# 朴素贝叶斯模型在天猫双十一复购预测中的应用

柯金宏 21213131

**摘要** 商家通过举办促销活动吸引新买家之后，需要确定哪些新买家可以转化为重复购买者，并对其进行精细化营销。为了对新买家在未来一段时间内再次购买该商品的概率进行预测，本文根据用户行为日志进行特征工程，构建了用户、商家、品牌、商品类别和商品画像，并运用感知机、朴素贝叶斯、决策树、Logistic 和 SVM 模型进行训练和预测，用进行模型融合。结果表明朴素贝叶斯对测试集的预测结果 AUC 最高，模型拟合时间较短，在比赛官网的成绩为 0.63。本研究确定了复购预测的重要变量，模型以及解决方案，对提升商家进行精细化营销，降低促销成本，提升促销的投资回报率具有实际意义。

**关键词** 复购，电商，用户行为日志

双十一是电商推出促销活动，吸引消费者的一个促销节日。众多商家在这一天迎来巨额的销量，但只有很少一部分消费者会在促销节日之后再次购买该商家的产品，给商家带来持续的收益。如何识别消费者会在节日过后再次购买哪一家商家的产品成为商家持续获益的关键，如果可以准确的预测消费者的复购行为，那么可以对消费者进行精准营销，减少商家的宣传开支。用户的复购行为也反映了用户忠诚度，忠诚用户可能会向亲朋好友推荐商品，因此维持忠诚用户可以给商家增加新的客源。除此之外，电商平台也可以根据复购用户的预测结果改进推荐算法，降低用户购物时的信息检索困难程度，带给用户更好的体验，提升平台的竞争力。因此，对复购行为的预测可以实现用户、商家和电商平台共同获益。

如何从双十一节日购买用户中区分出会再次购买的用户，是一个分类问题。商家可以挖掘历史的用户行为日志数据，得到用户是否再次购买的信息。并且通过用户浏览、购买、收藏商品的一系列行为构建用户画像特征，以及商品和品牌的销量特征评估商品和品牌对用户的吸引力，运用有监督学习的机器学习方法，对用户是否复购进行预测。下面介绍复购行为的影响因素以及复购行为预测两方面的研究，并提出本文研究思路和创新点。

# 1 文献综述

## 1.1 用户复购行为影响因素

用户复购行为的研究设计管理学、心理学、经济学等多学科知识，可以从用户自身感受、商家服务、商品种类、商品品牌等方面进行研究，有如下理论对用户复购行为进行分析：

感知服务质量理论认为，商家提供服务的好坏会改变商品在消费者心中的质量，进而影响消费者的复购行为[1]。如果消费者与网站的互动过程中出现问题，或者商家提供的售后服务不周到，便会体验到再次购买可能存在的风险，进而影响用户的再次购买行为。而商家的服务质量可以从商家网页的被访问次数、商家所有商品的整体销量中反映。感知交易成本理论的研究发现，信息搜索成本对复购意愿成正相关[2]。

感知价值理论认为，只有商品给消费者带来有价值的感受，消费者才会再次购买。价值可能有多个维度，比如可以将感知价值分为功利和享乐两种类型的价值[3]。前者与商品类型有关，实用的日常消耗品可以给消费者带来功利性价值。享乐性价值与情感相关，如果商品的品牌和用户之间存在情感关系，可以给用户带来尊贵、或是亲切的感受，则用户更可能再次购买该品牌的商品。

消费者主观决策对复购行为占主要影响。消费者根据自身需求和偏好，对商品复购做出行为决策，过程一般经过需求识别、信息检索、选择评估、购买决定、购后评价五个阶段[4]。需求识别、信息检索、选择评估阶段主要反映在消费者对商品页面的点击次数上，而消费者的购后评价反映在给出好评、将商品加入收藏或购物车以及购后点击商品的次数上。

## 1.2 用户复购行为预测研究

传统的购买行为预测模型主要有 SMC 模型、BG/NBD 模型和 RFM 模型，前两个模型通过验证历史的购买率、流失率等指标符合某个分布的模型假设，得到复购的分布和期望(基于 SMC 模型的航空公司常旅客活跃度分析)，后一个模型根据最近购买时间、购买的频率和购买金额对用户进行划分，识别重要价值客户，进而预测购买行为。

在国内，有研究基于决策树进行算法改进，构建新的用户购买行为预测模型[5]。有研究以前 26 周的用户消费行为数据预测第 27 周的用户购买率[6]。另外，

有研究以某时间段内消费者和商品的交互行为为分析对象，提出基于时序的数据预处理方法和基于 SSP 算法的特征选择方法，构建出基于 Bagging 模型融合策略的 Xgboost 混合模型，提升了模型的预测精度[7]。

