中国大数据算法大赛-用户购买时间预测

队伍名称：WTF

演讲者：梁策远

2018.07.19

JDATA

目录

JDATA

团队介绍

梁策远

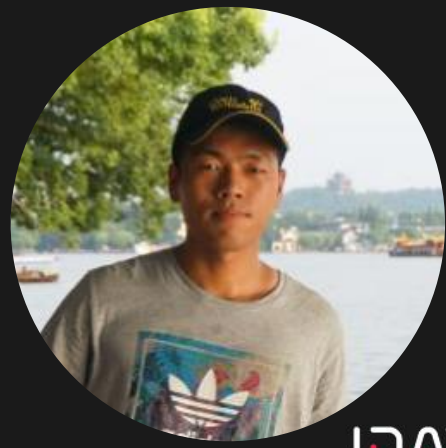
福大研二

冠军9人次，亚军4人次，季军4人次（覆盖jdata，kaggle，kdd，天池，dc等平台）

许文超

北邮研二

崔世文

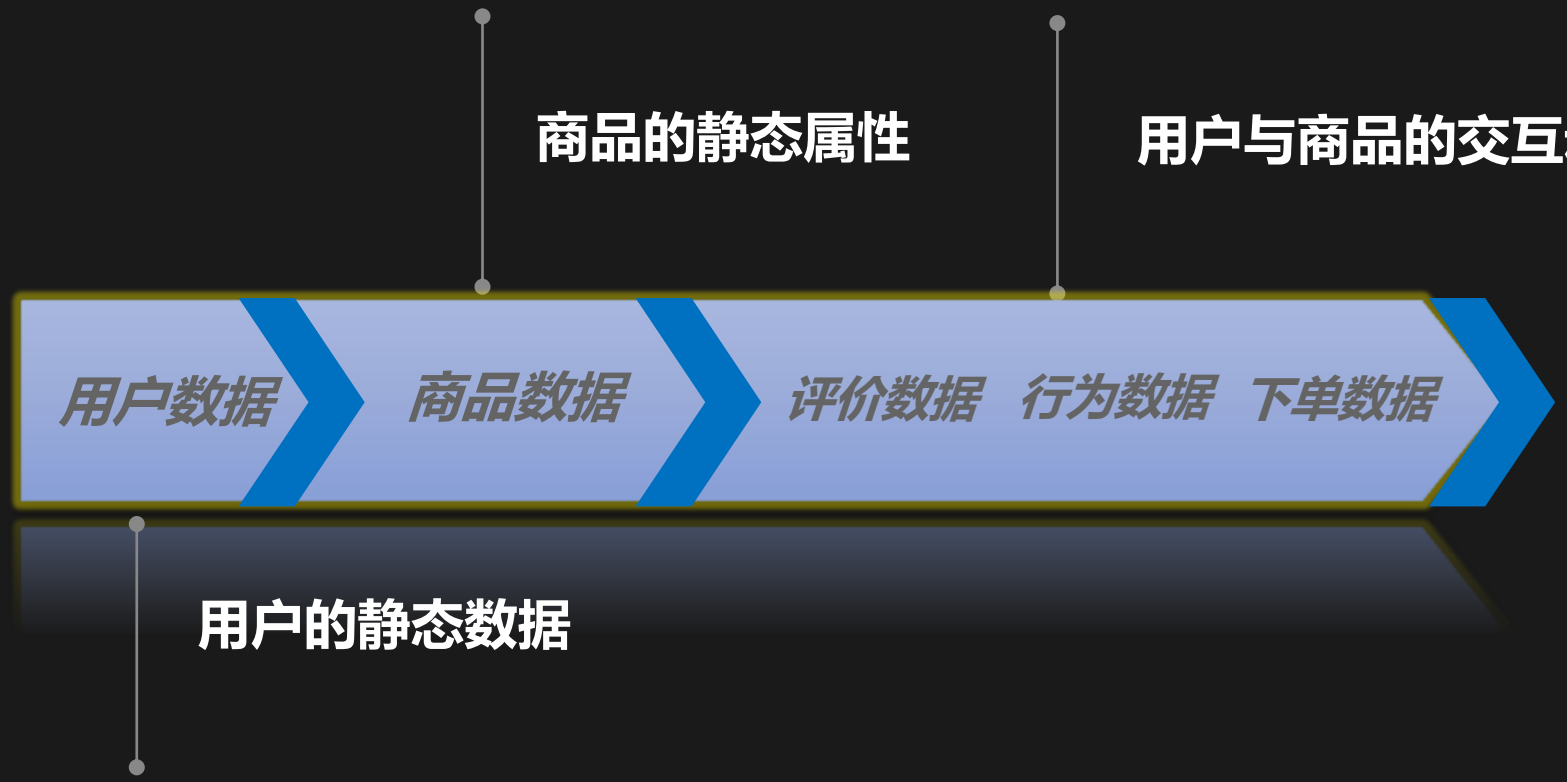


