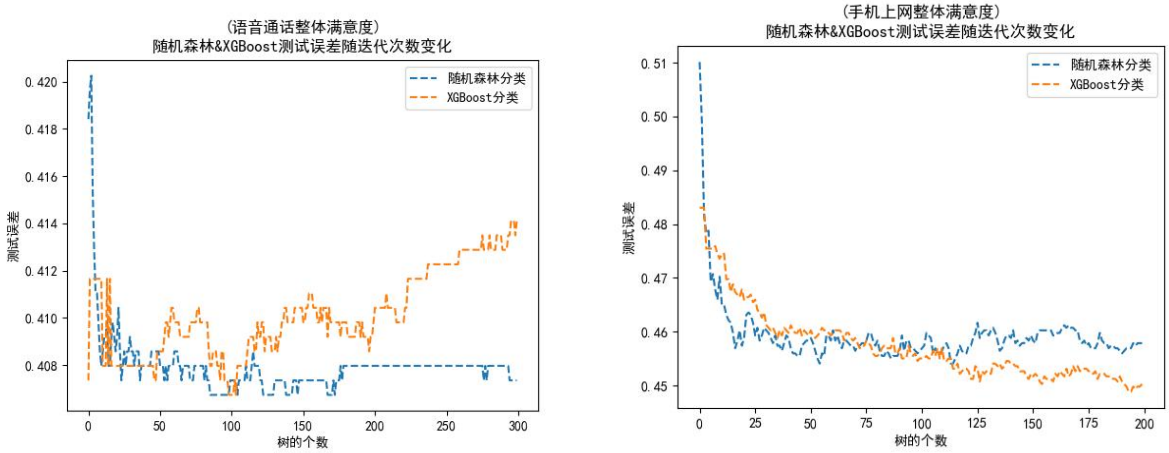


图 22　随机森林和 XGBoost 测试误差随迭代次数变化图

由图分析可以看出：对于语音通话和手机上网整体满意度数据的预测，随机森林的收敛速度较之 xgboost 会更快地达到一个较低的测试误差，其最低测试误差区段出现于 50 次迭代之后，而 xgboost 则位于 100~200 次之间，确定最佳迭代次数区段之后，对于两类模型迭代次数参数 n_estimators 通过网格搜索和五折交叉验证可找到对应打分数据预测的最佳参数值。

最后利用选取到的最优参数组合构建出最优的模型，并计算出调参前后各个模型的精度，具体的模型参数和调参前后模型的精度见**表 11-表 14**：（由于篇幅受限，本文只给出语音通话整体满意度和手机上网整体满意度两个打分项的详细数据，详细参数数据见附录问题二代码：模型训练&打分数据预测）：

表 11　随机森林模型参数与调参前后语音通话整体满意度的训练集和测试集精度

参数	值	调参前训练集精度	调参前测试集精度	调参后训练集精度	调参后测试集精度
max_depth	5				
n_estimators	86				
max_features	6	1	0.5583	0.5927	0.5896
min_samples_leaf	5				
min_samples_split	2				

表 12 随机森林模型参数与调参前后手机上网整体满意度的训练集和测试集精度

参数	值	调参前训练集精度	调参前测试集精度	调参后训练集精度	调参后测试集精度
max_depth	8				
n_estimators	55				
random_state	22	1	0.4998	0.6156	0.5413
max_features	10				
min_samples_leaf	1				
min_samples_split	8				

表 13　XGBoost 模型参数与调参前后语音通话整体满意度的训练集和测试集精度

参数	值	调参前训练集精度	调参前测试集精度	调参后训练集精度	调参后测试集精度
n_estimators	99				
max_depth	1				
eta	0.1				
objective	‘multi:softmax’				
random_state	22	0.9747	0.5662	0.5864	0.5906
subsample	0.7				
min_child_weight	14				
min_delta_step	0.9				
reg_lambda	0.4				

表 14　XGBoost 模型参数与调参前后手机上网整体满意度的训练集和测试集精度

参数	值	调参前训练集精度	调参前测试集精度	调参后训练集精度	调参后测试集精度
n_estimators	71				
max_depth	2				
eta	0.16				
objective	‘multi:softmax’				
random_state	22	0.9978	0.5074	0.5675	0.5422
subsample	0.4				
min_child_weight	14				
min_delta_step	0				
reg_lambda	0.4				

可以看出，经过模型超参数优化后，模型的测试集精度都得到了提高，模型的过拟合现象得到缓解，泛化性能得到了一定的提升。

6.5 模型融合

根据上述建立的三个模型，采用融合的方式将三个模型进行融合，采用软投票加权融合的方式。具体的模型结果如下：

$$result = \beta_1 \bullet result_1 + \beta_2 \bullet result_2 + \beta_3 \bullet result_3$$

其中，$\beta_1, \beta_2, \beta_3$ 为权重系数，$result$ 表示被分到该结果的概率，$result_1, result_2, result_3$ 分别表示在 CART 模型、随机森林、XGBoost 模型中被分到该结果的概率，经过不断调试，得到最好的权重组合为 $\beta_1 = 1$， $\beta_2 = 2$ $\beta_3 = 3$。然后将算出的 $result$ 进行归一化处理，得到了最终分类的概率。

之后计算出了 CART 模型、随机森林、XGBoost 模型、模型融合的测试误差，测试误差的可视化图见**图 23** 和**图 24**：

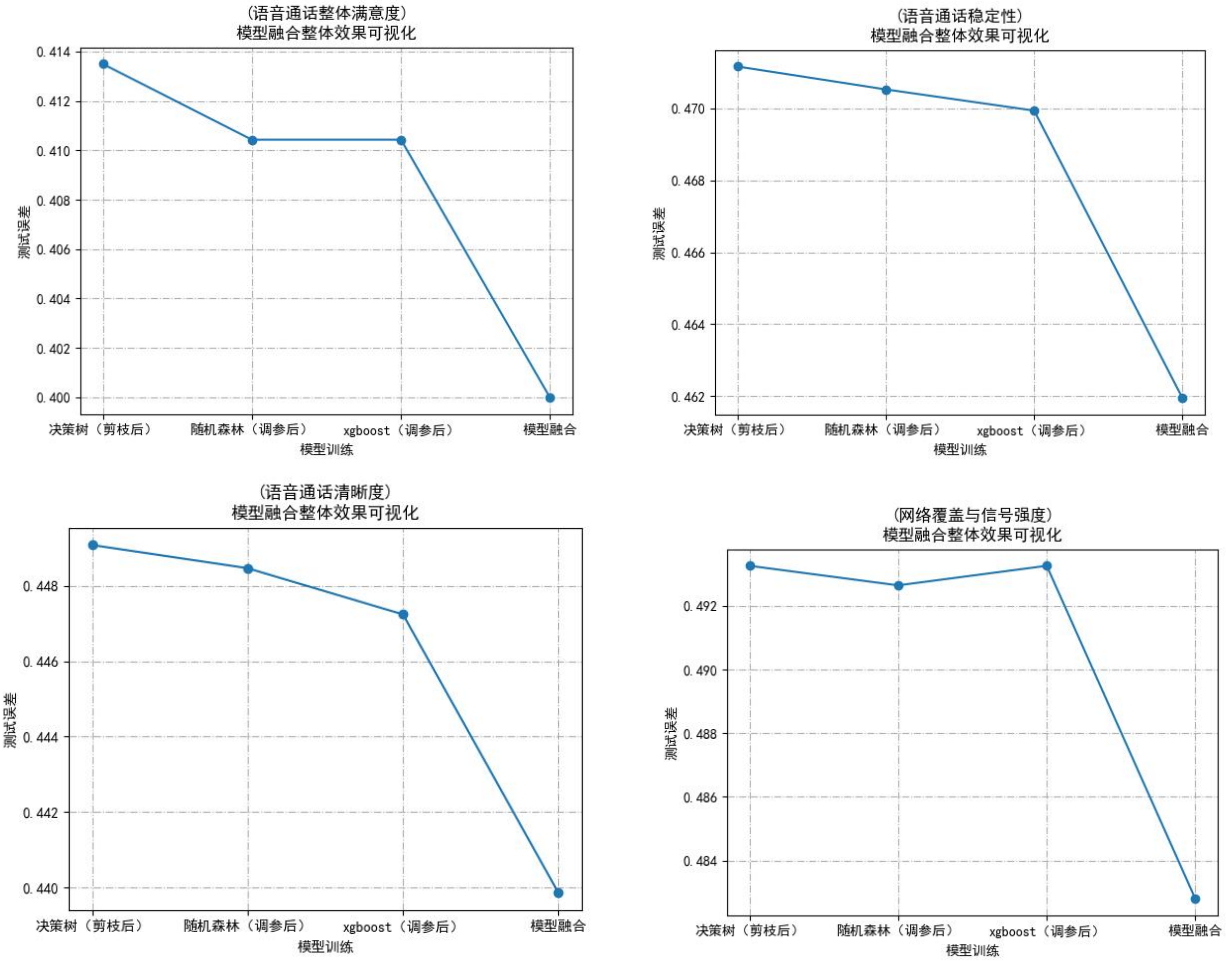


图 23　附件 1 中 4 个打分项目模型测试误差

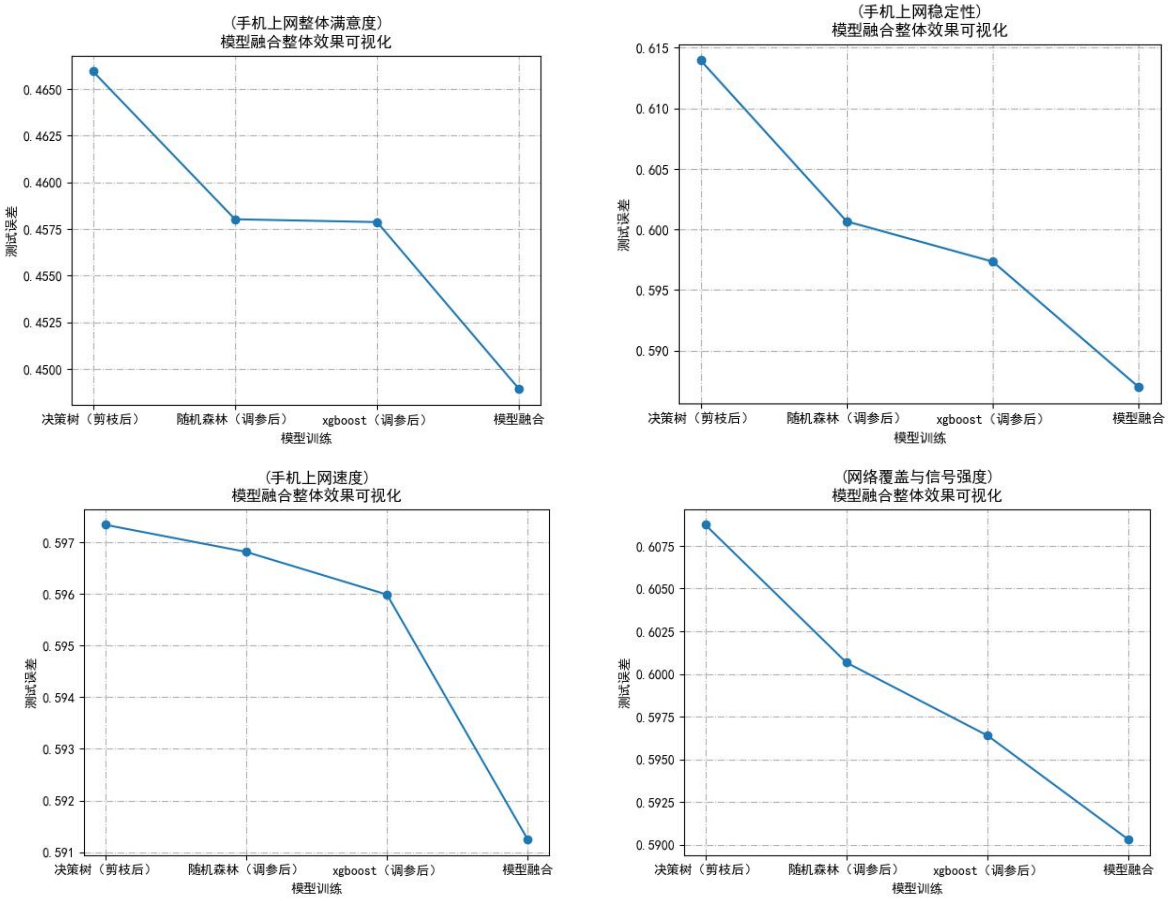


图 24　附件 2 中 4 个打分项目模型测试误差

可以看出，模型融合后测试误差降到了最低，即模型融合后的预测结果更准确

6.6 基于决策树模型的预测及结果合理性

将数据导入模型得到了预测结果，并保存在 result.xlsx 中。

上文的结果表明，模型最终的测试集精度在 0.4-0.6 之间，这个结果并不是很高，但结合问题一中对数据的描述性统计，发现数据中大部分指标存在异常值，即客户打分的主观性很强，在相同的因素下，由于不同的人需求不同，这也导致了打分存在不一致的情况。经过对模型不断的学习调试与融合，也确保了测试集的精度在不断提高。由此可知，最终的预测结果符合实际，是合理的。

七 、模型评价

1. 模型优点

（1）本文运用多个模型，对数据集使用充分，对数据集的处理比较完整。

（2）模型调参过程中使用了网格搜索和交叉验证，提高了模型的泛化能力。

（3）采用软投票加权融合的方法，提高了整体的预测能力。

2. 模型缺点

（1）特征工程的特征挖掘仍有不足，后续可以对特征工程更好的处理。

（2）由于数据存在异常值，加之客户打分十分主观，模型在测试集上的表现一般，但是预测出的打分结果与实际结果相差不大。

八、模型推广

本文所建立的模型，可以投入到现实生活中使用，可以针对其它地区移动用户满意度的判别提供参照，并且该模型可以推广到生活中更多场景的预测问题中。

参考文献

[1]周志华.（2016）.机器学习.清华大学出版社，北京
[2]武炜杰.随机森林的应用与优化方法研究[D].江南大学，2021
[3]吕晓令 宋捷.大数据挖掘与统计机器学习.中国人民大学出版社

队伍编号	MCB2202628
赛道	B

基于机器学习的北京移动用户满意度分析

摘 要

本文基于北京移动用户数据进行用户满意度分析，主要研究两方面的问题，一是量化分析用户满意度影响因素的重要性，寻找重要影响因素；二是在考虑变量重要性基础上建立预测模型，对用户满意度评分进行预测。

文章主要包含三大部分。**第一部分主要为数据预处理和探索性分析**，首先对数据进行变量筛选和剔除、缺失值和异常值处理、类别变量编码等，在尽量不损失信息的前提下简化特征；其次，在此基础上进行数据探索性分析，通过统计用户评分分布情况以及相关性分析等初步探究数据特征。

第二部分为影响因素重要性量化分析，采用随机森林进行特征重要性评估。结果显示，对语音业务四个方面打分而言，最重要的前三个变量均为：GPRS 总流量（>11%）、前三个月 MOU（>10%）、当月 ARPU（>10%）；而同一变量对上网数据四个方面打分影响程度略有不同，但从总体来看，重要变量为：年龄、当月 GPRS 资源使用量（GB）、当月 MOU，重要性均>6%。在此基础上，我们对重要性排序靠前的部分变量进行了深入分析，通过直方图直观对比不同评分下该变量的分布情况，并基于此对北京移动公司提出建议。

第三部分为用户满意度预测，首先进行特征选取，对于语音业务数据在考虑变量重要性的基础上，仅将预处理后的附件 1 和附件 3 共同含有的变量作为模型学习的特征，对上网业务数据做相同的处理；其次，针对两个数据集共 8 个方面的评分分别构建预测模型，选取决策树、随机森林、Light-GBM 三个机器学习模型，采用 10 折交叉验证法划分训练集和测试集，并通过网格搜索寻找最优参数，基于附件 1 和附件 2 数据训练模型并测试模型性能；最后，考虑满意度评分为定比数据，以 MSE 为模型评价指标，选取最优模型进行附件 2 和附件 4 的预测，并输出结果文件。

本文最后，综合所有分析结果，对北京移动公司就如何提升用户满意度提出部分建议，竭力提升北京移动的用户满意度，让企业在 5G 时代下争取到更多市场份额。

关键词：用户分析；机器学习；满意度预测；网格搜索寻优

目录

第一章 问题重述 **1**
1.1 问题背景 **1**
1.2 研究问题 **1**
第二章 问题分析 **2**
2.1 针对问题一的分析 **2**
2.2 针对问题二的分析 **2**
第三章 数据探索性分析 **3**
3.1 数据预处理 **3**
3.1.1 变量的筛选与剔除 3
3.1.2 缺失值、异常值处理 4
3.1.3 类别变量编码 5
3.1.4 小结 5
3.2 用户满意度评分分布情况 **6**
3.3 变量相关性分析 **8**
第四章 用户分析之满意度影响因素分析 **13**
4.1 特征选择方法简介——随机森林 **13**
4.2 语音业务满意度影响因素分析 **14**
4.2.1 变量重要性量化分析 14
4.2.2 部分重要变量分析 15
4.3 上网业务满意度影响因素分析 **17**
4.3.1 变量重要性量化分析 17
4.3.2 部分重要变量分析 18
第五章 用户分析之满意度预测 **21**
5.1 分类方法简介 **21**
5.1.1 决策树 21
5.1.2 随机森林 21
5.1.3 Light-GBM 22
5.1.4 交叉验证 22
5.1.5 网格搜索 23
5.2 建立预测模型 **23**
5.3 结果分析与模型评价 **24**
5.3.1 评价指标 24
5.3.2 模型评价和结果展示 24
5.3.3 结果的比较性分析 25
第六章 建议 **27**
参考文献 **29**
附录 **30**

第一章 问题重述

1.1 问题背景

移动通信技术迅猛发展，网络覆盖愈加完善，人们越来越离不开移动通信技术带来的各种便捷。移动运营商越来越重视客户的网络使用体验，以此提升网络服务质量。客户满意度即客户对运营商产品服务的满意程度，反映了客户期望与实际感知的产品服务之间的差异，是各大运营商市场运营状况的重要体现。

根据客户投诉，对影响用户体验的问题逐点解决，是传统提升客户满意度的方法。但随着用户数量的大幅增加，移动产品的种类越来越丰富，客户的需求越来越高，传统的方法已经难以有效提升客户的满意度。数字经济时代，各大运营商需要运用数字经济的管理理念和技术手段，建立客户体验生态的全方位系统性测评体系，实现客户满意度评测的数字化转型，让客户体验赋能商业决策，让商业决策真正服务客户，共同推动移动网络高质量可持续发展。

中国移动通信集团北京公司，让客户根据自身在网络覆盖与信号强度、语音通话清晰度和语音通话稳定性三个方面的体验进行打分，同时还让客户根据语音通话的整体体验进行语音通话整体满意度的打分，并统计整理影响客户语音业务体验的因素，以此分析客户语音业务满意度的主要影响因素，并提升客户语音业务满意度。同时，对于上网数据业务，中国移动北京公司让客户根据自身在网络覆盖与信号强度、手机上网速度、手机上网稳定性三个方面的体验进行打分，同时还让客户根据手机上网的整体体验进行手机上网整体满意度的打分，并统计整理影响客户上网体验的因素，以此分析影响客户上网业务体验的主要因素，并提升客户的上网体验。本研究拟通过分析影响用户满意度的各种因素，为决策提供依据，从而实现更早、更全面提升用户满意度。

1.2 研究问题

基于以上研究背景以及所给的数据集，本文解决如下两个问题：

问题 1：研究影响客户语音业务和上网业务满意度的主要因素，并给出各因素对客户打分影响程度的量化分析和结果。

问题 2：结合问题 1 的分析，对于客户语音业务和上网业务分别建立客户打分基于相关影响因素的数学模型，并据此对客户打分进行预测研究。

第二章 问题分析

2.1 针对问题一的分析

问题一主要为影响因素重要性分析，不仅要对影响因素进行重要性排序，还需对重要性程度进行量化分析。首先需要进行数据预处理，清洗数据简化特征；其次进行包含相关性分析在内的数据探索性分析，初步探究变量相关性和重要性。对于问题一，注重考虑机器学习模型特征重要性评估方法，此种方法不仅可以给出重要性排序，还可以量化重要性程度，具有较好的效果。

2.2 针对问题二的分析

问题二为用户满意度预测，主要考虑采用机器学习模型，将 1-10 共 10 个打分数据视作 10 个类别，由此将满意度评分预测转化为类别预测。这一部分需要着重考虑两个方面的问题：一是如何选取有效的特征，需要综合问题一的结果，同时也要考虑原始数据集（附件 1、附件 2）和预测数据集（附件 3、附件 4）所含变量的不同，在训练模型时需特别注意选取附件 1 和附件 3 同时包含的变量作为机器学习模型的特征，附件 2 和附件 4 做相同处理；二是模型评价指标的选取，满意评分属于定比尺度，在建立模型时将评分视作 10 个类别将会将数据划分为定类尺度，此时采用准确率等指标进行模型评估有失偏颇，需要寻找更加符合数据的评价指标。

第三章 数据探索性分析

3.1 数据预处理

3.1.1 变量的筛选与剔除

本文包含两份原始数据集，一份为语音业务用户满意度数据集（附件 1），共含有 5433 个用户的 54 个相关信息变量（包含四个打分数据），另一份则为上网业务用户满意度数据集（附件 2），共含有 7020 个用户的 124 个信息（包含四个打分数据）。观察数据可发现，两份数据集数据量适中，但变量均较多，尤其是上网业务用户满意度数据集，且数据呈现形式多样、非规范的特点。一方面，较多的数据信息有助于挖掘更多的信息，但另一方面，变量过于冗杂则会影响分析的效果。故首先对数据进行简单分析，根据常识剔除无关变量、合并相关变量，在尽量不损失信息的情况下简化数据集。

针对语音业务用户满意度数据集，首先，两个用户描述变量为文本数据，且该变量是作为一种补充信息对满意度分析影响不大，可以删除；其次，终端品牌类型、是否实名登记用户、当月欠费金额、前第 3 个月欠费金额这 4 个变量为无关变量，故而可以剔除；此外，关于用户是否为 5G 用户这一信息，有三个相关变量，分别为：4\5G 用户、是否 4G 网络客户（本地剔除物联网）、是否 5G 网络客户，这三个变量信息重复，故仅保留一个，考虑到信息的完整性，在此保留“4\5G 用户”这一变量；最后，经过简单的数据分析发现“是否遇到过网络问题”这一变量涵盖了包括居民小区、办公室...手机没有信号等 16 个变量信息，即这 16 个变量为“是否遇到过网络问题”，以及在何处、遇到了何种网络问题的具体描述，具体而言，若 16 个变量均为“否”，则变量“是否遇到过网络问题”记为“否”，其他情况均记为“是”，在对语音业务进行满意度分析时，我们不再对具体问题进行区分，故将 16 各变量统称为“是否遇到过网络问题”。综上，经过初步变量剔除、合并后共得到 5433 个用户的 31 个相关信息（包含四个打分数据）。

针对上网业务用户满意度数据集，首先，三个用户描述变量为文本数据作为种补充信息对满意度分析影响不大，可以删除，操作系统、终端品牌、终端品牌类型、畅享套餐名称为无关变量，可以直接剔除；其次，王者荣耀使用次数、游戏类 APP 使用次数、王者荣耀使用天数以及游戏类 APP 使用天数含有重复信息，考虑到次数更具代表性，故仅保留次数相关信息，同理，终端制式和是否为 5G 网络用户仅保留终端制式，码号资源-激活时间与码号资源-发卡时间选择激活时间，另外，综合是否校园套餐用户、以及是否为校园卡绑定合约用户为一个变量——校园卡套餐情况（1：非校园卡套餐用户，

2：校园卡套餐未绑定合约，3：校园卡套餐并绑定合约）；此外，与语音业务用户满意度数据集存在同样的情况，上网业务用户满意度数据集存在多个变量信息包含的情况，采用同样的方法，剔除具体描述变量，保留整体描述变量。此步骤完成后，共得到 7020 个用户的 51 个相关信息（包含四个打分数据）。

3.1.2 缺失值、异常值处理

在数据预处理的环节中，缺失值的处理是十份重要与关键的，缺失值的存在可能对研究分析会产生较大的影响。且导致模型难以运行，一般会采用空缺值填充或者删除的方式对缺失值进行处理。而异常值指在数据集中不合理的值，又称离群点，异常值会影响模型分析的效果，甚至导致模型失败，常见的异常值处理办法是删除法、替代法（连续变量均值替代、离散变量用众数以及中位数替代）、插补法（回归插补、多重插补）等，此外，可以先把异常值变成缺失值，然后进行后续缺失值补齐等。在本文的数据中，由于用户差异较大，较多变量存在离群值数据，这里异常值仅设定为错误数据。

分别对语音业务用户满意度数据集和上网业务用户满意度数据集进行空缺值和异常值检验。根据表 3-1 所示的结果，语音业务用户满意度数据集未检测出异常值，共有 5 个变量含有缺失值。经分析，外省流量占比空缺的原因是该用户总流量为 0，故该变量缺失值用 0 填充；终端品牌和客户星级标识缺失值数量不多，且缺失这两个变量信息的用户相同，故删除这 5 个用户；重定向次数与重定向驻留时长含有较多缺失值，且无法进行替代或插补，采用删除法进行处理。经过此步骤后，语音业务数据集共包含 5428 个用户的 29 个相关信息.

表 3-1 语音业务数据集缺失值、异常值检测结果

类别	空缺值					异常值
变量	外省流量占比	终端品牌	客户星级	重定向次数	重定向驻留时长	/
数量	2971	4	5	4260	4260	/

同理，对上网业务用户数据集进行空缺值处理和异常值检验。结果显示（详见表 3-2），码号资源-激活时间、王者荣耀使用次数、高频通讯分公司、其他游戏类 APP 使用次数以及终端类型存在缺失值，其中，码号资源-激活时间缺失值较多，无法替补，删除这一变量；有 16 位用户缺失高频通讯分公司这一信息，将这 16 个样本删除，并将高频通讯分公司进行整合将城区一分公司、城区二分公司、城区三分公司统称为城区，剩余类型分公司统称为其他（最终，城区：4203，其他：2801）；有 14 位用户缺少终端类型这

一信息，进一步统计发现 7020 位用户中有 7002 位用户（约 99.7%）终端类型为手机，这一变量在分析中失去意义，可剔除；最后是对王者荣耀使用次数和其他游戏类 APP 使用次数的处理，这两类变量缺失值较多，考虑到已经保留游戏类应用使用流量这一变量，故剔除王者荣耀使用次数和其他游戏类 APP 使用次数这两个变量。

此外，在对上网业务用户数据集中发现异常值，有 18 位用户年龄信息为 0，显然错误，删除这 18 位用户。

经过此步骤后，上网业务数据集共包含 6986 位用户的 49 个相关信息。

表 3-2 上网业务数据集缺失值、异常值检测结果

类别	空缺值					异常值
变量	码号资源-激活时间	王者荣耀使用次数	高频通讯分公司	其他游戏类 APP 使用次数	终端类型	年龄
数量	4150	414	16	414	14	18

3.1.3 类别变量编码

在很多机器学习任务中，特征不全是连续值，而有可能是分类值，而机器学习一般仅对数字特征友好，无法直接利用类别特征进行学习，我们在将数据输入算法进行训练前需要对类别特征进行编码处理，将其转换成数字特征。一是转换后可以提高模型运算效率，二是对于一些模型，比如逻辑回归或计算距离时，无法对分类值直接进行计算。然而，直接转换为数字也会带来一些问题，如转换为数字后，默认为连续变量，违背最初设计，影响效率；转换后的值会影响同一特征在样本中的权重。因此，需要更好的编码方式对特征进行转换。

目前最常用的编码方式有 One-Hot Encoding、Label Encoding、Ordinal Encoding 等。One-hot 编码是类别变量编码中较为常用的一种编码方式，One-hot 方法会创建出对应的 N 列特征，其中每列代表该样本是否为该特征的某一种取值。每个类别映射到一个包含 1 和 0 的矢量，表示该特征的存在与否，向量的数量取决于特征类别的数量。但由于附件 1、2 中如客户星级标识、终端品牌等离散特征的特征值过多，导致生成特征的数量太多且过于稀疏，大大降低运行速度和效率。考虑到多数类别间为无序关系，故本文最终采用 Ordinal Encoding 序列编码，把所有的相同类别的特征编码成同一个值，对特征进行随机排序，减少人为误差和影响。

3.1.4 小结

综上，前两部分数据预处理主要为变量的筛选和缺失值、异常值处理，变量筛选部

分遵循两个原则，一是去重原则，筛选出具有重复信息的变量，保留其中含有信息最全面、缺失值最少的变量；二是去繁留简原则，剔除具体描述性的变量，仅保留概括性信息变量。缺失值、异常值处理部分主要采用删除法，若某个变量缺失值过多，考虑剔除该变量；若某个变量缺失值较少，考虑剔除相应的用户（样本）。第三部分预处理则是为模型做准备。

整个数据预处理过程大致如下：

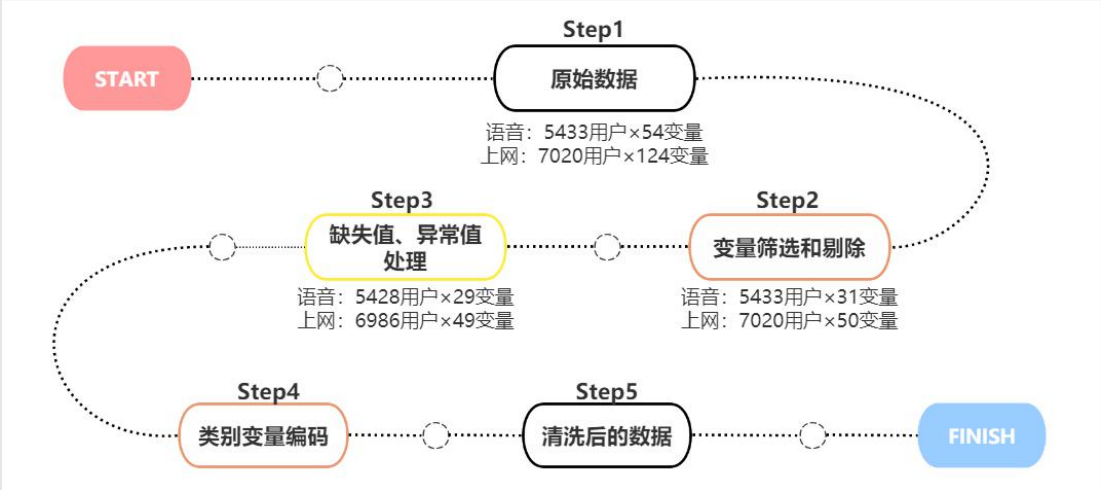


图 3-1 数据预处理过程

3.2 用户满意度评分分布情况

如图 3-2 所示，作出语音通话整体满意度、网络覆盖与信号强度、语音通话清晰度、语音通话稳定性四项指标的个数柱状图。该图描述出语音业务客户 1—10 打分的频繁程度，四项指标分布近似，10 分均占比最大，占总评分人数的 50%以上，10-2 分所占人数呈现阶梯式递减现象，在 2 分处出现转折，1 分人数相比于 2—4 分处较多。由此，可以看出，一半以上的客户对中国移动北京公司的语音业务非常满意，多数客户对语音业务较为满意，少部分客户对其语音业务不满意，不满意的客户中，1 分占比最大。

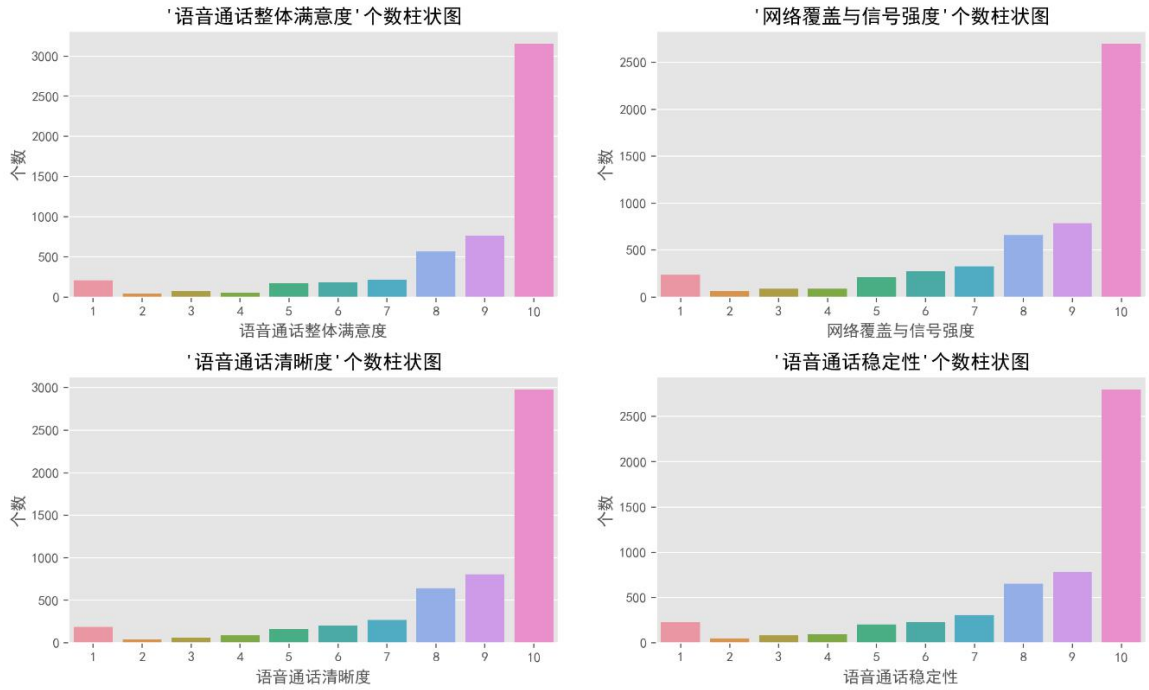


图 3-2 “语音业务”客户打分的分布情况

同理，作出手机上网整体满意度、网络覆盖与信号强度、手机上网速度、手机上网稳定性四项指标的个数柱状图（见如下图 3-3）。与语音业务客户打分情况相类似，上网业务打分情况中，10 分占比最多，超出 50%，其次打分最多的是 8 分，9 分的数量位于 8 分后，7-4 分递减，3 分的数量相比于 4 分回升，2 分处骤减，1 分大幅度增加。可以看出，上网业务的整体满意度不如语音业务，1 分非常不满意的比例有明显提升，较为满意的比例有所下降。

