一、数据假设

根据题目介绍，用户分析的数据来源与两个表格，即已知的商家名单文件 Data_Seller_Product.csv 和半年时间内每一所有买家的行为数据文件 Data_Action_20160101.csv（假设具体日期为 2016 年 1 月 1 日至 2016 年 6 月 30 日），最终结果文件保存在 Data_Seller_Product_User.csv 文件中，各文件内数据格式如下：

（1）、Data_Seller_Product.csv

seller_id	商家 ID
product_id	商品 ID

说明：一个商家 ID（seller_id）可以对应多种商品 ID（product_id）。

（2）、Data_Action_20160101.csv

user_id	用户 ID
product_id	商品 ID
s_num	浏览数量
c_num	收藏数量
r_num	推荐数量
b_num	购买数量
Date	日期

（3）、Data_Seller_Product_User.csv

seller_id	商家 ID
product_id	商品 ID
user_id	用户 ID

说明：推荐结果中 product_id-user_id 数据对是唯一的，会为每种商品（product_id）推荐 20 个（user_id）潜在的消费者。

二、解决方案

1、概述

1.1、问题分析

为商家推荐一定数量的用户，便于其进行针对性的营销活动。可以根据过去半年内的历史行为数据，推测用户在接下来一段时间（取 10 天）内购买某个商品的可能性大小，选取可能性前 N（20）的用户进行推荐，对每个商家的每种商

品的推荐用户进行汇总，得到每个商家所需要的推荐用户名单。本方案的目标是根据用户对商品历史行为数据来预测用户在未来一段时间（10 天）内购买该商品的概率。

问题转换成 2 分类问题，从用户行为数据表（Data_Action_20160101.csv）中构建特征，同时提取训练数据集和测试数据集，训练模型。

1.2、整体思路

最终预测为（product_id，user_id）的商品-用户对，具体分析有两种情况：

1、用户与商品有交互行为，可以再 action 表格中获取一个月内有过交互行为的商品-用户对；

2、用户与商品无交互行为，无法通过 action 表格获取此种商品-用户对，且 action 表格中没有商品类别信息，无法通过同类商品的交互行为进行推荐，故最终舍弃用户无商品无交互记录的商品-用户对。

方案首先进行数据清洗，对 action 表格中错误无效的数据记录进行剔除；其次进行特征工程，包括用户特征(user_feat)、商品特征（product_feat）和用户-商品特征（user_product_feat）；然后进行特征选择，采用 xgboost 模型训练，在训练过程中可获得特征重要性的排序结果；最后是模型训练与预测，xgboost 可以对模型进行继续训练，故先后通过 6 个月中的 5 份训练数据对模型进行训练，然后对预测数据集进行预测，获取预测结果。

1.3、代码说明

代码运行说明如下：

1、分别运行 2_user_feat.py、2_product_feat.py 和 2_user_product_feat.py 获取用户特性、商品特征和用户_商品特征，运行 2_labels.py 为测试数据集获取标签；

2、运行 2_merge_all.py 获取测试数据集和预测数据集；

3、运行 2_train_predict.py 通过测试数据集训练模型，然后获取预测结果。