京东高潜用户购买意向预测

数据挖掘课程期末报告 第 12 组 - 2018 年 12 月 31 日

组长：王嬰 18210980068
组员：栗书敬 18210980007
张继丹 18210980019
张媛媛 18210980079
张云华 18210980080
柳素问 18210980090

京东高潜用户购买意向预测

摘要：本次期末报告以预测京东高潜用户的购买意向为研究目标，通过数据探索、描述分析、特征工程以及建模分析等多方面，深入理解当今我国网购市场用户的购买行为，并基于用户真实购买行为信息，利用 Logistic 回归、随机森林、XGBoost 等模型，预测用户在未来 5 天内对某个目标品类下商品的购买意向，最终构建了 XGboost 和 LightGBM 的集成模型实现了对高潜用户购买意向预测效果的提升。

一、背景介绍

近年来随着人民物质消费水平的提高以及新一代消费群体的逐渐崛起，我国网络购物市场正持续高速发展。而互联网产业技术的不断提高和发展更是为广大消费者提供了越来越便利的网络购物环境。据中国互联网络信息中心发布报告的统计数据显示，在 2018 年上半年中国网络购物用户规模及占比情况中，中国网络购物用户规模为 56892 万人，与 2017 年末相比增长 3560 万人，网购人数占整体网民比例高达 71%。如图 1-1 所示，我国的网络购物用户规模正处于持续稳定增长的状态。拥有广大的用户群体为基础，网络购物市场无疑成为了充满商机的行业领域。

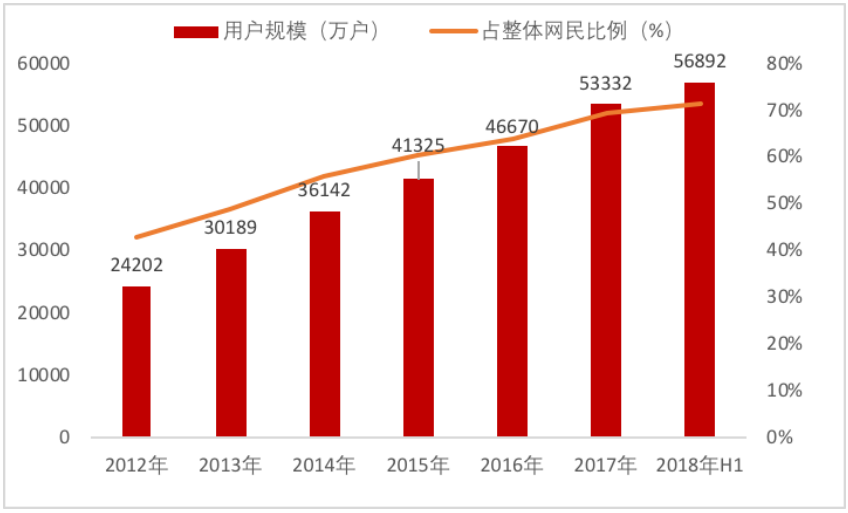

图 1-1 2012-2018 上半年中国网络购物用户规模及占比情况

在网购电商平台中，竞争更是逐年愈发激烈。京东商城作为全国综合网络零售商，中国电子商务领域受消费者欢迎和具有影响力的电商平台之一，多年来在保持高速发展的同时，沉淀了数亿的忠实用户，积累了海量的真实数据。基于培养人才、激发创新和释放数据价值等目标，京东在 2018 年上半年举办了京东 JDATA 算法大赛，提供了开源的脱敏后的京东真实用户历史行为数据。基于此，我们通过利用京东开源的数据，不仅能够将尝试从历史数据

中找出规律，将本门课程所学习到的知识运用其中，将大数据的思维和方法应用于市场实务中，尝试突破精准营销的关键问题。对于我们每个人而言，从研究网购用户历史行为数据的角度出发，更是对自身的日常网购行为的一种全新理解和认识。

在这样的背景下，我们将本次研究的目标设为：预测用户在未来 5 天内对某个目标品类下商品的购买意向，并输出高潜用户和目标商品的匹配结果。力图通过对原始数据特征的提取和模型的构建，更准确地预测未来一段时间内可能发生购买行为的用户-商品组合，发掘高潜用户的购买意向，提高电商平台精准营销、发掘高潜用户的能力，为最终产品销售量的提高创造条件。

二、数据说明

本次研究使用的数据来自京东 JDATA 算法大赛提供的脱敏后的用户历史行为记录，预测用户在未来 5 天内对某个品类下商品的购买意向，输出高潜用户和目标商品的匹配结果。原始数据为在 2016 年 2 月 1 日-2016 年 4 月 15 日时间跨度内用户基本信息数据、商品基本信息数据、商品评论信息数据以及用户对商品行为等数据。

对于四部分的原始数据，我们通过数据清洗的过程将这四个表的信息进行整合，将用户与商品联系起来。首先，我们根据用户 ID，将用户的基础信息与用户行为表进行整合，考虑到公司运营方面对转化率的重视，因此我们利用下单与其他各行为的比值，在对数处理后构建了各类行为的下单转化率，并根据浏览转化率和点击转化率值的大小判断用户是否为惰性用户，由此剔除了这两个值小于 0.0005 的用户数据。同理，我们将商品基础信息与评论信息，行为信息进行整合，依据同样的方法求出商品转化率的信息。最后根据用户累计行为和商品累计行为信息用来提取累计行为的特征，并选取三天、一星期、半个月、一个月为时间窗口，以此计算用户和商品累计行为的平均数。

在特征提取后，我们对分类变量采用 one-hot 编码，选取对类别 8 的商品的下单行为作为研究对象，[1]共筛选整合成 5038 条数据。基于此，预测 2016 年 4 月 16 日到 2016 年 4 月 20 日时间区间内用户对该类商品的下单情况。具体的变量说明如表 2-1 所示。

[1] 在描述性分析的数据探索部分，会对选取类别 8 商品作为研究对象的原因展开说明。

表 2-1 数据变量说明表

变量类型		变量名	详细说明	取值范围	备注
因变量		是否下单	分类变量	0,1	行为下单,品类 8 的标签为 1
自变量	用户因素	用户编号	数值变量	ID 值	脱敏
		年龄	分类变量	1-6	-1 表示未知
		性别	分类变量	0,1,2	0-男, 1-女, 2-保密
		用户等级	分类变量	1-5	级别枚举
		用户行为类别	分类变量	1-6	1-浏览, 2-加入购物车, 3-购物车删除, 4-下单, 5-收藏, 6-点击
		用户行为转化率	连续变量		分别为浏览转化率、加入购物车转化率、购物车删除转化率、收藏转化率、点击转化率
	商品因素	商品编号	数值变量	ID 值	脱敏
		属性 1	分类变量	1-3	枚举，-1 表示未知
		属性 2	分类变量	1,2	枚举，-1 表示未知
		属性 3	分类变量	1,2	枚举，-1 表示未知
		品类编号	分类变量		脱敏，研究 8
		品牌编号	分类变量		脱敏，枚举
		商品行为	分类变量	1-6	1-浏览,2-加入购物车, 3-购物车删除, 4-下单,5-收藏,6-点击
		商品行为转化率	连续变量		分别为浏览转化率、加入购物车转化率、购物车删除转化率、收藏转化率点击转化率
		评论数	连续变量	0-4	0-无评论, 1-有 1 条评论, 2-有 2-10 条评论, 3-有 11-5 条评论, 4-有大于 50 条评论
		是否有差评	分类变量	0, 1	0-无, 1-有
		差评率	连续变量	[0, 1]	
	累积因素	某个时间段内用户行为计数	连续变量		6 种用户行为分别计数
		某个时间段内用户行为平均数	连续变量		6 种用户行为分别计平均数
		某个时间段内商品行为计数	连续变量		6 种商品行为分别计数
		某个时间段内商品行为平均数	连续变量		6 种商品行为分别计平均数
		时间窗口	分类变量	3,7,15, 30	单位：天