国外首先有研究比较 SVM 预测模型和 Logistic 回归模型的预测准确率，评估模型的优劣势[8]。有研究通过 SVM 模型处理消费者在超市中消费所产生的日志数据，发现比线性回归等预测模型相比，SVM 模型显著提高预测消费者购买行为的准确率[9]。有研究使用了多种机器学习模型，对天猫“双十一”获取的新客户数据进行分析，构建了 100 余个特征，得到了相对较好的预测结果[10]。有研究进行聚类分析，通过忠诚客户的个人信息数据，定位潜在客户属性，通过关联规则分析检测客户对热销商品种类的兴趣[11]。有研究利用商品之间的关联建立商品集，然后分析用户对商品特征的偏好通过模型训练进行学习，得到未来用户最可能购买的前 n 个商品集[12]。

### 1.3 本文主要研究内容与创新点

由于数据量较大（有 5 千万条行为数据），本研究首先将数据分批导入 MySQL 数据库，然后进行数据清洗和简单的分析。然后，本研究根据感知服务价值、感知价值、感知交易成本和主管决策等多个理论，从用户行为日志数据中构建相应的买家特征、卖家特征、买家-卖家关系特征。接着利用比较 5 种经典的有监督分类算法在该数据集上的预测效果，并进行模型融合。最后对研究如何投入实际应用进行解释。

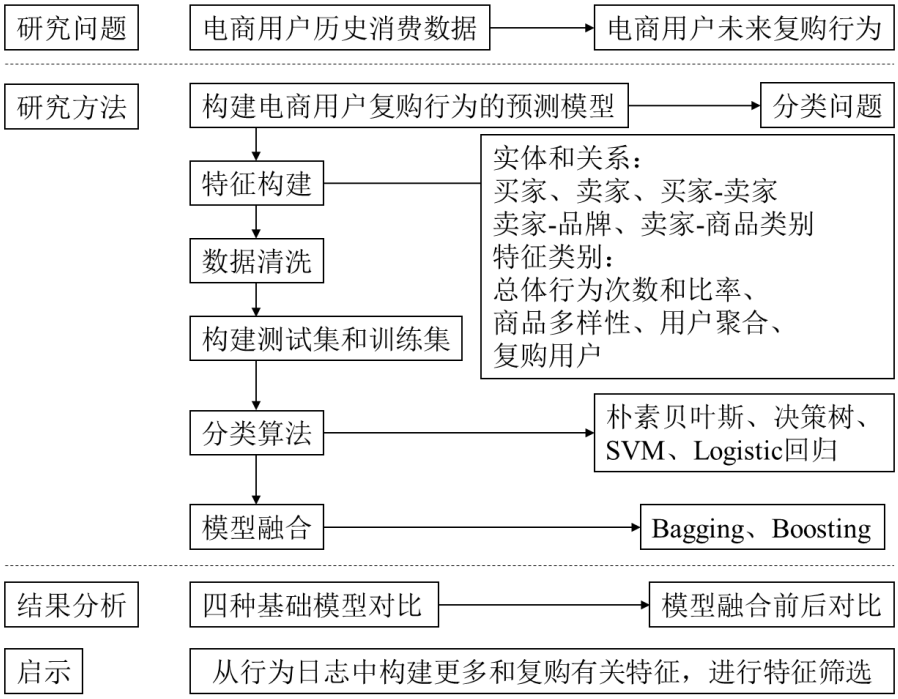



图 1 本研究的逻辑框架

## 2 研究设计

### 2.1 问题描述

研究问题是根据电商用户过去 5 月至双 11 的网上购物消费的数据，预测双 11 过后的复购行为。复购行为指的是在双十一当天购买了某个商家的商品，并在双 11 之后近一段时间内的某天再次购买该商家商品的行为。忠诚用户是持续关注某品牌商品的用户，有较高的购物需求，会长期、重复购买某品牌的商品。潜在用户是忠诚用户的候选人，虽然以往没有强烈购买需求，但有较大概率会重复购买。电商平台的任务是维系忠诚用户，挖掘并转化潜在用户。

该问题就是区分所有用户中哪些用户是可能复购的用户，是机器学习中的经典二分类问题，根据复购和未复购作为标签，划分数据集中的正负样本。在保证训练集、测试集和验证集用户的行为特征相同的基础上，在训练集数据上进行模型训练，在测试集上检测模型效果，最后对验证集用户进行分类预测。

