**算法说明书**

目录

[一. 数据集的标注 1](#_Toc8414_WPSOffice_Level1)

[二. 数据预处理 1](#_Toc26453_WPSOffice_Level1)

[三. 特征工程 2](#_Toc17855_WPSOffice_Level1)

[四. 模型构建思路 2](#_Toc12324_WPSOffice_Level1)

[五. 模型后处理 3](#_Toc13303_WPSOffice_Level1)

[六. Tricks 3](#_Toc9124_WPSOffice_Level1)

1. **数据集的标注**

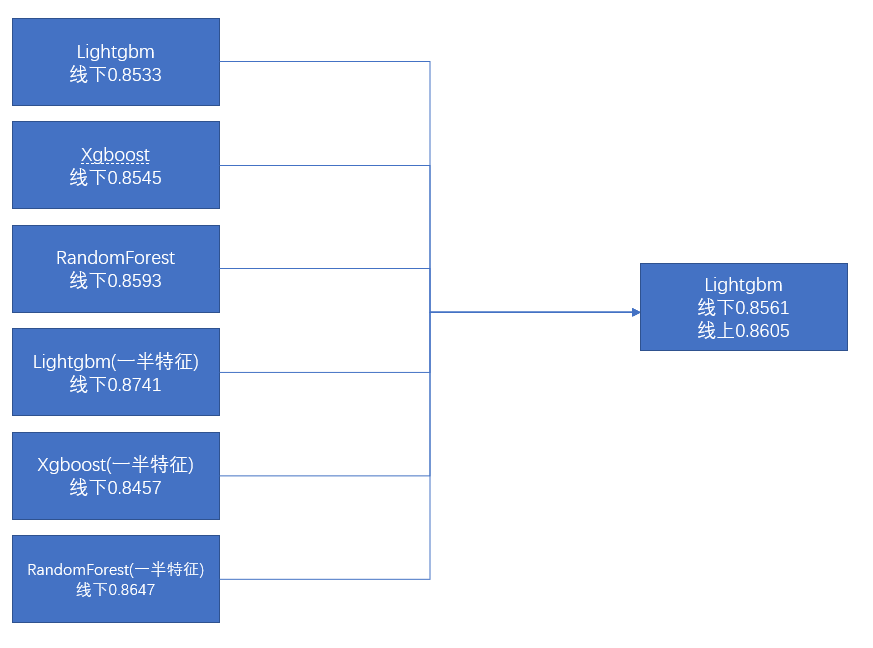
将2013年分为上半年和下半年两个部分，把上半年的数据作为训练集，根据其是否在下半年的数据中出现给其相应Label，以此生成本次比赛的数据集。

1. **数据预处理**
2. 函数std\_data将数据归一化
3. 函数chi\_square\_test进行卡方特征选取
4. 删除类别唯一的特征
5. 删除某一类别占比超过90%的列
6. 删除异常值
7. **特征工程**

由于我们所取的数据集是前半年的所有购买记录，故会出现一位用户对应多条购买记录的情况，所以我们需要对这种情况对相应特征做相关处理。下面是我们此次比赛提取的特征。

1. 用户购买次数
2. 用户的所有父订单的付款金额的最大值，最小值，平均值
3. 用户的会员状态
4. 用户性别
5. 用户所在省份，以及相应省份2013年的人均GDP(数据来自网络)
6. 用户购买商品时间与该商品上架的时间差
7. 商品上架时间
8. 用户购买的订单中商品总数
9. 用户购买的父订单所含商品数
10. 用户购买层级(以上半年所有付款金额的均值为分割)
11. 用户这次购买与上次购买时间的时间差
12. **模型构建思路**

我们将上面提取出的原始特征送入RandomForest, Xgboost, Lightgbm进行训练，得到了三个模型，然后对所选特征进行卡方特征选取，选择一半的特征同样送入RandomForest, Xgboost, Lightgbm进行训练，进而在得到三个模型。由此以来一共获得了6个模型。将这6个模型的输出作为6维的特征送入Lightgbm进行进一步训练，对上述6个模型进行融合，相关流程图如下：



1. **模型后处理**

我们对2013年全年购买次数超过10次的用户(共2051名用户)，在预测他们的购买概率的时候对其购买概率统一加0.1，线上提升0.0004。最终A榜0.86008.

1. **Tricks**

由于我们的数据集中的样本比例不平衡，以及本题对召回率的要求，故我们对所有相应模型的loss function进行了修改，对把1预测成0的进行了惩罚。对模型有较大的提升，线上分数有大概0.03的提升。