生活大实惠：O2O优惠券使用预测

----------基線----------

LodaLi　BRYAN　桑楡

2016.12.24

**队伍介绍 (**----------基線----------**)**

|  |  |  |
| --- | --- | --- |
| 队员姓名 | 参赛背景与本次竞赛负责部分 | 参赛经历 |
| 李凯东(LodaLi) | 北京凯来科技有限公司  负责赛题分析与数据分析 | 阿里云安全算法竞赛 1/940  机场客流时空分布预测2/3038  融360金融风控大数据竞赛 3/868 |
| 周耀(BRYAN) | 重庆邮电大学/腾讯  负责特征工程与模型融合 | 微额借款用户人品预测1/1556  阿里云安全算法竞赛 3/940  Bytecup国际机器学习竞赛5/1005  菜鸟需求预测分仓规划 6/2810 |
| 姬天戲  (桑榆) | 純規則研究中心  负责数据准备与特征工程 | 淘宝穿衣搭配算法 1/2100  阿里云24小时极限挑战赛 1/121  资金流入流出预测 2/4868  Bytecup 国际机器学习竞赛 2/1005 |

**方案提纲**

1. 问题分析
2. 数据准备
3. 特征工程
4. 模型设计
5. 总结回顾

**1问题分析**

给定用户在1-6月线下领券和线上点击消费行为数据，来预测用户在7月领券之后是否会消费。也就是需要我们分析用户过去和当前的状态信息，来预测用户领券之后会消费的概率，这是一个二分类问题，评估指标为AUC，对单个优惠券计算AUC，然后求平均值。

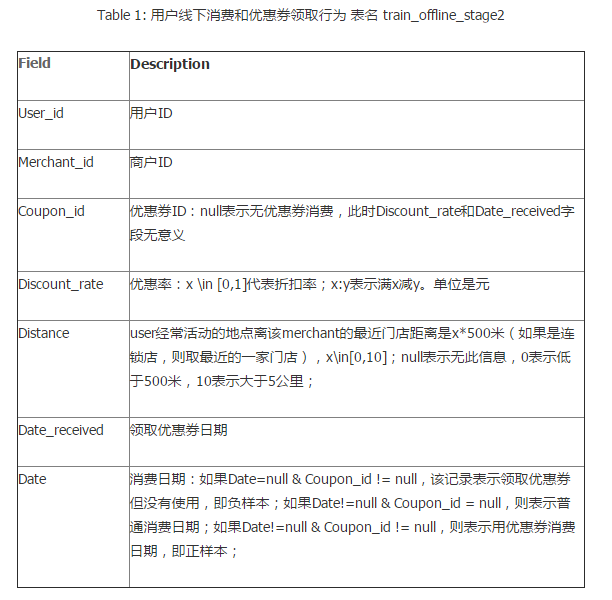
我们需要解决的主要问题有3个

Q1：如何划分数据集，样本如何表示

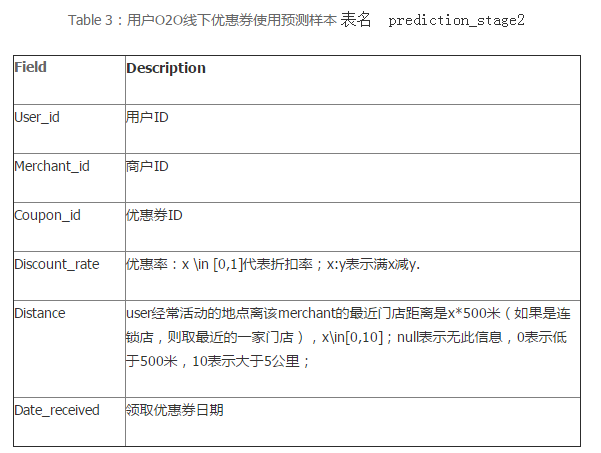
Q2：如何全面的构建特征工程来表达样本信息

Q3：评价方式AUC的本质是什么

**2.1数据准备**







**2.2数据简单探查**

如图1所示，蓝线表示领券之后消费数量，橙线表示没有领券消费数量，横轴表示时间轴。可以看到蓝线每个一段时间就会有个突增趋势，而且都是在中下旬，那么我们所预测的7月是否会有同样趋势呢？