JDATA

目录

JDATA

赛题分析



S1:二分类

未来1个月内最可能购买目标品类的用户

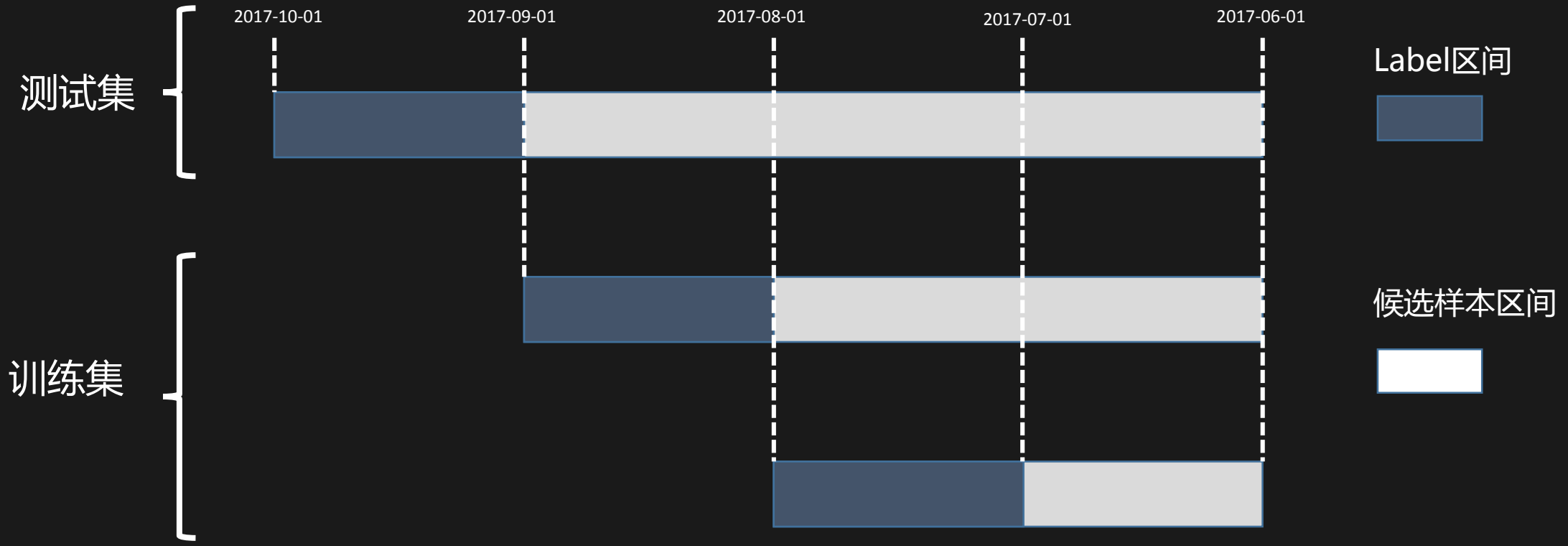
购买用户的首次下单日期

S2: 回归

任务描述： 参赛者需要根据用户基本信息、SKU基本信息、用户行为信息、用户下单信息及评价信息，自行设计数据处理相关操作、训练模型、预测未来1个月内最有可能购买目标品类的用户，并预测他们在考察时间段内的首次购买日期。

