学校代码：10730
分类号：O213 密级：公开

兰州大学

硕士学位论文

（专业学位）

论文题目(中文)	基于 LightGBM 的用户购买行为预测研究
论文题目(外文)	Research on User Purchase Behavior Prediction Based on LightGBM
作者姓名	王予涵
学科专业	应用统计
研究方向	机器学习
教育类型	学历教育
指导教师	焦桂梅 副教授
论文工作时段	2020 年 3 月至 2021 年 3 月
论文答辩日期	2021 年 5 月

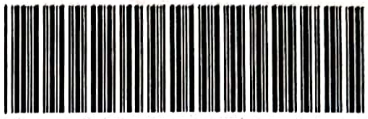
校址：甘肃省兰州市城关区天水南路 222 号

学　　院：数学与统计学院　　学　　号：220180921080

学生姓名：王予涵　　导师姓名：焦桂梅

学科名称：应用统计・应用统计

论文题目：基于 LightGBM 的用户购买行为预测研究



原 创 性 声 明

本人郑重声明：本人所呈交的学位论文，是在导师的指导下独立进行研究所取得的成果。学位论文中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究成果做出重要贡献的个人和集体，均已在文中以明确方式标明。

本声明的法律责任由本人承担。

论文作者签名：王予涵　　日　期：2021.6.3

关于学位论文使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学。本人完全了解兰州大学有关保存、使用学位论文的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本学位论文。本人离校后发表、使用学位论文或与该论文直接相关的学术论文或成果时，第一署名单位仍然为兰州大学。

本学位论文研究内容：

☑可以公开

□不宜公开，已在学位办公室办理保密申请，解密后适用本授权书。

（请在以上选项内选择其中一项打“√”）

论文作者签名：王予涵　　导师签名：焦桂梅

日　　期：2021.6.3　　日　　期：2021.6.3

基于 LightGBM 的用户购买行为预测研究

中文摘要

电子商务的概念愈发火热, 各种电商平台纷纷涌现, 越来越多的人加入网购的大军, 但当电商平台发展到一定程度后, 流量的增加终究会停止, 提高流量转化率无疑是一个重要且紧迫的课题. 目前各电商平台都引入了推荐算法, 为用户推荐其喜好的商品, 提高用户体验, 而预测是推荐的基础, 提前预测出用户的购买倾向无疑会大大提高推荐算法的效果, 这是一项极具意义的工作. 基于此, 本文选取京东算法大赛的数据来对用户购买行为预测进行研究, 主要工作内容及成果如下:

1. 确定预测目标: 在一段时间内有行为记录的用户-品类-店铺组合 (称为 F_1ID) 中, 预测未来 7 天会产生购买行为的 F_1ID, 这是预测中的二分类问题.

2. 确定训练集和预测集样本. 选取 2018-03-19 到 2018-04-01 按照正负样本比为 1:30 负采样后共 517049 个有行为记录的 F_1ID 作为训练样本, 其中在未来 7 天发生购买的 F_1ID 的标签为 1, 其余为 0; 选取 2018-03-26 到 2018-04-08 共 1792209 个有行为记录的 F_1ID 作为预测样本, 其中在未来 7 天发生购买的 F_1ID 的标签为 1, 其余为 0.

3. 构建基于时间滑动窗口的特征. 本文从基本特征、累积特征、时间滑动窗口特征 3 个方面在用户、品类、店铺、用户-品类、用户-品类-店铺 5 个维度构建了 564 维特征, 并对特征在缺失值等方面做了相应的处理.

4. 构建模型并选择最终的预测模型. 本文利用 LR、RF、GBDT、XGBoost、LightGBM 在 517049*565 的训练集上训练模型, 并在 1792209*565 的预测集上预测, 从 AUC、F_1 分数、训练时间等方面比较分析, 最终选取 LightGBM 作为最终的模型.

关键词: 用户购买预测, 时间滑动窗口, XGBoost, LightGBM

Research on User Purchase Behavior Prediction

Based on LightGBM

Abstract

The concept of e-commerce is becoming more and more popular. Various e-commerce platforms are emerging one after another. More and more people join the army of online shopping. When the e-commerce platform develops to a certain extent, the increase of traffic will eventually stop. Improving the traffic conversion rate is undoubtedly an important and urgent topic. At present, all e-commerce platforms have introduced recommendation algorithms to recommend their favorite products for users and improve user experience. Prediction is the basis of recommendation. Predicting the purchase tendency of users in advance will undoubtedly greatly improve the effect of recommendation algorithm, which is a very meaningful work. Based on this, this paper selects the data of Jingdong algorithm contest to study the prediction of users' purchasing behavior. The main work and results are as follows:

1. Determine the prediction target: in the user category store combination (called F_1ID) with behavior records in a period of time, it is predicted that the F_1ID of purchase behavior will be generated in the next seven days, which is a binary classification problem in prediction.

2. Determine the samples of training set and prediction set. A total of 517049 F_1IDS with behavior records from March 19, 2018 to April 1, 2018 were selected as training samples after negative sampling with a positive and negative sample ratio of 1:30, in which the tag with purchase F_1ID in the next 7 days was 1, and the rest was 0. A total of 1792209 F_1IDS with behavior records from March 26, 2018 to April 8, 2018 were selected as prediction samples, in which the tag with purchase F_1ID in the next 7 days was 1, and the rest was 0.

3. Construct features based on time sliding window. This paper constructs 564 dimension features from three aspects of basic features, cumulative features and time sliding window features in five dimensions of user, category, store, user category and user category store, and deals with the missing values of features.

4. Build the model and select the final prediction model. In this paper, LR, RF, GBDT, XGBoost and LightGBM are used to train the model on the 517049*565 training set, and to predict on the 1792209*565 prediction set. From the AUC, F_1 score, training time and other aspects, LightGBM is selected as the final model.

Keywords: User purchase forecast, time sliding window, XGBoost, LightGBM.

目 录

第一章 绪论

在互联网的背景下, 电子商务的概念是一个热点, 一系列电商平台应运而生, 随之吸引了大批用户的使用, 并且用户数日益增加, 这使得大量的行为数据产生, 如果我们能够利用这些真实的行为数据, 再结合精准高效的预测算法, 准确预测出用户可能购买的品类或店铺, 相应的为用户推荐该品类或店铺下的商品, 那么这在为平台提升收益、为用户提供便利方面将具有重大意义. 本章将介绍 “用户购买预测” 课题的研究背景与意义、该课题国内外的研究成果, 以及本文的组织架构.

1.1 研究背景及意义

电子商务发展至今已经有二十多年的历史. 在这期间, 我国电子商务不断涌现出丰富的创业创新模式, 电商平台提供线上交易的商品和环境, 在线支付技术实现货币交换, 快递物流行业将商品从商家手中运输到用户手中, 这种一体化的电子商务生态系统已经形成[1]. 电商平台几乎涵盖了所有线下能买到的商品, 并且品类丰富多样, 给用户提供了更多的选择, 线上即可支付, 购买方便, 直接可以送至家门, 网购的消费模式越来越流行. 据统计, 上网用户中 74.8%的用户有网购的经历, 约 6.39 亿, 这还仅是截至 2019 年 6 月的统计数字[2].

然而当电商平台发展到一定程度, 流量的增长却不会继续, 这时提高流量转化率对一个电商平台来说尤为重要[2]. 有几种方式可以提高转化率, 一是增加商品的多样性, 这样可以为少数用户的特定需求提供可能, 从而覆盖更多的人群, 二是增加用户对平台的喜好度, 加强老用户的粘性, 吸引更多的新用户[2].

大数据技术为电商的发展注入了新的动力. 电商平台每日能产生数以万计的行为日志, 运用相应的统计方法对这些行为日志进行分析, 可以充分了解用户的需求, 挖掘出用户的购买偏好和消费心理, 进而可以有针对性地制定相应的商品推广或促销策略, 促进平台的转化率, 提高市场竞争力, 用户的体验感也会大幅提升[3].

电商大数据的发展使得信息过载的问题愈发严重, 海量的数据中包含了太多冗余的信息, 干扰了我们对有价值信息的提取, 而基于大数据的个性化推荐系统可以缓解这个问题[4], 它可以从与用户相似度比较高的信息中挖掘出用户的偏好, 从而为用户实施精准推荐. 目前, 大部分电商平台都引入了个性化推荐系统,

如淘宝的首页会向你展现最近浏览过的商品或者关注的商家等, 待收货页面会展现你可能还喜欢的商品.

用户购买预测是个性化推荐的基础, 再加之人工智能发展火热, 将人工智能算法应用到电商领域成为必然, 可以为个性化推荐系统提供优化的空间, 从而为用户提供更优质的服务. 目前, 很多研究人员都将机器学习算法应用到了用户购买预测的研究中, 效果显著. 根据预测的结果, 可以找到一批有购买意愿的目标群体, 然后有针对性地设计相应的推荐算法, 这能为电商平台挖掘出巨大的收益空间[5].

1.2 国内外研究现状

电商迅猛发展, 使得平台的流量越来越多, 但达到一定量后, 会不再增长, 这时候提高平台的流量转化率才是关键, 于是, 个性化推荐的研究成为热点, 通过推荐系统, 可以将用户可能喜好的商品优先呈现在商品首页上, 提高用户使用电商平台的体验感, 缩短购物时间, 也使得用户发生购买行为前的行为次数减少, 提高流量转化率. 用户购买的预测是推荐系统的基础, 在这个研究方向, 许多国内外学者进行了大量的探索, 并取得了有效的收益.

1.2.1 推荐系统的研究现状

传统推荐系统的实践过程中, 涌现出了方方面面的问题, 许多研究学者在这些问题上进行了研究, 在传统的推荐系统的基础上加以改进, 并提出了多种多样的推荐算法.

传统的协同过滤是根据用户历史数据预测其偏好从而进行推荐, 但是用户的偏好是在变化的, 为了提高推荐的准确率, Yang Nihong et al.[6]根据记忆抑制理论, 提出了一种将追溯抑制理论与传统基于物品的协同过滤算法相结合的方法, 以准确探索用户偏好的演变.

为保护用户隐私, 数据管控日趋严格, 获取训练集的难度增加, 联邦推荐系统可以联合多方数据解决该问题, 联邦协同过滤是联邦推荐系统中一个比较经典的算法, 但其冷启动问题一直没有得到解决, 基于此, 王健宗等[7]研究出了安全内积的方法来解决联邦协同过滤的冷启动问题.

协同过滤在数据稀疏性、可扩展性和计算物品与目标用户的相关性等方面一直面临着巨大的挑战, Bansal Saumya[8]提出了一种 Bi-MARS 算法, 在 MovieLens 数据集上与 8 种方法比较, MARS 的精度提高了 66.3%.

1.2.2 用户购买预测的研究现状

用户购买预测的研究主要分为两方面，一方面是样本不均衡处理和特征选择，另一方面是模型的选择.

