2023 年第二届钉钉杯大学生大数据挑战赛初赛题目

初赛 A：智能手机用户监测数据分析

一、问题背景：

近年来，随着智能手机的产生，发展到爆炸式的普及增长，不仅推动了中国智能手机市场的发展和扩大，还快速的促进手机软件的开发。近年中国智能手机市场品牌竞争进一步加剧，中国超越美国成为全球第一大智能手机市场。手机软件日新月异，让人们更舒适的使用手机，为人们的生活带来很多乐趣，也产生了新的群体“低头一族”。手机软件进入人们的生活，游戏、购物、社交、资讯、理财等等APP吸引着、方便着现代社会的人们，让手机成为人们出门的必备物品。

该数据来自某公司某年连续30天的4万多智能手机用户的监测数据，已经做了脱敏和数据变换处理。每天的数据为1个txt文件，共10列，记录了每个用户（以uid为唯一标识）每天使用各款APP（以appid为唯一标识）的起始时间，使用时长，上下流量等。具体说明见表1。此外，有一个辅助表格，app_class.csv，共两列。第一列是appid,给出4000多个常用APP所属类别（app_class）,比如：社交类、影视类、教育类等，用英文字母a-t表示，共20个常用得所属类别，其余APP不常用，所属类别未知。

表 1

变量编号	变量名	释义
1	uid	用户的id
2	appid	APP的id（与app_class文件中的第一列对应）
3	app_type	APP类型：系统自带、用户安装
4	start_day	使用起始天，取值1-30（注：第一天数据的头两行的使用起始天取值为0，说明是在这一天的前一天开始使用的）
5	start_time	使用起始时间
6	end_day	使用结束天
7	end_time	使用结束时间
8	duration	使用时长（秒）
9	up_flow	上行流量
10	down_flow	下行流量

二．解决问题

1.聚类分析

（一）根据用户常用所属的20类APP的数据对用户进行聚类，要求至少给出三种不同的聚类算法进行比较，选择合理的聚类数量K值，并分析聚类结果。

（二）根据聚类结果对不同类别的用户画像，并且分析不同群体用户的特征。（用户画像定义：根据用户的属性，偏好，行为习惯等信息对用户打标签，用以描述不同群体的用户行为，从而针对不同群体的用户推荐不同所属类别的APP产品。）

2.APP使用情况预测分析：要研究的问题是通过用户的APP使用记录预测用户未来是否使用APP（分类问题）及使用时长（回归问题）

（一）对用户使用APP的情况进行预测，根据用户第1~11天的a类APP的使用情况，来预测用户在第12~21天是否会使用该类APP。给出预测结果和真实结果相比的准确率。（注：测试集不能参与到训练和验证中，否则作违规处理）

（二）对用户使用APP的情况进行预测，根据用户第1~11天的a类APP的使用情况，来预测第12~21天用户使用a类APP的有效日均使用时长。评价指标选用NMSE。

$$\mathrm{NMSE}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y}_i)^2}}$$

式中，y_i表示使用时长的实际值；$\hat{y}_i$表示使用时长的预测值；$\bar{y}$表示所有用户的实际使用时长的平均值。给出预测结果和真实结果之间的NMSE。（注：测试集不能参与到训练和验证中，否则作违规处理）