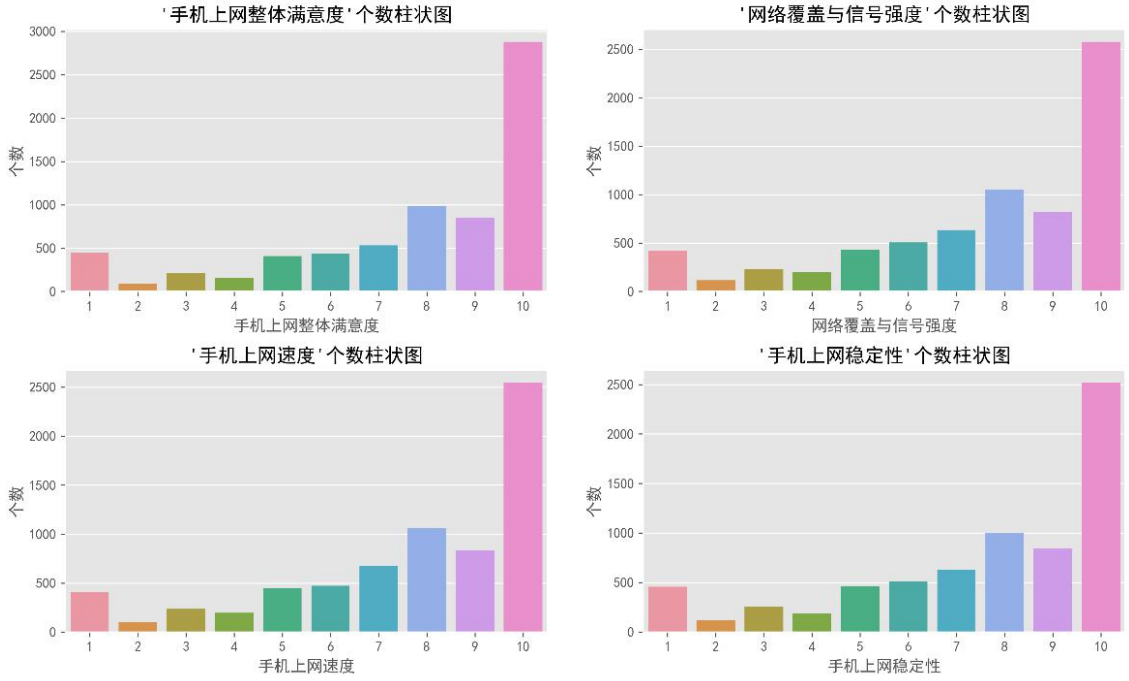


图 3-3 “上网业务”客户打分的分布情况

综上，我们对客户在语音业务、上网业务两方面的分布进行特征分析，可以初步得到客户的满意度趋势。中国移动北京公司在提升两方面满意度时，应略微侧重提升上网业务，其上升空间更大，效果会更加显著。中国移动北京公司在保证其非常满意的忠诚客户利益的同时，要更加重视那些对某些业务存在不满但对公司的总体满意度较高的用户，重视其不满和建议，以免失去该部分客户。

3.3 变量相关性分析

相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度，我们常用的是皮尔森相关系数和斯皮尔曼相关系数。其中，皮尔森相关系数（Pearson correlation coefficient, PCC）是衡量两个连续型变量的线性相关关系；Spearman 相关用于计算等级数据之间的关系，这类数据的特点是数据有先后等级之分，但连续两个等级之间的具体分数差异却未必都是相等的。即 Pearson 处理变量的数据原始值，而 Spearman 处理数据排序值。

基于上述变量相关性的基础知识，我们将语音业务、上网业务用户满意度数据中各影响因素分别划分为数值型数据和类别型数据，见表 3-3。数值型数据求出 Pearson 相关系数，类别型数据求出 Spearman 相关系数，并做出对应的相关系数热力图。

表 3-3 语音、上网业务数据集字段的划分

语音业务影响因素		上网业务影响因素	
数值型字段	类别型字段	数值型字段	类别型字段
mos质差次数	是否遇到过网络问题	年龄	居民小区
未接通掉话次数	是否去过营业厅	王者荣耀使用次数	办公室
家宽投诉	4\5G用户	游戏类APP使用次数-->改为其他	高校
资费投诉	语音方式	通信类应用流量	商业街
ARPU（家庭宽带）	终端品牌	游戏类应用流量	地铁
套外流量（MB）	是否关怀用户	网页类应用流量	农村
套外流量费（元）	客户星级标识	音乐类应用流量	高铁
外省语音占比		视频类应用流量	其他场景
语音通话-时长（分钟）		邮箱类应用流量	网络信号差/没有信号
省际漫游-时长（分钟）		当月GPRS资源使用量（GB）	显示有信号上不了网
当月ARPU		畅享套餐档位	上网过程中网络时断时续或时快时慢
当月MOU		主套餐档位	手机上网速度慢
前3月ARPU		当月MOU	其他网络问题
前3月MOU		近3个月平均消费	看视频卡顿
外省流量占比		本年累计消费	打游戏延时大
GPRS总流量（KB）		套外流量（MB）	打开网页或APP图片慢
GPRS-国内漫游-流量（KB）		套外流量费（元）	下载速度慢
			手机支付较慢
			其他APP通用问题
			高单价超套客户（集团）
			高频高额超套客户（集团）
			客户星级标识
			性别
			是否全月漫游用户
			是否不限量套餐到达用户
			终端制式
			当月高频通信分公司
			校园卡用户

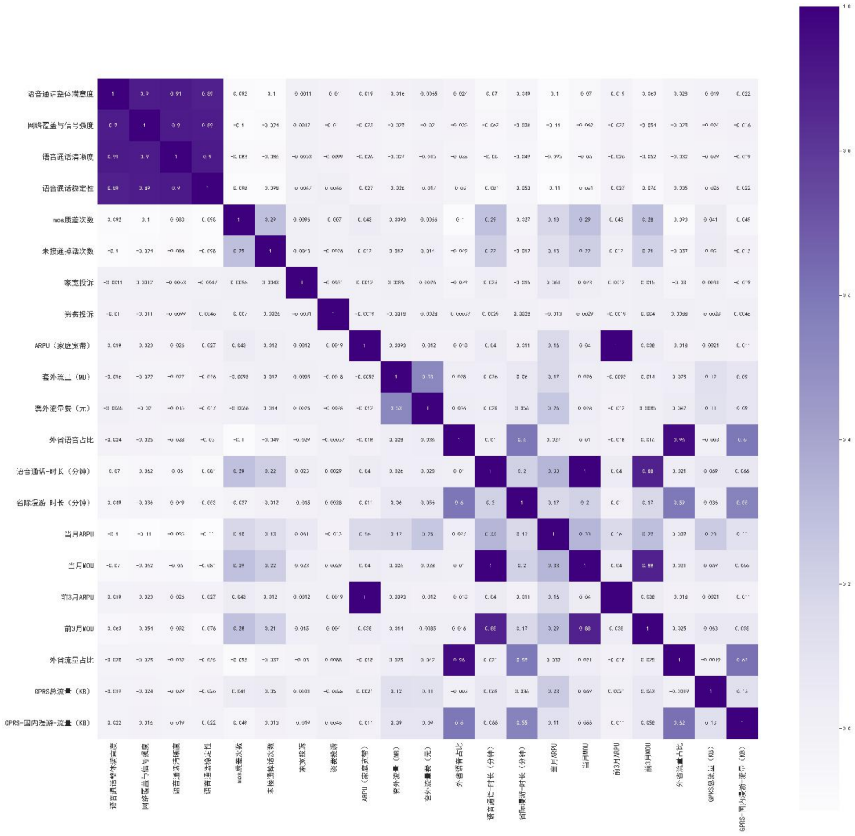


图 3-4 “语音业务”数值型指标相关系数热力图

在影响语音通话整体满意度的数值型指标中，我们可以看到，其 Pearson 相关系数的值均较低，网络覆盖与信号强度、语音通话清晰度、语音通话稳定性三者与语音通话整体满意度有很强的正相关性，均约为 0.9，但具体的影响因素指标与语音通话整体满意度的相关性均较低，其中当月 ARPU、未接通掉话次数与语音通话整体满意度的相关系数相比于其他指标较大，但仅为-0.1，有一定的负相关性，其余数值型指标影响较为微弱。

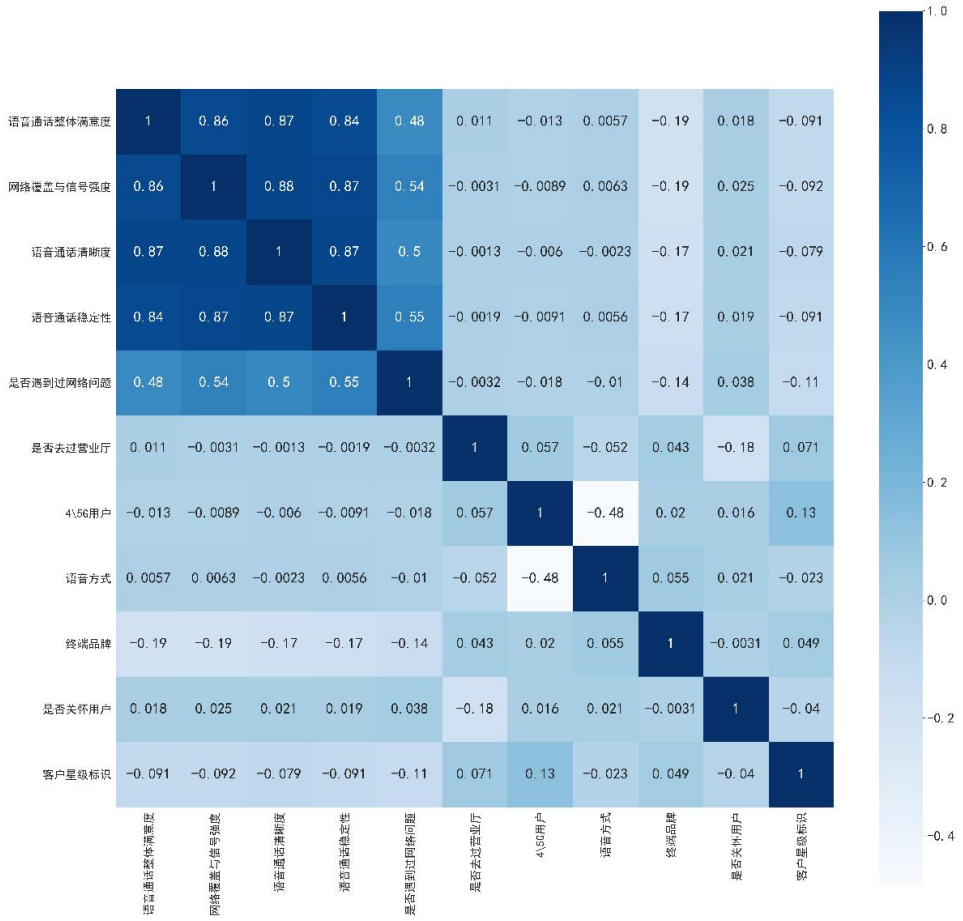


图 3-5 “语音业务”类别型指标相关系数热力图

在影响语音通话整体满意度的类别型指标中，我们可以看到，网络覆盖与信号强度、语音通话清晰度、语音通话稳定性三者与语音通话整体满意度的 Spearman 相关系数较高，有很强的正相关性，但具体的影响因素指标中，是否遇到过网络问题与语音通话整体满意度的相关系数为 0.48，说明其对语音通话整体满意度的影响较大。这里要注意说的是，数据预处理中，将类别变量随机编码，针对是否遇到过网络问题这一指标，python 随机用 0 表示是，1 表示否，故该相关系数为正值。此外，与其他剩余指标相比，终端品牌与语音通话整体满意度的相关系数为-0.19，认为如华为、苹果、小米等终端品牌对语音通话整体满意度有一定影响，其余指标的影响较为微弱。

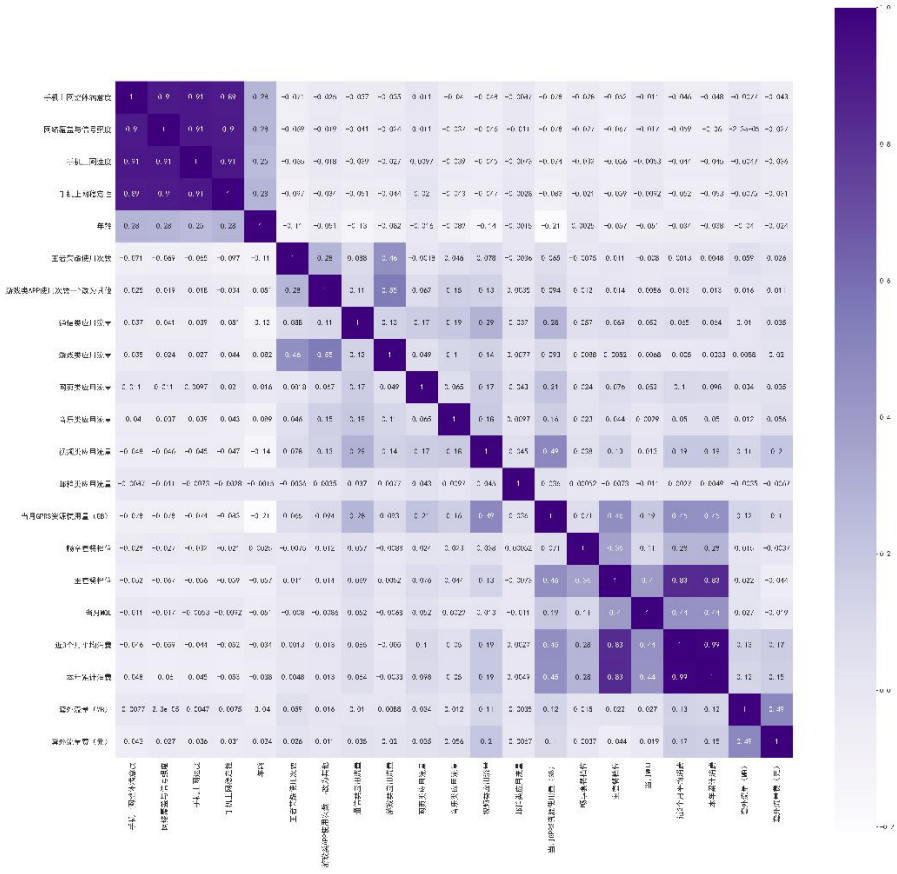


图 3-6 “上网业务”数值型指标相关系数热力图

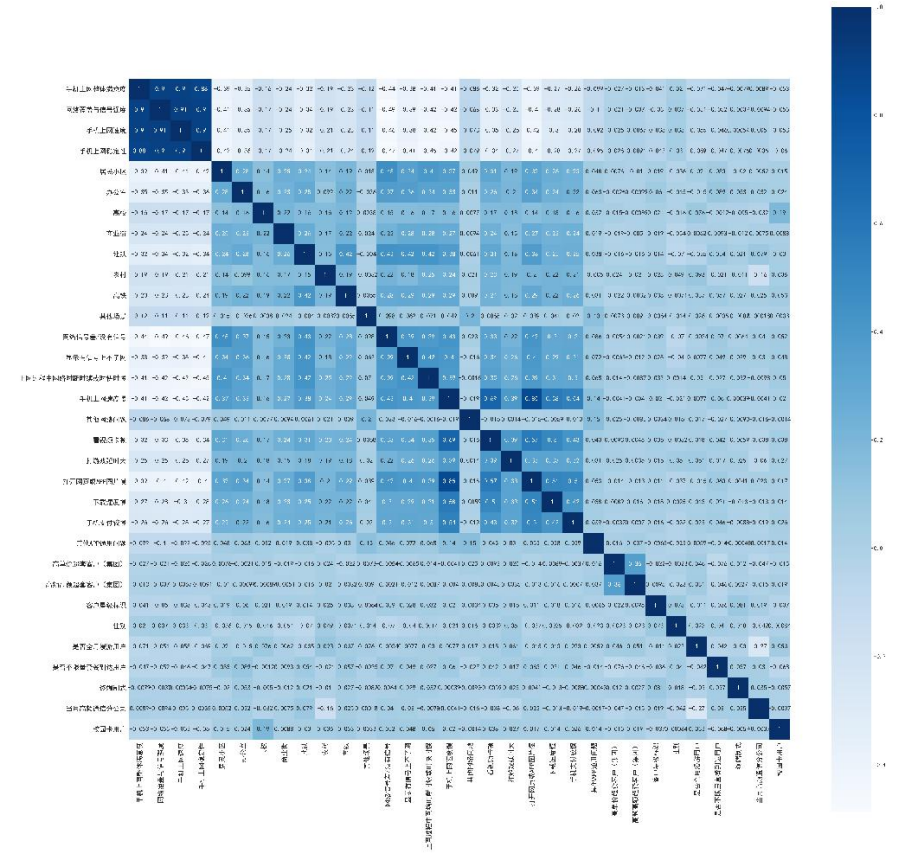


图 3-7 “上网业务”类别型指标相关系数热力图

同理，做出上网业务数值型、类别型指标的相关系数热力图（分别见图 3-6、图 3-7），可以看出，在数值型指标中手机上网整体满意度与年龄、当月 GPRS 资源使用量（KB）的相关系数分别为 0.28 和-0.078，相比于其他指标具有明显的相关性；类别型指标中我们发现手机上网整体满意度与网络信号差/没有信号、上网过程中网络时断时续或时快时慢、手机上网速度慢、打开网页或 APP 图片慢、居民小区、显示有信号上不了网的相关系数分别为-0.44、-0.41、-0.41、-0.39、-0.39、-0.38，存在多项负相关性较强的指标。

综上，我们先对预处理过的数据进行相关系分析，初步挖掘各指标变量之间的相关关系、大致了解影响用户在语音、上网两方面满意度的因素，为后面具体的满意度影响因素分析打下基础。

第四章 用户分析之满意度影响因素分析

4.1 特征选择方法简介——随机森林

当一个数据集具有充足数量的特征时，如何筛选出对结果影响最大的特征，即缩减建立模型时的特征数是尤为关键的。有效而准确的筛选特征不仅可以提高模型的效率同时也可以提升模型的准确度。用于特征选择的方法有很多，如多元分析方法主成分分析、因子分析等，基于正则化的变量筛选方法 lasso、Adaptive Lasso 等。除此之外，有很多机器学习方法在分类的过程中可以进行特征选择。其中，随机森林算法是最常用的一种特征选择方法。随机森林是以决策树为基础学习器的集成学习算法，在分类和回归上均具有很强的性能，也被誉为“代表集成学习技术水平的方法”。他的基本原理是，从原数据集中有放回的采样获得若干个子集，基于每个子集训练出不同的基分类器，再通过基分类器的投票获得最终的分类结果。

随机森林进行特征重要性评估方法主要两种，分别为 Decrease GINI 和 Decrease Accuracy[1]。

（1） Decrease GINI

对于分类问题，以 Gini 值作为评判标准，定义 Gini 值为：

$$Gini=1-\sum(P(i)*P(i)) \qquad （式 4-1）$$

其中，P(i)为当前节点上数据集中第 i 类样本的比例。一般来说，类别分布越平均，Gini 值越大，类分布越不均匀，Gini 值越小。在寻找最佳的分类特征和阈值时，寻找最佳的特征 f 和阈值 th，使得当前节点的 Gini 值减去左子节点的 Gini 和右子节点的 Gini 值最大。即评判标准为：

$$argmax(Gini-GiniLeft-GiniRight) \qquad （式 4-2）$$

对于回归问题，则更加简单，以 argmax(Var-VarLeft-VarRight)作为评判标准，即当前节点训练集的方差 Var 减去减去左子节点的方差 VarLeft 和右子节点的方差 VarRight 值最大。

（2） Decrease Accuracy

Decrease Accuracy 的基本思想为：如果一个变量 j 足够重要，那么改变它会极大的增加测试误差；反之，如果改变某一变量并没有引起测试误差增大，则说明该变量不是重要变量。

具体而言，对于一棵树 Tb(x)，采用 OOB 样本得到测试误差 1；然后随机改变 OOB 样本的第 j 列：保持其他列不变，对第 j 列进行随机的上下置换，得到测试误差 2。至此，可以用测试误差 1 与测试误差 2 的差值来刻画变量 j 的重要性。

4.2 语音业务满意度影响因素分析

4.2.1 变量重要性量化分析

在数据预处理基础上用随机森林方法对语音业务数据进行特征重要性评估。语音业务数据共含有用户 4 个方面的打分数据，分别为语音通话整体满意度、网络覆盖与信号强度、语音通话清晰度以及语音通话稳定性，每项打分区间为 1-10。经过数据预处理后，语音业务数据集含有 29 个变量，这里仅展示前 25 个变量重要性排序情况（详见图 4-1）。

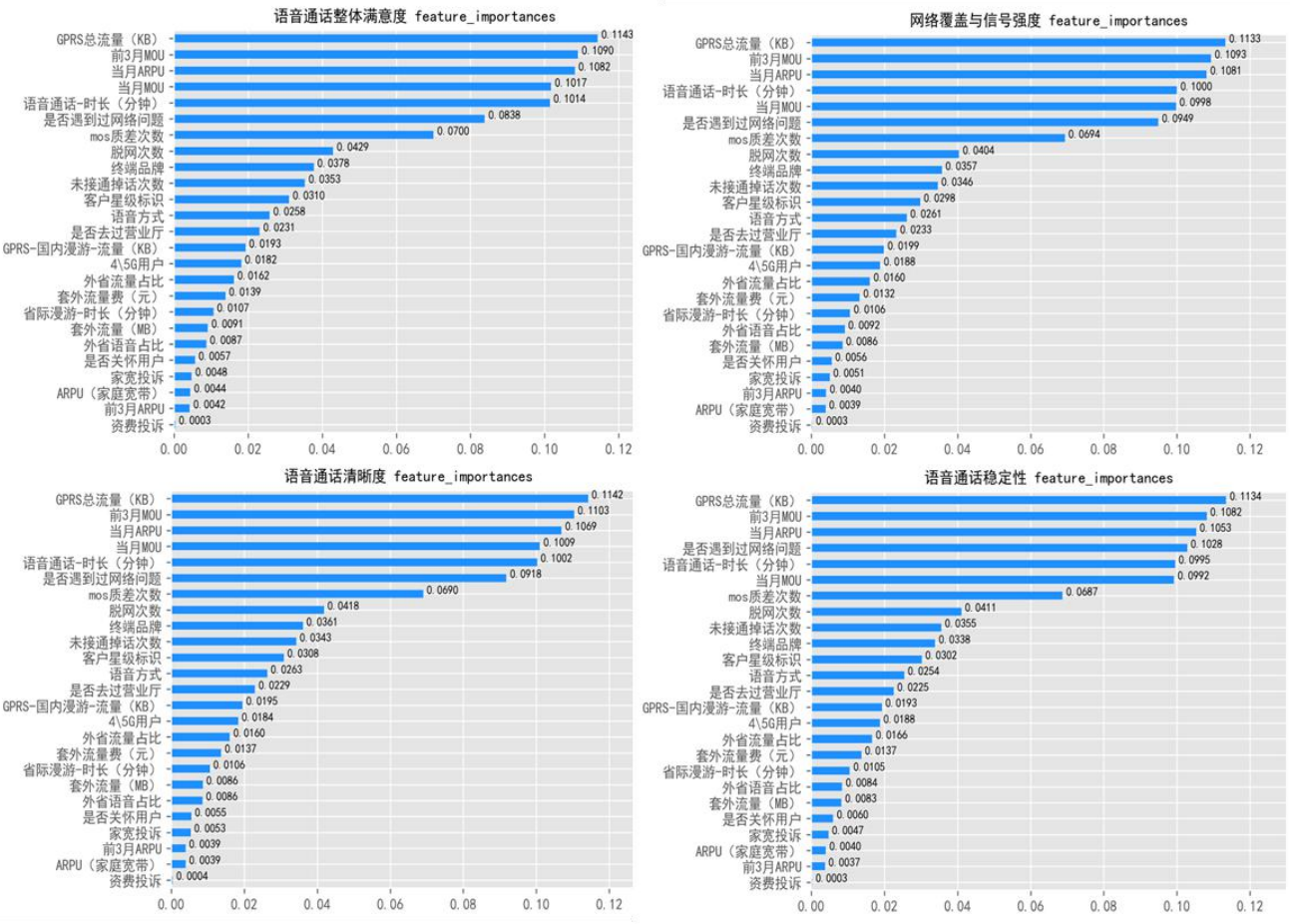


图 4-1 “语音业务”变量重要性排序

基于变量重要性排序结果图，可以得到以下分析结果：

（1）25 个变量对 4 项打分的重要性排序大致相同，即说明 4 个方面的打分基本受到相同变量的影响，这也导致了 4 各方面的打分呈现较强的相关性。

（2）最重要的三个变量为：GPRS 总流量、前三个月 MOU（前三个月语音通话时长）、当月 ARPU（家庭宽带）费用，这三个变量在 4 项打分重要性排序中均位于前三名，重要程度均大于 10%，在后期的分类预测中可能发挥很大作用；最不重要变量为资费投诉，这一变量对打分预测而言为无关变量。

（3）比较重要的变量为：是否遇到过与网络问题、语音通话时长、当月 MOU，这三个变量重要性程度为 10%左右，其次是质差次数、脱网次数，重要程度在 5%左右,这些变量对打分有很强的影响，可以作为打分预测的重要特征。

（4）不太重要的变量还有：ARPU (家庭宽带)套餐费用、前三个月 ARPU 费用、家宽投诉次数、是否为关怀用户等，这些变量对打分影响不大，在进行打分预测时可以剔除。

4.2.2 部分重要变量分析

选取语音通话整体满意度、网络覆盖与信号强度、语音通话清晰度、语音通话稳定性四项打分重要性最强的几个因素，同时将打分分成低（1-4 分）、中（5-7）、高（8-10）三个等级，进行排序分组，绘制直方图，进行重要变量重要性分析，最终选取对得分影响最为显著的因素前 3 月 MOU 即当月前面三个 3 月的通话时长进行展示和分析。

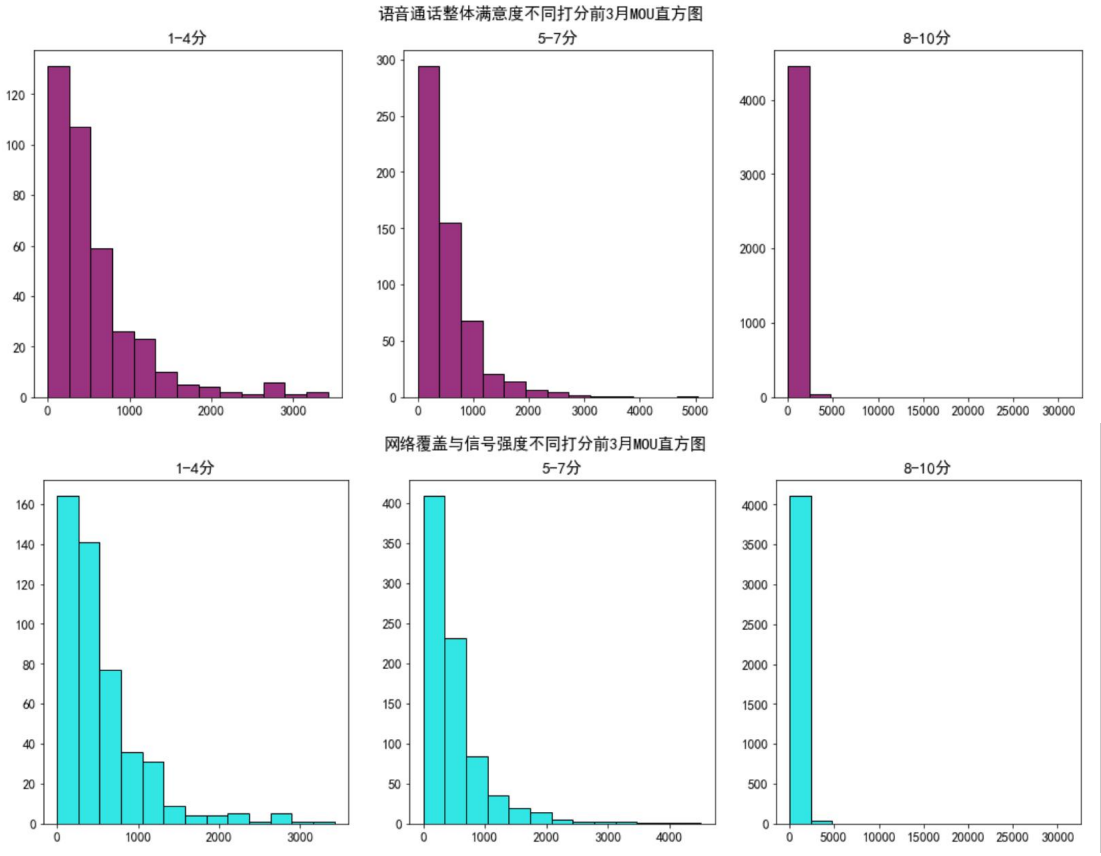


图 4-2 语音通话整体满意度、网络覆盖与信号强度前 3 月 MOU 直方图

通过图 4-2 我们可以看出，语音通话整体满意度、网络覆盖与信号强度前 3 月 MOU 的分布趋势较为一致，打分在 1-4 分、5-7 分、8-10 分之间的用户，其前 3 月 MOU 分别主要分布在 0-800 分钟、客户 0-1200 分钟、0-2500 分钟，三个等级的人数越来越高、前 3 月 MOU 逐渐增加。由此可以推断出，前 3 月 MOU 对用户在语音通话整体满意度、网络覆盖与信号强度的打分有显著影响，前 3 月 MOU 越大的用户，越偏向于较高的打分。

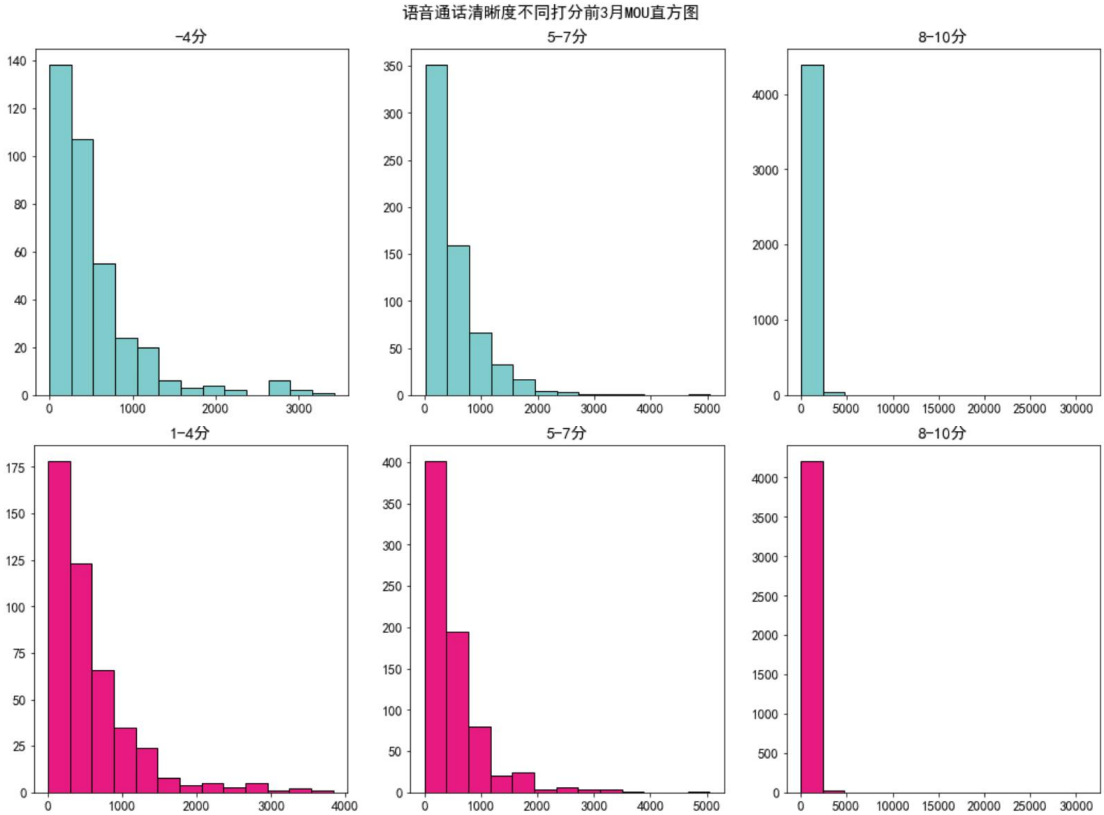


图 4-3 语音通话清晰度、语音通话稳定性前 3 月 MOU 直方图

同理，语音通话清晰度、语音通话稳定性前 3 月 MOU 的分布也是如此，同样认为前 3 月 MOU 对用户在语音通话清晰度、语音通话稳定性的打分有显著影响，前 3 月 MOU 越大的用户，越偏向于较高的打分。

通过上述分析，可以看出中国移动北京公司的语音通话体系较为完善，网络覆盖与信号强度、语音通话清晰度、语音通话稳定性较好，可以满足通话时间长的用户满意，从而获得该类客户的满意与支持。

4.3 上网业务满意度影响因素分析

4.3.1 变量重要性量化分析

在数据预处理基础上用随机森林方法对上网业务数据进行特征重要性评估。上网业务数据共含有用户 4 个方面的打分数据，分别为手机上网整体满意度、网络覆盖与信号强度、手机上网速度以及手机上网稳定性，每项打分区间为 1-10。经过数据预处理后，上网业务数据含有 49 个变量，这里同样仅展示前 25 个变量重要性排序情况（详见图 4-4）。

