|  |
| --- |
|  |





硕士学位论文

**THESIS OF MASTER DEGREE**

|  |  |
| --- | --- |
| 论文题目： | 基于用户分析的智慧出行 |
| （英文）： | **Smart Travel Based On User Analysis** |
|  |  |
|  |  |
|  |  |
| 作者： | 刘建富 |
| 指导教师： | 余力 副教授 |
|  |  |
|  |  |
|  | 2020年 01月14日 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 中国人民大学  硕士学位论文 | | | | |
|  | |  | | |
| （中文题目） | | 基于用户分析的智慧出行 | | |
| （英文题目） | | Smart Travel Based On User Analysis | | |
|  | |  | | |
|  | 资格卡号： | | 61310141 |  |
|  | 作者姓名： | | 刘建富 |  |
|  | 所在学院： | | 信息学院 |  |
|  | 专业名称： | | 计算机应用技术 |  |
|  | 导师姓名： | | 余力 |  |
|  | 论文主题词：  （3-5个） | | 共享电单车；RBF神经网络；最小元素法模型；调度再平衡 |  |
|  | 论文提交日期： | | 2020年01月14日 |  |

**独 创 性 声 明**

本人郑重声明：所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得中国人民大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者： 刘建富 日期： 2020年01月14日

**关于论文使用授权的说明**

本人完全了解中国人民大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

论文作者： 刘建富 日期：2020年01月14日

指导教师： 刘建富 日期：2020年01月14日

摘要

随着互联网、大数据时代的快速发展，共享经济这一模式十分流行，已经渗透到人们日常生活中的方方面面。其中共享交通出行领域无疑是共享经济在全球范围内影响最广的一个领域。作为2017年共享出行领域的热门话题，共享电单车也得到了快速的发展。“低碳环保、绿色出行、灵活便捷”的出行理念，使共享电单车成功解决“出行最后一公里”的难题，也成为打通城市毛细血管的有效交通工具。共享电单车的出现，极大的化解了城市中短途出行的难题。然而，一方面共享电单车在完善城市交通，方便居民出行的同时，另一方面共享电单车也会产生车辆分布不均衡、运营成本过高、乱停乱放、车辆投放大而服务水平低等不良现象。

本文针对以上现有问题，经过分析，认为高效合理的车辆调度能够有效改善车辆分布不均衡、降低运营成本，提高人民出行效率，节约人力和时间成本。因此本论文重点将围绕如何解决区域内共享电单车分布不均衡的问题，使用RBF神经网络对北京市People Go最近6个月的用户骑行历史数据进行训练和预测，并利用最小元素法模型寻找实现电单车在站点间调度再平衡的代价方案，通过使用闭回路法进行路线的优化调度，找出最优调度策略，验证优化调度模型的可行性。

本文的研究内容主要包括：

1. 夜间站点车辆再平衡策略；？？？？使得发送发士大夫撒放水电费水电费水电费？？？？
2. http://cdmd.cnki.com.cn/Article/CDMD-10004-1019204795.htm
3. 寻找白天高峰时段的调度策略；？？？？？？Zfsdfsdsfsdfdsfsdsdfdsfds

（3）预测站点在高峰时段的共享单车存量和净流入量；？？？？？？？？Fds

**关键词**：共享电单车；RBF神经网络；最小元素算法模型；调度再平衡

ABSTRACT

With the development of data technology, companies are increasingly paying attention to the value of data. Data technology helps companies to optimize their business operations, marketing analysis and product improvement. It also helps making better business decisions. By mining data on some business indicators, not only business value will be increased, but also potential problems will be identified. From the perspective of improving business, data technology can help identify opportunities in their own industries and businesses. Data mining can find out what users are paying more attention to and whether there is a strong demand for subscription. Therefore marketing methods can be adopted to increase the purchase rate of users. In this way, invalid marketing advertising and related costs are reduced. Accurate forecasting of service subscription is an important goal pursued by the operational department.

Based on the behavior data of a website user's access and subscription, this paper uses the historical data in 2 months to predict whether the user will order the same product in the next month. From the perspective of business, this paper proposes some novel features for forecasting. These features are extracted from the analysis of the main factors that affects user service ordering. Then they applied to Logistic Regression, Decision Tree and XGBoost models. These classification prediction models are used for comparison and verification. The results support the effectiveness of these novel features. In addition, XGBoost has achieved good predictive results in several sets of experiments. Furthermore, automatic feature extraction method is compared with these novel features proposed in this paper. The results show that the novel features proposed in this paper are more effective.

The research in this paper includes:

(1) Novel features design: According to the characteristics of the business, the feature extraction method is used to manually extract the multi-dimensional user features. Compared with the automatic feature generation method, it is found that the manual feature engineering structure is more meaningful and complete and has better effect. The automatic feature processing method is relatively simple. The features constructed in this paper are interpretable and can provide some references for operation and promotion work.

(2) Classification prediction model comparison experiment:Logistic Regression, Decision Tree and XGBoost algorithm are used to train the novel features generated by manual and automatic methods. Experiments show that XGBoost has better subscription prediction results.

(3) Resolving imbalance of sample data: SMOTE oversampling algorithm is used to achieve good performance on the business sample unbalanced data set.

**Keywords:** Subscription forecasting; Data mining model; XGBoost algorithm

目 录

[第1章 绪论 1](#_Toc28690167)

[1.1 研究背景 1](#_Toc28690168)

[1.2 研究目的和意义 2](#_Toc28690169)

[1.2.1 研究目的 2](#_Toc28690170)

[1.2.2 研究意义 3](#_Toc28690171)

[1.3 研究方法和结构安排 4](#_Toc28690172)

[1.3.1 研究方法 4](#_Toc28690173)

[1.3.2 组织结构 4](#_Toc28690174)

[1.4 本文的主要贡献 5](#_Toc28690175)

[第2章 相关理论研究与主要算法 6](#_Toc28690176)

[2.1 网站购买行为的影响因素 6](#_Toc28690177)

[2.2 网站购买行为预测研究现状 7](#_Toc28690178)

[2.3 机器学习的预测算法和评估指标 11](#_Toc28690179)

[2.3.1 逻辑回归算法 11](#_Toc28690180)

[2.3.2 决策树算法 12](#_Toc28690181)

[2.3.3 XGBoost算法 13](#_Toc28690182)

[2.3.4 三种算法比较 14](#_Toc28690183)

[2.3.5 模型评估指标 15](#_Toc28690184)

[第3章 研究框架与特征工程 20](#_Toc28690185)

[3.1 基于机器学习的订购业务分析和预测研究框架 20](#_Toc28690186)

[3.2 数据来源 21](#_Toc28690187)

[3.3 人工特征工程 23](#_Toc28690188)

[3.3.1 数据分析及可视化探索 23](#_Toc28690189)

[3.3.2 特征提取 28](#_Toc28690190)

[3.3.3 特征数据处理 38](#_Toc28690191)

[3.4 自动化特征工程 41](#_Toc28690192)

[第4章 模型应用及实验分析 43](#_Toc28690193)

[4.1 实验条件 43](#_Toc28690194)

[4.2 准备数据集 44](#_Toc28690195)

[4.2.1 设计特征对照组 44](#_Toc28690196)

[4.2.2 设计数据集划分策略 45](#_Toc28690197)

[4.2.3 样本不均衡数据处理 46](#_Toc28690198)

[4.3 订购业务预测分析 47](#_Toc28690199)

[4.3.1 模型参数调整 47](#_Toc28690200)

[4.3.2 模型效果评估检验 50](#_Toc28690201)

[4.3.3 特征重要性分析 58](#_Toc28690202)

[4.3.4 实验结果与对比 63](#_Toc28690203)

[第5章 结论与展望 66](#_Toc28690204)

[5.1 全文总结 66](#_Toc28690205)

[5.2 研究局限与展望 67](#_Toc28690206)

[参考文献 69](#_Toc28690207)

图表目录

[图2-1 逻辑回归模型的假设函数 11](#_Toc28690208)

[图2-2 基尼系数和熵之半的曲线 13](#_Toc28690209)

[图2-3 XGBoost分类问题举例 14](#_Toc28690210)

[图2-4标准二分类的混淆矩阵 16](#_Toc28690211)

[图2-5 ROC曲线示例 17](#_Toc28690212)

[图3-1 基于机器学习的业务订购预测框架图 21](#_Toc28690213)

[图3-2用户分省分布情况 23](#_Toc28690214)

[图3-3 业务产品订购分省分布情况 24](#_Toc28690215)

[图3-4 业务产品订购情况箱线图 24](#_Toc28690216)

[图3-5 历史业务订购频次 25](#_Toc28690217)

[图3-6 历史业务订购量趋势图 25](#_Toc28690218)

[图3-7 用户访问情况箱线图 26](#_Toc28690219)

[图3-8 分月访问人数（去重）趋势图 26](#_Toc28690220)

[图3-9 分月访问人数（去重）、订购次数趋势图 27](#_Toc28690221)

[图3-10 有访问行为的人隔几天购买 27](#_Toc28690222)

[图4-1 人工特征选取对照组三种机器学习算法预测情况 50](#_Toc28690223)

[图4-2 人工特征选取对照组三种机器学习算法二分类预测情况 52](#_Toc28690224)

[图4-3 人工特征选取对照组三种机器学习算法的ROC曲线 53](#_Toc28690225)

[图4-4 人工特征选取对照组三种机器学习算法的AUC值 53](#_Toc28690226)

[图4-5 人工特征选取对照组三种机器学习算法的Logloss值 54](#_Toc28690227)

[图4-6 人工特征选取对照组三种机器学习算法的运行花费时间 55](#_Toc28690228)

[图4-7 人工特征与自动特征的ROC曲线对比 57](#_Toc28690229)

[图4-8 人工特征与自动特征的评估指标对比 57](#_Toc28690230)

[图4-9 基于XGBoost的人工特征重要性排序 58](#_Toc28690231)

[图4-10 基于XGBoost的自动特征重要性排序 62](#_Toc28690232)

[表2-1 用户购买指标的维度拆分 7](#_Toc28690254)

[表2-2 三种分类算法特点对比 14](#_Toc28690255)

[表2-3 模型的评估指标 16](#_Toc28690256)

[表3-1 用户属性表 22](#_Toc28690267)

[表3-2 产品信息表 22](#_Toc28690268)

[表3-3 业务订购表 22](#_Toc28690269)

[表3-4 用户行为表 22](#_Toc28690270)

[表3-5 统计类特征-基本特征 29](#_Toc28690271)

[表3-6 统计类特征-订购统计特征 30](#_Toc28690272)

[表3-7 统计类特征-订购统计特征 31](#_Toc28690273)

[表3-8 比率特征-访问行为比率特征 32](#_Toc28690274)

[表3-9 时间特征-订购行为时间分布 32](#_Toc28690275)

[表3-10 时间特征-访问行为时间分布 33](#_Toc28690276)

[表3-11 时间特征-订购时间差、访问时间差 34](#_Toc28690277)

[表3-12 访问和订购组合特征 36](#_Toc28690278)

[表3-13 季节特征 37](#_Toc28690279)

[表3-14 自动化生成的新特征举例示意 41](#_Toc28690280)

[表4-1 实验条件列表 43](#_Toc28690281)

[表4-2 业务订购预测模型人工特征和自动特征对照组 44](#_Toc28690282)

[表4-3 样本数据集划分策略 45](#_Toc28690283)

[表4-4 逻辑回归模型需要的主要参数设置 47](#_Toc28690284)

[表4-5 决策树模型需要的主要参数设置 48](#_Toc28690285)

[表4-6 XGBoost模型需要的主要参数设置 49](#_Toc28690286)

[表4-7 基于XGBoost人工特征重要性得分和特征名称前5名示意 59](#_Toc28690287)

[表4-8 基于XGBoost人工特征排名和平均贡献度占比 61](#_Toc28690288)

[表4-9 基于XGBoost人工特征和自动特征重要性得分和特征名称前10名对比示意 62](#_Toc28690289)

[表4-10 基于XGBoost自动特征排名和平均贡献度占比 63](#_Toc28690290)

# 绪论

## 研究背景与意义

### 研究背景

近年来随着我国经济发展迅速，全国各地人们生活水平的提高，共享经济已经渗透到房屋出租、美食、物品交易以及交通出行等日常生活中的方方面面，共享经济这一模式变得十分流行。共享经济给人们的工作和生活带来了前所未有的便利，极大地提升了消费者的体验和资源的使用效率。共享电单车就是一个典型。

共享电单车是指运营企业在居民区、商业区和公共服务区投放电动单车供人使用的共享服务，是一种分时租赁模式，也是一种新时代下的新型环保共享经济，最大化的利用了公共道路资源。

共享电单车以“共享经济”为出发点和核心，近年来悄无声息地走进我们的生活，校园里、大街上、旅游景点、商业区等重要地段出现了一排排颜色和样式相同、摆放整齐规范的自行车。人们可以通过下载手机APP软件，进行注册、交押金、扫码开锁等一套便捷的流程，骑上单车，开始自己的行程。在到达目的地之后，进行锁车，停放在指定的位置，等待下一个人的使用。

共享电单车具有巨大的经济效益和社会效益，包括缓解决城市拥堵，提高交通效率，缓解空气污染，促进社会和自然和协调持续发展。为人们的日常生活和出行提供了便利和舒适，使城市的人文景观、文化气息和社会氛围更加和谐，营造了一个慢节奏、享受生活的环境。共享单车作为交通基础设施的一部分，广大市民通过单车骑行来深入城市的风光与文化，可以说，共享电单车已经成为城市的一道靓丽风景线，甚至有外国友人称其称为中国新时代的“四大发明”之一。

但是，任何事物都具有两面性。我们在看到共享电单车这种新型租赁式单车为经济发展和社会进步所带来的优势和好处之外，共享电单车在运营过程中存在很多有待解决的问题，比如，随着共享单车在中国的进一步普及和拓展，

车辆分布不合理的状况也逐步显现，过大的投放量使得大量共享电单车处于闲置状态，堆积在道路两旁，对其它交通产生一定干扰，同时其他不足车辆站点，用户无法使用出行。这与北京市区某些站点的People Go所面临的车辆时空分布不均的状况相似。

### 研究意义

所以本文通过研究北京People Go的使用现况，研究解决共享电单车时空分布不均衡问题的对策，同时也可以为国内共享电单车行业调度管理的改进提供参考，方便用户高效绿色出行，缓解交通压力等具有重要意义。

## 研究目的和意义

### 研究目的

随着大数据时代的到来，数据化运营已成为流行模式，企业需要以用户为中心，根据人群和场景提供个性化服务，不断提升管理、营销手段实现精细化运营。不同的用户喜好不同，可以面向不同喜好的用户定制不同的运营策略，快速的满足客户对产品不断变化的需求。因为产品反映的是人的需求，是用户价值的延伸，所以商业运营要随时关注“人、货、场”发生的变化[[[1]](#endnote-0)]，在网络环境下找到更有意义的影响消费者购买决策的因素，是制定营销策略的关键[[[2]](#endnote-1)]。需要借助人工智能技术，对顾客需求进行更深层次的挖掘，来有效促进企业和顾客互动合作，从而开发出更满足用户需求的产品，增加消费欲望，也能提升顾客体验和满意度。

利用各种技术手段实现精细化运营是必然的趋势。根据麻省理工学院对数字业务的研究，结果表明那些更倾向于以数据为依据来驱动决策的公司，比其他的公司收益会高6%，生产率要高4%[[[3]](#endnote-2)]。今年年初，BCG（波士顿咨询公司）与谷歌公司携手开展了一项研究，结果显示：通过大规模使用人工智能和高级分析，可以帮助消费品公司实现超过10%的营收增长。实现利润增长与企业对数据挖掘的重视是密不可分的，特别是对转化率的重视，如果把每个用户都转化为可以带来销售的目标客户，不但可以达成用户消费，也同时提升了用户满意度。如果企业使用自有数据采用机器学习的方法，对独立的用户进行跟踪的方式，我们就能定位到数据背后的每个人，也能记录用户在网站上的行为，丰富用户的业务标签和策略标签。只有对用户和产品进行更深入的追踪和分析，实现精准的预测，才能减少无效或过度的营销，从而促进业务和用户的长期增长[[[4]](#endnote-3)]。

对传统网站业务来说，在大数据的利用方面还有较大的优化空间，通过挖掘数据背后的价值是企业提升自身竞争力的有效方式。充分利用大数据可以实现对用户的精准营销，帮助运营了解用户的行为习惯和关注的内容，帮助企业进一步探索更精准的用户画像，建立针对每个用户的个性化的服务体系。通过技术手段不断给业务进行助力，最终会保持用户持续的增长。

正是在大数据发展的背景下，本文打算通过数据挖掘技术，来弥补企业对自身业务和用户行为挖掘的不足之处。首先在用户和业务层面做足够的分析，然后根据历史数据提取有效的业务特征，选用高效率的机器学习方法，来进行业务订购预测的分析和研究，从而实现用户和企业的双赢。

### 研究意义

本文针对某网站业务订购预测任务，主要研究对比了两种特征提取方法，并对采用的分类模型进行了预测效果评估，具体有以下几点意义：

本文的理论意义在于，通过学习电商等领域已有预测类问题的成功经验，并结合本身业务数据的特点，给出了业务订购预测模型的一般框架，也为后续业务数据标签库的研究打下了一定基础。其次，给出了业务订购预测模型的人工和自动化两种特征选择方法，通过实验对比人工构造的特征更有意义更完备，自动化特征虽然效率比较高，但相对来说特征比较简单，这些都为继续拓展相关业务预测提供了一定的参考。

本文的现实意义是，通过研究挖掘验证了哪些重要特征对用户订购起到了决定性作用，它们都是基于自身商业场景而发现的相关关系。这些具有可解释性的特征能够为后续运营工作提供重要参考依据，从而避免一些主观的、有认知偏见或基于本能导致的错误决策；让所有的运营决策更合理更有方向，也为运营真正实现数据驱动迈出了第一步；有效的特征也可以帮助产品进行迭代，提升运营效率，实现数据智能化为业务赋能。

## 研究方法和结构安排

### 研究方法

本文采用文献法、调研法和实验法来解决论文中提出的关键预测问题，通过文献法发现商品购买预测研究的现状和存在问题，通过调研法与有经验的人士咨询各种方法的优缺点，通过实验法来验证预测模型的有效性。

1、文献法，借鉴以往文献或学者的研究，如商品重复订购的研究，更精准的预测用户在哪天购买的研究等；查看一些比较重要的比赛或会议资料，如Kaggle（数据科学竞赛）、KDD-CUP（国际知识发现和数据挖掘竞赛）、阿里云天池比赛、GitHub等平台有很多优秀的文章供广大研究者进行学习。在总结归纳的基础上，分析得出了适合本文业务的研究方法。

2、调研法，通过咨询专业内人士，学习讨论如何进行业务预测的探索，从而知道一些方法的优缺点，利用参加讲座等机会了解相关行业的商品预测案例。

3、实验法，进行统计性分析和实验设计，提出针对预测模型特征的可行性假设，验证提出方法的有效性，并在实践中不断修正。

### 组织结构

第一章：引言部分，对本文研究背景、意义、研究内容等进行综合性阐述。

第二章：介绍网站购买行为影响因素，购买预测问题在相关行业中国内外的研究现状、采用的解决方法等。

第三章：介绍本文的预测模型研究框架，具体业务场景下的数据预处理方法，以及人工特征和自动化特征两种方法的实现过程。

第四章：介绍本文分类模型的应用和参数说明，以及对设计的特征对照组进行指标对比评估，得出实验结果。

第五章：对全文进行总结，阐明本次实验不足之处和对今后工作的展望。

## 本文的主要贡献

目前基于传统网上用户购物消费的预测，大多只根据产品相关度进行推送，或根据流行度预测用户感兴趣的商品来给用户进行推荐，无法预测具体哪些用户会进行消费购买。本文利用运营数据中产生的业务订购历史数据，来推测近期一段时间内用户对业务产品的购买意愿，即推测用户是否购买，可以更精准的为运营产品迭代和促销提供参考数据。

本文贡献之处在于：

（1）基于运营用户的历史业务数据，使用了两种方法对比进行特征构造。人工方法假设了5个对照组，将不同维度指标数据结合进行了特征衍生，通过实验验证了生成的特征是对模型有重要贡献意义、有效性比较高的新特征；另外，对比使用了自动化特征方法，验证了它大大缩短了在特征工程上花费的时间，也具有不错的预测效果，但自动特征相对简单，对未知数据来说容易产生过拟合的风险。结果表明，人工特征具有更强的解释性和可调节性，对实际工作更有意义。

（2）本文的业务数据属于样本不均衡数据，通过对比更换采样周期和过采样SMOTE算法，发现过采样算法有效解决了不平衡问题，提高了预测效果。

（3）利用本文设计的业务订购预测研究框架，后续运营时可以针对不同特征的用户制定不同的策略，应用到运营推广和投放中去，起到一定的积极指导作用，助力实现精细化运营的目标。

# 相关理论研究与主要算法

## 网站购买行为的影响因素

影响消费者购买的因素有很多，有为了满足物质或精神需要的动机，有对特定情境或外界刺激的反映需要，有对是否喜欢某事物的主观态度评价，也有消费者基于以往购买和使用商品的经验积累。消费者购买时究竟如何思考？一些研究发现决定消费者购买的重要因素集中在用户信任度[[[5]](#endnote-4)]、满意度、价值、价格、服务质量、转换成本这几方面上[[[6]](#endnote-5)]。而消费者在线购买时会更关注产品的使用性，对使用性放心之后，消费者才会关注商品的外围属性，是否能达到其满意度[[[7]](#endnote-6)]。在线的商品信息呈现出的可视性、交互性功能也会影响消费者最终的购买倾向，并且良好的购物体验对消费者购买意愿起到一定的促进作用[[[8]](#endnote-7)]。在线购买体验是指用户通过网站注册、登录、浏览、比对产品等环节时的心里状态，包括一系列的心里和情绪感受[[[9]](#endnote-8)]，比如还没决定购买时反复看产品介绍是否符合预期，比对商品价格，决定购买后才会快速的下单。

网上购买的一般流程是诱发需求、消费者搜集对比信息、做出购买决定。对于网站运营工作来说，需要从消费者购买动机、购买行为轨迹等角度来分析发现问题，一方面它可以指导运营推广工作帮助企业达成销售目标，另一方面也能从用户行为数据角度发现产品是否符合用户需求，网站的UI(User Interface)或功能设计是否有缺陷等等，从而促进运营工作的顺利开展。

当用户访问企业网站就会产生一系列的行为，系统日志会将用户的各种行为都记录下来，变成了一个个接触的记录，如果将所有的记录串在一起，就产生了一条条的用户行为流，再把用户行为流放在一起，就帮我们描绘了一个个所能触达的整个用户群体。我们如何找到这些相似用户群体呢？

