腾讯社交广告
Tencent Social Ads

高校算法大赛

团队：Threeldiots

成员：辛超 黄伟鹏 梁晓

团队介绍

辛超

北京大学

黄伟鹏

北京大学

梁晓

中国科学技术大学

目录

问题描述

数据来源：腾讯社交广告系统

数据内容：

根据第17天到第30天中广告日志中的信息以及部分用户信息和用户安装app信息来预测在第31天中App广告被用户点击后激活的概率。

用户信息

广告信息

上下文信息

历史流水

问题描述

用户信息

经过算法修正或者用户注册填写的性别、年龄、学历、常驻地、婚姻状况、App安装列表等信息。

广告信息

广告主ID、推广计划、推广App、广告素材、App类别等信息。

上下文信息

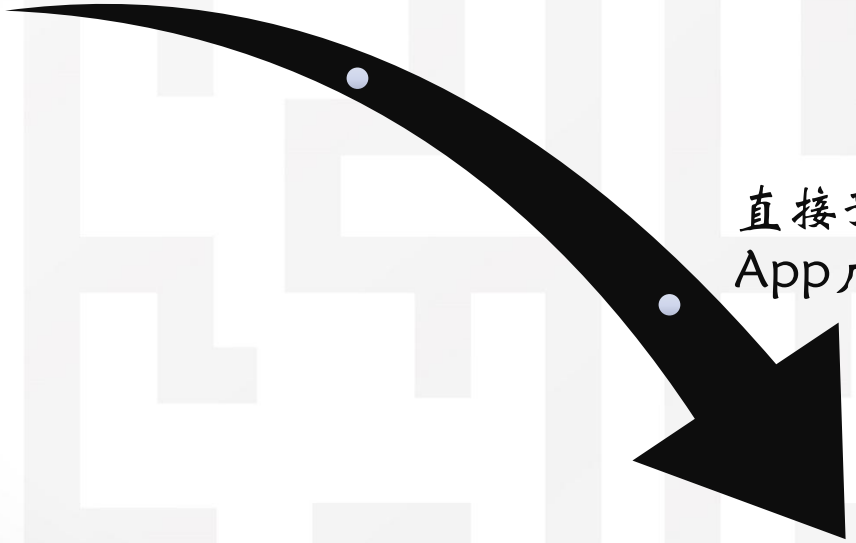
广告位类型、运营商、广告位置等信息。

历史流水

用户17-30天中对App广告点击与激活情况。

问题描述

应用场景：移动App广告

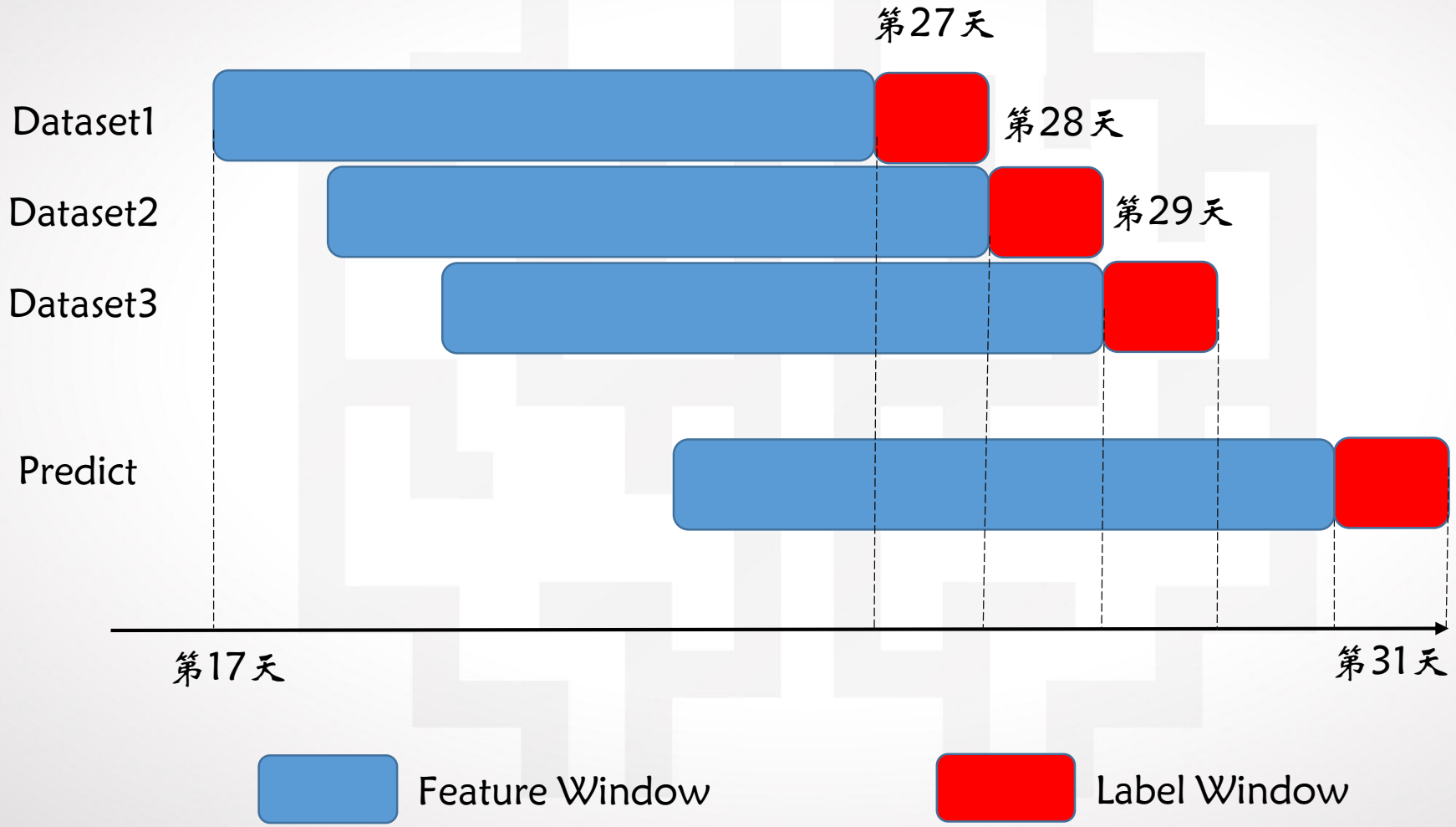


直接预测目标：用户点击App广告后是否激活

二分类问题，评价指标：logloss

算法框架

数据划分


第27天
Dataset1
第28天
Dataset2
第29天
Dataset3
Predict
第17天
第31天
Feature Window
Label Window

特征工程

特征工程

用户特征

◆ 描述用户的偏好、行为

用户信息

用户的个人信息
用户预先安装App数目
用户安装不同类别App数占比
用户安装当前类别App个数
……

用户的偏好信息，以及对用户本身的刻画

历史流水

用户最近一次安装App时间差
用户最近一次安装同类App时间差
用户同类别广告浏览统计
用户最近几天安装App统计
用户历史浏览position hash
用户历史浏览App hash
