中国大数据算法大赛-用户购买时间预测

队伍名称：皱眉可达鸭

2018/07/19

JDATA

目录

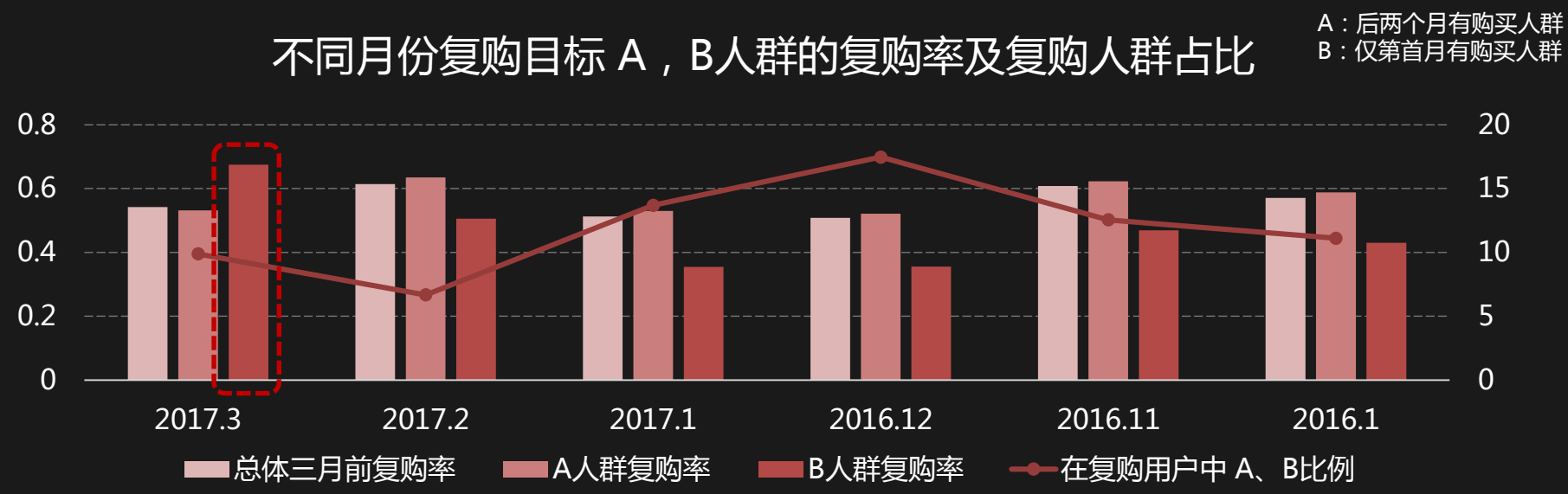
JDATA

目录

JDATA

数据分析结论：S1训练集划分方式

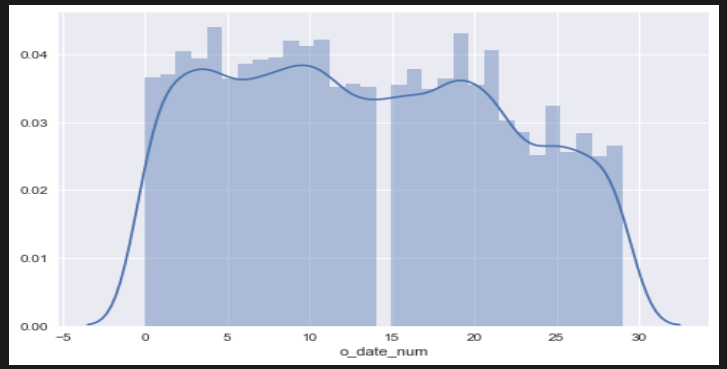
- S1训练数据集构建：
  - 训练集的构建需要保证不会导致结果泄露
  - 只关注最近两个月有购买用户的复购准确性，S1 Score上限大概为（预测2月复购为例）0.8，因此放弃仅首月有购买用户的复购率估计是合理的
  - 当然，也尝试构建了历史上更早时间的近首月购买用户的特征，作为额外数据补充，但由于分布差异，反而影响了整体的准确性