### 2.2 数据获取

研究使用 2014 年天猫双 11 购物活动的销售数据，该数据集最初于 2015 年国际人工智能联合会议中的一场重复购买预测竞赛中发布[13]，现可在天池数据库中下载[14]。由于数据量较大，先将数据分批导入 MySQL 数据库中。

这份数据由 4 个基本表组成，分别为训练表，测试表，用户画像表和用户

行为日志表。每个表格包括的字段如表 1，其中训练表有标签，可以用于划分训练集和测试集，测试表用于预测及提交比赛结果。

表 1 各基本表的字段

| 基本表名 | 字段 |
| --- | --- |
| 训练表 | 买家 id，卖家 id，是否复购 |
| 测试表 | 需要预测的买家 id，卖家 id |
| 用户画像表 | 买家 id，性别、年龄 |
| 用户行为日志表 | 买家 id，商品 id，商品类别 id，卖家 id，销售日期（月日），<br>行为（0-点击、1-加入购物车、2-购买、3 收藏） |

训练表和测试表的描述统计如表 2 所示，训练表和测试表卖家数量相同，在训练表中，复购的成对买家-卖家数量比例仅占 6.12%。

表 2 训练集和测试集

| | 训练表 | 预测表 |
| --- | --- | --- |
| 买家数量 | 212062 | 212108 |
| 卖家数量 | 1993 | 1993 |
| 成对买家-卖家的数量 | 260864 | 261477 |
| 复购对数 | 15952 | 未知 |
| 复购率 | 6.12% | 未知 |

用户画像表中年龄和性别分布情况如图 2 所示，原始数据已对年龄进行了分箱，有 22.43%的买家年龄未知，在已知年龄的买家中，大部分买家集中于 25 至 34 岁。对于性别，女性占 68.86%，男性占 28.68%，性别的缺失值仅占 2.46%，且未知性别用户的年龄分布情况和图 2 类似，因此图 2 中未显示缺失性别群体的年龄分布情况。对于缺失的性别，填充为女性，缺失的年龄信息，采用均值填充。