在数据处理及特征选择方面，段海龙[9]对多数类样本进行 k 均值聚类，根据各个簇中样本数量的比例，将要剔除样本的数量分配到每个簇中，以此来达到样本均衡的目的，有效的避免了随机欠采样方法会导致的信息丢失问题.

在模型构建方面，很多研究人员尝试了单一预测模型，如 Tang et al.[10]通过基于萤火虫算法的 SVM 模型提出了一个具有优化参数的购买行为预测框架，取得了优于 SVM 模型的效果.

但是单一模型在用户行为数据上的特征解释能力弱、准确性也不高，很多研究证明集成学习模型的效果往往优于单一模型，如 Qian Guo et al.[11]基于用户行为数据建立了逻辑回归和 XGBoost 两个模型，经网络数据集的验证，XGBoost 模型的效果更好.

传统的集成学习模型视角比较单一，学习能力往往弱于神经网络，于是一些用户购买行为预测的神经网络模型相继提出: Ling et al.[12]用全连接的长短期记忆网络 (FC-LSTM) 对用户购买行为进行了预测.

现如今，集成学习模型发展迅速，取得了巨大的进步，出现了多个不同模型的组合，包括神经网络间的组合、集成模型间的组合、Stacking 思想等，这种模型包含的基模型丰富多样，吸收了各个模型的优点，从而可以提升模型的性能，提高整体的预测精度，如胡晓丽等[13]提出了一种分段下采样 + CNN-LSTM 组合网络的模型，经电商平台的真实数据集验证，与基准模型相比，该模型的 F_1 值平均提高了 7%~11%; Huibing Zhang[14] 分别建立了 XGBoost、LightGBM 和级联森林的用户购买行为预测模型，然后采用 FCV-堆栈 (FCVS) 方法整合三种预测模型，形成最终的集成学习预测模型，通过数据验证，该方法在精确率、召回率等各项评价指标中都有较好的分数.

1.3 本文组织架构

本文旨在对用户购买行为进行更精准地分析和预测，为精准营销提供高质量的目标群体，选取 2019 京东算法大赛提供的用户、商家、商品等多方面数据进行研究，主要研究内容分为两个方面: 特征工程和模型构建，分成六章来论述.

第一章介绍了所选课题的背景、意义、国内外研究成果以及本文的组织架构.

第二章归纳总结了所选课题会用到的理论知识，包括单一的机器学习算法、

集成学习算法等.

第三章对所选数据进行了探索分析.

第四章构建出了训练集和预测集, 包括训练集涵盖时间范围的确定、特征工程、数据清洗.

第五章进行了模型的构建和选择, 本文尝试多种模型的构建, 包括单一模型、集成学习模型, 模型调优后, 在各个评估指标、运行效率等方面进行比较, 选择一个最优的模型作为最终的模型.

第六章总结了本文的研究内容及成果, 反思了模型存在的问题以及未来的优化空间.

第二章 机器学习理论概述

机器学习是一门涵盖多个领域多个专业的交叉学科, 其致力于研究人工智能, 能够利用计算机工具模拟人类的学习行为, 从过往数据或经验学习其中蕴含的行为规律, 并重新更新现有的知识架构, 实现自动改进算法性能的效果[15]. 近几十年来, 机器学习的研究工作迅猛发展, 在很多领域得到了广泛应用, 电商平台的用户购买预测也不例外, 目前, 传统机器学习的研究方向也拓宽了, 包含单一模型、集成学习和人工神经网络, 并且通过大量经验证明, 在条件合适的情况下, 集成学习算法往往比单一模型算法有更高的预测精度.

本章将介绍本文所选课题涉及到的机器学习算法, 包括传统的单一模型分类算法: Logistic Regression、Decision Tree; 集成算法: Random Forest、GBDT、XGBoost、LightGBM.

2.1 逻辑回归

2.1.1 逻辑回归的思想

逻辑回归 (Logistic Regression, 简称 LR) 算法原理简单, 可以并行实现, 速度快, 且泛化能力比较强. 基本思想: 首先对输入数据做线性回归拟合, 然后将其输出通过 sigmoid 函数映射到[0,1]区间, 实现二分类的作用. 二分类 LR 模型的表示为

$$P(Y=1|x)=\frac{\exp(\omega\bullet x+b)}{1+\exp(\omega\bullet x+b)},$$

其中 $x\in R^n$ 是输入, $Y\in\{0,1\}$ 是输出, $\omega\in R^n$ 和 $b\in R$ 是参数.

若 $P(Y=1|x)>0.5$, 则分为第一类, 否则分为第二类.

2.1.2 模型最优参数的求解

对于给定的训练集 $T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$, 可以用极大似然估计法来求解模型参数[16].

设

$$P(Y=1|x)=\pi(x),\ \ P(Y=0|x)=1-\pi(x),$$

则似然函数为

$$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i},$$

对数似然函数为

$$L(\omega)=\sum_{i=1}^{N}[y_i\log\pi(x_i)+(1-y_i)\log(1-\pi(x_i))],$$

所以求解模型最优参数的问题可以转化为对数似然函数最大化的最优化问题.

2.1.3 求解最优化问题的常用方法

LR 中常用的求解最优参数的方法为梯度下降法. 梯度下降法 (gradient descent) 的基本思想为: 给定目标函数和初始位置, 由于在负梯度方向上目标函数下降最快, 所以每一次更新参数时都在当前位置的基础上, 沿当前位置的负梯度方向走一定的步数, 如此不断走下去, 直到找到目标函数的局部最小值, 其中每一步走的步长即为梯度下降的速率, 称为学习率, 用α表示.

由梯度下降的思想知, LR 每次迭代更新参数的公式为

$$\omega=\omega+\alpha\sum_{i=1}^{N}[(y_i-\pi(x_i))x_i].$$

2.2 决策树

2.2.1 决策树的思想

决策树 (Decision Tree) 模型是基于特征对样本进行分类的树形结构, 树形结构包括根结点、内部结点和叶结点, 根结点包含全部样本集[16]. 在进行分类时, 从根结点开始, 首先选择一个最佳特征和其最优取值作为分裂条件, 然后根据分裂条件将每一个样本分配到对应的内部结点中, 每个内部结点代表某个 (些) 特征取值范围的样本集合, 如此递归的分裂下去, 直到没办法进行分裂或者达到分裂阈值, 最终的结点称为叶结点, 叶结点中大多数样本所属的类为叶结点的取值. 决策树的结构图如图 2-1 所示:

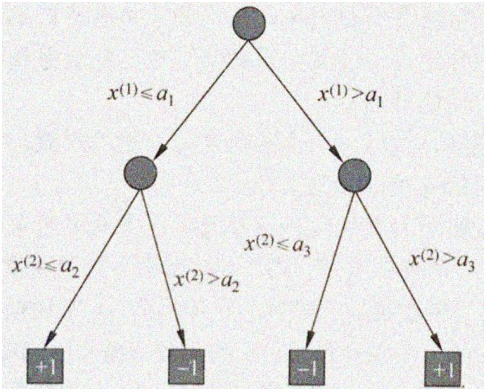


图 2-1 决策树结构图

2.2.2 特征选择准则

怎样选择一个最佳特征和其最优取值作为分裂条件呢？主要有三种准则：信息增益、信息增益比、基尼指数.

1. 信息增益

在信息论中，用信息熵 (information entropy) 来度量信息量的大小. 一个信源含有多少信息量即信息熵可以根据其所有可能发送的符号的平均不确定性来度量[17,18]，计算公式为

$$H = -\sum_{i=1}^{n} p_i \log p_i .$$

其中 $p_1, p_2, ..., p_n$ 为各符号发送的概率，且各符号是否发送之间相互独立，式中对数一般取 2 为底，单位为比特.

把信息熵运用到数据集中，数据集的信息熵[16] $H(D)$ 是对数据集 D 进行分类时的不确定性的度量，熵越大，不确定性越大，设有 K 个类 $C_1, C_2, ..., C_K$，$|C_k|$ 表示 C_k 中的样本数，$|D|$ 表示数据集中的样本数，则

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|},$$

条件熵[16] (conditional entropy) $H(D|A)$ 表示在已知特征 A 的信息下对数据集 D 进行分类的不确定性，设 A 有 d 个取值 $a_1, a_2, ..., a_d$，则

$$H(D|A) = \sum_{i=1}^{d} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{d} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}.$$

信息增益[16] (information gain) $g(D, A)$ 表示在已知特征 A 的信息下，使得对数据集 D 进行分类时不确定性减少的程度，计算公式为

$$g(D, A) = H(D) - H(D|A).$$

2. 信息增益比

在实际决策树生成的过程中, 取值较多的特征的信息增益一般高于取值较少的特征, 这会影响决策树学习的精度, 为了消除特征取值个数带来的影响, 使得在对每一个结点分裂时, 每一个特征的选择是公平的, 引入了信息增益比 (information gain ratio) 这一准则, 其计算公式为

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)},$$

其中,

$$H_A(D)=-\sum_{i=1}^{d}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}.$$

3. 基尼指数

基尼指数 (Gini index) 也是衡量数据集不确定性的一种方法. 从数据集中随机抽取两个样本, 其类别不一致的概率即为基尼指数[19], 基尼指数越小, 数据集纯度越高, 不确定性越小, 基尼指数一般用于二叉树的构建, 其计算公式为

$$Gini(D,A)=\sum_{i=1}^{2}\frac{|D_i|}{|D|}Gini(D_i)=\sum_{i=1}^{2}\frac{|D_i|}{|D|}[1-\sum_{k=1}^{K}(\frac{|D_{ik}|}{|D_i|})^2].$$

2.3 随机森林

随机森林 (Random Forest) 是集成学习中 Bagging 思想的一个经典算法, 它的基学习器为决策树, 基学习器之间没有关联, 可以并行生成, 还结合了 Bootstrap 抽样方法[19], 具体步骤如下:

1. 假设原始数据集有 N 个样本, 从 N 个样本中有放回取样出相同数量的样本作为子训练集;

2. 在每个结点进行分裂时, 从所有特征中随机选取 k 个特征作为候选分裂特征集, 再从中选取最佳特征进行分裂;

3. 重复以上步骤 m 次, 即生成 m 棵决策树, 形成随机森林;

4. 对于一个新样本, 经过每棵树决策, 投票决定分到哪一类.

2.4 梯度提升树

梯度提升树 (Gradient Boosting Decision Tree, GBDT) 是集成学习中 Boosting 思想的算法, 基学习器是 CART 回归树[19], 它实际上是提升树 (Boosting Decision Tree) 的推广, 二者的区别在于损失函数, 提升树的损失函数是特定的, 梯度提升树的损失函数更具一般性, 可以自定义, 它在刚开始被提出时就是公认的泛化能力较强的算法.