我们可以把影响每个用户购买的指标进行各种维度的拆分，分为用户维度、商品维度、渠道维度[[[10]](#endnote-9)]，详见表2-1。

表2-1 用户购买指标的维度拆分

|  |  |
| --- | --- |
| **维度** | **指标** |
| 1、用户维度 |  |
| 自然属性 | 年龄、性别、所在省份、教育水平、出生日期、职业等 |
| 兴趣属性 | 内容偏好、产品偏好、品牌偏好等 |
| 社会属性 | 工作情况、家庭状况、婚姻情况、社交情况等 |
| 消费属性 | 收入水平、购买能力、购买频率、购买渠道等 |
| 2、商品维度 |  |
| 商品定位 | 生活/非必须品、高端/低端、一线/非一线城市等 |
| 商品属性 | 功能、价格、质量、颜色等 |
| 3、渠道维度 |  |
| 信息渠道 | 信息获取来源是哪、百度、微信、微博、小程序等 |
| 购买渠道 | 在哪个渠道购买的、APP、官网、微信、京东、淘宝等 |

分析这些指标目的是让用户最终能产生购物的行为[[[11]](#endnote-10)]，对本文基于网站业务订购中影响用户订购行为的实验研究分析过程也有所帮助。本文后续会在特征工程部分，对用户的一些操作行为特征进行假设检验，实验结果也会从数据的角度验证这些影响因素的有效性。

## 网站购买行为预测研究现状

从网上购物的发展路径来看，以一些大的电商网站为代表的探索性尝试都具有非常重要的意义。早期亚马逊上架新商品时，发现购买者非常在意商品评价，就通过一些免费的试用活动，来鼓励用户完善评价。根据马太效应，商品评价越多越好，越有可能被购买。因为用户摸不到看不到，所以商品评论直接影响用户是否能够购买，也直接关系到企业的利益。并且亚马逊的推荐系统，能够精确的把握住用户的需求和购买倾向，向用户推荐需要的商品，对当今的商品推荐系统具有重大的贡献。在国内，阿里巴巴的双十一，在短短几分钟之内就突破交易的亿元大关，都是因为企业非常重视数据，及时对消费者进行商品推送、进行消费行为的分析，以及对未来可能购买商品进行的预估，是否有周期性重复订购的需求，都是交易提升的重要手段[[[12]](#endnote-11)]。对比传统的消费者购买预测只停留在定性的阶段，现在我们可以运用大数据分析技术来追踪消费者的购物行为，并且针对消费者个性化的需求提供相应商品，提升消费者的忠诚度。比如，京东等企业运用深度学习方法来深入分析研究用户，探索未来消费趋势，并从深层次挖掘消费者的购买需求，从而实现客户细分[[[13]](#endnote-12)]和精准营销，这些都对推动电商的不断进步，提高资源的有效利用具有重要的意义。网购行为算是一个跨学科的研究方向，目前研究方法有，将用户的显式与隐式[[[14]](#endnote-13)]参与行为结合进行分析。用户的显式行为是指通过购买、评价等公开的表达自己的意愿的方式，隐式行为则是指通过浏览、分享、收藏等手段表达自己态度的方式。现阶段，如协同过滤[[[15]](#endnote-14)]等用户的各种显式行为已经得到了广泛应用。而如猜你喜欢等隐式推荐虽然各个网站上也随处可见，但有的并未使用相似度[[[16]](#endnote-15)]而是基于近期浏览记录进行推荐，效果仍需要不断进行优化。

虽然现在很多预测分析技术已经在各行各业进行了运用，但是针对不同业务场景来有效进行用户行为预测还在不断的研究中。比如，面对金融风控、广告预测、推荐系统、客户流失分析等问题时，客户和企业期望特征有较好的可解释性，更多的使用机器学习的方法来进行研究。一些常见的预测问题包括广告的点击率预测[[[17]](#endnote-16)]或客户流失预测。预测用户是否点击了广告的分类问题中，FM（Factorization Machines）[[[18]](#endnote-17)]和FFM（Field-aware Factorization Machine）模型都是最近几年的新模型，适合特征比较稀疏的大数据量情况，屡次在CTR（Click-Through-Rate）即点击通过率预估比赛中取得优秀的成绩。为了有效的投放广告，2018年，Harakawa等人[[[19]](#endnote-18)]使用了FFM实时预测的模型，可以实时的获取用户输入的数据进行个性化推荐。2016年，Backiel等人[[[20]](#endnote-19)]为了识别客户中的忠实客户和流失客户，当用户取消订阅流失后，企业利用生存分析模型继续观察电信用户行为一段时间。在网站购买预测研究中，2010年，Park C H和Park Y H[[[21]](#endnote-20)]利用顾客在线购买的时间模式，并将该行为加入到重复购买的行为预测中，验证了他们相关性并进行了精准的预测。2015年，Silahtaroglu和Donertasli[[[22]](#endnote-21)]根据收集的用户网购历史数据并结合人口统计信息，利用决策树和神经网络来预测消费者会不会来购买购物车的商品。国内的研究有，2016年，马姝[[[23]](#endnote-22)]以某电商数据为基础，使用逻辑回归和GBDT分别建立预测模型，选择已购买过的用户群，预测用户对7天内有过交互的产品，在接下来的一天是否会进行购买。2017年，马倩[[[24]](#endnote-23)]使用逻辑回归、GBM和XGBoost等模型融合的方法，对双十一在某商家购买过的新用户，预测其未来6个月是否会发生重复购买行为。2017年，刘潇蔓[[[25]](#endnote-24)]利用特征选择和模型融合（逻辑回归和支持向量机）的方法对某电商平台的用户购物和访问数据进行了预测，用前30天的数据预测用户第31天是否会购买，发现了用户购买前一天的访问数据为关键特征。2018年，张子实[[[26]](#endnote-25)]利用逻辑回归和决策树双模型融合的方法，来预测哪些用户未来5天内，会对某个目标品类下商品进行消费进行了研究。2018年，李国凤[[[27]](#endnote-26)]用“唯品会用户购买行为预测”赛题为基础数据，利用基本属性、电商经营档期、购物行为等进行特征衍生，用Relief算法[[[28]](#endnote-27)]进行过滤式特征选择，最后采用Logistics、决策树和朴素贝叶斯算法，针对前30天有登录行为的90后高端消费女性人群，其第31天是否有购买进行了预测。2018年，李海丽[[[29]](#endnote-28)]基于Logistics和XGBoost算法对网站上的用户信息，用户订单等二十多个数据指标进行流失分析，构建了用户流失预测模型。2018年，李梦[[[30]](#endnote-29)]使用协同过滤推荐算法分析，发现用商品重复购买周期因素，可以对用户再次购买的时间范围进行预测等。2019年，艾金金[[[31]](#endnote-30)]结合RFM模型[[[32]](#endnote-31)]利用XGBoost算法建立了电商客户流失预警模型，附带流失客户名单，并得到影响用户流失的重要因素。

在图像识别、语音分类和自然语言处理等非结构化数据问题中，以及在百万级以上这种大数据量的场景，更倾向于选择深度学习来进行研究。而近几年在结构化数据方面，深度学习结合实体嵌入法在一些比赛和学术文献中也取得了比较好的效果。分类变量的实体嵌入是指，把函数逼近问题中的分类变量映射到欧几里得空间中。2015年，de Brébisson等人[[[33]](#endnote-32)]在Kaggle比赛获胜方案中使用了实体嵌入方法拟合神经网络，处理了每一辆车的分类元数据，预测出了出租车行驶距离。2016年，Guo和Berkahn[[[34]](#endnote-33)]使用了一个实体嵌入的简单神经网络作为分类变量的方法，其中分类变量的种类超过了1000个类别，作为第三个解决方案实现了德国Rossmann商店销售任务预测。

自动化技术也是近几年比较新的方法，如AutoML、Auto-Keras[[[35]](#endnote-34)]、TPOP、AUTO-SKLEARN、H2O等，替代了人工设计模型，使用了自动优化和随机搜索[[[36]](#endnote-35)]等方法，通过查找匹配数据集的最优模型，简化了模型选择和机器学习调优过程。自动化特征工程方面有深度特征合成FeatureTools[[[37]](#endnote-36)]，它可以从一组相关数据表中自动提取有用且有意义的特征，不但减少了特征工程花费的时间，还创建了可解释的特征，并能通过过滤与时间相关的数据来防止数据泄漏。自动化特征工程改进了传统人工特征工程的标准流程，比人工特征工程更高效，可重复性更高，这项技术并不完美，但能显著地提高效率，使人们在机器学习的其它方面可以花费更多的时间，从而能够更快地构建更好的预测模型。

综上所述，经过对国内外的研究梳理，预测算法有协同过滤推荐、FM、逻辑回归、决策树、朴素贝叶斯、XGBoost集成方法、模型融合、神经网络、自动化方法等等。没有单一的模型可以完美地解决所有问题，并且不同的数据模型根据操作任务的不同会表现不同[[[38]](#endnote-37)]。神经网络学习是通过时空位置进行的建模，除了要做一些数据预处理工作，他会自动完成在机器学习中需要实现的特征工程任务，特别适合于图像、语音等有多变量的高纬数据，但是模型结构会比较复杂，训练比较麻烦，虽然深度学习可以帮助我们省去特征工程这个繁琐的过程，但是却失去了对特征的认识，如不能对特征的重要性进行解释。对于自动化方法有实验表明，使用ResNet架构和Google的Auto-Keras在相同的数据集上进行训练，达到相似的准确度，ResNet架构不到1小时，而自动化方法要运行24小时[[[39]](#endnote-38)]，需要付出更高的时间成本。

从本文网站业务特点来，看我们更关注具体哪些特征对模型预测会更有利，机器学习的方法会更合适，因为它在特征工程方面有非常好的可解释性。我们可以利用业务现有的特征进行特征衍生，可以使用人工和自动两种方法对比进行特征工程，也可以纳入一些业务系统外的特征。从前述的购买预测类文献来看融合模型有比较高的准确度，研究也发现基于树模型的XGBoost集成算法可以很好处理变量较少的表格类数据，运行速度比较快，并且具有神经网络没有的特性，如特征的可解释性，输入数据不变性，易于进行调参等。正是因为XGBoost具有这些优点，比较适合本文的研究问题，所以选择了XGBoost作为预测模型。

## 机器学习的预测算法和评估指标

本文主要选择机器学习方法中的逻辑回归、决策树、XGBoost算法，分别进行算法原理的概述和对比分析，并对二分类评估指标进行了说明。

### 逻辑回归算法

逻辑回归（Logistic Regression）是一种分类算法，通过把线性函数的结果映射到连续性的Sigmoid函数中，将线性回归的问题转化为分类问题[[[40]](#endnote-39)]。即先找到预测函数，通过极大似然法推导构造损失函数，再用梯度下降法最小化损失函数，达到了对数据进行分类的目的。逻辑回归模型简单易理解常被作为性能基准。

预测函数表示为

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

上式也被称为逻辑回归模型的假设函数，图2-1为其函数图像[[[41]](#endnote-40)]。

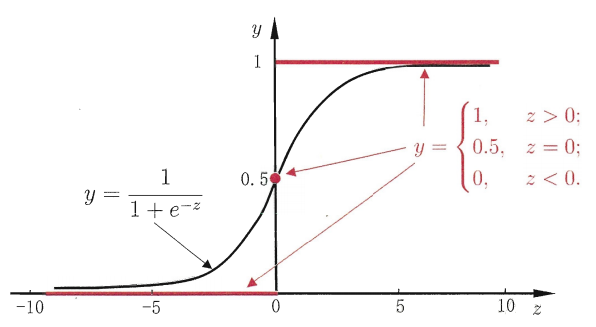


图2-1 逻辑回归模型的假设函数

训练数据预测值和实际数据的偏差，记为损失函数表示为

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

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

越小预测越准确，最小化损失函数表示为

|  |  |  |
| --- | --- | --- |
|  |  | (2-4) |

### 决策树算法

决策树（Decision Tree）可以分为两大类，分类树和回归树[[[42]](#endnote-41)]。分类树对离散变量做决策树，回归树对连续变量做决策树。决策树构造过程包括特征选择、树的生成（ID3、C4.5、CART等）、树的剪枝。决策树可以采用信息增益、信息增益比、基尼（GINI）系数来划分数据集的纯度和不确定性，不纯度越低越好。CART算法建立的是二叉树[[[43]](#endnote-42)]，使用GINI系数作为特征选择的度量，GINI系数越小，代表特征越好。因为使用了贪心算法，树容易产生过拟合，通过剪枝可以极小化决策树整体的损失函数，从而来减小模型复杂度。

对于样本D，如果有K个类别，将第k个类别的数量设为，则样本D的GINI系数表达式如下[[[44]](#endnote-43)]：

|  |  |  |
| --- | --- | --- |
|  |  | (2-5) |

对于样本D，根据特征A条件下某个值，把D分成、两个部分，那么样本D的GINI系数表达式为：

|  |  |  |
| --- | --- | --- |
|  |  | (2-6) |

基尼系数表示集合D的不确定性，基尼系数越大不确定性越大，这与熵相似，如图2-2[47]，横轴表概率，纵轴表损失，GINI系数和熵之半的曲线彼此很接近，因此可以把GINI系数做为熵模型的一个近似替代。

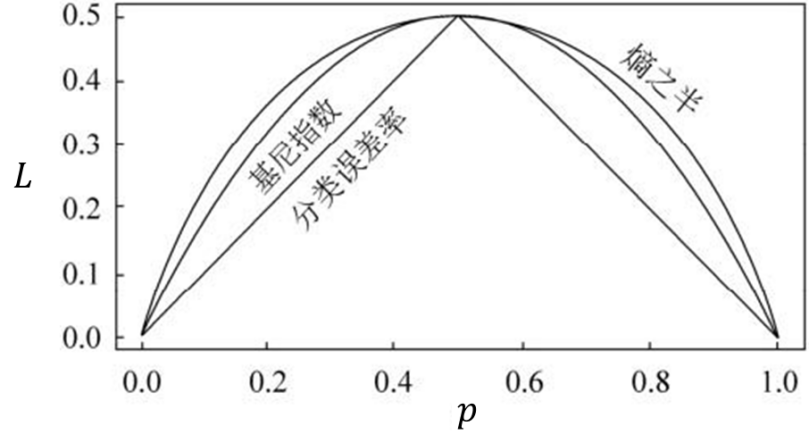


图2-2 基尼系数和熵之半的曲线

### XGBoost算法

XGBoost全名为极端梯度提升（Extreme Gradient Boosting）[[[45]](#endnote-44)]，它是boosting算法[[[46]](#endnote-45)]其中的一种，主要是利用了迭代思想，把很多弱分类器转化成为强分类器，用来实现分类的效果。XGBoost算法是梯度提升树算法（GBDT）的改进型算法，是一种集成学习算法，也是一种监督学习算法。它用到的树模型为CART回归树模型，在损失函数里加入正则项，防止过拟合并提高了泛化能力，它对代价函数进行二阶泰勒展开，并引入了一阶和二阶导数。

XGBoost算法的思想，就是基于单棵决策树选择一些特征进行切分得到结果，然后不断的添加树进行训练，每次都会学习一个拟合上次预测残差的新函数。根据不同的特征分裂产生k棵树，样本在每棵树上都会有一个对应的叶节点，并且会得到一个分数，再把样本在k棵树的分数总和相加，就是该样本的预测值[[[47]](#endnote-46)]。

举例如图2-3[48]，我们预测一个人喜欢玩电脑游戏的可能性，输入年龄、性别、是否每天使用电脑等特征，根据数据集训练出2棵决策树，把小男孩在所有叶节点的分数相加，就是小孩的预测分数，爷爷的预测分数也同理。

目标函数为：

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

经过变换目标函数写成为关于叶子结点分数的一个一元二次函数，我们把最优的代入目标函数，得到树结构的打分函数：

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

其中，为一阶导数；，为二阶导数。

结构分数越小，代表树的结构越好，也就是损失值越小越好。

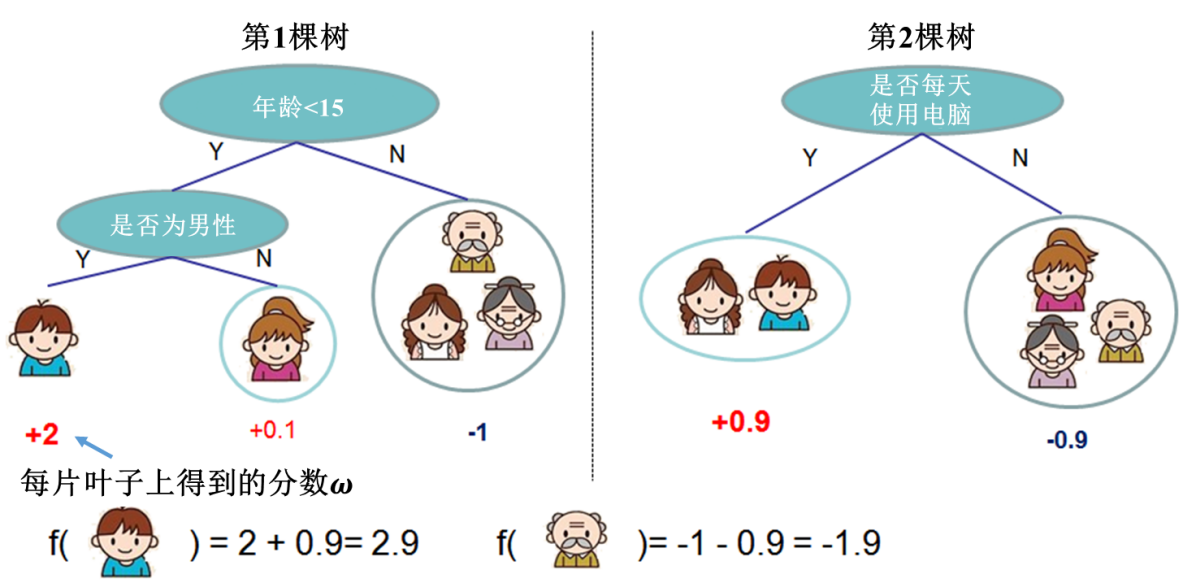


图2-3 XGBoost分类问题举例

### 三种算法比较

本文主要目的是根据历史数据来选择并建立一个易用性好、泛化能力强等各方面都比较理想的模型，接下来我们将目前工业和学术界中评价均比较好的XGBoost算法和经典的逻辑回归和决策树算法进行一下比较，详见表2-2。

表2-2 三种分类算法特点对比

| **对比项** | **逻辑回归** | **（CART）决策树** | **XGBoost** |
| --- | --- | --- | --- |
| 算法类型 | 单一算法 | 单一算法 | 集成算法 |
| 预测准确率 | 比较好 | 比较好 | 比基分类器高 |
| 收敛时间 | 比较快 | 比较快 | 比基分类器慢 |
| 特征的可解释性 | 有可解释性 | 有可解释性 | 有可解释性 |
| 容易过拟合程度 | 容易 | 容易 | 存在 |
| 缺失数据敏感度 | 敏感 | 不敏感 | 不敏感 |
| 处理数据类型 | 适合特征和  标签线性关系  比较强的数据 | 适合特征之间  关联性不强的数据 | 可以适应  复杂多样的数据 |
| 训练数据量 | 无要求 | 小数据集 | 大数据集 |

XGBoost对比其它两个算法，优缺点如下：

优点：

1、XGBoost是集成算法，逻辑回归和决策树是单一算法，在实际业务中如果数据量比较大，数据结构比较复杂，使用单一的分类方法准确率不够高。

2、XGBoost集成学习是对基分类器进行的综合，它提升了分类效果和更好的解决了过拟合的问题，比单一分类方法要好。

3、如果处理有缺失的数据集，基于树模型的方法比逻辑回归方法要好，XGBoost是由CART树构成，处理缺失数据集效果比单一的CART树要更好。

缺点：

XGBoost集成算法预测速度比单一分类器时间要长，随着基分类器数目的增加，存储空间也不断增加。

综合以上对比结果，XGBoost算法拥有较高的预测准确率、具有特征的可解释性、能很好处理过拟合问题、对缺失数据不敏感、能处理各种类型的数据的特点。又因为XGBoost本身是基于树模型的，树模型的计算速度本身就比较快，集成算法运行时间也不会太差，存储空间问题可以通过硬件扩容等方式来解决。所以本文选用XGBoost研究业务预测问题预计会达到比较理想的效果。

### 模型评估指标

本文实验将用到的分类问题的评估标准，详见表2-3。

表2-3 模型的评估指标

|  |  |
| --- | --- |
| **评估指标** | **指标名称** |
| 检验结果的预测准确率 | 准确度（Accuracy） |
| 召回率（Recall） |
| 精确度（Precision） |
| F1-score |
| 分类的效果程度 | AUC面积 |
| 预测模型损失值 | 对数损失函数Logloss |
| 运行花费时间 | 模型运行时间 |

以上评估指标除了后两个，其它都是基于混淆矩阵而得出。混淆矩阵是用来可视化算法性能的一种特殊矩阵表示法，是针对二分类问题的多维度量的指标系统，如图2-4。在混淆矩阵中，我们将少数类作为正例，将多数类作为负例，当正负样本不均衡时时也很有用。通常在基于树模型的分类算法里，少数类是1表示正例，多数类是0表示负例。它主要包含6种指标：准确度、精确度、召回率、F1-score、特异度、假正率。

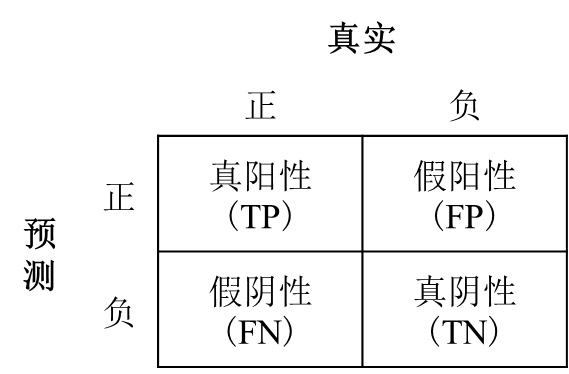


图2-4标准二分类的混淆矩阵

TP = True Postive = 真阳性； FP = False Positive = 假阳性

FN = False Negative = 假阴性； TN = True Negative = 真阴性

评估指标作用：

1、准确度（Accuracy）：用来确定模型的整体效果，表示所有准确预测的所有样本除以总样本，通常越接近1越好。

2、精确度（Precision）：又叫查准率，表示所有预测为少数类的样本中，实际少数类所占的比例。

3、召回率（Recall）：又叫敏感度(Sensitivity)，真正率，查全率，表示所有实际为1的样本中，预测正确的样本所占的比例。我们一般使用精确度来判断大多数类是否正确，使用召回率来评估少数类是否正确。

4、F1-score或F1-measure：是同时考虑准确性和召回率，生成调和平均数作为两者之间平衡的综合指标。F1-score分布在[0,1]之间，越接近1越好，比较高的F1-score，才可以实现更好的精确度和召回率。