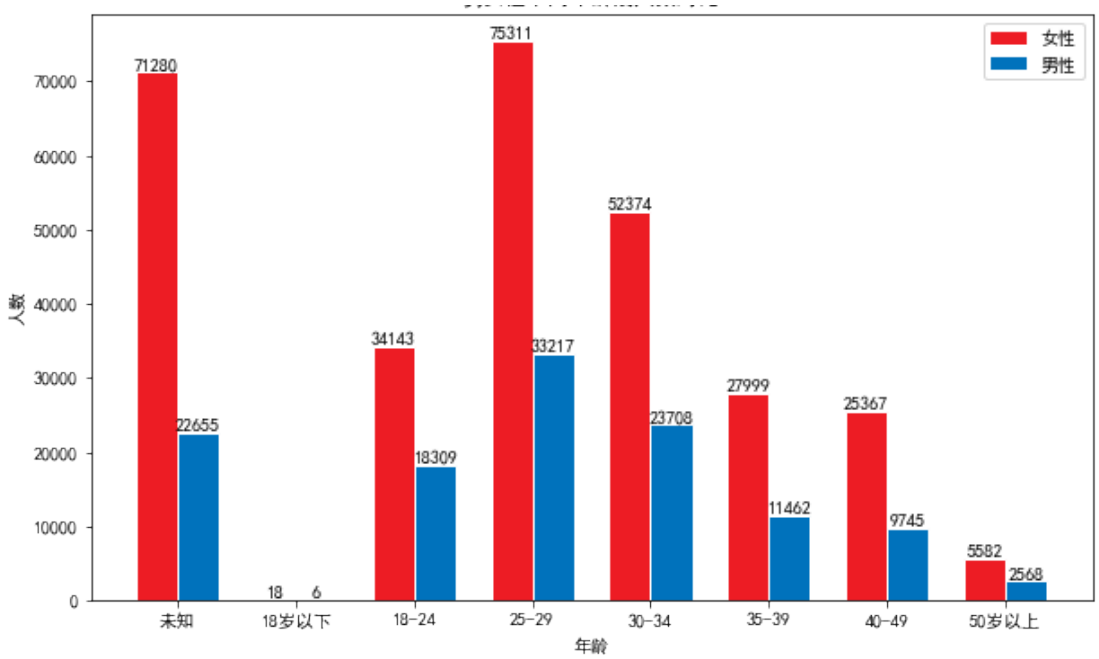



图 2 买家年龄和性别分布

用户行为日志表中包含最多的信息，除了用户行为这个字段之外，其他字段均为 id。表 3 汇总了各 id 的去除重复后的数量，可以看出商品 id 数量是最多的，在这个数据集中，卖家和商品 id 之间是一对多的关系，卖家之间没有公共的商品 id，同一个型号的商品即使在不同卖家手里也会被编码为不同商品 id。除了这一对关系之外，其他的实体间关系均为多对多的关系，比如同一件商品可能被标注为不同品牌、不同商品类别，同一品牌下可能有多个商品类别，多件商品。

表 3 用户行为日志中的各种 id 去除重复后的数量

| | 数量 |
|---|---|
| 总记录数 | 54925330 |
| 买家数量 | 424170 |
| 卖家数量 | 4995 |
| 商品 id 数量 | 1090390 |
| 商品类别数量 | 1658 |
| 品牌数量 | 8443 |

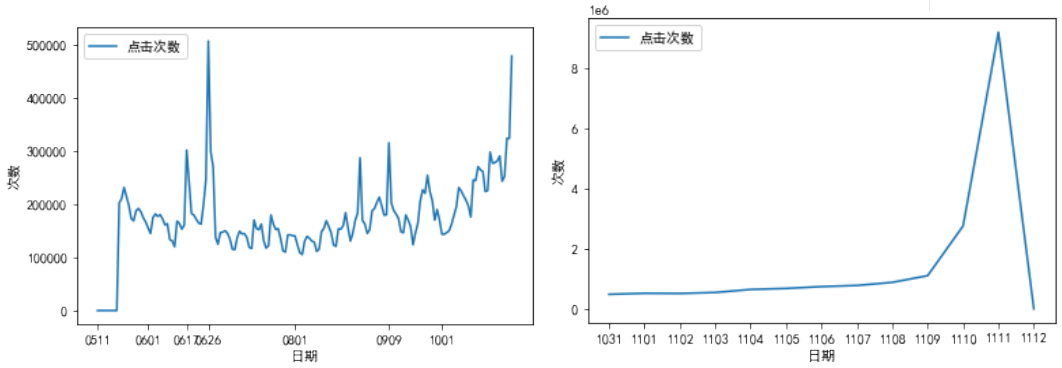
用户行为次数及占比如表 4 所示，可以看出大部分行为都是点击，极少数行为是加入购物车(0.14%)，因此后续特征构建将加入购入车和点击合并为一类，统

称为点击，不再区分加入购物车和点击的区别。

表 4 用户行为占比

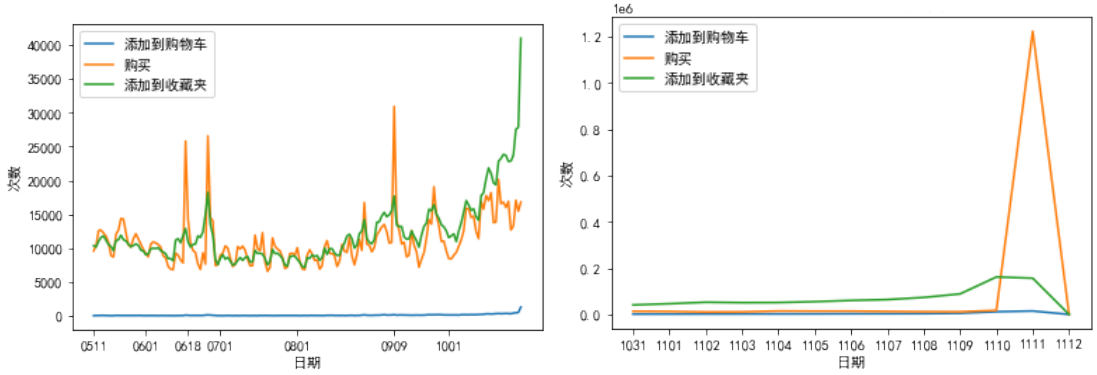
| 用户行为 | 次数 | 占比 |
|---|---|---|
| 点击 | 48550713 | 88.39% |
| 加购物车 | 76750 | 0.14% |
| 购买 | 3292144 | 5.99% |
| 收藏 | 3005723 | 5.47% |