三、描述分析

在对用户购买商品意向进行模型探究预测之前，首先对用户基础信息以及用户的商品转化率等变量进行描述性分析，并对用户购买行为以及商品销售的情况进行数据探索，以初步判断对用户未来一段时间内购买意向的影响因素，为后续研究做铺垫。

（一） 用户基础信息

在分析用户购买行为信息之前，首先对我们研究所选取的对品类为 8 的商品具有行为的 5038 名用户的基础信息进行描述。在 5038 名用户对品类为 8 的商品的行为中，有 1329 名用户具有下单行为，占比 26.38%，没有下单行为的用户数量为 3709，占比 73.62%。进一步根据用户注册填写的性别信息，如图 2-1 可知，总体样本中填写“保密”为性别的用户占比最多，男性用户次之，注册信息填写为“女性”的用户最少。我们初步推测，在庞大的性别“保密”的用户群体里，绝大多数可能是女性，处于对自身隐私的保护等原因，她们并没有在京东注册时填写性别信息，这也符合我们对网络购物市场消费者群体的基本认知。

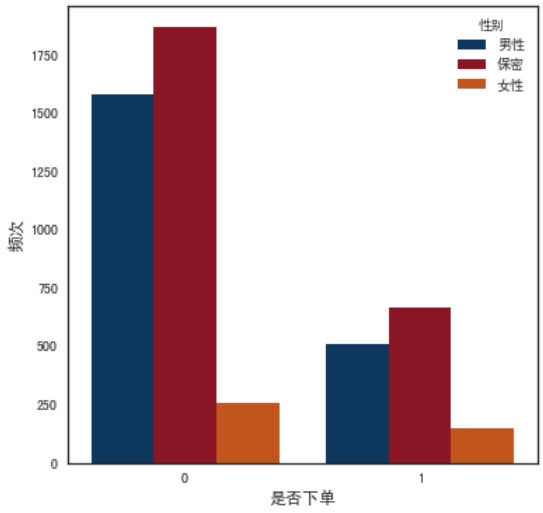

图 3-1 用户性别分布柱形图

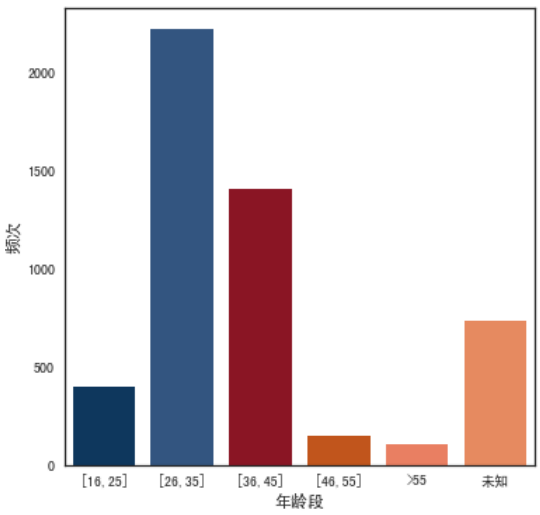

图 3-2 用户的年龄段分布柱形图

通过对用户年龄段绘制的柱形图 2-2 可以看出，本研究样本用户年龄主要集中在 26-45 岁之间。其中 26-35 岁的用户共计 2225 人，占总样本的 44.16%；36-45 岁的用户占总样本的 27.99%，共计 1410 人。超过 3/4 的用户在此年龄范围内，也显示出了如今网购市场消费主力群体的年龄层。

（二） 数据探索

对于用户购买行为的分析我们从原始用户行为数据开始，通过对原始属于全部用户和商品去重后的结果按照购买数量的星期和具体日期分布初步探索用户购买商品行为的轨迹。我们从用户行为数据中提取出行为类型四，即用户对商品下单购买的行为数据，按购买日期

（星期一至星期日）分类，计算去重后的用户数和商品数分布情况，绘制如图 3-3 所示的用户和商品的购买行为次数的星期分布图。

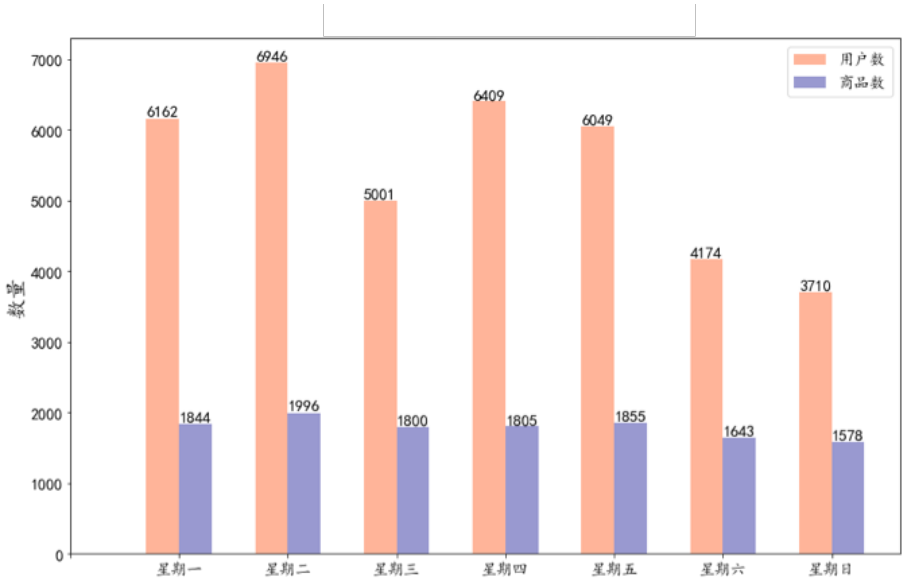

图 3-3 用户和商品的购买行为次数的星期分布图

从图 3-3 中可以看出，星期二有购买行为的用户数和商品数最多，而周六、周日数量较少。我们初步判断周末网购用户数量和商品销售数量较是主要是由于周末更多的人会选择利用休息时间外出逛街购物休闲娱乐，而不再是网上购物这种简单便捷的方式。

在月度购买行为数据分析方面，我们首先针对 2016 年 2 月的购买情况，同样分为用户数和商品数，进行户和商品的购买行为次数的统计。如下图 3-4 中可以看出，2016 年 2 月 1 日至 15 日之间用户购买商品以有明显的购买量下降又回升的趋势，可能的原因是，该时间段处于春节前后，很多商家不再发货，快递也不再营业，网购数量减少，节后又慢慢恢复以往的状态。