5、特异度(Specificity)：表示所有真实为0的样本中，正确预测为0的样本所占的比例。

6、假正率（False Positive Rate）：特异度表示一个模型正确判断多数类的能力，而假正率（1-特异度）就是模型错误判断多数类的能力。

7、ROC曲线

ROC（Receiver Operating Characteristic Curve）叫做受试者工作曲线，是一种组合评价指标，它分别使用假正率（FPR）和真正率（TPR）作为曲线的横轴和纵轴，如图2-5[[[48]](#endnote-47)]。

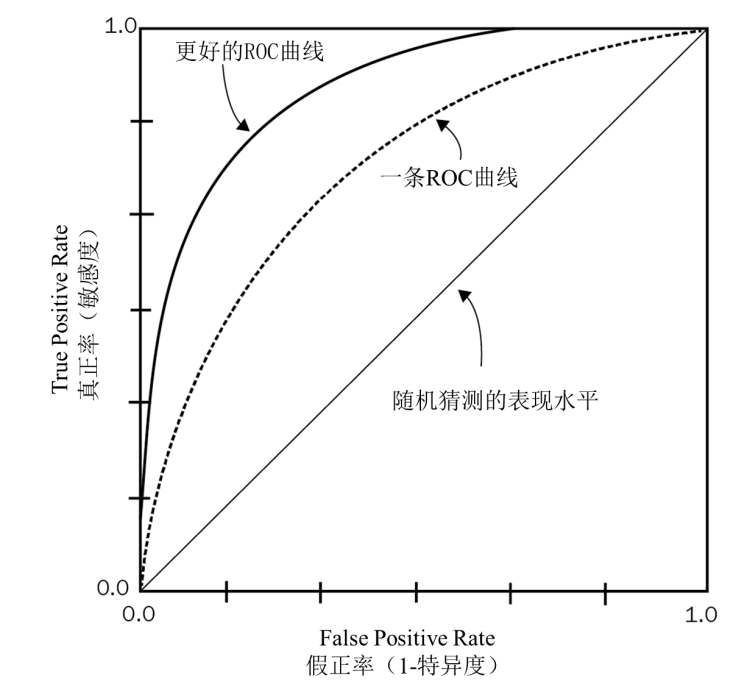


图2-5 ROC曲线示例

FPR横轴：FPR越大表示预测正类中实际负类越多。

TPR纵轴：TPR越大表示预测正类中实际正类越多。

ROC曲线的作用[[[49]](#endnote-48)]：

（1）ROC曲线的每个点对应于一个阈值，所以容易查出对学习器泛化性能最好的阈值。

（2）当ROC曲线[[[50]](#endnote-49)]无限靠近左上角模型查全率越高，并且分类错误最少，越偏离对角线45度越好，因为AUC为0.5的时候没有研究意义。

（3）可以把不同学习器的ROC曲线绘到一起，通过可视化的方法鉴别哪个学习器准确性更高，它是一个检验准确性的综合指标。

8、AUC面积

AUC（Area Under ROC Curve）就是ROC曲线下的面积[[[51]](#endnote-50)]，它是衡量二分类模型好坏的一种性能评价指标。AUC面积的取值范围一般在0.5~1之间，不会大于1，因为对应的ROC曲线要在y=x直线的上方才有意义，当值为0.5时模型没有实际意义。

一般通过看ROC曲线中更靠近左上角的那条曲线学习器性能更好，但是实际任务中，ROC曲线存在着交叉或重叠，不能清晰的分辨出哪个性能更优，此时比较合理的判断依据是使用AUC值作为评价标准，对应AUC值更大的分类器效果更好[[[52]](#endnote-51)]。

另外，AUC同时考虑了学习器对于正例和负例的分类能力，即使在样本不平衡的情况时候，仍然能够对分类器做出比较合理的评价。AUC对样本是否均衡并不敏感，所以AUC通常也被用来评价不均衡样本的学习器性能。

9、对数损失函数Logloss

对数损失函数，即似然损失函数（Log-likelihood Loss）[[[53]](#endnote-52)]，它可以通过惩罚错误的分类，来实现量化分类器准确度的作用，一般来说值越小越好[[[54]](#endnote-53)]。对于二分类其对数损失函数公式如下：

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

其中，表示数据的实际类别，表示预测数据属于类别1的概率。整体的对数损失是由所有样本对数损失的均值计算而来的。

10、模型运行时间：就是通过time库来计算模型执行消耗的时间，可以由此判断模型易用程度。

综上，逻辑回归通过Sigmoid的函数来进行二分类、决策树通过信息纯度不断分枝来分类、XGBoost实质上是用很多CART树集成不断迭代的思想来进行分类。另外，分类算法评估的核心，就是利用混淆矩阵计算得到相关的指标来判断分类的效果。本文将利用这些算法和评估指标来进行实验。

# 数据获取与特征分析

## 数据抓取过程

共享电单车的数据主要来源于用户注册APP的个人基本信息和用户骑行时产生的订单信息。数据来自已有的数据库，不需要交给爬虫工程师去采集来提供。数据获取的流程大概有以下5步：

1.用户下载APP，注册个人信息

2.用户登录单车APP (APP将登录信息存储到mysql库中)，查看附近是否有车辆，附近没有，则继续移动到其它位置进行找车骑行，否则下一步

3.扫车身上的二维码或者输入编号，就会远程开锁。（车身有两处二维码，一处是车龙头上面，一处是锁上面）

4.开完锁，APP开始计费，用户开始骑行，到达目的地，把车子停在政府划线的停放点。

5.锁车，扫码付款，APP将骑行订单信息存储到mysql库中。

如图3-1，共享电单车数据抓取流程图。



图3-1 共享电单车数据抓取流程图

## 数据集介绍

本文采用是2019年北京市People Go的历史骑行记录数据，使用了从2019年7月到2019年12月People Go的电单车使用记录，以及7-12月北京市的天气数据。为了更好的保护用户敏感隐私信息，将原始数据中用户的敏感信息进行了脱敏处理。

本研究主要用到了用户订单骑行表、天气表、用户表。具体表结构说明

详见表3-1到3-3。

表3-1 用户订单骑行属性表

| **字段名称** | **解释说明** |
| --- | --- |
| orders\_bike\_id | 订单车辆主键id |
| order\_no | 订单号 |
| bike\_no | 车牌号 |
| take\_latitude | 取车维度 |
| take\_longitude | 取车经度 |
| take\_station\_id | 取车站点id |
| take\_gps\_time | 取车gps定位时间 |
| take\_address\_bike | 取车地址 |
| return\_latitude | 还车维度 |
| return\_longitude | 还车经度 |
| return\_station\_id | 还车站点 |
| return\_gps\_time | 还车gps定位时间 |
| return\_address\_bike | 还车地址 |
| take\_mileage | 取车里程数 |
| return\_mileage | 还车里程数 |
| order\_mileage | 订单里程数 |
| order\_date | 订单日期 |

表3-2 城市天气表

| **字段名称** | **解释说明** |
| --- | --- |
| city\_weather\_id | 天气主键id |
| ad\_code | 地区编码 |
| create\_time | 日期 |
| day\_weather | 白天天气气象 |
| night\_weather | 晚上天气气象 |
| night\_temp | 晚上温度 |
| day\_temp | 白天温度 |
| day\_wind | 白天风向 |
| night\_wind | 晚上风向 |
| day\_power | 白天风力 |
| night\_power | 晚上风力 |
| max\_temp\_f | 当日最高气温 |
| min\_temperature\_f | 当日最低气温 |
| precipitation\_rain\_in | 降雨量 |
| precipitation\_snow\_in | 降雪量 |
| precipitation\_ice\_in | 结冰量 |
| ultraviolet\_rays | 紫外线强度 |

表3-3 用户表

| **字段名称** | **解释说明** |
| --- | --- |
| user\_id | 用户id |
| nick\_name | 用户昵称 |
| mobile | 手机号 |
| real\_name | 用户真实姓名 |
| email | 邮箱 |
| sex | 性别 |
| register\_time | 注册时间 |
| register\_city\_code | 注册城市 |
| last\_login\_time | 最后登录时间 |
| last\_login\_city\_code | 最后登录城市 |

## 数据预处理

在我们通过观察、调查收集到初步的样本数据集后，接下来要考虑的问题是：样本数据集的数量和质量是否满足模型构建的要求？数据质量是数据分析挖掘中数据准备过程的重要一环，是数据预处理的前提，也是数据分析挖掘结论有效性和准确性的基础，没有可信的数据，数据挖掘构建的模型将是空中楼阁。数据质量分析的主要任务是检查原始数据中是否存在脏数据，脏数据一般是指不符合要求，以及不能直接进行相应分析的数据。在常见的数据分析挖掘工作中，脏数据包括缺失值、异常值、不一致的值、重复数据以及含有特殊符号（如#、￥、\*）的数据。

在数据分析挖掘中，海量的原始数据中存在着大量不完整（有缺失值）、不一致、有异常的数据，严重影响到数据挖掘建模的执行效率，甚至可能导致分析结果的偏差，所以进行数据清洗就显得尤为重要，数据清洗完成后接着进行或者同时进行数据集成、转换、规约等一系列的处理，该过程就是数据预处理。数据预处理一方面是要提高数据的质量，另一方面是要让数据更好地适应特定的挖掘技术或工具。统计发现，在数据分析的过程中，数据预处理工作量占到了整个过程的60％。

数据预处理的主要内容包括数据清洗、数据集成、数据变换和数据规约等过程。处理过程如图3-2所示。

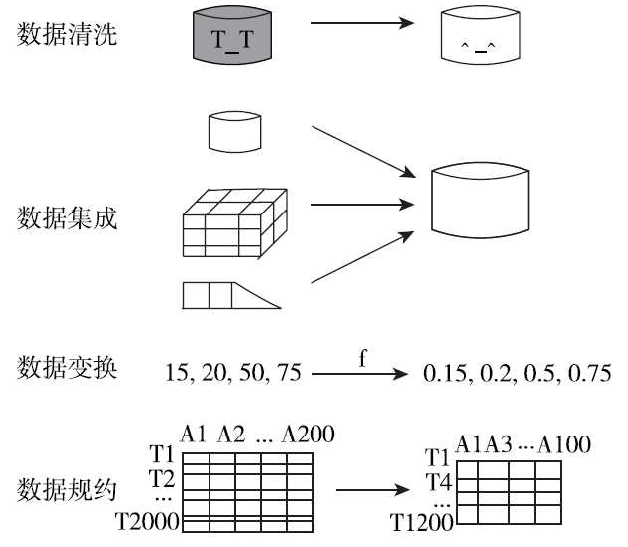


图3-2数据预处理过程图

### 数据清洗-缺失值处理

数据的缺失值主要包括记录的缺失和记录中某个字段信息的缺失，两者都会造成分析结果的不准确，以下从缺失值产生的原因及影响等方面展开分析。

（1）缺失值产生的原因

1）有些信息暂时无法获取，或者获取信息的代价太大。

2）有些信息是被遗漏的。可能是因为输入时认为不重要、忘记填写或对数据理解错误等一些人为因素而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障等非人为原因而丢失。

3）属性值不存在。在某些情况下，缺失值并不意味着数据有错误。对一些对象来说某些属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入等。

（2）缺失值的影响

1）数据分析挖掘建模将丢失大量的有用信息。

2）数据分析挖掘模型所表现出的不确定性更加显著，模型中蕴涵的规律更难把握。

3）包含空值的数据会使建模过程陷入混乱，导致不可靠的输出。

（3）缺失值的分析

使用简单的统计分析，可以得到含有缺失值的属性的个数，以及每个属性的未缺失数、缺失数与缺失率等。

从总体上来说，缺失值的处理分为删除存在缺失值的记录、对可能值进行插补和不处理3种情况。

如果通过简单的删除小部分记录达到既定的目标，那么删除含有缺失值的记录的方法是最有效的。然而，这种方法却有很大的局限性。它是以减少历史数据来换取数据的完备，会造成资源的大量浪费，将丢弃了大量隐藏在这些记录中的信息。尤其在数据集本来就包含很少记录的情况下，删除少量记录可能会严重影响到分析结果的客观性和正确性。一些模型可以将缺失值视作一种特殊的取值，允许直接在含有缺失值的数据上进行建模。

其中常用的数据插补方法，见表3-4。

表3-4 常用插补法表

| **插补方法** | **方法介绍** |
| --- | --- |
| 均值/中位数/众数插补 | 根据属性值的类型，用该属性取值的平均数/中位数/众数进行插补 |
| 使用固定值 | 将缺失的属性值用一个常量替换。比如北京市一个工厂外来人员的基本工资属性的空缺值可以用外来人员的工资标准3000/月 |
| 最近临插补 | 在记录中找到与缺失样本最接近的样本的该属性值插补 |
| 回归方法 | 对带有缺失值的变量，根据已有数据和与其有关的其它变量（因变量）的数据建立拟合模型来预测缺失的属性值 |
| 插值法 | 插值法是利用已知点建立合适的插值函数f（x），未知值由对应点求出的函数值近似代替 |

本文重点介绍拉格朗日插值法。其他的插值方法还有牛顿插值法、Hermite插值、分段插值、样条插值法等。

（1）拉格朗日插值法

根据数学知识可知，对于平面上已知的n个点（无两点在一条直线上）可以找到一个次多项式，使此多项式曲线过这n个点。

1）求已知的过n个点的n-1次多项式：



将n个点的坐标（，），（，）…（，）代入多项式函数，得

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

解出拉格朗日插值多项式为：

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

2）将缺失的函数值对应的点x代入插值多项式得到缺失值的近似值L（x）。

拉格朗日插值公式结构紧凑，在理论分析中很方便，但是当插值节点增减时，插值多项式就会随之变化。牛顿插值法也是多项式插值，但采用了另一种构造插值多项式的方法，与拉格朗日插值相比，具有承袭性和易于变动节点的特点。从本质上来说，两者给出的结果是一样的（相同次数、相同系数的多项式），只不过表示的形式不同。因此，在Python的Scipy库中，只提供了拉格朗日插值法的函数（因为实现上比较容易），如果需要牛顿插值法，则需要自行编写函数。

本文在收集天气温度数据发现了很多缺失值的情况，主要有空值，通常要根据实际业务情况来缺失值填补，比如用户所在省存在NAN，可以填补成“未知”。对于空值NAN当不影响特征含义的时候，可以填补成fillna(0)，即用0填补。如表3-5为7月份中一周的天气气温表，

表3-5 7月份一段时间天气温度表

| **日期** | 2019/07/15 | 2019/07/16 | 2019/07/17 | 2019/07/18 | 2019/07/19 |
| --- | --- | --- | --- | --- | --- |
| **温度（°C）** | 35 | 30 |  | 29 | 32 |
| **日期** | 2019/07/20 | 2019/07/21 | 2019/07/22 | 2019/07/23 | 2019/07/24 |
| **温度（°C）** |  |  | 36 | 34 | 37 |

其中17、20、21日数据缺失，用拉个朗日插值对缺失值进行插补，关键核心代码如下：

代码清单3-1 拉格朗日法进行插补

1. #拉格朗日插值代码，index插入的位置，df为列向量，k为取前后的个数
2. **def** deal\_empty\_column(index, df, k=5):
3. **print**(index)
4. #取插补值索引前，后各5个数
5. y = df[list(range(index - k, index))
6. + list(range(index + 1, index + 1 + k))]
7. y = y[y.notnull()]
8. #调用拉格朗日插值函数
9. **return** lagrange(y.index, list(y))(index)

应用拉格朗日插值法算对表3-5中的缺失值进行插补，使用缺失值前后各5个未缺失的数据参与建模，得插值结果如表3-6所示。

表3-6 7月份一段时间天气温度插补表

| **日期** | 2019/07/15 | 2019/07/16 | 2019/07/17 | 2019/07/18 | 2019/07/19 |
| --- | --- | --- | --- | --- | --- |
| **温度（°C）** | 35 | 30 | 31 | 29 | 32 |
| **日期** | 2019/07/20 | 2019/07/21 | 2019/07/22 | 2019/07/23 | 2019/07/24 |
| **温度（°C）** | 34 | 35 | 36 | 34 | 37 |

利用拉格朗日插值对这17、20、21日的数据进行插补，结果是31℃、34℃、35℃，7月份正值是夏季并且这几天为晴天，所以插值结果比较符合实际情况。

### 数据清洗-异常值处理

异常值分析是检验数据是否有录入错误以及含有不合常理的数据。忽视异常值的存在是十分危险的，不加剔除地把异常值包括进数据的计算分析过程中，对结果会产生不良影响；重视异常值的出现，分析其产生的原因，常常成为发现问题进而改进决策的契机。

异常值是指样本中的个别值，其数值明显偏离其余的观测值。异常值也称为离群点，异常值的分析也称为离群点分析。

（1）简单统计量分析

可以先对变量做一个描述性统计，进而查看哪些数据是不合理的。最常用的统计量是最大值和最小值，用来判断这个变量的取值是否超出了合理的范围。如客户年龄的最大值为199岁，则该变量的取值存在异常。

（2）3σ原则

如果数据服从正态分布，在3σ原则下，异常值被定义为一组测定值中与平均值的偏差超过3倍标准差的值。在正态分布的假设下，距离平均值3σ之外的值出现的概率为P（｜x-μ｜>3σ）≤0.003，属于极个别的小概率事件。

如果数据不服从正态分布，也可以用远离平均值的多少倍标准差来描述。

（3）箱型图分析

箱型图提供了识别异常值的一个标准：异常值通常被定义为小于QL－1.5IQR或大于QU＋1.5IQR的值。QL称为下四分位数，表示全部观察值中有四分之一的数据取值比它小；QU称为上四分位数，表示全部观察值中有四分之一的数据取值比它大；IQR称为四分位数间距，是上四分位数QU与下四分位数QL之差，其间包含了全部观察值的一半。

箱型图依据实际数据绘制，没有对数据作任何限制性要求（如服从某种特定的分布形式），它只是真实直观地表现数据分布的本来面貌；另一方面，箱型图判断异常值的标准以四分位数和四分位距为基础，四分位数具有一定的鲁棒性：多达25％的数据可以变得任意远而不会很大地扰动四分位数，所以异常值不能对这个标准施加影响。由此可见，箱型图识别异常值的结果比较客观，在识别异常值方面有一定的优越性，如图3-3所示。

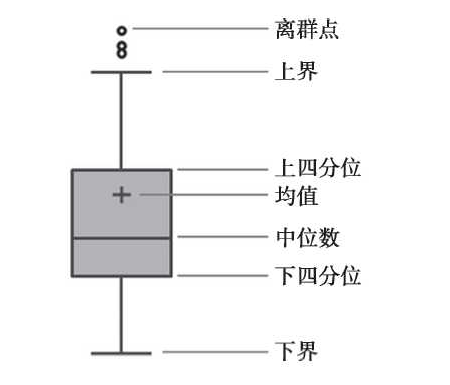


图3-3箱图检测异常值

在数据预处理时，异常值是否剔除，需视具体情况而定，因为有些异常值可能蕴含着有用的信息。异常值处理常用方法见表3-7。

表3-7 异常值处理常用办法

|  |  |
| --- | --- |
| **异常值处理方法** | **方法介绍** |
| 删除含有异常值的记录 | 直接将含有异常值的记录删除 |
| 视为缺失值 | 将异常值视为确实值，利用缺失值处理的方法进行处理 |
| 平均值修正 | 可用前后两个观测值的平均值修正该异常值 |
| 不处理 | 直接在具有异常值的数据集上进行挖掘建模 |

将含有异常值的记录直接删除的方法简单易行，但缺点也很明显，在观测值很少的情况下，这种删除会造成样本量不足，可能会改变变量的原有分布，从而造成分析结果的不准确。视为缺失值处理的好处是可以利用现有变量的信息，对异常值（缺失值）进行填补。

在很多情况下，要先分析异常值出现的可能原因，再判断异常值是否应该舍弃，如果是正确的数据，可以直接在具有异常值的数据集上进行挖掘。

在数据集的订单车辆表中，取车、还车的经纬度可能会出现缺失值和异常值，如表3-8中数据所示。

表3-8 异常值部分数据表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **订单号** | **车辆号** | **取车纬度** | **取车经度** | **还车维度** | **还车经度** |
| 794497300283320270 | 305022237 | 39.905317 | 116.475589 | 39.894669 | 116.498 |
| 794501403901911640 | 305020105 | 39.930351 |  | 39.937714 | 116.450554 |
| 794511215466090270 | 305004591 | 40.010702 | 116.467717 | 116.462273 | 40.002928 |
| 794546671997232230 | 305046175 | 39.931089 | 116.449807 |  |  |
| 318760400383541420 | 716453 |  |  |  |  |
| 318822471939330270 | 712219 | 9.47292 | 106.622878 | 9.47275 | 106.622801 |

从数据中可以发现，其中有部分数据是缺失以及经度的值不合理的情况，因为数据记录和属性较多，使用人工分辨的方法是不切实际的，本文这里先使用编写程序来检测出含有异常值以及缺失率。

在Python的Pandas库中，只需要读入数据，然后使用describe（）函数就可以查看数据的基本情况。其中count是非空值数，通过len（data）可以知道数据记录为48008970条，缺失值数为1735条。另外，提供的基本参数还有平均值（mean）、标准差（std）、最小值（min）、最大值（max）以及1/4、1/2、3/4分位数（25％、50％、75％）。为了更直观地展示这些数据，并且可以检测异常值，本文使用箱线图。其核心关键检测代码如代码清单3-2所示。

代码清单3-2 订单车辆数据异常值检测代码

1. #以下参数都是经过调试的，需要具体问题具体调试
2. **for** i **in** range(len(x)):
3. # 处理临界情况， i》0时
4. **if** i>0:
5. plt.annotate(y[i],xy=(x[i],y[i]),xytext=(x[i] +0.05-0.8 / (y[i] - y[i - 1]), y[i]))
6. **else**:
7. plt.annotate(y[i],xy=(x[i], y[i]), xytext=(x[i] + 0.08, y[i]))
8. plt.show()

经过上面代码的检测后得到，异常值的箱线图，如图3-4中数据所示。

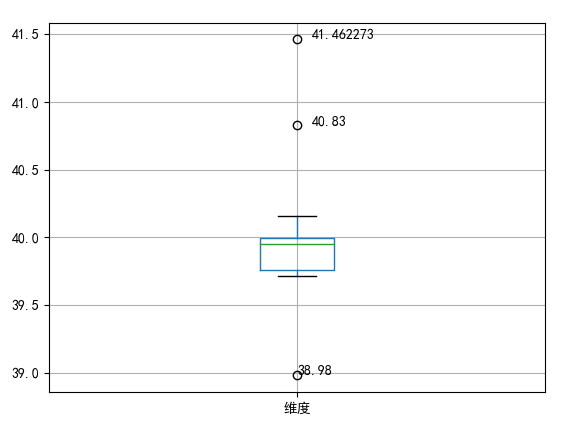


图3-4异常值检测箱型图

从图3-4中可以看出，箱型图中的超过上下界的3个维度数据可能为异常值。结合北京地图五环内以内的纬度（39.755-40.032）可以把41.462273、40.83、38.98归为异常值。最后确定过滤规则为：维度在39.713以下40.172以上则属于异常数据。经分析，这些异常数据可能是程序发布版本或者第三方网络速率不佳，小概率情况造成的异常数据，结合本数据集数据量比较大，异常数据相对比较少（48008970>1735），且异常数据的值不具有任何意义，所以这里针对异常数据采取的办法是：编写过滤程序，删除这里异常数据。