通俗理解: 假设我们希望预测一个人的年龄, 他的实际年龄为 30 岁, 首先用一个模型预测为 20 岁, 发现离真实值还有 10 岁差距, 为了不改变原有模型的参数, 于是想到在原有模型的基础上做一些改善来弥补差距, 建立一个新的模型来拟合未完全拟合的部分即这 10 岁的差距, 继续迭代下去, 最终将预测的岁数相加就是模型输出的结果. 具体步骤如下:

1. 初始化模型为一常数, 使得在所有常数中, 为损失最小的值;
2. 每一次迭代时, 计算损失函数关于当前模型的负梯度, 进而可以得到每一个样本当前的负梯度值, 将其作为残差的近似值 r, 这时需要拟合的数据集变为 $\{(x_1,r_1),(x_2,r_2),...,(x_N,r_N)\}$, 拟合该数据集学习一个回归树;
3. 将训练出的回归树加到当前模型上, 更新当前模型;
4. 重复以上步骤 m 次, 输出最终的模型.

2.5 XGBoost

如果数据量很大或者特征个数很多, GBDT 的训练速度很慢, 2015 年, 陈天奇博士提出了 XGBoost (eXtreme Gradient Boosting) 算法[19]对 GBDT 算法做了改进, 大大提升了训练速度. 主要创新点如下:

2.5.1 近似算法

结点分裂时, 传统算法的基本思想为: 遍历每一个特征, 根据特征值的大小对结点中样本进行升序排列, 分别将每一个特征值作为分割阈值, 阈值左边的样本分到左结点, 右边的样本分到右结点; 计算增益, 若增益增大, 则更新目前的分割特征和阈值; 所有特征遍历完后, 输出最终的特征和阈值.

当数据量太大时, 传统算法将数据加载到内存或者分布式环境时会超出, 传统算法将不再适合. 于是 XGBoost 提出了一种近似算法 (Approximate Algorithm) 解决这个问题. 具体思想为: 根据每个特征的分布, 找到候选分割点集合, 在结点分裂时, 遍历分割点集合中的值计算增益即可, 不需要遍历全部样本的值, 加

快训练速度. 这里有两种方式, 一种是 global 的, 每次建树之前找出, 每次结点分裂时都遍历固定的候选分割点; 一种是 local 的, 在每次结点分裂时都重新找候选分割点集合, 所以每次结点分裂时遍历的候选值不一样.

2.5.2 稀疏特征的分裂算法

若样本某个特征缺失, 结点分裂遍历该特征时, 没办法对样本进行左右结点的分配, 基于此, XGBoost 提出了 Sparsity Aware Split Finding 算法, 通过该算法可以自动为样本分配一个使增益最大的分裂方向, 这加快了模型的训练速度. 具体步骤如下:

1. 对于每个特征, 将该特征存在缺失值的样本全部分到右结点, 其余有数值样本的分配方式与传统算法一致, 计算增益. 具体地, 选出特征值不缺失的样本, 根据特征值大小将样本升序排列, 依次将每一个特征值作为分裂阈值, 该阈值左边的样本分到左结点, 其余样本 (包含缺失值样本) 分到右结点, 计算增益.

2. 对于每个特征, 将该特征存在缺失值的样本全部分到左结点, 其余有数值样本的分配方式与传统算法一致, 计算增益. 具体地, 将结点中不含缺失值的样本按照特征值降序排列, 依次将每一个特征值作为分裂阈值, 该阈值左边的样本分到右结点, 其余样本分到左结点, 计算增益.

3. 特征遍历完后, 输出增益最大的特征及其分割点.

2.5.3 工程上的优化

XGBoost 可以在特征粒度上实现并行化, 加快训练速度. 在决策树构建的过程中, 最耗时的步骤是寻找最优分割点, XGBoost 提出了 Block 结构来存储提前按特征值排好序的数据, 节约了时间.

2.5.4 其他方面的优化

1. 正则项: 为防止模型太过复杂使模型将数据中的噪音也学习到, 影响后续模型的预测能力, 故在目标函数中加入正则项;

2. 二阶导数: 对损失函数进行二阶泰勒展开, 相比一阶泰勒展开更能近似损失函数.

3. 列采样: XGBoost 采用了 RF 中的列采样技术, 降低过拟合的风险, 也可以减少计算量;

4. Shrinkage (缩减) : Shrinkage 会将新训练的提升树用一个因子η进行缩放后再加到现在的模型中, 减少每棵树的影响, 使得可以有更大的空间留给后续的模型去学习.

2.6 LightGBM

XGBoost 在很多领域都得到了广泛的应用，并且效果显著，但在数据量大、特征维数多的情况下，仍然存在训练时间长、占用内存大等问题，主要原因是，在寻找最佳分割点时，每个特征都需要扫描一遍全部样本来计算所有分割点的信息增益，效率很低. 基于此，微软亚洲研究院 (MSRA) 于 2017 年提出了名为 LightGBM 的 Boosting 框架[20]，其基本原理与 XGBoost 一样，都是以决策树为基学习器的算法，但是训练速度和内存方面做了优化，LightGBM 主要提出了以下新技术来提升训练速度:

1. 在选择每个特征的候选切分位置时，使用直方图算法代替 XGBoost 的预排序算法构建的数据结构，虽然有损失一定的精度，但大大提高了训练速度，节省了内存空间的消耗.

2. 采用 GOSS 算法对数据进行行采样，移除梯度较小的样本，用其余更有优化空间的样本训练模型，提高了模型拟合的效率.

3. 采用 EFB 算法对特征采样，将互斥特征集合为一个特征，达到降维的目的，降低了选择分割点的时间复杂度.

2.6.1 直方图算法

直方图算法 (Histogram-based Algorithm) 的基本思想如下:

1. 在一个要分裂的结点上，为每一个特征构建直方图. 具体地，先将特征值做分箱处理，然后按照分箱值构造一个直方图. 遍历结点中的每一个样本，在直方图中累积每个 bin 的样本数和样本的梯度之和，当遍历完一次数据后，直方图就累积了需要的统计量，直观的过程如图 2-2 所示:

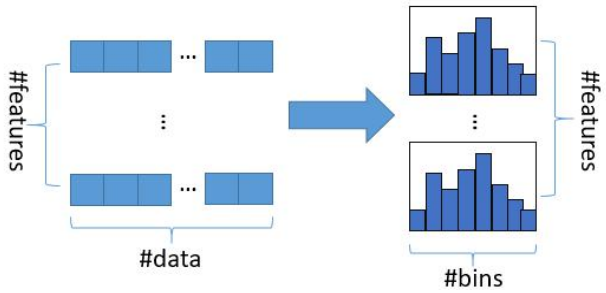


图 2-2 直方图构建过程

2. 对于每个特征，根据构建的直方图，遍历每一个 bin 值寻找最优分裂特征及 bin 值. 具体地，依次将每一个 bin 值作为分割阈值，该 bin 左侧样本分到左结点，右侧样本分到右结点，累加左结点的梯度值和样本数，从而计算出左结点的

直方图, 右结点的直方图用父结点的直方图减去左结点的直方图得到, 基于此计算增益, 若增益增大, 更新叶结点、特征、增益情况.

2.6.2 基于梯度的单边采样

样本的梯度可以反映样本的重要程度, 前文中提到, 样本的梯度可以近似预测值与真实值之间的误差, 梯度越小, 误差越小, 基于此, LightGBM 提出了基于梯度的单边采样 (Gradient-based One-Side Sampling, GOSS) 来减少样本的个数, 进而提高训练速度, 具体步骤如下:

1. 将数据集按照梯度的绝对值降序排列;
2. 选择前 $a\times100\%$ 的样本作为大梯度的样本集合 A;
3. 在剩下的 $(1-a)\times100\%$ 样本中, 随机抽取 $b\times100\%$ 的数据作为小梯度的样本集合 B;
4. 合并样本集合 A、B, 构成新数据集;
5. 计算增益时, 将样本集 B 中的梯度扩大 $\frac{1-a}{b}$ 倍, 来保持数据集分布不变.

2.6.3 互斥特征捆绑

当数据集的特征维数较高时, 这些特征往往是稀疏的, 而在稀疏的特征中, 很多特征的取值不会同时为非 0 值, 称这些特征是互斥的, 如果能够把互斥特征合并为更稠密的特征, 就可以降低特征维数, 提高建直方图的效率, 进而提升训练速度, 基于此, LightGBM 提出了 Exclusive Feature Bundling (EFB) 算法, 这个算法又包括两个步骤, 分别解决了两个问题: 怎样找出互斥的特征去合并、找出互斥的特征后怎样合并.

2.6.3.1 Greedy Bundle

对于第 1 个问题, 通过将特征作为顶点并在每两个特征不互斥的情况下为每条特征添加边, 从而将最优捆绑问题简化为图着色问题. 然后使用贪婪算法, 得到良好的捆绑结果.

具体步骤如下:

1. 构造带权图 G, 具体地, 将每个特征作为顶点, 在每两个特征不互斥的前提下为每个顶点添加边, 边的权重代表两个特征之间冲突的数量, 即两个特征同时不为非 0 值的的样本数;
2. 对每个特征按度降序排序, 记为 feature_sorted;
3. 依次遍历 feature_sorted 中的特征, 在不超过一定冲突量的前提下, 将其加入冲突数最小的 bundle 中, 否则自成一个 bundle.

2.6.3.2 Merge Exclusive Features

对于第 2 个问题, 可以将适合的偏移量加到原始特征的取值上, 使得互斥特征的取值能够分布在不同的 bin 中, 这样即使合并为一个特征, 也能够识别出这个 bundle 中的原始特征, 如一个 bundle 中有两个特征 1 和 2, 特征 1 的取值范围为[0, 20), 特征 2 的取值范围为[0, 40), 若特征 2 的每个取值都加 20, 则特征 2 取值变为[20, 60), 这时就可以合并特征 1 和 2, 新特征的取值为[0, 60).

第三章 用户购买数据的探索分析

在得到一个预测的研究课题时，首先要明确其预测目标是什么，可以抽象为机器学习中的什么问题，其次要对数据集有一个初步的了解，包括数据来源、数据集现有字段描述等，接下来要对数据进行探索分析，帮助我们对数据集有一个全面的认识，包括各字段的缺失值、分布情况等，数据探索分析是建模过程中必不可少的一个环节，可以为之后数据处理和构建特征环节指引方向，因此本章将对数据探索分析的结论做简要说明.

3.1 预测目标

本文希望根据用户的历史数据信息，运用机器学习算法，建立用户购买的预测模型，输出未来 7 天内会发生购买行为的用户、品类、店铺的匹配结果，为精准营销提供高质量的目标群体. 这是一个二分类问题.

3.2 数据介绍

对于任何一个数据集，首先要确保数据的真实性，真实的数据才能刻画真实的分布规律，用真实数据训练出的模型才会有更好的预测结果，其次数据量的大小会影响模型的稳定性. 这两点都需要关注，故本节将对数据的来源、数据的详细情况做一个介绍.

3.2.1 数据来源

本文数据集来自京东 JDATA 算法大赛，在 JDATA 智汇平台上可以找到完整数据集，数据集有 5 个底层表，包含了从 2018-02-01 到 2018-04-15 时间范围内共 37214269 条行为记录、1608707 个用户的基本信息、10399 个店铺的基本信息、352539 种商品的基本信息、1774233 条评论数据，数据来源靠谱，真实可靠且数量庞大，有利于我们进行模型算法的研究，为广大数据挖掘爱好者提供了便利.