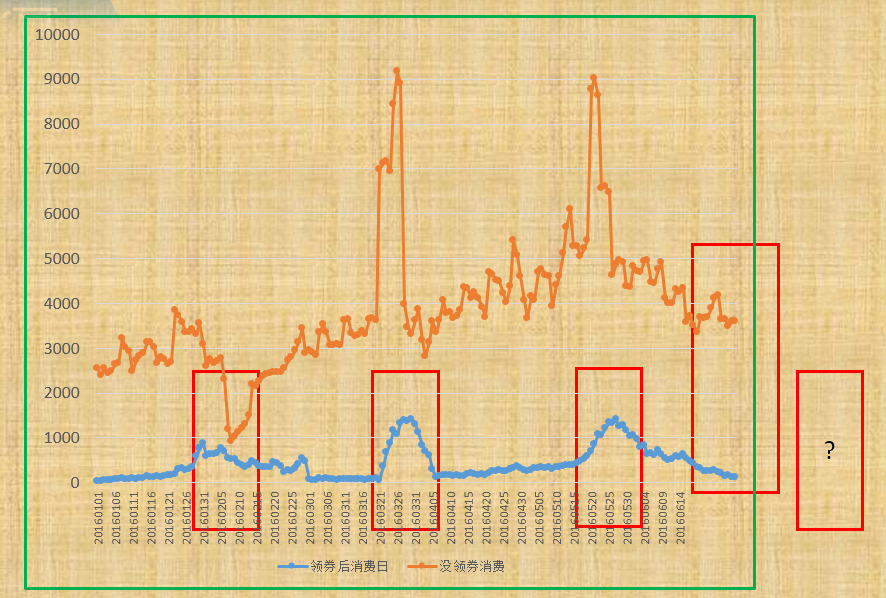


图1

如图2所示，横轴为时间轴4-6月，纵轴为领券之后被消费次数，可以明显的看到在每周四周五会有个小波峰，也就是说用户是否偏好在将近周末的时候领券被消费的概率会更大？

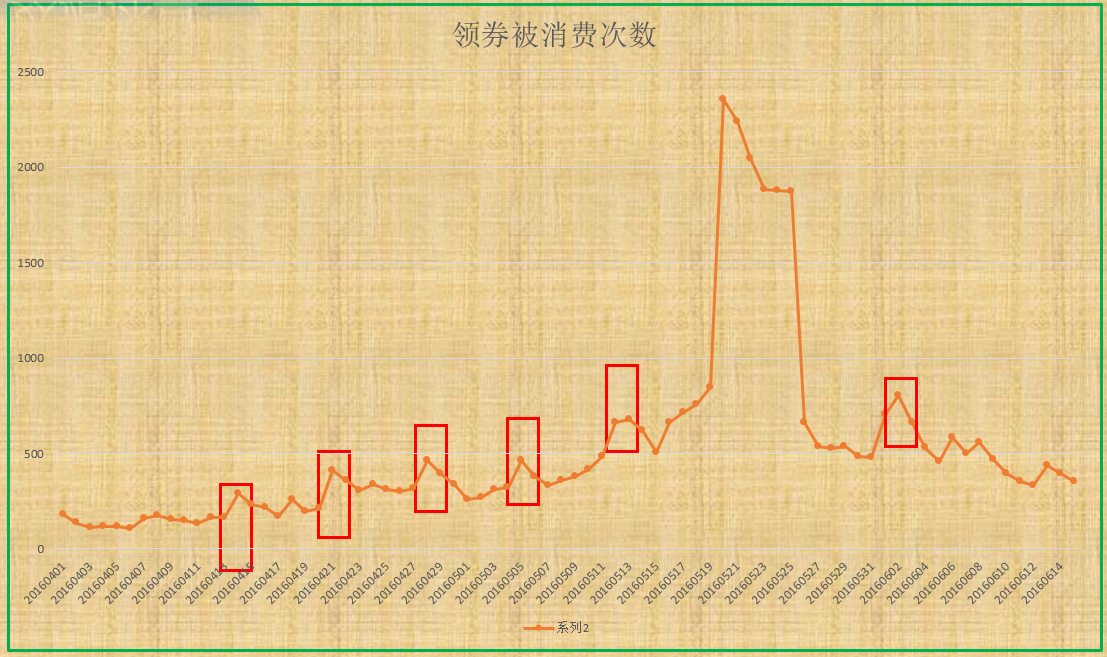


图2

**2.3数据准备-table1**

如图3所示，表示table1数据集的一个关于时间轴的分布，在6.16-6.30这段时间内，只有消费行为，没有领券行为。



图3

我们所需要预测的是7月的领券行为之后是否有消费行为，按时间区间计算特征与是否领券之后消费的相关性如图4所示。可以看到在最近一个月的相关性较大，最近2-3个月相关性很小。