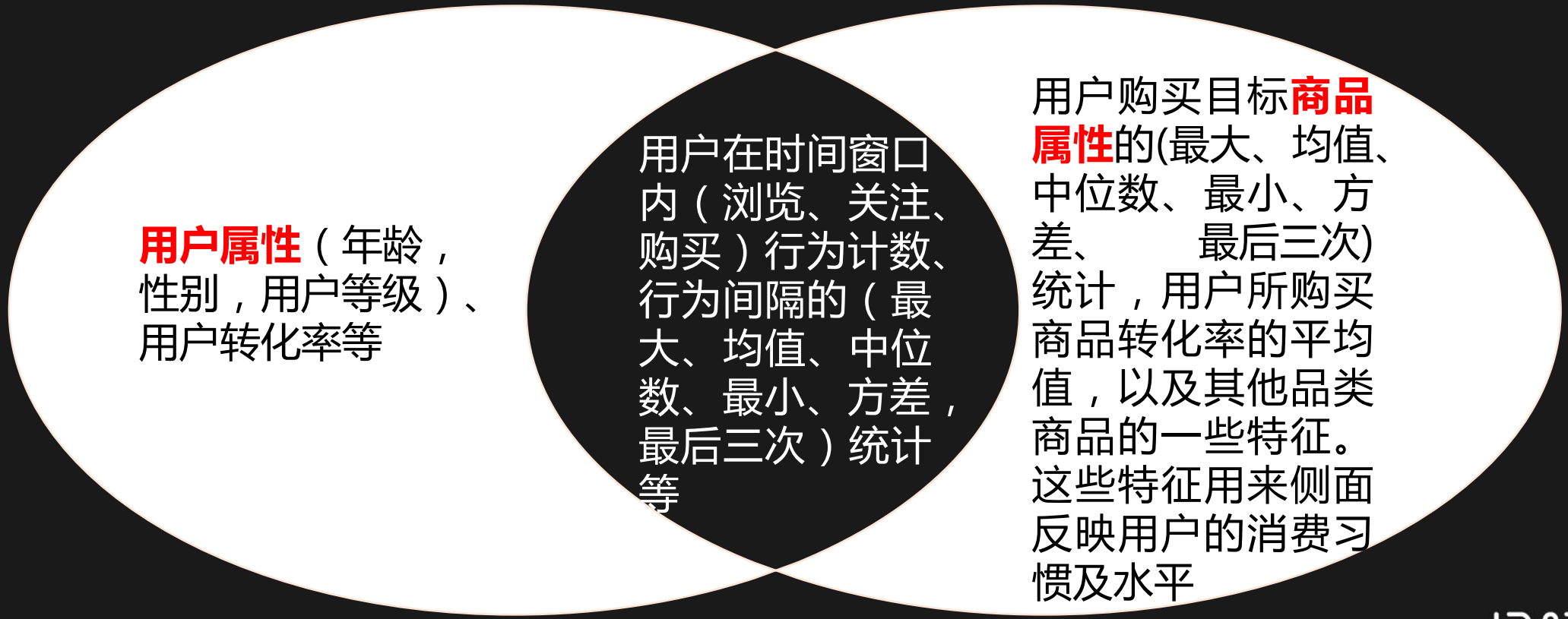
JDATA

S1样本构建

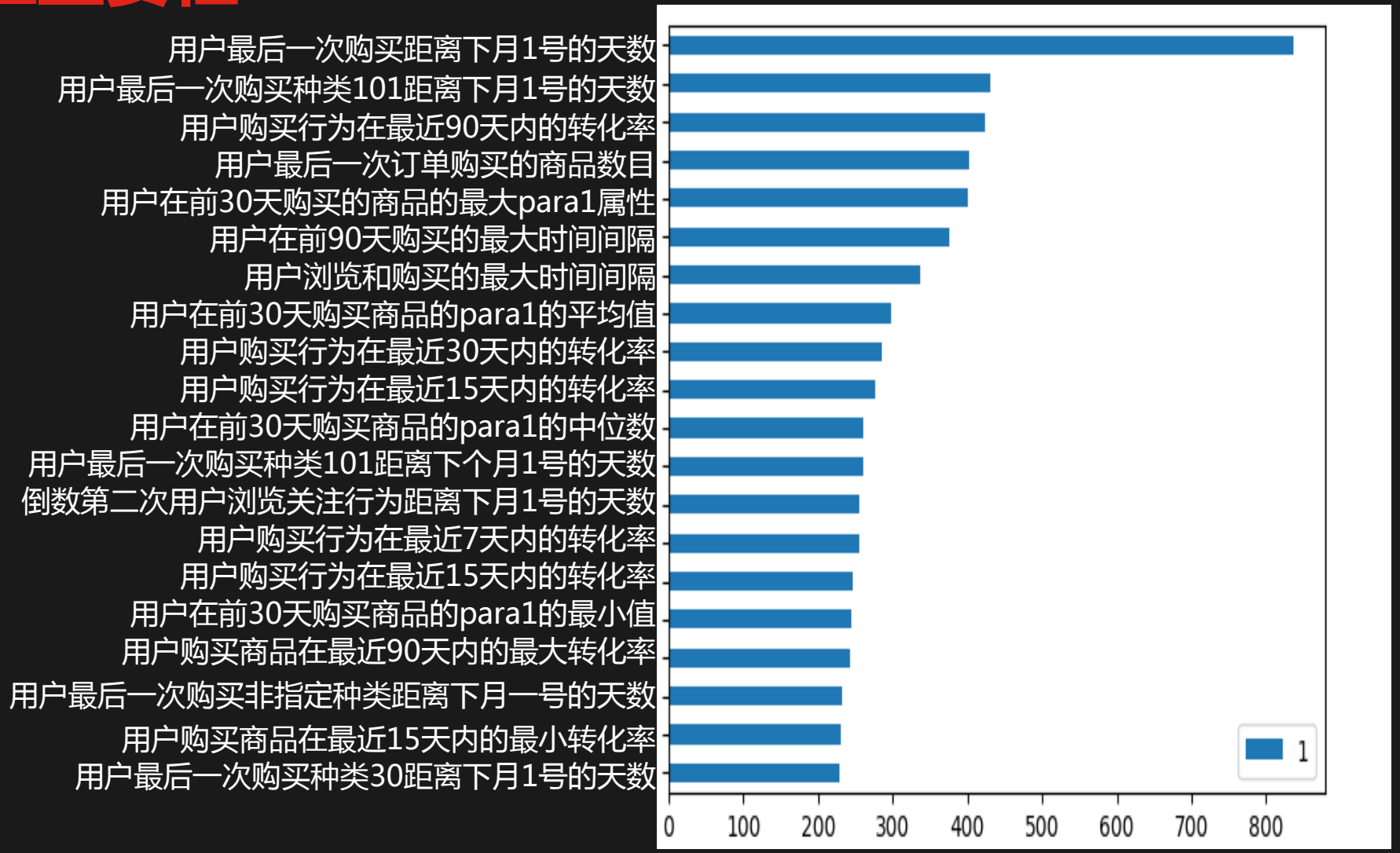

2017-10-01
2017-09-01