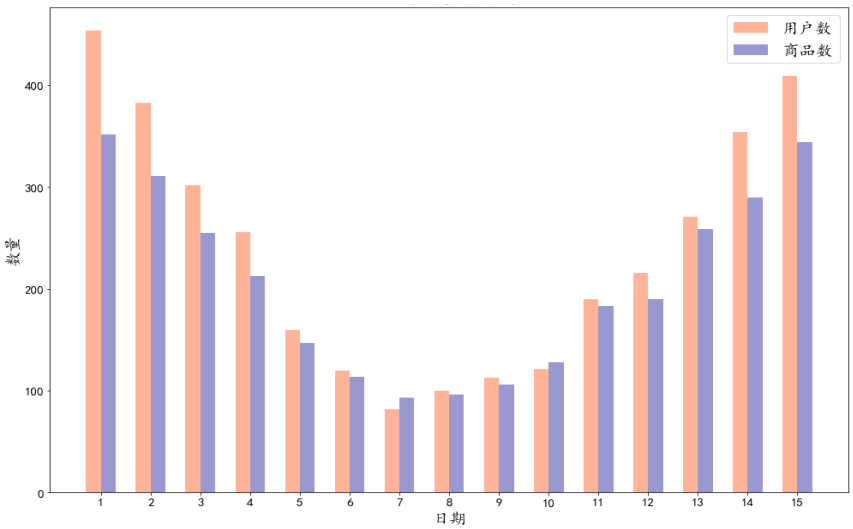

图 3-4 2016 年 2 月用户和商品的购买行为次数的日期分布图

而在 2016 年 3 月绘制的用户和商品的购买情况分布图 3-5 中可以看出，3 月 15 日附近购物量急剧增多，带动 3 月 15 日前后的购物量均有所提高。

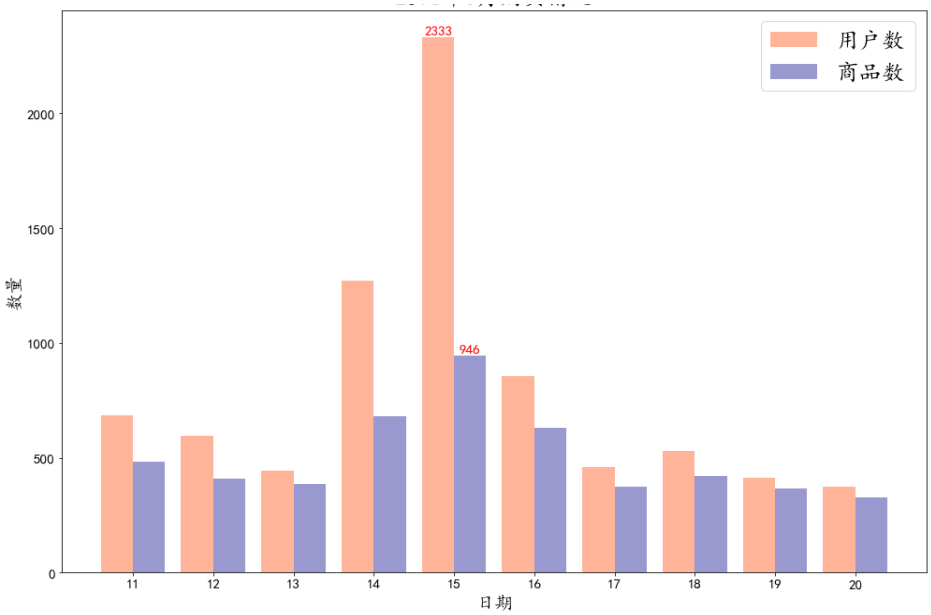

图 3-5 2016 年 3 月用户和商品的购买行为次数的日期分布图

我们初步判断可能是因为“3.15 消费者权益日”的因素影响，在该日前后京东商城出台了若干对消费者有优惠的活动所导致的。综合 2016 年 2 月和 3 月份的数据俩看，3 月份体现购买行为的用户数和商品数多于 2 月份。

（三） 商品信息

在分析完商品数变化趋势后，我们根据商品不同类别进一步分析不同类别商品的销售情况，如图 3-6 所示，类别 8、类别 5 和类别 4 是分别是本研究中销售量排名前三的类别。而类别 8 的商品销售量基本保持持续的领先水平，其销售量的变化趋势与整体全类型商品的变化趋势相似，总体商品数变化主要是由类别 8 的商品销售情况引起的。

基于这样的发现，我们明确了类别 8 的商品在整个用户购买行为数据中的重要地位，以及验证我们选择以预测用户对类别 8 商品未来的购买意向作为本研究核心目标的合理性与价值。

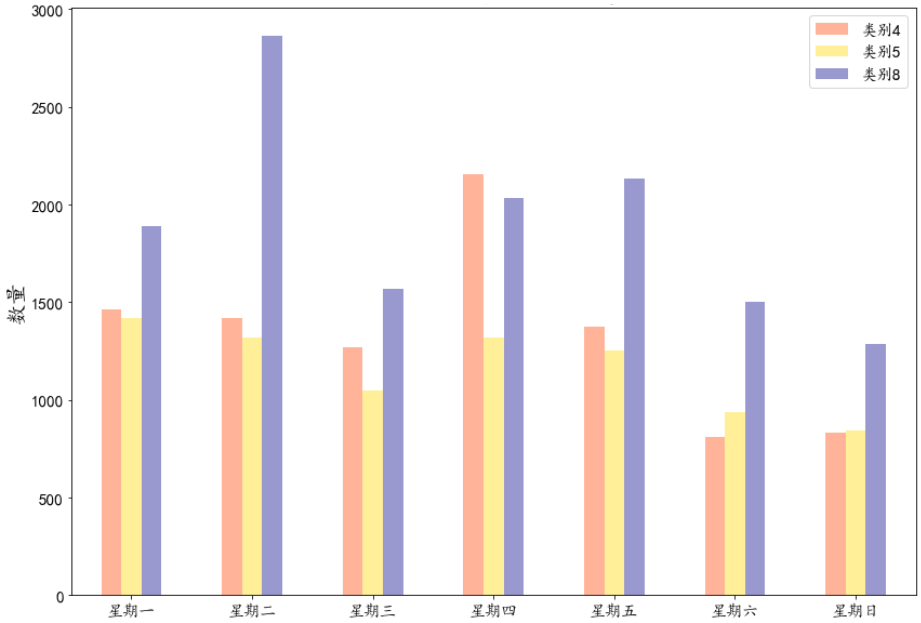

图 3-6 类别商品销售数量情况星期分布图

进一步，我们单独分析类别 8 商品在 2016 年 2 月、3 月和 4 月的销售量变化趋势。如图 3-7 所示，我们发现在 2016 年 2 月，类别 8 商品的购买意愿普遍偏低，相比之下 3、4 月份大多数情况下销售量相近，但 8 日、9 日和 15 日有较明显的差距，尤其是 15 日，类别 8 的销售量达到了一个巅峰，将近占了 3 月 15 日商品总销售量的半数左右。分析可能的原因，主要是由于电商促销等原因导致的。

