

**实验报告册**

|  |  |
| --- | --- |
| 学年学期： | 2023 -2024学年 第1学期 |
| 课程名称： | 数据工程综合实践 |
| 学生学院： | 计算机科学与技术学院/人工智能学院 |
| 专业班级： | 04012104 |
| 学生学号： | 2021212151 |
| 学生姓名： | 周昭光 |
| 联系电话： | 15826301649 |

**重庆邮电大学教务处制**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学生姓名 | 周昭光 | 学生学号 | 2021212151 | 专业班级 | 04012104 |
| 校内指导教师 | 王进 | | | | |
| 实验日期 | 2023.12 | 实验地点 | 综合实验楼B501/B502 | | |
| 实验名称 | 数据工程综合实践-任务二 | | | | |
| 评阅人签字 | 王进 | | | 成绩 |  |
| **一、实验目的和要求** | | | | | |
| 完成O2O 优惠券使用预测阿里天池平台线上测评 | | | | | |
| **二、实验设备** | | | | | |
| 1、操作系统：Windows  2、基本硬件配置要求：处理器Intel i3、内存8G；  3、Python及其组件版本：  Python 3.6.5  NumPy 1.14.3  Pandas 0.23.0  Scikit-learn 0.19.1  XGBoost 0.6  LightGBM 2.1.0 | | | | | |
| **三、实验内容** | | | | | |
| 1. **赛题简介**   **1.1问题背景**  随着移动设备的普及和移动互联网的发展，O2O（Online to Offline）消费领域备受关注。据不完全统计，O2O行业估值上亿的创业公司至少有10家，也不乏百亿级别的巨头。O2O行业涉及数亿消费者，各类APP每天记录了超过百亿条用户行为和位置记录，形成了庞大的数据资源，成为大数据科研和商业化运营的重要场景之一。以优惠券盘活老用户或吸引新客户进店消费是O2O的一种重要营销方式。然而随机投放的优惠券对多数用户造成无意义的干扰。对商家而言，滥发的优惠券可能降低品牌声誉，同时难以估算营销成本。个性化投放是提高优惠券核销率的重要技术，它可以让具有一定偏好的消费者得到真正的实惠，同时赋予商家更强的营销能力。  **1.2目标**  本次比赛的目标是针对O2O场景，通过分析建模，精准预测用户是否会在规定时间内使用相应的优惠券。   1. **数据集介绍**   **2.1数据来源**  本赛题提供用户在2016年1月1日至2016年6月30日之间真实线上线下消费行为（数据均已作匿名处理、有偏采样和必要过滤），预测用户在2016年7月领取优惠券后15天以内的使用情况。  **2.2数据描述**  本次比赛数据共有四张表，两张训练数据表，一张测试数据表，一张提交样例表。  表字段如下：  Table 1: 用户线下消费和优惠券领取行为   |  |  | | --- | --- | | **Field** | **Description** | | User\_id | 用户ID | | Merchant\_id | 商户ID | | Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 | | Discount\_rate | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元 | | Distance | user经常活动的地点离该merchant的最近门店距离是x\*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里； | | Date\_received | 领取优惠券日期 | | Date | 消费日期：如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用，即负样本；如果Date!=null & Coupon\_id = null，则表示普通消费日期；如果Date!=null & Coupon\_id != null，则表示用优惠券消费日期，即正样本； |   Table 2: 用户线上点击/消费和优惠券领取行为   |  |  | | --- | --- | | **Field** | **Description** | | User\_id | 用户ID | | Merchant\_id | 商户ID | | Action | 0 点击， 1购买，2领取优惠券 | | Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 | | Discount\_rate | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元 | | Date\_received | 领取优惠券日期 | | Date | 消费日期：如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用，即负样本；如果Date!=null & Coupon\_id = null，则表示普通消费日期；如果Date!=null & Coupon\_id != null，则表示用优惠券消费日期，即正样本； |   Table 3：用户O2O线下优惠券使用预测样本   |  |  | | --- | --- | | **Field** | **Description** | | User\_id | 用户ID | | Merchant\_id | 商户ID | | Coupon\_id | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 | | Discount\_rate | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元 | | Distance | user经常活动的地点离该merchant的最近门店距离是x\*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里； | | Date\_received | 领取优惠券日期 |   Table 4：选手提交文件字段  其中user\_id,coupon\_id和date\_received均来自Table 3,而Probability为预测值   |  |  | | --- | --- | | **Field** | **Description** | | User\_id | 用户ID | | Merchant\_id | 商户ID | | Date\_received | 领取优惠券日期 | | Probability | 15天内用券概率，由参赛选手给出 |  1. **数据预处理** 2. 