### 数据集成

数据挖掘需要的数据往往分布在不同的数据源中，数据集成就是将多个数据源合并存放在一个一致的数据存储（如数据仓库）中的过程。

1.实体识别

实体识别是指从不同数据源识别出现实世界的实体，它的任务是统一不同源数据的矛盾之处，常见形式如下。

（1）同名异义

数据源A中的属性ID和数据源B中的属性ID分别描述的是车辆编号和订单编号，即描述的是不同的实体。

（2）异名同义

数据源A中的order\_dt和数据源B中的order\_date都是描述订单日期的，即A.order\_dt=B.order\_date

（3）单位不统一

描述同一个实体分别用的是国际单位和中国传统的计量单位。检测和解决这些冲突就是实体识别的任务。

2.冗余属性识别

数据集成往往导致数据冗余，例如，

1）同一属性多次出现；

2）同一属性命名不一致导致重复。

仔细整合不同源数据能减少甚至避免数据冗余与不一致，从而提高数据挖掘的速度和质量。对于冗余属性要先分析，检测到后再将其删除。

本文数据集都是用户骑行的数据存储到mysql中，针对用户骑行订单车辆表数据量太大，按照订单号末位取模的策略进行了水平分表，水平分库分表能够有效的缓解单机和单库的性能瓶颈和压力，突破IO、连接数、硬件资源等的瓶颈。使用的时候需要进行union all将小表合并查询实用。如图3-5中数据所示。

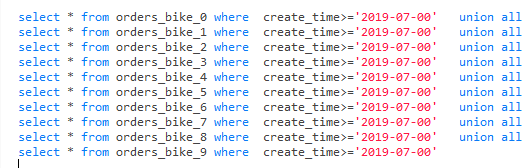


图3-5数据集使用图

## 数据特征分析

“数据决定了机器学习的上限，而算法只是尽可能逼近这个上限”，这句话很好的阐述了数据在机器学习中的重要性。大部分直接拿过来的数据都是特征不明显的、没有经过处理的或者说是存在很多无用的数据，那么需要进行一些特征处理，特征的缩放等等，满足训练数据的要求。

特征是数据中抽取出来的对结果预测有用的信息，可以是文本或者数据。

特征工程是使用专业背景知识和技巧处理数据，使得特征能在机器学习算法上发挥更好的作用的过程。过程包含了特征提取、特征构建、特征选择等模块。

特征工程的目的是筛选出更好的特征，获取更好的训练数据。因为好的特征具有更强的灵活性，可以用简单的模型做训练，更可以得到优秀的结果。“工欲善其事，必先利其器”，特征工程可以理解为利其器的过程。互联网公司里大部分复杂的模型都是极少数的数据科学家在做，大多数工程师们做的事情基本是在数据仓库里搬砖，不断地数据清洗，再一个是分析业务不断地找特征。 例如，某广告部门的数据挖掘工程师，两周内可以完成一次特征迭代，一个月左右可以完成模型的小优化，来提升auc。

最初的原始特征数据集可能太大，或者信息冗余，因此在机器学习的应用中，一个初始步骤就是选择特征的子集，或构建一套新的特征集，减少功能来促进算法的学习，提高泛化能力和可解释性。例如：你要查看不同地域女性的穿衣品牌情况，预测不同地域的穿衣品牌。如果其中含有一些男性的数据，是不是要将这些数据给去除掉。

### 特征归一化

数据规范化（归一化）处理是数据挖掘的一项基础工作。不同评价指标往往具有不同的量纲，数值间的差别可能很大，不进行处理可能会影响到数据分析的结果。为了消除指标之间的量纲和取值范围差异的影响，需要进行标准化处理，将数据按照比例进行缩放，使之落入一个特定的区域，便于进行综合分析。如将工资收入属性值映射到[-1，1]或者[0，1]内。

数据规范化对于基于距离的挖掘算法尤为重要。

（1）最小-最大规范化

最小-最大规范化也称为离差标准化，是对原始数据的线性变换，将数值值映射到[0，1]之间。

转换公式如下：

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

其中，max为样本数据的最大值，min为样本数据的最小值。max-min为极差。离差标准化保留了原来数据中存在的关系，是消除量纲和数据取值范围影响的最简单方法。这种处理方法的缺点是若数值集中且某个数值很大，则规范化后各值会接近于0，并且将会相差不大。若将来遇到超过目前属性[min，max]取值范围的时候，会引起系统出错，需要重新确定min和max。

（2）零-均值规范化

零-均值规范化也称标准差标准化，z-score 标准化(zero-meannormalization)也叫标准差标准化，经过处理的数据符合标准正态分布，即均值为0，标准差为1，其转化函数为:其中mean为所有样本数据的均值，σ为所有样本数据的标准差，标准差为1。转化公式为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-4) |

其中x为原始数据的均值，σ为原始数据的标准差，是当前用得最多的数据标准化方法。

（3）小数定标规范化

通过移动属性值的小数位数，将属性值映射到[-1，1]之间，移动的小数位数取决于属性值绝对值的最大值。

转化公式为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-4) |

考虑到零-均值规范化适应于在已有的样本足够多的情况下比较稳定，适合嘈杂的数据场景，本文使用零-均值规范化对数据集中天气数据的风力、降雨量等特征进行归一化。对原始的数据矩阵进行零-均值规范化后的结果（部分展示结果），如图3-6所示。

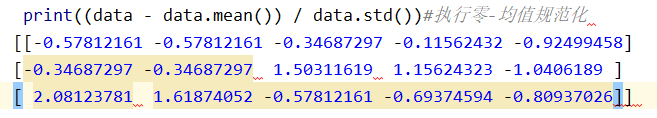


图3-6数据集天气数据特征归一化结果图

### 特征降维-PCA（主成分分析）

特征选择就是单纯地从提取到的所有特征中选择部分特征作为训练集特征，特征在选择前和选择后可以改变值、也不改变值，但是选择后的特征维数肯定比选择前小，毕竟我们只选择了其中的一部分特征。而对于特征选择来说，有很多方法：

1. Filter(过滤式)：VarianceThreshold
2. Embedded(嵌入式):正则化、决策树
3. Wrapper(包裹式)

其中过滤式的特征选择后，数据本身不变，而数据的维度减少。而嵌入式的特征选择方法也会改变数据的值，维度也改变。Embedded方式是一种自动学习的特征选择方法。

PCA（Principal component analysis）：主成分分析，顾名思义，就是找出数据里最主要的方面，用数据里最主要的方面来代替原始数据。具体的，假如我们的数据集是n维的，共有m个数据。我们希望将这m个数据的维度从n维降到n’维，希望这m个n’维的数据集尽可能的代表原始数据集。我们知道数据从n维降到n’维肯定会有损失，但是我们希望损失尽可能的小。那么如何让这n’维的数据尽可能表示原来的数据。通俗点讲：假设100个特征，就是将100个特征，通过PCA降维到10个特征，这10个特征就能代表了这个100个特征的功能。

主成分分析的计算步骤如下。

1. 设原始变量的n次观测数据矩阵为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-5) |

2）将数据矩阵按列进行中心标准化。为了方便，将标准化后的数据矩阵仍然记为X。

3）求相关系数矩阵R，R=（）p×p，的定义为：

|  |  |  |
| --- | --- | --- |
|  |  | (3-6) |

其中，=，=1

4）求R的特征方程det（R-λE）＝0的特征根λ1≥λ2≥λp＞0。

|  |  |  |
| --- | --- | --- |
|  |  | (3-6) |

5）确定主成分个数m：，α根据实际问题确定，一般取80％。

6）计算m个相应的单位特征向量。

7）计算主成分：

|  |  |  |
| --- | --- | --- |
|  |  | (3-7) |

在Python中，主成分分析的函数位于Scikit-Learn下：

sklearn.decomposition.PCA(n\_components = None, copy = True, whiten = False)

在数据集订单车辆表中每个样本有20个特征，各个特征的数据结构信息，如图3-7中所示。

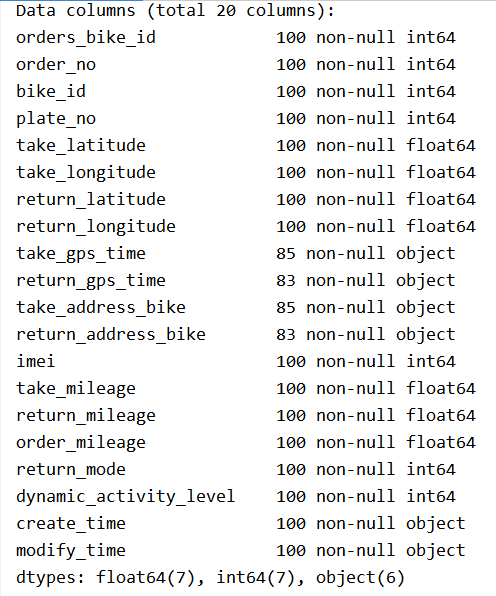


图3-7订单车辆表数据样本特征图

本论文使用PCA对数据集订单车辆表的样本进行降维，为了保证每个维度特征的重要性等同，在运用PCA前做了特征标准化，保存数据集中对方差影响最大的那些特征。本文使用95%的特征降维数目，寻找出最小的属性子集并确保新数据子集的概率分布尽可能地接近原来数据集的概率分布。返回样本的各个特征向量如图3-8所示。

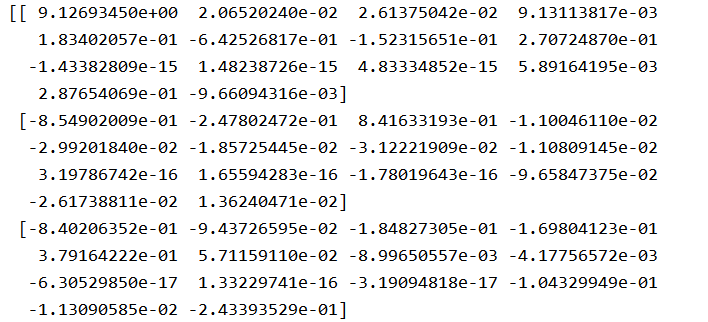


图3-8 PCA降维样本特征向量图

从上面的结果可以得到特征方程det（R-E）＝0有14个特征根、对应的14个单位特征向量以及各个成分各自的方差百分比（也称为贡献率）。其中，方差百分比越大，说明向量的权重越大。

### 特征分析

数据特征分析要求我们在数据挖掘建模前，通过**频率分布分析**、对比分析、帕累托分析、**周期性分析、相关性分析**等方法，对采集的样本数据的特征规律进行分析，以了解数据的规律和趋势，为数据挖掘的后续环节提供支持。

## 人工特征工程

特征工程是把原始数据转化为对预测结果有意义信息的过程，它的目的就是根据本领域的业务逻辑构造新特征，从而提升模型的表现或者更好地解释模型。在进行数据建模之前需要找出有效的特征变量，当研究问题的数据集中没有足够有用的特征时，就要通过创建新特征来为模型的目标变量提供更多的信息，来帮助模型性能的提升。本文目的是预测用户是否会订购，由于个人消费和行为的差异，预测结果也是不同的，接下来会采用人工特征工程方式构建新特征，并使用自动特征工程方法进行对比分析。

### 数据分析及可视化探索

在人工特征构建之前，需要对数据进行初步分析，本文使用描述性统计，来查看数据是否异常，观察数据分布情况，探索数据之间的相关性。我们对可能的积极影响因素做出假设猜测来帮助理解问题，试图挖掘出对准确预测有价值、有意义的信息，为构造特征工程和模型建立做好准备。

1、关于用户表的探索和假设：

用户表共记录了40301个用户及其所在省份，未发现明显的异常值，但用户归属省份有缺失值，将缺失值填补为未知。

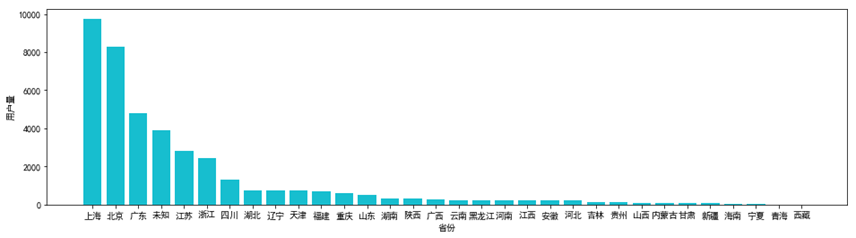


图3-2用户分省分布情况

如图3-2，查看用户所在省份的频数分布情况，加和共计32个省份（含未知），通过可视化图表来看，用户省份分布存在着偏态分布情况，部分省份的用户比较多如北、上、广等城市。

所以我们假设，省份信息可能是比较重要的特征指标，因为位于一线城市的人群消费水平更高，人口稠密程度更大，所以需求和订购量可能更多。

2、关于产品表的探索和假设：

产品主要面向6个大洲方向200多个国家和地区。如图3-3结合不同省份的业务订购频数分布情况来看，没有发现明显的不均衡现象。

所以我们假设，地区可能是比较重要的特征，因为顾客根据个人目的地需要，只能订购固定区域的产品。

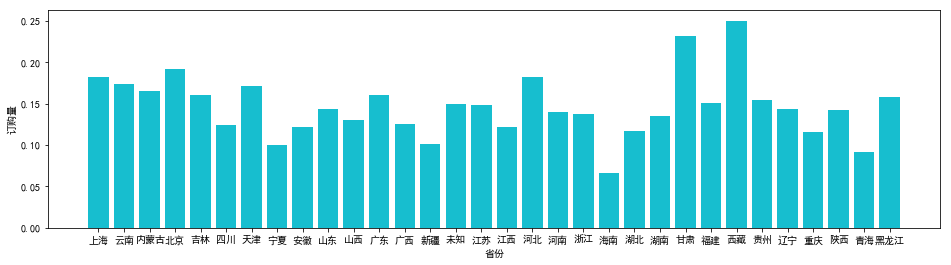


图3-3 业务产品订购分省分布情况

3、关于订购表的探索和假设：

（1）历史订购共有20643笔，根据用户订购情况来绘制箱线图，如图3-4，观察到个别用户订购量过高，属于非正常现象，需要将这部分异常值去掉。

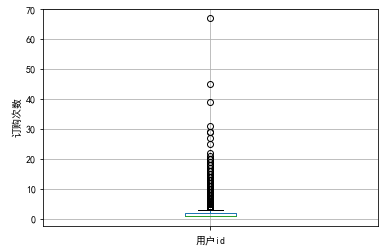


图3-4 业务产品订购情况箱线图

（2）根据历史数据查看业务订购频次，如图3-5，约有60%左右的用户购买一次，剩余40%的用户会有重复订购行为，多次订购次数主要集中在5次以内。从特点来看，可以初步推断业务属于中低频购买类产品，订购时间周期间隔会相对较长。

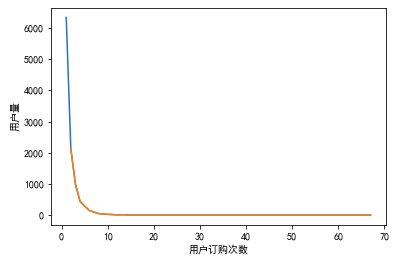


图3-5 历史业务订购频次

（3）查看业务分月订购情况，如图3-6，从趋势来看存在周期性现象。

（注：因9月数据源不完整，所以图中看到数据有显著减少，实际分析时不使用9月，图3-6只是示意业务订购分布而显示）。

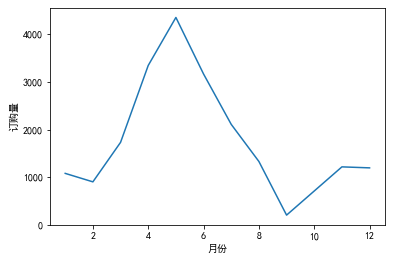


图3-6 历史业务订购量趋势图

根据趋势情况我们假设，订购量、订购频次、订购月份等会是影响用户是否订购的关键特征，可能会在模型当中起到重要作用。

4、关于用户行为表的探索和假设：

（1）历史访问记录共有1261555条，根据用户访问情况绘制箱线图，如图3-7，发现个别用户访问量过高，属于非正常现象，需要去掉这部分离群值。

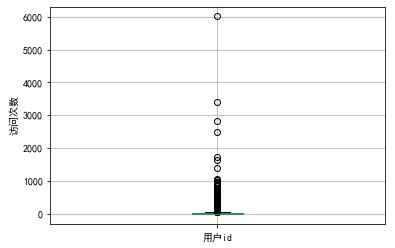


图3-7 用户访问情况箱线图

（2）查看用户分月访问情况，如图3-8，从趋势来看也存在着周期性访问现象，春季4-6月访问量出现显著波峰，9月-10月出现小波峰。

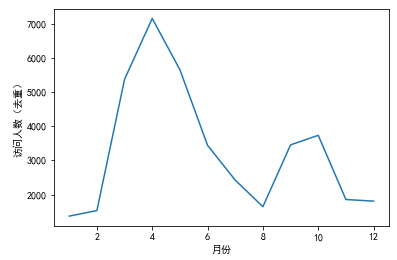


图3-8 分月访问人数（去重）趋势图

（3）将分月的访问和订购分布情况合并对比，如图3-9，发现访问人数和购买订单量的走势呈现成正相关（注：图3-9中下方折线9月订购量数据不完整，此处只为显示时间趋势而展示）。

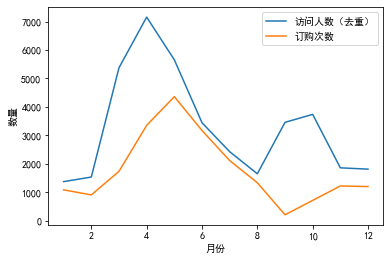


图3-9 分月访问人数（去重）、订购次数趋势图

（4）进一步分析历史访问数据，如图3-10，有访问行为的人隔几天购买，发现一些潜在的规律，约90%的用户访问后会在0~25天内购买，如果不买再买的概率较低。

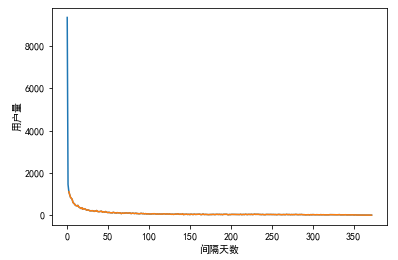


图3-10 有访问行为的人隔几天购买

所以我们假设，访问次数、访问频次、访问月份等可能是关键特征，并且用户在近0~25天内的行为比更早的行为在特征中更重要。另外，我们发现访问人次和订单量可能是正相关，那么对于某些只有访问记录没有订购记录的用户来说，是否也能进行准确的预测呢，本文后边将会设计对照实验来进行研究。

### 特征提取

一般传统特征指的是原始数据中已存在的用户基本属性如所在城市等。但有时我们会发现原始数据中的特征数量非常少，或者特征数量虽然足够多的但是还想要进一步提升模型的性能。这时候我们就要以原始特征为基础，衍生新特征来增加特征的数量，通常效果都会比原来要好。特征衍生：就是指利用业务思维和统计方法衍生出新特征，可以使用的方法有逻辑关联、增量、频率分析等[[[55]](#endnote-54)]。特征组合：指将两个或多个输入特征通过数学的加减乘除运算进行组合。

为什么本文要进行特征提取，因为每一个用户，可能对应多个订购记录，每一次订购对应多条访问轨迹，即对应多行数据。但是每一个用户，是不是订购过只对应于一个标签。也就是说，本文特征工程的单位是基于用户，需要为每一个用户构建一个特征向量。但本文的订购和访问记录中基本特征信息量比较少，因为在很大的程度上，预测模型的性能取决于该模型特征的质量[[[56]](#endnote-55)]，为了模型性能的提升就要提取新特征。比如，订购类数据包含订购量、订购频率等信息，访问类数据包含访问次数，访问频率，访问间隔等信息，这些指标对于预测可能具有一定的价值，因此需要进行一定的转换来提取用户特征：

1、基于单个特征的衍生特征：统计类特征、比率特征、时间特征等；

2、基于两个特征的组合特征：基于业务逻辑增加的新组合特征通常比单个特征效果要好；

3、基于季节性因子的季节特征：从本文业务每个月订购情况来看，存在着季节性波动现象。

下面我们根据业务数据特点，开始特征提取的工作，选择什么样的特征变量能帮助模型效果提升，可以参照第2.1节中影响消费者网上购买行为因素，以及第3.3.1节可视化探索发现的规律来进行选择，具体可以细分为5类：

**一、统计类特征**

统计类特征可以反映出研究对象的数量性和整体性，下面将从数量、频次等方面生成相关统计类特征变量，可以大概推断出不同用户的消费能力、对产品需求范围、购买能力和购买频度、访问时购买动机是否强烈等。

1、用户基本特征：属于用户购买的自然属性信息，本业务涉及到用户省份，可以反映用户的地理标识分布，因为不同省市用户消费能力和人口数量不同，所以是潜在重要特征。

2、产品基本特征：属于用户购买的商品功能属性，包括产品所属地区，产品适用范围有多少国家能使用含有一些信息量。因为用户个体对不同区域和范围内产品会有不同的需求，所以也是比较重要的特征，详见表3-5。

表3-5 统计类特征-基本特征

|  |  |  |  |
| --- | --- | --- | --- |
|  | **特征分类** | **数据类型** | **特征名称** |
| 1、用户基本特征 | 基础特征 | 分类变量 | 用户所在省份 |
| 2、产品基本特征 | 基础特征 | 分类变量 | 不同的产品使用区域 |
| 基础特征 | 分类变量 | 不同的产品使用范围 |

3、订购统计类特征：用户根据自身的使用性需求进行购买，因为从第3.3.1节分析结果知道用户多次订购次数主要集中在5次以内，所以我们需要对每个用户所有业务的订购量进行汇总统计来作为参考，判断用户在这段周期内是否还有购买诉求。

从订购量角度来看，订购频数反映了用户的消费属性，可以发现用户的购买偏好，即用户会高频次订购哪些产品。每个用户的业务订购最大次数，指的是用户订购最多的是哪个业务定了多少次；业务订购个数，是找出用户一共订购了几种类型的业务产品；业务使用区域最大次数，是找出用户订购时，定的最多的是哪个区域的产品，定了多少次。业务使用区域个数，是找出用户一共定了几种区域，反映了用户的购买需求是否多样。