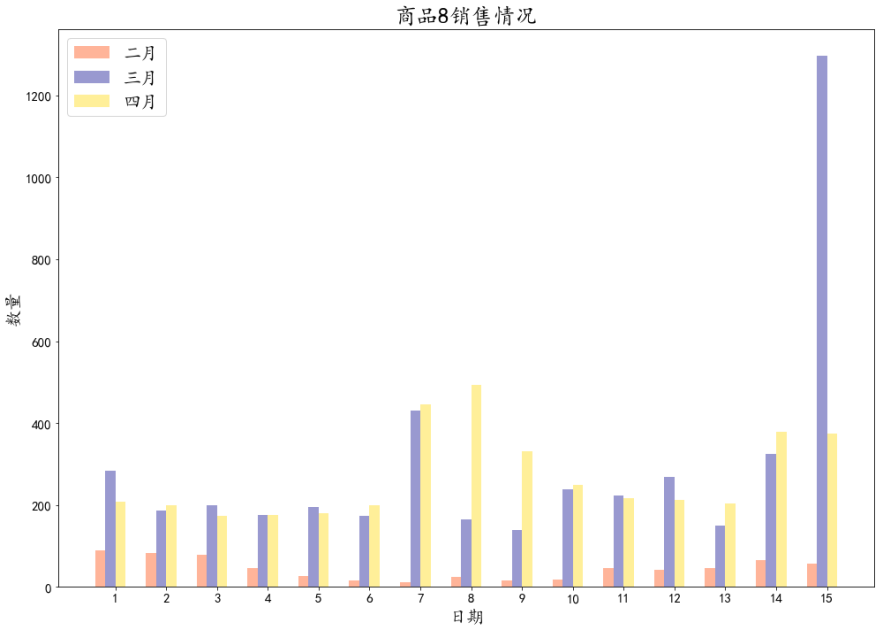

图 3-7 类别 8 商品销售数量情况日期分布图

类别 8 商品具有较高的销售数量，相对更受平台用户的欢迎。针对这样的情况，我们对影响其高销售量的因素展开了进一步的探究。我们对类别 8 商品的评论数和差评率数据进行分析，将对类别 8 商品的评论数划分为 0，1，2-10，11-50 以及大于 50 这样的评论数分类变量，绘制了如图 3-8 的柱形图。

如图可以看到，绝大多数的类别 8 商品的评论数在 50 条以上，说明对该类别商品有购买行为的用户数量较多，且该类商品能够获得较多的反馈信息，有助于提高商品的知名度。通过对类别 8 商品的差评率绘制直方图可以看到，在图 3-9 中类别 8 商品的差评率相对较低，平均在 2.94%左右，说明类别 8 商品的评论数较多，且多数为对商品质量和满意度的正反馈信息，两者结合来看，商品评论数的提高会对商品的销售数量产生一定的正向促进作用。

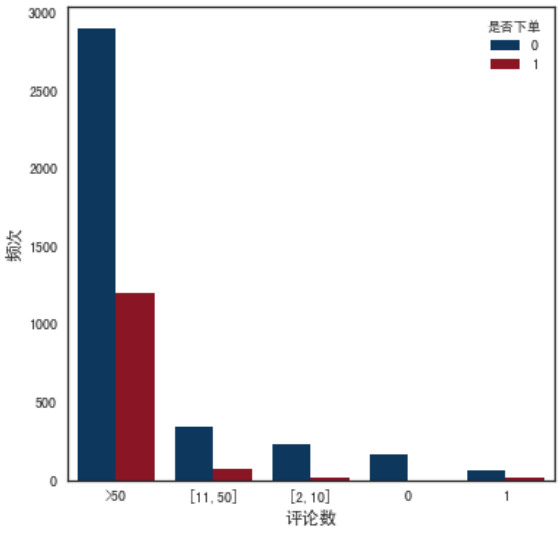

图 3-8 商品评论数分布柱形图

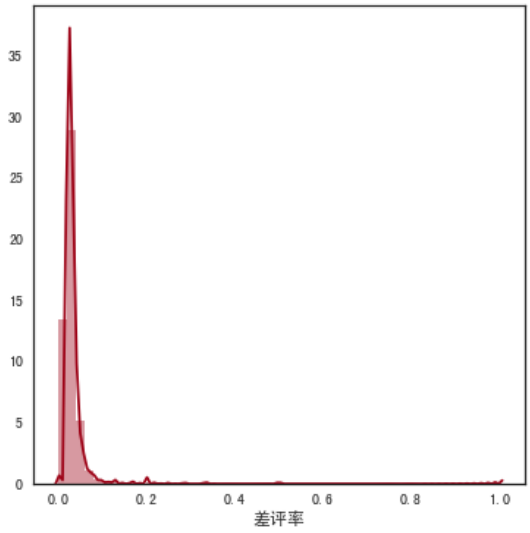

图 3-9 商品差评率分布直方图

（四） 行为信息

基于用户对类别 8 商品的不同行为信息，我们利用下单与其他各行为的比值，在对数处理后构建了各类行为的下单转化率，如图 3-10 所示，分别对“下单-浏览”、“下单-加购物车”、“下单-购物车删除”、“下单-收藏”以及“下单-点击”的转化率进行描述。由图可以得出，浏览点击行为的转化率较低，加购物车，从购物车删除，收藏的转化率较高。购物车删除行为的转化率高，我们初步分析原因可能是由于：当用户决定买某个商品时，会把相似的商品从购物车删除。