复制原始数据，以免修改原始数据 3. 处理 Discount\_rate 列     - 添加新列 'is\_off'，如果 Discount\_rate 中包含 ':' 则设置为 1，否则设置为 0  - 添加新列 'min\_cost'，若 Discount\_rate 中包含 ':' 则提取出最低消费金额，否则置0  - 将 Discount\_rate 转换为折扣率，如果 Discount\_rate 中包含 ':'，则计算折扣率   1. 处理 Distance 列     - 计算 Distance 列的均值，并用均值填充空值   1. 处理 Date\_received 和 Date 列     - 将这两列转换为日期格式，格式为 '%Y%m%d'   1. **数据划分**   数据划分过程包括以下步骤：     1. 定义了三个时间段，分别用于训练、验证和测试数据集 2. 根据时间段将原始数据集划分为训练、验证和测试特征数据集 3. 如果原始数据集包含 'Distance' 列，则额外创建训练和验证目标数据集，目标数据集是根据 'Date\_received' 列在特定时间段内的数据 4. 返回划分后的训练特征数据集、训练目标数据集、验证特征数据集、验证目标数据集和测试特征数据集   数据划分采用的区间：     1. **特征工程** 2. 获取主要特征 3. 各商家 用户购物数      1. 各商家 用户数      1. 各商家 用户用卷数      1. 各商家 发卷数      1. 各商家 用户用卷消费折扣均值      1. 各商家 用户用卷消费折扣最小值      1. 各商家 用户用卷消费折扣最大值      1. 各商家 用户用卷消费距离均值      1. 各商家 用户用卷消费距离最小值      1. 各商家 用户用卷消费距离最大值      1. 各优惠卷 使用次数      1. 各优惠卷 发卷数      1. 各用户 消费数      1. 各用户 领卷数      1. 各用户 领卷消费数      1. 获取趋势度量特征     主要步骤：   1. 从给定数据中获取使用优惠券消费的样本 2. 计算被使用优惠券的折扣率的均值、最小值、最大值和中位数 3. 计算用户使用优惠券消费的距离的均值、中位数、最小值和最大值 4. 将上述计算结果存储在一个字典中 5. 返回包含趋势度量特征的字典 6. 获取日期特征      1. 获取线上特征   1. 复制给定数据  2. 提取用户、商家和优惠券特征  3. 提取用户线上特征：  - 计算每个用户的线上消费次数    - 计算每个用户的线上使用优惠券消费次数    - 计算每个用户线上领取优惠券次数    4. 将提取的用户线上特征与用户特征合并，并将其存储在一个数据集中。  5. 将需要归一化处理的列添加到列表中    6. 返回用户特征、商家特征和优惠券特征   1. **构造特征数据集**   通过整合不同来源的特征，并对数据进行清洗和处理，生成用于训练模型的特征数据集:   1. 调用函数提取主要特征、线上特征、趋势度量特征和日期特征      1. 获取样本集合      1. 根据样本集合中的用户ID、商家ID和优惠券ID，从主要特征中筛选出对应的商家特征、优惠券特征和用户特征，并将它们合并到数据集中      1. 根据样本集合中的用户ID，从线上特征中筛选出对应的用户线上特征，并将其合并到数据集中      1. 将趋势度量特征的值逐一添加到数据集中      1. 删除无用的属性列      1. 用均值填补数据集中的空值      1. 将数据集中的Coupon\_id和Distance列转换为整数类型      1. 将数据集中的Date\_received列转换为整数类型，并按照指定的格式进行格式化 2. 去除数据集中的重复样本，并重新设置索引      1. **模型训练与评估**   **6.1模型选择**  XGBoost是一种强大而灵活的机器学习工具，适用于各种数据挖掘和分析任务，并且能够提供高性能和准确度的预测结果，故本次实验采用次模型。  **6.2模型训练**  1. 数据预处理  从训练和测试数据中分离出特征和目标变量，并对特征进行z分数规范化    2. 模型初始化和参数设置  使用`xgb.DMatrix`将数据转换为XGBoost的数据格式    设置XGBoost模型的参数    3. 模型训练  使用`xgb.train`方法训练XGBoost模型，指定训练数据、迭代次数、监视列表等    4. 模型预测  使用训练好的模型对测试数据进行预测，得到预测概率  5. 结果处理  将预测结果和测试数据中的用户ID、优惠券ID、领取日期等信息合并    6. 特征权重计算  使用`model.get\_score()`获取特征的重要性，将其存储在DataFrame中    **6.3模型评估**  阿里云天池账号：    历史提交截图，最高得分为0.7466        打印出的特征： | | | | | |
| **四、学习心得** | | | | | |
| 参与天池新人实战赛的经历让我对数据挖掘有了更深入的理解，也让我体会到了参与比赛的乐趣和收获。本次比赛的主题是《生活大实惠：O2O优惠券使用预测》，是一个典型的数据挖掘问题，要求通过分析建模，精准预测用户是否会在规定时间内使用相应优惠券。  首先，我完成了一系列学习内容和要求，包括Python机器学习库的安装、机器学习基本概念、基础和中级Python编程、第三方Python库等。这些学习内容为我提供了必要的工具和知识基础，让我能够更好地进行后续的数据分析和建模工作。  在学习过程中，我特别关注了O2O优惠券使用预测算法的基本框架和项目解读。通过理解项目背景和目标，我对问题的核心思想有了更清晰的认识，也对如何处理数据、构建模型有了初步的想法。同时，我还学习了数据分析与预处理方法、数据划分与打标方法等实用技能，这些技能对于处理真实数据具有重要意义。  在实际比赛中，我首先对数据进行了初步的探索和分析，了解了数据的特征和分布情况。接着，我根据学习到的算法和技巧构建了多个预测模型，并对模型进行了调参和优化。在模型训练和评估过程中，我不断尝试不同的方法和策略，提高模型的准确性和泛化能力。  通过参与天池新人实战赛，我收获了很多。首先，我对数据挖掘和机器学习的理论知识有了更深入的理解，学会了如何将理论知识应用到实际问题中。其次，我提高了数据分析和建模的能力，学会了如何处理真实数据、构建预测模型。最重要的是，我体会到了团队合作的重要性，通过与其他参赛选手的交流和学习，我不断完善自己的方法和思路，取得了比赛的进步和成绩。  总的来说，参与天池新人实战赛是一次宝贵的学习和锻炼机会，我收获了知识、技能和经验，也激发了我对数据科学和机器学习的兴趣。我会继续努力学习和实践，不断提升自己在数据挖掘领域的能力，为未来的学习和工作打下坚实的基础。 | | | | | |
| **五、教师评阅意见** | | | | | |
|  | | | | | |