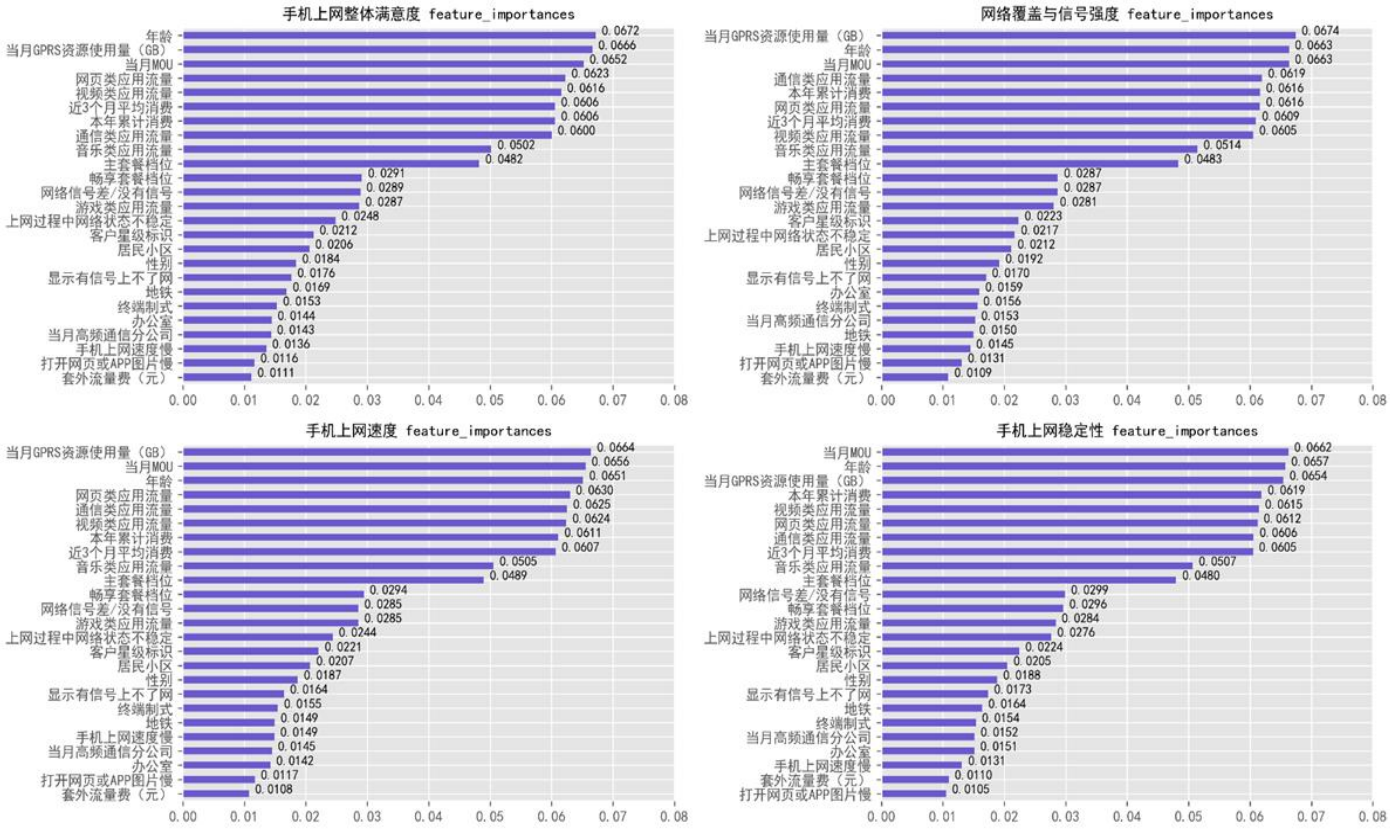


图 4-4 “上网业务”变量重要性排序

基于变量重要性排序结果图，可以得到以下分析结果：

（1）对上网业务 4 项打分而言，前 25 个最重要的变量大致相同，但变量重要性排序有所不同，

（2）对上网整体满意度而言，最重要的前十个变量依次是：年龄、当月 GPRS 资源使用量（GB）、当月 MOU、网页类应用流量、视频类应用流量、近 3 个月平均消费、本年累计消费、通信类应用流量、音乐类应用流量以及主套餐档位，这十个变量的重要程度均超过 4%，前八个变量重要程度超过 6%，这些变量对于手机上网整体满意度预测有很重要的影响；相对而言，不太重要的变量为：套外流量费、打开网络或 APP 图片慢、手机上网速度慢、当月高频通讯公司、是否在办公室遇见网络问题以及终端制式等，

这些变量重要程度小于 1.5%，对整体满意度影响较小。

（3）对网络覆盖与信号强度满意度而言，最重要的前十个变量依次是：当月 GPRS 资源使用量（GB）、年龄、通信类应用流量、本年累计消费、网页类应用流量、近 3 个月平均消费、视频类应用流量、音乐类应用流量以及主套餐档位，这十个变量的重要程度均超过 4%，前八个变量重要程度超过 6%，这些变量对网络覆盖与信号强度满意度预测有很重要的影响；相对而言，不太重要的变量为：套外流量费、打开网络或 APP 图片慢、手机上网速度慢、是否在地铁遇见网络问题、当月高频通讯公司以及终端制式等，这些变量重要程度小于 1.7%，小于平均水平，对整体网络覆盖与信号强度满意度影响较小。

（4）对手机上网速度满意度而言，最重要的前十个变量依次是：当月 GPRS 资源使用量（GB）、当月 MOU、年龄、网页类应用流量、通信类应用流量、视频类应用流量、本年累计消费、近 3 个月平均消费、音乐类应用流量以及主套餐档位，这十个变量的重要程度均超过 4%，前八个变量重要程度超过 6%，这些变量对手机上网速度满意度预测有很重要的影响；相对而言，不太重要的变量为：套外流量费、打开网络或 APP 图片慢、是否在办公室遇见网络问题、当月高频通讯公司等，这些变量重要程度小于 1.5%，小于平均水平，对手机上网速度满意度影响较小。

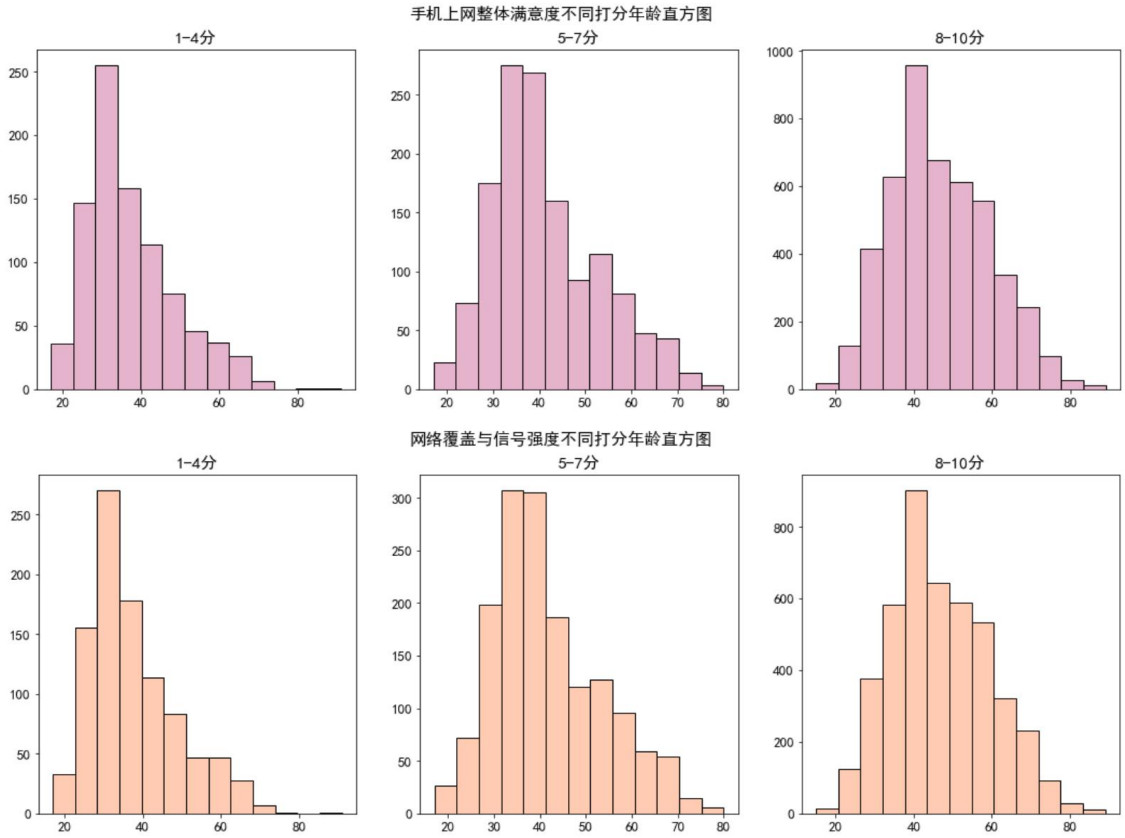
（5）对手机上网稳定性满意度而言，最重要的前十个变量依次是：当月 MOU、年龄、当月 GPRS 资源使用量（GB）、本年累计消费、视频类应用流量、网页类应用流量、通信类应用流量、近 3 个月平均消费、音乐类应用流量以及主套餐档位，这十个变量的重要程度均超过 4%，前八个变量重要程度超过 6%，这些变量对手机上网稳定性预测有很重要的影响；相对而言，不太重要的变量为：打开网络或 APP 图片慢、套外流量费、手机上网速度慢、是否在办公室遇见网络问题、当月高频通讯公司等，这些变量重要程度小于 1.5%，小于平均水平，对手机上网稳定性满意度影响较小。

4.3.2 部分重要变量分析

选取手机上网整体满意度、网络覆盖与信号强度、手机上网速度以及手机上网稳定性四项打分重要性最强的几个因素，同样将打分分成低（1-4 分）、中（5-7）、高（8-10）三个等级，进行排序分组，绘制直方图，进行重要变量重要性分析，最终选取对得分影响最为显著的因素年龄进行展示和分析。

通过手机上网整体满意度、网络覆盖与信号强度、手机上网速度以及手机上网稳定性的年龄直方图（下图 4-5）,我们可以看出，手机上网整体满意度、网络覆盖与信号强

度、手机上网速度以及手机上网稳定性年龄的分布趋势较为一致，打分在 1-4 分、5-7 分、8-10 分之间的用户，其年龄分别主要分布在 22-38 岁、32-42 岁、38-60 岁，低分的群体中，青年人的比例较高，高分群体中，中老年人比例较高。由此可以推断出，年龄对用户在手机上网整体满意度、网络覆盖与信号强度的打分有显著影响，年龄越大的用户，越偏向于较高的打分。



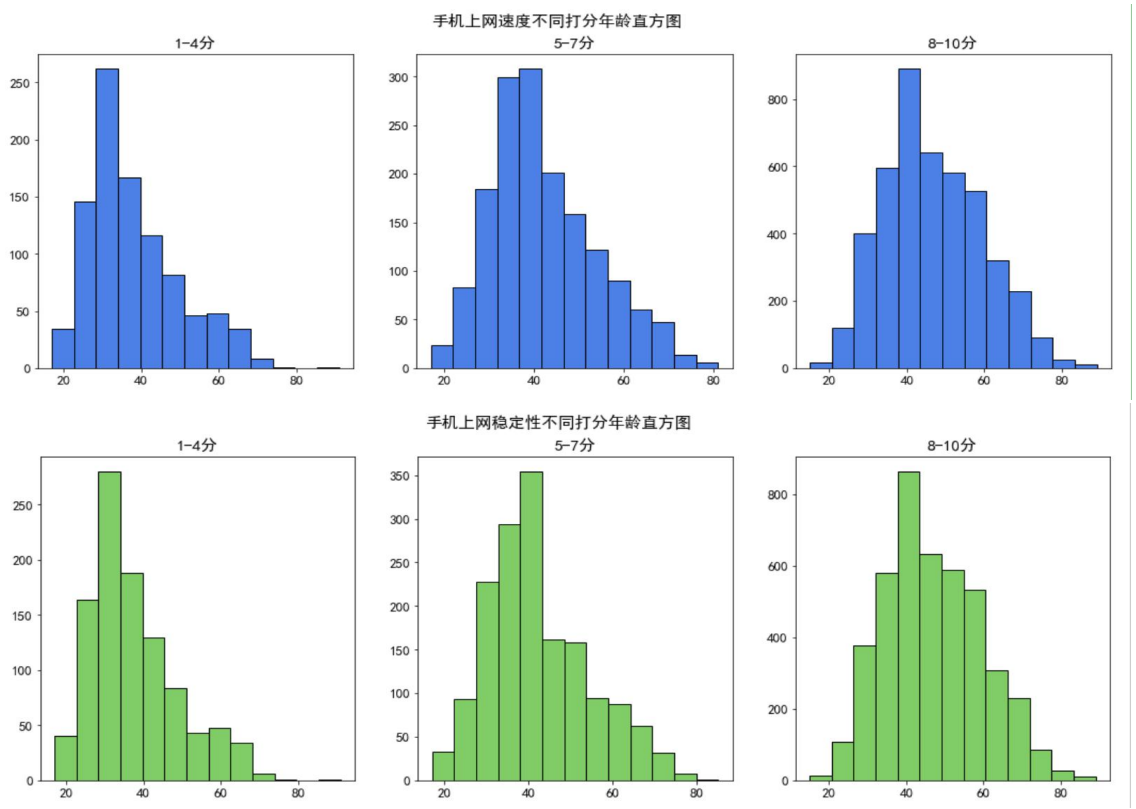


图 4-5 上网打分的年龄直方图

通过上述分析，可以看出相比于语音通话体系，中国移动北京公司的网络体系略有不足，青年群体的在手机上网整体满意度、网络覆盖与信号强度、手机上网速度以及手机上网稳定性的打分普遍低于中老年人。造成该现象的原因主要有，相较于中老年人，年轻人的上网时长更多、使用的 APP 数量更多更复杂、网络出现问题的概率更高，这会大大降低用户对手机上网的体验感，故青年人在低分中占比较大。

第五章 用户分析之满意度预测

5.1 分类方法简介

5.1.1 决策树

决策树是一个树结构。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果[2]。

一棵决策树的生成过程主要分为以下 3 个部分：

（1）特征选择：特征选择是指从训练数据中众多的特征中选择一个特征作为当前节点的分裂标准，如何选择特征有着很多不同量化评估标准标准，从而衍生出不同的决策树算法。

（2）决策树生成： 根据选择的特征评估标准，从上至下递归地生成子节点，直到数据集不可分则停止决策树停止生长。 树结构来说，递归结构是最容易理解的方式。

（3）剪枝：决策树容易过拟合，一般来需要剪枝，缩小树结构规模、缓解过拟合。剪枝技术有预剪枝和后剪枝两种。

决策树适用于数值型和标称型（离散型数据，变量的结果只在有限目标集中取值），能够读取数据集合，提取一些列数据中蕴含的规则。在分类问题中使用决策树模型有很多的优点，决策树计算复杂度不高、便于使用、而且高效，决策树可处理具有不相关特征的数据、可很容易地构造出易于理解的规则，而规则通常易于解释和理解。

5.1.2 随机森林

随机森林是一种基于决策树的集成学习算法，它由多棵决策树组成，每棵决策树具有一个投票权，随机森林最后的分类结果就取决于这些决策树的投票结果。随机森林中每棵决策树的生成方式如下[2]:

（1）从训练集中用有放回抽样法随机抽取一定数量的样本组成新训练集;

（2）在特征集中随机选择一定数量的特征;

（3）利用建立决策树的方法，依次计算最优的特征作为当前节点进行建树。

上述过程重复执行 n 次，每次抽取的样本数量、特征数量都固定不变，建立的 n 棵树就组成了随机森林。由于每次对样本和特征都是随机抽取，就会构造出许多不同的决

策树。因此，随机森林中的所有决策树对同一个样本进行分类，会得到许多分类结果，每棵决策树对其分类结果进行一次投票，通常取被投次数最多的类作为该样本所属的类。

由于随机森林选取训练样本和特征的随机性，以及多次对测试样本进行分类的特性，应用此算法进行分类可以较好地解决单棵决策树泛化能力弱的问题，一定程度上弥补了决策树容易出现过拟合的缺陷。

5.1.3 Light-GBM

Light-GBM（Light Gradient Boosting Machine）是一款基于决策树算法的分布式梯度提升框架，可以实现 GBDT(Gradient Boosting Decision Tree)算法，是一种可以进行快速高效、低内存占用、高准确度、支持并行和快速处理海量数据处理的数据科学工具[3]。GBDT 的主要思想是利用弱分类器迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT 不仅在工业界应用广泛，通常被用于多分类、点击率预测、搜索排序等任务。

然而 GBDT 在每一次迭代时，都需要遍历整个训练数据多次。若将整个训练数据装进内存，则会限制训练数据的大小；不装进内存，反复地读写训练数据又会消耗非常大的时间。在 Light-GBM 提出之前，XGBoost 是 GBDT 的有利工具，它是基于预排序方法的决策树算法。但其存在空间消耗大、在遍历每一个分割点的时候，都需要进行分裂增益的计算，消耗的代价大、对 cache 优化不友好等缺点。因此，Light-GBM 提出的主要原因就是为了解决 GBDT 在海量数据遇到的问题，让 GBDT 可以更好更快地用于工业实践。它的设计思路主要为：

（1）减小数据对内存的使用，保证单个机器在不牺牲速度的情况下，尽可能地用上更多的数据。

（2）减小通信的代价，提升多机并行时的效率，实现在计算上的线性加速。

5.1.4 交叉验证

交叉验证的基本思想是将原始数据进行分组，一部分做为训练集来训练模型，另一部分做为测试集来评价模型。k 折交叉验证一般步骤为将数据集分成 k 份，轮流将其中 k-1 份作为训练数据，1 份作为测试数据，进行试验。每次试验都会得出相应的正确率（或差错率）。k 次的结果的正确率（或差错率）的平均值作为对算法精度的估计，一般还需要进行多次交叉验证，再求其均值，作为对算法准确性的估计[4]。

相比较一般划分训练集和测试集的方法，交叉验证的优点表现在：

（1）交叉验证用于评估模型的预测性能，尤其是训练好的模型在新数据上的表现，可以在一定程度上减小过拟合。

（2）可以从有限的数据中获取尽可能多的有效信息。

5.1.5 网格搜索

为了最大化模型性能，了解如何优化超参数至关重要。人为手动调参依赖大量的经验，耗时且性能未必最佳，因此发展出了许多自动化超参数优化方法。网格化寻优（Grid Search）网格化寻优可以说是最基本的超参数优化方法，它是指定参数值的一种穷举搜索方法，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。即将各个参数可能的取值进行排列组合，列出所有可能的组合结果生成“网格”。使用这种技术，只需为所有超参数的可能构建独立的模型，评估每个模型的性能，并选择产生最佳结果的模型和超参数。

5.2 建立预测模型

现共有语音业务和上网业务两个数据集，每个数据集有四个方面的打分情况需要预测，则共需要预测 8 个指标。按照 5.1 方法简介，针对每个预测指标建立 3 个分类预测模型，最终选取预测效果最好的模型对附件 3 和附件 4 进行预测。

建立模型的过程可以概括为以下几个步骤：

（1）处理附件 1 和附件 3 两个文件，仅保留附件 1 和附件 3 共同含有的变量，处理后每个数据集含有 36 个特征；对附件 2 和附件 4 进行同样的操作，处理后附件 2 和附件 4 每个数据集含有 26 个特征。

（2）采用 10 折交叉验证法分别对附件 1 和附件 2 划分训练集和测试集，用训练集训练分类预测模型，并通过测试集模型测试模型性能，同时采用网格搜索法寻找最优参数。

（3）对比模型预测效果，选取预测效果最佳的模型对附件 3、附件 4 进行预测并输出预测结果文件

满意度预测技术路线图下：

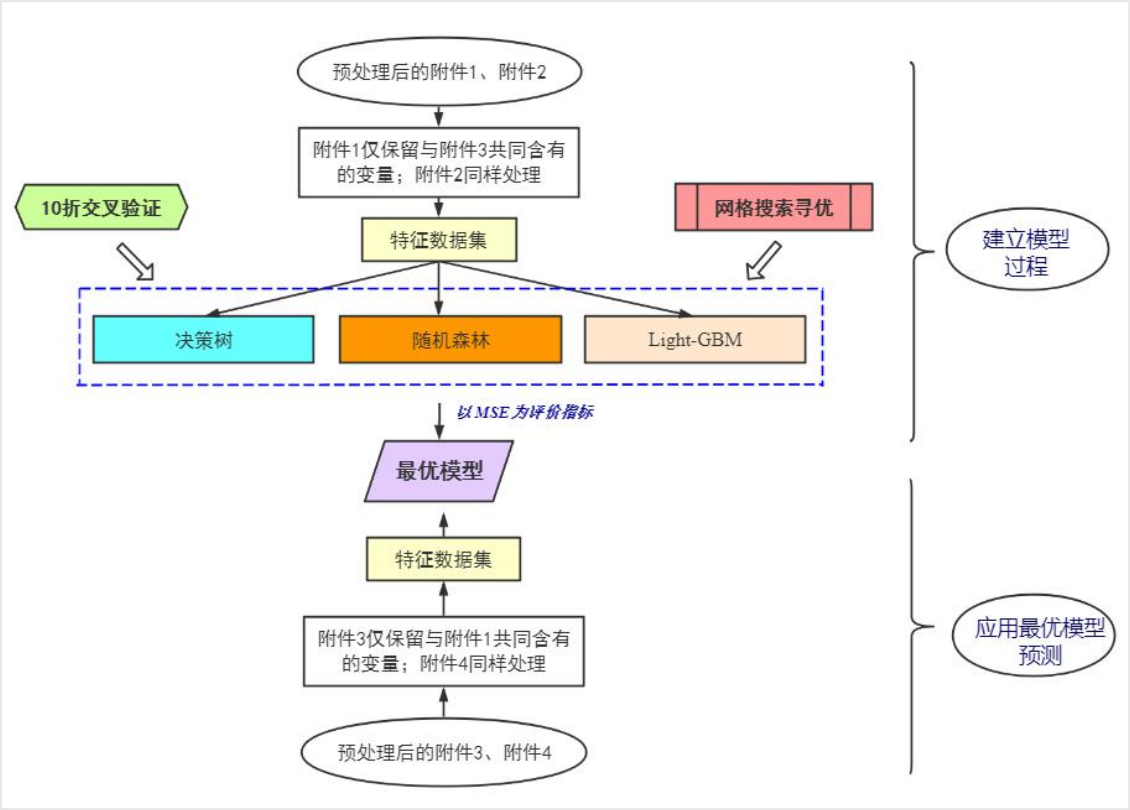


图 5-1 满意度预测技术路线

5.3 结果分析与模型评价

5.3.1 评价指标

对于一个分类预测模型，通常通过混淆矩阵观测预测结果，并计算准确率等指标来评估模型性能，但对于本数据而言，预测指标为满意度评分，在分类预测中，虽然我们将 1-10 评分视为 10 个类别，但是这 10 个类别并不是并列的关系，一般我们认为若将真实评分为 1 错误预测为 10，要比将真实评分为 1 错误预测为 4 更加严重，若采用准确率作为评价指标有失偏颇，故，我们采用均方误差（Mean Squared Error, MSE）作为模型评价指标，定义 MSE 为：

$$\text{MSE}= \frac{1}{n}\sum_{1}^{n}(y_{pred}-y_{test})^2 \qquad \text{（式 5-1）}$$

其中，n 为测试集样本量，y_{pred}为预测评分结果，y_{test}为真实评分结果。

5.3.2 模型评价和结果展示

对于语音业务和上网业务数据集，共 8 个满意度指标，每个指标分别建立决策树，随机森林以及 light-GBM 预测模型，统计模型结果可以得知（详见表 5-1），对于语音

数据集而言，4 个指标预测效果最好的模型均为决策树模型，对于语音整体满意度而言，随机森林 MSE 为 5.945；对于网络覆盖与信号强度，随机森林 MSE 为 6.340；对于语音通话清晰度，随机森林 MSE 为 5.515；最后，对语音通话稳定性预测中，随机森林 MSE 为 6.512，以上随机森林模型最优参数 n_estimators 均为 50。

表 5-1　语音业务数据各模型 MSE

模型	语音通话整体满意度	网络覆盖与信号强度	语音通话清晰度	语音通话稳定性
决策树	6.839	9.347	6.733	6.557
随机森林	**5.945**	**6.340**	**5.515**	**6.512**
Light-GBM	6.492	7.492	6.508	6.887

上网数据模型结果显示，对于手机上网整体满意度而言，决策树的 MSE 最小，为 8.613，最优参数设置为：max_dept= 3, min_samples_leaf= 2, min_samples_split= 2；对于网络覆盖与信号强度，仍然是决策树 MSE 最小（7.543）,最优参数为：max_dept= 4, min_samples_leaf= 2, min_samples_split= 2；对于语音通话清晰度，随机森林 MSE 为 8.491，为三个模型中中的最小值，n_estimators=50；最后，针对手机上网稳定性预测，Light-GBM 效果最好，MSE 为 8.743。

表 5-2　上网业务数据各模型 MSE

模型	手机上网整体满意度	网络覆盖与信号强度	手机上网速度	手机上网稳定性
决策树	**8.613**	**7.543**	9.249	9.556
随机森林	9.195	8.529	**8.491**	8.927
Light-GBM	12.088	8.856	9.421	**8.743**

5.3.3 结果的比较性分析

将 MSE 作为模型评价指标，上述已经对 3 个模型性能进行了比较分析，分别得出 8 个预测指标最佳分类预测模型，比较 8 个指标在各自最佳预测模型下的预测效果（详见图 5-2）。



图 5-2 语音业务和上网业务最优预测效果对比

对比分析结果主要有以下几点：

（1）语音业务数据集的预测整体效果优于上网数据；

（2）针对语音业务数据，预测效果排序依次是：语音通话清晰度 > 语音通话整体满意度 > 网络覆盖与信号强度 > 语音通话稳定性;

（3）针对上网业务数据，预测效果排序依次是：网络覆盖与信号强度 > 手机上网速度 > 手机上网整体满意度 > 手机上网稳定性。

究其原因，可能是因为语音业务数据集选取的特征更好地保留了用户特征，而上网业务数据保留的特征损失掉了部分样本信息，导致预测效果略差于语音数据。

第六章　建议

基于上述对影响用户语音、上网业务满意度因素的分析，本章提出针对性的整改措施，正确处理语音、上网业务短板问题，竭力提升北京移动用户满意度，让企业在在 5G 时代下争取到更市场份额。

（1）注重客户基础通信需求的满足

客户对现阶段语音、网络质量的满意度存在一定不足，用户在日常使用过程中，无论是固网宽带还是移动网，如果出现上网卡看视频加载时间久、拨打电话掉线、有杂音等都是网络质量不好的体现，网量的好与坏会影响到客户对网络甚至运营商的选择，故而需要保证语音信号与上网信号在覆盖过程中具有稳定性与全面性，保证室外信号在客户使用过程中稳定，提高上网速度，提升上网感知。

中国移动通信集团北京公司应在成功构建网络运行体系的必要前提下，持续性进行工作流程的优化，完善工作内容，促进基站整体服务质量实现新提升。工程建设中心应依据城市内及周边县郊区域的人口和网络需求进行尽可能泛的基础网络设施建设，为中国移动实现更广泛更强的信号覆盖而努力。与此同时，还需针对性的对投诉较多的区域进行集中区域信号质量改善，如居民小区、办公室、高校等人口集中区域，把控好网络运行质量，推进网络信号治理，打消消费者对移动网络是否流畅稳定的疑虑，促使消费效果向着最优化方向转变。对客户联网需求和网络故障投诉予以最高效最有效地满足，减少投诉情况的发生，一步步树立好移动网络优质的相关口碑，从而稳定移动存量用户。

（2） 提高投诉处理能力

对客户语音、联网需求和故障投诉予以最高效最有效地满足，减少投诉情况的发生，一步步树立好中国移动通信集团北京公司语音、网络优质的相关口碑，从而稳定移动存量用户。服务端服务工作也会影响客户满意度，警如:网络质量投诉受理等待过程的长短。开展室分测试优化相关工作，通过定期主动走访活动了解重要单位及小区出现的室分系统问题，做到问题及时发现，及时处理,积极推进维护优化工作落实。实施网络优化测试，做好不同线路网络信号监测工作，保证所有基站在运行过程中符合要求，其倾角与方位角都为正确定，进一步提升网络质量，具体措施有:频点调整、下倾角调整、方位角调整及相关设备维护和升级等。

（3） 提升产品类型适配度

首先，基于大数据展开客户使用习惯的深化分析，提高产品开发与客户使用习惯的匹配度,通过引导帮助客户选择更符合自身需求的资费方式，实现公众客户满意度提高。

实践发现客户在产品与资费的选择过程中更倾向于简洁明了的，这个更好的降低双方交易成本，有助于增强客户的使用感受。

其次，为特定人群制定专属产品。现阶段中国老龄化问题较为突出，该类人群常常只需要基础通讯，故而日常产生的资费较低，与此同时，00 后消费正在快速崛起，该类人群会对宽带及流量有较高的需求。教育行业应受到疫情的影响出了更多的线上教育产品，出现了通讯服务需求。通讯公司可以做好人群细分工作，进行特殊产品的针对性制推动企业差异化营销策略，从不同层面进行客户满意度的提升。

参考文献

[1] https://blog.csdn.net/csdn_lzw/article/details/80159735

[2] https://blog.csdn.net/weixin_40449129/article/details/102970339

[3] https://www.jianshu.com/p/104c3e0622d3

[4] https://blog.csdn.net/m0_37805255/article/details/97106652

[5] 张蓓蓓,胡敏.基于网格搜索改进随机森林的顾客满意度预测[J].北京信息科技大学学报(自然科学版),2021,36(04):50-53+58.DOI:10.16508/j.cnki.11-5866/n.2021.04.010.

[6] 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,No.491(03):33-36.DOI:10.13999/j.cnki.scyj.2020.03.011.

[7] 朱晓晨,尹奇志,赵福芹,钱巍文,赵奎奎.基于 LightGBM 的船舶航速预测模型研究[J/OL].大连海事大学学报:1-9[2023-01-19].

[8] 王维虎,刘艳超,杨雷,纪慎思.基于决策树的家庭垃圾分类模型的研究[J].湖北工程学院学报,2022,42(06):68-71.

[9] 胡玉红,张明辉,鲍晓青,孙炜,鲍勇,郭丽君.基于有序 Logistic 回归与决策树的上海市中医家庭康复满意度调查及影响因素分析[J].中国康复理论与实践,2022,28(10):1159-1168.

[7] 颜佳,黄一,王晓娜.基于交叉验证梯度提升决策树的管道腐蚀速率预测[J].腐蚀与防护,2021,42(11):68-74.

[8] 吴菲菲. 中国移动石河子分公司顾客满意度提升策略研究[D].石河子大学,2022.DOI:10.27332/d.cnki.gshzu.2022.001202.

[9] 徐知彬. 基于客户满意度的 H 市联通公司服务质量提升策略研究[D].石家庄铁道大学,2022.DOI:10.27334/d.cnki.gstdy.2022.000466.

[10] 唐忠川. 中国移动 T 市公司客户满意度提升策略研究[D].电子科技大学,2021.DOI:10.27005/d.cnki.gdzku.2021.005255.