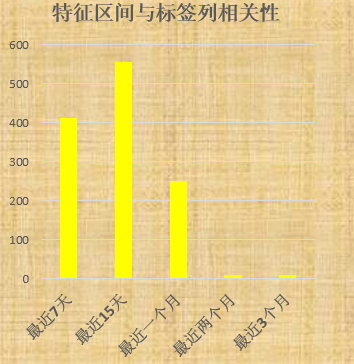


图4

所以我们可以按如下方式划分线上预测集，线下训练集，线下验证集，并严格保持数据集分布一致。



图5 线上预测集



图6 线下验证集

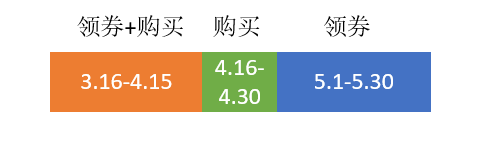


图7 线下训练集

**2.4数据准备-样本表示**

如图8所示，我们将用户，商家，优惠券，领券时间，折扣，距离作为key，将提取到的特征关联到key上，label作为标签，表示领券15天之内是否消费。领券15天内消费表示正样本，未消费表示负样本。



图8

训练集由两部分组成，其中一份正负样本比例1:8，主要用于xgboost，sparse-xgboost，sparse-lr模型。另一份数据集正负样本比例1:1，主要用户gbdt模型。