从业务产品订购的受欢迎角度来看，它反映了用户的兴趣属性。我们通过历史数据对业务产品订购量进行排序，找到TOP5的业务产品，和TOP3的订购区域。通过计算可以发现哪些用户订购了这些产品，即每个用户是否订购了TOP5，如果订购的话，得到订购的业务数量。是否订购了TOP3的区域，如果订购的话，得到订购的业务数量，详见表3-6。

表3-6 统计类特征-订购统计特征

|  |  |  |  |
| --- | --- | --- | --- |
|  | **特征分类** | **类型** | **特征名称** |
| 3、订购统计特征 | 订购特征 | 数值变量 | 业务订购量 |
|  | 订购特征 | 数值变量 | 业务订购最大次数 |
|  | 订购特征 | 数值变量 | 业务订购个数 |
|  | 订购特征 | 数值变量 | 业务使用区域最大次数 |
|  | 订购特征 | 数值变量 | 业务使用区域个数 |
|  | 订购特征 | 数值变量 | 订购TOP5业务次数 |
|  | 订购特征 | 类别变量 | 是否订购TOP5业务 |
|  | 订购特征 | 数值变量 | 订购TOP5业务数量 |
|  | 订购特征 | 数值变量 | 订购TOP3区域次数 |
|  | 订购特征 | 类别变量 | 是否订购TOP3区域 |
|  | 订购特征 | 数值变量 | 订购TOP3区域数量 |

4、访问统计类特征：当消费者有浏览行为时，在一定程度上反映了消费者是否具有购买动机，本文按照用户的正常访问订购流程来分，将访问的关键页面和节点，转化为编码的形式，1展现入口，2栏目首页，3产品使用区域介绍页4业务列表页，5业务介绍页，6点击在线购买，7购买确认。其中行为1-4没有访问顺序要求，但行为5-6有先后顺序关系。并且1-5的页面可以根据营销和市场的需求，进行不同类型的推广，所有访问行为记录的多少也会受相关活动因素的影响。虽然目前没有办法获取到推广数据，但是推广类数据作为渠道类信息，也是决定用户是否订购的重要特征因素。

用户访问会产生访问量或点击量，把用户访问和操作行为次数进行汇总，可以得到每个用户所有行为总次数。再进行细分，行为1-5可以累计为与支付无直接关系的非支付行为总次数，行为6-7可以累计为支付行为总次数。还可以具体到每个用户的1-7种行为，得到每个用户每一种操作行为的总次数，对比每个用户各种行为点击量多少，用户购买动机是不是足够强烈。

另外，查看用户在网站的访问深度，用户最后一个行为是什么，是完成了订购结束的，还是没有到最后的订购提交环节就提前离开了，根据最后几个行为可以发现大部分用户都到了哪一步了，离支付的远近程度是多少，是否算作目标客户，因此可以得到每个用户，最后一个行为，倒数第二个行为，倒数第三个行为分别是什么，对比每个用户的行为走到了哪一步，对特征来说很有意义，详见表3-7。

表3-7 统计类特征-订购统计特征

|  |  |  |  |
| --- | --- | --- | --- |
|  | **特征分类** | **类型** | **特征名称** |
| 4、访问统计特征 | 访问特征 | 数值变量 | 所有行为总次数 |
|  | 访问特征 | 数值变量 | 非支付行为总次数 |
|  | 访问特征 | 数值变量 | 支付行为总次数 |
|  | 访问特征 | 类别变量 | 最后一个行为是什么 |
|  | 访问特征 | 类别变量 | 倒数第二个行为是什么 |
|  | 访问特征 | 类别变量 | 倒数第三个行为是什么 |
|  | 访问特征 | 数值变量 | 浏览展现入口(1)次数 |
|  | 访问特征 | 数值变量 | 浏览栏目首页(2)次数 |
|  | 访问特征 | 数值变量 | 浏览产品使用区域介绍页(3)次数 |
|  | 访问特征 | 数值变量 | 浏览业务列表页(4)次数 |
|  | 访问特征 | 数值变量 | 浏览业务介绍页(5)次数 |
|  | 访问特征 | 数值变量 | 点击在线购买(6)次数 |
|  | 访问特征 | 数值变量 | 点击购买确认(7)次数 |

**二、比率特征**

比率特征可以把对比分析的数据变成相对数值，反应了他们相互对比关系。

访问行为比率特征：在用户访问过程中，如果访问深度越深的行为所占比例越高，越有利于我们判断用户具有购买意向。具体方法是我们可以统计每个用户的各种行为所占的比率作为其特征。另外，可以求出每个用户的非支付行为占比，支付行为占比，以及每个用户的7种操作行为各自的占比，详见表3-8。

表3-8 比率特征-访问行为比率特征

|  |  |  |  |
| --- | --- | --- | --- |
|  | **特征分类** | **类型** | **特征名称** |
| 访问行为比率特征 | 访问特征 | 数值变量 | 浏览展现入口(1)占比 |
|  | 访问特征 | 数值变量 | 浏览栏目首页(2)占比 |
|  | 访问特征 | 数值变量 | 浏览产品使用区域介绍页(3)占比 |
|  | 访问特征 | 数值变量 | 浏览业务列表页(4)占比 |
|  | 访问特征 | 数值变量 | 浏览业务介绍页(5)占比 |
|  | 访问特征 | 数值变量 | 点击在线购买(6)占比 |
|  | 访问特征 | 数值变量 | 点击购买确认(7)占比 |
|  | 访问特征 | 数值变量 | 非支付行为占比 |
|  | 访问特征 | 数值变量 | 支付行为占比 |

**三、时间特征**

本文业务具有中低频购买的特性，所以预计订购时间间隔会较长。但对于时间类型来说，在分类模型中无法直接使用，我们可以根据业务特点，将其转换为能反映出用户购买或访问间隔等特征，可以更好的帮助模型预测。

1、订购行为时间分布：从业务订购的时间间隔角度，可以帮助我们发现每个用户对业务产品平均订购周期大概是多久。均值和方差分别反映了订购的平均程度和集中程度，详见表3-9。

表3-9 时间特征-订购行为时间分布

|  |  |  |  |
| --- | --- | --- | --- |
|  | **特征分类** | **类型** | **特征名称** |
| 1、订购行为时间分布 | 订购特征 | 数值变量 | 订购时间间隔均值 |
|  | 订购特征 | 数值变量 | 订购时间间隔方差 |

2、访问行为时间分布：可以查看用户所有操作行为分布的平均程度和集中程度。用均值计算每个用户所有行为的平均时间间隔，反应了每个用户访问所有页面之间间隔平均时间的长短，观察平均时间比较长还是比较短，如果很短说明用户对产品购买期望可能比较大；时间间隔方差，反应了每个用户访问操作行为的集中程度，时间长的多还是时间短的多，如果时间短的多说明购买期望可能比较大。

另外，还可以计算出每种行为类型下，每个用户操作时间间隔的均值、方差。以及计算出每个用户倒数三个行为时间间隔均值和方差，看一下最后几个关键行为的平均程度和集中程度，如果最后关键行为时间比较短也比较集中，说明用户购买可能性比较大，详见表3-10。

表3-10 时间特征-访问行为时间分布

|  | **特征分类** | **类型** | **特征名称** |
| --- | --- | --- | --- |
| 2、访问行为时间分布 | 访问特征 | 数值变量 | 所有行为时间间隔均值 |
|  | 访问特征 | 数值变量 | 所有行为时间间隔方差 |
|  | 访问特征 | 数值变量 | 浏览展现入口(1)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 浏览展现入口(1)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 浏览栏目首页(2)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 浏览栏目首页(2)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 浏览产品使用区域介绍页(3)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 浏览产品使用区域介绍页(3)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 浏览业务列表页(4)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 浏览业务列表页(4)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 浏览业务介绍页(5)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 浏览业务介绍页(5)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 点击在线购买(6)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 点击在线购买(6)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 点击购买确认(7)行为用户每次的操作时间间隔均值 |
|  | 访问特征 | 数值变量 | 点击购买确认(7)行为用户每次的操作时间间隔方差 |
|  | 访问特征 | 数值变量 | 倒数三个关键行为时间间隔均值 |
|  | 访问特征 | 数值变量 | 倒数三个关键行为时间间隔方差 |

3、订购时间差：即用户购买时间间隔，可以反映用户隔多久购买一次。

4、访问时间差：因为用户访问轨迹也可以反映出用户购买动机。我们可以计算出每个用户所有行为最后时间与每种行为最后时间的间隔是多少，最近一次动作是哪种类型的动作，看用户是否已经非常接近于进行最终订购了。

计算出每个用户最后一个行为的时间间隔，倒数第二个行为时间间隔，以及倒数第三个行为时间间隔，锁定用户是否为关键操作行为，因为最后三项行为决定了用户是否能最终完成订购。另外，也可以计算每个用户时间间隔的最大值和最小值，看用户多久操作一次，用户最近访问过相关页面还是已经好久没有操作过，判断用户是否有可能近期进行购买，详见表3-11。

表3-11 时间特征-订购时间差、访问时间差

|  | **特征分类** | **类型** | **特征名称** |
| --- | --- | --- | --- |
| 3、订购时间差 | 订购特征 | 数值变量 | 订购平均时间间隔 |
| 4、访问时间差 | 访问特征 | 数值变量 | 最近一次动作与最近一次浏览展现入口(1)\_时间间隔 |
|  | 访问特征 | 数值变量 | 最近一次动作与最近一次浏览栏目首页(2)\_时间间隔 |
|  | 访问特征 | 数值变量 | 最近一次动作与最近一次浏览产品使用区域介绍页(3)\_时间间隔 |
|  | 访问特征 | 数值变量 | 最近一次动作与最近一次浏览业务列表页(4)\_时间间隔 |
|  | 访问特征 | 数值变量 | 最近一次动作与最近一次浏览业务介绍页(5)\_时间间隔 |
|  | 访问特征 | 数值变量 | 最近一次动作与最近一次点击在线购买(6)\_时间间隔 |
|  | 访问特征 | 数值变量 | 最近一次动作与最近一次点击购买确认(7)\_时间间隔 |
|  | 访问特征 | 数值变量 | 浏览展现入口(1)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 浏览展现入口(1)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 浏览栏目首页(2)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 浏览栏目首页(2)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 浏览产品使用区域介绍页(3)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 浏览产品使用区域介绍页(3)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 浏览业务列表页(4)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 浏览业务列表页(4)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 浏览业务介绍页(5)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 浏览业务介绍页(5)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 点击在线购买(6)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 点击在线购买(6)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 点击购买确认(7)行为用户每次的操作时间间隔最小值 |
|  | 访问特征 | 数值变量 | 点击购买确认(7)行为用户每次的操作时间间隔最大值 |
|  | 访问特征 | 数值变量 | 最后一个行为时间间隔 |
|  | 访问特征 | 数值变量 | 倒数第二个行为时间间隔 |
|  | 访问特征 | 数值变量 | 倒数第三个行为时间间隔 |
|  | 访问特征 | 数值变量 | 所有行为时间间隔最小值 |
|  | 访问特征 | 数值变量 | 所有行为时间间隔最大值 |

**四、组合特征**

对于既有访问又有订购的用户，两种特征组合后放入模型可能会比直接把两种原始特征变量放入模型的解释力要更好。比如，所有行为\_订购占比=每个用户所有动作总次数/每个用户业务总订购量，就可以得到这种关系。反映了用户是访问了很多页面才达成订购，还是浏览了很少的页面就迅速达成了订购，程度是有差异的，详见表3-12。

表3-12 访问和订购组合特征

|  | **特征分类** | **类型** | **特征名称** |
| --- | --- | --- | --- |
| 访问和订购组合特征 | 组合特征 | 数值变量 | 浏览展现入口(1)\_订购占比 |
|  | 组合特征 | 数值变量 | 浏览栏目首页(2)\_订购占比 |
|  | 组合特征 | 数值变量 | 浏览产品使用区域介绍页(3)\_订购占比 |
|  | 组合特征 | 数值变量 | 浏览业务列表页(4)\_订购占比 |
|  | 组合特征 | 数值变量 | 浏览业务介绍页(5)\_订购占比 |
|  | 组合特征 | 数值变量 | 点击在线购买(6)\_订购占比 |
|  | 组合特征 | 数值变量 | 点击购买确认(7)\_订购占比 |
|  | 组合特征 | 数值变量 | 所有行为\_订购占比 |
|  | 组合特征 | 数值变量 | 非支付行为\_订购占比 |
|  | 组合特征 | 数值变量 | 支付行为\_订购占比 |

**五、季节特征**

业务具有购买周期相对较长的特点，根据第3.3.1节中图3-9用户订购和访问趋势来看，用户购买受季节（如春季假期、国庆假期等）的影响，订购量出现了季节性波动，可能对于特征工程具有重要意义，所以将季节性因子的影响纳入考虑范围。

季节特征：按照历史上每个月的订购量和按月的访问量，计算得出每个月订购量和访问量的占比作为权重如下：

历史上每个月订购权重：0.05,0.04,0.08,0.15,0.19,0.14,0.09,0.06,0.05,0.06,0.05,0.05

历史上每个月访问权重：0.03,0.04,0.14,0.18,0.14,0.09,0.06,0.04,0.09,0.09,0.05,0.05

然后再将每个用户每个月的订购权重和访问权重分别进行加权平均，得到每个用户的季节性订购权重和访问权重的值，详见表3-13。

表3-13 季节特征

|  |  |  |  |
| --- | --- | --- | --- |
|  | **特征分类** | **类型** | **特征名称** |
| 季节特征 | 季节特征 | 数值变量 | 订购权重 |
|  | 季节特征 | 数值变量 | 访问权重 |

通过以上的特征提取和生成过程，一共发现并生成潜在特征变量93个，并将它们加入到特征矩阵中。林林总总的特征类别有这么多，预期对模型效果提升都是有一定帮助的。

在工业界有一种说法说明了特征工程的重要性：特征工程决定了学习结果的上界，而选择的分类方法只决定了是否能接近这个上界。我们根据实际业务情况，选择或构建的特征越相关越合理，后期就可以很大程度上减少寻找最优模型而调整参数的时间，也能降低模型的复杂度，让模型更趋于简单对于大多数模型和算法来说，在好的数据特征下表现的性能都不会太差。

### 特征数据处理

利用人工特征抽取生成以上特征体系后，还需要分别对生成的特征进行数据预处理。数据预处理是指处理或删除有问题的数据，对不准确或不适合模型训练的数据进行处理。因为数据可能还存在很多问题，如：数据类型不同（数值、文本、日期等），数值型又分为连续型和离散型；数据质量很差，可能存在异常、噪音、重复、缺失等；数据量纲不统一、数据是偏态、数据样本不平衡、数据量太大或太小等。数据预处理的目的，就是使数据类型和格式更符合模型的需求，更适应模型的使用。

**一、缺失值填补**

其实在收集业务数据以及特征选择和计算的过程中，已经发现了很多缺失值的情况，主要有空值NAN，通常要根据实际业务情况来填补缺失值，比如用户所在省存在NAN，可以填补成“未知”。对于一些空值NAN当不影响特征含义的时候，可以用0填补。另外还有INF表示无穷大，当我们计算比例的时候，通常不能将其直接替换成0，这样会改变特征代表的含义，可以根据情况将其替换成NAN表示数据不存在。

**二、处理数值型特征：无量纲化**

数据无量纲化，指机器学习里一般要把数据从多种不同的规格更改为相同规格的数据，或将数据从不同的分布转换为符合特定分布的数据。无量纲的原理就是使用不同的计算方法，对数据进行不同方式的缩放。它的作用是，加快模型收敛速度，提高模型的精度，并可以避免一些很大的值对距离计算的影响。常用的无量纲方法是数据归一化和数据标准化。

数据归一化(Normalization，又称Min-Max Scaling)：是把数据x以最小值进行中心化后，然后按极差（即最大值与最小值间的差距）进行缩放，相当于把数据移动了最小值个单位，收敛到[0,1]区间内，从而使数据遵循正态分布。数据归一化对异常值较为敏感，适合不太符合正态分布且不涉及到距离的数据，常用于数字图像处理等问题中。其公式为：

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

数据标准化(Standardization，又称Z-score Normalization)：是把数据x以均值进行中心化后，然后按标准差进行缩放，使数据遵循均值为0，标准差为1的正态分布。数据标准化对异常值不太敏感，适用于基于正态分布的算法，使用距离来衡量相似性的数据时效果更好，常用于逻辑回归、聚类、神经网络或PCA进行降维等问题中。其公式为：

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

观察第3.3.2节中特征指标中的所有数值变量，大都不符合模型需要的无量纲化要求，除了那些表示分类型的变量和表示占比的变量，将所有的数值型变量都采用数据标准化来进行无量纲化。

**三、处理分类型特征：编码与哑变量**

在机器学习中大部分算法（比如逻辑回归等）只能处理数值型的数据，不能处理文本型数据。（注：本文采用的是Sklearn库，在算法训练时只能导入数组或矩阵，不支持文本数据的输入。）但实际案例中，我们提取到的特征和标签数据类型是多种多样的，不能都以数字形式显示。（举例，银行办理的“银卡，金卡，白金卡”等，就属于分类型数据）。为了使数据符合算法和库的要求，必须对数据进行编码，把那些文本型数据转换为数值型数据。

1、分类型变量有三种分类方式：名义变量、有序变量、有距变量

（1）名义变量：比如性别（男、女），彼此相互独立，没有完全联系。

（2）有序变量：比如学历（小学、中学、大学），性质上有小学<中学<大学这样的联系，但是取值之间是不可以计算的。

（3）有距变量：比如重量（1斤、2斤、3斤），如1斤+2斤=3斤，分类之间可以通过数学计算进行转换。

我们对分类型变量的特征进行编码的时候，如果单纯的把文本数据转换为数值型，比如重量（1斤、2斤、3斤）转换为数字[0,1,2]，默认情况下数值就获得了加减计算的数学性质。但对于性别和学历来讲，它们不属于有距变量，只有有距变量才具有加减的关系。当遇到名义变量或有序变量转换为数值型数据时，其含义是不一样的，需要进一步处理去掉数值的这些数学特性，才能保证我们建模的准确性。

2、对于不同分类型变量转换为数值型数据处理方法如下：

有距变量：使用LabelEncoder，能够直接将分类数据转换为分类数值；

有序变量：使用OrdinalEncoder，能够将分类特征转换为分类数值；

名义变量：使用OneHotEncoder，创建哑变量，使用独热（one-hot）编码。

3、哑变量（Dummy Variable），也叫作虚拟变量，目的是用1和0将不能定量处理的分类变量进行量化，如果存在n个分类属性，1就是所选的分类，剩下的n-1个0就是哑变量[[[57]](#endnote-56)]。（注：分类变量首先转换为整数值，然后再转换为二进制表示形式，标记为1或0。如果字符形式直接进行one-hot操作则会报错。）离散的分类变量进行one-hot编码后，可以消除数据中的噪声并加快模型的收敛速度，才能将分类信息以最准确的形式传递给算法。举例，可以把学历用哑变量表示为：

小学 100

中学 010

大学 001

4、观察第3.3.2节生成的特征指标中，发现那些具有独立特性的分类变量都属于名义变量，如：

（1）用户基本特征中用户所在省份，目前属于字符型，需要先通过LabelEncoder方法将分类转换为数值型，然后在通过OneHotEncoder进行独热编码变成哑变量。

（2）目前已经用数值型表示了的分类变量，如第几次访问行为，对其直接进行OneHotEncoder独热编码变成哑变量即可。

## 自动化特征工程

自动化特征工程指的是通过数据集自动创建候选特征，来进行训练的一种方式。本文采用的FeatureTools是一个执行自动特征工程的开源框架，它是基于深度特征合成来叠加多个转换和聚合操作，能够把相关联的数据集转换为特征矩阵。它能够快速推进特征的生成过程，从而把时间放在构建机器学习模型的其他方面。主要包含三个组件：

1、实体和实体集：一个实体就是一张表，一个实体集是指多个表以及他们之间所有的关联规则，关联规则是指定连接两张表的变量，比如用户表和订购表通过user\_id进行变量连接。

2、特征基元：是我们用来构造新特征的操作，包括转换和聚合。

转换：对一张表中一列或多列完成的操作。比如，对一个表中的时间字段取月份，或对表中两列时间取时间差。

聚合：根据父与子（一对多）的关联完成的操作，也就是根据父亲来分组计算儿子的统计量。比如，根据user\_id对订购表分组求每个user\_id的订购量。

3、深度特征合成DFS（Deep Feature Synthesis）：指的是把特征基元应用于实体集中用实体关系来创建特征，它与深度学习无关。

本文通过自动化特征工程的方法，（1）建立用户实体集，实体包含用户表、产品表、订购表、行为表。（2）添加实体之间的关系：通过user\_id创建用户表和订购表之间的关系，通过user\_id创建用户表和行为表之间的关系，通过product\_id创建产品表和订购表之间的关系。（3）没有人为指定，通过完全自动的方式，让特征工具为我们选取特征，聚合特征后，共生成37个新特征，举例示意详见表3-14。

表3-14 自动化生成的新特征举例示意

| **自动化生成的新特征** | **特征的解释** |
| --- | --- |
| prov | 用户所在省份 |
| COUNT(order\_info) | 所有业务订购量 |
| NUM\_UNIQUE(order\_info.product\_id) | 订购不同产品的数量 |
| MODE(order\_info.product\_id) | 订购最多的产品 |
| COUNT(action\_info) | 用户所有行为总次数 |
| …… | …… |
| MODE(action\_info.actionType) | 用户操作最多的行为 |

# 模型应用及实验分析

## 实验条件

本次实验从样本数据预处理、特征工程、预测模型应用到模型评估等环节需要具备的主要实验条件，详见表4-1。

表4-1 实验条件列表

|  |  |
| --- | --- |
| **笔记本** | **配置** |
| 处理器 | Intel(R) Core(TM) i5-8250U CPU @1.60GHz 1.8GHz |
| 内存 | 8G |
| 操作系统 | Windows 10（64位） |
| JDK | Java version 1.8.0 |
| Python | Python version 3.7.1 |
|  | Anaconda3 (64-bit) |
|  | Numpy version 1.16.1 |
|  | Pandas version 0.23.4 |
|  | Matplotlib version 3.0.2 |
|  | SciPy version 1.1.0 |
|  | Scikit-learn version0.20.1 |
|  | Imblearn version0.0 |
|  | Xgboost version 0.90 |
|  | Featuretools version 0.11.0 |

