D 国反击战解题思路

A 榜排名第 5，B 榜排名第 9.

S1 概述

采用的模型

采用的主体模型为 lightgbm 算法，其中本 s1 答案由两个模型进行归一化后平均，每个模型都对残差进行了二次预测。

重要的特征和特征清洗方法

重要的特征主要包括：

1. 最后一次购买时间+9 个月平均购买时间间隔（目标商品）
2. 最后一次购买时间+用户在最后 3 个月里购买商品的平均购买间隔（目标商品）
3. 最后一次购买的总价格（目标商品）
4. 最后一次购买的时间（目标商品）
5. 最后一次评论距离预测日期的距离（目标商品）
6. 最后一次的第一参数（目标商品）
7. 用户在最后 3 个月里购买商品最小的 sku 售出总数（目标商品）
8. 9 个月购买时间间隔的标准差（目标商品）
9. 最后一次评论时间（目标商品）
10. 最后一次行为时间+9 个月平均行为时间间隔（目标商品）
11. 商品价格中位数（非目标商品）

特征是否经过处理

部分特征进行额外分支建立，如为缺失值进行赋值-999，同时对部分特征，如第二参数与第三参数中的空缺值赋 nan

使用的工具

使用的工具主要为 lightgbm，sklearn，numpy 以及自己开源的特征选择工具 MLFeatureSelection。

S1 数据处理

该题目的数据处理，或曰数据构造方法最为关键。针对 s1 题目我们进行了两种不同的数据处理方法，包括：

1. a 模型：选取 6，7 月份有过购买行为的用户，并对该部分用户在 8 月的行为进行检测，以是否进行购买作为 label 并进行训练，验证与预测。其中线下检验是通过 8 月份用户标签构成的样本进行 5 折检验，以 auc 作为评价指标进行建模。其中该模型最重要的特征为用 7 月用户购买行为作为训练集进行训练，并对 8 月进行预测的特征。
    a. 这种方法需要注意的点是所有特征均需要为相对特征，因为训练集只有 8 月份的特征，而需要预测的则是用户 9 月份的行为。
2. b 模型：选取前 5 个月的数据进行数据集构造，其中用前 2 个月对目标商品有交互的用户，举例说明为通过 6，7 月份有交互的用户在 8 月购买行为作为标签，5，6 月份有交互的用户在 7 月份购买行为作为标签，因此得到了 3，4，5，7，8 月份的训练数据，其中由于 6 月份为京东 618 大促，具有较大的噪声，因此不选择该月数据作为训练集。在
    a. 这种方法在构造的时候我们使用部分绝对时间，并增加预测时间作为标签来表征时间维度。
    b. 线下验证是我们用 8 月进行验证

这两个数据构造的方法构建了我们最终 s1 两个预测结果，归一化平均后形成我们最终线上提交结果。

本代码使用了主办方提供的所有列表，包括最重要的订单列表----直接表征用户购买行为，行为列表----表征用户的需求和日常浏览行为，评论列表----表征了用户再次活跃行为以及对商品的满意度，还有用户属性行为。

并无对数据进行特殊的处理，曾尝试进行异常值检测与提出，但线上线下均无效果。

S1 特征选择与获取

哪些特征是关键特征

上文已列出前 10 主要特征，关键特征主要为行为特征以及商品属性特征。

特征的构思与获取

主要的特征构建是通过日常购物理解，其中包括初级特征，2 级特征以及 3 级特征（经植物 ijcai2018 比赛分享代码启发）。

1. 初级特征：用户等级，性别，年龄，最后下单时间等；
2. 2 级特征：用户平均购买时间间隔，用户平均行为时间间隔，用户购买商品平均价格，平均特性 1 等；
3. 3 级特征：用户购买多个 sku 的平均 sku 售出总数，用户最小 sku 在 3 月售出总数等。

其中代码进行多个滑窗，包括 1，3，9 个月以及个别特征更精细的滑窗。

最后还有用户对上 5 次的购买记录信息，包括对上 5 次的购买总数，购买时间，购买总价钱，行为类型，以及对上 11 个月的购买日期和购买数量，购买天数。在获得所有的特征后进行特征筛选，特征筛选主要使用笔者开源的特征选择代码 MLFeatureSelection，其中依次进行了根据重要性进行筛选，根据相关性进行筛选以及根据线下验证进行贪心算法和退火算法的筛选，提交代码中后缀为 log 的文件即特征选择后的日志。具体的特征选择代码细节可参考 https://github.com/duxuhao/Feature-Selection。其中贪心算法和退火算法的筛选流程如图 1 所示。