**2.5数据准备-预处理**

日期离散化：将日期转换为星期几，是否节假日，是否节前，是否节后

距离离散化：将距离转换成0-10，null这12个离散值

折扣率离散化：将折扣率转换成满多少，减多少，折扣率离散到5个区间

**3特征工程**

主要使用Map-Reduce的方式提取，方便使用key-value的方式进行计算key的特征，少部分特征使用SQL提取。

特征主要包括：基础特征群，组合特征群，交叉特征群，业务特征群，GBDT离散特征群，对于长尾特征采用取log的方式降低噪声。

**3.1基础特征群**

主要使用7,15,30天这3个区间，对用户，商家，优惠券，日期，折扣，距离进行特征提取。



图9

**3.2组合特征群1**

如图10所示，主要使用 用户与商家，优惠券，日期，折扣，距离进行组合，将组合后的字段作为key传入Map-Reduce进行提取组合特征。

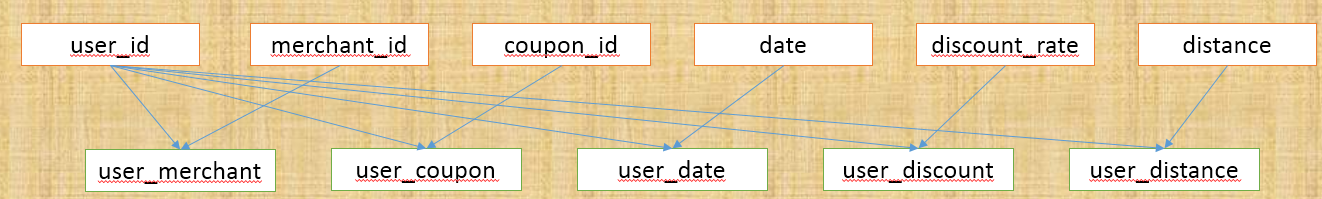


图10

其中，

user-merchant 特征群主要刻画用户对商家的偏好

user-coupon 特征群主要刻画用户对优惠券的偏好

user-date 特征群主要刻画用户对时间的偏好

user-discount 特征群主要刻画用户对折扣的偏好

user-distance 特征群主要刻画用户对距离的偏好

**3.3组合特征群2**

如图11所示，主要使用商家与优惠券，日期，折扣，距离进行组合，将组合后的字段作为key传入Map-Reduce进行提取组合特征。

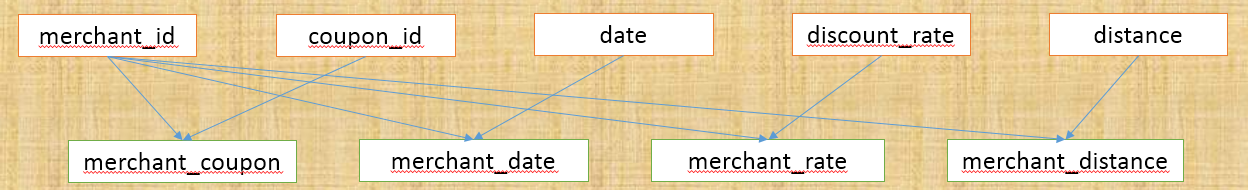


图11

其中，

merchant-coupon 特征群主要刻画优惠券在商家下的热度

merchant-date 特征群主要刻画不同时间对商家的影响

merchant-discount 特征群主要刻画不同折扣对在商家下的热度

merchant-distance 特征群主要刻画不同距离对商家的影响

**3.4组合特征群3**

如图12所示，主要使用<用户-商家>与优惠券，日期，折扣，距离进行组合，将组合后的字段作为key传入Map-Reduce进行提取组合特征。这些特征群刻画了不同因素对<用户-商家>组合的竞争。

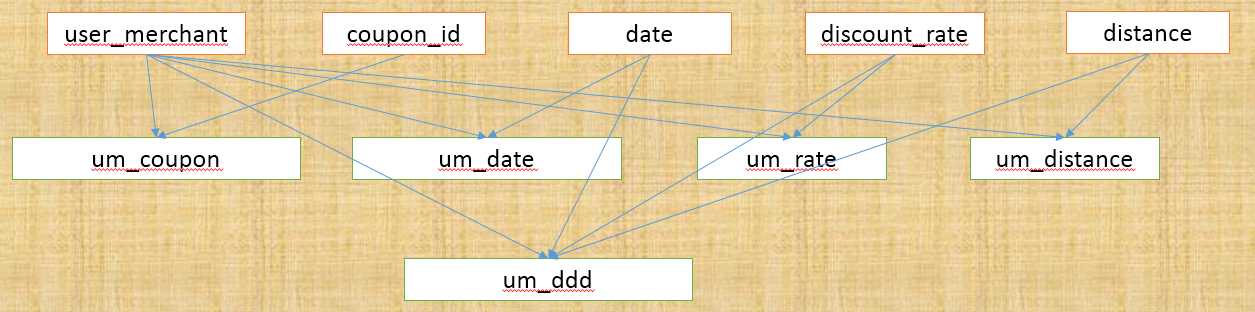


图12

**3.5交叉特征群**

如图13所示，主要使用与用户组合的特征群与用户大特征群交叉相除，作为交叉特征，这部分特征在SQL中进行提取。这些特征群刻画了不同因素对用户的竞争。如图14所示，主要使用与商家组合的特征群与商家大特征群交叉相除，作为交叉特征，这些特征群刻画了不同因素对商家的竞争。