## 准备数据集

### 设计特征对照组

在第3章特征工程构造时，我们采用了人工和自动化两种方法，那么他们的效果如何，通过设置如下对照实验来观察：

1、人工特征设置五个对照组进行对比实验

对于本文业务数据，在第3.3.1节可视化折线图中，我们发现用户访问和购买行为可能存在相关系，因为可以假设一般访问量高的时候订购量也高，所以在第3.3.2节构建了用户的购买和浏览行为的组合关系。那么对于那些暂时没有购买而有浏览行为的，只能用户特征+访问特征进行预测，但有了订购相关的特征后是否会更精准的进行预测呢？以及假设去掉访问特征，只有用户特征+订购特征，是否可以有好的预测结果，再继续添加访问和订购组合特征，以及添加季节性特征，在构建特征时会不会预测更精准呢？为了让我们模型的预测准确性更高，是否要添加这些特征进去，我们设计如下实验对照组来进行验证。

2、自动特征和人工特征对照组中最好的一组进行对比实验

自动化特征在构建速度上比人工特征工程更高效，但是生成的特征是否有效，对模型预测效果是否精准的呢，我们也通过设置对比实验来进行验证，详见表4-2。

表4-2 业务订购预测模型人工特征和自动特征对照组

|  |  |  |
| --- | --- | --- |
| **分类** | **分组** | **特征名称** |
| 人工特征 | 对照组1 | 用户基本特征+访问特征 |
| 对照组2 | 用户基本特征+订购特征 |
| 对照组3 | 用户基本特征+访问特征+订购特征 |
| 对照组4 | 用户基本特征+访问特征+订购特征+访问和订购组合特征 |
| 对照组5 | 用户基本特征+访问特征+订购特征+访问和订购组合特征+季节特征 |
| 自动特征 | 对照组6 | 自动特征集合 |

说明：

对照组1：目的是人工创建特征时，对用户是否购买进行预测，假设去掉订购记录，只有基本特征和访问特征时，看没有订购记录时模型预测效果如何。

对照组2：目的是人工创建特征时，对用户是否购买进行预测，假设去掉访问记录，只有基本特征和订购特征时，看没有访问记录时模型预测效果如何。

对照组3~5：目的是人工创建特征时，因为数据中有很多是只访问了目前暂未购买的用户，所以在对照组1（用户基本特征+访问特征）的基础上，依次增加订购特征、组合特征以及季节特征，对用户是否购买进行预测，看加入新特征的模型效果是否更好，以及新特征的贡献度情况。

对照组6：目的是当采用自动化方法创建特征时，对用户是否购买进行预测，看模型表现效果如何，以及新特征的贡献度情况。

如上所述，经过一系列特征提取和分组，为后续的实验做好了特征准备。

### 设计数据集划分策略

根据第3.3.1节中分析出的数据特点，并且每个月订购分布不均匀的现象，一种方式是利用更多的历史数据来进行预测，目前无法实现更早数据的提取。故换一种方式，因为约90%的用户访问后会在0~25天内购买，所以取用户访问和购买的时间差一般在25天以内的数据进行预测，考虑到样本太少可能会不准确，所以用近2个月的数据，对下一个月用户是否购买进行预测。划分数据集的策略，详见表4-3。

表4-3 样本数据集划分策略

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **特征构造方法** | **对照组** | **数据集** | **特征提取区间** | **用户集提取区间**  **（有订购或**  **有访问的用户）** | **标签**  **区间** |
| 人工特征工程 | 对照1~5 | 训练集 | 18年11月~19年6月 | 5月、6月 | 7月 |
| 测试集 | 18年12月~19年7月 | 6月、7月 | 8月 |
| 自动特征工程 | 对照组6 | 训练集 | 18年11月~19年6月 | 5月、6月 | 7月 |
| 测试集 | 18年12月~19年7月 | 6月、7月 | 8月 |

根据业务特点进行数据集划分，训练集和测试集提取方式如下：

训练集，选择预测期7月前2个月（5-6月）有订购或有访问的用户作为对象，并在2018年11月~2019年6月之间进行特征提取，然后再关联7月用户是否订购的信息作为标签列。

测试集，选择预测期8月前2个月（6-7月）有订购或有访问的用户作为对象，并在2018年12月~2019年7月之间进行特征提取，然后再关联8月用户是否订购的信息作为标签列。

我们按照第4.2.1节的特征分类方法，来提取构建特征矩阵，并且针对六个对照组的样本数据分别生成训练集和测试集。

### 样本不均衡数据处理

样本不平衡是指在一组数据集中，标签的某一类天生占有很大的比例，本文中业务订购预测数据，正样本为订购业务，负样本为未订购业务，通过对正负样本比例进行计算，正样本不足10%，但我们业务订购预测的目标是捕捉少数类，属于正负样本显著不均衡的数据集。在尝试使用模型评估的过程时发现预测能力特别差，这样训练的模型因为数据倾斜导致模型不准。故需要增加正样本比例，算法采样时可采用的方法有上采样、下采样、混合采样[[[58]](#endnote-57)]等。

上采样，指以数据量多的一方的样本数量为标准，把样本数量较少的类的样本数量生成和样本数量多的一方相同，称为上采样。可以使用的方法有随机上采样和SMOTE（Synthetic Minority Oversampling Technique）算法等，SMOTE采用人工合成少数类的数据策略，把少数类样本分析合成新样本后添加到原数据集中，这样可以使生成样本的数量和分类样本较多的数据量保持一致[[[59]](#endnote-58)]，且在处理样本不均衡问题时性能比较高。下采样，指对于一个不均衡的数据，让目标值(如0和1分类)中的样本数据量相同，且以数据量少的一方的样本数量为准。可采用的方法有随机下采样和Informed undersampling采样技术等。混合采样法，是一种结合上采样和下采样的方法。

本文中的标签比例不均衡会影响建模效果，但业务数据本身并不多，所以放弃了下采样的方法，使用了imblearn库的SMOTE上采样算法，处理后使数据集中的正负样本各占一半，以提高预测效果。

## 订购业务预测分析

本次业务订购预测实验中所有对照组都使用逻辑回归（LR）、决策树(DT)、XGBoost(XGB)三种机器学习预测算法进行对比实验。主要步骤是：获取对照组的实验数据，确定参数和模型检验方法，并对结果进行评估。

### 模型参数调整

先根据第4.2.3节方法将数据集进行上采样使目标样本量相同，再通过实验效果画模型参数学习曲线（Learning curve）等方式进行模型参数调整。

1、逻辑回归模型需要的主要参数设置，详见表4-4。

表4-4 逻辑回归模型需要的主要参数设置

| **主要参数** | **参数意义和取值** |
| --- | --- |
| Penalty | 正则化方式，支持的惩罚项有“L1”或“L2”两种正则化方式，不填写默认“L2”，本文选择“L1”，因为当设置为“L2”模型在数据集上的效果表现很差，通过画学习曲线对比后本文选择了L1正则化。 |
| solver | 求解器的选择，当选“L1”正则化时，参数solver仅能够使用求解方式“liblinear”和“saga”，本文solver设置为坐标下降法“liblinear”，“liblinear”是二分类专用，也是默认的求解器。 |
| C | C是正则化强度的倒数，必须是一个大于0的浮点数，不填写默认为1，即默认正则项与损失函数的比值是1:1。C越小，损失函数会越小，模型对损失函数的惩罚越重，正则化的效力越强，参数会逐渐被压缩得越来越小。通过画学习曲线对比后本文选择设置C为0.5。 |
| max\_iter | 梯度下降重要参数，逻辑回归目的是求解能让模型最优化，并且拟合程度最好的参数的值，即求解能够让损失函数最小化的值。用梯度下降法(Gradient Descent)求解二元逻辑回归，设置参数max\_iter最大迭代次数来代替步长，帮助我们控制模型的迭代速度并适时地让模型停下。max\_iter越大步长越小代表模型迭代时间越长，反之，步长设置很大代表模型迭代时间很短。为了让逻辑回归在限定的步数内找到损失函数的最小值max\_iter，本文max\_iter设为500。 |

2、决策树模型需要的主要参数设置，详见表4-5。

表4-5 决策树模型需要的主要参数设置

|  |  |
| --- | --- |
| **主要参数** | **参数意义和取值** |
| Criterion | 特征选择标准，用来衡量分枝质量的指标，有“gini”基尼系数和“entropy”信息增益两种，本文使用“gini”不纯度决定划分。 |
| Splitter | 特征划分标准，在特征的所有划分点中找出最优的划分点，所有特征上进行递归，适用于数据集较小的时候，本文使用best。 |
| max\_depth | 决策树最大深度，切分循环次数，常用来防止过拟合，本文设为5，使模型拟合度和准确度均达到最优。 |
| min\_samples\_split | 子数据集内部需要切分的最小样本量，默认是2，本文取2。 |
| min\_samples\_leaf | 叶节点最小样本数，否则会被剪枝，本文取1。 |

3、XGBoost模型需要的主要参数设置，详见表4-6。

表4-6 XGBoost模型需要的主要参数设置

| **主要参数** | **参数意义和取值** |
| --- | --- |
| 1、通用参数 |  |
| Booster | 弱评估器，一种是gbtree，一种是gbliner，本文我们选用树模型gbtree。 |
| Nthread | 表示多线程所使用的CPU核数，本文我们设置为8。 |
| Silent | 是否打印迭代信息，本文我们改为1不出现，因为默认是0默认开启打印。 |
| 2、Booster参数 |  |
| Eta | 集成中的学习率，又称为步长，默认值为0.3，可以控制每一步权重，减少模型鲁棒性，典型取值0.01~0.2，本文我们使用0.025。 |
| max\_depth | 表示决策树最大深度，切分循环次数，常用来防止过拟合，本文我们设为5。 |
| min\_child\_weight | 表示最小叶子节点的样本权重和，它的作用是避免过拟合，当它的值较大的时候，可以避免模型到局部的特殊样本，但过高了又欠拟合，本文我们设置为2。 |
| Subsample | 随机抽样的样本比例，默认是1，减少的话会趋于保守，典型取值0.5~1，本文我们取平均值0.75。 |
| colsample\_bytree | 它和GBM里的max\_feature参数类似，控制每一颗随机采样列数的占比，适当调整可以防止过拟合，典型取值0.5~1，本文取平均值0.75。 |
| lambda | 是L2正则化项的系数，有减少过拟合的用途，本文我们使用默认值1。 |
| 3、学习目标参数 |  |
| Objective | 定义了最小化的损失函数，预测目标概率，因为预测的是二分类问题，本文用binary:logistic逻辑回归概率。 |
| eval\_metric | 评估概率分数，本文我们用auc。 |

### 模型效果评估检验

将所有实验分类下的特征输入三种机器学习算法模型，然后对预测结果进行分析，并从检验结果的预测准确率、分类的效果程度（AUC）、预测模型损失值（Logloss）、运行花费时间，这四个方面进行比较。

**一、人工特征各对照组对比**

**（一）、检验结果的预测准确率**

1、人工特征5个组都分别调用混淆矩阵中的评估指标，返回四种指标：准确度（Accuracy）、召回率（Recall）、精确度（Precision）和F1-score，如图4-1。

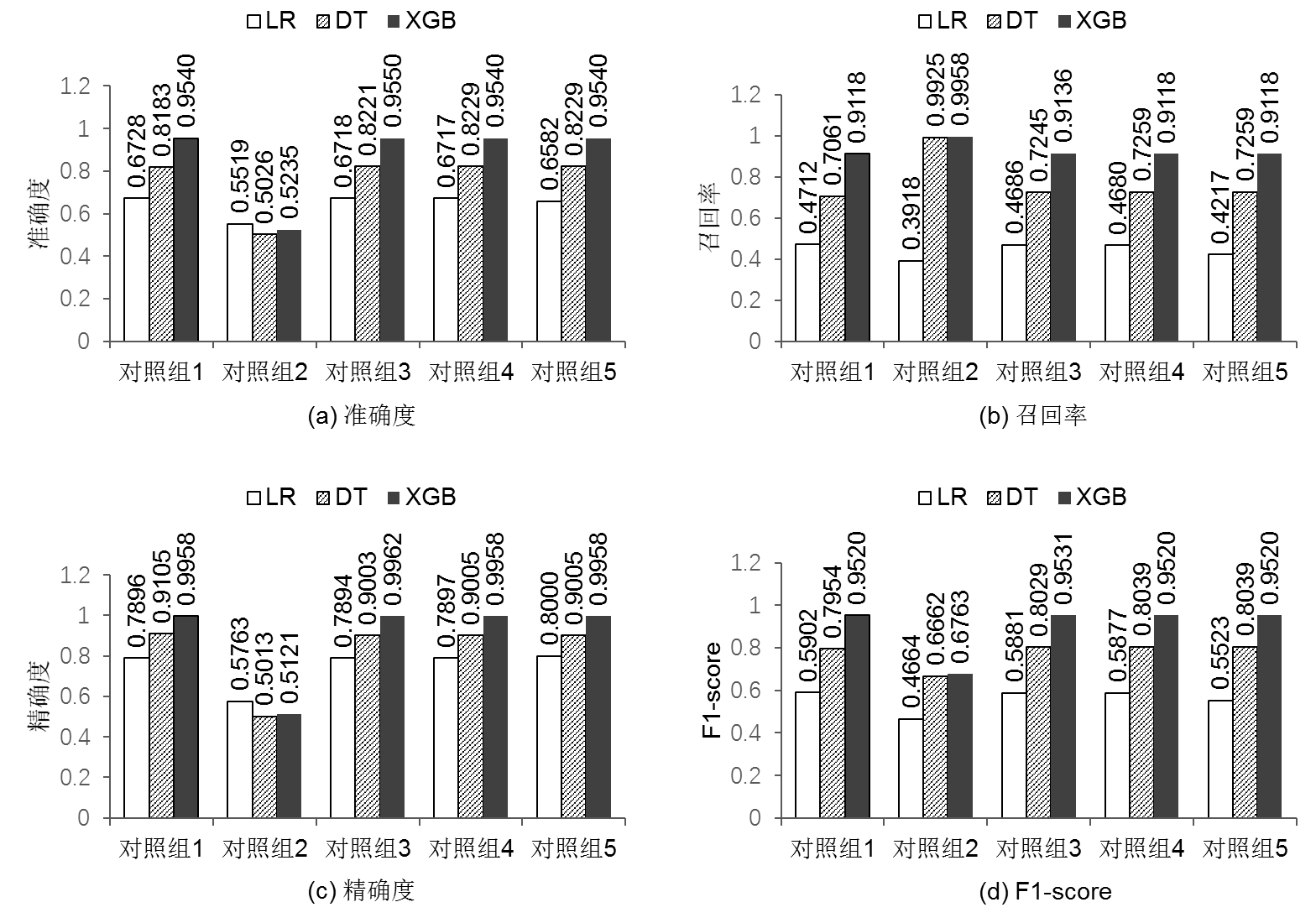


图4-1 人工特征选取对照组三种机器学习算法预测情况

从图4-1来看，对于准确度（Accuracy）指标，XGBoost预测模型在对照组3时，值为0.955是各组中最高的。对于召回率（Recall）指标，XGBoost在对照组2时，值为0.9958是各组中最高的，它说明XGB算法在实际的小样本用户中，预测对小样本用户所占的比例。对于精确度（Precision）指标，XGBoost在对照组3时，值为0.9962是各组中最高的，表示模型预测的小样本用户中，实际的小样本用户所占的比例。对于F1-score指标，XGBoost在对照组3时，值为0.9531是各组中最高的，且它是综合考虑召回率和精确度率后的调和值。

综上，并且根据F1-score调和指标来看，三种算法在对照组1（用户基本特征+访问特征）只有访问特征时表现比较好，在对照组2（用户基本特征+订购特征）当只有订购特征时表现最差，对照组3（用户基本特征+访问特征+订购特征）综合了访问和订购特征后，除LR外各个评分指标都有一定程度提升，但对照组4和5继续增加访问和订购组合指标以及季节因子后对评分值的提升不是特别显著。初步表明访问特征比订购特征对模型贡献更大，但两种特征都加入时比单一特征效果要更好一些。

2、使用Sklearn的classification\_report函数，对各组数据二分类的准确率进行计算，输出分类指标的文本报告，包括精确度、召回率和F1-score值，对比在人工特征分组下，算法预测结果的准确率，如图4-2。

图4-2细化了正样本（Class1）和负样本（Class0）的评分指标，从F1-score调和指标来看，三种算法在对照组1（用户基本特征+访问特征）只有访问特征时表现比较好，在对照组2（用户基本特征+订购特征）当只有订购特征时各项得分都比较差，增加特征后的对照组3、4、5，对于正负样本各项评分指标都没有显著变化，只有决策树在对照组5的正样本召回率比对照组1有0.02分的提升。对于XGB增加访问相关指标，访问和订购组合指标以及季节因子后对正样本和负样本评分值增长不是特别显著。以上正负样本细化评分结果表明，访问特征比订购特征对模型贡献更大，对于决策树多种特征都加入时比单一特征效果要更理想。

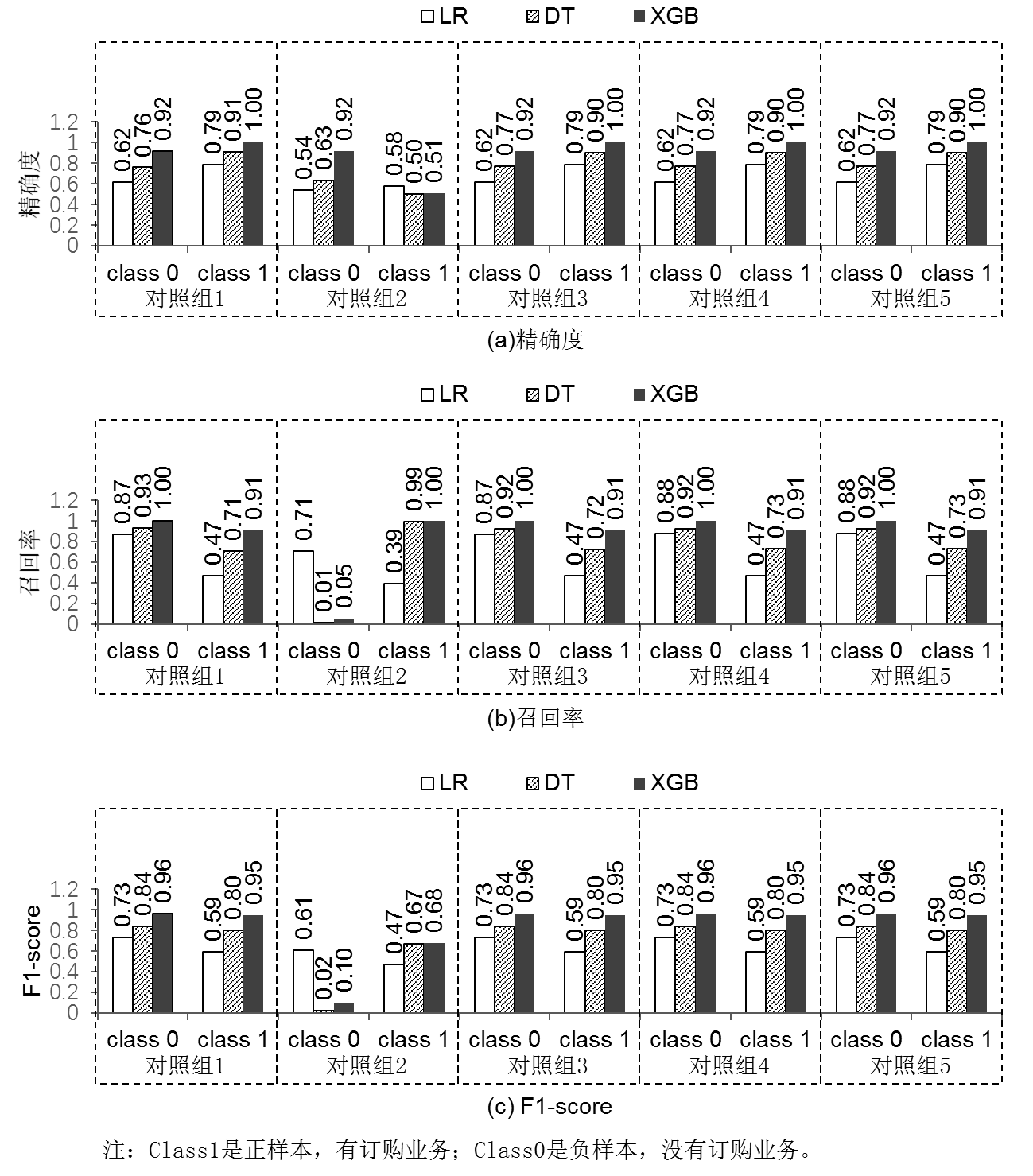


图4-2 人工特征选取对照组三种机器学习算法二分类预测情况

**（二）、分类的效果程度（AUC）**

在人工特征对照组下，对三种机器学习算法的预测结果的ROC曲线进行对比，如图4-3。

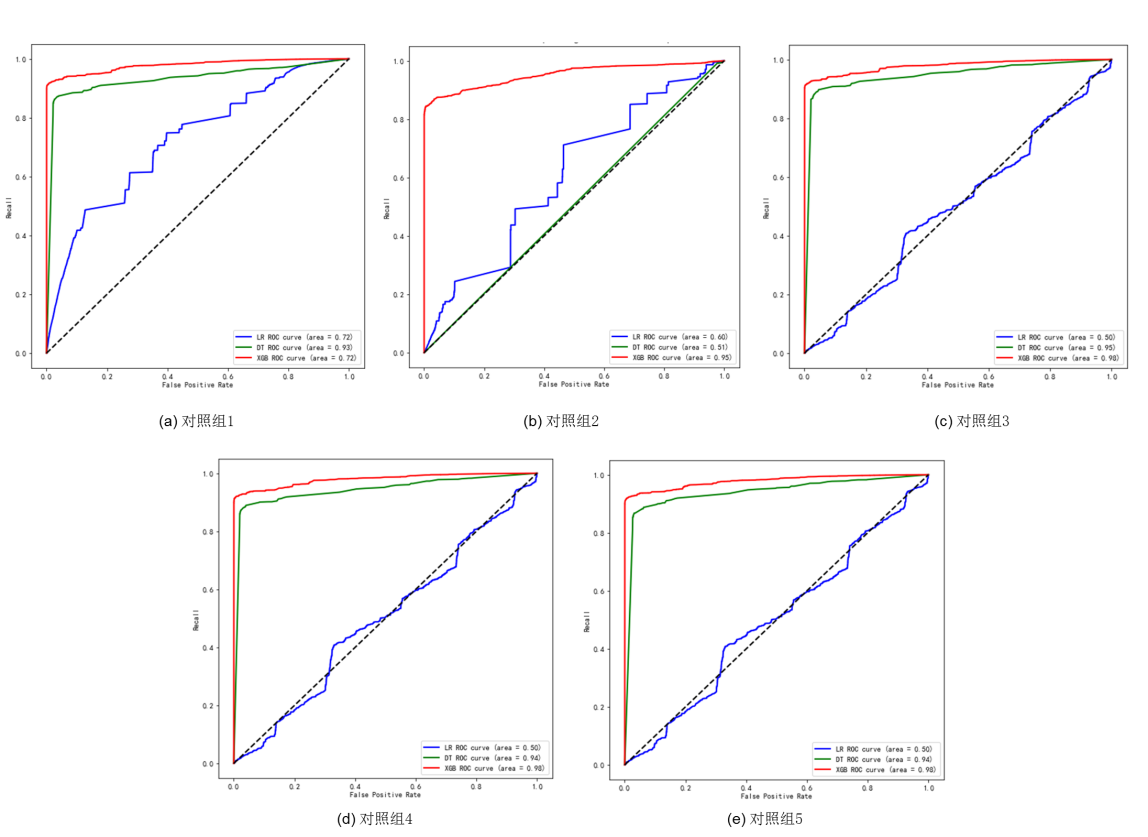


图4-3 人工特征选取对照组三种机器学习算法的ROC曲线

AUC指的是ROC曲线下的面积，它是平衡了真正率（True Positive Rate, TPR）和假正率（False Positive Rate, FPR）后的结果，值越大说明分类的效果越好。在人工特征对照组下，对三种机器学习算法的预测结果的AUC值进行对比，如图4-4。