队伍编号	MCB2100090
赛道	A

基于多模型 Stacking 融合的二手车定价销售策略研究

摘要

随着中国二手车交易体量日趋增长，一个完善合理的二手车资产价值评估体系亟需构建。但由于二手车具有“一车一况”的特殊属性且新车价格的波动会带来联动影响，二手车的价值衡量标准难以准确设定。本文旨在建立一个二手车资产价值评估的普适性模型，为二手车价格估算提供一种辅助解决方案，并探究缩短门店成交周期的有效手段，为缓和二手车市场信息不对称、完善交易评估体系出一份绵薄之力。

针对问题一，第一步(数据清洗)：本文运用**统计指标、K 近邻算法(KNN)**等填充方法，逐个对附件 1、附件 2 进行缺失值、异常值处理。第二步(特征工程)：本文对部分类别特征数据进行**标签编码**和**独热编码**，然后进行**特征构造**并删除重复信息的字段，同时对部分数值特征进行**分箱**处理。第三步(可视化)：利用 **pandas-profiling** 工具生成研究报告，并通过绘制分布图、热力图、三维柱状图、三维散点图等，对特征变量进行统计解释。第四步(模型准备)：本文依据题干信息及统计学理论，建立并采用**评测标准、调整决定系数R^2、平均相对误差、5%误差准确率**这四项指标作为本文的模型评估标准。第五步(模型建立)：本文先建立**神经网络**模型评判数据集预测情况，然后利用 **XGBoost** 算法、**随机森林**、SFS **序列前向选择**筛选得到 **34 个特征**，据此分别建立 3 个线性回归模型及 7 个树模型，并根据上述 11 个模型的评估结果，决定采用测评表现最优的 **XGBoost** 算法、**LightGBM** 算法、**CatBoost** 算法作为本文的预测模型，同时利用**贝叶斯优化调参**对其进行优化。第六步(模型融合)：为进一步提升模型的预测效果，本文采用 **Stacking 法**将上述选择的 3 个模型进行**模型融合**，其结果表明融合后模型波动更小且预测更为稳定。

针对问题二，本文先在特征工程中对附件 4 数据集进行**特征构造**，同时将其与附件 1 数据集进行拼接处理。通过分析数据集，本文构建车辆的总成交周期由总调价周期和最佳成交周期两个阶段组成的模型，并利用 **XGBoost** 算法判断影响二手车总成交周期的关键因素。针对如何采取行之有效的手段加快门店在库车辆的销售速度，本文决定采用问题一预测出的二手车交易价格作为**指导价**，针对上架价格进行定价处理；同时通过减小调价周期、增大降价幅度、增加调价次数对总调价周期进行干预；针对最佳成交周期，本文先抛弃调价因素，再利用 **XGBoost** 算法进行特征重要性排序，以此得到了多个对最佳成交周期产生显著性影响的特征，并利用**数据可视化**和 **K-Means 聚类分析**探究如何利用这些因素来改善最佳成交周期进而缩减总成交周期。

针对问题三，本文从提高利润率，扩宽客户受众，拓展业务广度，增添附加值业务四个方面入手，通过与互联网公司合作**绘制用户画像**，针对不同地区不同的用户偏好，因地制宜进行展销，并且通过布局二手新能源汽车业务，如二手电池产业，充电桩产业以拓展业务广度，并从二手车检测认证，翻新，质保，保险四个方面增添附加值业务。将有限资源，精准投入到重点领域，增强业绩稳定性，实现良性发展。

关键词：特征工程、神经网络、Boosting 算法、Stacking 融合、K-Means 聚类

目录

一．问题重述 1

1.1 研究背景 1
1.2 研究问题 1

二．问题分析 1

2.1 针对问题一的分析 2
2.2 针对问题二的分析 2
2.3 针对问题三的分析 3

三．原始数据集介绍 4

四．问题一 模型的建立和求解 4

4.1 数据清洗 4
4.2 特征工程 7
4.3 可视化分析 8
4.4 模型准备 10
4.5 神经网络 10
4.6 树模型 11
4.7 模型融合——Stacking 法 19

五．问题二 模型的建立与求解 20

5.1 特征工程 20
5.2 挖掘车辆成交周期的关键因素 23
5.3 寻求加快在库车辆销售速度的有效手段 24

六．问题三 模型的建立与求解 30

6.1 提高利润率 30
6.2 扩宽客户受众 30
6.3 拓宽业务广度 31
6.4 增添附加值业务，增厚利润空间 31

七．参考文献 32

一．问题重述

1.1 研究背景

新冠肺炎疫情的不期而至为全球汽车行业带来严峻挑战。但是，除 2020 年受疫情影响下跌 4%外，中国二手车交易量在过去的 18 年中持续增长。根据中国汽车流通协会统计数据显示，2021 年 1 月至 11 月，中国二手车交易量达 1596 万辆，2021 年全年更是突破了 1700 万辆，同比增长 20%。另外，我国“十四五”规划也深刻影响着汽车产业，加速了二手车“内循环”的构建，促进二手车买卖体系的进一步成熟与完善。除此之外，年度政府工作报告提出“取消对二手车交易的不合理限制”，意在顺畅二手车交易的脉络，鼓励竞争，反对垄断。

纵然十几年间年交易量基本居于增长态势，但由于二手车“一车一况”的特殊属性及新车价格的波动会带来联动影响，我国目前尚未出台有关二手车资产价值衡量的相关政策。因此，市场上二手车价值评估往往来源于二手车交易平台或第三方估价平台。此外，线上收集用户信息结合线下门店展示销售的 020 门店模式是当前二手车的主要零售方法，然而尚未完善的资产评估标准体系可能会导致门店的滞销甚至打包批发。

综上，完善二手车资产价值评估体系是打通二手车市场“任督二脉”的关键所在。道阻且长，行则将至。

1.2 研究问题

鉴于以上背景，本文研究解决问题如下：

问题 1：选择估价方法并预测零售价格

依据附件 1“估价训练数据”所提供的训练用二手车交易样本数据及题目文本涉及的其他相关资料，基于机器学习及统计学理论知识，联系结合各特征变量，训练并测试模型，以选择二手车交易的最优估价方法并预测附件 2“估计验证数据”的二手车零售交易价格，最后将预测结果保存在附件 3“估价模型结果”文件中。

问题 2：挖掘车辆成交周期的影响因素并探究加快销售速度的手段

结合附件 4“门店交易训练数据”所提供的车辆销售各阶段的时间数据以及题目文本涉及到的其他相关资料，结合问题一的结论，深入探究挖掘车辆交易周期的影响性因素，并分析加快二手车销售速度的有效方法及其应用前提和预期效果。

问题 3：思考其他研究问题并阐述思路

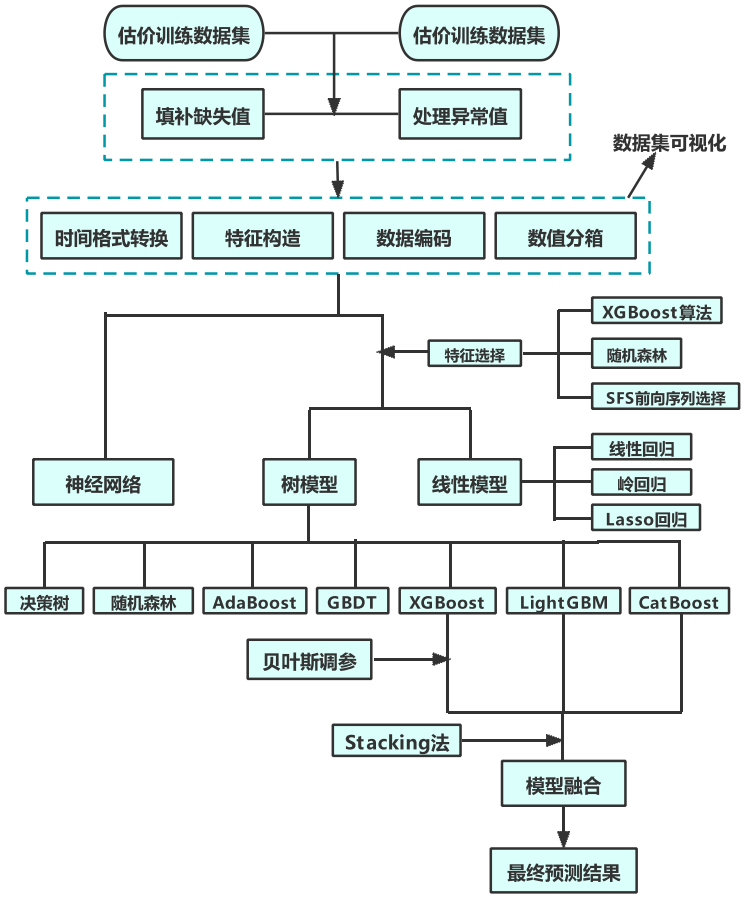
结合题目所提供的数据集及问题一、问题二的研究结果，思考其他值得研究的二手车行业的问题，并阐述相应思路。

二．问题分析

基于重述及附件数据，本文对问题进一步分析如下：

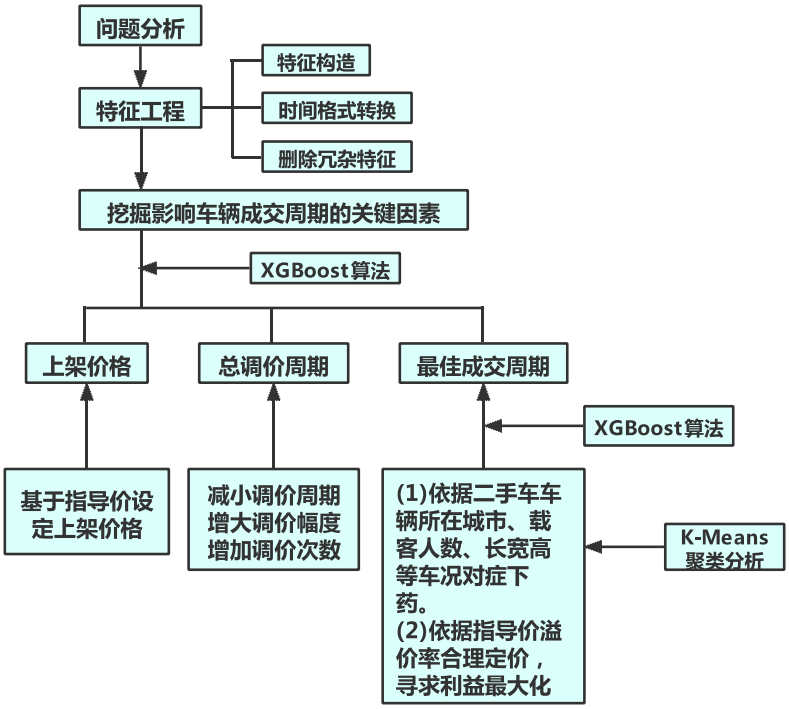
2.1 针对问题一的分析

本文首先对附件 1 与附件 2 的数据集进行预处理，利用统计学手段、K 近邻算法等方法处理原始数据集中的缺失值和异常值，再进行特征工程和数据可视化。此外，本文构建模型的评价指标，以从不同算法中遴选出多个最佳预测模型，并利用 Stacking 法对其进行模型融合，达到进一步稳定预测、减小波动的目的。



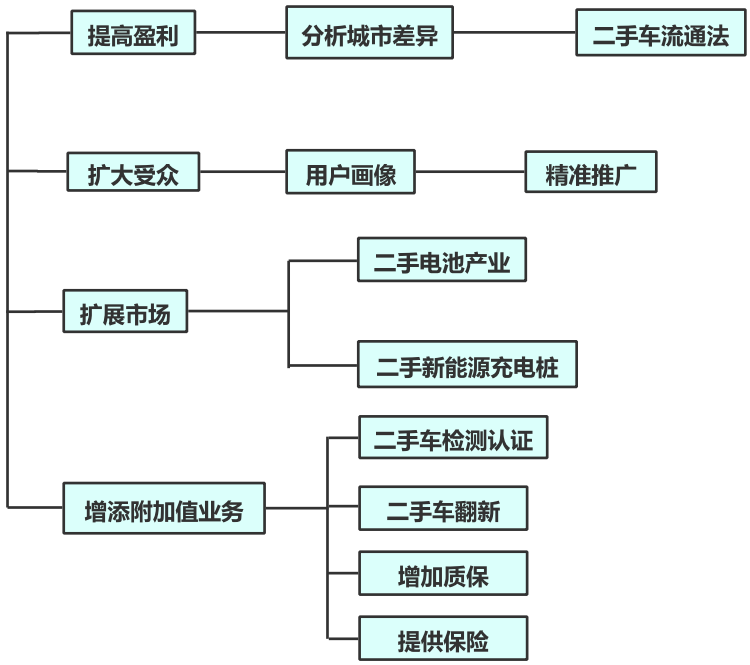
2.2 针对问题二的分析

本文首先对数据集进行特征工程，并利用 XGBoost 算法判断影响二手车总成交周期的关键因素。而针对如何采取行之有效的手段加快门店在库车辆的销售速度，本文以问题一预测得到的二手车交易价格作为指导价，对上架价格进行定价处理；同时通过减小调价周期、增大降价幅度、增加调价次数对总调价周期进行干预；针对最佳成交周期，本文利用 XGBoost 算法进行特征重要性排序，以此获得多个对最佳成交周期产生显著性影响的特征，并利用数据可视化和 K-Means 聚类分析探究如何利用这些因素来改善最佳成交周期进而缩减总成交周期。



2.3 针对问题三的分析

本文从提高利润率，扩宽客户受众，拓展业务广度，增添附加值业务四个方面入手，通过与互联网公司合作绘制用户画像，针对不同地区不同的用户偏好，因地制宜进行展销，并且通过布局二手新能源汽车业务，如二手电池产业，充电桩产业以拓展业务广度，并从二手车检测认证，翻新，质保，保险四个方面增添附加值业务，从而将有限资源，精准投入到重点领域，增强业绩稳定性，实现良性发展。



三．原始数据集介绍

附件 1：原始数据集由 30000 条来源于二手车交易平台的训练用样本数据组成。该数据集中涉及的二手车展销时间跨度长达 19 个月，自 2020 年 1 月 1 日至 2021 年 7 月 31 日。此外，本原始数据集一共包含了 36 个字段，其中 20 列字段主要为二手车车辆基础配置信息、交易时间及价格信息等，15 列字段为匿名特征，另有 1 列字段为本文预测目标——二手车交易价格估值。

附件 2：原始数据集由 5000 条来源于二手车交易平台的验证用样本数据组成。该数据集中涉及的二手车展销时间跨度为 5 个月，自 2021 年 8 月 1 日至 2021 年 12 月 7 日。此外，本原始数据集一共包含了 35 个字段，其中 20 列字段主要为二手车车辆基础配置信息、交易时间及价格信息等，15 列字段为匿名特征。

附件 3：待填充预测值的数据集。

附件 4：原始数据集由 10000 条门店交易数据组成。该数据集一共有 6 个字段，除车型 id 以外，其他字段均与二手车交易过程中各阶段时间及价格有关。

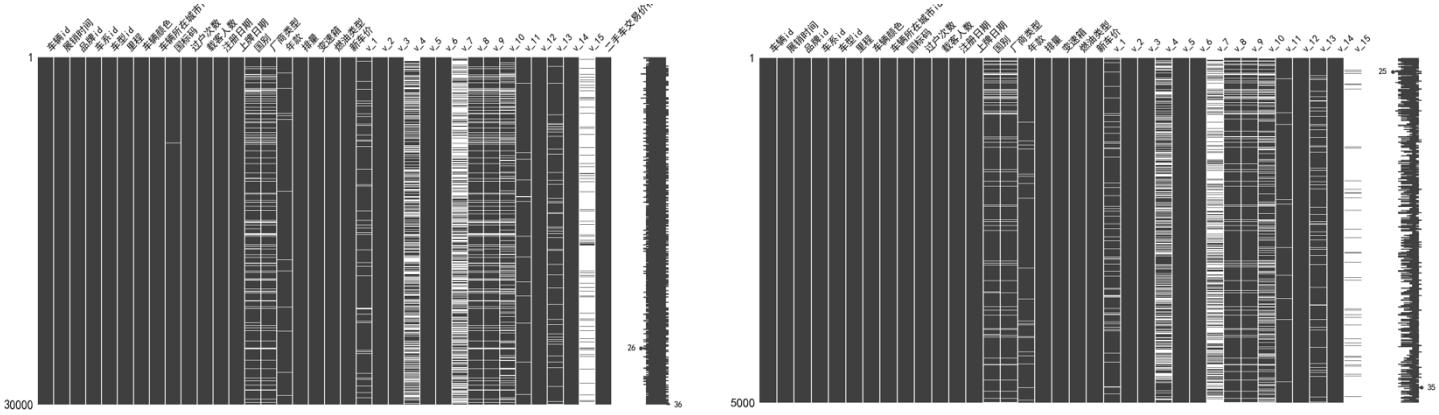
四．问题一 模型的建立和求解

4.1 数据清洗

附件 1、附件 2 原始数据集中，介于时间跨度长、基础配置信息纷繁复杂、车辆品牌型号涉及较广等因素，可能会导致原始数据集出现缺失值、异常值等现象，影响本文数据挖掘及分析的结果，因此本文将先对原始数据集进行初步的数据清洗。

4.1.1 查找缺失值并计算占比

首先，本文先进行缺失值可视化分析，其中，空白值越多代表缺失情况越严重。分别绘制附件 1、附件 2 的缺失值数据密度图如下：



由二图可知，匿名变量 7 及匿名变量 15 空白值明显较多，此外半数的特征变量存在缺失值的情况。为进一步定量处理缺失问题，本文分别查找计算了原始数据集的各字段缺失及其占比情况，如下表所示：

字段名	缺失值数量		缺失值占比	
数据集	训练集	验证集	训练集	验证集

国标码	9	0	0.0300%	0
国别	3757	396	12.5233%	7.92%
厂商类型	3641	375	12.1367%	7.50%
年款	312	106	1.0400%	2.12%
变速箱	1	0	0.0033%	0
匿名变量 1	1582	340	5.2733%	6.80%
匿名变量 4	12108	1863	40.3600%	37.26%
匿名变量 7	18044	3315	60.1467%	66.30%
匿名变量 8	3775	416	12.5833%	8.32%
匿名变量 9	3744	413	12.4800%	8.26%
匿名变量 10	6241	1231	20.8033%	24.62%
匿名变量 11	461	73	1.5367%	1.46%
匿名变量 12	0	1	0	0.02%
匿名变量 13	1619	260	5.3967%	5.20%
匿名变量 15	27580	4719	91.9333%	94.38%

易知，有 15 个字段存在缺失值，其中匿名变量 15 的缺失值占比高于 90.00%，该字段所含缺失值过多而丧失填补价值，因此本文将其剔除并不进行后续数据分析。

4.1.2 处理国标码字段缺失值及异常值

根据国家质量监督检验检疫总局发布的规定，国标码取决于汽车型号、汽车使用燃料及汽车排气污染物排放限值。由此，同一车型 id 对应的载客人数、排量及燃油类型相同，相应的，其国标码也应当相同。所以，在不将 NaN 计入总数的情况下，本文通过分别计算同一车型 id 各国标码所含样本数目，经比较得其众数作为填充值，以对缺失值及异常值进行替换处理。

4.1.3 处理国别字段缺失值及异常值

易知，同一品牌 id 对应同一国别（如，宝马品牌仅对应德国）。本文以品牌 id 为依据进行分组可知，国别字段同时存在缺失值及异常值。对此，在不将 NaN 计入总数的情况下，本文通过分别计算同一品牌 id 各国别所含样本数目，经比较得其众数作为填充值，对缺失值及异常值进行替换处理。

此外，本文分别基于以下两种情况进行特殊填补处理：

1) 当某品牌 id 对应国别均为空值时，则将其填补为所有国别字段的众数，即为 779412。
2) 已知国别区间为 779411-779419 及 779421，而品牌 id 为 37 的字段对应国别均为 0，故将其填补为 779420。

此外，介于国别变量是一个类别特征，而考虑到较大数值对于后续模型的影响，本文利用 Label Encoder 对其进行编码。

4.1.4 处理厂商类型字段缺失值及异常值

检索数据集可知，同一品牌 id 对应厂商类型相同。对此，在不将 NaN 计入总数的情况下，本文通过分别计算同一品牌 id 各厂商类别所含样本数目，经比较得其众数作为填充值，对缺失值及异常值进行替换处理。

特殊的，若某品牌 id 对应厂商类别均为空值，则将其填补为所有厂商类型字段的众数，即为 2。

4.1.5 处理年款字段缺失值

易知，展销时间、注册时间、上牌时间这 3 个字段均为 object 格式，为后续展开比较和计算，本文利用 pandas 库中的方法 dt 对其进行高效操作，将上述 3 个字段转化为 datetime64 格式。

据此检索发现，年款基本居于注册日期与上牌日期附近。因此，本文决定以注册日期的年份填补年款缺失值，而数据集原本的已填年款不做改动。

由于年款变量也是一个类别特征，所以本文同样利用 Label Encoder 对其进行编码处理。

4.1.6 处理变速箱字段缺失值

由于变速箱字段仅存在 1 个缺失值，且该缺失值对应的车系 id 为 459。经检验，车系 id 为 459 的所有记录对应变速箱均为 15，因此本文填补缺失值为 15。

4.1.7 处理匿名变量 7

本文观察数据集易知匿名变量 7 为时间变量，过户次数和匿名变量 7 存在联系，且匿名变量 7 的时间均先于展销时间，据此本文猜测匿名变量 7 为前一次过户时间。通过进一步观察，本文对此归结了以下两种异常情况并作出处理：

1) 当过户次数为 0 时，几乎所有匿名变量 7 为空值。对此，本文将过户次数为 0 的所有记录处理为空值。
2) 当过户次数不为 0 时，仅存在 5 个变量不为空值。对此，本文将这 5 条记录的过户次数处理为 0。

4.1.8 处理匿名变量 11

由于难以判断匿名变量 11 的特征类型，因此本文用该匿名变量的众数对其进行填补。

4.1.9 处理匿名变量 12

易知匿名变量 12 是车辆的长宽高数据，而附件 2 中该特征变量存在 1 个缺失值，其对应的车型 id 为 1866，而附件 1、附件 2 中不存在其他车型 id 为 1866 的记录。因此本文根据该缺失值的车系 id4，选取车系 id4 对应车辆长宽高数据的众数来进行缺失值填充。

4.1.10 处理匿名变量 13

易知，匿名变量 13 的时间基本在注册时间之前，且相差不超过 5 个月，本文猜测该变量为出厂日期。由于该特征的信息解释和注册时间重复性及相关性较高，因此本文仅保留注册时间，对匿名变量 13 进行删除处理。

4.1.11 处理匿名变量 1、4、8、9、10

通过观察数据集，匿名变量 1、4、8、9、10 均为类别变量，因此本文采取 K 近邻(K-NN)算法对样本进行判别归类。K 近邻主要思想为少数服从多数，寻找待分类样本周围最近的 K 个已分类样本，若 K 个已分类样本大多数属于类别 A，则将该待分类样本归为类别 A。本文具体步骤如下：

1) 选取特征

本文选取“品牌 id”、“车系 id”、“车型 id”、“国别”、“厂商类型”、“排量”、“变速箱”及“燃油类型”这 8 个字段作为特征变量。

2) 计算欧式距离

此时本文利用欧式距离公式计算待分类样本点与已分类样本点之间的距离，即不同样本之间的相似程度。

由于存在 8 个特征变量，因此本文假设待分类样本点 A 的坐标为$(A_1, A_2, \cdots, A_8)$，已分类样本点 B 的坐标为$(\mathrm{B}, B_2, \cdots, B_8)$，那么 A、B 两点的欧式距离为：

$$|AB| = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \cdots + (A_8 - B_8)^2}$$

3) 根据 K 值判定类别

计算得到各已分类样本与待分类样本的欧式距离后，即可根据距离远近对已分类样本进行升序排序。通过选取前 K 个已分类样本并核算得到其中的最多类别，从而实现未知样本的归类，完成匿名变量 1、4、8、9、10 的缺失值处理。

4.1.12 处理二手车交易价格异常值

根据本文的模型假设 1，认定实际保值率大于等于 150%为异常值，据此查找计算二手车交易价格的异常值占比大约为 0.033%。例如车辆 id 为 49878 的车辆新车价格为 15.28 万，而其二手车交易价格为 109000 万，明显为异常值，本文猜测是录入时小数点缺失导致。对于一般汽车而言，二手车交易价格往往低于新车价格，因此本文将小数点前移 4 位，修改其价格为 10.90 万元。其他异常值同理。

4.2 特征工程

4.2.1 特征构造

步骤一：构造特征——使用时长、使用年限

本文利用“展销时间”减去“上牌时间”得到新特征“使用时长”，再除以 365 天并调用 Math 库向上取整得到新特征“使用年限”。

步骤二：构造特征——距离上次过户时长

本文利用展销时间减去匿名变量 7 得到新特征“距离上次过户时长”，其中没有过户次数的记录统一填充为 0。

步骤三：构造特征——交易淡旺季

本文根据展销时间择选得到，每年的 2、3 月份及 6、7、8 月份是二手车交易的淡季，而年初、年末及 9 月份为旺季。

结合资料，淡季可能是由于 2、3 月份正值春节或春节刚过，汽车行业适才开业；而 6、7、8 月渐入夏季，炎热的天气导致人们外出频率的降低，且暑假的来临让学生家庭接送孩子的需求也相应下降，可能因此延后购车。而旺季可能是因为年初、年末人们日常薪资与年终奖到位，未来消费预期也比较乐观，且此时正值春运，购入汽车回家便捷；而 9 月学生开学，学生家庭购车需求增强，且气温稍降，人们日常出行频率变高。

步骤四：构造特征——实际保值率、实际折旧率、年实际折旧率、每公里实际折旧率

除此之外，本文仅对附件 1 训练集构造了上述 4 个特征，其计算公式分别如下：

$$\begin{cases} \text{实际保值率} = \dfrac{\text{二手车交易价格}}{\text{新车价}} \\ \text{实际新车价} = \dfrac{\text{新车价} - \text{二手车交易价格}}{\text{新车价}} \\ \text{年实际折旧率} = \dfrac{\text{折旧率}}{\text{使用年限}} \\ \text{每公里实际折旧率} = \dfrac{\text{折旧率}}{\text{里程}} \quad （若 > 1 \text{ 则计作 } 1） \end{cases}$$

步骤五：特征构造——v_11_n

本文将匿名变量 11 进行独热编码，为去除编码后“匿名变量 11”变量名中的“+”，本文对该变量重新命名如下：

原变量	重命名后变量
v_11_1	v_11_1
v_11_1+2	v_11_2
v_11_1+2,4+2	v_11_3

v_11_1,3+2	v_11_4
v_11_3+2	v_11_5
v_11_5	v_11_6

步骤六：特征构造——长、宽、高

观察易知匿名变量 12 是一个三维数据，本文猜测这是车辆的长宽高，因此本文将其拆分为“长”、“宽”、“高”三个特征变量。

步骤七：保存数据集

本文将此时的训练集和验证集分别保存为“估价训练数据整理_可视化.xlsx”和“估价验证数据整理_可视化.xlsx”，用作后续可视化分析使用。

步骤八：删除冗杂特征——展销时间、注册日期、上牌日期、匿名变量 7、12

由于在上述步骤中已充分利用“展销时间”、“注册日期”、“上牌日期”、“匿名变量 7”计算得到了新的时间特征变量，因此为了避免信息的重复解释并对数据集降维，本文对这 4 个特征进行删除处理。

此外，由于“匿名变量 12”被拆分为 3 个新特征，因此该变量也需要进行删除。

步骤九：保存数据集

本文将此时的训练集和验证集分别保存为“估价训练数据整理_NN.xlsx”和“估价验证数据整理_NN.xlsx”，用作后续神经网络模型分析使用。

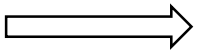
4.2.2 数据转换

步骤一：独热编码(One Hot Encoder)

易知，本数据集中存在多个离散变量，且变量的取值之间不存在大小比较的意义。因此，为了扩充特征作用以解决属性数据的离散值问题，本文对“车辆颜色”、“厂商类型”、“燃油类型”、“交易淡旺季”以及匿名变量 1、2、3、6、8、9、10、14 进行独热编码，将上述 12 种类型的数据存储在垂直空间中。

以“匿名变量 14”的部分样本为例：

车辆 id	匿名变量 14
1	1
2	2
5	2
6	2
7	1
10	1
12	2
……	



车辆 id	匿名变量 14_1	匿名变量 14_2	匿名变量 14_3
1	1	0	0
2	0	1	0
5	0	1	0
6	0	1	0
7	1	0	0
10	1	0	0
12	0	1	0
……			

步骤二：“里程”分箱

本文将“里程”进行降序排列，然后以(0,3]、(3,6]、(6,9]、(9,12]、(12,15]、(15,45]为区间对数据进行分箱处理，再分别对其进行独热编码并命名“里程”标签，结果如下：

里程标签	区间	居于该区间内的样本个数
里程_1	(0,3]	8364
里程_2	(3,6]	7940
里程_3	(6,9]	5135
里程_4	(9,12]	4466
里程_5	(12,15]	2546

里程_6	(15,45]	1549

步骤三：“使用时长”分箱

此外，本文将“使用时长”进行降序排列，然后以(0,1000]、(1000,2000]、(2000,3000]、(3000,4000]、(4000,6000]为区间对数据进行分箱处理，再分别对其进行独热编码并命名“使用时长”标签，结果如下：

使用时长标签	区间	居于该区间内的样本个数
使用时长_1	(0,1000]	6539
使用时长_2	(1000,2000]	10145
使用时长_3	(2000,3000]	7451
使用时长_4	(3000,4000]	4196
使用时长_5	(4000,6000]	1669

步骤四：保存数据集

经上述特征工程，本文构造了 121 个特征变量以供后文中的树模型进行训练，并将此时的训练集和验证集分别保存为“估价训练数据整理_TreeModel.xlsx”和“估价验证数据整理_TreeModel.xlsx”。

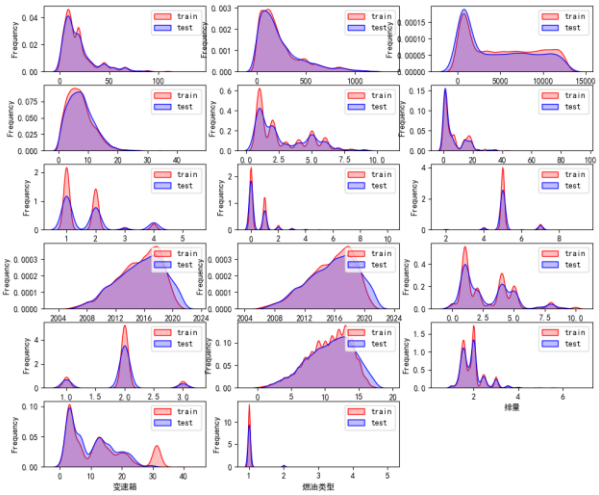
4.3 可视化分析

4.3.1 使用 pandas-profiling 工具生成数据报告(详见“数据报告.html”)

报告显示，经处理后数据集不存在缺失值。对单变量进行分析，里程的频数分布直方图呈正态分布，最大里程数为 44.74 公里；车辆过户次数基本都在 2 次以内；载客人数基本都为 5 人，这意味着小型五人座私家车为二手车交易的主要车型；二手车交易年款以 2016 年款居多，可能是因为 2016 年购入的新车截止到目前为止，里程和使用年限均处于二次交易的理想范围；排量主要分布在 1～2 之间，可能因为国家限制碳排放政策的落实；燃油类型主要为 1 类型，本文猜测 1 类型为汽油，满足绝大部分私家车的燃油需求，2 类型为柴油，主要由客车、货车使用；新车价主要集中在 30 万以内，这是绝大部分家庭所能承受的私家车新车价格；使用年限为 4 年居多，主要分布在 3～5 年之间；车辆的长和宽趋于正态分布；高相对集中在一个明显高峰和一个小高峰，可能是因为绝大部分两厢私家车的高度趋于相同，而一部分高顶车型为一部分人群的受众偏好。

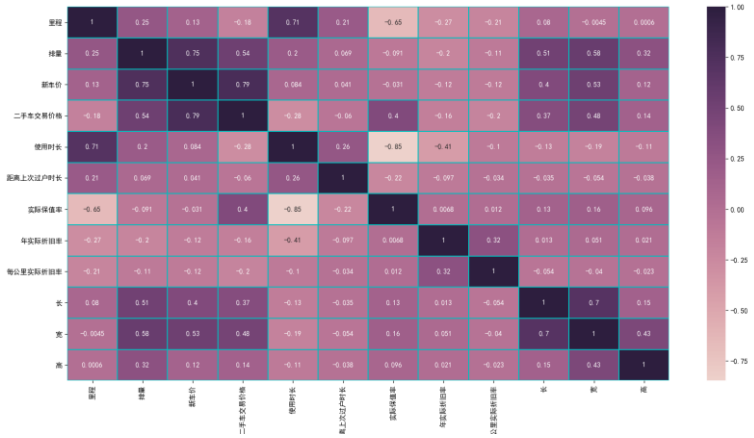
4.3.2 查看数据分布

通过绘制题给训练集与验证集所有特征字段的分布，得出二者分布基本趋于一致，表明后续通过将模型对训练集进行训练并预测验证集是行之有效的策略。



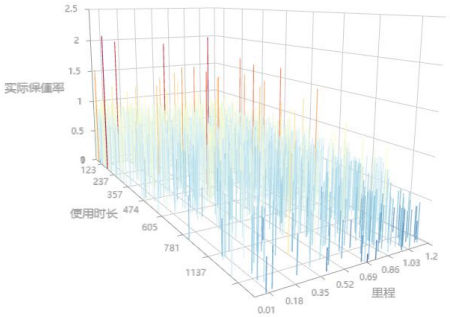
4.3.3 连续数字特征可视化

将二手车的里程、排量、新车价、二手车交易价格、使用时长、距离上次过户时长、实际保值率、年实际折旧率、每公里实际折旧率、长、宽、高这些数字特征绘制热力图如下，同时本文绘制配对网格图见所附文件“配对网格图.png”。

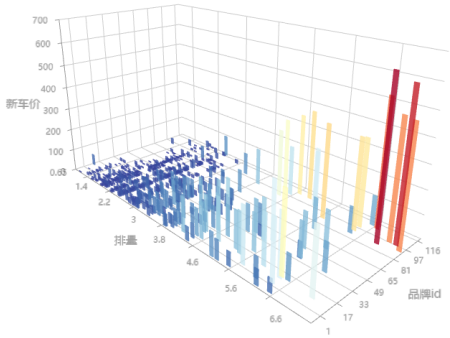


4.3.4 绘制三维柱状图

1) 本文以二手车的使用时长为 X 轴，里程为 Y 轴，实际保值率为 Z 轴绘制了三维柱状图，如右所示。由图可知，随着里程和使用时长的增加，柱体变短，表明二手车的实际保值率相应下降；使用时长和里程接近于 0 的相对较新的二手车，其实际保值率极高。因为实际保值率与二手车资产价值直接挂钩，所以里程和使用时长可能是二手车价格评估的重要因素。

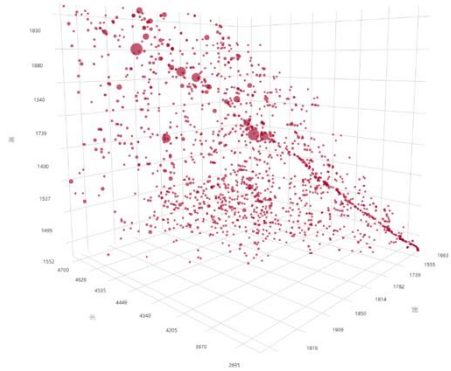


2) 本文以二手车的排量为 X 轴，品牌 id 为 Y 轴，新车价为 Z 轴绘制了三维柱状图，如右所示。由图可知，以同一品牌 id 的车辆为前提，排量越大，代表新车价的柱体越高，故排量与新车价呈正相关，对后续预测二手车价格起到至关重要的作用。



4.3.5 绘制三维散点图

本文以二手车的长为 X 轴，宽为 Y 轴，高为 Z 轴，新车价为散点体积大小，绘制了三维散点图，如右所示。



由图可知，绝大部分二手车的长宽高呈多元线性关系，且新车价较高的样本也基本聚集在该线附近，这意味着二手车体积按相同比例放大或缩小，可能存在一个相对舒适且利用率合理的行车空间。

4.3.6 图片见附件

1) 本文绘制了二手车交易价格预测值的分布情况，将约翰逊分布、正态分布、对数正态分布与预测值进行比对，判断其预测值满足经过约翰变换后服从正态分布概率的随机变量的概率分布。
2) 本文绘制了二手车各特征变量的偏度、峰度图，反映了该统计数据分布的偏斜方向和陡缓程度的度量。