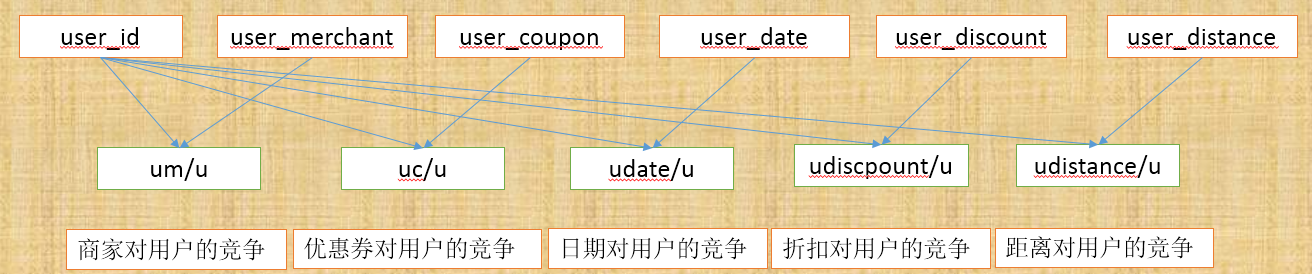


图13

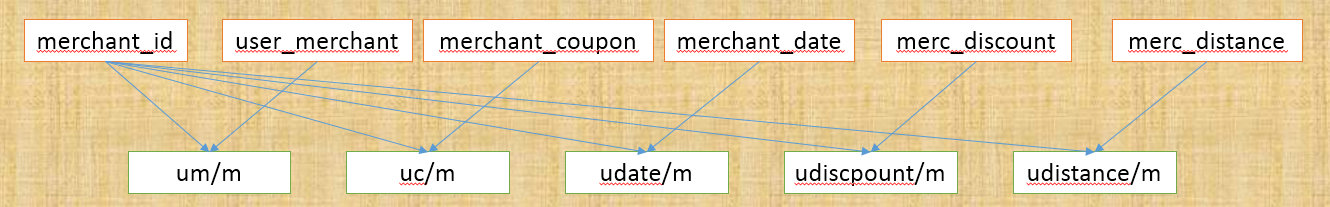


图14

**3.6交叉特征群-线上行为表table2**

如图15所示，在线下table1表中凸起user-date的组合特征，在线上table2表中提取user-date-action和user-date-discount特征群，然后将特征群关联起来，通过小类特征群除以大类特征群作为交叉特征，这部分特征群主要刻画了用户在相同时间段线上线下行为的对比。

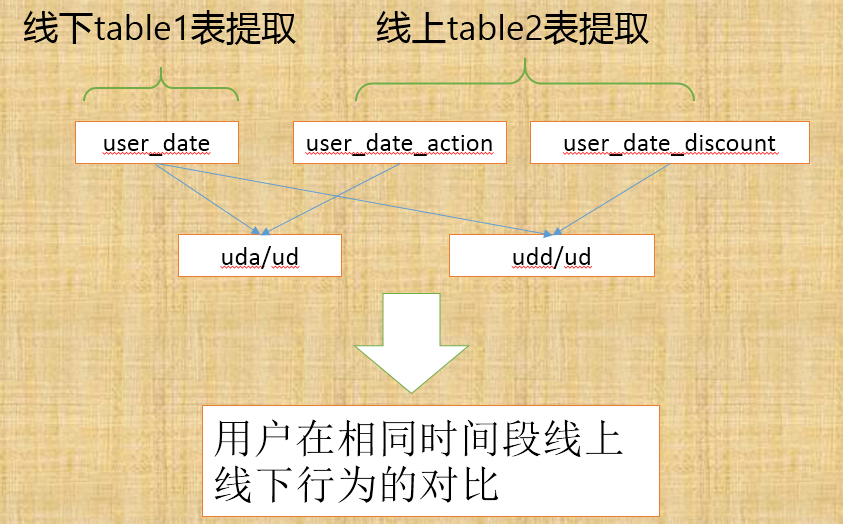


图15

**3.7业务特征群-领券月提取**

如图16所示，在需要预测的领券月提取业务相关特征，主要提取的特征如图17所示。



图16

|  |  |
| --- | --- |
| **上个月是否领过该券** | **上次领券是折扣率还是减满** |
| 本月领券次数 | 当次领券与上次领券优惠力度对比 |
| 当次领券时间排名 | 同一商家不同用户消费率排名 |
| 当次领券折扣率排名 | 同一商家不同用户当月领券数排名 |
| 当次领券距离排名(null当做0排序) | 用户最近一次消费时间与当前时间距离 |
| 当次领券不同商家的消费率排名 | 用户平均消费星期与当前星期距离 |
| 同一优惠券不同用户的消费率排名 | 用户平均消费距离与当前距离差距 |

图17

**3.8 GBDT离散特征**

如图18所示，这部分特征主要使用算法平台组件提取特征，提取原理大致如下：首先根据训练集训练一个GBDT模型，然后将训练集和测试集合并输入到模型中，每个样本将会落到不同子树的某个叶节点上，然后将当前样本所落到的不同子树的叶节点进行编码，这样一个样本的离散特征就构造好了，其余样本相同。