同一个position处用户浏览不同素材个数
……

对用户近期行为的刻画，广告的关注度以及兴趣程度

特征工程

广告特征

◆ 描述广告本身或App的推广度

广告信息

同一个App对应的推广计划数目
同一个App对应的素材数目
距离creative第一次出现的时间
距离App第一次出现的时间
距离position广告第一次被点击的时间
……

对于App的推广程度，不同素材或者是广告位的使用时间

特征工程

Label窗特征

◆ 测试集中提取的特征，描述用户当日行为

当日数据

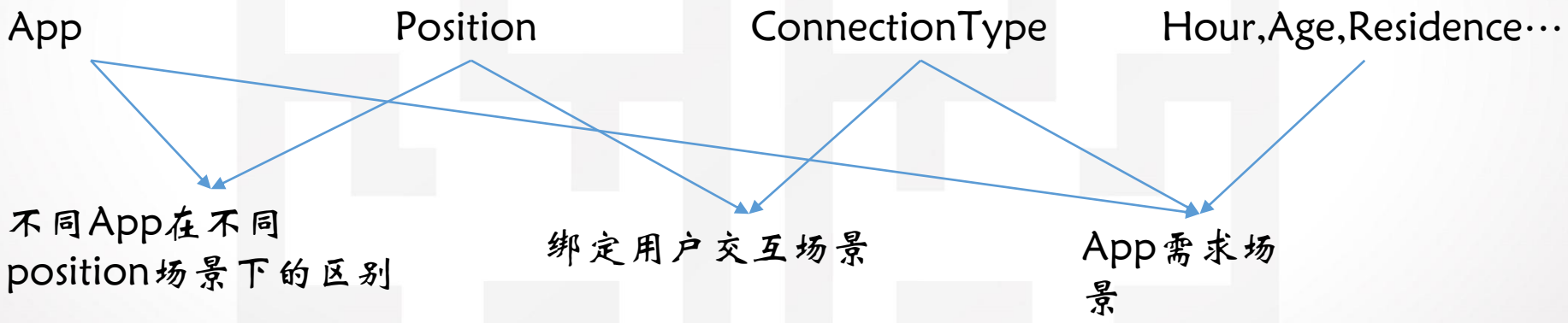
距离上一次和下一次点击广告的时间差
距离当日第一次和最后一次点击广告的时间
当日点击广告的持续时间
短时间内重复点击广告的计数以及次序

描述当日用户的点击广告情况，记录上下文信息。

特征工程

CVR特征

除去对单个要素计算CVR，还有部分组合CVR。



特征工程

其他特征

计算P(install App|用户属性)

根据历史安装App用户信息计算与当前用户的cosine distance

确定不同人群对App的需求度和App对个人的符合程度

模型融合

- 由于我们三个人有各自分别的模型，所以这里融合是直接针对三个人的模型进行融合。

$$最终结果=0.4*max+0.4*median+0.2*min$$

总结

问题：

内存

On-disk训练，训练时尽量避免使用Pandas、Numpy，如需使用要即时释放内存。

效率

流式或分块统计。

总结

反思：

没有尝试使用更多天数训练

比赛后期时间规划不合理

个人的模型特征没有进行汇总

模型单一

总结

收获：

接触到了接近工业实际的数据

锻炼了个人能力，综合考虑收益和时间成本去解决问题

锻炼了团队协作的能力

感谢比赛的组织方和项目出题方的付出！

腾讯社交广告
Tencent Social Ads

高校算法大赛

THANKS