3.解题思路

3.1．用户模型

1）概述

A．采用的模型

GBDT（xgboost,应该不是GBDT 你看雪伦大大的博客）

B．最重要的特征

用户活跃天数、用户等级。

用户统计量

C．使用的工具

Python+Pandas。

2）数据处理

A. 数据的处理

筛选出仅和cate8类商品交互的行为数据。（你不是用了局部信息和全局信息么，写一下）

B. 为什么要使用这些数据？数据的作用？

因为需要预测在商品子集上的用户购买行为，因此仅在这部分数据上进行采样和提取相关特征更加精准。

3）特征选择与获取

A．哪些特征是关键特征

用户活跃天数；用户等级；用户六种行为的统计和；用户对商品的行

为交互量的最大、最小值，均值，方差；用户的购买转化率……

用户类别活动天数， 商品行为交互量的最大、最小值，均值，方差，

商品特征, 用户点击14的model\_id

B．特征是如何想到和获取的

来自对业务的理解，对数据的分析，相关比赛的经验分享……

C．特征之间是否有相关关系

有，比如转化率就来自购买量/其余的交互量。

商品流行度根据各个统计量加权平均

D．特征是否经过处理？

否。（个别特征映射成0，1）

4）模型选择与训练

A. 为什么选择这个模型？

准确率相对较高，对数据的要求不高（如标准化等），它的一个实现（xgboost）训练速度快，且效果很不错。

B. 模型的训练方式？

输入特征矩阵直接训练

C．是否进行了模型融合？模型的融合方式？

我们测试了bagging，blending(基于权重，平均)， multi-stacking等融合方式。

在user模型中，融合效果不佳，最终未采用

在user-sku模型中最终选择了基于特征选择的blending ------ 进行特征重要性分析， 将重要特征单独训练一个模型， 在其余的非重要特征中， 随机选取10次组成10个不同特征， 加权评价

5）有趣的发现

A. 使用的小技巧？

分步预测（先预测哪些用户会买，后预测这些用户买了什么）提升很大

B．对于特殊、异常数据，如何处理它们？

user模型 未处理

user-sku模型 去掉了过年7天的数据

C．你觉得你最突出的优势是什么？

User模型中单模型取得了较高的成绩

User-sku模型，充分利用特征信息， 并防止f12过拟合

6）其他你想分享的

代码托管，code review很重要， 团队合作，队友很赞