本次训练GBDT模型使用参数为500棵树，子树最大10个叶子节点，最大深度8，学习率0.02。

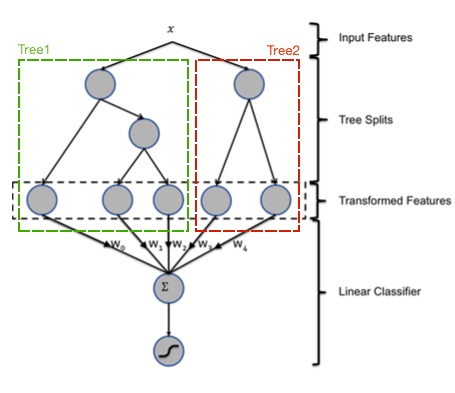


图18

**3.9特征重要性得分**

通过上述特征构造过程，我们生成了6个基础特征群，14个组合特征群，10个交叉特征群，1个业务特征群，一个离散特征群。这些特征群的重要性如图19所示。从中我们可以得到一些业务逻辑，比如靠前的基本都是用户商家的组合与交叉特征，那么我们是否可以从中取挖掘忠诚用户呢，另外在后面还有根据折扣组合的特征，从中我们是否可以思考如何用加个去诱导非忠诚用户呢？

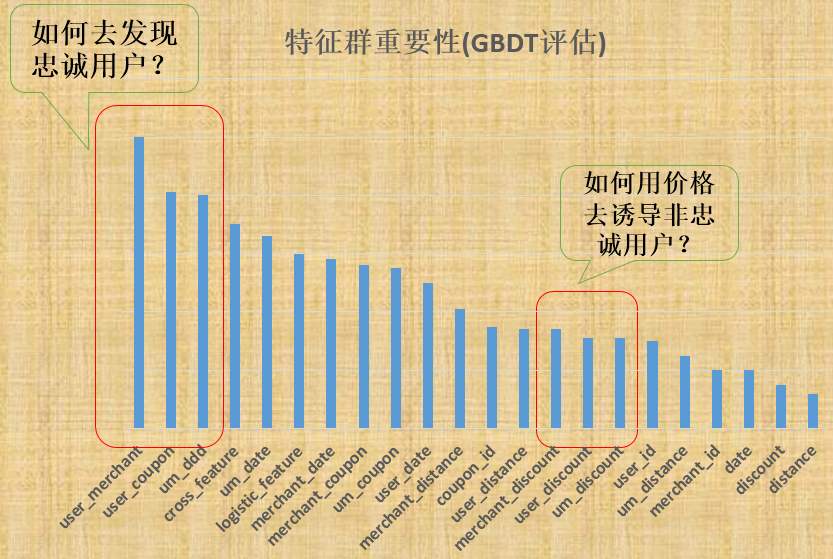


图19

**4模型设计**

本次竞赛我们主要基于3个单模型，GBDT，XGBOOST，LR。前两者为树模型，后者为线性模型，其中单模型效果XGBOOST>GBDT>LR。LR效果不如前两者主要原因是LR更需要离散的特征，也就是更加容易线性可分的特征，受此启发，我们将原始特征使用GBDT离散化之后追加到原始特征上，以此训练出两个新的单模型，Sparse-xgboost和Sparse-LR 。

其中GBDT和XGBOOST单模型还可以进一步进行bagging融合出新的模型，我们采用的是将特征评分，然后对评分取余(feature-importance%3)将特征划分为3部分分别训练模型，然后平均融合形成新的模型XGBOOSTs和GBDTs 。