3.2.2 数据集变量描述及初步分析

数据集含有的 5 个表分别为行为表、评论表、商品表、店铺表、用户表. 为避免隐私泄露，相关字段已做脱敏处理. 各数据表样例如下所示:

表 3-1 用户行为表数据样例

user_id	sku_id	action_time	module_id	type
937922	357022	2018-02-04 08:28:15	8107857	1
937922	73	2018-02-04 08:27:07	8107857	1
937922	29583	2018-02-04 08:26:31	8107857	1
937922	108763	2018-02-04 08:26:10	8107857	1
1369473	331139	2018-02-03 21:55:49	3712240	1

在用户行为表中, user_id 表示用户 ID, sku_id 表示商品 ID, action_time 表示用户行为时间, module_id 若为购买记录表示订单 ID, 否则表示会话 ID, 三个 ID 字段均进行了脱敏处理, type 表示用户行为类型, 其中 1 为浏览, 2 为购买, 3 为关注, 4 为评论, 5 为加购物车.

本文的预测目标就是要根据用户的历史行为规律来预测用户未来的购买行为, 所以对该表数据的分析是重点.

表 3-2 商品评论表数据样例

dt	sku_id	comments	good_comments	bad_comments
2018-03-15	340522	1	1	0
2018-03-18	340522	1	1	0
2018-04-13	340522	1	1	0
2018-02-07	340522	1	1	0
2018-02-23	340522	1	1	0

在商品评论表中, dt 表示商品评论日期, comments、good_comments、bad_comments 分别表示商品评论数、好评数、差评数, 该表每条记录关于[dt, sku_id]唯一.

商品的评论情况与用户的购买行为的关系初步分析可能为: 在有一定评论量的基础上, 商品好评率越高, 用户越愿意购买该商品, 购买体验越好, 该初步分析也为今后构建特征指引了一个方向.

表 3-3 商品信息表数据样例

sku_id	brand	shop_id	cate	market_time
226519	6302	2399	79	2015-07-02 11:19:04.0
63114	9167	4216	79	2016-07-08 14:29:12.0
372345	2748	7125	79	2016-04-07 16:21:40.0
366931	2698	10252	79	2016-09-11 15:00:22.0
174979	8368	871	79	2017-12-06 17:56:17.0

在商品信息表中, brand 表示商品所属品牌 ID, shop_id 表示店铺 ID, cate 表示商品所属品类 ID, 三个字段均进行了脱敏处理, market_time 表示商品的上市时间.

商品信息表只包含了商品的基本信息, 表中的字段貌似与用户购买行为没有太大的关系, 只是指明了用户购买的商品所属的品类和店铺, 在之后分类问题的预测上用于与用户 ID 做绑定.

表 3-4 店铺信息表数据样例

vender_id	shop_id	fans_num	vip_num	shop_reg_tm	shop_score
3666	5330	0	0	NaN	0.0
7288	1047	762	1017	2014-11-03 10:45:58.0	-1.0
4674	2265	2294	21925	2014-10-17 18:25:58.0	-1.0
5217	3200	9527	53231	2014-10-25 09:17:59.0	-1.0
8057	443	211	336	2014-10-29 10:19:58.0	-1.0

在店铺信息表中, vender_id 表示商家的唯一标识, fans_num 表示店铺的粉丝数, vip_num 表示店铺的 VIP 数, shop_reg_tm 表示开店时间, shop_score 表示店铺评分.

店铺信息的情况与用户购买行为的关系初步猜测为: 店铺的粉丝数和会员数越多, 说明店铺越优质, 用户会更愿意在该店铺消费, 店铺的评分也是用户在购物时会关注的一个指标, 所以店铺信息表中可以在进一步分析下 fans_num、vip_num、shop_score 这三个字段与用户购买的相关强度.

表 3-5 用户信息表数据样例

user_id	age	sex	user_reg_tm	user_lv_cd	city_level	province	city
381581	NaN	NaN	2018-01-21 21:03:37.0	1	NaN	NaN	NaN
478401	NaN	NaN	2009-07-17 17:18:07.0	7	NaN	NaN	NaN
581131	NaN	NaN	2012-12-10 13:02:25.0	7	NaN	NaN	NaN
1154151	NaN	NaN	2014-02-01 16:08:00.0	5	NaN	NaN	NaN
1603505	NaN	NaN	2012-03-03 15:04:36.0	5	NaN	NaN	NaN

在用户信息表中, age 表示用户的年龄段, 取值 1-6, sex 表示用户性别, user_lv_cd 表示会员等级, city_level 表示常用收获地址所在城市等级, province、city 表示用户所在省市编号, user_reg_tm 表示用户的注册时间.

用户信息表的字段与用户购买的关系初步推测为: 年龄维度上, 年轻人相比老年人购物更频繁, 会发生复购的概率更大; 性别维度上, 相比男生, 女生的购物频次更高; 会员等级维度上, 等级越高, 说明用户的消费越多, 那么其再次消费的概率就越高; 城市等级维度上, 等级越高, 由于等级高的城市交通便利, 物流更顺畅, 消费群体的收入相对高, 所以用户购买的次数也会更多. 用户信息表的这几个字段与用户购买的关系可以再继续挖掘一下, 之后都可以做处理后入模.

3.3 数据探索分析

用户在网上购物时, 会产生各种各样的行为操作, 比如浏览、关注等. 不同用户的购买习惯不同, 比如, 有的用户习惯挑选几个意向的商品然后比较择一购买, 这种用户在购买前需要有一定的浏览量, 有的用户可能浏览当天没有购买意愿, 但是会将喜好的商品加入购物车或关注, 一定的时间间隔后才会发生购买行为; 对于不同品类的商品, 用户的购买频率和购买周期也不同, 比如家电类商品, 用户可能购买一次后, 下次购买间隔会很久, 对于生活类用品, 用户可能定期会发生一次购买行为; 部分用户总是在特定的几个店铺内购物.

数据中蕴含着用户的行为规律, 通过探索分析, 加之可视化, 可以更直观的呈现出这种规律, 从而可以更有针对性的、全面多角度的构建特征.

另外, 数据中可能会存在一些冗余数据, 对后续建模会有较大影响, 通过探索分析可以提前发现并提前删除, 避免冗余数据对模型带来的负面影响.

3.3.1 各数据表之间的关系

数据集含有 5 个数据表, 由于数据库中表设计的三大范式, 每个数据表的字段都与主键直接相关, 比如用户行为表的主键是 user_id, 该表中不包含用户购买的商品所属的品类和店铺等信息, 在数据探索时, 势必要了解用户购买的商品的情况, 这时需将用户行为表和商品信息表按照 sku_id 字段关联, 所以事先理清各表之间的关联关系是非常有必要的, 具体的关联关系如图 3-1 所示:

图 3-1 数据表关联图

3.3.2 数据可视化

1. 每日行为分布分析

观察行为分布，可以帮助我们从整体上了解数据分布的稳定性情况，为之后筛选训练/预测集提供思路，也可以反映电商平台的流量程度，于是从用户行为表中统计电商平台每天的行为次数，可视化如图 3-2 所示：

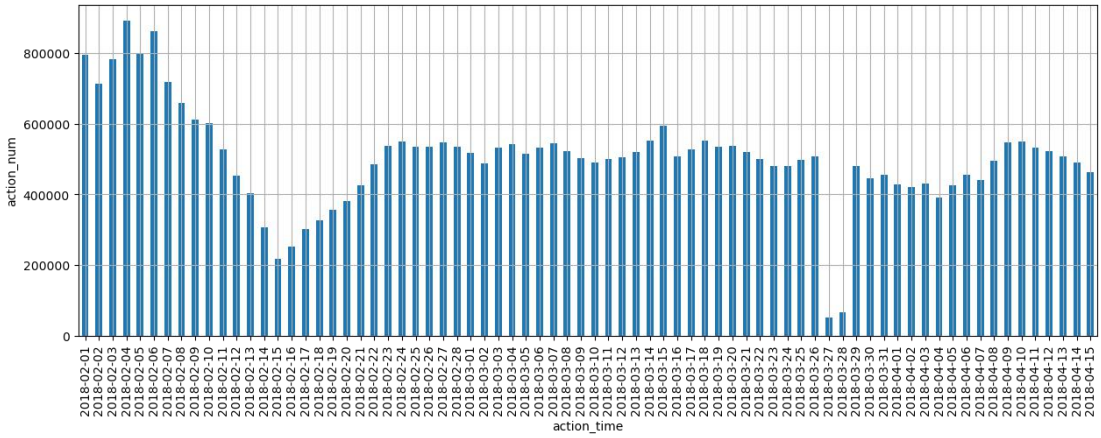


图 3-2 每日行为次数分布图

2018-02-16 为春节，由图 3-2 知，春节前后数据分布不一致. 春节前后几天用户行为数最低，春节前两周用户行为数远高于平常日期，这符合常理，但本文要预测的是平常日期的用户购买行为，为保持数据分布的一致性，构造训练集和预测集的时间范围应选在 2018-02-22 之后.

从图 3-2 还可以知道，03-27 和 03-28 的数据异常，这可能是异常采样的结果，可以再进一步细分探究一下，如图 3-3 至 3-7 所示:

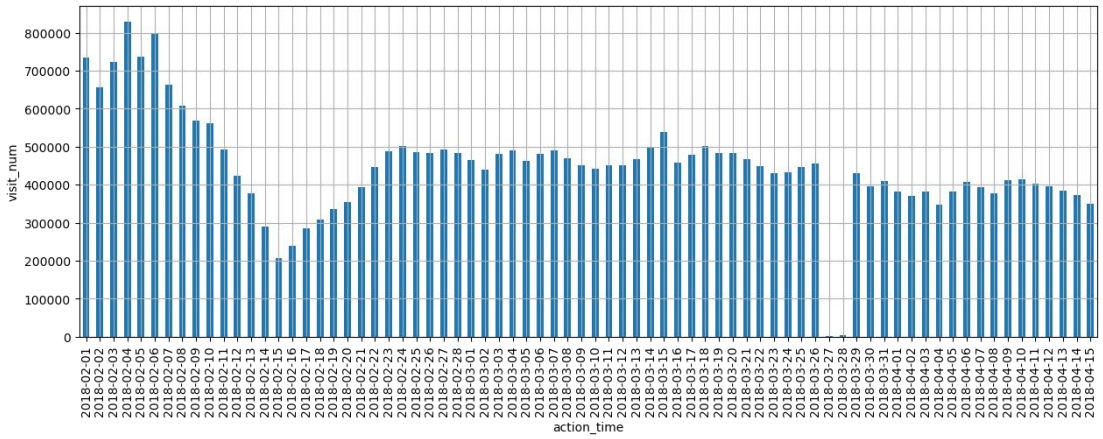


图 3-3 每日浏览次数分布图

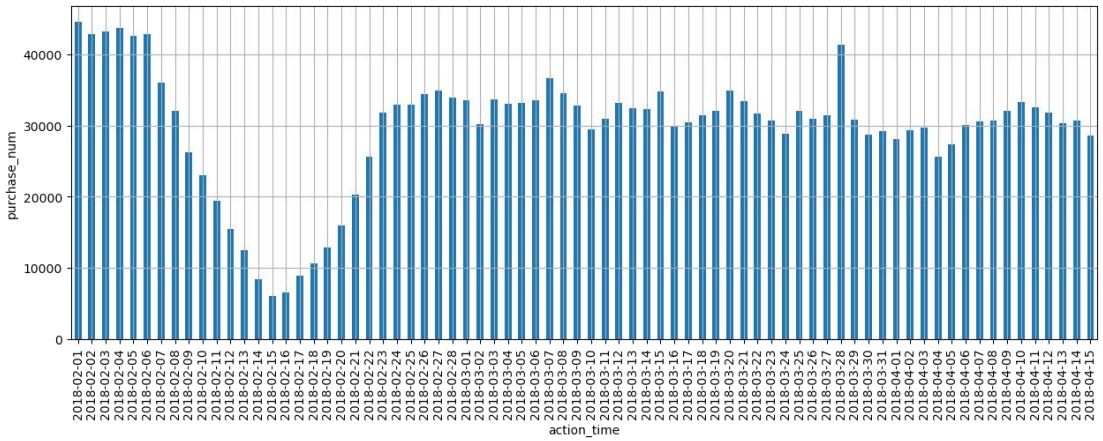


图 3-4 每日购买次数分布图

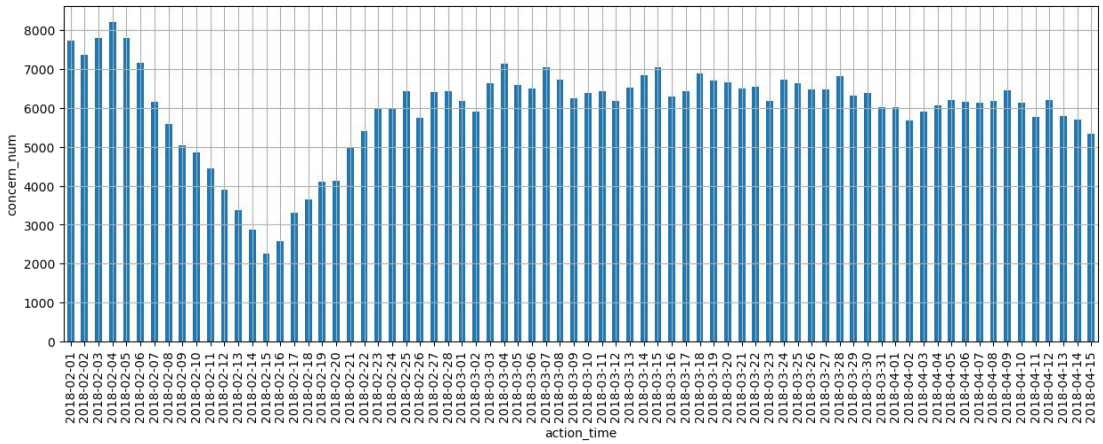


图 3-5 每日关注次数分布图

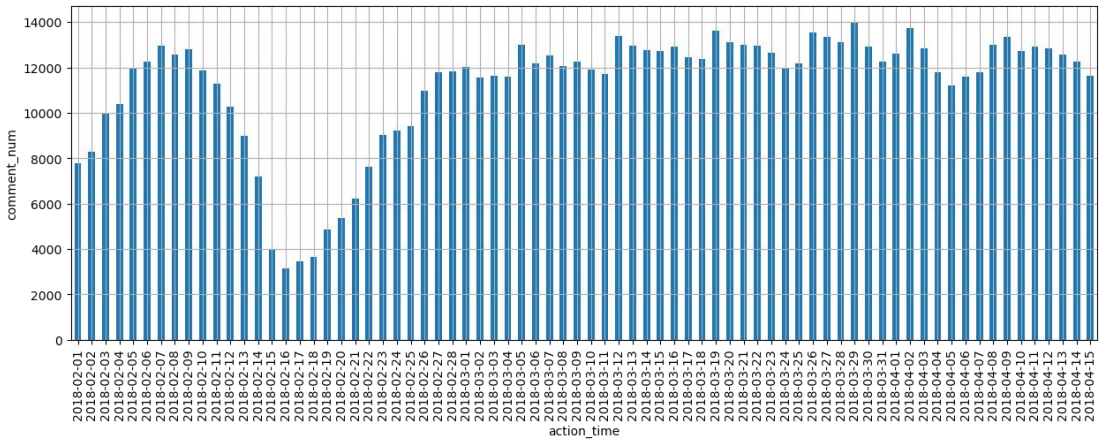


图 3-6 每日评论次数分布图

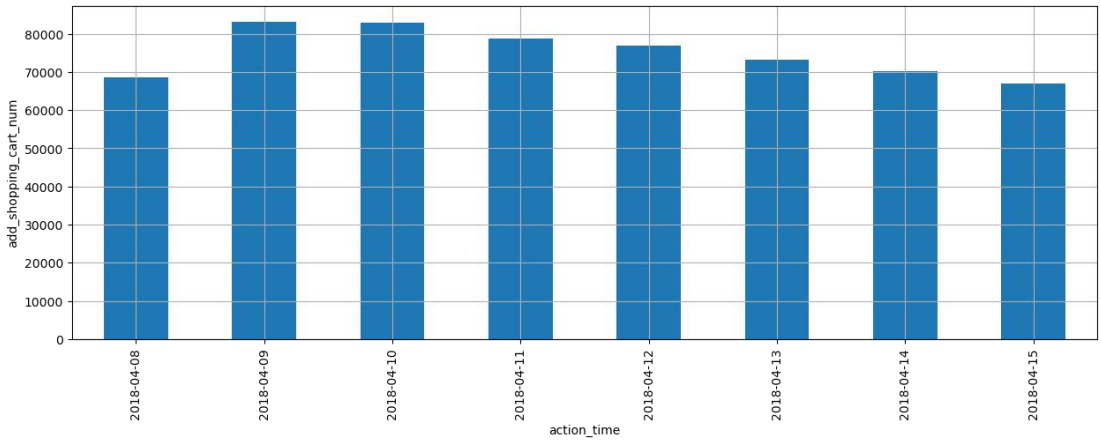


图 3-7 每日加购物车次数分布图

从图 3-3 至 3-7 可以看出, 3 月 27 日和 3 月 28 日数据异常是由于这两天的浏览次数异常导致; 用户加入购物车这一行为只有在 4 月 8 日至 4 月 15 日这 8 天出现, 有可能也是数据采样异常的结果; 用户每日浏览次数 40w~50w, 购买次数 3w 左右, 关注次数 6000 左右, 评论次数 12000 左右, 初步推算, 用户产生 14~17 次其他行为次数, 才会产生一次购买行为.

2. 用户行为分布分析

用户在过去一段时间的行为分布可以反映行为的稀疏度和用户的活跃度, 购买行为分布可以反映用户的购买稠密度, 如图 3-8、3-9 所示:

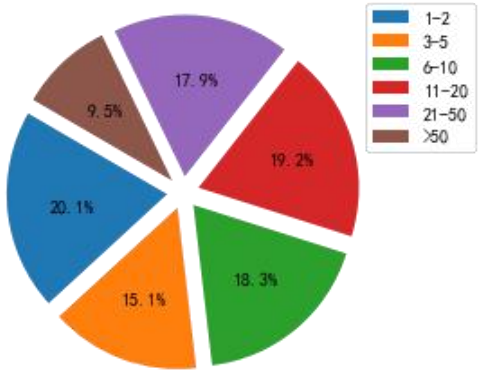


图 3-8 用户行为分布图

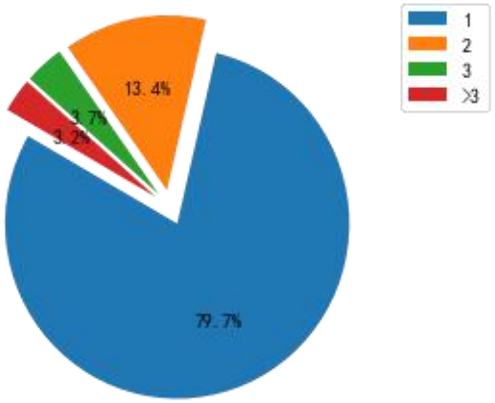


图 3-9 用户购买行为分布图

由图 3-8、3-9 可知，有一半用户的行为次数在 10 次及以下，79.7%的用户只发生一次购买行为，说明大部分用户的活跃度不高，数据比较稀疏，有可能是行为数据只有 2 个多月的原因.

3.3.3 其他探索结论

1. 从店铺信息表中发现，vender_id == 3666 关联了 645 个 shop_id，其他都只关联了 1 个，通过查看这些 shop_id 的行为记录，发现其占全部记录的 38%，且全部是非购买记录，故未来在这些 shop_id 上大概率也不会发生购买行为，在之后构建数据集时可考虑将其删除，不做预测.

2. 用户行为表中存在未在商品信息表中出现的 sku_id. 具体地，用户行为表中有 378457 个 sku_id，商品信息表中有 352539 个 sku_id，且全部能关联上，存在 25918 个 sku_id 关联不上，也就是这部分 sku_id 之后在构建特征时不会有店铺、品类的相关特征，故也不做预测.

3. 用户行为表关联商品信息表时，发现在 cate == 13 上也没有发生购买行为，故未来在 cate == 13 上大概率也不会发生购买行为，在之后构建数据集时考虑将其关联数据删除.

第四章 基于时间滑动窗口的用户购买特征构建

以往预测问题研究的实践证明, 数据的好坏决定着最终模型的预测精度的上限, 数据的好坏也就是构建的特征的好坏, 拿本文研究课题用户购买预测举例, 一组好的特征能够尽可能涵盖购买行为的信息规律的所有可能情况, 也就是如果有一个分布能够衡量用户的行为, 构建的数据集的分布要符合该分布, 这样当模型去学习的时候, 才能将所有购买的行为规律都学到, 从而提升预测精度. 构建的特征好, 即使用简单算法训练的模型, 也能有很高的精度, 构建的特征涵盖的信息少, 那么即使用复杂的模型拟合, 精度也提升不大. 因此, 在建模流程中, 构建特征是最关键的一个环节. 本章将详细介绍构建的特征情况.

4.1 数据处理及特征工程的方法

数据处理的方法如下:

1. 异常值处理

连续型特征会存在数据值异常的情况, 即某些值特别突兀, 会拉偏特征的整体分布, 可以通过长尾截断或画箱线图删除异常值的方法处理, 也可以用中位数或均值等替换异常数据.