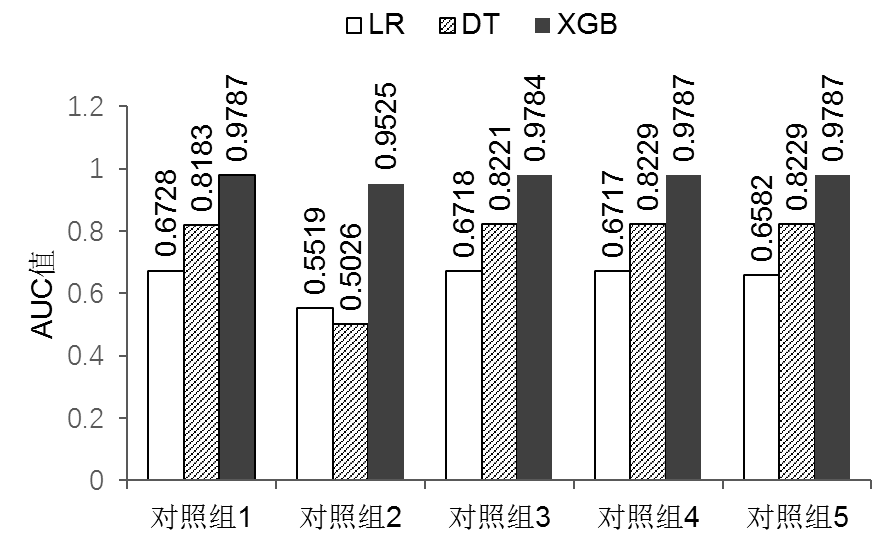


图4-4 人工特征选取对照组三种机器学习算法的AUC值

（1）从特征角度对比分析，通过查看ROC曲线和AUC值，三种算法在对照组1（用户基本特征+访问特征）只有访问特征时表现比较好，在对照组2（用户基本特征+订购特征）当只有订购特征时表现最差，增加特征后的对照组3、4、5，AUC值没有明显变化，只有决策树在对照组5比对照组1有0.0046分的提升。对于XGB增加访问相关指标，访问和订购组合指标以及季节因子后对评分值增长不是特别显著。故，从AUC值的对比也表明，访问特征比订购特征对模型贡献更大，对于决策树多种特征都加入时比单一特征效果要好一些。

（2）从算法角度对比分析，因为当AUC值大于0.5时才有意义，第2组的LR和DT都在0.5左右，所以不做考虑。综合除第2组外其他各组的情况，三种模型AUC值从好到坏综合排序为XGB>DT>LR。

**（三）、预测模型损失值（Logloss）**

在人工特征对照组下，对三种机器学习算法预测结果的Logloss值进行对比，如图4-5。

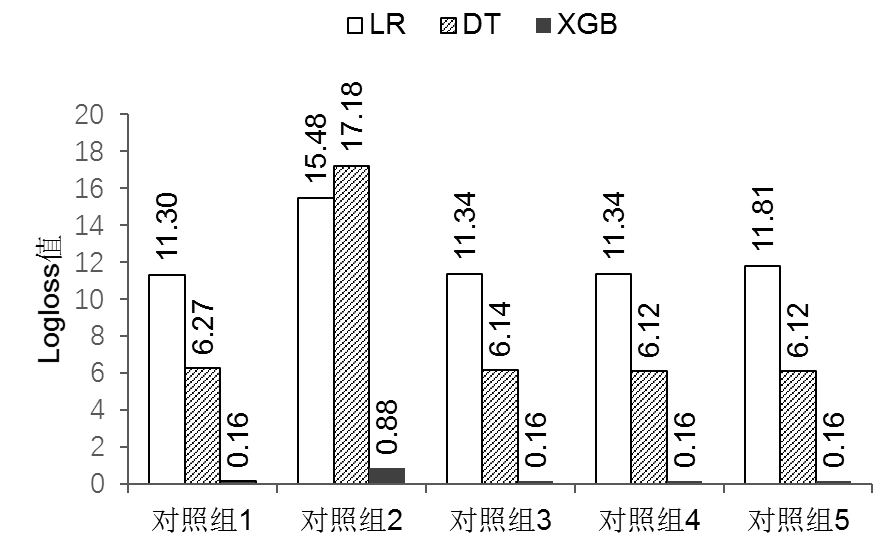


图4-5 人工特征选取对照组三种机器学习算法的Logloss值

（1）从特征角度对比分析，三种算法在对照组1（用户基本特征+访问特征）只有访问特征时损失值比较小，在对照组2（用户基本特征+订购特征）只有订购特征时损失值最大，增加特征后的对照组3、4、5的Logloss值和对照组1相比没有显著变化，只有决策树在对照组5比对照组1损失值减少了0.15。对于XGB增加访问相关指标，访问和订购组合指标以及季节因子后对Logloss值变化不是特别显著。故，从Logloss值的结果也表明，访问特征比订购特征损失值小，因此对预测模型贡献更大，对于决策树多种特征都加入时比单一特征效果要好一些。

（2）从算法角度对比分析，Logloss值可以惩罚错误的分类，值越小越好，模型预测准确度越高，综合各组的Logloss值情况进行大小排序XGB<DT<LR。

**（四）、预测模型运行花费时间**

在人工特征对照组下，对三种机器学习模型训练的运行时间进行对比，如图4-6。

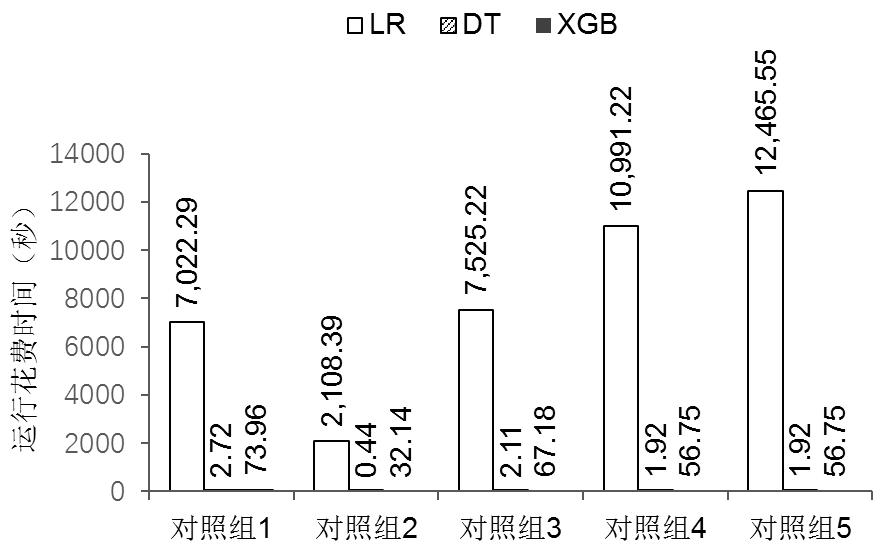


图4-6 人工特征选取对照组三种机器学习算法的运行花费时间

逻辑回归模型的运行时间[[[60]](#endnote-59)]取决于梯度下降的时间复杂度O(N\*C\*I)，N表示样本数量（N=1为随机梯度下降，N=mini-batchsize为mini-batch梯度下降），C表示单个样本的计算量（取决于梯度计算公式），I表示控制的迭代次数（取决于收敛的速度）[[[61]](#endnote-60)]。

决策树的运行时间取决于分枝的时间复杂度O(N\*M\*D)，N表示样本量大小，M表示特征的个数，D表示生成树的深度。当CART树生长时，需要把所有特征内的值都作为一个分裂的候选，并为它计算一个评价指标，例如信息增益、增益比率、GINI系数等[[[62]](#endnote-61)]，所以每一层的时间复杂度为O(N\*M)，D层树总的时间复杂度就是O(N\*M\*D)。XGBoost对应的模型就是一堆CART树，原理就是将每棵树的预测值加到一起作为最终的预测值。

（1）5个组根据特征数量从小到大排序为：对照组2<对照组1<对照组3<对照组4<对照组5，所以每组运行时间的结果也随特征数增加成正比例增长。

（2）三种算法运行时间从小到大排序为：DT <XGB <LR，从运行时间长短来看，DT是最快的，XGB次之，LR的运行时间最长。对于逻辑回归除第2组约1个小时左右，其余每组的平均运行时间都在2小时以上。对于决策树本文为防止过拟合将CART决策树的max\_depth设置为5，从结果来看决策树能够快速分枝，运算速度相对较快，除第2组约0.44秒左右，其它各组平均2秒左右运行完毕。XGBoost运行时间相当于多棵树的叠加，除第2组约半分钟左右，其它各组平均运行1分钟左右出结果。

综合以上四项评估结果，在人工特征的5个分组中，第5组整体预测效果较为理想；预测模型中效果最优的是XGBoost，各项评分均比较高、损失值最小、模型运行时间比较短，其次是决策树，而对于逻辑回归各项指标都不太理想。

**二、人工特征和自动特征对比**

选择人工特征表现最好的第5组和自动化方法生成的特征，基于三种机器学习算法分别画ROC曲线进行对比，如图4-7。

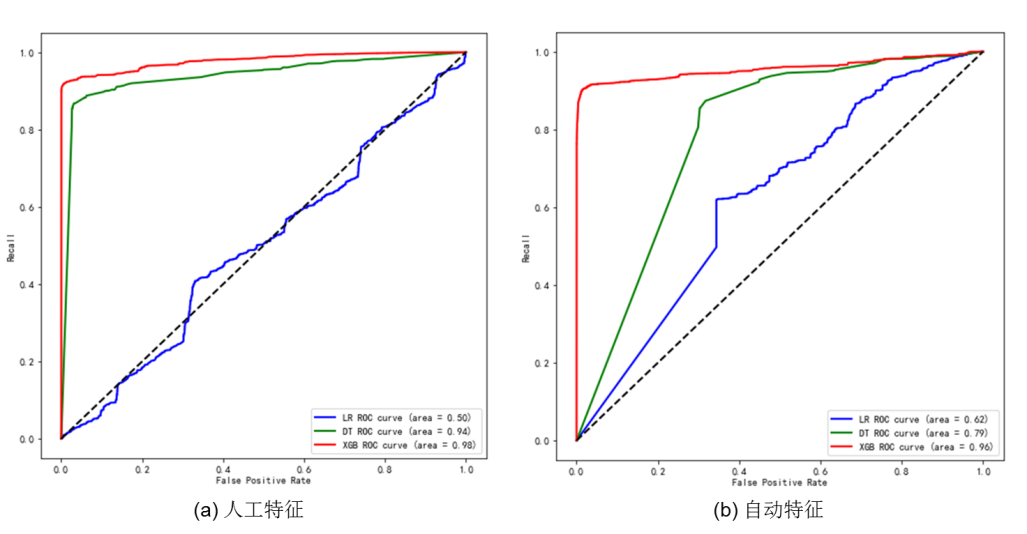


图4-7 人工特征与自动特征的ROC曲线对比

将人工特征对照组整体表现最好的第5组与自动特征，在F1-score、AUC值、Logloss值、运行花费时间这几个方面，基于三种机器学习算法分别进行得分对比，如图4-8。

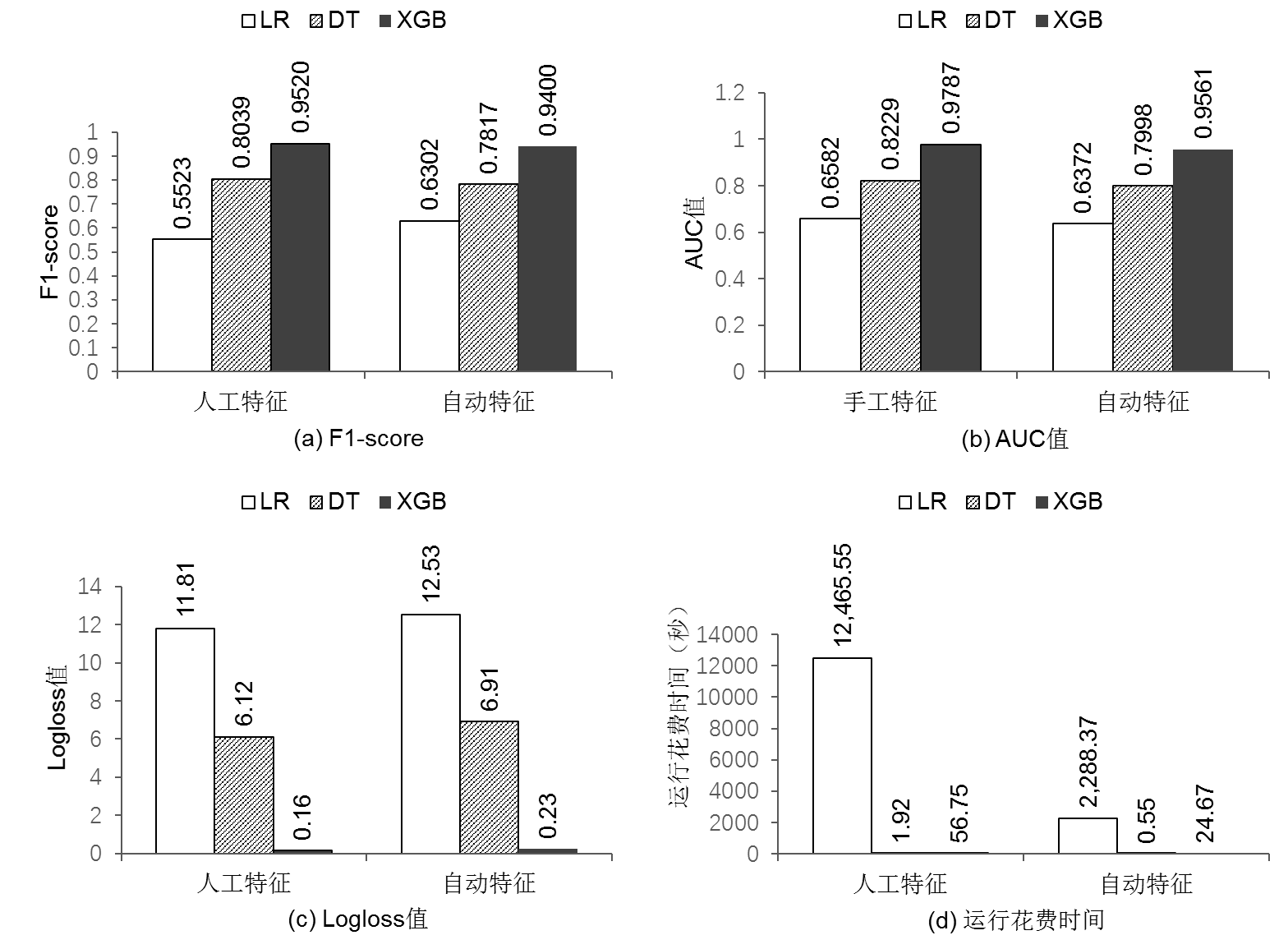


图4-8 人工特征与自动特征的评估指标对比

从图4-7和图4-8的结果来看，两种特征选择方式下，模型的整体效果从好到坏排序都为XGB>DT>LR。两种特征在各项评估指标下的表现如下：

1、在检验结果的预测准确率（F1-score）方面：XGB在人工特征组比自动化特征组中得分整体要高出0.01左右。

2、分类的效果程度（AUC）方面：XGB在人工特征组比自动化特征组中得分整体要高出0.02左右。

3、预测模型损失值（Logloss）方面：XGB在人工特征组比自动化特征组中整体损失值要小0.07左右。

4、运行花费时间方面：因自动化生成的特征个数明显少于人工特征，所以三种模型运行时间都比人工特征组约少一半以上的时间。

### 特征重要性分析

**一、基于XGBoost的人工特征重要性排序**

特征重要性分析可以发现哪些特征变量对模型的表现效果影响更大，从第4.3.2节实验模型评估结果来看，XGBoost的预测效果确实是比较理想的。接下来，针对XGBoost模型训练的人工特征，根据特征重要性进行排序，如图4-9，绘制特征重要性条形图。

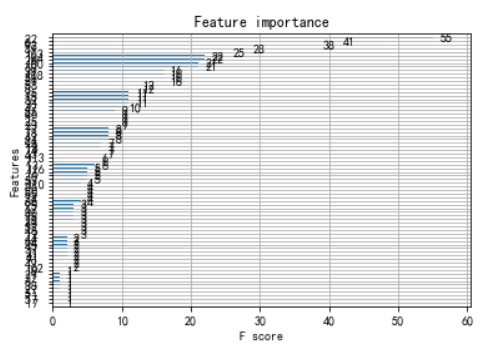


图4-9 基于XGBoost的人工特征重要性排序

从图4-9中看出预测模型的表现通常受重要性较强的特征影响较大，我们取出人工特征和其特征重要性得分进行观察，详见表4-7。

表4-7 基于XGBoost人工特征重要性得分和特征名称前5名示意

|  |  |  |
| --- | --- | --- |
| **序号** | **对照组1\_特征重要性得分** | **对照组1\_特征名称** |
| 1 | 0.14116 | 所有行为时间间隔均值 |
| 2 | 0.11396 | 最后一个行为时间间隔 |
| 3 | 0.08527 | 非支付行为占比 |
| 4 | 0.07661 | 点击购买确认(7)次数 |
| 5 | 0.06716 | 最近一次动作与最近一次浏览业务介绍页(5)\_时间间隔 |
| …… | …… | …… |
| **序号** | **对照组2\_特征重要性得分** | **对照组2\_特征名称** |
| 1 | 0.121842 | 用户基本特征-用户所在省份 |
| 2 | 0.120252 | 用户基本特征-用户所在省份 |
| 3 | 0.098908 | 用户基本特征-用户所在省份 |
| 4 | 0.090388 | 订购时间间隔方差 |
| 5 | 0.086431 | 业务使用区域个数 |
| …… | …… | …… |
| **序号** | **对照组3\_特征重要性得分** | **对照组3\_特征名称** |
| 1 | 0.11781 | 最后一个行为时间间隔 |
| 2 | 0.10945 | 所有行为时间间隔均值 |
| 3 | 0.08815 | 非支付行为占比 |
| 4 | 0.06942 | 最近一次动作与最近一次浏览业务介绍页(5)\_时间间隔 |
| 5 | 0.04815 | 点击购买确认(7)次数 |
| …… | …… | …… |
| **序号** | **对照组4\_特征重要性得分** | **对照组4\_特征名称** |
| 1 | 0.14116 | 所有行为时间间隔均值 |
| 2 | 0.11396 | 最后一个行为时间间隔 |
| 3 | 0.08527 | 非支付行为占比 |
| 4 | 0.07661 | 点击购买确认(7)次数 |
| 5 | 0.06716 | 最近一次动作与最近一次浏览业务介绍页(5)\_时间间隔 |
| …… | …… | …… |
| **序号** | **对照组5\_特征重要性得分** | **对照组5\_特征名称** |
| 1 | 0.14116 | 所有行为时间间隔均值 |
| 2 | 0.11396 | 最后一个行为时间间隔 |
| 3 | 0.08527 | 访问权重 |
| 4 | 0.07661 | 点击购买确认(7)次数 |
| 5 | 0.06716 | 最近一次动作与最近一次浏览业务介绍页(5)\_时间间隔 |
| …… | …… | …… |

根据人工特征重要性排序情况，分析人工对照组中依次增加的特征重要性得分的多少，以此判断特征的有效性和贡献度。

（1）人工对照组2（用户基本特征+订购特征）除了省份，在只有订购特征时，比较有意义的特征是第4名“订购时间间隔方差”，重要性得分占0.09分左右，这个特征从时间角度反映了用户订购集中程度，对于有过多次订购记录的用户，如果用户重复购买的时间间隔比较集中，那么他再次购买的可能性也比较高。但是如果用户没有订购过，那么预测用户是否购买的特征就只有省份了，所以从特征贡献角度，也说明了为什么在第4.3.2节模型效果评估时，对照组2的平均预测效果都是最差的。

（2）在人工对照组1（用户基本特征+访问特征），对照组3（用户基本特征+访问特征+订购特征），对照组4（用户基本特征+访问特征+订购特征+访问和订购组合特征）情况下，前三名都包含“所有行为时间间隔均值”、“最后一个行为时间间隔”、“非支付行为占比”属于访问行为类特征，比如看用户访问行为时间平均时间间隔，如果时间很短比较频繁，说明用户预期进行购买的可能性比较高。这3个访问行为类特征在各组贡献得分都比较高，而且对照组1、3、4在第4.3.2节模型效果评估时得分较高，预测效果也比较好。一方面说明了我们第3.3.2节中构造的时间特征、比率特征比较有意义，另一方面也说明了访问特征比订购特征确实对模型预测起到了更重要的贡献作用。

（3）人工对照组5（用户基本特征+访问特征+订购特征+访问和订购组合特征+季节特征），和对照组1、3、4相比，前三名除了访问行为类特征“所有行为时间间隔均值”、“最后一个行为时间间隔”外，新增的季节性特征“访问权重”排到了第3名，重要性得分占0.085分左右，一方面说明季节性因子对模型预测起到了重要贡献作用，另一方面说明特征维度最多的人工对照组5，各类特征都对模型起到了一定贡献意义。

接下来，根据计算得到人工特征排名和平均贡献度占比进行观察，详见表4-8。

表4-8 基于XGBoost人工特征排名和平均贡献度占比

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **特征排名** | **前3名** | **前5名** | **前10名** | **前15名** | **前20名** | **前25名** | **前30名** |
| 贡献度占比 | 30%以上 | 43%以上 | 63%以上 | 78%以上 | 84%以上 | 88%以上 | 90%以上 |