下面分析用户行为的时间序列。点击次数序列在 6 月 26 日、9 月 9 日达到小高峰，在 11 月 11 日前 15 天不断攀升，至 11 日达到最高峰。购买次数的趋势和点击次数相近（图 4）。在 11 月，收藏次数和点击次数开始上升日期先于购买次数。图 4 中显示添加购物车次数极少，因此后续分析将其归为点击次数对预测结果应该没有太大影响。




(a) 5 月至 10 月　　(b) 11 月

图 3 不同用户点击次数的时间序列



(a) 5 月至 10 月　　(b) 11 月

图 4 不同用户加入购物车、购买和收藏次数的时间序列

### 2.3 特征构建

参考以往研究中所构建的重要特征[10]，根据用户日志表可以构建如表 5 的特征。其中，总体行为次数和比率是买家和卖家之间的活动的次数以及某种活动的次数占比，次数越多，购买比率越大买家越可能买该卖家的商品。商品多样性从商品的角度看待复购，如果一个买家从卖家那里观看、购买或者收藏了多样商品，那么该买家更可能对该卖家进行复购。用户聚合特征反映某商家、品牌和商品类别的用户历史购买习惯，如果该用户历史上购买天数较少，那么他可能只在双十一当天购买，是一次性买家，如果该用户历史上购买天数较多，则该用户可能为长期忠实客户。复购用户特征反映卖家、品牌和商品类别的历史复购情况，如果历史上买家购买次数越多，则买家越可能复购该卖家的、该品牌的或商品类别的商品。先计算出表格中的特征，然后在用户日志表中找出双 11 当天买家购买卖家商品 id，及对应的品牌 id、商品种类 id，如果购买多件商品，选择最小的一个商品 id，然后根据 id 将特征和买家-卖家相连接。

表 5 根据用户日志表构建的特征

| 特征类型 | 特征名称 | 含义 |
|---|---|---|
| 总体行为次数 | um_total_click_action_num | 某个买家点击某个卖家的链接的总次数 |
| | um_total_buy_action_num | 某个买家购买某个卖家的商品的总次数 |
| | um_total_favor_action_num | 某个买家收藏某个卖家的商品的总次数 |
| 总体行为比率 | um_total_click_action_ratio | 某个买家点击某个卖家的链接次数占该买家对该卖家所有行为次数之比 |
| | um_total_buy_action_ratio | 某个买家购买某个卖家的商品次数占该买家对该卖家所有行为次数之比 |
| | um_total_favor_action_ratio | 某个买家收藏某个卖家的商品次数占该买家对该卖家所有行为次数之比 |
| 商品多样性 | um_click_item_num | 某个买家点击某个卖家的不同商品个数 |
| | um_buy_item_num | 某个买家购买某个卖家的不同商品个数 |
| | um_favor_item_num | 某个买家收藏某个卖家的不同商品个数 |
| 用户聚合 | mb_user_buy_day_num_avg | 某对卖家-品牌下，所有用户购买天数的平均数 |
| | mb_user_buy_day_num_std | 某对卖家-品牌下，所有用户购买天数的标准差 |
| | mc_user_buy_day_num_avg | 某对卖家-商品类别下，所有用户购买天数的平均数 |

| | | |
|---|---|---|
| | mc_user_buy_day_num_std | 某对卖家-商品类别下，所有用户购买天数的标准差 |
| | m_user_buy_day_num_avg | 对于某个卖家，所有用户购买天数的平均数 |
| | m_user_buy_day_num_std | 对于某个卖家，所有用户购买天数的标准差 |
| 复购用户 | mb_repeat_buy_day_ratio | 某对卖家-品牌下，所有用户复购天数除以总购买天数 |
| | mc_repeat_buy_day_ratio | 某对卖家-商品类别下，所有用户复购天数除以总购买天数 |

## 2.4 数据预处理

对于缺失值，经检查性别、年龄和与品牌相关的变量存在缺失值。由于年龄经过分箱处理，属于类别变量，因此性别和年龄分别按众数填充，缺失的性别填充为女性，年龄填充为 25-29 岁的区间。与品牌相关的变量属于连续变量，按均值进行填充。

为解决不同变量间量纲大小不统一的问题，对年龄以及其他连续变量进行了最大最小归一化。

各特征的相关分析如图 5 所示，可以发现各指标和标签（是否复购）的相关性较低。其中和标签相关最高的变量是 mb_user_buy_day_num_avg 某对卖家-品牌下，所有用户购买天数的平均数，相关系数为仅为 0.11。用户聚合特征之间可能存在共线性。