2. 缺失数据处理

对缺失值的处理有以下三种情况:

（1） 删除. 可以删除列, 也可以删除行, 当某一个特征大部分样本都缺失时, 该特征可以考虑删除, 若某一个样本其大部分特征都缺失, 该样本可以删除;

（2） 填充. 可以填充均值、众数、中位数等, 也可以根据实际情况, 填充常数值, 比如类别型变量的缺失值可以填充-1, 用于表示特征缺失, 也可以用随机森林等算法根据未缺失样本值预测缺失样本的值来填充.

（3） XGBoost、LightGBM 算法有处理缺失值的机制, 所以可以不做处理.

3. 格式内容错误处理

数据中可能由于人工采样等原因存在数据格式不一致、不统一的情况, 这时候要统一, 比如省份字段, 有山东、山东省两种表示同一省份的取值.

4. 逻辑错误处理

逻辑错误数据指的是某一字段取值与我们的认知不符的情况, 比如年龄字段, 若取值 200 岁或为负数, 则可能该数据存在问题, 考虑对其用均值替代等方法处理.

5. 删除重复数据

按照唯一标识删除数据集中的重复行.

6. 删除唯一特征和与模型训练无关字段

唯一特征指的是样本的唯一标识, 比如 user_id、cate、shop_id 等, 不能刻画数据的分布规律, 故删除; 与模型训练无关字段, 比如 type, 将该字段转化为其它字段后, 该字段就无用了, 在模型训练前要将其删除.

特征工程的方法如下:

1. 若类别型变量的取值具有大小关系, 可以转化为标量, 比如 city_level 等; 若不具有大小关系, 做 one-hot 处理, 比如 sex 等;

2. 交叉特征: 将多个特征融合为一个特征, 若 A 特征的取值为 $\{a_1,a_2\}$, B 特征的取值为 $\{b_1,b_2\}$, 则融合后的特征取值为 $\{(a_1,b_1),(a_1,b_2),(a_2,b_1),(a_2,b_2)\}$, 可以用于两个类别型变量融合为一个类别型变量.

3. 特征分箱, 即连续型变量离散化, 之后做 one-hot 处理或者构建统计特征, 比如 age 可以离散化为类别型变量, 也可以统计不同年龄段下购买的用户数等.

4. 数据变换, 增加的新特征为某一列特征的函数映射, 比如要想某个特征的分布更集中, 可对其进行 log 变换.

5. 统计特征: 按类别型变量的每个类别计算统计特征, 比如统计每个品牌下的商品数、每个年龄段购买各个品类的次数.

6. 特征缩放

当特征值之间的量级相差比较悬殊时, 就可以通过缩放来减少量纲的影响, 比如常见的标准化、归一化等.

7. 特征选择

若特征维数太高, 为防止过拟合, 可以减少特征的数量. 常见的方法有去除低方差特征、通过算法选出重要性比较高的特征、PCA 降维等.

4.2 数据集样本 ID 及标签的确定

本文的预测目标是建立预测的二分类模型, 在某段时间内有过行为记录的 user_id + cate + shop_id 组合 (简称 F_1ID) 中, 找到未来 7 天内会发生购买行为的 F_1ID, 这段时间内有过行为记录的 F_1ID 就是数据集的样本 ID 集合, 其中未来 7 天内有过购买的 F_1ID 的标签为 1, 其余为 0. 那么, 怎么确定 “某段时间” 的时间范围呢?

无疑，如果数据集的样本 ID 集合能极大程度的覆盖未来 7 天内实际发生购买行为的F_1ID，那么选取的样本ID是比较好的，即购买的召回越高越好，形象化的理解如图 4-1 所示:

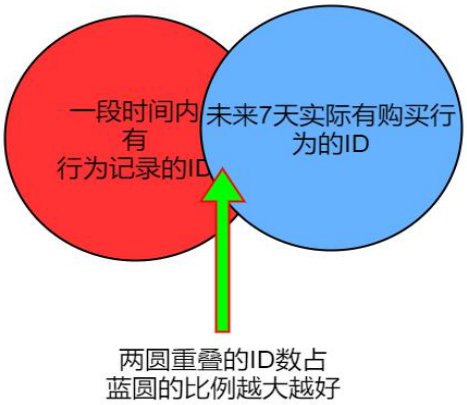


图 4-1　数据集时间范围确定的形象化理解图

为了确定这一时间范围，选取 03-27 至 04-02 这 7 天的购买 F_1ID，观察购买的召回率关于与 3 月 27 日的间隔天数的变化趋势，如图 4-2 所示:

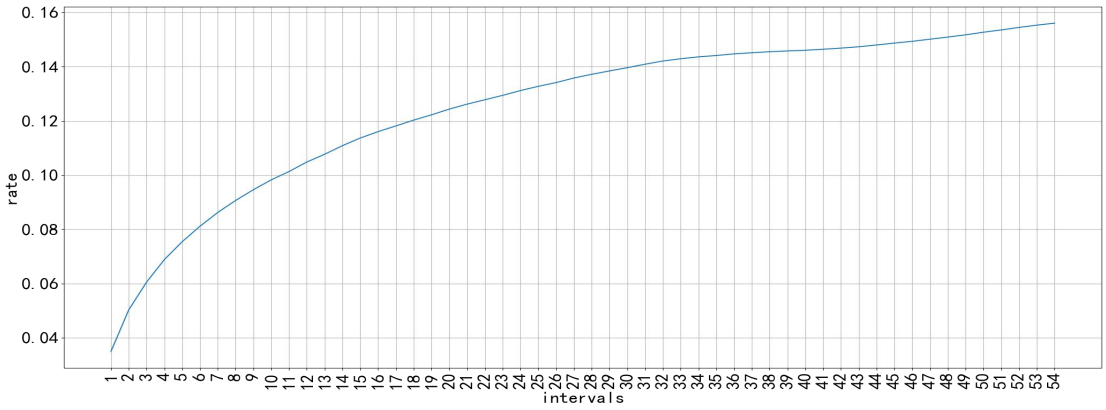


图 4-2　购买召回率曲线图

由图 4-2 知，购买的召回是有上限的，当时间范围为 2018-02-01 至 2018-03-26 时最高，为 15.6%. 经验证，其余 84.4%的 F_1ID 无历史行为记录，在未来 7 天内才有行为记录，可能是临时起意产生购买需求，也可能是数据集涵盖时间短的原因.

由图还知，召回曲线逐渐平缓，历史一周数据的召回为 8.6%，历史两周的召回为 11.1%，增长了 2.5%，历史三周的占比为 12.6%，一周时间仅增长了 1.5%，考虑到时间跨度大会导致数据量多，占用内存大，现有单机系统跑不起来，而时间跨度小覆盖的购买用户占比不够，所以最终选择历史两周的时间范围来筛选数据集的样本 ID.

由第 3 章数据探索分析得到的结论知，加购物车行为只发生在 4 月 8 日之后，vender_id == 3666 的 shop_id 和 cate == 13 无购买行为，各表之间存在未关联上的数据，根据 4.1 节介绍的数据处理方法，为不影响数据分布，数据更加简洁，本文将这部分行为数据当作脏数据删除.

根据以上分析结果确定训练集和预测集的样本 ID 集合及标签如下:

1. 训练集

样本 ID 集合: 2018-03-19 到 2018-04-01 之间去除脏数据后，有行为记录的不重复 F_1ID 集合.

标签: 样本 ID 在 2018-04-02 到 2018-04-08 之间有过购买行为的 F_1ID 标签为 1，其余 ID 标签为 0.

2. 预测集

样本 ID 集合: 2018-03-26 到 2018-04-08 之间去除脏数据后，有行为记录的不重复 F_1ID 集合.

标签: 样本 ID 在 2018-04-09 到 2018-04-15 之间有过购买行为的 F_1ID 标签为 1，其余 ID 标签为 0.

具体时间轴如图 4-3 所示:

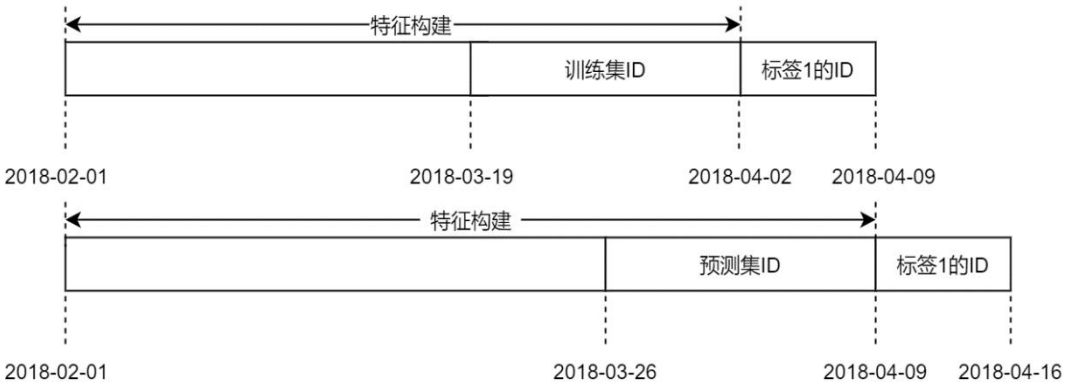


图 4-3 数据集构建时间轴

4.3 样本不均衡的处理

根据 4.2 节，本文得到了训练样本 1923146 个，预测样本 1792209 个. 通过对训练集的购买和未购买样本进行统计，发现比例接近 1:100，样本不均衡，这会影响训练的模型的泛化能力.

在数据层面，常见的处理样本不均衡的方法有: 过采样，通过有放回抽样增加数据集中的少数类样本; 欠采样，剔除部分多数类样本; 增加少数类样本的权

重, 即一个样本当作多个样本使用. 在算法层面, 一些算法的设计机制可以缓解样本不均衡的情况, 比如随机森林.

本文采取欠采样的方法, 从未购买样本中随机抽取量级为购买样本数量 30 倍的样本作为训练集中的未购买样本, 最终, 重新采样后, 训练集样本为 517049 个.

4.4 特征构建

以上训练集和预测集的样本 ID 及标签已确定好, 本节将构建这些样本的特征. 第 3 章数据探索分析为本节构建特征指明了很多方向, 再结合 4.1 节数据处理和特征工程的方法, 本文基于 5 个数据底层表从基本特征、累积特征、时间滑动窗口特征三个方面构建了 564 个特征, 并进行了相应的数据处理, 下面将具体介绍构建的特征及相应的数据处理方式.

4.4.1 基本特征

1. 用户基本特征 (27 个)

来源于用户信息表, 除表中已有特征 age、sex、user_lv_cd、city_level 外, 还将会员注册时间字段量化为了可度量的会员注册时间距今时长特征, 按省份、城市维度统计了用户数, 具体特征描述如表 4-1 所示:

表 4-1 用户基本特征描述

特征字段	特征含义
age	年龄段
sex	性别
user_lv_cd	用户会员等级
city_level	用户所在城市等级
user_reg_days	用户注册时间距今时长
province_user_num	每个省的用户数
city_user_num	每个城市的用户数

数据处理方式: 类别型变量 age、sex、user_lv_cd、city_level 存在缺失值用 -1 填充后做 one-hot, 统计特征每个省份/城市的用户数, 当用户的省份/城市缺失关联不上时, 用-1 填充. 处理完后用户基本特征共 27 个.