人工特征前三名特征的贡献度之和都在30%以上，以此类推，前30名特征贡献度占比已经达到了90%以上，从贡献度占比角度，也说明了人工选择的特征比较有效。

**二、基于XGBoost的自动特征重要性排序以及和人工特征重要性对比**

XGBoost模型训练的自动特征，根据特征重要性进行排序，如图4-10，绘制特征重要性条形图。

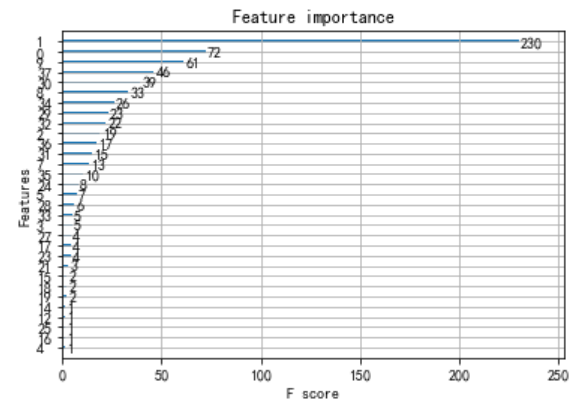


图4-10 基于XGBoost的自动特征重要性排序

在自动特征中，以及第4.3.2节评估效果最好的人工特征第5组中，分别抽取出重要性排前10名的特征进行对比观察，详见表4-9。

表4-9 基于XGBoost人工特征和自动特征重要性得分和特征名称前10名对比示意

| **序号** | **人工特征\_**  **重要性得分** | **人工特征\_特征名称** | **自动特征\_**  **重要性得分** | **自动特征\_特征名称** |
| --- | --- | --- | --- | --- |
| 1 | 0.14116 | 所有行为时间间隔均值 | 0.136197 | 访问最多的天 |
| 2 | 0.11396 | 最后一个行为时间间隔 | 0.105481 | 最多的行为名称 |
| 3 | 0.08527 | 访问权重 | 0.102868 | 所有行为总次数 |
| 4 | 0.07661 | 点击购买确认(7)次数 | 0.091675 | 访问最多的月份 |
| 5 | 0.06716 | 最近一次动作与最近一次浏览业务介绍页(5)\_时间间隔 | 0.080217 | 用户所在省份 |
| 6 | 0.04647 | 点击购买确认(7)占比 | 0.055903 | 所有业务订购量 |
| 7 | 0.04642 | 点击购买确认(7)\_订购占比 | 0.04729 | 访问最多的年份 |
| 8 | 0.04097 | 最近一次动作与最近一次浏览业务列表页(4)\_时间间隔 | 0.042678 | 不同年份的访问量 |
| 9 | 0.0334 | 是否订购TOP3区域 | 0.041944 | 不同年份的订购量 |
| 10 | 0.02932 | 最后一个行为是什么 | 0.033424 | 产品定的最少的区域 |
|  | …… | …… | …… | …… |

（1）自动特征中，排名前3的分别是“访问最多的天”、“最多的行为名称”、“所有行为总次数”，也都属于访问行为类特征，并且自动特征生成主要是基于统计分布的众数、聚合函数的求和等方式产生。

（2）针对人工和自动特征贡献前几名的特征进行对比观察发现，人工特征是基于挖掘业务特点而设计的，维度更多更全面。自动特征都是利用统计和聚合函数等方法自动批量构建的，虽然也具有可解释性，但是相对于人工特征稍显简单。

接下来，根据计算得到自动特征排名和平均贡献度占比进行观察，详见表4-10。

表4-10 基于XGBoost自动特征排名和平均贡献度占比

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **特征排名** | **3名** | **5名** | **10名** | **15名** | **20名** | **25名** | **30名** |
| 贡献度占比 | 33%以上 | 51%以上 | 69%以上 | 72%以上 | 91%以上 | 94%以上 | 95%以上 |

自动特征前三名特征的贡献度之和都在33%，并且前30名特征贡献度占比已经达到了95%以上，从贡献度占比角度，说明的自动特征构建的特征也比较有效。

### 实验结果与对比

1、三种算法模型选择方面

在准确率、F1分数、AUC值、Logloss值、训练时间等方面进行综合比较：无论是人工特征还是自动特征，平衡算法表现和运算速度，效果最优的都是XGBoost，它各项评分均比较高、损失值最小、模型运行时间比较短，其次是决策树，而对于逻辑回归各项指标都不太理想。所以最终选择XGBoost作为业务订购预测模型的分类器最合适。

2、特征的贡献度对比方面

（1）人工特征对照组内对比：

除了基础特征外，对照组1（用户基本特征+访问特征）只有访问特征，对照组2（用户基本特征+订购特征）只有订购特征。实验结果发现当只有订购信息时，预测准确性最差，因为分析业务特点时发现订购过的客户短时间内再次订购率较低，所以只有订购指标时模型的预测准确率不会很高。当只有访问信息时，模型预测准确率明显提升，用户行为特点决定了用户近期是否有购买需求，说明了访问信息属于更关键的信息。

对照组3、4、5中都已经包含了访问类特征，设置这几个对照组的目的，就是在访问特征基础上，分别增加订购、订购和访问组合特征和季节性因子等特征，来验证新增加特征对于模型的贡献度。从模型效果评估指标得分来看，各组都比只有访问特征时，预测准确性有微增。从特征重要性排序来看，新增加特征的贡献度也都占有一定比重，特别是第5组增加季节性特征后，它排在了第3名对模型贡献度比较显著。

从以上结果来看，每个对照组中增加的每一类型特征对模型都起到了一定贡献作用，验证了包含所有人工特征的照组5更具有完备性。

（2）人工特征和自动特征对比：

利用上述评估效果最好的第5组人工特征与自动特征进行比较：

人工特征数量为93个，前三个最重要特征分别是“所有行为时间间隔均值”、“最后一个行为时间间隔”、“访问权重”，说明了临近几天的访问行为与预测结果相关程度很高。自动化特征采用完全自动的方式生成37个特征，前三个最重要特征分别是“访问最多的天”、“最多的行为名称”、“所有行为总次数”，说明了具体的日期和最多的行为与预测结果相关程度较高。

从数量上对比，人工特征工程构造的特征更完备，自动特征的更简洁；特征类型上对比，两者生成的前三个重要特征均为行为、时间相关类型的特征；基于XGBoost算法预测效果对比，人工特征比自动特征准确率要高2%，损失值小7%。时间上人工特征运行约1分钟，自动特征运行约半分钟。

综上所述，人工特征对模型预测结果具有更强的解释性和可调节性。可解释性让我们知道了输出预测结果的主要原因，以及特征的重要程度，并且特征更有意义，当出现问题的时候，我们可以进行原因的分析。可调节性是指当出问题或删除特征的时候，可以随时对模型进行修正来满足新需求。对比来讲自动特征虽然构建时间更快，但准确率差一点，自动化特征工程生成的特征较为简单，并且“访问最多天数”作为最核心特征，对于未知数据集来说容易产生过拟合的风险。

# 结论与展望

## 全文总结

本文从实际运营业务的需求角度出发，重点对比和讨论了业务的人工和自动特征提取方法，两种方法下订购预测模型准确性和可用性，以及针对业务特征贡献值来判断特征提取策略的有效性等问题。

现对本文总结如下：

（1）本文针对业务特点，在人工特征的贡献程度方面，先用可视化的方法发现数据集中历史数据规律和特性，并假设了5个对照组，依次增加要验证的特征，通过实验证明人工设计的特征比较完备，在预测模型上得到了有较高的特征重要性得分。又因为业务产品具有中低频购买的特点，相对于高频购买的产品来讲，购买周期相对较长，并且实验表明用户购买季节性会更加显著，中低频用户有更倾向于在某些季节内消费的特点。可见人工特征有较强的解释性和可调节性，可以对特征指标贡献程度高的指标进行利用，完善营销策略以符合客户的期望，也为运营和管理工作提供了重要的参考。

（2）采用了人工和自动特征工程对比分析的方式，利用特征衍生法构建了适用于预测模型的人工特征矩阵，利用深度特征合成DFS把数据集根据关联关系聚合为自动生成的新特征，通过实验两种方法均达到了良好的预测效果。对比来说，自动特征更节约时间，也具有一定的解释性，但生成的新特征比较简单；人工特征具有更强的解释性和可调节性，可以根据结果随时来修正模型，花费时间通常较长，但是对实际工作更有意义。

（3）在解决数据集的样本不平衡问题上，通过增加或缩短预测周期的长度发现效果不理想，实验中用SMOTE算法进行上采样后达到了理想的预测效果。

（4）为了使业务预测效果更好，采用了三种机器学习算法进行对比试验，预测效果最优的是XGBoost，其次是决策树，而对于逻辑回归方法AUC评分比较低，并且模型运行时间较长约2~3个小时，所以验证了基于增强树模型的XGBoost比较适合作为业务订购预测模型的分类器。

## 研究局限与展望

本文通过实验发现了历史业务数据的价值，并为后面的运营工作提供了订购预测的支持。目前各行业运用数据挖掘技术提升企业运营效率的例子越来越多了，许多商业价值都蕴含在数据之中，如果进行很好的挖掘与使用，意义将是巨大的，希望这些有潜在价值的数据未来可以为运营工作提供有效的支撑。

本次实验还存在一些局限和改进空间如下：

（1）对于分类算法，特征的选择是非常重要的，因为特征工程是DM/ML基石，它决定了预测效果能达到的高度。由于客观条件限制，本次实验利用的特征范围比较少，比如缺少推广引流量等特征，希望未来可以融入业务中更多有价值的指标和数据来做特征，并且随着特征的增多尝试学习利用特征过滤算法，让模型的准确性和效率得到进一步提升。

（2）本文使用人工特征工程，基于业务知识挖掘了一些周期性、趋势性的复杂性特征。并且通过人工和自动化生成的特征进行对比发现，自动化特征虽然效率比较高，但是训练时过分突出了访问最多的天数作为最关键特征，因为每年主要的节假日（清明、五一等）不在同一个工作日，容易出现过拟合现象，以后可以尝试利用人工发现有意义特征，来自定义自动化特征将方法进行结合，来提升模型的准确性和提高特征工程效率。

（3）本文曾尝试用多层感知机（MLP）的方法来进行模型预测，但效果非常不理想，所以没有作为本文预测方法，随着新技术发展，会继续学习发现更好的预测模型，如自动化算法或深度学习的神经网络（如CNN、RNN、DNN）等方法来提高自动化程度以及预测准确度，从而在控制运营成本的同时，提升运营效率。

（4）业务订购类数据很容易出现样本不平衡现象，针对本文不平衡问题XGBoost算法虽然有着不错的预测效果，但后续会遇到各种复杂类型的数据，要根据实际业务做对比实验来寻找更优的解决方案。

# 参考文献

1. [] 王先庆, 雷韶辉. 新零售环境下人工智能对消费及购物体验的影响研究——基于商业零售变革和人货场体系重构视角[J]. 商业经济研究, 2018(17): 5-8. [↑](#endnote-ref-0)
2. [] 赵巧. 网络环境下消费者购买决策行为影响因素分析[J]. 现代商贸工业, 2016(7): 66-68. [↑](#endnote-ref-1)
3. [] 杨萍. 构建数据化运营体系很重要[N]. 中国计算机报, 2015-08-10(023). [↑](#endnote-ref-2)
4. [] 杜刚, 黄震宇. 大数据环境下客户购买行为预测[J]. 管理现代化, 2015, 35(1): 40-42. [↑](#endnote-ref-3)
5. [] Chen, Y., Yan, X., Fan, W., et al. The joint moderating role of trust propensity and gender on consumers’ online shopping behavior[J]. Computers in Human Behavior, 2015, 43: 272-283. [↑](#endnote-ref-4)
6. [] 黄媛. B2C网络消费者再购买影响因素研究[D]. 华南理工大学, 2017. [↑](#endnote-ref-5)
7. [] 王曼琳. 基于消费者情绪的在线购买意愿影响因素研究[D]. 武汉理工大学, 2017. [↑](#endnote-ref-6)
8. [] 喻昕. 在线商户商品信息呈现对消费者购买意愿影响的研究[D]. 吉林大学, 2017. [↑](#endnote-ref-7)
9. [] 刘遗志, 汤定娜. 感知价值对消费者移动购物意愿的影响研究——基于TAM和VAM理论模型[J]. 兰州学刊, 2015(4): 169-175. [↑](#endnote-ref-8)
10. [] 都在说用户画像, 你真的了解透了吗?. https://cloud.tencent.com/developer/article/1184424. 最后访问于2019-12-25. [↑](#endnote-ref-9)
11. [] 周杰, 邓馨洁, 付佳乐, 等. 网购过程中服务接触对消费者购买行为的影响因素分析[J]. 经济师, 2017(4): 53-54. [↑](#endnote-ref-10)
12. [] 丁一楠. 基于深度学习的消费者行为分析及应用研究[D]. 重庆工商大学, 2018. [↑](#endnote-ref-11)
13. [] Wu, R.S., Chou, P.H.. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach[J]. Electronic Commerce Research and Applications, 2011, 10(3): 331-341. [↑](#endnote-ref-12)
14. [] 卢军, 张天凡. 基于隐式用户行为的推荐系统研究[J]. 湖北工程学院学报, 2016(3). [↑](#endnote-ref-13)
15. [] Grishman, R.. Information Extraction: Capabilitis and Challenges[Z]. Notes Prepared for the 2012 International Winter School in Language and Speech Technologies, Rovira i Virgili University, Tarragona, Spain, 2012. [↑](#endnote-ref-14)
16. [] 王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7): 66-76. [↑](#endnote-ref-15)
17. [] Bucklin, R.E., Sismeiro, C.. Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing[J]. Journal of Interactive Marketing, 2009, 23(1): 35-48. [↑](#endnote-ref-16)
18. [] Rendle, S.. Factorization Machines[C]. IEEE International Conference on Data Mining. IEEE, 2011. [↑](#endnote-ref-17)
19. [] Harakawa, R., Takehara, D., Ogawa, T., et al. Sentiment-aware personalized tweet recommendation through multimodal FFM[J]. Multimedia Tools and Applications, 2018, 77(2): 1-19. [↑](#endnote-ref-18)
20. [] Backiel, A., Baesens, B., Claeskens, G.. Predicting time-to-churn of prepaid mobile telephone customers using social network analysis[J]. Journal of the Operational Research Society, 2016, 67(9). [↑](#endnote-ref-19)
21. [] Park, C.H., Park, Y.H.. Investigating purchase conversion by uncovering online visit patterns[J]. Marketing Science, 2016, 35(6): 894-914. [↑](#endnote-ref-20)
22. [] Silahtaroglu, G., Donertasli, H.. 2015 IEEE International Conference on Big Data (Big Data) - Analysis and prediction of Ε-customers’ behavior by mining clickstream data[C]. IEEE International Conference on Big Data. IEEE, 2015: 1466-1472. [↑](#endnote-ref-21)
23. [] 马姝. 基于数据挖掘的消费者购买预测的研究[D]. 云南财经大学, 2016. [↑](#endnote-ref-22)
24. [] 马倩. 基于机器学习的电子商务平台重复购买客户预测[D]. 兰州大学, 2017. [↑](#endnote-ref-23)
25. [] 刘潇蔓. 基于特征选择和模型融合的网络购买行为预测研究[D]. 北京交通大学, 2017. [↑](#endnote-ref-24)
26. [] 张子实. 电子商务平台基于用户行为数据的消费预测研究[D]. 北京邮电大学, 2018. [↑](#endnote-ref-25)
27. [] 李国凤. 电商平台移动端网络购买预测研究[D]. 暨南大学, 2018. [↑](#endnote-ref-26)
28. [] Kira, K., Rendell, L.A.. A practical approach to feature selection[J]. Machine Learning Proceedings, 1992, 48(1): 249-256. [↑](#endnote-ref-27)
29. [] 李海丽. 基于数据挖掘的电子商务客户流失分析[D]. 对外经济贸易大学, 2018. [↑](#endnote-ref-28)
30. [] 李梦. 考虑商品重复购买周期的推荐方法研究[D]. 武汉科技大学, 2018. [↑](#endnote-ref-29)
31. [] 艾金金. 电商平台客户流失预警分析及应用研究[D]. 南京大学, 2019. [↑](#endnote-ref-30)
32. [] 张宁, 范崇睿, 张岩. 一种基于RFM模型的新型协同过滤个性化推荐算法[J]. 电信科学, 2015, 31(9): 103-111. [↑](#endnote-ref-31)
33. [] de Brébisson, Alexandre, Simon, et al. Artificial Neural Networks Applied to Taxi Destination Prediction[J]. Computer Science, 2015. [↑](#endnote-ref-32)
34. [] Guo, C., Berkhahn, F.. Entity Embeddings of Categorical Variables[J]. 2016. [↑](#endnote-ref-33)
35. [] Jin, H., Song, Q., Hu, X.. Auto-Keras: Efficient Neural Architecture Search with Network Morphism[J]. 2018. [↑](#endnote-ref-34)
36. [] Bergstra, J., Bengio, Y.. Random Search for Hyper-Parameter Optimization[J]. Journal of Machine Learning Research, 2012, 13(1): 281-305. [↑](#endnote-ref-35)
37. [] Kanter, J.M., Veeramachaneni, K.. Deep feature synthesis: Towards automating data science endeavors[C]. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2015. [↑](#endnote-ref-36)
38. [] 芦思雨. 数据挖掘中分类算法的比较分析[D]. 天津财经大学, 2016. [↑](#endnote-ref-37)
39. [] Rosebrock, A.. (2019.01.07). Auto-Keras and AutoML: A Getting Started Guide.https://www.pyimagesearch.com/2019/01/07/auto-keras-and-automl-a-getting-started-guide/. 最后访问于2019-12-25. [↑](#endnote-ref-38)
40. [] Chapelle, O., Manavoglu, E., Rosales, R.. Simple and Scalable Response Prediction for Display Advertising[J]. ACM Transactions on Intelligent Systems and Technology, 2014, 5(4): 1-34. [↑](#endnote-ref-39)
41. [] 周志华. 机器学习[M]. 北京：清华大学出版社, 2016: 57. [↑](#endnote-ref-40)
42. [] 蔡萌萌, 张巍巍, 王泓霖. 大数据时代的数据挖掘综述[J]. 价值工程, 2019, 38(05): 155-157. [↑](#endnote-ref-41)
43. [] Breiman, L., Friedman, J., Stone, C.. Classification and Regression Trees. Wadsworth, 1984 [↑](#endnote-ref-42)
44. [] 李航. 统计学习方法[M]. 北京：清华大学出版社, 2012: 74. [↑](#endnote-ref-43)
45. [] Chen, T., Guestrin, C.. Xgboost: A scalable tree boosting system[C]. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794. [↑](#endnote-ref-44)
46. [] Friedman, J., Hastie, T., Tibshirani, R.. Additive logistic regression: a statistical view of boosting(with discussions).Annals of Statistics, 2000, 28: 337–407 [↑](#endnote-ref-45)
47. [] 吴锦华, 王志生, 刘重阳, 等. 特征选择方法在信用评分系统中的应用[J]. 信息与电脑(理论版), 2019, 426(08): 125-126. [↑](#endnote-ref-46)
48. [] Ciaburro, G., Venkateswaran, B.. Receiver Operating Characteristic curve. https://www.oreilly.com/library/view/neural-networks-with/9781788397872/bf8d6e44-8ea1-4697-8268-995bac7867bd.xhtml. 最后访问于2019-12-25. [↑](#endnote-ref-47)
49. [] 汪云云, 陈松灿. 基于AUC的分类器评价和设计综述[J]. 模式识别与人工智能, 2011, 24(01): 64-71. [↑](#endnote-ref-48)
50. [] Zhu, D., Cui, Y.. Understanding random guessing line in ROC curve[C].2017 2nd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2017. [↑](#endnote-ref-49)
51. [] Bradley, P.. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms[J]. Pattern Recognition, 1996, 30: 1145-1159. [↑](#endnote-ref-50)
52. [] Fawcett, T.. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861-874. [↑](#endnote-ref-51)
53. [] 对数损失函数(Logarithmic Loss Function)的原理和Python 实现. https://www.cnblogs.com/klchang/p/9217551.html. 最后访问于2019-12-25. [↑](#endnote-ref-52)
54. [] Making Sense of Logarithmic Loss. https://www.r-bloggers.com/making-sense-of-logarithmic-loss/. 最后访问于2019-12-25. [↑](#endnote-ref-53)
55. [] 特征工程二：特征抽象、特征衍生. https://www.cnblogs.com/yyy-blog/p/10420175.html. 最后访问于2019-12-25. [↑](#endnote-ref-54)
56. [] Domingos, P.. A few useful things to know about machine learning[J]. Communications of the ACM, 2012, 55(10): 78. [↑](#endnote-ref-55)
57. [] 赵婷婷. 基于算法融合的客户流失预测方法研究[D]. 东北财经大学, 2018. [↑](#endnote-ref-56)
58. [] 向鸿鑫, 杨云. 不平衡数据挖掘方法综述[J]. 计算机工程与应用, 2019, 55(04): 1-16. [↑](#endnote-ref-57)
59. [] 石拓, 魏新蕾, 邵旭芬. 基于ID3—SMOTE结合算法的社会群体性事件预警模型[J]. 中国传媒大学学报(自然科学版), 2017, 24(06): 9-15. [↑](#endnote-ref-58)
60. [] 机器学习常见算法时间复杂度. https://blog.csdn.net/WASEFADG/article/details/88981492. 最后访问于2019-12-25. [↑](#endnote-ref-59)
61. [] 刘亚. 基于海量数据分析的银行担保圈风险识别方法的研究[D]. 内蒙古大学, 2019. [↑](#endnote-ref-60)
62. [] 王凯. 基于图计算的用户行为建模关键技术研究[D]. 南京邮电大学, 2018.

    # 致 谢

    三年的求学生活暂告段落，回想整个过程虽然不易，却让我摒弃浮躁，汲取了知识的营养、锻炼了意志，也给了我很多思考和启迪。本论文的顺利完成，首先，感谢我的导师孙彩虹副教授，孙老师悉心指导我进行文章的选题、研究思路的探索、以及遇到问题如何寻找突破点，比如如何用正确的思路找到方法解决正负样本不均衡问题，如何进行特征方法的选择和模型的平衡等等。让我从入学时对机器学习可以说不太了解的人，到现在能够根据研究问题提出假设，并通过不断尝试验证得出结论，学会了一套探索和研究问题的思考方法。通过这次论文撰写的机会，能够尝试做出业务预测模型，对自己来说是一个很大的挑战，既开阔了眼界也丰富了实践，如果没有孙老师的谆谆教诲，可能这些都不会实现。孙老师学识渊博，治学态度严谨，对学生平易近人，对我产生了很大的影响，在此向您表示衷心的感谢！

    同时，我要对参与评审的专家和教授们致以诚挚的敬意，感谢各位老师百忙之中提出的宝贵建议，让我的论文更加严谨。

    然后，还要感谢在我论文写作过程中给予过我帮助的良师益友和同学们，谢谢大家的相互帮助和支持。

    最后，在这里再次感谢老师、家人、同学和朋友们在我学习过程中对我无私的关怀和鼓励，致以深深的谢意！ [↑](#endnote-ref-61)