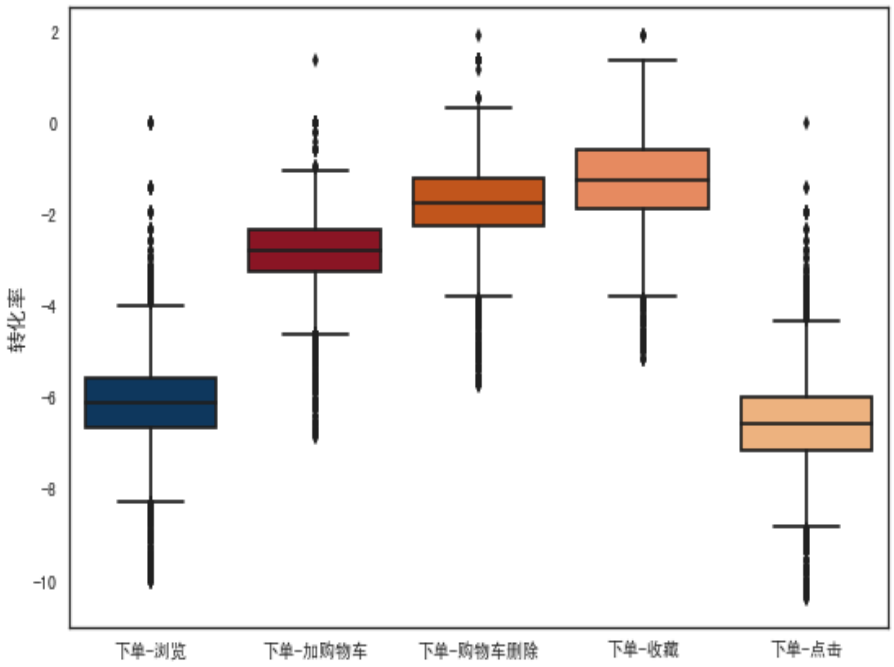

图 3-10 用户不同行为的转化率箱线图

在对用户的购买行为有初步的了解后，接下来对用户进行分类，分为高潜用户和非高潜用户。高潜用户的特征是[2]：（1）必须有购买行为（2）对一个商品购买和其他五种交互行为(浏览，点击，收藏等)的时间差多于一天，这样的用户更容易在近期内再次产生购买行为。基于这样的定义，我们对样本数据中的用户和商品进行了进一步的分类，并以用户编号为 295413 的用户，在购买商品编号为 18585 的商品时为例，因为此用户-商品组合满足高潜用户的条件，绘制该用户关于该商品的行为轨迹图。

如图 3-11 所示，该用户共产生了 3 次购买行为，第一次购买（4 月 13 日）前，4 月 2 日该用户已经将商品加入购物车，并进行了点击和浏览，但是没有下单，8 天后（4 月 10 日）又点击浏览该商品，可能觉得哪里不合适或者遇到了更合适的商品，将其从购物车中删除，但一天后又点击浏览，加入购物车，并在多次点击和浏览后下单，也比较符合大众挑选商品时比较纠结的心态，第一次购买至第二次购买（4 月 14 日）之间，由于对商品已经有一些了解，用户只有点击和加入购物车的行为，第二次下单后又进行浏览，可能发现了这个商品的更多优势，又在当天产生第三次购买行为。

[2] 高潜用户定义参考 https://github.com/daoliker/JData

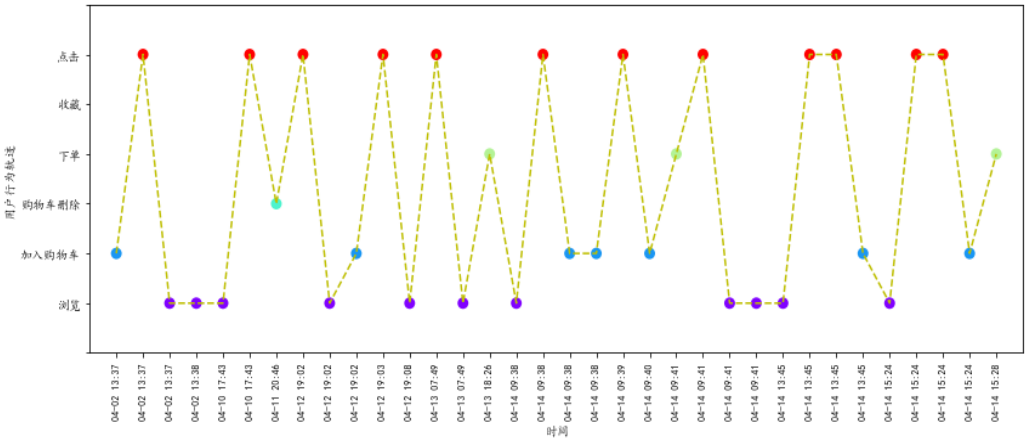

图 3-11 某一高潜用户行为轨迹示意图

用户编号为 257591 的用户，在购买商品编号为 96536 的商品时，表现为非高潜用户，绘制该用户关于该商品的行为轨迹图，如图 3-12 所示。我们发现该用户从开始有交互行为至下单只隔了 6 个小时，且前期只有加入购物车和点击行为，并无浏览行为，后期对商品进行了点击浏览后很快下单，很可能是 4 月 15 日当天突然对某种商品有需求，到京东上查询相关商品并快速下单，但短期之内可能不会再有类似的购买行为，不是京东的常驻用户。

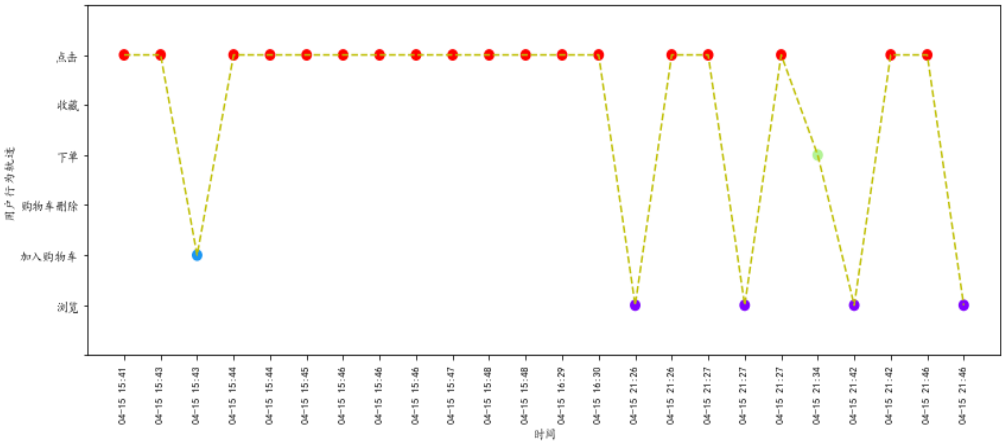

图 3-12 某一非高潜用户行为轨迹示意图

四、特征工程与建模分析

基于对用户购买行为信息和商品销售情况等的了解，我们在特征工程与模型建立部分尝试采用 Logistic 回归、随机森林、XGBoost 等模型，并在此基础上进一步改进算法，提高预测效果。由于数据预处理后，我们得到 78 个变量，数据集维度较高，极有可能存在变量冗余的情况。所以我们在采用 logistic 回归进行建模时，选择了三种惩罚机制，分别为 Ridge、Lasso 和 Elastic Net。针对随机森林和 XGBoost 这两种算法，我们通过进行特征工程，采用了四种方法，进而比较特征选取前后的模型的预测结果。随后，我们基于初步的模型预测效果，通过改进和集成算法进一步提升模型的预测结果。

（一） Logistic 回归

数据预处理后，我们所选取的变量共有 78 个，包括 77 个自变量和一个 label。去除用户 ID 和商品 ID 后，仍剩余 75 个自变量。这是一个较高维的数据，因此在我们所造出来的特征中也不可避免存在变量相关和冗余的情况，故我们首先考虑采用带有惩罚项的 Logistic 回归进行建模。这里的惩罚机制包括 Ridge、Lasso 和 Elastic Net，即为一范数、二范数和两者的混合。

首先将数据集按 80%和 20%的比例拆分成训练集和测试集。由于 Ridge 和 Lasso 均只有一个参数 lambda，故我们首先对训练集采取交叉验证的方式找出最佳的 lambda 值，再将最优的 lambda 值带入模型中从而得到 label 的估计值。交叉验证的结果如图 4-1 和 4-2 所示。