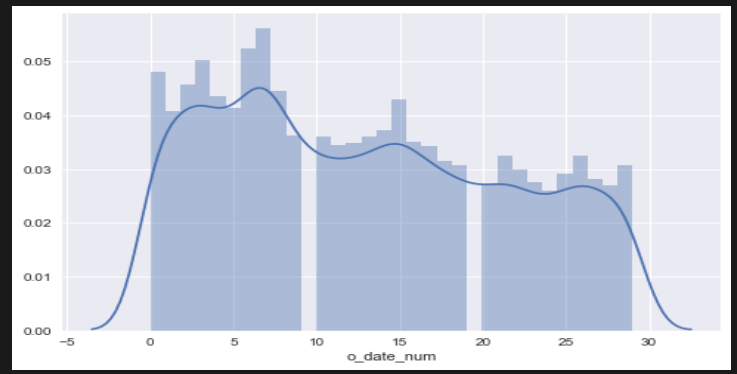
数据分析结论：采样应对S2不同月份分布差异

- 由下图可以看出，两个月份的第一次购买日期(S2-label)的用户数分布有明显差异
- 为了避免单一月份数据带来的不确定性，通过滑动窗口和取中位数的方式，得到稳定的label 分布数据
- 融合四个时间段的S2训练数据集，进一步降低可能的偏差
- 最后利用得到的分布结果，对数据集进行降采样，形成最终S2 训练集

2017-07-01未来30天 各天首次购买分数分布



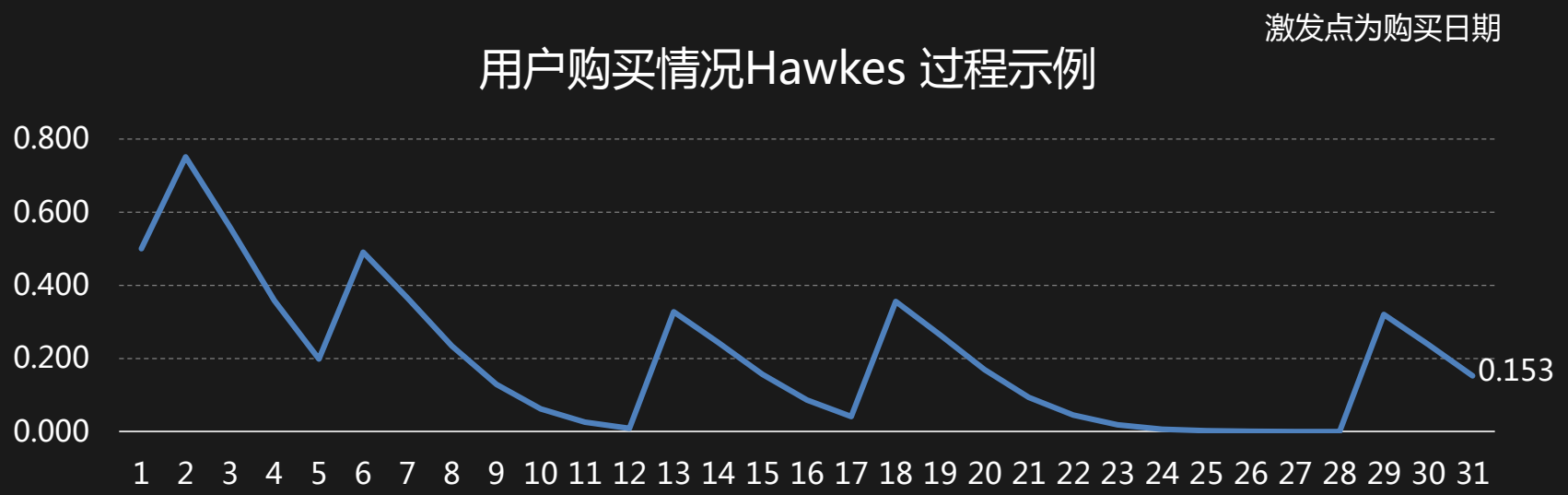
2017-08-01未来30天 各天首次购买分数分布



JDATA

重要特征：生存分析问题 Hawkes过程建模应用

- 生存分析问题，需要关注在实验期内的事件发生的频率间隔对期末的影响
- Hawkes过程能够将一段时间的行为所产生的影响，在期末综合地表达出来
- 在本赛题中，对分品类的购买，action，评论等行为应用Hawkes过程建模，从而能够在期末得到从用户以上各行为所反映出的对目标品类的关注度，有助复购预测



JDATA

重要特征：业务理解的交叉特征效果出众

- SKU消耗时长
  - 假设目标品类是具有周期性消耗特征的商品，那么用户历史上的SKU消耗时长，将能够很好地预测下一次的购买时间
- Label Baseline
  - 通过划分不同时间段，将历史上用户在不同时间段对目标品类的复购情况作为该用户的复购行为baseline参考
- 分组Label Baseline
  - 不同等级，不同地区的用户复购情况差异明显，因而再利用历史复购数据，统计不同等级、不同地区用户的人群情况
- 此外，还对用户对促销偏好特征进行了提取，不过没有取得明显的效果