2. 品类基本特征 (7 个)

来源于商品评论表和商品信息表, 通过商品的评论特征来统计品类的评论

特征, 由商品信息表统计品类的商品数、品牌数、店铺数, 具体特征描述如表 4-2 所示:

表 4-2 品类基本特征描述

特征字段	特征含义
cate_comments_num	品类评论数
cate_good_comments_num	品类好评数
cate_bad_comments_num	品类差评数
cate_good_comments_rate	品类好评率
cate_product_num	品类下的商品数
cate_brand_num	品类下的品牌数
cate_shop_num	品类下的店铺数

数据处理方式: 若品类下商品的评论特征存在缺失值, 则认为该品类没有被评论, 填充 0.

3. 店铺基本特征 (8 个)

来源于店铺信息表和商品信息表, 包括店铺信息表中原有字段特征: 店铺的粉丝数、会员数店铺评分, 数值计算的特征: VIP 转化率 = 会员数/粉丝数、开店时间距今时间间隔, 根据商品信息表统计的特征: 每个店铺下的商品数、品类数、品牌数, 共构建了 8 个特征, 具体特征描述如表 4-3 所示:

表 4-3 店铺基本特征描述

特征字段	特征含义
fans_num	粉丝数
vip_num	会员数
shop_score	店铺评分
vip_rate	VIP 转化率
shop_reg_days	店铺注册时间距今时长
shop_product_num	商品数
shop_cate_num	品类数
shop_brand_num	品牌数

数据处理方式: shop_reg_days 存在缺失值时, 认为该店铺是新店铺, 填充 0; 店铺的商品数、品类数、品牌数缺失时用-1 填充.

4.4.2 累积特征

累积特征也可以称为长期特征, 刻画的是从数据集的开始日期累积到截止日期的行为规律. 本文从用户、品类、店铺、用户-品类、用户-品类-店铺 5 个维度来构建从 2018-02-01 至数据集最后一天的累积特征.

1. 历史行为统计特征 (92 个)

历史行为统计特征是指行为次数或行为次数之间的运算的特征, 具体特征如图 4-4 所示:

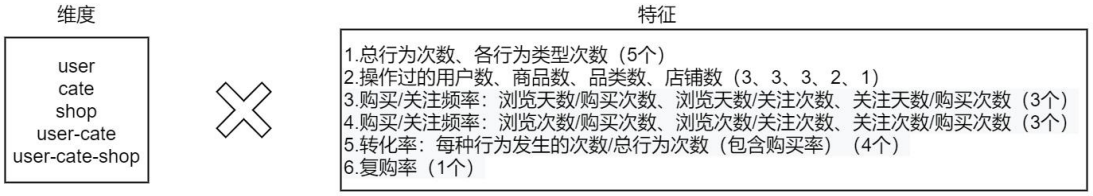


图 4-4 历史行为统计特征

其中复购指的是购买多于 1 次, user、user-cate、user-cate-shop 维度的复购率 = 购买大于 1 次的商品数/购买的商品数, cate、shop 维度的复购率 = 购买大于 1 次的用户数/购买的用户数.

数据处理方式: 维度下未发生的行为类型, 行为次数填充为 0; 当维度下购买/关注行为未发生时, 购买/关注频率产生缺失值, 填充为 0.

2. 交叉特征 (3 个)

这里的交叉特征指的是各维度的特征之间的交叉组合, 具体特征如图 4-5 所示:

图 4-5 交叉特征

3. 时间特征 (85 个)

时间特征是描述各维度的时间规律的特征, 具体特征如图 4-6 所示:

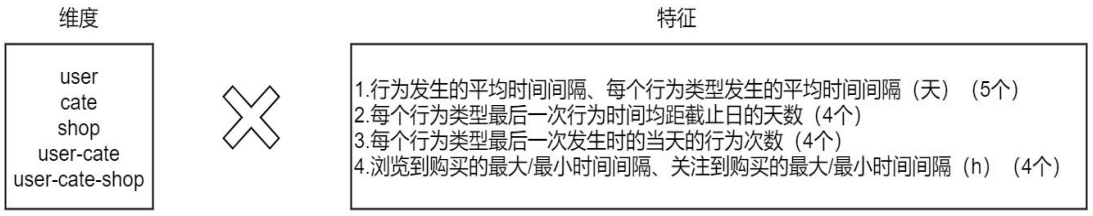


图 4-6 时间特征

其中，平均时间间隔的计算为：以 user 维度为例，设用户产生了 n 次行为，第 $i-1$ 次行为和第 i 次行为的时间间隔为 Δt_i，则平均时间间隔 $\overline{t}=\frac{1}{n-1}\sum_{i=2}^{n}\Delta t_i$.

浏览/关注到购买的最大/最小时间间隔的计算为：计算 user-cate-shop 下的浏览/关注到购买的时间间隔 = 最后一次购买时间 – 第一次浏览/关注的行为时间，按各个维度取最大、最小间隔.

数据处理方式：若维度下只有一条记录或某行为类型没有发生时，平均时间间隔会产生缺失值，认为可能在比 2018-02-01 还早的时间发生过行为，故填充一个比较大的值；若维度下某行为类型没有发生时，最后一次行为时间距截止日的天数也会产生缺失值，由于 2018-02-01 到截止日 2018-04-01/2018-04-08 最长为 67 天，故填充 67；最后一次行为发生时当天的行为次数缺失时，填充 0；若维度下未发生浏览/关注/购买行为，则浏览/关注到购买的最大/最小时间间隔有缺失，认为不经浏览/关注直接购买，故填充 0.

4. 用户维度补充特征 (12 个)

活跃时长(h)：每种行为的最晚时间 – 最早时间 + 1 (4 个)

用户行为频率：每种行为的次数/活跃时长 (4 个)

购买天数；活跃天数；最后一次行为距今时长(天)；最后一次购买距今时长(天) (4 个)

数据处理方式：用户未产生的行为类型，该行为类型的活跃时长、行为频率填充为 0；用户没有产生购买行为时，购买天数填充为 0，最后一次购买距今时长填充为 67.

4.4.3 时间滑动窗口特征

时间滑动窗口指的是一个固定的时间区间，通过滑动窗口可以统计不同时间范围的累积特征. 用户在购买前，可能会对某些品类、店铺有一些行为操作，最近的行为操作对用户未来的行为影响最大，故本文构建基于时间滑动窗口的特征来刻画用户购买的行为规律，滑动窗口逐渐宽泛，共构建了 330 个特征，具体特征如图 4-7 所示：

图 4-7 滑动窗口特征

数据处理方式：由浏览/评论/关注次数为 0 导致的比值特征缺失，填充为 0.

第五章 用户购买预测模型的建立

本文研究课题为预测用户未来 7 天是否发生购买行为, 是一个二分类问题, 以上, 特征已构建完, 并且对特征缺失等情况做了相应的处理, 训练集和预测集准备完成, 本章将用训练集训练模型并对预测集进行预测. 本章选取多个模型训练数据, 通过从各个角度比较选出一个最佳模型.

5.1 模型建立前的准备

在模型建立前, 首先要根据实际场景选择适合的评估模型性能的指标, 其次确定购买召回率的上限, 第三明确建立模型的程序的运行环境.

5.1.1 常用的分类模型性能评估指标

对于二分类问题, 根据样本的真实标签和模型预测出的样本标签, 以本文研究课题为例, 统计出的混淆矩阵如表 5-1 所示:

表 5-1 混淆矩阵

真实标签	预测结果	
	购买	未购买
购买	TP	FN
未购买	FP	TN

其中,

TP: 预测和真实情况都为购买的样本数;

FP: 预测为购买但现实是没有购买的样本数;

FN: 预测为未购买但现实是购买的样本数;

TN: 预测和真实情况都为未购买的样本数.

根据混淆矩阵, 可以得到以下评估指标:

1. 准确率

准确率 (Accuracy) 是指预测结果和真实情况一样的样本数占总样本数的比例, 计算公式如下:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

2. 精确率

精确率 (Precision) 是指在所有预测为购买的样本中预测正确的比例，计算公式如下:

$$precision = \frac{TP}{TP + FP}.$$

3. 召回率

召回率 (Recall) 是指在所有实际购买的样本中预测正确的比例，从“召回”一词形象地说，召回率指从所有正样本中“捞回”的比例，计算公式如下:

$$recall = \frac{TP}{TP + FN}.$$

4. AUC

AUC 是 ROC 曲线与 x 轴、x = 1 所围的图形的面积. ROC 曲线的生成方式为: 通过模型预测出每个样本发生购买的概率，并降序排列，遍历每一个概率值，若大于该值则分为正样本，否则为负样本，每次计算一对假正例率和真正例率作为绘制 ROC 曲线的点，最终将得到的每一对点绘制成阶梯型的折线图或光滑的曲线即得到 ROC 曲线，如图 5-1 所示:

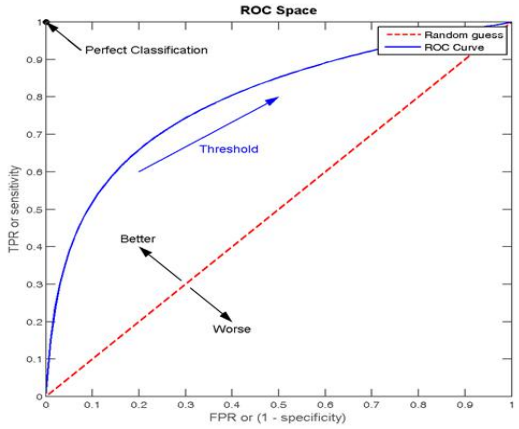


图 5-1 ROC 曲线样例图

其中假正例率 FPR 表示在所有未购买样本中预测为购买的样本数的比例，真正例率 TPR 表示在所有购买样本中预测为购买的样本数的比例，计算公式为

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}, FPR = \frac{FP}{FP + TN} = \frac{FP}{N}.$$

5.1.2 模型评估指标的选择

根据预测用户购买的应用场景，希望尽可能多的召回可能发生购买行为的 ID 组合，从而可以对相应用户进行精准推荐，节省用户购买时间，增加购买成交次数，也要保证预测出的购买 ID 组合有一定的准确率，尽可能的少推荐给用户

无用信息，提高用户使用体验感，故本文综合考虑召回率和精确率两个评估指标，将模型评估指标定为

$$F_1 = \frac{5*recall*precision}{2*recall+3*precision}.$$

这里的*recall*指的是未来实际购买的召回率，等价于模型的召回率与购买召回率上限的乘积.

由 4.2 节，本文得到 1792209 个预测样本，其中标签 1 的样本有 19227 个，而未来 7 天实际发生购买的样本有 199680 个，因此召回的上限为 19227/199680 = 9.63%.