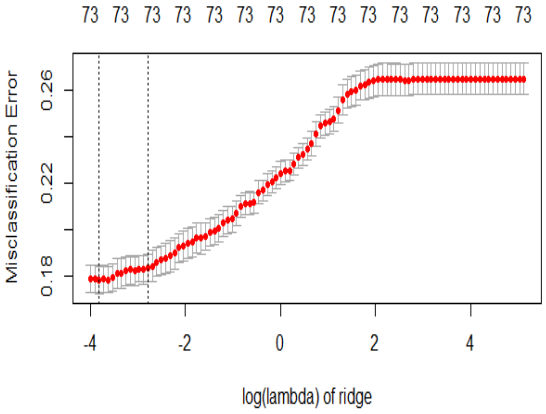

图 4-1 Ridge 交叉验证结果

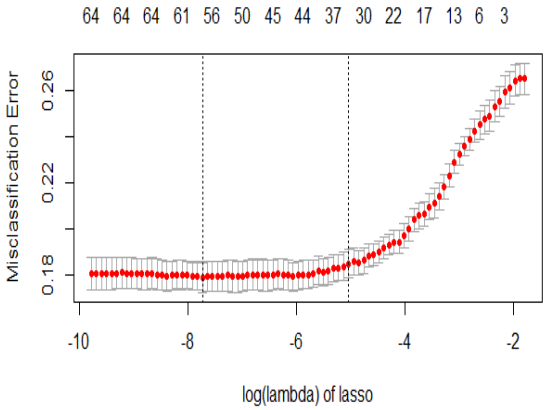

图 4-2 Lasso 交叉验证结果

我们注意到，在所记录的数据中，最后实际发生购买的不足 10%，同时由于我们的目标是希望将高潜用户尽可能多地预测出来，故在建模中我们会更侧重考虑召回率，也会适当的减小 threshold 的值。根据测试集上的输出结果，我们将该值设置为 0.4，最终得到 ridge 和 lasso 的训练集和测试集上的结果如下表 4-1 所示。

表 4-1 Logistic 回归预测结果

Ridge	召回率	准确度	Lasso	召回率	准确度
训练集	0.8425	0.7206	训练集	0.8285	0.7233
测试集	0.8359	0.7063	测试集	0.8397	0.7093

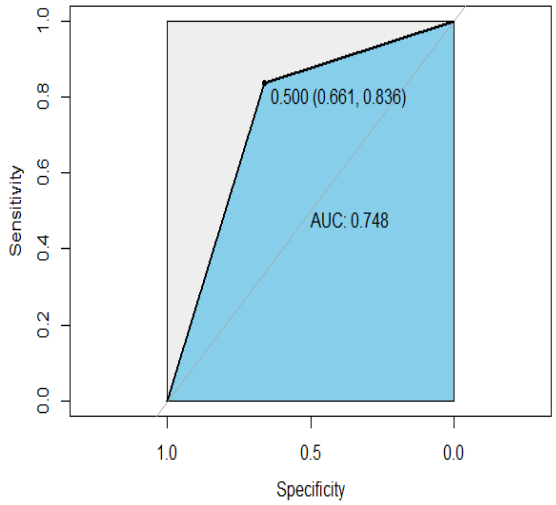

图 4-3 Ridge 方法下模型预测结果

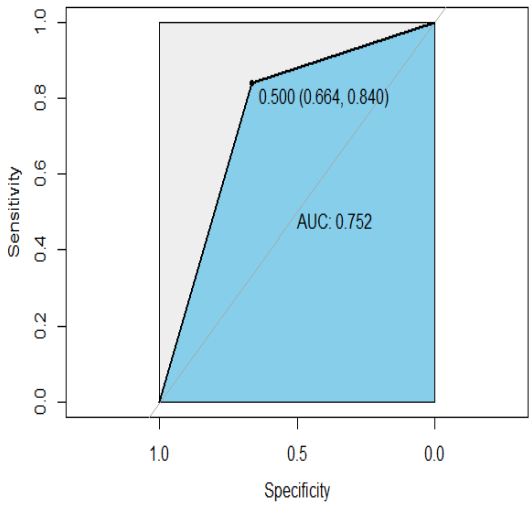

图 4-4 Lasso 方法下模型预测结果

接下来我们用 lasso 和 ridge 的混合，即 elastic net 来进一步建模分析。lasso 和 ridge 其实是 elastic net 的特殊情况，即在 alpha=0 和 alpha=1 的时候，elastic net 就退化成了 lasso 和 ridge。而通过绘制 lasso 方法下的准确度和召回率在不同 alpha 值下的曲线可以看到，随着 alpha 从 0 到 1，accuracy 波动幅度约为 0.4%，变化不大，故最终直接取 alpha=0，即 Ridge 回归的结果即可。

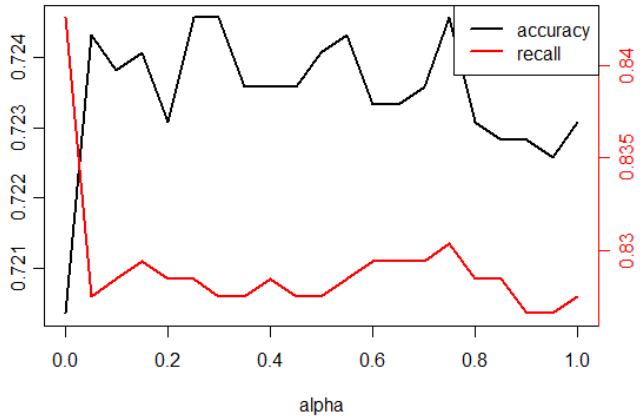

图 4-5 Lasso 方法下的准确度和召回率在不同 alpha 值下的曲线

（二） 随机森林

接下来我们选择随机森林算法，对用户购买意向进一步预测。我们通过进行特征工程，采用了四种方法，进而比较特征选取前后的模型的预测结果。通过对 5038 个数据(其中训练集占据 70%，测试集占据 30%)进行测试集训练集的划分。通过对 70%的训练集调参建立随机森林的模型进行分类，并利用得到的最优结果在 30%的测试数据上进行检验，我们得到了以下的结果：1.分类准确率为 80.95%；2.召回率为 71.28%。

（三） XGBoost

我们通过利用 XGboost 模型，对残差的不断学习来是最大化目标函数，提高用户购买意向预测模型的拟合效果。由于该算法是基于对残差的学习，因此较容易发生过拟合，故我们需要对参数进行一定的调整。由于该模型的参数较多，故在建模中，我们对其几个比较主要的参数 eta、min_child_weight、max_depth 采取网格搜索和交叉验证的方法来获得近似最优参数，在参考近似最优参数的基础上设定参数来进行建模。根据训练集的输出结果，我们将 threshold 取了一个较小的值 0.15。最终得到的训练集和测试集上的结果如表 4-2 所示，XGboost 模型给出的树结构如图 4-6 所示。XGboost 在测试集上的召回率为 0.8611，准确度为 0.8095，可以看到，这相比于 ridge、lasso 方法下的 Logistic 回归以及随即森林模型的预测效果来说，均有了很大的提升。