本文将上述图片保存在附件中。

4.4 模型准备

4.4.1 模型评测标准

根据题干，本文建立了模型评测标准如下：

$$0.2 \times (1 - Mape) + 0.8 \times Accuracy_5$$

其中，Ape代表相对误差，公式为：

$$Ape = \frac{|\hat{y} - y|}{y}$$

$Mape$代表相对误差，假设真实值为$y = (y_1, y_2, \cdots, y_m)$, 模型预测值$\hat{y} = (\widehat{y_1}, \widehat{y_2}, \cdots, \widehat{y_m})$，其为公式为：

$$Mape = \frac{\sum_{i=1}^{m} Ape_i}{m}$$

$Accuracy_5$代表 5%误差准确率，其公式为：

$$Accuracy_5 = \frac{count(Ape \leq 0.05)}{count(total)}$$

其中，$count(Ape \leq 0.05)$代表Ape在 5%以内的样本个数，而$count(total)$意味着样本总数量。

4.4.2 调整决定系数R^2

本文运用调整决定系数R^2(Adjusted R-squere)对后文回归模型结果进行评估，其公式如下所示：

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

R为决定系数，其公式为：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

其中 SSR 为回归平方和，SST 为离差平方和，SSE 为残差平方和。

决定系数往往能够用于描述回归模型的拟合优度，但是当样本体量较大时会存在一定程度上的误差，而调整决定系数能够消除样本体量和特征维度带来的误差影响。由于本题的训练集涉及到 30000 条样本记录，因此本文也选择使用调整决定系数来对后文中回归模型进行拟合优度评估。

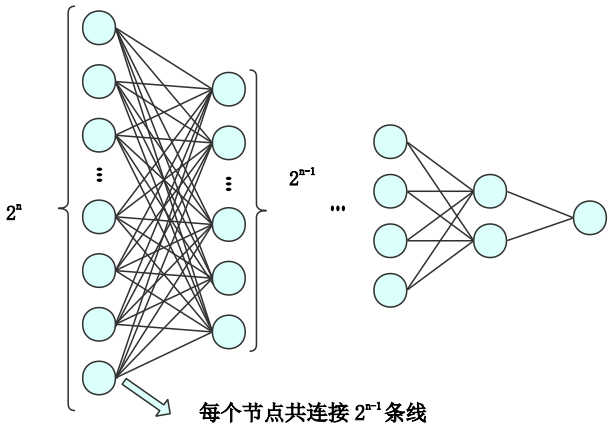
4.5 神经网络

4.5.1 Sequential 模型

神经网络模型的思想来源于模仿人类大脑思考的方式。神经元是神经系统最基本的结构和功能单位，人类大脑在思考时，神经元会接受外部的刺激，当传入的冲动使神经元的电位超过阈值时，神经元就会从抑制转向兴奋，并将信号向下一个神经元传导。神经网络的思想是通过构造人造神经元的方式模拟这一过程。

实际应用中，常常采用多层神经网络，在多层神经网络模型中，输入层和输出层间可以有多层隐藏层，层与层之间互相连接，信号通过线性变换和激活函数的复杂映射，不断地进行传递。

Keras 是一个高层神经网络库。Keras 序列模型可以很容易地拟合回归数据并预测测试数据，Keras 分为两种类型的模型，顺序模型(Sequential)和泛型模型(Model), 我组考虑利用 keras 中的 Sequential 模型对二手车交易价格进行预测。Sequential 通过多个网络层的线性堆叠达到回归目的，具有极好的实用效果。其原理如下图：



4.5.2 模型评价

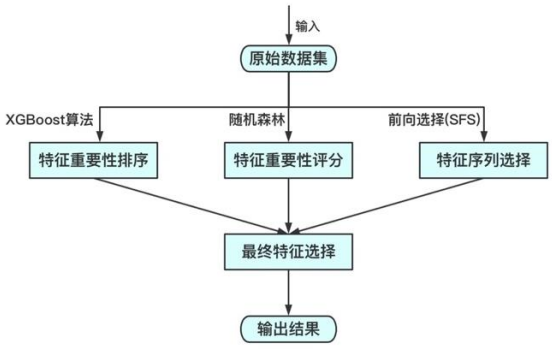
	调整决定系数R^2	平均相对误差	5%误差准确率	评测标准
模型评价	0.93242	0.17508	0.25480	0.36882

由此可知，神经网络平均相对误差高达 15%以上，考虑到二手车交易价格定价的严谨性，本文暂不采纳该神经网络模型。

4.6 树模型

4.6.1 特征选择

为增强模型的泛化能力并减少过拟合，本文将利用包裹式(Wrapper)特征选择法完成对各变量的遴选。相较于过滤式方法(Filter)，包裹式方法能够筛选出规模相对较小的优化特征子集。而机器学习本身具有的特征打分机制，能够被运用到特征选择任务上以达到筛选的目的。本文的特征选择流程如下图所示：



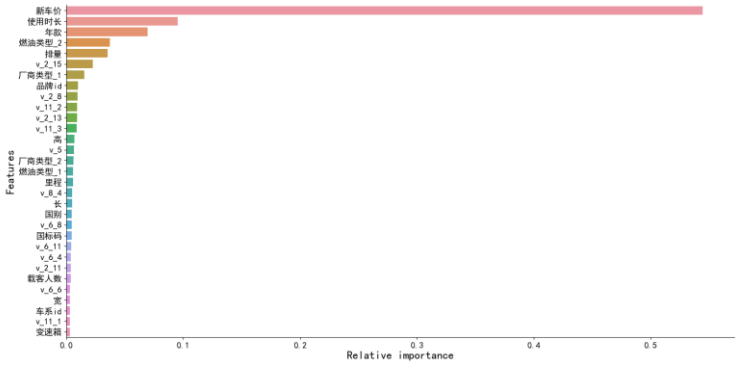
步骤一：特征重要性排序——XGBoost 算法

树模型通常用于机器学习算法自动进行的特征筛选。其获取各个特征重要性分数的核心在于信息增益的计算。信息增益即为某一条件划分下信息不确定性减少的程度，而树模型正是利用信息增益准则来对特征变量进行选择。因此，本文利用树模型中的 XGBoost 算法对特征变量进行遴选，其信息增益公式为：

$$IG(X,Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x) \cdot p(y|x) \cdot logp(y|x) - \sum_{y} p(y) \cdot logp(y)$$

在机器学习中，熵意味着变量不确定的度量。对于上述公式，经验熵$H(Y)$代表数据集Y的不确定性，而经验条件熵$H(Y|X)$代表在特征变量X这一给定条件下数据集Y的不确定性，二者之差即为信息增益，代表特征变量X作用下数据集Y不确定性减少的程度。由于不同特征变量往往能带来不尽相同的信息增益，因此通过计算各特征变量的信息增益并将其排序，即可得到回归能力最优的特征。

本文将特征重要性大于 0.3%的变量绘制成特征重要性分数条形图如下所示：

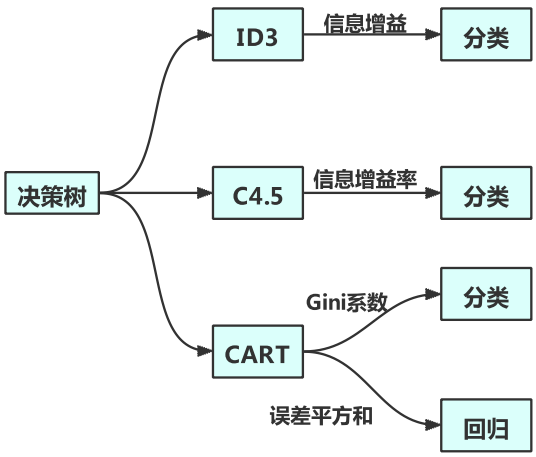


由图可知，新车价是二手车交易价格最重要的特征影响因素，其特征重要性分数高达 54.447%；其次是使用时长，其特征重要性分数为 9.543%，可能是因为除限量车型等特殊情况以外，一般性车辆的硬件设施及行驶品质会随时长增加而下降，其资产价值也相应下滑；年款对二手车零售价格也具有一定影响，其特征重要性分数为 6.963%，可能是因为汽车款式的迭代更新会带来更加新颖和高档的驾驶体验，价格也随之上涨；燃油类型_1 和排量的特征重要性分数分别为 3.735%和 3.523%，这可能与近些年来新能源汽车的市场占有率上涨有所挂钩，且使用汽油的私家车价格差异更为明显（相比较而言，代表柴油的燃油类型_2 特征重要性分数较低，可能是因为使用柴油的客车及货车相较于私家车的社会功能不同导致其二手车价格大同小异），同时，排量对价格的影响可能还来源于政府节能减排政策的落实以及绿色环保节能出行的社会倡议。

步骤二：特征重要性评分——随机森林

随机森林(Random Forest，RF)是一种将多棵互不关联的决策树(Decision Tree)加权集合构成的模型算法，是集成学习中 Bagging 方法的一种，具有随机属性且适用于较高维度的数据集。随机森林的核心在于样本采样和特征采样的随机性，其基本原理是输入待预测的X个原始数据，利用集成的N棵决策树执行判断和分类任务并得到N个分类结果，而随机森林通过集成汇总所有分类结果以最终输出投票次数最多的类别。

而决策树主要分为 ID3、C4.5 以及 CART 三种类型的算法，前两者分别以信息增益和信息增益率解决分类问题，但 CART 算法既能够以基尼系数(Gini Coefficient)为依据应用在分类问题上，也可以用于建立回归树模型。决策树各类型算法如右图所示：



而本文正是基于 CART 算法建立随机森林回归模型，其基本原则为最小均方差(MSE)。CART 回归树是针对连续的目标变量，选取能够最优划分特征变量的值，再通过将数据与该值比较以达到树分裂的目的。其基本思想为：针对随机划分的特征变量A，将其所对应的随机划分点s划分为数据集D_1和D_2，并求出这两个数据集的最小均方差之和的最小值所对应的特征变量及特征值划分点，其表达式为：

$$\underbrace{min}_{A,s}[\underbrace{min}_{c_2}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+\underbrace{min}_{c_2}\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2]$$

其中，c_1是数据集D_1的最小均方差，c_2是数据集D_2的最小均方差。

本文利用基于 CART 算法建立的回归随机森林对所有特征变量进行重要的的打分，其步骤如下：

1、使用训练数据训练模型；

2、计算训练数据在随机森林上依据某一特征变量预测结果和实际结果的决定系数 R 方，记作得分，并使用 k 折交叉验证取该特征变量得分的平均值；

3、遍历训练数据集中的每一列特征变量，在原训练数据集的基础上通过对比得分将对应的特征变量进行筛选操作，然后使用模型得到筛选后特征变量。

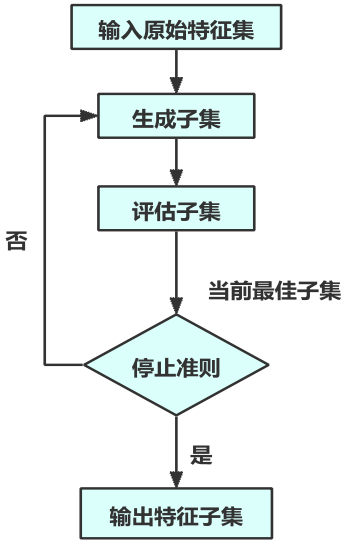
由于篇幅有限，本文选取随机森林打分排名前 10 的特征变量，具体如下表所示：

排名	特征变量名	随机森林打分情况
1	新车价	0.616
2	排量	0.452
3	宽	0.343
4	品牌 id	0.337
5	v_5	0.313
6	变速箱	0.261
7	长	0.248
8	v_11_5	0.166
9	高	0.152
10	国别	0.14

步骤三：特征序列选择——前向选择(SFS)

由于高纬度特征往往存在特征冗余及特征无关的情况，因此当特征维度过高时，通常会导致分类器性能的下降。为了避免上述的“维灾难”，本文将利用特征序列选择（Sequential Feature Selection）来去除冗余特征及地相关性特征的影响，改善精度的同时减少模型训练所需时间。

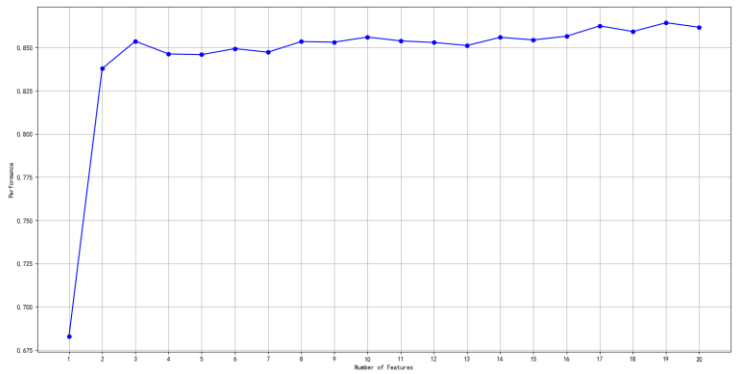
特征序列选择的原理是从初始特征中筛选出高相关性的最优特征以达到降低数据集维度的目的，其基本步骤如右图所示：



此外，特征序列选择主要有以下三种：

特征序列选择种类	具体方法
序列前向选择	向一个空集中逐步添加高相关性最优特征。
序列向后消除	将一个全集中逐步删除低相关性的最差特征。
序列前向选择和后向消除组合	逐步择选高相关性的最优特征的同时，删除剩余特征中低相关性的最差特征。

本文采用了序列前向选择的方法，从 121 个特征变量的原始数据集中遴选出 20 个特征变量，并绘制了特征重要性的边际效应图如下：



步骤四：最终特征选择

本文通过联合比较 XGBoost 算法中特征重要性分数大于 0.3%的变量与随机森林回归模型重要性评分大于 0.03 的变量，取其交集得到二者共同遴选出的特征变量，再将之与步骤三中特征前向选择的结果取并集，并绘制关键词标签云如右图所示，最终筛选得到 34 个特征变量如下：



序号	特征变量名	序号	特征变量名
1	品牌 id	18	车辆颜色_8
2	车系 id	19	厂商类型_1
3	里程	20	燃油类型_1
4	国别	21	v_1_3
5	年款	22	v_2_4
6	排量	23	v_2_8
7	变速箱	24	v_2_11
8	新车价	25	v_2_13
9	v_5	26	v_2_15’
10	使用时长	27	v_3_3
11	使用年限	28	v_3_6
12	交易旺季	29	v_6_7
13	v_11_1	30	v_6_12
14	v_11_3	31	v_8_4

15	长	32	v_10_2
16	宽	33	里程_2
17	高	34	使用时长_3

4.6.2 线性回归模型

步骤一：建立模型

模型 1——线性回归(Linear Regression)

本文采用的线性回归模型是假设目标值与特征变量线性相关的基础上，以误差均方函数为损失函数，并令损失函数最小以确定参数的回归方法。

模型 2——岭回归(Ridge Regression)

岭回归在损失函数中增加了二阶正则项的最小二乘，也叫 L2 范数，具有降维作用的同时，它还以限制模型参数对异常样本的匹配程度，处理相关性高的数据集，进而提高模型面对多数正常样本的拟合精度。我组通过 RidgeCV 针对正则化强度 alpha 进行调参，得出在 alpha 为 14 时可以达到较好的拟合效果。

模型 3——Lasso 回归(Lasso Regression)

Lasso 回归类似于上文提到的岭回归，也通过构造惩罚函数来处理特征变量的共线性问题。但是，相较于岭回归，Lasso 回归可以通过使惩罚项由 L2 范数变为 L1 范数将相对不显著的特征变量系数压缩为 0，达到剔除变量的目的；而岭回归仅仅只在一定程度上压缩特征变量系数且会保留回归模型的全部变量。

步骤二：模型评估

根据上文模型准备中所详细阐述的评测标准和调整决定系数R^2，本文对线性回归模型、岭回归模型、Lasso 回归模型进行评估，其结果如下表所示：

模型	线性回归模型	岭回归模型	Lasso 回归模型
调整决定系数R^2	0.76108	0.76121	0.75069
平均相对误差	0.52198	0.52298	0.53002
5%误差准确率	0.13317	0.13317	0.13317
评测标准	0.20214	0.20194	0.20214

由表可知，三种线性回归模型的调整决定系数R^2均小于 0.8，这意味着模型的拟合优度表现一般；平均相对误差均在 52%左右，表现不佳；而 5%误差准确率基本在 13%左右，相对较高；除此之外，模型的测评标准相对较低，均在 0.2 左右。因为，本文暂不采用线性回归模型。

4.6.3 树模型

步骤一：建立模型

模型 1——决策树模型

决策树已在上文特征选择中有过阐述，当它执行回归任务时，其结构为是否问题的二叉树结构，是从根部节点伊始，通过单一特征变量进行测试，以测试结果为依据将样本划分至子节点。

模型 2——随机森林

随机森林已在上文特征选择中详细阐述。基于 CART 算法建立随机森林回归模型，其基本原则为最小均方差(MSE)。为了保证模型的泛化能力，随机森林在建立每棵树的时候，往往会遵循数据随机、特征随机这两个基本原则。CART 回归树再针对连续的目标变量，选取能够最优划分特征变量的值，并通过将数据与该值比较以达到树分裂的目的。

模型 3——AdaBoost 算法

AdaBoost 算法是一种有效而实用的 Boosting 算法，它以一种高度自适应的方法顺序地训练弱学习器。其算法步骤如下：初始化各个样本点权重$w_i\ (i = 1,2, \dots, N)$，使各权重相等，并根据误差率e_m计算公式构造误差率最小的弱学习器$F_m(x)$：

$$e_m = \sum_{i=1}^{N} w_{mi} I(F_m(x) \neq y_i)$$

在得到误差率之后，调整弱学习器的权重，计算弱学习器$F_m(x)$的系数$\alpha_m = \frac{1}{2} ln\left(\frac{1-e_m}{e_m}\right)$，得到第$m$次迭代的强学习器$f_m(x)$：

$$f_m(x) = \sum_{i=0}^{m} a_i F_i(x)$$

进而，利用弱学习器的权重，更新原来样本点的权重。增大错误分类点的权重，减少正确分类点的权重：

$$w_{m+1,i} = \frac{w_{m,i}}{Z_m} \exp\left(-a_m y_i F_m(x_i)\right)$$

$w_{m+1,i}$是本次更新的样本点权重，$w_{m,i}$是上次的样本点权重，各样本权重之和为1。y_i是样本i的实际值，$F_m(x_i)$是弱学习器$F_m(x)$所预测的样本i的分类，Z_m是一个规范化因子，其公式如下：

$$z_m = \sum_{i=1}^{N} w_{mi} \exp\left(-a_m y_i F_m(x_i)\right)$$

最后，将上面的过程反复迭代，代直到误分类数达到阈值或达到最大迭代次数。M次迭代后得到最终强学习器：

$$sign[f_M(x)] = sign\left[\sum_{i=1}^{M} \alpha_i F_i(x)\right]$$

其中sign(x)是符号函数。同时，在上述过程中为防止 Adaboost 过拟合，可以向模型加入正则化项，即为每个弱学习器的权重缩减系数 v，也即学习率(Learning Rate)。

AdaBoost 充分考虑的每个分类器的权重，很好的利用了弱分类器进行级联，具有很高的精度。

模型 4——GBDT 算法

GBDT 算法是一种由多棵树组成加法模型的梯度提升决策树算法，GBDT 算法则将损失函数的负梯度作为残差的近似值，不断使用残差迭代和拟合回归树生成最终的强学习器。简单来说，AdaBoost 算法是调整权重，而 GBDT 算法则是拟合残差。其中残差计算公式为：

$$r_{mi} = y_i - f_{m-1}(x_i)$$

GBDT 算法每一次的残差计算都增大了分错样本的权重，而分对的权重都趋近于 0，因此泛化性能比较好。并且预测精度高，适合低维数据，可以灵活处理各种类型的数据。

模型 5——XGBoost 算法

XGBoost 回归模型已在上文详细阐述，其核心思想在于计算信息增益，来反映各特征变量减少信息不确定性的程度。它是经过优化的分布式梯度提升库，旨在高

效、灵活且可移植。作为 GBDT 算法的高效实现，XGBoost 算法在如下两个方面做了优化：

1) 算法本身的优化：XGBoost 算法的损失函数，除了本身的损失，还加上了正则化部分，可以防止过拟合，泛化能力更强。XGBoost 算法的损失函数是对误差部分采用二阶泰勒展开，相较于 GBDT 算法的损失函数只对误差部分做一阶泰勒展开，更加准确。
2) 算法运行效率的优化：对每个弱学习器，如决策树建立的过程做并行选择，找到合适的子树节点分裂特征和特征值，从而提升运行效率。

模型 6——LightGBM 算法

LightGBM 模型是一种根据直方图的离散值来遍历数据以寻找最优分割点的决策树算法。和 XGBoost 算法一样是对 GBDT 的高效实现，原理和 XGBoost 类似，都采用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。

1) 基于 leaf-wise 的决策树生长策略

LightGBM 使用的 leaf-wise 策略，每次在当前叶子节点中，找出分裂增益最大的叶子节点进行分裂，而不是所有节点都进行分裂，以此提高精度。

2) 直方图算法

先对特征值进行装箱处理，把连续的浮点特征值离散化成 k 个整数，形成一个一个的箱体，同时构造一个宽度为 k 的直方图，在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。

3) 并行学习

LightGBM 支持特征并行和数据并行两种。传统的特征并行主要思想是在并行化决策树中寻找最佳切分点，在数据量大时难以加速，同时需要对切分结果进行通信整合。LightGBM 则是使用分散规约，它将直方图合并的任务分给不同的机器，并利用直方图做加速训练，降低通信和计算的开销。

模型 7——CatBoost 算法

CatBoost(Categorical Boosting)是基于完全对称的决策树且能够高效处理类别型特征变量的 GBDT 框架算法，具有较高的准确性和较优的泛化能力。

CatBoost 算法的工作原理是首先对所有样本进行随机排序，然后针对某个类别型特征，每个样本的该特征转为数值型时都是基于排在该样本之前的类别标签取均值，同时加入了优先级和优先级的权重系数，降低了类别特征中低频次特征带来的噪声，从而避免梯度估计的偏差，进而解决预测偏移的问题。其优先级的权重系数公式如下：

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j}, k = x_{\sigma_p}\right] Y_{\sigma_j} + a \cdot P}{\sum_{j=1}^{p-1}[x_{\sigma_j}, k = x_{\sigma_{p,}} k] + a}$$

同时，在 CatBoost 算法中，还将浮点型特征、统计值以及 One Hot 编码进行特征二值化并放入向量中，然后在打分过程中利用二值化的特征计算模型的输出。

步骤二：模型评估

根据上文模型准备中所详细阐述的评测标准和调整决定系数R^2，本文对决策树模型、随机森林、AdaBoost 算法、GBDT 算法、XGBoost 算法、LightGBM 算法以及 CatBoost 算法进行评估，其结果如下表所示：

模型	决策树模型	随机森林	AdaBoost 算法	GBDT 算法
调整决定系数R^2	0.59135	0.69540	0.44358	0.94921
平均相对误差	0.47981	0.46154	1.78527	0.09320

5%误差准确率	0.09900	0.11167	0.05300	0.42033
评测标准	0.18324	0.19703	-0.11465	0.51763

模型	XGBoost 算法	LightGBM 算法	CatBoost 算法
调整决定系数R^2	0.95005	0.93953	0.94227
平均相对误差	0.08956	0.09324	0.08918
5%误差准确率	0.46533	0.45467	0.45183
评测标准	0.55435	0.54509	0.54363

易知，XGBoost 算法最优，其调整决定系数R^2达到了 0.95005，为 7 个模型最高，这意味着该模型拟合优度良好；平对相对误差为 8.956%，其评测标准为 0.554333，相对最为理想。除此之外，LightGBM 算法以及 CatBoost 算法稍逊于 XGBoost 算法，二者的调整决定系数R^2均在 0.94 左右，评测标准也在 0.38 上下。因此，本文暂时采用 XGBoost 算法、LightGBM 算法以及 CatBoost 算法为当前阶段的回归模型。

步骤三：模型调参——贝叶斯优化调参方法

网格优化是将一个参数空间内的所有点进行验证以返回最优损失函数的方法，然而该方法往往在计算量与计算时间上有着不可避免的缺陷，因此本文决定采用基于高斯过程建模的贝叶斯优化调参方法（BOA, Bayesian Optimization Algorithm），以争取效率和精准的双赢。

贝叶斯优化调参对给定优化的目标函数，在搜索的每次迭代中,从域空间为每个超参数选择一个样本点，利用机器学习模型计算该超参数组在验证集上的损失，并通过高斯过程不断地添加样本点来更新目标函数的后验分布,直到后验分布基本贴合于真实分布，使目标函数（每组超参数在验证集上的损失）最小。

当我们进行 GridSearch 网格搜索时,是从域空间形成的网格中为每个超参数随机选择样本点;而在贝叶斯优化中,是按照域空间超参数样本点的多元高斯概率分布进行选择，其使用了不断更新的贝叶斯概率模型,通过推断过去的结果来“集中”有希望的样本点。

对于一个 N 维向量 X，其多元高斯概率分布为：

$$p(x_1,\cdots,x_N)=\frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\cdot e^{-\frac{1}{2}[(\vec{X}-\vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{X}-\vec{\mu})]}$$

其中，μ为一个 N 维的均值向量，Σ为一个 N*N 维的协方差矩阵，|Σ|表示其行列式。进一步的，超参数θ的后验分布$p(\theta|X,Y)$为:

$$p(\theta|X,Y)=\frac{L(Y|X,\theta)p(\theta)}{\int L(Y|X,\theta)p(\theta)d\theta}$$

其中$L(Y,X|\theta)=\sum_{i=1}^{N}p(y_i,x_i|\theta)$是基于该超参数生成的机器学习模型中的似然函数，而当$p(y,x|\theta)$是高斯概率分布时，对数似然函数的闭式解存在。

经过贝叶斯调参进行优化，本文采用的三个数模型超参数取值如下：

1) XGBoost 算法超参数含义及其取值：

参数	意义	取值
learning_rate	弱分类器的权重缩减系数	0.05
max_depth	决策树的最大深度	12
n_estimators	弱学习器最大迭代次数	5000

2) LightGBM 算法超参数含义及其取值：

参数	意义	取值

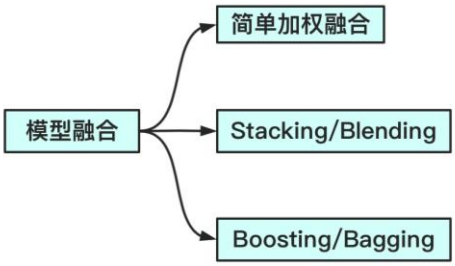
learning_rate	弱分类器的权重缩减系数	0.06
max_depth	决策树的最大深度	11
n_estimators	弱学习器最大迭代次数	5300
num_leaves	最大叶子数量	1000

3) CatBoost 算法超参数含义及其取值：

参数	意义	取值
iterations	决策树的最大树数	5000
learning_rate	弱分类器的权重缩减系数。	0.06
depth	决策树的深度	11
l2_leaf_reg	L2 正则数	3
loss_function	损失函数	MAE
eval_metric	用于过拟合检验和最佳模型选择的损失函数	MAE
random_seed	训练过程的随机种子	35

4.7 模型融合——Stacking 法

模型融合的主要思想是通过训练对多个模型进行组合以改善整体表现，在机器学习中被广泛使用。模型融合主要有以下几种方法：



本文将采用 Stacking 法对树模型进行融合，再利用加权融合法将其与神经网络模型进行融合。

4.7.1 Stacking 法

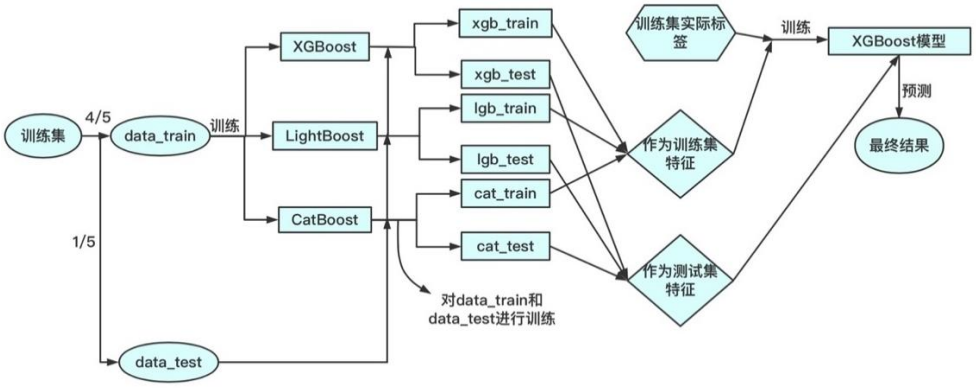
首先，本文采用堆叠法(Stacking)来对上述选择的 3 个树模型进行融合。Stacking 方法旨在减小模型的泛化误差，本质是“堆叠”的分层式结构，优势在于提高预测的准确性。本文运用两阶段 Stacking 方法，具体步骤如下图及下文所示：

第一层：

1) 本文以 4:1 的比例将前文保存的“估价训练数据整理_TreeModel.xlsx”训练集拆分为“data_train”和“data_test”，分别包含了 24000 条数据和 6000 条数据。
2) 利用 XGBoost 算法、LightGBM 算法以及 CatBoost 算法对“data_train”进行训练。
3) 用步骤 2 训练完成的 3 个模型对“估价训练数据整理_TreeModel.xlsx”训练集进行预测，其结果分别保存为“y_xgb”、“ y_lgb”以及“y_cat”。

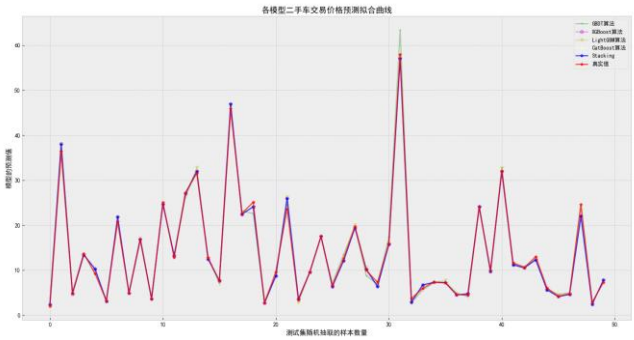
第二层：

1) 将第一层步骤 3 的三个预测结果数据集作为新的训练集特征，与“估价训练数据整理_TreeModel.xlsx”训练集的实际特征一起，选取第一层中单一模型评测标准较高的 XGBoost 算法作为元模型进行次级训练预测。
2) 用训练完成的 XGBoost 算法对“估价验证数据整理_TreeModel.xlsx”验证集进行预测，得到最终结果。



4.7.2 拟合后模型评估

本文作出 Stacking 后的模型及单一模型对随机抽取的 50 辆二手车交易真实价格预测结果的拟合曲线。



由图可知，Stacking 融合后模型的拟合曲线更贴切真实交易价格曲线，模型预测精度更高。在局部预测差异较大的区域，Stacking 融合后模型表现出的波动较小且预测效果更为稳定。在模型评价方面，下表展示了 Stacking 在四种指标下的模型预测效果：

	调整决定系数R^2	平均相对误差	5%误差准确率	评测标准
模型融合后评估	0.94987	0.08952	0.46500	0.55410

与上文单一模型对比，本文发现 Stacking 后的模型平均相对误差略微降低，且评测标准有所提高。

4.7.3 保存数据集

1） 利用 Stacking 融合后的模型对估价训练数据进行预测，将预测值保存为“估价训练数据指导价.xlsx”。

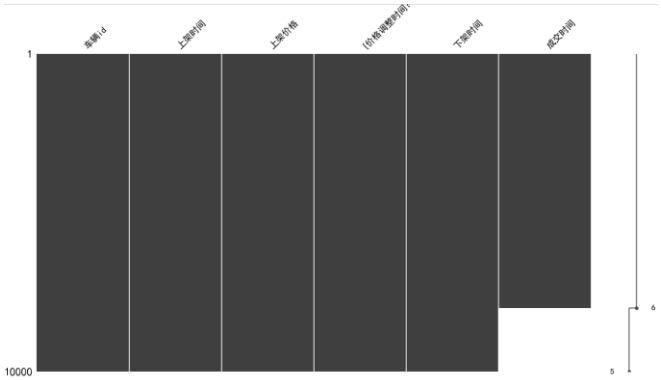
2） 利用 Stacking 融合后的模型对估价验证数据进行预测，将预测值保存至附件 3“估价模型结果.txt”中。

五．问题二 模型的建立与求解

5.1 特征工程

5.1.1 特征构造——是否成交

首先，本文绘制附件 4 数据集的缺失值数据密度图如下：



由图易知，仅“成交时间”这一字段缺失 2000 条数据。结合题干表格，成交车辆下架时间与成交时间相同，这意味着存在 2000 辆二手车未成交就已被下架。因此本文构造新特征变量“是否成交”，其中成交车辆填充为 1，未成交车辆填充为 0。

5.1.2 特征构造——价格调整时间 n、调整后价格 n、调价次数

易知特征“{价格调整时间：调整后价格}”中包含二手车的每一次调价时间和每一次调整价格。为利于后续对该变量所含信息进行深入分析挖掘，本文采用正则表达式，将每一次价格调整时间和每一次调整后价格分离处理，由于最多调整次数为 12 次，因此本文将它们分门别类地存储于 24 个新变量“价格调整时间 1～12”和“调整后价格 1～12”中，并计算价格调整的次数，保存在新字段“调价次数”中，其中未进行调价的时间、价格计作空值。

5.1.3 处理时间格式

检索数据集可知，“上架时间”、“下架时间”、“成交时间”、“价格调整时间 1～12”这 15 个字段均为 object 格式，为后续进一步开展时间特征的数据分析，本文利用 pandas 库中的方法 dt 对其进行高效操作，将上述 15 个字段转化为 datetime64 格式。

5.1.4 特征构造——平均降价幅度、最大降价幅度、最小降价幅度

本文根据价格调整时间 n、调整后价格 n 构造了上述 3 个特征，利用公式分别进行计算填充。

1） 平均降价幅度：

$$\begin{cases} 0, & \text{调价次数} = 0 \\ \dfrac{\text{上架价格} - \text{调整后价格}_1}{\text{上架价格}}, & \text{调整次数} = 1 \\ \dfrac{\dfrac{\text{上架价格} - \text{调整后价格}_1}{\text{上架价格}} + \sum_{i=1}^{n-1} \dfrac{\text{调整后价格}_i - \text{调整后价格}_{i+1}}{\text{调整后价格}_i}}{n}, & \text{调整次数} = n > 1 \end{cases}$$