部分特征间有较高的相关性，在相关性筛选时我们队伍以 0.95 为阈值来进行筛选。

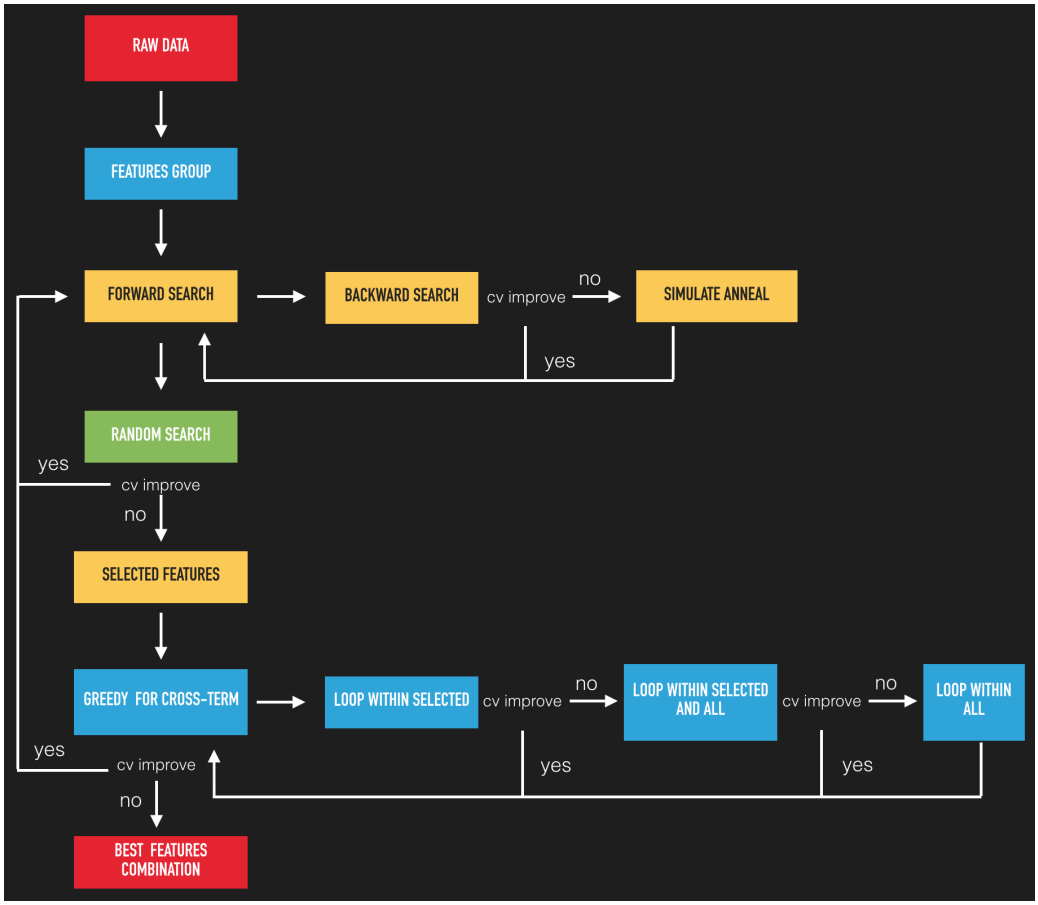

图 1. 贪心算法和退火算法的筛选

特征是否进行过处理

并无对特征进行特殊的处理，曾尝试进行异常值检测与提出，但线上线下均无效果。

S1 模型的选择与训练

为什么选择这个模型

本算法主要在 lightgbm 为基础进行展开， 选择 lightgbm 的原因是训练速度较快且内存消耗更少。

模型的训练方法

本预测结果结合了两个相对独立的模型，a 模型和 b 模型，这个两个模型不仅从数据构造上不一致，特征也是有所差别。其中预测的流程图如图 2 和图 3 所示。a 模型仅采用 8 月数据进行训练，在训练前用 7 月训练的模型进行 8 月概率预测，达到一定知识迁移的效果，然后原特征集加上 7 月模型预测结果构成总的特征集在用同样的算法进行训练获得 9 月预测概率。对于 b 模型则先通过特征组合 b1 进行全训练集和测试集的预测，然后将预测结果结合另外的特征组合 b2 进行回归预测。