表 4-2 XGboost 模型预测结果

XGBoost	召回率	准确度
训练集	0.977	0.9283
测试集	0.8611	0.8095

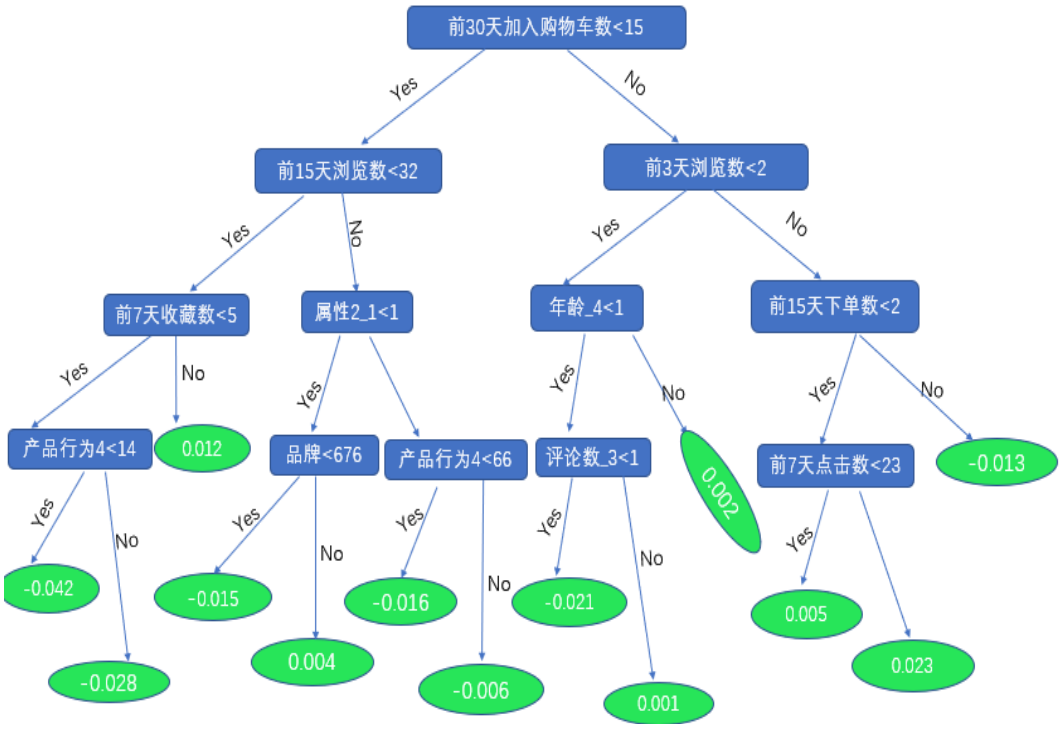

图 4-6 XGboost 模型结果下的树结构

（四） 特征工程

考虑到数据集维度较高，极有可能存在变量冗余的情况。因此我们针对随机森林和XGBoost 这两种算法，采用了四种特征选择的方法，降低数据维度，进而比较特征选取前后的模型的预测结果。特征选择方法具体如下：（1）data95：去掉 95%以上为零的变量剩下的变量，剩余 52 个变量；（2）fearture_rf：原数据集，按照 GINI important > 0.01 筛选出的变量，剩余 27 个变量；（3）fearture_rf_1：data95：按照 GINI important > 0.01 筛选出的变量，剩余 25 个变量；（4）fearture_lasso：lasso 方法筛选出的 lasso 系数大于 0.00001 的变量。特征选择后的随机森林和 XGBoost 模型的预测结果如表 4-3 所示。

表 4-3 特征选择后的模型预测结果

特征选择	评估准则	随机森林	XGBoost
Data95%	AUC	96.07%	95.83%
	ACC	93.57%	93.64%
	召回率	77.77%	76.74%
Feature_rf	AUC	96.08%	95.82%
	ACC	93.55%	93.52%
	召回率	79.57%	75.61%
Feature_rf_1	AUC	94.78%	95.82%
	ACC	93.55%	93.52%
	召回率	79.57%	75.61%
Feature_lasso	AUC	94.92%	94.74%
	ACC	92.70%	92.82%
	召回率	70.46%	68.81%

通过特征工程之后用模型对测试集对 AUC、ACC、召回率得分的计算，并和未作特征工程之前的随机森林和 XGBoost 进行比较发现，没有做特征工程的数据集的各种得分还是偏高的，这与集成算法的优势有密切的关系，通过比较，我们还是选择利用特征选择前的数据集上建模型。

而通过对比 Logistic 回归、随即森林、XGBoost 这三种模型的预测结果，我们也可以看到 XGBoost 模型的预测效果最好。因此我们将基于初步的模型预测效果，继续通过改进和集成算法进一步提升模型的预测结果。

（五） LightGBM

基于 Boosting 方法，我们还尝试利用 LightGBM 模型利用二阶梯度来对节点进行划分，在分裂的过称中较少对低信息增益节点的考虑，较大程度地提高训练速度。与前面用 XGBoost 建模过程相似，我们对 LightGBM 的几个比较重要的参数 num_leaves，learning_rate，

n_estimators 采用网格搜索交叉验证来找到近似最优参数，其它的参数则采用默认值。下图是随着树的数量不断增多，训练集的 AUC 越来越接近 1，测试集的最终也稳定在 97.5%左右。但是根据实际情况，出于重在预测高潜用户的目标，我们同样设置了一个较小的 threshold，根据训练集的结果在着重考虑召回率的基础上综合考虑准确率、AUC 等指标，选取 threshold 为 0.2，得到的训练集和测试集上的结果如表 4-4 所示，可以看到，LightGBM 模型在各个指标上都略优于 XGBoost。

通过观察该模型的特征重要性，如图 4-7 所示，对用户下单行为影响最显著的特征变量依次是：差评率，删除转化率，加入购物车转化率，产品点击率，浏览转化率等。对于商品的差评率重要性的理解，我们认为这或许就是网上之所以这么多商家费尽心思刷好评的原因。在网购中，我们不能直接接触到商品，因而其它买家的评论就尤为重要。接下来就是购物车删除数的转化率，一但买家把商品从购物车中删除，很大可能就是已经在别的店够买了商品或者是对该商品已经不需要或不感兴趣了，购买机率大大减小。