5.1.3 程序运行环境及配置

本文是用 Python 语言来实现所有建模过程的，程序运行的配置条件如表 5-2 所示:

表 5-2 程序运行配置表

软硬件条件	配置情况
操作系统	Windows10
开发平台	Jupyter Notebook
开发语言	Python3. 8
CPU	AMD R7 4800H
内存	16G

5. 2 模型建立

本文在第 4 章构建了大小为 517049*565 的训练集和大小为 1792209*565 的预测集，现基于此数据集进行模型训练和预测，本节通过超参数的调整、多个模型的比较，确定最优模型的框架.

5.2.1 模型超参数的设置

模型超参数的不同，模型的表现也是不同的，为了找到性能最高的模型，需要对超参数调优，最终，各模型的部分超参数选择如表 5-3 所示:

表 5-3 各模型部分超参数设置表

模型名称	超参数含义	超参数	参数取值
LR	正则项系数	C	0.1
	正则化项	penalty	L_2
RF	基学习器数量	n_estimators	250
	结点分裂准则	criterion	entropy
	树的深度	max_depth	5
	叶结点最小样本数	min_samples_leaf	20
GBDT	基学习器数量	n_estimators	50
	树的深度	max_depth	6
	列采样比例	subsample	0.8
	分裂的最小样本数	min_samples_split	900
XGBoost	树的深度	max_depth	6
	叶子节点最小样本权重和	min_child_weight	5
	学习率	eta	0.01
	列采样比例	subsample	0.8
	行采样比例	colsample_bytree	0.8
	迭代次数	num_boost_round	3000
	早停轮数	early_stopping_rounds	100
LightGBM	叶结点的数量	num_leaves	31
	叶结点最小样本数	min_data_in_leaf	20
	学习率	learning_rate	0.01
	列采样比例	feature_fraction	0.8
	行采样比例	bagging_fraction	0.8
	迭代次数	num_boost_round	5000
	早停轮数	early_stopping_rounds	100

5.2.2 模型结果对比分析

各模型的最优情况在不同指标上的得分如表 5-4 所示:

表 5-4 各模型比较结果

算法名称	预测购买量	召回率	精确率	F_1	AUC	运行时间
LR	130000	0.6278	0.0928	0.07025	0.8829	0:00:35.37
RF	130000	0.6315	0.0934	0.07067	0.8830	0:04:20.19
GBDT	110000	0.6197	0.1083	0.07274	0.8936	0:19:15.22
XGBoost	110000	0.6320	0.1105	0.07418	0.9017	0:23:40.24
LightGBM	110000	0.6349	0.1110	0.07453	0.9023	0:02:20.92

由表 5-4 知, LightGBM 在各个指标上均优于其他模型, 因此本文选取 LightGBM 作为最终的模型.

第六章 总结与展望

6.1 总结

用户购买行为预测是商品推荐系统研究的重点问题之一，提前预测出用户的购买倾向会大大提高推荐算法的效果，这是一项极具意义的工作.

本文对用户购买的预测的流程如下:

1. 数据探索分析，对数据有一个充分的认知，对之后数据构建和数据处理打好基础，挖掘用户购买习惯和行为规律，为构建特征提供方向.

2. 基于时间滑动窗口的特征构建，首先确定了训练和预测样本，其中训练集存在样本不均衡问题，进行了负采样处理；其次在用户、品类、店铺、用户-品类、用户-品类-店铺 5 个维度上分别构建了基本特征、累积特征、时间滑动窗口特征，在特征生成过程中一并对特征缺失等情况做了处理，最终构建了大小为 517049*565 的训练集和大小为 1792209*565 的预测集.

3. 模型建立，首先明确模型的评估指标和程序运行的环境及配置，其次运用了逻辑回归、随机森林、GBDT、XGBoost、LightGBM 进行预测，通过比较分析选出最优模型.

6.2 展望

本文从多个维度构建特征，为训练模型提供了充分的信息，研究多种机器学习的算法在该课题上的效果，得到了一些有益结论，但仍有一些想法没有来得及探索和实现，也存在一些技术问题需要攻破.

1. 本文没有对行为时间 action_time 做相应的特征，之后可以对 action_time 进行分段处理，比如分为上午、下午、晚上，再对其进行特征统计，看是否会进一步提升模型精度.

2. 由于数据中加购物车的行为只在 4 月 8 日之后出现，所以本文在构建数据集时将加购物车的数据当作脏数据删除了，之后可以在加购物车这一维度上探索分析挖掘一下特征.

3. 本文出于程序运行内存的限制和时间开销的考量，数据集时间范围选择了两周，之后可考虑在一些开源平台，比如阿里云等内存大的平台上运行程序，这时可以尝试数据集时间范围放到一个月，特征也可以再丰富些.

4. 模型的超参数只进行了部分调试，之后可以再尝试其他超参数的组合.

5. 为了提升模型精度, 可以再建立一个预测 5 天购买的模型, 7 天模型与 5 天模型的融合结果作为最终的输出结果.

参考文献

[1] 鞠雪楠, 欧阳日辉. 中国电子商务发展二十年: 阶段划分、典型特征与趋势研判[J]. 新经济导刊, 2019, (03):26-33.
[2] 苏鸣立. 1997-2019: 电商 22 周年发展历程及未来[J]. 计算机与网络, 2019, 45(19):8-10.
[3] 谭超颖. 大数据技术在电子商务系统中的应用研究[J]. 无线互联科技, 2020, 17(23):113-114.
[4] 任敏. 大数据个性化推荐分析[J]. 物联网技术, 2019, 9(11):62-64+67.
[5] Hua Chen. Personalized recommendation system of e-commerce based on big data analysis[J]. Journal of Interdisciplinary Mathematics, 2018, 21(5):1243-1247.
[6] Yang Nihong, Chen Lei, Yuan Yuyu. An Improved Collaborative Filtering Recommendation Algorithm Based on Retroactive Inhibition Theory[J]. Applied Sciences, 2021, 11(2):843.
[7] 王健宗, 肖京, 朱星华, 李泽远. 联邦推荐系统的协同过滤冷启动解决方法[J]. 智能系统学报, 2021:1-9.
[8] Bansal Saumya, Baliyan Niyati. Bi-MARS: A Bi-clustering based Memetic Algorithm for Recommender Systems[J]. Applied Soft Computing, 2020, 97(PA) .
[9] 段海龙. 数据平衡与模型融合的用户购买行为预测研究[D]. 南昌大学, 2020.
[10] L. Tang, A. Wang, Z. Xu, et al. Online-purchasing behavior forecasting with a firefly algorithm-based SVM model considering shopping cart use[J]. Eurasia J Math Sci Technol Educ, 2017, 13(12):7967–7983.
[11] Qian Guo, Chun Yang, Shaoqing Tian. Prediction of Purchase Intention among E-Commerce Platform Users Based on Big Data Analysis[J]. IIETA, 2020, 34(1):95-100.
[12] C. Ling, T. Zhang, Y. Chen. Customer purchase intent prediction under online multi-channel promotion: a feature-combined deep learning framework[J]. IEEE, 2019, 7(8):112963-112976.
[13] 胡晓丽, 张会兵, 董俊超, 吴冬强. 基于 CNN-LSTM 的用户购买行为预测模型[J]. 计算机应用与软件, 2020, 37(06):59-64.
[14] Huibing Zhang, Junchao Dong. Application of sample balance-based multi-perspective feature ensemble learning for prediction of user purchasing behaviors on mobile wireless network platforms[J]. EURASIP Journal on Wireless Communications and Networking, 2020, 2020(1):148-154.
[15] 耿凌霄. 基于 MKSVM 的发酵过程动态建模方法研究及其应用[D]. 北京工业大学, 2014.
[16] 李航. 统计学习方法 (第 2 版) [M]. 北京:清华大学出版社, 2019.
[17] 梁栋, 张兴. 信息论简明教程[M]. 北京:北京邮电大学出版社, 2009.
[18] 李梅. 信息论基础教程[M]. 北京:北京邮电大学出版社, 2005.
[19] 周志华. 机器学习 (第 1 版) [M]. 北京:清华大学出版社, 2016.
[20] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System, 2016.
[21] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Weichen, Weidong Ma, Qi Wei Ye, Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 2017.
[22] 付红玉. 基于异源集成算法的用户购买行为预测研究[D]. 山东大学, 2020.

[23] 张建彬, 霍佳震. 基于 Stacking 模型融合的用户购买行为预测研究[J]. 上海管理科学, 2021, 43(01):12-19.
[24] 吴非. 基于特征工程的用户购买预测模型研究[D]. 长安大学, 2019.
[25] 蔡一凡. 基于用户聚类和特征选择的在线购买行为预测研究[D]. 华中科技大学, 2019.
[26] Li Q, Gu M, Zhou K, et al. Multi-Classes Feature Engineering with Sliding Window for Purchase Prediction in Mobile Commerce[C]. IEEE, 2015: 1048-1054.
[27] Jing Zhou, Wei Li, Jiaxin Wang, Shuai Ding, Chengyi Xia. Default prediction in P2P lending from high-dimensional data based on machine learning[J]. Physica A:Statistical Mechanics and its Applications, 2019, 534.
[28] 慕钢, 张宏烈, 党佳俊, 李广峰. 基于 LightGBM 模型的二手房房价预测研究[J]. 高师理科学刊, 2020, 40(12):27-31.
[29] 方婷婷, 李全, 秦明远. 基于 LightGBM 的企业信用评估预测[J]. 信息技术与信息化, 2020, (12):17-19.
[30] 王克利, 邓飞其. 基于阿里巴巴大数据重复购买预测的实证研究[J]. 时代金融, 2018, (1): 237-239.
[31] 张李义, 李一然, 文璇. 新消费者重复购买意向预测研究[J]. 数据分析与知识发现, 2018, 2(11):10-18.
[32] 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.
[33] 李美其, 齐佳音. 基于购买行为及评论行为的用户购买预测研究[J]. 北京邮电大学学报(社会科学版), 2016, 18(04):18-25.

致 谢

时光飞逝，转眼间三年的研究生生涯已接近尾声，这三年，有心酸、有挫折，但更多的是收获与成长，在这里，我要向曾经帮助过我、关心过我的人表达最真挚的感谢.

首先，我要感谢我的导师焦桂梅老师，焦桂梅老师为人亲切，待人诚恳，读研期间，在学习上为我指引方向，在生活上对我给予了很大的帮助和鼓励，在实践上支持我，使我不断地进步，非常有幸能够遇到焦桂梅老师.

感谢我的父母，不论何时何地，都给予我温暖的关心，是我人生道路上最坚实的后盾.

感谢我的朋友们，在我低谷时能够听我倾诉，鼓励我走出阴霾，幸好有你们.

感谢 2018 级应用统计全体同学，相遇就是缘分，希望大家都越来越好.

感谢实习过程中遇见的所有人，从他们身上学习到了很多.

感谢兰州大学能够给我带来一份宝贵的校园回忆，我会带着这份回忆继续前行.