2017-08-01
2017-07-01
2017-06-01
测试集
训练集
Label区间
候选样本区间

特征工程思路

我们主要考虑三部分信息，用户自身信息，商品属性信息，用户-商品交互信息（由时间窗口控制短期和长期信息）

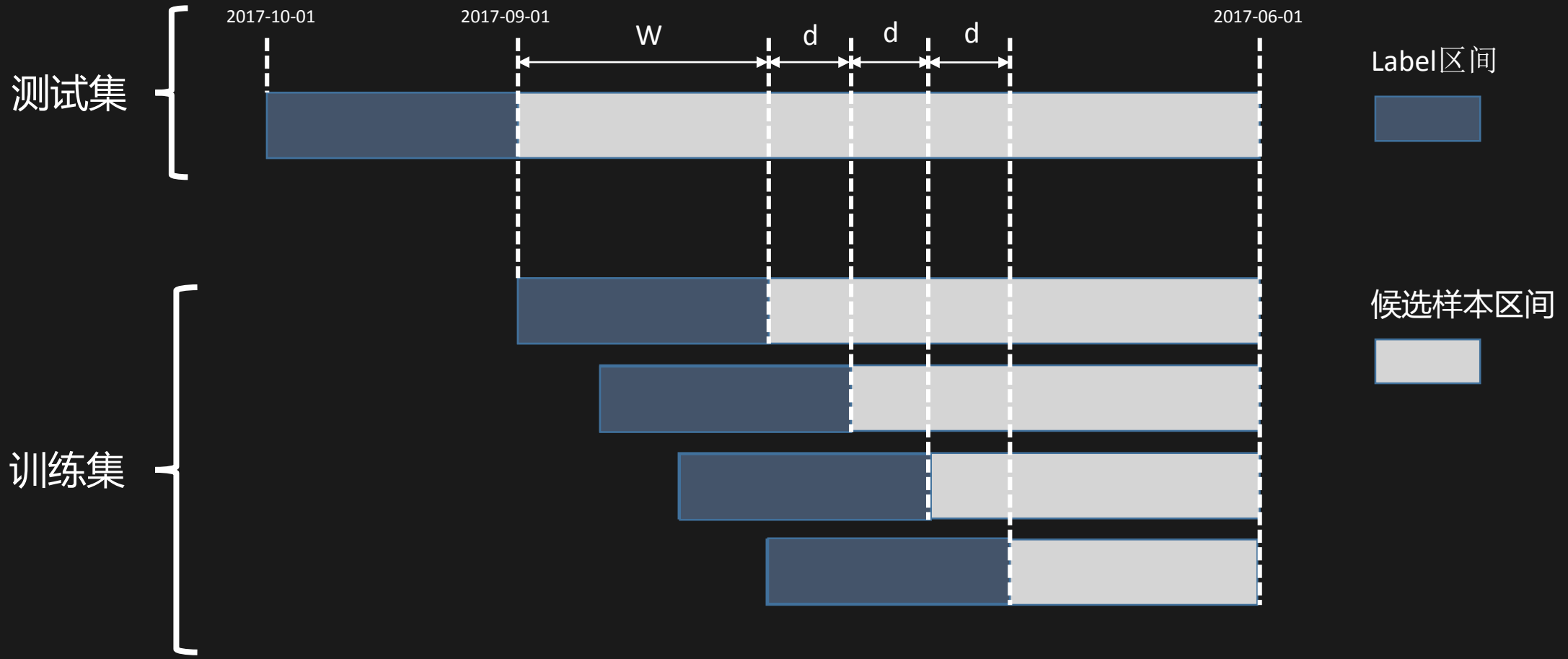


S1特征重要性



JDATA

S2样本构建

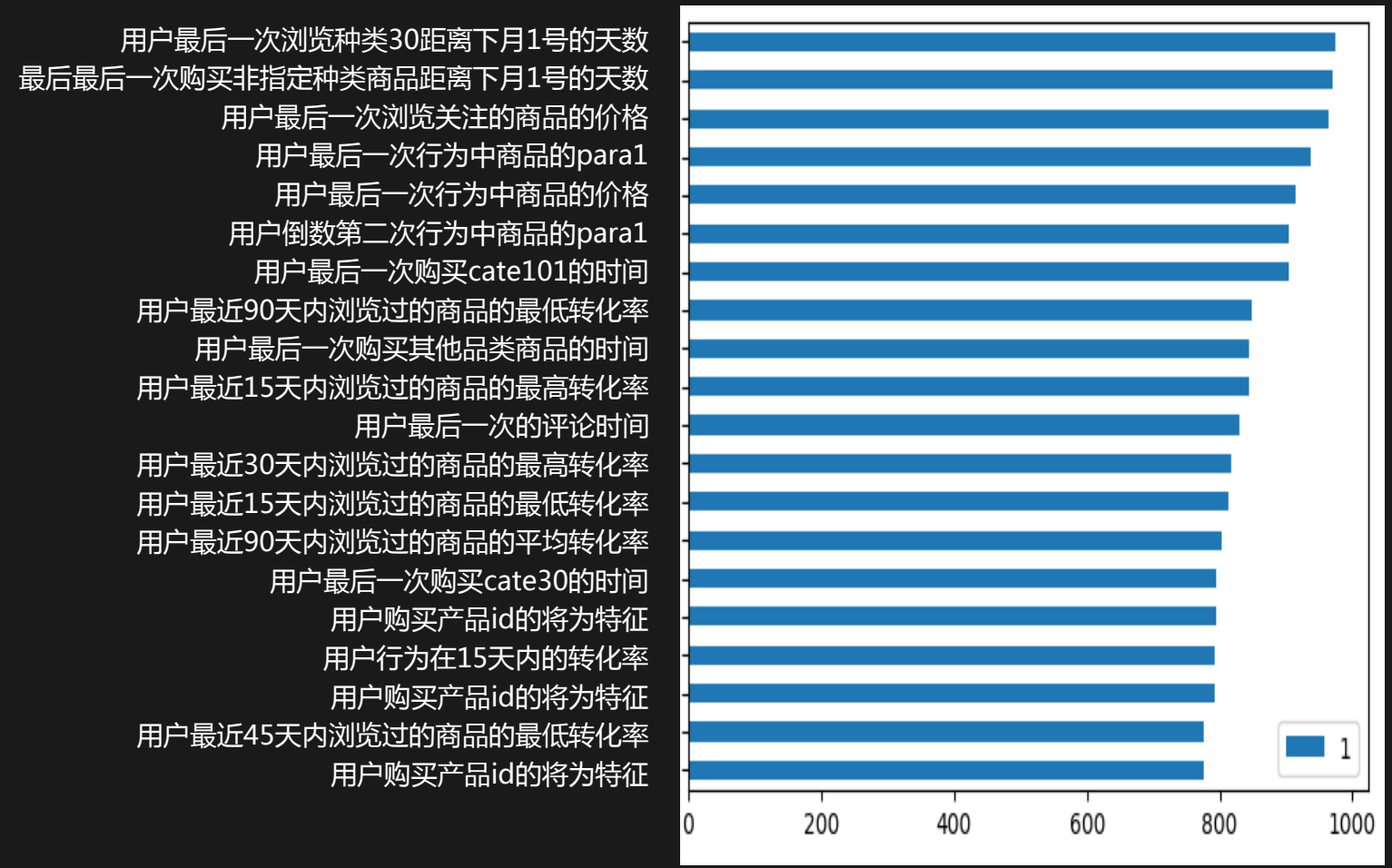


JDATA

S2建模及特征重要性

1、看作回归任务，目标变量Y表征用户在距标签起始日第Y天首次下单。

2、S2任务的评价指标的平方损失项位于分母，着重小值， 所以我们选取q<0.5的分位数回归，对预测值偏大的结果惩罚更大，于是会出现模型预测值普遍偏小的情况，而这更符合s2的评分取向。