表 4-4 LightGBM 模型预测结果

LightGBM	召回率	准确度
训练集	1	0.9164
测试集	0.8993	0.8145

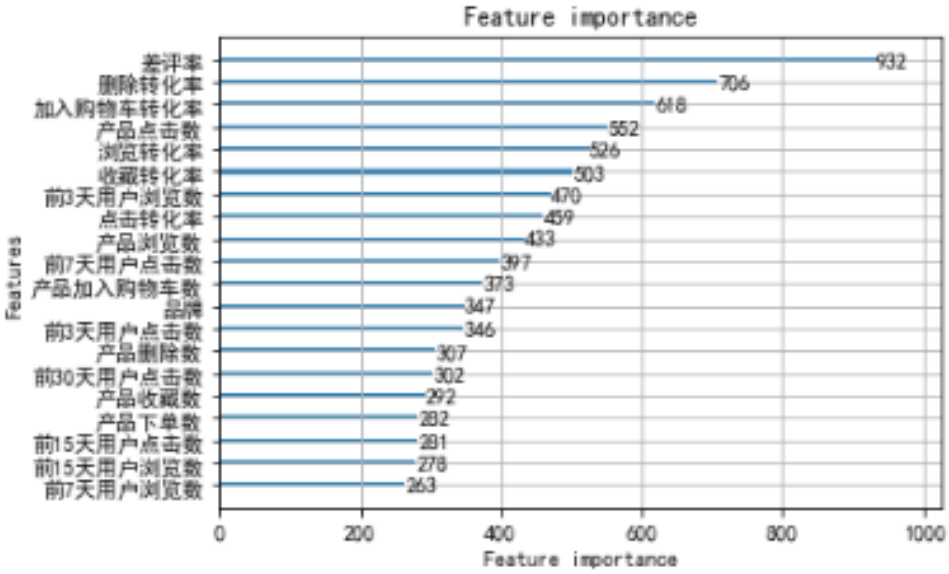

图 4-7 LightGBM 模型特征重要性

（六） LightGBM+XGboost 集成模型

鉴于以往经验，集成模型通常比单个模型效果更好。因此我们在前面模型的基础上，集成了 LightGBM+XGboost 的预测结果，得到了在不同比例下模型在训练集上的预测结果在

三个指标上的表现情况，如图 4-8 所示。

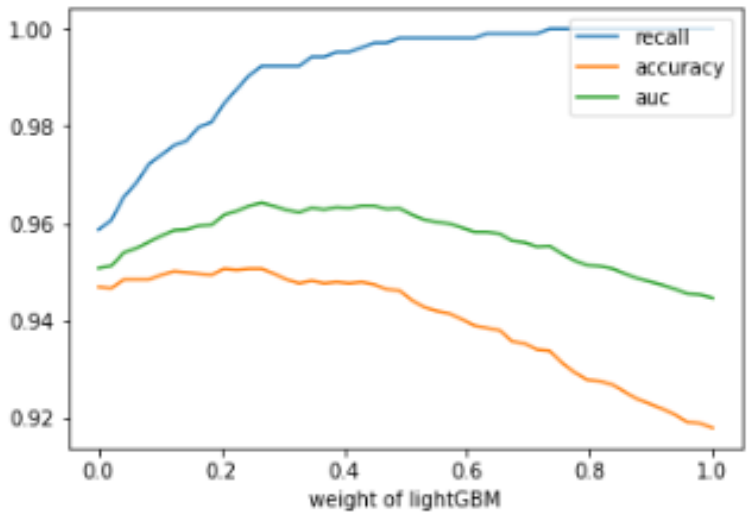

图 4-8 LightGBM+XGboost 集成模型的预测表现

可以看到，随着 LightGBM 的比例不断增加，召回率有继续提升的趋势，准确率和 AUC 呈现先升后降的趋势。当 LightGBM 的比例大于 40%以后，召回率基本不再发生变化，准确率和 AUC 转而持续下降。考虑到 LightGBM 在召回率上的优良表现，且召回率是我们在本次研究中着重考虑的目标，最后选取的是 50%XGboost 和 50%的 LightGBM 集成的模型，作为最终优化后的预测模型。如表 4-5 所示，在经过最终的模型预测结果比较后，XGboost 和 LightGBM 集成的模型在测试集上的召回率和准确度相对最高。

因此，我们最终选取 XGboost 和 LightGBM 集成的模型作为我们本次研究中对于用户未来购买意向预测的最优模型。

表 4-5 Boosting 方法模型预测结果比较

模型选择	召回率		准确度	
XGBoost	训练集	0.977	训练集	0.9283
	测试集	0.8611	测试集	0.8095
LightGBM	训练集	1	训练集	0.9164
	测试集	0.8993	测试集	0.8145
50%XGBoost+50%LightGBM	训练集	1	训练集	0.9275
	测试集	0.8924	测试集	0.8224

考虑到本研究是以类别 8 的商品作为用户购买的意向商品进行模型训练和测试的，结果可能存在一定的偶然性。因此，我们进一步扩展数据范围，将用户的购买意向不只是局限于类别 8 的商品，而是基于原始数据中所有类别的商品进行更近一步的模型检验。

我们基于上述具有较好效果的模型方法，如 XGboost、LightGBM 以及 XGboost 和 LightGBM 集成的模型，进一步预测用户对全类别商品的购买意向。如表 4-6 所示，模型的预测结果如下。我们可以看到，可能是由于数据量增多的原因，三种模型的效果均比仅针对类别 8 的商品的购买意向预测效果更好。而在这其中，50%XGboost 和 50%的 LightGBM 集成的模型在召回率和准确度上综合来看仍具有最好的表现。这也能从一方面说明，我们对用户购买行为的预测模型是相对稳健的。

表 4-6 基于全类别商品的 Boosting 方法模型预测结果比较

模型选择	召回率		准确度	
XGBoost	训练集	0.9888	训练集	0.942
	测试集	0.9111	测试集	0.9159
LightGBM	训练集	1	训练集	0.9442
	测试集	0.9228	测试集	0.9073
50%XGBoost+50%LightGBM	训练集	1	训练集	0.934
	测试集	0.9498	测试集	0.9032

五、结论与建议

本次期末报告使用京东 JDATA 算法大赛提供的脱敏后的用户历史行为记录，通过数据探索、描述分析的方法深入理解当今我国网购市场用户的购买行为，通过特征工程以及建模分析多种研究手段，基于用户真实购买行为信息，构建 Logistic 回归、随机森林、XGBoost 等模型，预测用户在未来 5 天内对某个目标品类下商品的购买意向。最终通过构建 XGboost 和 LightGBM 的集成模型，实现了对高潜用户购买意向预测效果的提升，得到如下结论：

1. 影响预测用户购买意向的主要因素有：（1）商品信息：差评率；（2）用户行为信息：删除转化率，加入购物车转化率，产品点击率，浏览转化率、前 3 天用户浏览数等。
2. XGboost 和 LightGBM 的集成模型针对我们研究的数据具有最好的预测效果。

由于近些年我国的网络购物用户规模正处于持续稳定增长的状态，网络购物市场仍是充满商机和活力的行业领域。对于如何利用已有信息和购买行为数据更准确地预测未来一段时间内可能发生购买行为的用户-商品组合，提高电商平台精准营销、发掘高潜用户的能力也将会成为持久热议的话题。因此，对于网购用户购买行为的研究，我们还需要拓展更多的维度，结合更多购买商品的实际内容，在预测模型中加入更多因素，如用户关注的社区、用户收藏店铺的信息等等，进一步丰富用户县骨干信息、添加购买商品信息的维度，以及对用户历史购买行为信息进一步整合。

六、小组分工

姓名	学号	小组分工
王罂	18210980068	数据处理，课堂汇报
栗书敬	18210980007	模型建立
张继丹	18210980019	模型建立
张媛媛	18210980079	特征工程
张云华	18210980080	数据探索
柳素问	18210980090	描述分析，报告撰写