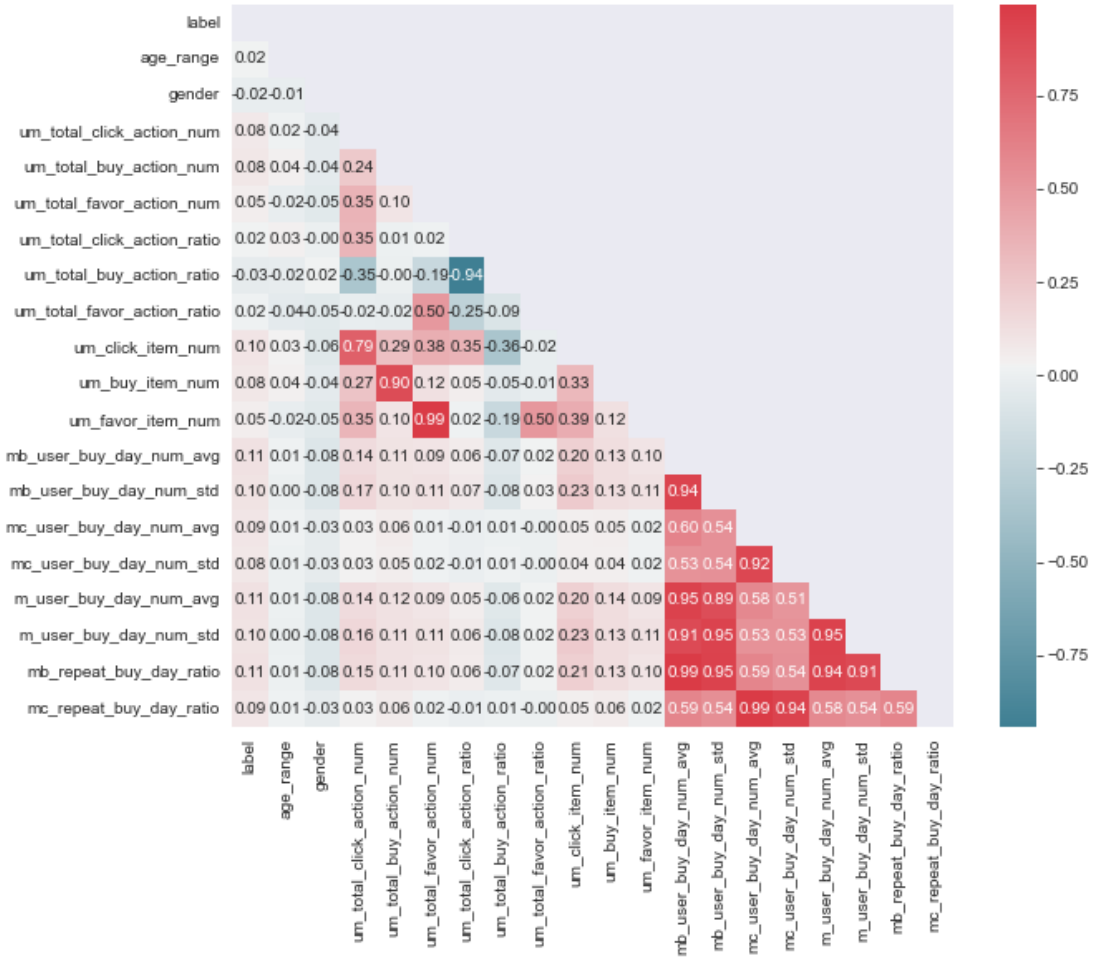



图 5 特征之间的 Person 相关系数矩阵

# 3 用户复购行为预测模型构建

## 3.1 前期准备工作

采用荣耀笔记本电脑，处理器为 AMD Ryzen 5 3500U，内存 8GB。安装 Jupyter Lab，Python，sikit-learn 软件包。模型评估指标采用正确率和 AUC (即 ROC 曲线下面积)作为模型评估指标。对有标签的数据按 8：2 的比例划分为训练集和测试集。

## 3.2 单个模型的预测效果对比

首先构建了朴素贝叶斯，Logistic 回归模型，决策树，SVM 这 4 种分类器进行训练和预测。其中，由于训练样本种复购人数很少，因此朴素贝叶斯采用 ComplementNB，应对正负性样本不平衡的问题。对于决策树模型，对复购类别进行赋权，复购的 class_weight＝0.99，非复购的 class_weight＝0.01，经过遍历，