最后我们考虑到评估指标为AUC，AUC的本质其实是模型对样本的排序能力，受此启发，我们采用 的方式对优惠券分组排序进行顶层融合，让最终结果排名更加稳定和精准。

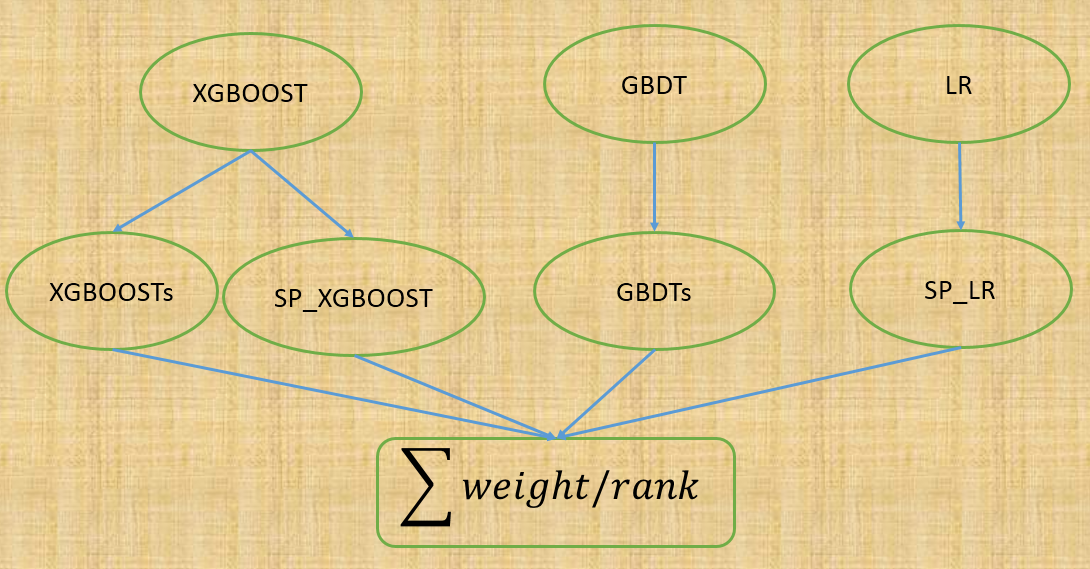


图20

如图21所示，分别为单模型，多模型融合，以及顶层模型融合的线下测试结果，和线上A榜B榜的结果。由于最后一次提交预测结果大于1导致得分为0.5，后来通过官方工作人员进行线下测试得到在B榜的结果为0.7979，验证了sparse-xgboost模型的有效性。

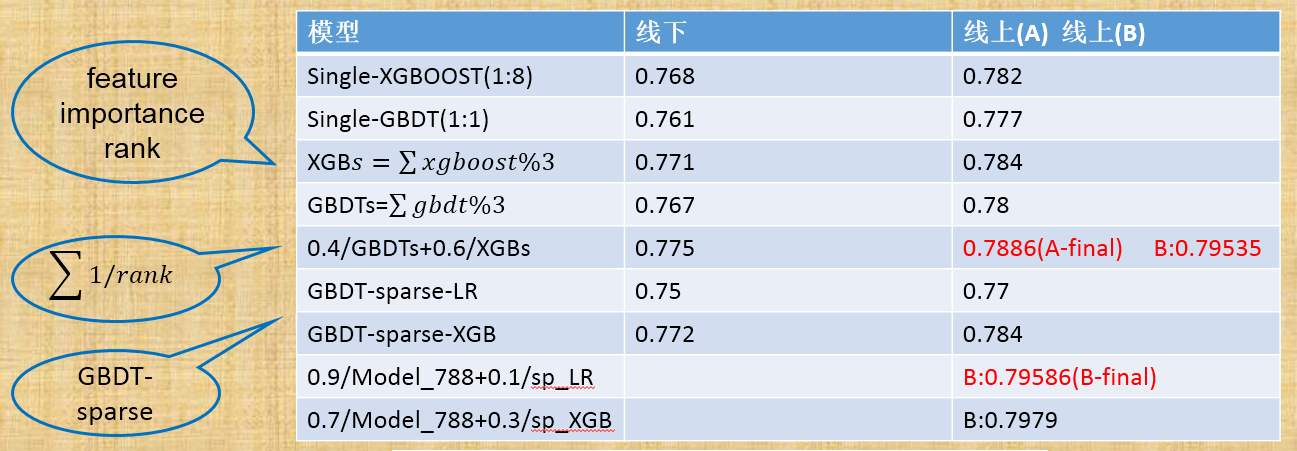


图21

**5总结回顾**

在本次报告中，我们提出了3个问题 ，现在一一得到解答。

Q1：如何划分数据集，样本如何表示？

A1：保持数据线上线下分布一致，使用key-value方式提取特征来关联特征与样本。

Q2：如何全面的构建特征工程来表达样本信息？

A2：充分利用所有信息，构建基础特征群，组合特征群，交叉特征群，业务特征群，离散特征群来全面的描述样本信息。

Q3：评价方式AUC的本质是什么？

A3：AUC的本质为模型对样本的排序能力，通过rank的方式进行多模型融合。

在本次竞赛中，初赛我们排名第二，复赛A榜排名第三，复赛B榜排名第三，有效的验证了我们设计的算法稳定性。





最后感谢各位专家领导聆听我们的答辩，感谢CCF和天池提供平台让我们学习，感谢指导老师王国胤教授的悉心指导，感谢各位队友的互相配合与不懈努力！