2） 最大降价幅度：

$$\begin{cases} 0, & \text{调价次数} = 0 \\ \dfrac{\text{上架价格} - \text{调整后价格}_1}{\text{上架价格}}, & \text{调整次数} = 1 \\ \max\{\dfrac{\text{上架价格} - \text{调整后价格}_1}{\text{上架价格}}, \dfrac{\text{调整后价格}_i - \text{调整后价格}_{i+1}}{\text{调整后价格}_i}(i = 1, \cdots, n-1)\}, & \text{调整次数} = n > 1 \end{cases}$$

3） 最小降价幅度：

$$\begin{cases} 0, & \text{调价次数} = 0 \\ \dfrac{\text{上架价格} - \text{调整后价格}_1}{\text{上架价格}}, & \text{调整次数} = 1 \\ \min\{\dfrac{\text{上架价格} - \text{调整后价格}_1}{\text{上架价格}}, \dfrac{\text{调整后价格}_i - \text{调整后价格}_{i+1}}{\text{调整后价格}_i}(i = 1, \cdots, n-1)\}, & \text{调整次数} = n > 1 \end{cases}$$

5.1.5 特征构造——平均调价周期、最大调价周期、最小调价周期

除此之外，本文根据价格调整时间 n、调整后价格 n 构造了上述 3 个特征，利用公式分别进行计算填充。

1） 平均调价周期

$$\begin{cases} 0, & \text{调价次数} = 0 \\ \text{价格调整时间}_1 - \text{上架时间}, & \text{调整次数} = 1 \\ \dfrac{\text{价格调整时间}_1 - \text{上架时间} + \sum_{j=2}^{n-1} \text{价格调整时间}_j - \text{价格调整时间}_{j-1}}{n}, & \text{调整次数} = n > 1 \end{cases}$$

2） 最大调价周期

$$\begin{cases} 0, & \text{调价次数} = 0 \\ \text{价格调整时间}_1 - \text{上架时间}, & \text{调整次数} = 1 \\ \max\{\text{价格调整时间}_1 - \text{上架时间}，\text{价格调整时间}_j - \text{价格调整时间}_{j-1}; (j = 2, \cdots, n)\}, & \text{调整次数} = n > 1 \end{cases}$$

3） 最小调整周期

$$\begin{cases} 0, & \text{调价次数} = 0 \\ \text{价格调整时间}_1 - \text{上架时间}, & \text{调整次数} = 1 \\ min\{\text{价格调整时间}_1 - \text{上架时间}，\text{价格调整时间}_j - \text{价格调整时间}_{j-1}; (j = 2, \cdots, n)\}, & \text{调整次数} = n > 1 \end{cases}$$

5.1.6 特征构造——成交价格

以前文“是否成交”这一特征为依据，对完成成交的二手车构造新特征“成交价格”。该字段的填充方式为：

1） 若该二手车未进行调价，则以“上架价格”中的数据对“成交价格”进行填充。

2） 若该二手车参与调价，则以最后一次“调整后价格”对“成交价格”进行填充。

5.1.7 特征构造——成交周期、下架周期

以前文“是否成交”这一特征为依据，构建新特征如下：

1） 若二手车完成成交，则用“成交时间”减去“上架时间”构造特征“成交周期”。

2） 若二手车未完成成交，则用“下架时间”减去“上架时间”构造特征“下架周期”。

5.1.8 特征构造——最佳降价幅度

以前文“是否成交”这一特征为依据，构建新特征如下：

1） 若二手车完成成交，且进行过调价，则利用如下公式构造特征“最佳降价幅度”：

$$\frac{\text{上架价格} - \text{调整后价格}n}{\text{上架价格}}, n = 1, \cdots, 12$$

2） 若二手车完成成交，但未进行调价，则直接填充为 0。

3） 若二手车未完成成交，则为空值。

5.1.9 特征构造——最佳成交周期

以前文“是否成交”这一特征为依据，构建新特征如下：

1） 若二手车完成成交，且进行过调价，则利用如下公式构造特征“最佳成交周期”。

$$\text{成交时间} - \text{价格调整时间}n\ ,n = 1,\cdots,12$$

2） 若二手车完成成交，但未进行调价，则用“成交时间”减去“上架时间”构造特征“最佳成交周期”。

3） 若二手车未完成成交，则为空值。

5.1.10 删除冗杂特征——上架时间、{价格调整时间：调整后价格}、下架时间、成交时间

由于在上述步骤中已充分利用“上架时间”、“ {价格调整时间：调整后价格}”、“ 下架时间”、“ 成交时间”这 4 个特征，且计算构造了新的时间特征变量，因此为了避免信息的重复解释并对数据集降维，本文对这 4 个特征进行删除处理。

5.1.11 保存数据集

本文将经上述处理的数据集保存为“门店交易训练数据.xlsx”。由于篇幅有限，本文摘取前 5 行数据如下表所示：

车辆 id	上架价格	是否成交	调价次数	平均降价幅度	最大降价幅度	最小降价幅度	平均调价周期
68603	3.98	1	0	0	0	0	0
12312	4.5	1	0	0	0	0	0
57655	23.9	1	1	0.037657	0.037657	0.037657	23
45688	20.5798	1	0	0	0	0	0
52081	12.28	1	1	0.030945	0.030945	0.030945	21

车辆 id	最大调价周期	最小调价周期	成交价格	成交周期	下架周期	最佳降价幅度	最佳成交周期
68603	0	0	3.98	0	空值	0	0
12312	0	0	4.5	31	空值	0	31
57655	23	23	23	26	空值	0.037657	3
45688	0	0	20.5798	3	空值	0	3
52081	21	21	11.9	53	空值	0.030945	32

5.2 挖掘车辆成交周期的关键因素

5.2.1 数据集拼接

由题干可知，车辆本身的性能与价格是否公道是成交的关键性因素，因此本文将问题一中保存的数据集“估价训练数据整理_NN.xlsx”、“ 估价训练数据指导价.xlsx”与问题二特征工程中保存的“门店交易训练数据.xlsx”进行拼接，用于后文进一步探究二手车交易的成交周期。

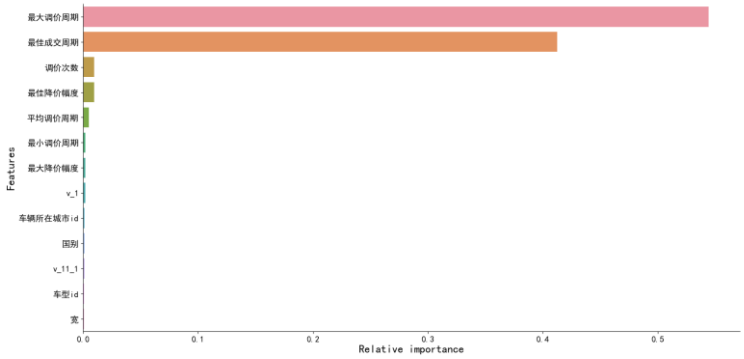
5.2.2 特征构造——指导价调价幅度

本文利用如下公式构造新特征“指导价调价幅度”：

$$\frac{\text{成交价格} - \text{指导价}}{\text{指导价}}$$

5.2.3 特征重要性排序——XGBoost 算法

本文利用上文已阐述的 XGBoost 算法对特征重要性进行排序，重要性前 13 位的特征如下：



根据上文特征构造的具体阐述，易知“成交周期=调价周期+最佳成交周期”，其中，成交周期代表从上架到最后一次调价所用时长，而最佳成交周期代表最后一次调价到成交所用时长，这意味着减短调价周期和最佳成交周期可能是减短成交周期的至关要素所在。据图分析如下

1） **最大调价周期**

最大调价周期对二手车成交周期的影响最为显著，其特征重要性分数为 54.433%。在二手车销售过程中，价格的调整往往是不可或缺的阶段。不可避免地，对于门店而言，更高的销售价格会带来更为可观的利润，但由于消费者通常对二手车购买预备有目标预算或可接受的价格区间，因此门店相应地按照消费者心理预期作出价格调整是难以避免的销售行为。对此，门店销售逐步调整价格，在以成交为最终目的的同时期待更高利润，但倘若能够压缩每一次调价的周期，即可在保证上述要求的同时减短最大调价周期，从而达到减短成交周期的目的。

2） **最优调价周期**

最佳成交周期间不再对价格进行调整，意味着该时间段内二手车价格已符合消费者的心理预期水平。由于门店销售的谈判技巧并非成交的根本因素，因此本文认为最佳成交价格与车辆本身的性能相关，这也是消费者在此期间内尚在考量犹豫的原因所在。

由图，最佳成交周期的特征重要性分数较高，为 41.239%，这意味着最佳成交周期是影响二手车成交的关键性因素。结合上段分析，本文认为最佳成交周期时长背后的根本因素在于二手车性能的好坏，这可以引申为二手车车况对推动其是否迅速成交具有强烈影响。比如，以同等零售价格为前提，二手车本身为普适度强且各功能状况良好的实用性车型，其受众程度更广，且消费者对其最终考量犹豫的时长可能会远远短于相比之下具有特殊分工或车况稍次的二手车。

3） **调价次数和最佳降价幅度**

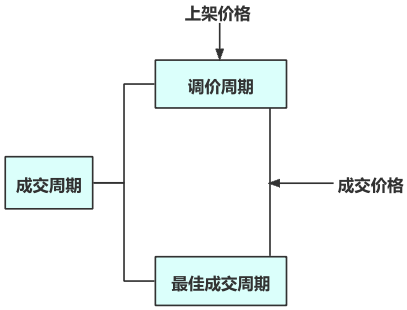
由图可知，调价次数和最佳降价幅度对成交周期具有一定的影响。其二者之间存在一定联系：降价幅度越大，则需要更少的调价次数，即可到达消费者心理预期的价格区间；相应地，调价次数越多，则无须大幅度降价，便可以达到消费者的预期价格。因此，最佳降价幅度增大或调价次数增多，二手车的成交周期都得到压缩。

4） **二手车性能**

由图可以判断，二手车性能对成交周期的影响微乎其微。本文认为，二手车性能直接影响的是二手车销售价格，仅通过前文所提到的最佳调价周期而对成交周期产生间接影响。

5.3 寻求加快在库车辆销售速度的有效手段

5.3.1 模型构建



为了加快门店二手车的销售速度、压缩成交周期，本文进行了以下思考：

1） 针对上架价格

根据问题一中 Stacking 后的模型预测出适宜的成交价格(指导价)，由于问题一中的模型搭建已经考虑到车辆性能、车况、受众环境等多方面因素，故而该指导价与消费者所预期的价格直接挂钩。门店定价师对消费者心理作出一定了解，对二手车适宜的成交价格有所预判后，可以参照指导价对初始上架价格进行设定。在指导价上浮 3%-7%左右设定上架价格，既可在接近消费者心理价位预期的情况下保证超额收益，也可根据短期内是否成交进行后续的调价行为。

2） 针对调价周期

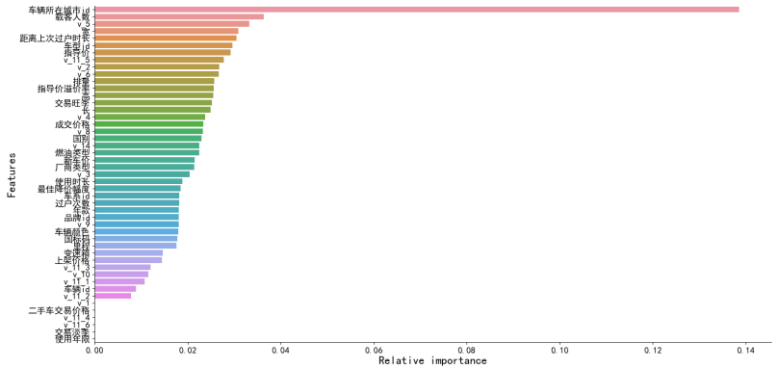
在上架价格无法满足消费者预期时，门店定价师可以通过调价的手段来刺激消费者购买。由 6.2.3 分析出的关键影响因素可知，尽可能地缩短每次调价之间调价周期、增大每次调价的降价幅度，能够使价格在相对较短的时间内达到消费者可接受的价格区间(指导价附近)，以减少从上架到最后一次调价所需要的全部时长(调价周期)，最终达到压缩总的成交周期的目的。

3） 针对最佳成交周期

由 6.2.3 已知最佳成交周期会直接影响到总的成交周期，而如何缩减最佳成交周期，亦是门店定价师需要采取的手段的重中之重，本文将在后续进一步探究缩减最佳成交周期的使用条件和有效手段。

5.3.2 最佳成交周期的特征重要性排序——XGBoost 算法

本文利用 XGBoost 算法对最佳成交周期的特征重要性进行排序如下：

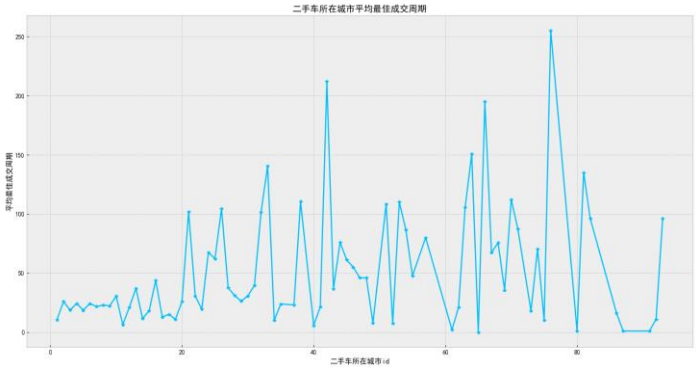


如图，“车辆所在城市 id”、“载客人数”、“匿名变量 5”以及“宽”为最佳成交周期的几个主要影响因素，其特征重要性分数如下表所示：

特征变量	重要性分数
车辆所在城市 id	13.855%
载客人数	3.638%
匿名变量 5	3.321%
宽	3.090%

本文将对二手车车辆所在城市、载客人数、长宽高、指导价溢价率这 6 个特征进行逐一讨论，探究如何利用这些特征采取行之有效的手段以缩减二手车的最佳成交周期。

5.3.3 根据车辆所在城市探究改善最佳成交周期的手段

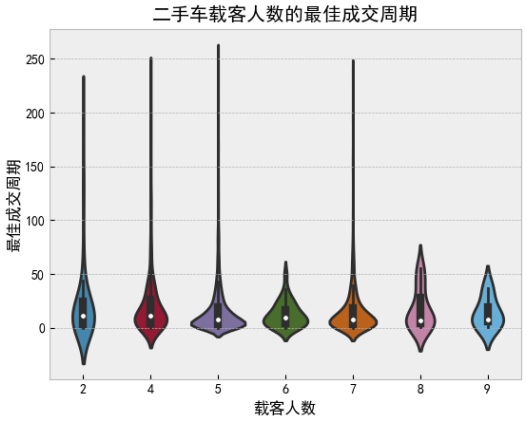


通过计算不同城市的平均车辆最佳成交周期并绘制折线图，发现不同城市的平均最佳成交周期从 10 天到 260 天不等。这可能是因为不同城市之间经济发展水平存在差异，二手车价格相较于收入的比例为人们带来不尽相同的考量；同时，不同地区人均消费水平及消费观念也有所不同，中小城市消费者相对传统的消费行风习惯于用积蓄购买车辆，而大城市消费者乐观的超前消费理念不会考虑过多的因素，达到心理预期价位即可成交；最为重要的是，相较于二三线城市，二手车在一线乃至超一线城市的受众面更广，二手车门店客流量更大，更多的顾客带来更多交易的可能性，故而导致其最佳成交周期更短。

采取手段：

根据第一问由车辆本身性能及车况预测出的指导价，同样的车辆，已降价至指导价附近时，针对大城市，应该给与其消费者更小的时间选择范围，即之后的待成交时间接近该城市平均最佳成交周期时，便可考虑再进行降价微调，因为大城市二手车门店客流量大，在接近平均最佳成交周期的时长内仍未被购买，表明此价格仍未达到消费者心理预期，故可提前降价；而面向中小城市，由于本身受众面较窄、门店客流量较少，故可适当放宽消费者的时间选择范围，例如可以以比该城市平均最佳成交周期更长的时间去等待合适的交易的发生，无需做额外的调价，从而既缩减了最佳成交周期，也保证了二手车交易平台的权益。

5.3.4 根据车辆载客人数探究改善最佳成交周期的手段



通过计算不同载客人数的车辆的最佳成交周期并绘制小提琴图，发现不同载客人数对车辆最佳成交周期存在较大影响，其中载客人数为 5，6，7 人的车辆最佳成交周期分布相对集中，且最佳成交周期的 95%置信区间较小，而其它载客人数的车辆最佳成交周期相对分散，且最佳成交周期的 95%置信区间较大。这可能是由于不同车辆的

受众存在大小差异导致，目前市场上主流的私家车一般都是 5 座的轿车，6，7 座的 SUV 为主，受众较广。而 2 座，3 座的汽车以价格较高的跑车为主，8 座，9 座的车以面包车为主，受众较小，所以最佳成交周期分布更分散，平均最佳成交周期相对更长。

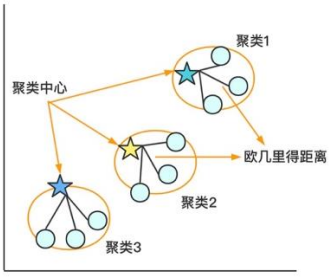
采取手段：

根据第一问由车辆本身性能及车况预测出的指导价，同样的车辆，已降价至指导价附近时，针对载客人数为 5，6，7 人的车辆，由于其受众广，门店定价师在其之后的待成交时间接近平均最佳成交周期时，便可考虑再进行降价微调，因为该类别二手车需求旺盛，在接近平均最佳成交周期的时长内仍未被购买，表明此价格仍未达到消费者心理预期，故可提前降价；针对载客人数为 2，3，8，9 人的车辆，由于其受众较小，门店定价师在其之后的待成交时间超过该载客人数车辆最佳成交周期的 95%置信区间后再降价微调，充分等待潜在客户的挑选。

5.3.5 基于 K-Means 聚类探究改善二手车最佳成交周期的手段

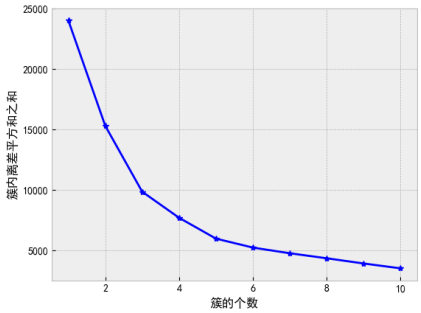
由特征重要性分数可知二手车的长宽高对最佳成交周期亦存在较大影响。

K-Means 算法是一种将数据划分到相应集群的聚类算法，具有高效处理大体量数据的优点，且聚类效果较好。K-Means 算法的基本思想是:先给定一个参数 K，随机选取 K 个质点作为聚类初始质点，计算各个质点到聚类中心的欧几里得距离，并将其分别分配到距离最近的聚类中心所在簇中;然后重新计算各个簇的中心，并再对所有质点进行分配;重复以上步骤，直至相邻两次迭代中聚类中心不再变化或聚类准则函数收敛，而上述往往采用误差平方和作为准则函数。



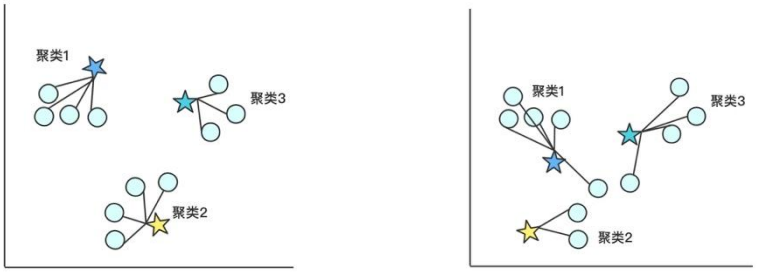
其基本步骤如下：

1) 由于本文所涉及数据集需要对二手车最佳成交周期进行分类，因此我组根据 SSE 簇内离差平方和拐点法，选择折线坡度骤减的点作为最优 K，如下图 K 为 3 时最优，确定将数据集划分为 3 类，并选取 3 个初始聚类中心为 (C_1, C_2, C_3)；



2) 计算样本x_i $(1 \le i \le 8000)$到初始聚类中心的欧几里得距离，并根据最小距离原则确定每个样本的聚类中心C_j $(1 \le j \le 3)$，将分别分配到相应的簇中；

3） 重新计算各簇的聚类中心，并重新分配各样本，取第一、二次分配绘图如下：

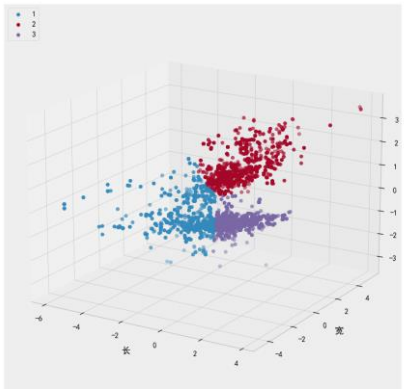


4） 重复上述步骤，迭代计算聚类中心并分配样本，直至聚类中心稳定不变或满足聚类准则函数收敛。其中，聚类准则函数的收敛情况通常以各样本到聚类中心的距离平方和是否最小来进行判断，其公式如下：

$$Var = \sum_{i=1}^{n} [min_{j=1,\cdots k} d(x_i, z_j)^2]$$

其中，若Var值收敛，则终止算法，完成聚类。

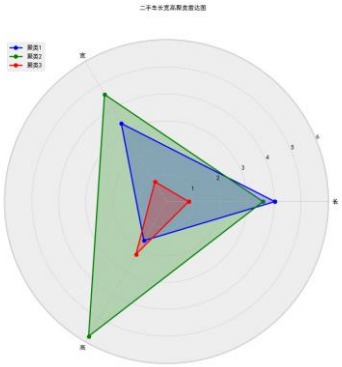
本文将可视化结果绘制成三维着色散点图如下：



该可视化结果以二手车的“长”为 X 轴，“宽”为 Y 轴，“高”为 Z 轴，通过不同的颜色将车辆分为了 3 类，其分别包含的样本数及聚类中心点如下表所示：

类别	所含样本数目	聚类中心点坐标
1	3487	(0.51476, 0.17706, -0.66106)
2	2258	(-1.08819, -1.07229, -0.36899)
3	2255	(0.29364, 0.79992, 1.39170)

此外，本文绘制了依据长宽高这三个显著性特征的聚类雷达图如下：



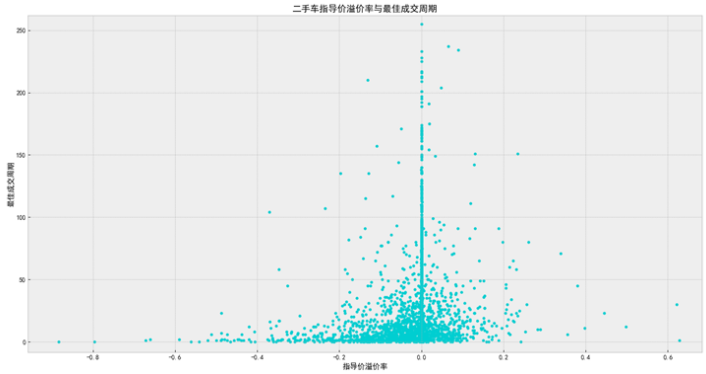
结合散点图、雷达图及表格，并比对数据集可知：

1) 聚类 1 包含样本数目最多，且该类型车辆长宽高趋于中等一致，本文认为聚类 1 可能是受众面较广的轿车。经检索数据集进行比对可知，该类别车辆平均最佳成交周期相对于后两种更短，表明该类别车辆市场需求量大。
2) 聚类 2 明显更趋向于高和宽有明显优势的车型，本文查阅专业文献得出聚类 2 可能是运动型多功能车(SUV)或多用途商务车(MPV，多为两厢式，不追求车长)，该聚类的车辆在涵盖聚类 1 轿车的城市功能基础上，又具有一定的特殊功能，如越野功能、旅行乘务功能或载物功能，但可能由于车型、价格等多方面考量，其与聚类 1 轿车相比，平均最佳成交周期较长，表明市场和需求都有所下降。
3) 聚类 3 明显缩减了车辆的整体空间，本文查阅专业文献得出聚类 3 可能属于跑车或 MINI 类型的单厢车。该类型的车意在追求车身的小巧，但往往仅适用于个人或两人小家庭，且不具备载货或商务功能，故而市场需求不大，因此该类别车辆的平均最佳成交周期较大。

采取手段：

根据第一问由车辆本身性能及车况预测出的指导价，同样的车辆，已降价至指导价附近时，针对聚类 1 车型的二手车，由于其为大众所普遍接受，门店定价师在其之后的待成交时间接近该聚类平均最佳成交周期时，便可考虑再进行降价微调，因为该类别二手车需求旺盛，在接近平均最佳成交周期的时长内仍未被购买，表明此价格仍未达到消费者心理预期，故可提前降价；对于聚类 2 车型的二手车，市场对于其接受程度适中，门店定价师可在其之后的待成交时间等于该聚类平均最佳成交周期时准时调价；而对于聚类 3 车型的二手车，受众面往往较窄，故而可延长时间再进行调价，充分等待合适的顾客进行挑选。

5.3.6 通过二手车指导价溢价率探究改善最佳成交周期的手段



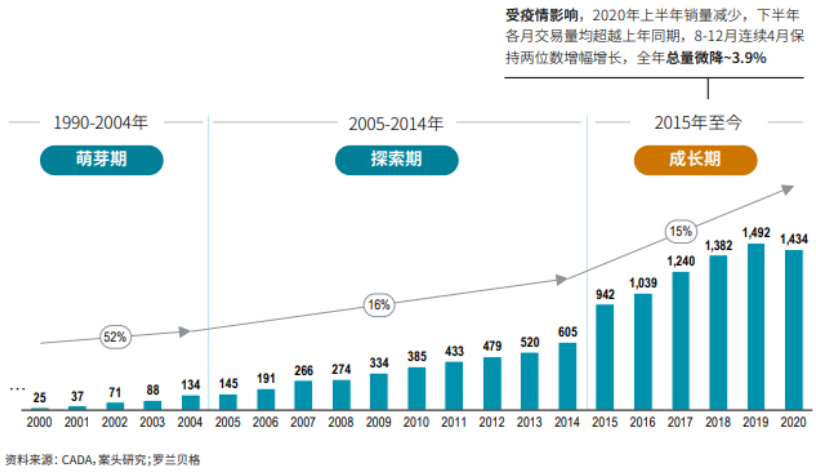
通过分析二手车实际成交价格与第一问预测出的成交价格(指导价)的溢价率，可以很明显的发现，当指导价溢价率为负值时(即以比预测的价格低估卖出)，98.3%的二手车最佳成交周期均小于 30 天，达到了很好的流通效果；而当指导价溢价率为正时，消费者受众的接受意愿相比之前明显下降，尤其是当溢价率超过 10%时，有接近 50%的二手车最佳成交周期超过了 30 天，表明在实际成交价格相对于指导价溢价过多时，消费者虽有购买需求但是其购买意愿并不会那么强，因此导致流通效果变差。

采取手段：

对于探究如何利用指导价溢价率来加快最佳成交周期的手段，门店定价师可结合上述分析的多方位重要因素进行考量。相较受众面较窄的二三线城市，一线乃至超一线城市针对普适型号的二手车可以考虑最终的定价相对于指导价溢价率较高一些；而二三线城市针对受众面较小的商务车、跑车等，可使最终的定价略低于指导价，以次加快门店在库二手车的销售速度。

六．问题三 模型的建立与求解

我国二手车发展历经萌芽期、探索期和成长期三个阶段。随着互联网玩家的快速扩张，以及国家支持性政策法规密集出台，我国二手车交易规模快速上升，行业整合淘汰加速。在过去几年中，中国二手车市场保持着快速增长的态势，2019 年突破交易量 1,492 万台。然而，与成熟市场相比，中国二手车市场仍处于早期发展阶段，整体市场二手车占新车的比率较低。



在供应端和消费端共同利好的影响下，未来二手车市场存在较大的发展潜力。因此，实现可持续性发展便显得尤为重要，本文从提高利润率，扩宽客户受众，拓展业务广度，增添附加值业务四个方面入手，研究如何实现可持续性发展。

6.1 提高利润率

由于 020 门店模式为线下实体门店对外展销和售卖，导致每个线下门店能展销的车辆有限，因此针对不同城市消费水平，客户偏好等差异进行展销尤为必要，可以通过分析历史销售数据，结合问题二中分析得出不同城市对不同类型二手车的最佳成交周期会有所不同，进一步得到不同城市对于不同类型汽车的需求量，从而可以在不同城市之间进行二手车的流通，保证二手车的销售效率，达到利润率最大化。

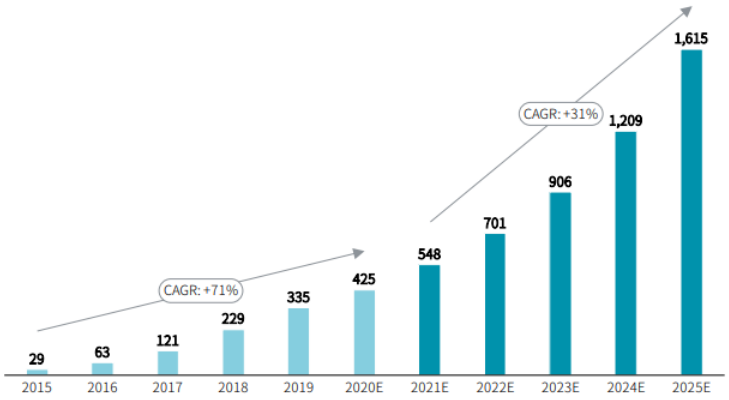
6.2 扩宽客户受众

传统经销商利用自身品牌效应获客，电商平台多利用互联网流量平台。二手车是低频消费，要想稳定客源就需要不断的吸引新客户。目前二手车市场竞争激烈，新车的价格波动也会冲击二手车市场的交易价格，容易造成客户流失，无论是线下传统交易还是二手车电商交易，均需要较高的获客成本。传统的大型经销商集团具有品牌效应，可以通过口碑营销等方式取得低获客成本的引流效果，二手车电商平台通过与互联网巨头的战略合作也有助于其引流。如优信通过与零售巨头淘宝在供应链、交易过程、金融服务以及售后物流等全方位的合作，有助于优信从超级流量平台获取流量资源优势。而中小二手车商既没有品牌效应，也难以接入流量平台引流，获客成本相对较高。由题意可知典型的二手车零售场景一般通过互联网等线上渠道获取用户线索，线下实体门店对外展销和售卖，因此可以通过与主流互联网运营商进行合作，寻找潜

在目标客户，并绘制用户画像，对潜在客户进行精准的二手车辆信息广告投放，从而扩宽客户受众的广度。

6.3 拓宽业务广度

目前汽车市场从增量市场转换为存量市场，因此二手车市场是逐年递增的，如下图所示，新能源汽车的市场占有率逐年提高，新能源车保有量的快速攀升将有效带动新能源二手车产业的蓬勃发展，也为二手车产业链带来全新的挑战。可以见得未来二手新能源汽车市场会不断增长。因此可以针对二手新能源汽车市场提供针对性业务模式，例如新能源汽车车电分离后二手电池产业，包括二手电池残值定价、二手电池交易，三电系统检测与整备修复，以及回收电池的梯次利用等等；以及二手新能源充电桩产业，包括新能源车置换过程中私桩充电桩的交易与回收等等。



资料来源:IEA;罗兰贝格

6.4 增添附加值业务，增厚利润空间

对二手车进行检测认证：与经销商通过以旧换新的收车价格相比，认证二手车的价格可以上升约 40%，远高于未得到认证车辆的价格。二手车认证不仅对二手车相关手续、车型和年款、以及对真实车况的认证，还会通过各交易市场的实际历史成交价格直接给出客观的建议交易价格，保障了消费者的合理权益。

对二手车进行翻新：二手车整备和翻新通常需要经过比较复杂的工序，其中涉及了更换零部件、清洗发动机舱、更换轮胎以及车身抛光等一系列工序。二手车商可以利用翻新工作提高自身对于二手车的定价，同时利用其规模和技术经验优势降低成本，带来较大的利润空间。

增加质保项目：随着二手车拍卖等新型交易的出现，抢占了二手车车源的部分市场份额，导致二手车价格被抬高。二手车商向消费者提供的质量保障方案进一步增强消费者的信心。

为消费者提供保险：为消费者提供保险销售和咨询服务、赚取代理佣金，可成为二手车商重要的盈利点。经销商需要根据用户需求变化趋势和不同用户的需求差异性，调整车险品种推荐方式，从而增加车险服务成交率。通过服务打包、会员制等多种创

新方式，经销商不仅可以完成保险销售，而且带动其他增值服务的开展，长期锁定用户进店消费。

七．参考文献

[1]姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报：工学版, 2014(1):5.
[2]胡仁涛. 基于贝叶斯优化调参的双频带天线设计[A]. 中国电子学会.2020 年全国微波毫米波会议论文集（下册）[C].中国电子学会:中国电子学会微波分会,2020:3.DOI:10.26914/c.cnkihy.2020.054779.
[3]彭星光, 高晓光, 魏小丰. 基于贝叶斯优化算法的 UCAV 编队对联合目标的协同攻击研究[J]. 系统仿真学报, 2008, 20(10):4.
[4]赵梦想, 胡敏. 基于 Stacking 融合的电商平台二手车定价研究[J]. 信息与电脑, 2021, 33(19):4.
[5] Chen T, Tong H, Benesty M. xgboost: Extreme Gradient Boosting[J]. 2016.
[6] Zheng H, Yuan J, Chen L. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation[J]. Energies, 2017, 10(8):1168.
[7] Wang D, Yang Z, Yi Z. LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients[C]// the 2017 International Conference. 2017.
[8]方匡南, 吴见彬, 朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3):7.
[8]李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4):8.
[9]曹正凤. 随机森林算法优化研究[D]. 首都经济贸易大学.
[10]刘波, 秦川, 鞠平,等. 基于 XGBoost 与 Stacking 模型融合的短期母线负荷预测[J]. 电力自动化设备, 2020, 40(3):7.
[11]王星. 一种 stacking 模型融合方法:, CN112163620A[P]. 2021.
[12]于玲, 吴铁军. 集成学习:Boosting 算法综述[J]. 模式识别与人工智能, 2004, 17(1):8.
[13] Jabeur S B, Gharib C, Mefteh-Wali S, et al. CatBoost model and artificial intelligence techniques for corporate failure prediction[J]. Technological Forecasting and Social Change, 2021, 166(8):120658.
[14]黄冠翎. CatBoost 算法在乳腺肿瘤诊断研究中的应用[D]. 上海师范大学, 2019.