发现树的深度在 29 的时候对测试集的预测效果最好，如图 6 所示。

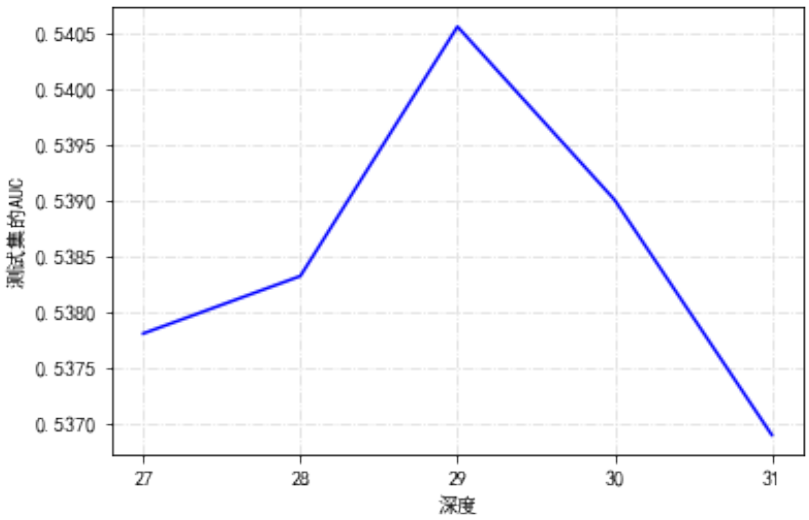



图 6 决策树的深度和测试集 AUC 的关系

图 7 根据决策树模型对特征的重要性进行了排序，对于每个卖家，其买家购买天数标准差是做重要的变量，该变量表明卖家所卖的商品受众类型较多，既有经常购买的用户，也有一次性购买的用户。其次是年龄、购买行为占比。

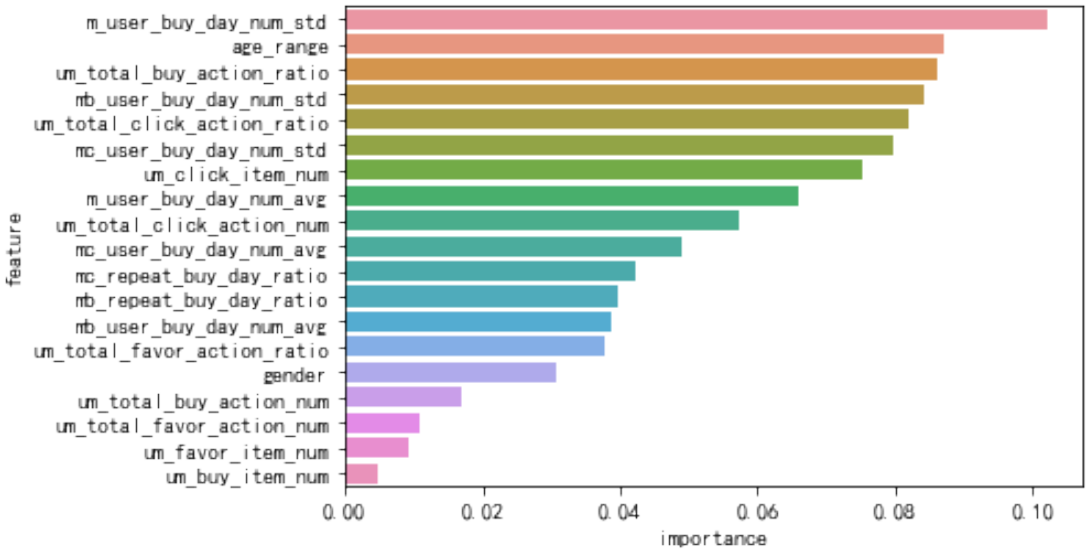



图 7 决策树模型对特征的重要性排序

如表 6 所示，以 AUC 为评价指标，朴素贝叶斯模型的效果最好，该模型的拟合时间最快，如果以正确率为评价指标，SVM 的效果最好。但是正确率没有

考虑虚报、误报，而这对于复购用户的预测较为重要，所以朴素贝叶斯模型是最好的。

表 6 决策树模型对特征重要性程度的排序

| 评价指标 | 朴素贝叶斯 | Logistic | 决策树 | SVM |
|---|---|---|---|---|
| trainAUC | **0.5967** | 0.5008 | 0.5008 | 0.5002 |
| testAUC | **0.5949** | 0.5007 | 0.5007 | 0.5002 |
| train 正确率 | 0.6389 | 0.9389 | 0.9389 | 0.9390 |
| test 正确率 | 0.6375 | 0.9381 | 0.9381 | 0.9382 |
| 拟合时间 | 0.0700 | 1.4400 | 1.4400 | 12.4700 |