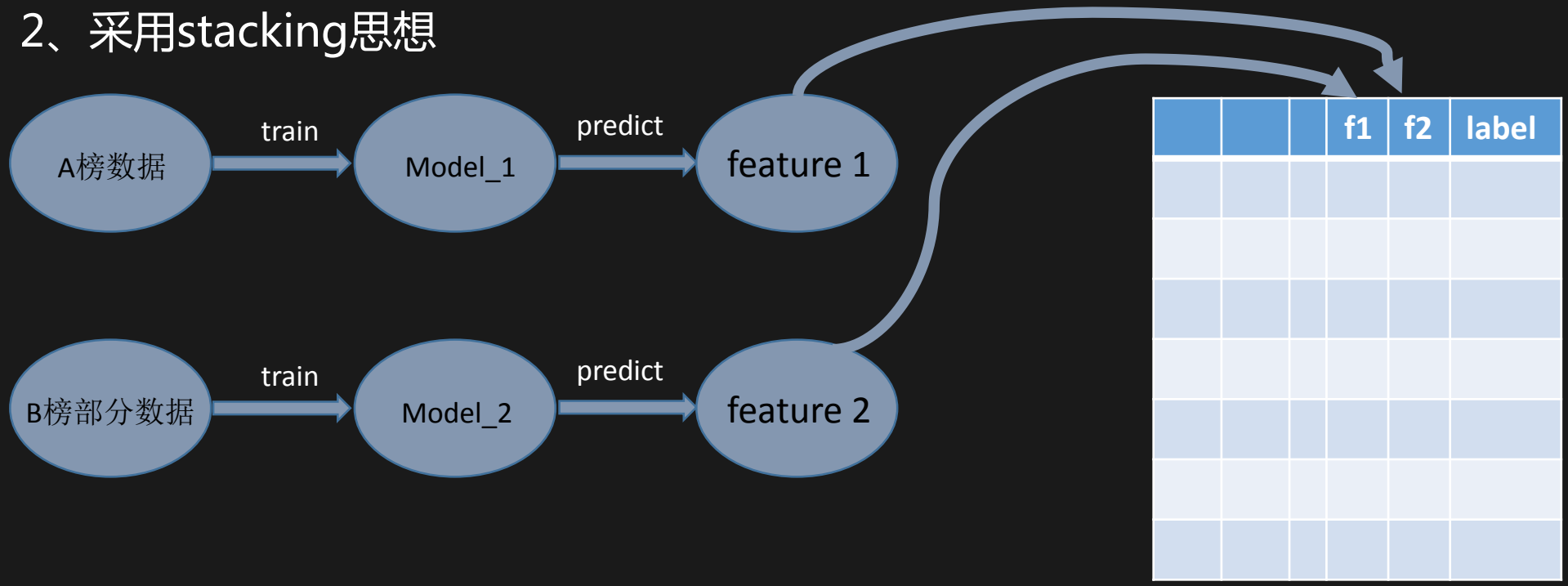


JDATA

特征工程技巧

1、利用用户和商品（商品属性）的共现信息，构建用户商品（商品属性）共现矩阵，通过矩阵分解算法，得到用户的K维度隐向量表示，表征用户商品购买偏好。

2、采用stacking思想



如何防止S1过拟合

问：过拟合原因？
答：由于本题特殊的抽样方式从最后3个月中抽取有购买的用户，所以存在一种特殊的现象，就是3个月前没有购买记录的用户一定会在最后三个月发生购买。

问：如何解决？
两种办法：
1、删除掉过拟合的特征，通过分析可知过拟合的直接特征是最后一次购买时间，所以删除掉所有跟最后一次购买时间相关的特征，线上分数也是提高的。
2、删除过拟合的样本，正常情况下为了保证线上线下一致，都要选取前3个月有过购买记录的用户作为候选样本，但是在6月份之前没有购买的用户都会发生训练集leak，所以我们在选取候选的时候只选取6月份之后的用户。
通过对比：发现方案二效果远远好于方案一，归其原因是由于最后一次购买时间是本题最重要的特征之一。

目录

JDATA

经验总结

1、要充分利用控制变量法的思想，去搞明白每一次提升和下降的原因。

2、队友之间的交流要多交流思路和想法，少交流细节，这样既可以更容易突破个人瓶颈，又能尽可能保持模型的差异。

3、保存每个版本的提交代码，这样代码出现错误后能够快速复原。

4、建立合理的线下测试机制。

5、赛后多学习其他队友的比赛经验，这是最佳的学习机会。

6、碰到新的评测指标，需要分析评测指标的性质，研究是否可以直接对评测指标进行优化。

谢谢！

JDATA