队伍编号	MCB2102542
赛道	（B）

信息流智能推荐算法中序列收益评估模型的构建与研究

摘 要

目前互联网信息蓬勃发展，用户面临信息过载，而推荐算法不仅满足用户个性化内容消费需求，而且提升了用户获取有用信息的效率。本文通过建立推荐序列总点击量和总时长收益评估模型，为推荐序列提供预测服务，从而达到更好地进行信息序列的推送和进行信息流整体收益评估。

针对问题一：基于训练集数据，**建立评估推荐序列总点击和总时长收益数学模型**，并进行**优化网格矩阵**计算，得出**不同排列组合的收益值**并计算**时间收益系数**。同时统计分析不同时段的点击量，计算不同时段对收益值的影响。

针对问题二：基于问题一模型，建立各用户的**优先网格矩阵并**获取相关数据，建立**隐语义预测模型**，获得**隐分类用户矩阵**和**隐分类物品矩阵**，得到未浏览得点击量和浏览时长，考虑时间、排序等因素，**预测**请求 ID 对应**推荐序列的点击量和时长**之和，并填入附件 result.csv 中，**预测精准度达到 93%**左右。

针对问题三：根据题目设计如下**剪枝策略**。统计分析每个请求 ID 的训练内容获得序列的最优长度；基于问题一的模型，建立每个 ID 的**优先网格矩阵，曲线拟合**获得不同内容排序带来的收益；通过每条内容的收益值设计每种分类内容的选择概率，再通过**轮盘赌**的方式选择推荐序列。经过检验证明**最优序列精准度达到 90%**左右。

综上所述，本文依据各题所给的条件较全面地分析并解决了信息流智能推荐算法中序列收益评估建模问题。经过分析验证，本文建立的模型合理有效，简便易行，便于推广到日后其他序列收益评估的建模中去，有一定的现实意义。

关键词：推荐算法 推荐序列 优化网格矩阵 隐语义预测模型 曲线拟合 剪枝策略 轮盘赌

目录

1.问题重述 1
1.1. 问题背景 1
1.2. 问题提出 1
1.3. 问题求解流程图 1
2.问题分析 2
2.1. 问题一的分析 2
2.2. 问题二的分析 2
2.3. 问题三的分析 3
3.符号说明 3
4.基本假设 3
5.数据处理 4
6.问题一：推荐序列总点击和总时长收益评估模型 4
6.1. 问题一的建模流程 4
6.2. 模型的建立及求解 5
7.问题二：预测推荐序列的总点击量和总时长 8
7.1. 考虑序列顺序的隐语义模型概述 8
7.1.1 基于用户的协同过滤算法 9
7.2.1 基于物品的协同过滤算法 10
7.3.1 隐语义模型 11
7.2. 考虑序列顺序的隐语义预测模型 13
7.3. 误差分析和结果写入 13
8.问题三：剪枝策略下精确评估最优序列收益 14
8.1. 问题三的建模流程 14
8.2. 模型的建立及求解 15
8.3. 剪枝策略复杂度分析 15
8.4. 剪枝策略的精准度检验 16
9.总结和展望 17
参考文献 18
附录 19

1. 问题重述

1.1. 问题背景

21 世纪是我们所生活的年代，是世界经济发展的时代，是科学力量加强的时代，也是一个信息化的时代，因此生活在这样的一个时代中，用户在使用互联网应用时时常面临着信息过载的问题。推荐算法的出现，满足了用户个性化的内容消费需求，提升了用户获取有用信息的效率，在互联网 APP 里已被广泛应用。

1.2. 问题提出

信息流作为推荐算法的主要应用场景，是用户触及互联网信息的主要入口，已经完全融入了人们的日常生活中，成为了人们了解世界的主要方式。传统推荐算法的核心思想是挖掘被推荐内容与用户兴趣的匹配关系，以及内容本身的优质程度，选择与用户最相关或者最优秀的内容推荐给用户。

但是，经过专家学者研究发现，除了内容本身因素以外，内容之间的排列组合关系，也会影响用户的浏览体验，进而影响推荐收益的大小。例如，相似内容的高度集中，往往会带来较差的结果反馈，即使它们都高度匹配用户兴趣或者具有较高的内容质量。于是，越来越多的研究集中在如何选择最优的内容排列组合上，而不仅仅是最优的内容上。因此本题目要求设计数学模型对给定的候选序列进行序列整体收益的评估。

根据题目提供附件中的训练集 train_data.txt、内容的基础属性 doc_info.txt、测试集 test_data.txt 以及数据格式示例和包括的涉及字段，进行信息流智能推荐算法中的序列评估问题的解决。这需要我们解决如下问题：

问题 1：建立评估推荐序列总点击收益和总时长收益的数学模型，以及如何根据点击收益和时长收益对综合收益进行量化。在序列评估模型设计中，需要详细阐述如何考虑不同排列组合对收益大小的影响。

问题 2：结合问题 1 设计的数学模型，预测测试集中推荐序列的总点击量和总时长（单位为秒），将预测结果写入 result.csv 并提交。

问题 3：假设有 N 条候选内容，从中选择长度为 K（N≥K）的最优推荐序列，需要参与收益评估的序列量为A_N^K。在真实推荐场景中，由于计算性能的考虑，系统无法对所有可能序列进行收益评估，往往需要先采用计算复杂度更低的方式对序列集合进行剪枝，圈定出少量候选序列进行精确收益评估。而剪枝策略的目标是保证候选序列集更大可能包含最优序列。请详细阐述你的建模思路，以及剪枝策略的精准度和时间复杂度。

1.3. 问题求解流程图

综合分析上述四个问题，我们的建模目标是根据提供的近一周时间内用户在信息流产品上的曝光历史、内容的基础属性以及在训练集时间之后的一部分用户推荐序列等信息，构建评估推荐序列总点击收益和总时长收益的数学模型，从而为更好的进行信息的推荐并达到收益最大化提供预测服务。基于此，给出本题的解题流程图如下所示：

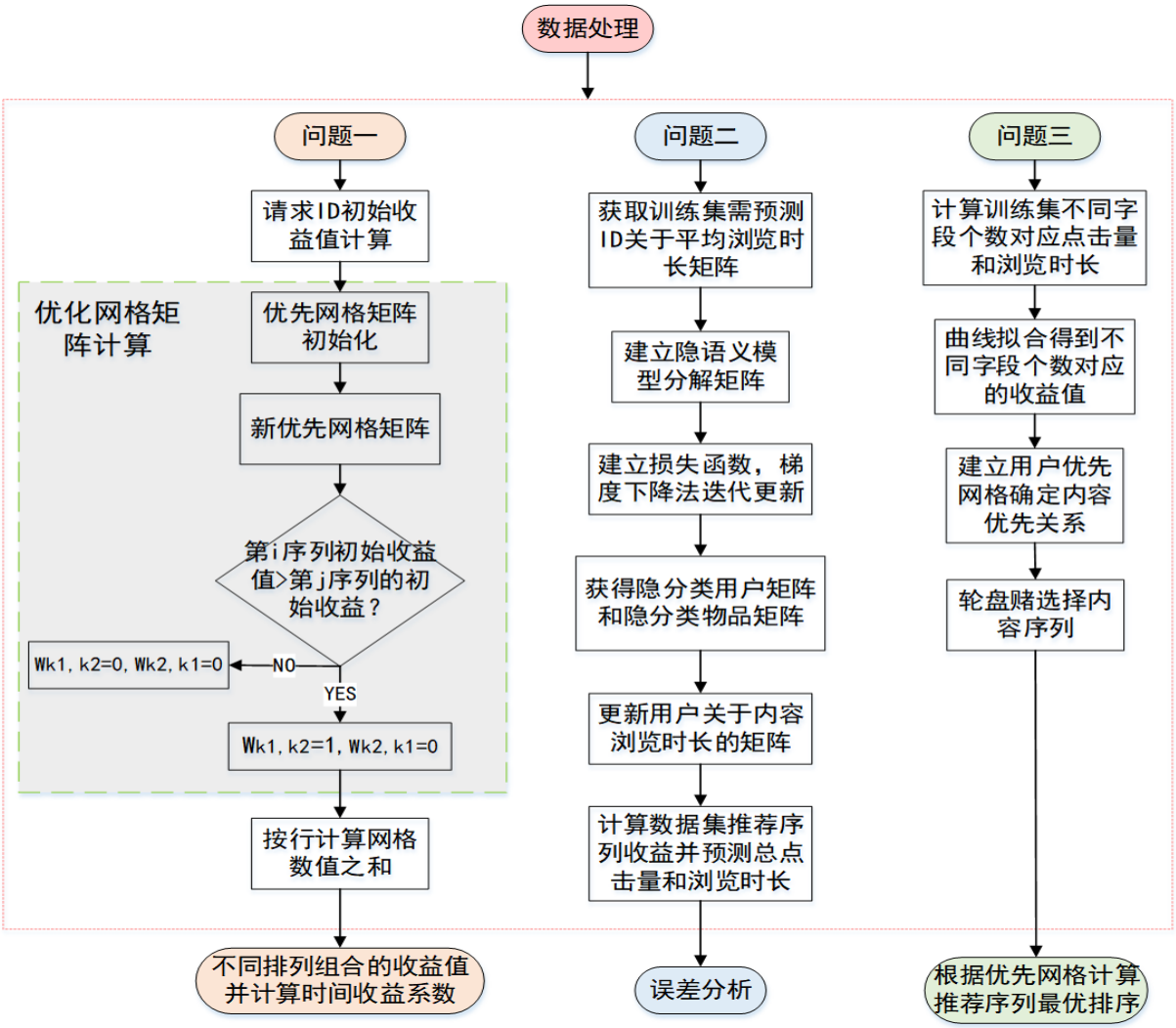


图 1：解题流程图

2. 问题分析

2.1. 问题一的分析

问题一要求建立评估推荐序列总点击收益（序列中单条内容的点击量之和）和总时长收益（序列中单条内容的浏览时长之和）的数学模型，以及如何根据点击收益和时长收益对综合收益进行量化，同时还需要考虑不同排列组合对收益大小的影响。

训练数据中可以得到每个 ID 的每一条推荐序列的点击量和浏览时长，在不考虑排列组合的情况下，每一条序列的初始收益值也是已知的。要确定不同排列组合对收益值的影响，可以考虑根据点击量和浏览时长产生的初始收益值更新推荐内容的排序，确定在已有的序列中，用户对那些推荐内容更加喜欢或者是更加喜欢这个推荐内容放在哪个位置，最终就可以得到不同排序对收益影响的大小。

同时训练数据中也给各个 ID 的推荐序列标准了时间，需要对不同时段的点击量进行统计分析，计算在不同时间对收益值的影响。

2.2. 问题二的分析

问题二要求基于问题一设计的数学模型，预测测试集（test_data.txt）中推荐序列的总点击量和总时长（单位为秒）。题目中训练数据的推荐内容包含了一级分类和二级分类内容，为了提高模型的

准确性，采用二级分类作为对各个内容分类的标签。

首先，按照问题一中的数学模型，建立各个用户的优先网格矩阵，获得各个用户对应的不同推荐内容各自的收益值；其次，根据训练数据获得各个用户每一个推荐内容的平均点击量和平均浏览时长；再者，建立隐语义模型，获得没有浏览记录的点击量和浏览时长；最后，考虑时间、排序等因素，预测每个请求 ID 对应推荐序列的点击量和时长之和，并填入附件 result.csv 中。

2.3. 问题三的分析

问题三中要求设计剪枝策略，目标是在 N 条候选内容中选择长度为 K（N≥K）的推荐序列，同时还需保证候选序列集更大可能包含最优序列，除此以外还需要分析模型的精准度和时间复杂度。若使用穷举法对每一条序列是收益进行评估，精准度会提升，但无疑增加了模型的计算量。因此需要设计一种剪枝策略，在保证精准度的同时，还要减少模型的计算量。

因此，首先需要对每个请求 ID 的训练内容进行统计分析，获得序列的最优长度；其次，根据问题一的模型，建立每个 ID 的优先网格矩阵，获得不同内容排序带来的收益；最后，通过每条内容的收益值设计每种分类内容的选择概率，再通过轮盘赌的方式选择推荐序列。同时，分析模型的时间复杂度，并检验模型的有效性。

3. 符号说明

下表列出了一部分符号和简称并给予了说明，文中出现的其它符号或简称会在出现的相应位置给予解释。

表 1：符号/简称说明

符号/简称	符号说明
CT^{I}	（content type）内容类别一级分类
$\mathrm{CT}^{\mathrm{II}}$	内容类别二级分类
N^{I}	一级分类个数
N^{II}	二级分类个数
ID_{ct}	内容 ID
ID_{us}	用户 ID
ID_{ps}	请求 ID
W	优先网格
E_0	只考虑点击次数和观看时长的收益
E_1	考虑推荐序列排序的收益
CN	点击数量
WT	观看时长
CN_{max}	用户序列的最大点击量
WT_{max}	用户序列的最大观看时长
RS	推荐序列
F	字段个数
$F_{P_i}^{k}$	第P_i个推荐序列的第 k 个字段
P	推荐序列个数
P_i	第 i 个推荐序列

4. 基本假设

假设 1：异常值处理后的数据均为准确数据；

假设 2：用户单次请求推荐服务提供的内容观看时长为内容的最大时长；

假设 3：一旦点击内容，观看时长大于 0；没有点击内容，观看时长效应 0；

假设 4：所有用户没有重复浏览相同的内容；

假设 5：每个内容只有一个一二级分类；

假设 6：内容的基础属性第一个位置数据为一级分类。

5. 数据处理

基于题目提供内容的基础属性（doc_info.txt），可以发现数据中存在多个异常数据，为更好的进行问题的解决，需要采取措施进行数据处理。同时，数据异常有几种表现形式，诸如：(1)内容类别没有一级分类和二级分类；(2)内容类别只有一级分类，没有二级分类；(3)单个内容存在多个一二级分类；(4)内容的基础属性存在异常符号。

针对上述异常值情况，对应做如下处理操作：

1、针对异常情况(1)，考虑数据的完整性，将一级和二级内容均设置为其他（处理之后分类为其他/其他）；

2、针对异常情况(2)，将二级分类设置为其他（处理之后分类为***/其他）；

3、针对异常情况(3)，取第一个分类作为单个内容的一二级分类（处理之后分类为：）；

4、存在异常符号的类别，也作为一个单独的分类，并按照异常情况(2)、(3)做数据处理。

按内容的基础属性的基础属性进行分类，二级分类一共 290 个项目（图形）、一级分类一共***个，按照一二级分类建立网络关系，网络关系图如下所示：

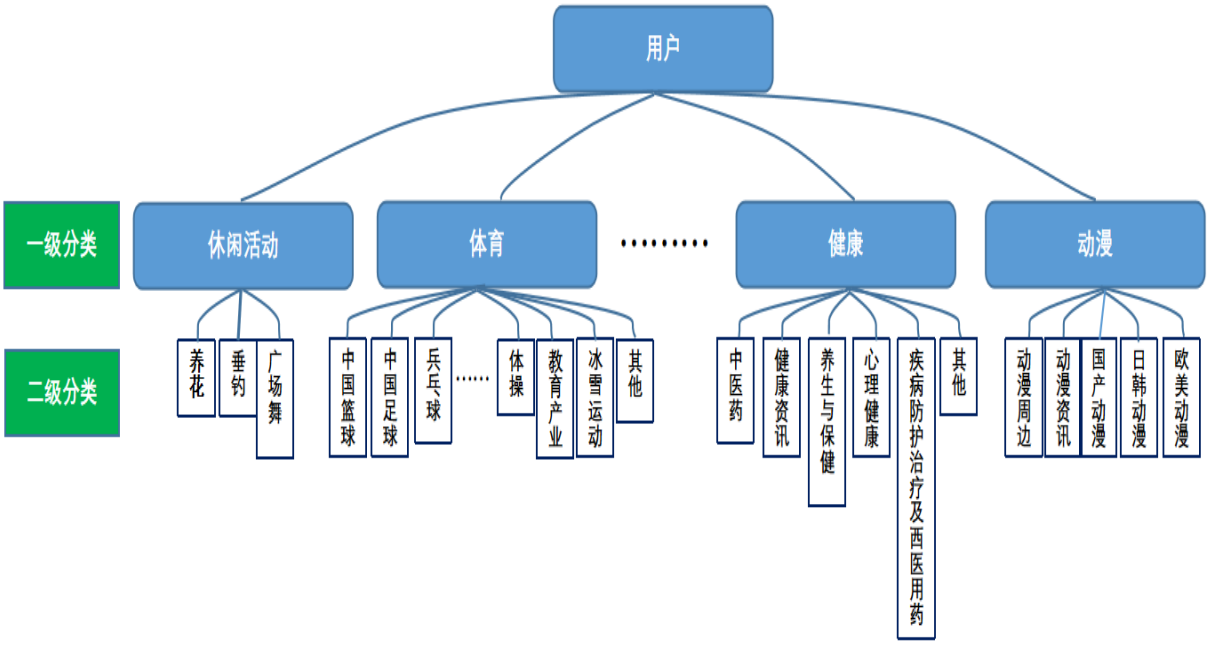


图 2：内容分类网络关系图

6. 问题一：推荐序列总点击和总时长收益评估模型

6.1. 问题一的建模流程

对于单个用户来说，考虑不同排列组合收益的大小，推荐序列总点击收益数学模型的建模流程如下：

表 2：问题建模流程

步骤编号	详细内容
1	按照二级分类将项目分成 290 个项目
2	290*290 的优先网格初始化 W

3	推荐服务返回的内容列表的点击量和观看时长
4	计算推荐序列的累计最大时长和最大点击量
5	将每一条内容列表的点击量和观看时长分别除以最大时长和最大点击量并相加并除以2，获得初始收益值 pi
6	更新优先网格，获得序列的优先关系，步骤如下： 6.1 初始化，用第一条序列更新优先网格，网格的初始收益值为 p1 6.2 如果 pk > pi（k= 1:290、i= 1:290） 6.2.1 获取内容 ID 的项目 6.2.2 根据内容分类优先关系，更新网格 6.3 如果 pk <=pi（k= 1:290、i= 1:290） 6.2.1 获取内容 ID 的项目 6.2.2 根据内容分类优先关系，把不存在优先关系的项目位置更新（W(j1,j2)=W(j2,j1)=0）
7	根据优先网格计算 290 个项目的优先关系
8	计算序列优先顺序得分需要交换才能得到的最优序列次数 N 除以（（n+1）*n/2）
9	上述结果加上初始收益值 pi 再除以 2

6.2. 模型的建立及求解

按二级分类对字段内容进行整理，整理公式如下：

$$N^{\text{II}} = \text{set}（\text{CT}^{\text{II}}）\tag{1}$$

其中，N^{II}表示二级分类个数 ；CT^{II}表示内容类别二级分类；set表示对字段内容进行去重。

经过计算，N^{II}的值为 290，并对内容进行依次编号。

初始化优先网格W如下式所示：

$$\text{W} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}_{290\times290}\tag{2}$$

矩阵的值表示优先关系，矩阵索引表示二级分类的内容，1 表示优先，0 表示不优先，若对称位置$W_{i,j} = W_{j,i} = 0$，表示两个内容之间没有优先关系。

推荐服务返回的内容列表的点击量和观看时长如式 3、4 所示：

$$CN_{P_i} = \sum_{j=1}^{P} CN_{F_{P_j}^k}\tag{3}$$

$$WT_{P_i} = \sum_{j=1}^{P} WT_{F_{P_j}^k}\tag{4}$$

计算推荐序列的累计最大时长和最大点击量如式 5、6 所示：

$$CN_{max} = \max\left(CN_{P_i}\right)\ \ (i = 1、2\cdots P)\tag{5}$$

$$WT_{max} = \max\left(WT_{P_i}\right)\ \ (i = 1、2\cdots P)\tag{6}$$

初始收益计算如下式所示：

$$\mathrm{E}_{0_i}=\frac{CN_{P_i}WT_{max}+WT_{P_i}CN_{max}}{2CN_{max}WT_{max}}\quad(i=1、2\cdots P) \tag{7}$$

更新优先网格，获得序列的优先关系。第P_i个推荐序列的第 k 个字段与内容类别的对应关系为：

$$F_{P_i}^{k}\leftrightarrow \mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k}\ （k=1、2、\cdots、\mathrm{F;i}=1、2、\cdots、\mathrm{P}） \tag{8}$$

其中$\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k}$表示第$P_i$个序列的第 k 个字段对应的优先网格位置，这就可以分为两种情况了。

（1）第一种情况：$\mathrm{E}_{0_{P_i}}>\mathrm{E}_{0_{P_j}}(j=1、2、\cdots、i)$，其中$\mathrm{E}_{0_{P_i}}$表示第 i 个序列的初始收益值；$\mathrm{E}_{0_{P_j}}$表示第 j 个序列的初始收益值。

$$W^{(k1,k2)}_{(\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k1},\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k2})}=1\ (k1=1、2、\cdots、F;k2=1、2、\cdots、k1;) \tag{9}$$

$$W^{(k1,k2)}_{(\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k2},\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k1})}=0 \tag{10}$$

$W^{(k1,k2)}_{(P_i,N)}$表示当字段$F_{P_i}^{k1}$在字段$F_{P_i}^{k2}$前面时，将字段对应内容的优先网格矩阵的位置（$\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k1},\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k2}$）变为 1。

（2）第二种情况：$\mathrm{E}_{0_{P_i}}\le\mathrm{E}_{0_{P_j}}(j=1、2、\cdots、i)$

$$W^{(k1,k2)}_{(\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k1},\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k2})}=1\ (W^{(k1,k2)}_{\left(\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k1},\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k2}\right)}=0\ \&\&\ W^{(k1,k2)}_{\left(\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k2},\mathrm{CT}^{\mathrm{II}}{}_{P_i}^{k1}\right)}=0) \tag{11}$$

根据优先网格计算 290 个项目的优先关系，计算每一行的累计数值，即为优先关系数值大小，再进行排序：

$$\mathrm{score}_i=\sum_{j=1}^{N^{\mathrm{II}}}W_{(\mathrm{i,j})} \tag{12}$$

其中score_i表示第 i 个二级分类内容的优先关系数值，即：$\mathrm{priority}=\mathrm{sort(score)}$，其中priority表示排序之后的二级分类内容编号。

$$E_e=1-\frac{N}{(F-1)^2} \tag{13}$$

$$E_1=\frac{E_1+E_e}{2} \tag{14}$$

其中，N表示序列通过移动编号达到最优排序的移动次数，E_e表示对序列进行交换的收益值；E_1表示序列的收益值。

随机选取 5 个内容，产生不同长度字段，随机取大小不同的初始收益值，作为案例，具体信息如下表：

表 3：案例信息

推荐序列编号	推荐序列(数值表示内容编号)	初始收益值
1	{2、1、3}	0.6
2	{5、3、1、2 、1}	0.8
3	{1、3、5、4}	0.5

其变化过程为$W_0\rightarrow W_1\rightarrow W_2\rightarrow W_3$，$W_0$表示初始化的优先网格矩阵，$W_1$、$W_2$、$W_3$分别表示新的推荐序列对优先网格矩阵的作用关系，则优先关系数值为$\mathrm{score}=[1\ 1\ 3\ 0\ 4]$、则优先关系为{5、3，1、2、4}或{5、3，2、1、4}。具体变化示意图如下：

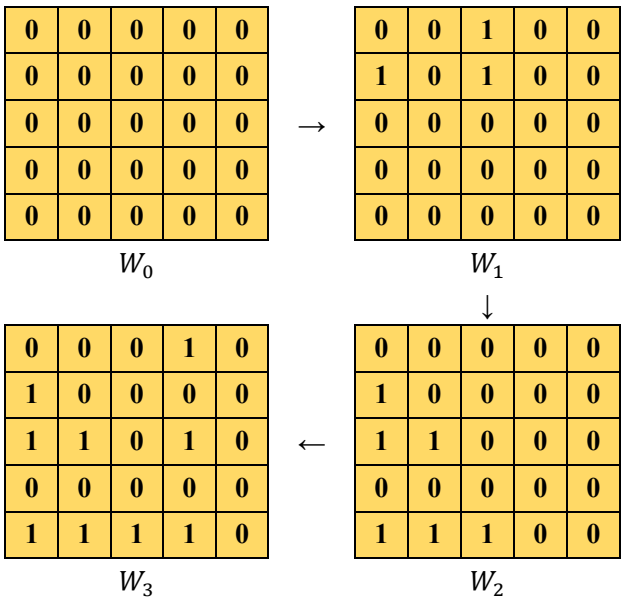


图 3：案例变化示意图

由于内容 1 和内容 2 的优先关系数值相同，故这两种排列组合的收益值是相同的。同时我们可发现$W_{4,2} = W_{2,4} = 0$，从score中我们可以发现，内容 2 是优于内容 4 的，但是差距不大。从推荐序列也可以发现，这是由于关于 4 的内容较少导致，只能由前面的推荐序列得到内容 4 的优先关系。

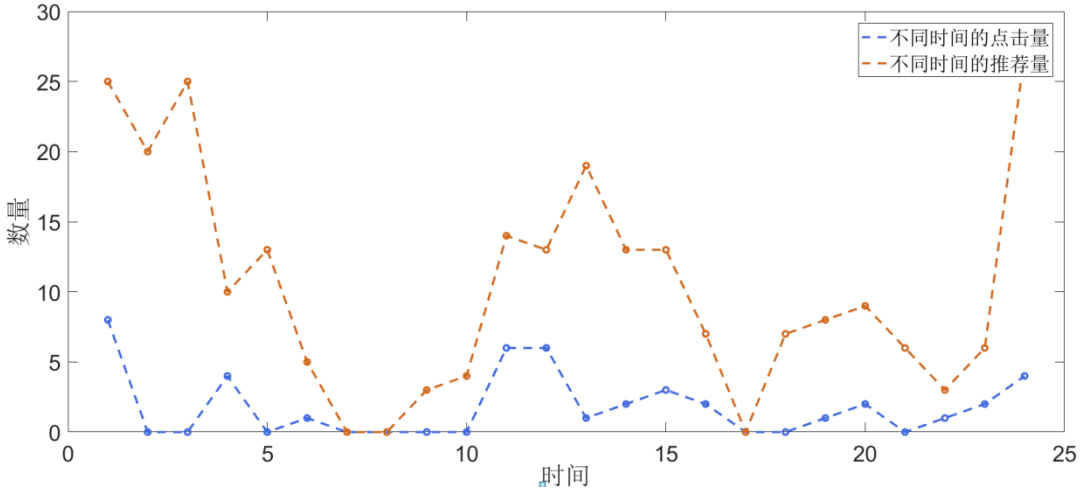


图 4：时间收益分布图

上图为 ID 号为 1385016772 从 0 点到 23 点总的推荐内容数量和总的点击数量，从中我们可以发现，不同时间的点击率是不同的，因此，我们需要设置时间收益τ_d来体现不同时间推送内容带来的收益。

$$P_i^d = \frac{CN_i^d}{RN_i^d}$$
$$\tau_{d_i} = \frac{P_i^d}{\max(P_i^d)} \quad (15)$$

其中，CN_i^d为时间为 i 时总的点击量，RN_i^d为时间为 i 时总的推荐量，P_i^d为对应的点击率，τ_{d_i}为时间为 i 是的收益值。

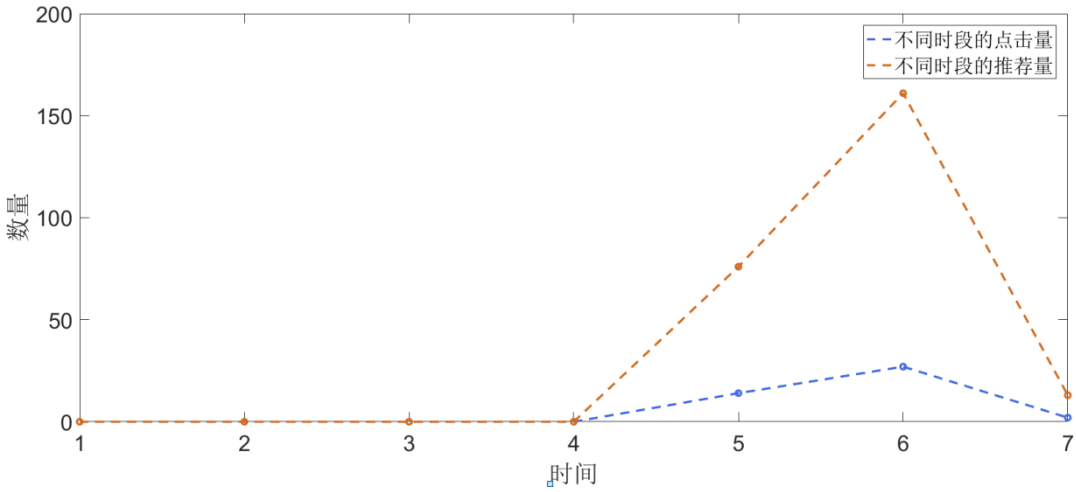


图 5：日期收益分布图

训练数据中为日期为 2021 年 10 月 20 日到 31 日之间的数据，我们把它的转化为以周为单位。以 ID 号为 1385016772 为例，数据如上图所示，其中横坐标编号代表周一至周天，我们可以发现，推荐内容的时间为周五、周六和周日，因此设置日期收益τ_w来体现不同时间推送内容带来的收益。

$$P_i^w = \frac{CN_i^w}{RN_i^w}$$
$$\tau_{w_i} = \frac{P_i^w}{\max(P_i^w)} \quad (16)$$

其中，CN_i^w为时间为周 i 时总的点击量，RN_i^w为时间为周 i 时总的推荐量，P_i^w为对应的点击率，τ_{w_i}为时间为周 i 是的收益值。

7. 问题二：预测推荐序列的总点击量和总时长

7.1. 考虑序列顺序的隐语义模型概述

一个完整的协同过滤推荐系统可以用下面的框架图：

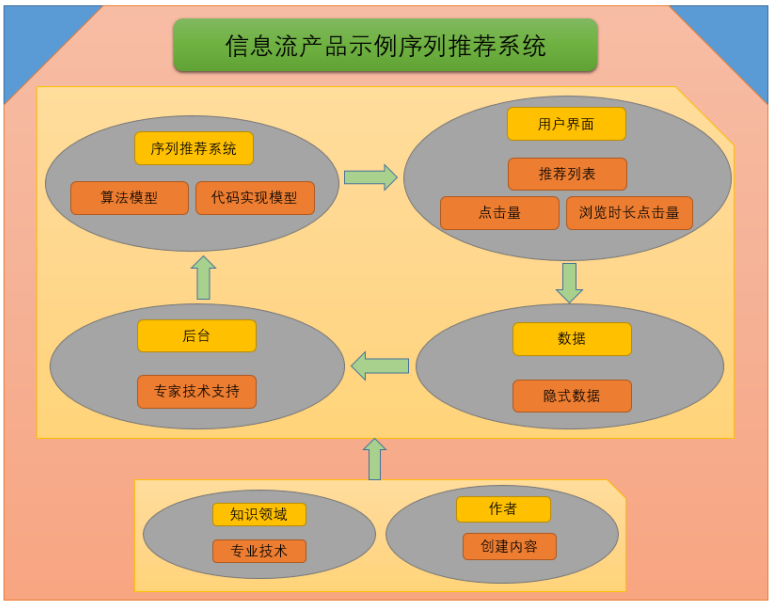


图 6：信息流产品示例序列推荐系统

一般可以根据算法原理把协同过滤算法分为三种，如表 4。可以看到基于用户的协同过滤算法(User-based)和基于物品的协同过滤算法(Item-based)是最经典领域模型算法。矩阵分解模型包括有

SVD、MF、LFM 以及 PMF 等算法。基于图的模型中研究最有代表性的是随机游走算法[15]。

表 4：协同过滤算法分类

模型	代表性算法
基于领域的模型	基于用户的协同过滤 基于物品的协调过滤
矩阵分解模型	奇异值分解 隐语义模型
基于图的模型	随机游走 Tangent、PathFusion

下面将介绍最通用的三种模型算法：基于用户的协同过滤、基于物品的协同过滤、隐语义模型。因此将算法说明中所需要的符号进行设定，表格如下：

表 5：使用到的公式符号

符号	定义
A(a)、B(b)、C(c)、u	表示用户 A(a)、用户 B(b)、C(c)、u
$\vec{a}$，$\vec{b}$	物品空间中的表示用户的用户向量
$R_{a,k}$	表示用户 a 对物品 k 的评分
I_a	表示用户 a 消费或评分的物品集合
I_{ab}	表示用户 a 和用户 b 共同消费或评分的物品集合
$\overrightarrow{R_a}$	表示用户 a 所有物品评分的平均值
i，j	表示物品 i 和物品 j
$\vec{\imath}$，$\vec{\jmath}$	用户空间中的表示物品的物品向量
$\overrightarrow{R_\imath}$	表示物品 i 评分的平均值
U_{ij}	表示物品 i 和物品 j 都已经评价过的用户集合
$Sim(a，b)$	表示用户 a 和 b 的相似度
$Sim(a，b)$	表示物品 i 和 j 的相似度
P，Q	隐语义模型矩阵分解后的 P 和 Q 参数矩阵
r_{ui}	表示用户 u 对 i 的评分
$\widehat{r_{u\imath}}$	表示用户 u 对 i 的与非评分
α	梯度下降法的学习率
λ	正则化参数
F	LFM 的隐因子个数

7.1.1 基于用户的协同过滤算法

基于用户的协同过滤算法(User-based Collaborative filtering)是最早提出的推荐算法之一，一开始应用于邮件过滤系统，后面被 Grouplens 应用于新闻推荐[14]。这种算法是协同过滤算法最直接的体现，作用过程示意图如图 7。



图 7：基于用户的协同过滤算法形象表示

从图中可以直观的看到，用户 A 和用户 C 同时对物品 A 和物品 C 感兴趣，可以将他们看作有着相似偏好的用户，用户 A 没有使用过物品 D，User-CF 推荐系统就会自动将物品 D 可以推荐给用户 A[14]。可以将基于用户的协同过滤算法表述为 3 个步骤：