对 b 模型如此操作的考量是因为在进行瓶颈分析时发现有相当部分的用户在过去数月都进行购买的前提下最后一个月没购买，而这批用户如果用过去几次的购买行为特征可一定程度的捕获，因此进行如此操作。最终模型在归一化后进行平均获得最后结果。

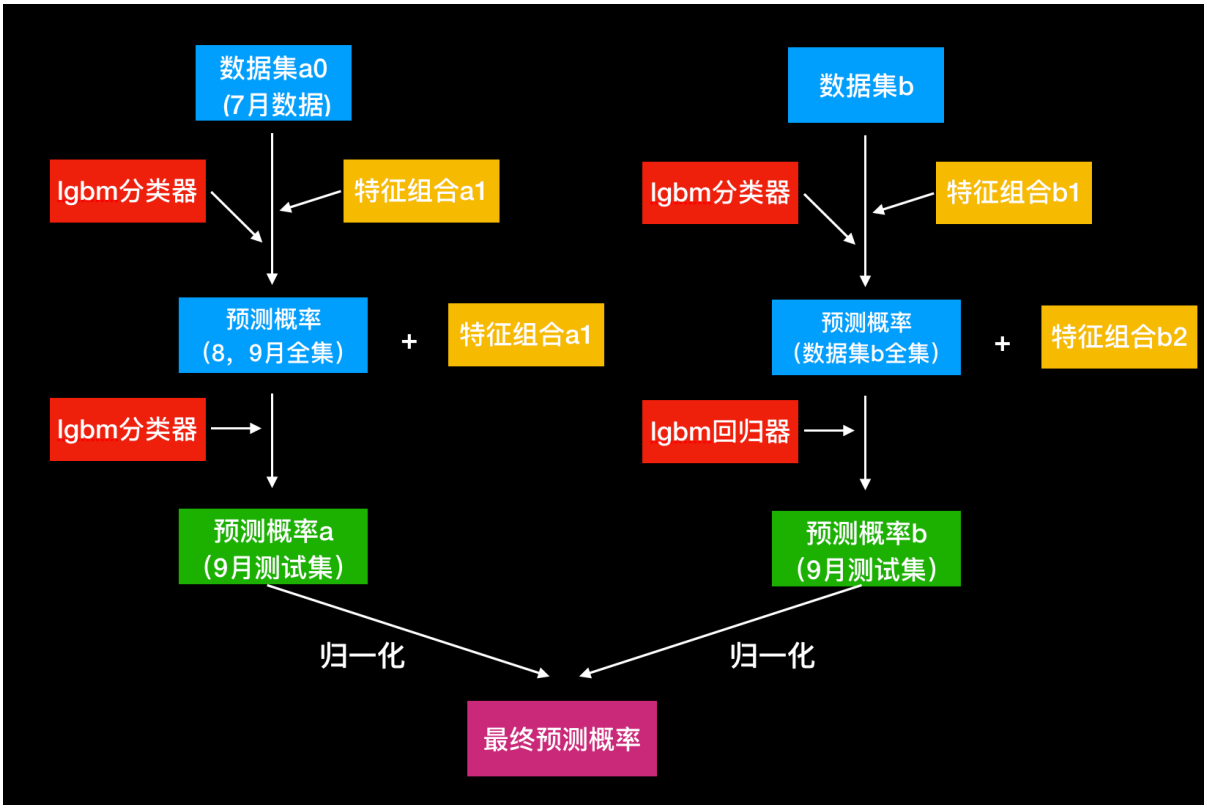

图 2. a,b 模型预测流程及最终融合结果

其他想分享的

这个比赛较于其他比赛门槛要稍微高一些，第一次感觉数据构造是如此有意思和重要的事情（可能是因为这次用户集选取的原因）。因为稍有不慎就会有信息泄漏（因为用户集选取是定向的），所以感觉一路做下来乐趣无穷！非常感谢大赛方举办的这次比赛！