模型选择及训练

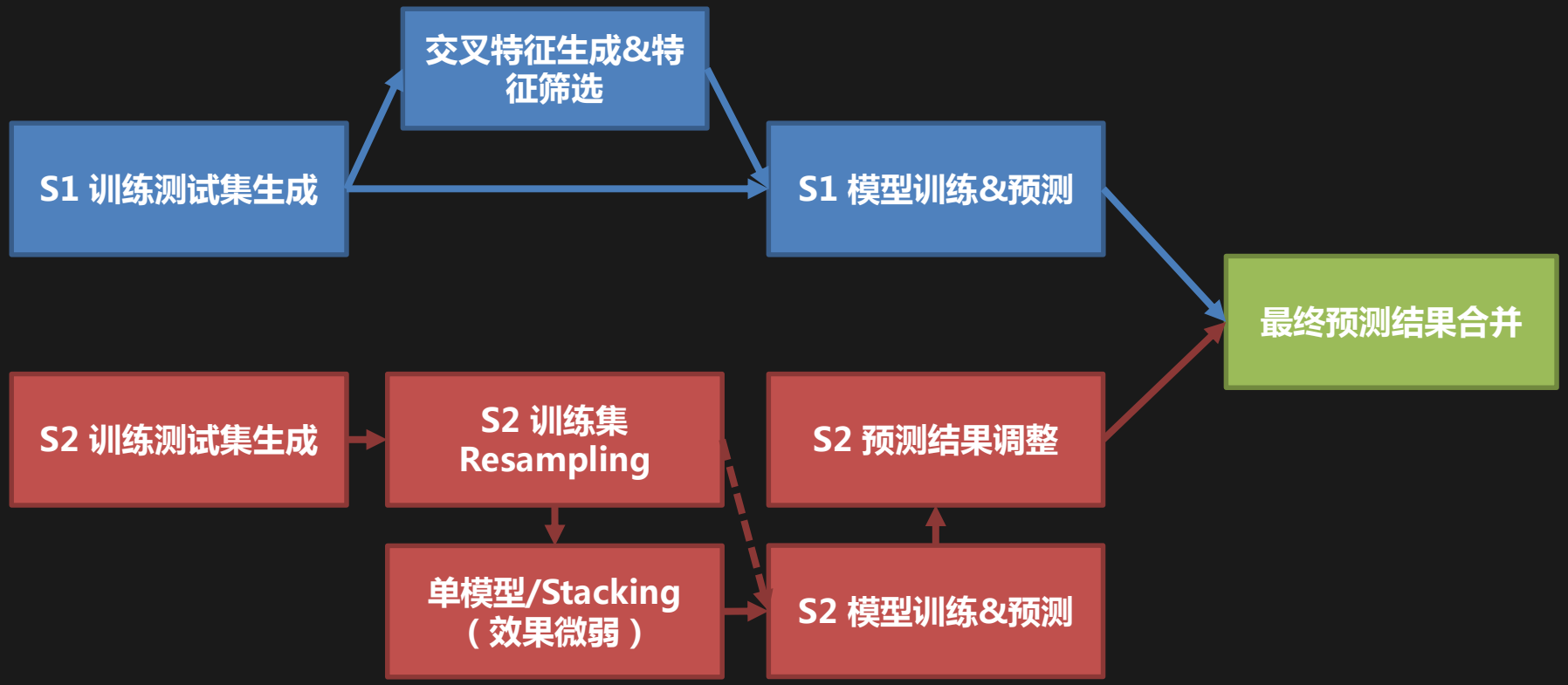
- 基础模型：XGBOOST，stacking中使用了LGB 、RF等

- Object function 选择：
  - S1 XGBOOST rank:pairwise, auc
  - S2 XGBOOST regression, log处理 ，rmse

- Stacking部分
  - Stacking S1效果不明显，最终S1部分没有使用stacking
  - S2部分stacking（仅略有效果）

- S2 预测后调整
  - 为了使得S2最终预测与之前训练集时间的分布更为一致，将5万条结果中，时间靠前60%记录减去1天，靠后10%记录增加1天

方案模型结构


交叉特征生成&特征筛选
S1 训练测试集生成
S1 模型训练&预测
最终预测结果合并
S2 训练测试集生成
S2 训练集 Resampling
S2 预测结果调整
单模型/Stacking（效果微弱）
S2 模型训练&预测

目录

JDATA

优势遗憾以及其他方案

- 优势 ：
- 成员背景各异，互相启发思路
- 对数据的前期分析，分布统计比较充分
- 熟悉电商业务场景

- 遗憾：
- 后期过于关注stacking，忽略了模型参数、模型目标函数的修正
- 更多的数据时间划分尝试，Hawkes过程建模的深层应用，更多精细的采样方法

- 更多的方案：
- 时间序列数据RNN方案DCM 细化等

Thanks for listening

皱眉可达鸭

2018.7.19

JDATA