步骤 1. 计算各用户与活动用户的相似程度；

步骤 2. 根据各用户与活动用户的相似度确定该活动用户的邻居；

步骤 3. 将邻居的得分的加权值作为活动用户的评分。

要找到指定用户的相似用户，需要用算法来衡量用户之间的相似程度。下面介绍 3 种经典的计算方式: 余弦相似性、相关相似性和修正的余弦相似性。

1、余弦相似性：用户 a 和用户 b 之间的余弦相似性为:

$$sim(\mathrm{a},\mathrm{b}) = \cos(\vec{a}\cdot\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{||a||_2\cdot||b||_2} = \sum_k \frac{R_{ak}}{\sqrt{\sum_{l\in I_a} R_{il}^2}}\cdot\frac{R_{bk}}{\sqrt{\sum_{l\in I_a} R_{bl}^2}} \tag{17}$$

式中分母用于归一化，这样打分项目较多的用户在计算式时候不会优先于其他的用户。

2、相关相似性：使用皮尔逊相关系数(Pearson)来度量两个用户之间的相似性:

$$\mathrm{sim}(\mathrm{a},\mathrm{b}) = w_{a,b} = \frac{\sum_{k=1}^{K}(R_{a,k}-\bar{R}_a)(R_{b,k}-\bar{R}_b)}{\sqrt{\sum_{k=1}^{K}(R_{a,k}-\bar{R}_a)^2}\sqrt{\sum_{k=1}^{K}(R_{b,k}-\bar{R}_b)^2}} \tag{18}$$

其中 K 这个变量是用户 a 和用户 b 相同的评分数量。

3．修正的余弦相似性：考虑一种情况，比如一些用户评分比较极端，而另一些用户评分比较温和，如果采用下面公式进行计算，这两种用户的差异会表现的很大，即 使两者的兴趣其实是很相似的。可以用下面的式子进行修正:

$$\mathrm{sim}(\mathrm{a},\mathrm{b}) = w_{i,j} = \frac{\sum_{k\in l_{a,b}}(R_{a,k}-\bar{R}_a)(R_{b,k}-\bar{R}_b)}{\sqrt{\sum_{k\in I_a}(R_{a,k}-\bar{R}_a)^2}\sqrt{\sum_{k\in I_b}(R_{b,k}-\bar{R}_b)^2}} \tag{19}$$

7.2.1 基于物品的协同过滤算法

在用户数目比物品数目少的情况下，基于用户的协同过滤算法运行得不错。但是随着一个在线系统的用户越来越多，基于用户的协同过滤算法在这种情况下时间复杂度非常高，寻找指定用户的邻居花费时间很长。为了解决这个问题，Linden 等人提出了基于物品的协同过滤算法(Item based)[15]。

基于物品的协同过滤算法和基于内容的推荐系统在形式上看起来很像，都是找寻相似的物品，但是两者有着本质上的不同。基于内容的推荐系统是通过物品的内在属性来寻找相似的物品。而基于物品的协同过滤算法是分析用户的行为数据来计算物品之间的相似程度。比如很多用户会同时购买《推荐系统实战》和《数据挖掘导论》两本书籍，那么就可以认为这两本书籍的相似度是很大的，因此可以为只买了《推荐系统实战》的用户推荐《数据挖掘导论》这本书[15]。形象表示如图 8。

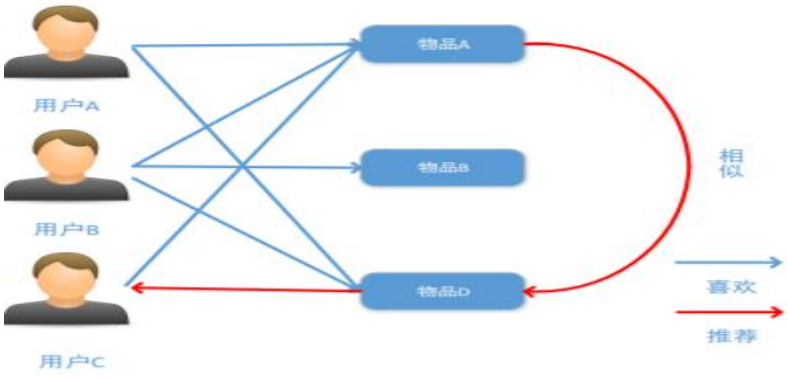


图 8：基于物品的协同过滤算法形象表示

Amazon 研究人员提出基于物品的协同过滤算法主要考虑到了在他们的数据库当中，物品的数量是相对不会有很大变化的，而且总的物品数量相对于不断增加的用户数量是比较少的，因此计算机系统可以使用较小的时间复杂度就可以计算得到物品之间的相似度[13]。而且在用户不断加入的时候，整个推荐系统的扩展性是很强的，因为不需要用大量的时间计算新增加的用户的最近邻居。所以这种算法效率较高，而且准确性也有很大的保证。一经提出，就成为了推荐系统的宠儿[14]。

基于物品的协同过滤系统一般也分为三个步骤：

步骤 1. 通过用户-物品矩阵求出不同物品之间的相似度；

步骤 2. 得到指定物品的相似度最大的 N 个邻居物品；

步骤 3. 为购买了指定物品的用户推荐指定物品的邻居物品，前提是用户没有买过该物品。

如基于用户的算法一样，步骤 1 中的计算不同物品之间的相似度也是基于物品的算法的核心步骤。选择不同的算法计算物品的相似度，会得到不同的推荐结果。计算两个物品的相似度，首先要分离出对这两个物品所有已经评分的用户，然后利用相似性度量算法计算得到两个物品的相似性。同样采用 3 种经典的计算方式进行计算，具体计算公式见下述若干式子。

1、余弦相似性：设有物品 i 和物品 j，他们之间的相似性如式所示:

$$sim(i,\mathrm{j}) = \cos(\vec{\imath}\cdot\vec{j}) = \frac{i\cdot\vec{j}}{||i||_2\cdot||j||_2} \tag{20}$$

2、相关相似性：这种算法一般采用皮尔逊系数作为物品之间的相似系数。

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_i)(R_{u,j}-\bar{R}_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_j)^2}} \tag{21}$$

3、修正的余弦相似性：各个用户的评分尺度不一样，需要对此进行修正。建议将用 户评分减去该用户所有评分的平均值，经过修正之后，不论用户是极端型的评分还是 温和型的评分，都可以得到使得计算更加准确。具体的计算公式如下所示:

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_u)(R_{u,j}-\bar{R}_u)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_u)^2}} \tag{22}$$

在得到物品的相似度之后，最后需要计算的用户 u 对物品 i 的评分的预测值，计算公式如下（21）:

$$S_{u,i} = \frac{\sum_{k\in N_u(i)} sim(k,j)\cdot R_{u,k}}{\sum_{k\in N_u(k)}|sim(k,j)|} \tag{23}$$

公式中$N_u(i)$表示表示与物品 i 相似度最高的 n 个物品的集合。这种方法在评分预测中是最常用的方法，被称为权相加法。

7.3.1 隐语义模型

矩阵分解模型是近年来推荐系统算法研究的热点方向。这个算法的主要思想是采用了矩阵论中的矩阵分解算法。推荐系统算法研究学者一直希望将矩阵奇异值分解(Singular value

decomposition ,SVD)引入到推荐系统中，但是传统的 SVD 有着难以克服的局限性，所以进展一直较慢[13]。

在 Netflix Prize 比赛中，Simon Funk 基于最优化理论对 SVD 算法进行改进，后来这个算法被比赛的冠军 Koren 称为隐语义模型。由于这个新算法在比赛中的优异表现，他也受到越来越多的研究人员的关注[14]。

在隐语义模型之前，SVD 算法一直是矩阵分解领域的研究重点。Sarwar 等人将其应用到了推荐系统中。但是 SVD 算法在大规模矩阵上分解的时间复杂度过高，一般是 $O(N^3)$，当前的商业公司都是千万级别的用户—物品矩阵，这个速度是不能容忍的。另外还需要对大规模稀疏矩阵进行补全，也需要耗费很大的人力资源[15]。这些缺点使得 SVD 算法很难在实际的推荐系统进行商用。

在 Netflix prize 开始后，为了解决上述 SVD 算法的缺陷，在 2006 年，Simon Funk 在提出了一个算法—Funk-SVD 算法，这个算法被后来的 Netflix Prize 冠军 Koren 称作隐语义模型(Latent Factor Model)[16]。

隐语义模型的核心公式是：

$$preference(u,i) = R_{ui} = P_u^T Q_i = \sum_{k=1}^{K} p_{u,k} q_{i,k} \tag{24}$$

这个分解过程可以用图对矩阵的分解进行形象化表示，如图 9 所示：

	Item 1	Item 2	Item 3	Item 4
User 1	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	$R_{1,4}$
User 2	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	$R_{2,4}$
User 3	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$	$R_{3,4}$

R

=

	Class 1	Class 2	Class 3
User 1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$
User 2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$
User 3	$P_{3,1}$	$P_{3,2}$	$P_{3,3}$

P

×

	Item 1	Item 2	Item 3	Item 4
Class 1	$Q_{1,1}$	$Q_{1,2}$	$Q_{1,3}$	$Q_{1,4}$
Class 2	$Q_{2,1}$	$Q_{2,2}$	$Q_{2,3}$	$Q_{2,4}$
Class 3	$Q_{3,1}$	$Q_{3,2}$	$Q_{3,3}$	$Q_{3,4}$

Q

图 9：矩阵分解示意图

其中 R 是用户—物品矩阵，矩阵值 R_{ij} 表示用户 i 对物品 j 的兴趣度。P 矩阵是用户—隐分类矩阵，矩阵值 P_{uk} 表示的是用户 u 对第 k 个隐类的关系。Q 矩阵是隐分类物品矩阵，矩阵值 Q_{ki} 表示的是第 k 个隐类和物品 i 之间的关系。

P 和 Q 矩阵的计算需要通过最优化损失函数理论算法对数据集进行计算得到。

数据预处理完成后，通过计算公式（23）来计算 P 和 Q 参数矩阵，在大多数的研究中，评测指标一般定为均方根误差 RMSE(root mean square error)，然后通过随机梯度下降法使得损失函数最小化。

损失函数的公式为：

$$C = \sum_{(u,i)\in K} (r_{ui} - \hat{r}_{ui})^2 = \sum_{(u,i)\in K} (r_{ui} - \sum_{k=1}^{K} p_{u,k} q_{i,k})^2 + \gamma ||p_u||^2 + \gamma ||q_i||^2 \tag{25}$$

公式（23）中，K 表示用户—物品集合，γ项表示正则化项，用来防止过拟合。γ 值一般是通过实验来进行确定。

接下来采用随机梯度下降法来最小化 RMSE。

首先对 p_{uk} 和 q_{ik} 分别求偏导，公式为：

$$\begin{aligned} \frac{\partial C}{\partial p_{uk}} &= -2q_{ik}e_{ui} + 2\gamma p_{uk} \\ \frac{\partial C}{\partial p_{ik}} &= -2q_{uk}e_{ui} + 2\gamma q_{ik} \end{aligned} \tag{26}$$

随机梯度下降法，将参数随着最快的下降方向进行下降，公式为：

$$\begin{aligned} p_{uk} &= p_{uk} + \alpha(q_{ik}e_{ui} - \gamma p_{uk}) \\ p_{ik} &= p_{ik} + \alpha(q_{uk}e_{ui} - \gamma q_{ik}) \end{aligned} \tag{27}$$

其中，α 这个参数是 SGD 的学习率，这个参数决定了每一步下降的速度，这个参数需要通过多次试验得到。每次迭代都需要将学习率减少，这是梯度下降法要求的，主要是为了使算法尽快收敛。

当求出 P 和 Q 矩阵之后，就可以对一个用户没有使用过的物品进行预测评分。这里生成的评分不单是用来最后的评测，前期的随机梯度下降也用到了这个预测评分公式。

7.2. 考虑序列顺序的隐语义预测模型

根据（train_data.txt）作为训练集计算每个用户对应每个内容的点击次数和浏览时长，通过隐语义模型计算用户没有观看历史的内容的点击次数和浏览时长，得到每一个用户对应每一个内容的点击次数矩阵和浏览时长矩阵，根据优先网格计算推荐序列的收益值，再通过公式（26），则可计算测试集每一个推荐序列的总点击量和点击时长。

$$CN_{p_i}^{S} = \tau_s \tau_w \tau_d \cdot \sum_{i=1}^{F_{p_i}} (\mathrm{CN}_{F_{P_i}^k} \cdot E_{1_{p_i}})$$

$$WT_{p_i}^{S} = \tau_s \tau_w \tau_d \cdot \sum_{i=1}^{F_{p_i}} (\mathrm{WT}_{F_{P_i}^k} \cdot E_{1_{p_i}}) \quad (28)$$

其中$CN_{p_i}^{S}$和$WT_{p_i}^{S}$分别表示第p_i个推荐序列的总点击量和总浏览时长，$\mathrm{CN}_{F_{P_i}^k}$表示第p_i个推荐序列的第 k 个字段的点击量，$WT_{F_{P_i}^k}$表示第p_i个推荐序列的第 k 个字段的浏览时长；τ_s表示序列顺序收益值系数；τ_w表示日期收益系数；τ_d表示时间收益系数。

将预测结果填入 result.csv 中给定位置的总点击量和总时长。

7.3. 误差分析和结果写入

在训练数据集中随机选取了 500 个推荐序列，其中包含了若干个用户，预测结果如图 10 所示。

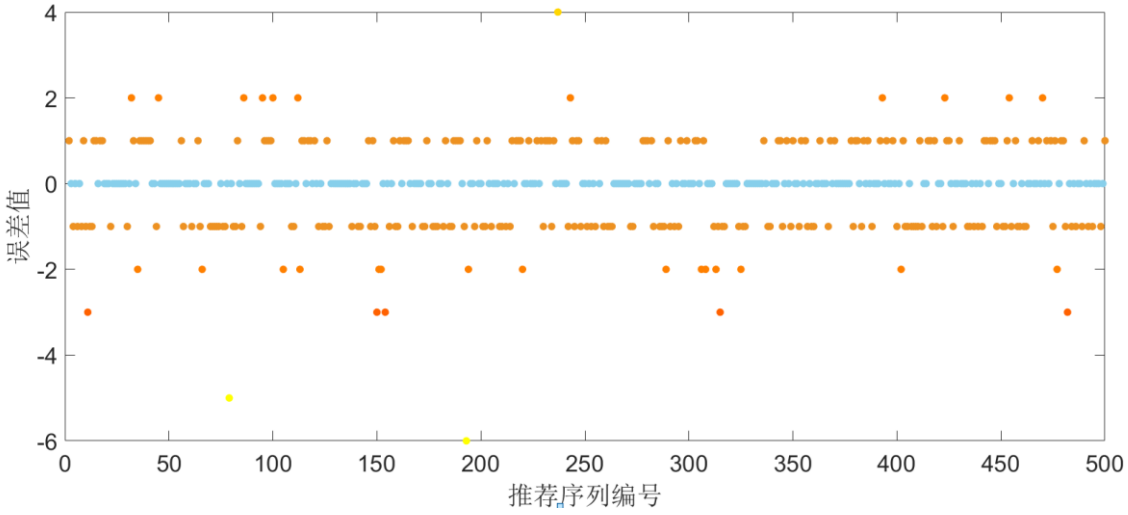


图 10：各推荐序列编号下的误差值

图中的数据为预测的点击量减去实际的点击量，大于 0 表示，预测的点击量多于实际的点击量，小于 0 表示预测的点击量少于实际的点击量，等于 0 表示预测的点击量和实际的点击量相同。从图中我们可以发现：93%的预测点击量与实际的点击量偏差是小于等于 1，误差在可控的范围之内，说明设计的模型具有有效性。

同时我们选取 1679 个数据进行统计分析各个误差值的数量分布，结果如表 6 所示，并用 MATLAB 的正态分布拟合工具对数据进行拟合，结果如图 11 所示。从图中我们也可以发现，大量的数据分布

在[-2,2]这个区间上，验证了我们模型的有效性

表 6：不同点击量的误差值

误差值	数量	误差值	数量	误差值	数量
-6	1	-1	398	4	11
-5	2	0	672	3	38
-4	7	6	2	2	77
-3	32	5	0	1	355
-2	84				

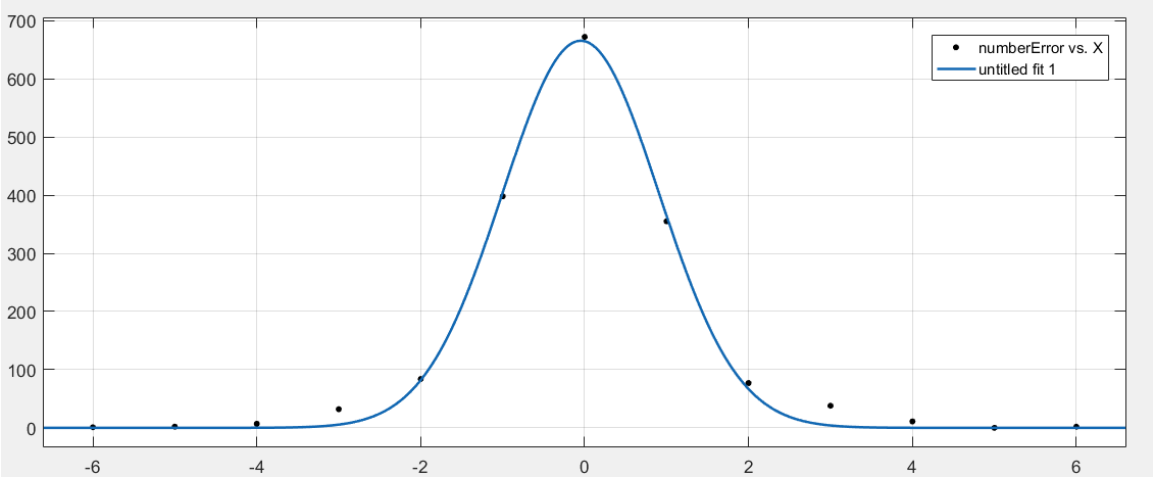


图 11：曲线拟合图

同时，基于问题一设计的数学模型，预测了测试集（test_data.txt） 中推荐序列的总点击量和总时长（单位为秒），并将预测结果写入 result.csv。部分预测数据结果表如下所示：

表 7：部分预测数据结果

请求 ID 编号	总点击量	总浏览时长
719708291_1635480753679_2960	1	339
679978594_1635491281923_1780	0	0
781051348_1635478916952_3923	0	0
1223262997_1635514428746_6191	1	33
1210159908_1635487197871_9372	1	48

8. 问题三：剪枝策略下精确评估最优序列收益

8.1. 问题三的建模流程

针对问题三的求解采用下述求解思路：

步骤 1：根据训练数据计算不同字段个数对应的点击数量和浏览时长；

步骤 2：用正态分布拟合数据；

步骤 3：得到不同字段个数对应的收益值 E2；

步骤 4：建立用户的优先网格确定内容的优先关系；

步骤 5：轮盘赌选择内容序列；

步骤 6：根据优先网格计算推荐序列的最优排序；

8.2. 模型的建立及求解

借助正态分布进行数据拟合需要给出正态分布的函数如式（29）：

$$\mathrm{f(x)}=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu^2)}{2\sigma^2}} \tag{29}$$

假设数据点{x_i、y_i}（i=1,2,3,...N）符合正态分布曲线，对其进行拟合（曲线拟合不同于分布拟合，需要乘以幅度 A），给出损失函数式（30），对损失函数 J0 求解即可实现参数估计。

$$J_0=\sum_{i=1}^{N}|f(x_i)-y_i|^2=\sum_{i=1}^{N}|\frac{A}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu^2)}{2\sigma^2}}-y_i|^2\mathrm{f(x)}=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu^2)}{2\sigma^2}} \tag{30}$$

由于求导比较复杂（可以借助 Mathmatica/Maple），因此这里换一个思路：如果$e^x\to y$，则$x\to lny$，则重新定义准则函数：

$$J_1=\sum_{j=1}^{N}\left|ln\frac{A}{\sqrt{2\pi}\sigma}-\frac{(x_i-\mu)^2}{2\sigma^2}-lny_i\right|^2=\sum_{i=1}^{N}\left|-\frac{x_i^2}{2\sigma^2}+\frac{\mu x_i}{\sigma^2}-\left(\frac{\mu^2}{2\sigma^2}-ln\frac{A}{\sqrt{2\pi}\sigma}\right)-lny_i\right|^2 \tag{31}$$

此时，变成了{x_i、$ln\,(y_i)$}的一元二次最小二乘拟合问题。假设拟合结果为：

$$p_1X_i^2+p_2X_i+p_3=lny_i \tag{32}$$

对应的参数估计为：

$$\begin{gathered}\sigma^2=\frac{-1}{2p_1}\\ \mu=\sigma^2p_2\\ \mathrm{A}=e^{\left(\frac{\mu^2}{2\sigma^2}+p_3\right)}\times\sqrt{2\pi\sigma^2}\end{gathered} \tag{33}$$

使用正态分布去拟合不同字段个数的点击数量，计算不同字段个数对应的曲线面积

$$S_k=\int_{k-1}^{k}f(k)\quad（\mathrm{k}=1、2、\cdots、\mathrm{F}） \tag{34}$$

其中S_k表示不同字段个数对应的概率值，k 的值为max (S_k)对应的 k 值。

在确定选取的字段个数之后，就需要选取字段对应的内容。假设有 N 条候选内容，二级分类内容可能有N_1条，优先网格只能确定用户不同二级内容的优先关系，无法确定同一种分类内容的优先关系，因此，这里我们通过轮盘赌的方式进行选取。建立N_1个分类内容的优先网格，确定内容的优先关系，确定 N 条候选内容的排名，记为$\mathrm{V}_i(i=1、2、\cdots N)$.

$$P_i=\left[\frac{\sum_{j=1}^{i-1}(\max(V)-V_j)}{\sum_{j=1}^{N}V_k}\quad\frac{\sum_{j=1}^{i}(\max(V)-V_j)}{\sum_{j=1}^{N}V_k}\right] \tag{35}$$

其中，P_i表示每一条内容的概率分布区间，即：

$$\mathrm{R}_{1\times k}=\left[rand_1()、rand_2()、\cdots rand_k()\right] \tag{36}$$

其中，$\mathrm{R}_{1,i}$对应的值落在P_i所属的区间即为选取的对应的内容。

8.3. 剪枝策略复杂度分析

在计算机科学中，算法的时间复杂度是一个函数，它定性描述该算法的运行时间。这是一个代表算法输入值的字符串的长度的函数。时间复杂度用O表述，不包括这个函数的低阶项和首项系数[5]。使用这种方式时，时间复杂度可被称为是渐近的，亦即考察输入值大小趋近无穷时的情况。在问题三中，我们考虑的是最坏时间复杂度，剪枝策略需要用到的模型如下表所示：

表 8：剪枝策略模型及其时间复杂度

模型	时间复杂度
曲线拟合拟合模型	$O(nlog(n))$
概率求解模型	$O(n)$
优先网格模型	$O(n^2)$
内容选取模型	$O(n)$

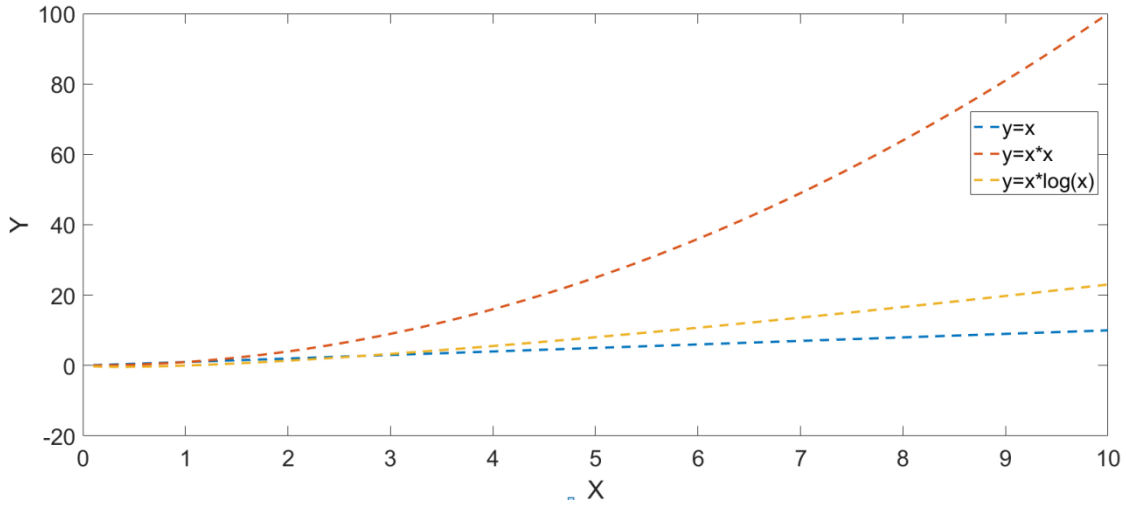


图 12：时间复杂度曲线图

从上图我们可以发现当n取值较大时，$n^2 > nlog(n) > n$，故剪枝策略的时间复杂度为$O(n^2)$。

8.4. 剪枝策略的精准度检验

为了检验剪枝策略的精准度，以 ID 号为 1385016772 为例，从 290 个二级分类的条候选内容中，在训练数据集中寻找收益值最大的序列，作为目标序列 K。

基于剪枝策略计算最优的序列K'以及序列的长度。若通过剪枝策略得到的序列包含序列 K，即$K \in K'$，则记为 0，否则记为 1。

判断结果如图 13 所示，我们可以发现：绿色表示找到了最优序列，红色空心圆圈则表示没有找到最优序列，在产生的 500 个序列中仅有 56 个没有包含给定的最优序列，由此计算出剪枝策略的精准度可以达到 90%左右。

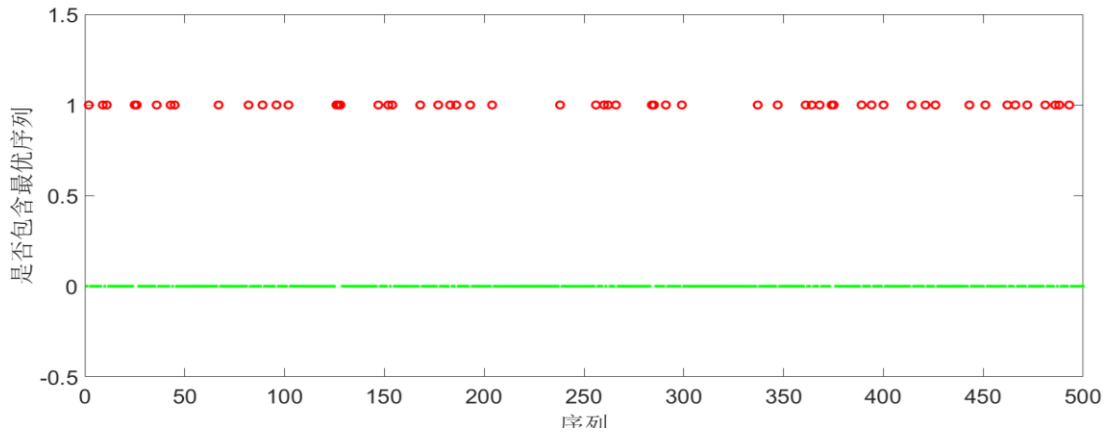


图 13：不同序列是否包含最优序列判断图

9. 总结和展望

近些年，推荐系统在商业应用中已经非常普遍。包括像电子商务公司、内容提供公司都有着自己研制的推荐系统，例如国外的 Amazon、youtube，国内的京东、腾讯视频等等，这些公司都通过自己的推荐系统获得了丰厚的商业回报[16]。因为有着强大的商业驱动力，推荐系统的发展如火如荼。

本文首先介绍了推荐系统的发展现状，并详细阐述了其中比较知名的算法，包括基于邻域的协同过滤算法、矩阵分解模型等。另外还介绍了推荐系统的评价标准以及用户数据获取和表示这两个推荐系统的重要步骤。

优先矩阵模型能够根据用户推荐序列的点击量，实时更新用户推荐序列中的内容的喜爱程度，同时还能分析不同的排列组合对用户的影响。

隐语义模型能够让用户对没有使用过的物品进行预测评分，但隐语义模型在实际使用中有一个困难，那就是它很难实现实时推荐。经典的隐语义模型每次训练时都需要扫描所有的用户行为记录，这样才能计算出用户对于每个隐分类的喜爱程度矩阵 P 和每个物品与每个隐分类的匹配程度矩阵 Q。而且隐语义模型的训练需要在用户行为记录上反复迭代才能获得比较好的性能，因此 LFM 的每次训练都很耗时，一般在实际应用中只能每天训练一次，并且计算出所有用户的推荐结果。从而隐语义模型不能因为用户行为的变化实时地调整推荐结果来满足用户最近的行为。隐语义模型和优先矩阵相结合，就能实时地根据用户的浏览情况，更新不同内容带来的收益，做到实时更新。最后设计剪枝策略，能够在 N 条候选内容中，快速地选择选择长度为 K（N≥K）的最优推荐序列，减小模型的计算时间。

总而言之，文中建立的推荐序列收益评估模型，能够快速地计算不同序列排列组合带来的收益，并且更够实时地更新，并快速地选取最优序列推荐给用户。

参考文献

[1] 温雯,梁方宇.基于用户潜在状态及依赖关系学习的时序行为推荐[J/OL].计算机应用:1-8[2022-01-20].

[2] 余皑欣,冯秀芳,孙静宇.结合物品相似性的社交信任推荐算法[J/OL].计算机科学:1-13[2022-01-20].

[3] 陈壮,邹海涛,郑尚,于化龙,高尚.基于用户覆盖及评分差异的多样性推荐算法[J/OL].计算机科学:1-10[2022-01-20].

[4] 王苑丞,韩林志.模因论视域下抖音平台信息流广告的创作研究[J].中国广告,2022(01):70-74.

[5] 陈靓,钱亚冠,何志强,关晓惠,王滨,王星.深度卷积神经网络的柔性剪枝策略[J/OL].电信科学:1-17[2022-01-20].

[6] 高秋英,王丽丽,王荣忠.最小二乘法曲线拟合及优化算法研究[J].工业控制计算机,2021,34(11):100-101.

[7] 刘欢.基于大数据的信息流营销探讨[J].中国管理信息化,2021,24(22):92-94.

[8] 王伟,蔡文生,邓科.利用最小二乘法拟合水位流量关系曲线探讨[J].陕西水利,2021(11):51-52+58.DOI:10.16747/j.cnki.cn61-1109/tv.2021.11.017.

[9] 郭静宁.深度卷积神经网络模型剪枝方法研究[D].北京交通大学,2020.DOI:10.26944/d.cnki.gbfju.2020.001651.

[10] 张镇. 卷积神经网络通道剪枝算法研究[D].西安电子科技大学,2020.DOI:10.27389/d.cnki.gxadu.2020.000642.

[11] 甘子美. 短视频 App 算法推荐智能化对用户心流体验的影响研究[D].广东外语外贸大学,2020.DOI:10.27032/d.cnki.ggdwu.2020.001056.

[12] 纪煜东. 对采用个性化推荐的移动端视频平台的优化策略[D].江南大学,2019.

[13] 王宁,沈正一,崔德龙,刘晴瑞.结合协同过滤与隐语义模型的视频推荐策略[J].湖南工程学院学报(自然科学版),2021,31(04):40-43.DOI:10.15987/j.cnki.hgbjbz.2021.04.007.

[14] 黄伟建,顾明星,黄远.结合隐语义模型与改进用户信任的协同过滤推荐算法[J].现代电子技术,2021,44(12):141-144.DOI:10.16652/j.issn.1004-373x.2021.12.030.

[15] 张小雷. 基于协同过滤和隐语义模型的图书推荐算法研究与实现[D].阜阳师范大学,2021.DOI:10.27846/d.cnki.gfysf.2021.000168.

[16] 周熙然. 基于隐语义模型的推荐算法研究与应用[D].西安科技大学,2021.DOI:10.27397/d.cnki.gxaku.2021.000884.

https://www.saikr.com/vse/bigdata2026

为了促进大数据人才培养，向各行业输送具有大数据创新能力和实践能力的高端人才，提升高校毕业生的就业竞争力，全面推动行业发展及人才培养进程，MathorCup数学应用挑战赛组委会决定在“MathorCup数学应用挑战赛”中增加大数据专题赛项——“大数据竞赛”。竞赛秉承研究会创始人华罗庚教授提出的数学与实践相结合的宗旨，通过大数据分析总结科技、经济与社会发展规律，利用算法工具来高质量解决实际问题。竞赛以企业真实场景数据为基础，面向全国高校研究生、本科生和专科生，共同探索数据科学的应用实践，推动产学研用协同发展。

第六届MathorCup数学应用挑战赛于2025年10月24日至10月31日举行，本次竞赛共有539所高校近1万名学生报名参赛。参赛高校覆盖清华大学、北京大学、南开大学、上海交通大学、复旦大学、南京大学、中国科学技术大学等35所985高校， 90所211高校。除中国大陆高校外，竞赛也吸引了来自南安普顿大学、格拉斯哥大学、俄罗斯国立管理大学、卡迪夫大学、伊利诺伊大学等境外高校参赛。

本竞赛的最大特点是结合企业实际问题出题，题目一般来源于各行业经过适当简化加工的实际问题，竞赛题目会和企业进行征稿，并每年举办"数学在企业中的应用研讨会"，邀请阿里巴巴、百度、浪潮等资深算法专家作报告，分享数学在实际问题中的应用。已经建立合作关系的企业有：京东、顺丰科技、杉树科技、阿里巴巴等企业。

竞赛设置金银铜奖，采用线上或线下的方式进行答辩，评出金奖1支队伍（奖金10000元税前）、银奖2支队伍（奖金5000元税前）、铜奖5支队伍（奖金2000元税前）。

竞赛具有较高的国际影响力，在国内高校中更是作为保研、综合测评、创新奖学金等评定竞赛之一。

联系我们

电话：182-1092-2591（郭老师，同微信）

"MathorCup数学应用挑战赛"公众号

微信二维码

关注可获取更多竞赛资讯