### 3.3 模型融合与结果分析

采用 Bagging 方法进行模型融合，结果如表 7 所示，可以看出 Bagging 方法对决策树模型有较大的提升，而对朴素贝叶斯、Logistic、SVM 几乎没有提升。

表 7 Bagging 方法对各学习器的融合效果

| 评价指标 | 朴素贝叶斯 | Logistic | 决策树 | SVM |
|---|---|---|---|---|
| trainAUC | 0.5962 | 0.5008 | 0.8707 | 0.5002 |
| testAUC | 0.5959 | 0.5007 | 0.5494 | 0.5001 |
| train 正确率 | 0.6362 | 0.9389 | 0.8473 | 0.9390 |
| test 正确率 | 0.6347 | 0.9381 | 0.7922 | 0.9382 |
| 拟合时间 | 0.7000 | 11.7700 | 20.3600 | 235.7000 |

### 3.4 模型的预测效果验证

将朴素贝叶斯模型的预测结果提交至天池比赛官网得到模型的最终预测效果为 0.6291287，在榜单上可以排在 910 名 (https://tianchi.aliyun.com/competition/entrance/231576/rankingList)。


(a)自己的成绩

| | | | | |
|---|---|---|---|---|
| 913 | dragoner203029 | 南京航空航天大学 | 0.629372 | 2022-01-07 |
| 914 | 3go3cacwe4pae | 中国科学院 | 0.629284 | 2021-11-26 |
| 915 | 柯金宏 | 中山大学 | 0.629129 | 2022-01-21 |
| 916 | NM_SZ2116121 | 南京航空航天大学 | 0.628649 | 2021-12-31 |

(b) 官网排名，从左往右依次为排名、昵称、学校、成绩、提交日期

图 8 比赛官网上的成绩

## 4 结论与对策建议

本研究选取了与复购预测相关的变量。在实际应用中，需要获取用户活动日志文件，并从日志文件从计算这些指标，才能对用户的复购进行预测。其中，每个商家的所有用户购买天数的平均值、标准差以及用户年龄对复购预测的作用更大。对于该数据集，目前发现朴素贝叶斯模型的预测效果相较于 Logistic 回归、决策树和 SVM 的效果更好。但是，少量的特征对模型预测效果并不理想，后续研究需要进行系统的特征工程，构建更多指标并从中筛选出有效的指标。

## 参考文献

[1]熊晓元. 基于互动和感知理论的网络重购行为研究 [D]; 西南交通大学, 2014.
[2]WU L-Y, CHEN K-Y, CHEN P-Y, et al. Perceived value, transaction cost, and repurchase-intention in online shopping: A relational exchange perspective [J]. Journal of Business Research, 2014, 67(1): 2768-2776.
[3]CHANG E-C, TSENG Y-F. Research note: E-store image, perceived value and perceived risk [J]. Journal of Business Research, 2013, 66(7): 864-870.
[4]YOON C, GONZALEZ R, BECHARA A, et al. Decision neuroscience and consumer decision making [J]. Marketing Letters, 2012, 23(2): 473-485.
[5]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.
[6]陈洁, 谢文昕, 杨升荣. 在线渠道消费者动态品牌选择购买率预测 [J]. 工业工程与管理, 2011, 16(3): 52-56.
[7]周成骥. 基于机器学习的商品购买行为预测模型设计 [D]; 广州大学, 2018.
[8]CUI D, CURRY D. Prediction in marketing using the support vector machine [J]. Marketing Science, 2005, 24(4): 595-615.

[9]ZUO Y, ALI A B M S, YADA K. Consumer purchasing behavior extraction using statistical learning theory [J]. Procedia Computer Science, 2014, 35: 1464-1473.
[10]LIU G, NGUYEN T T, ZHAO G, et al. Repeat buyer prediction for e-commerce [M]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 155-164.
[11]CHANG H-J, HUNG L-P, HO C-L. An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis [J]. Expert Systems with Applications, 2007, 32(3): 753-764.
[12]QIU J, LIN Z, LI Y. Predicting customer purchase behavior in the e-commerce context [J]. Electronic Commerce Research, 2015, 15(4): 427-452.
[13]IJCAI. Repeat buyers prediction after sales promotion[EB/OL]. (). https://ijcai-15.org/repeat-buyers-prediction-competition/.
[14]阿里巴巴. 天猫复购预测之挑战baseline[EB/OL]. (). https://tianchi.aliyun.com/competition/entrance/231576/information.