# 风险水位模型项目汇报-20200825

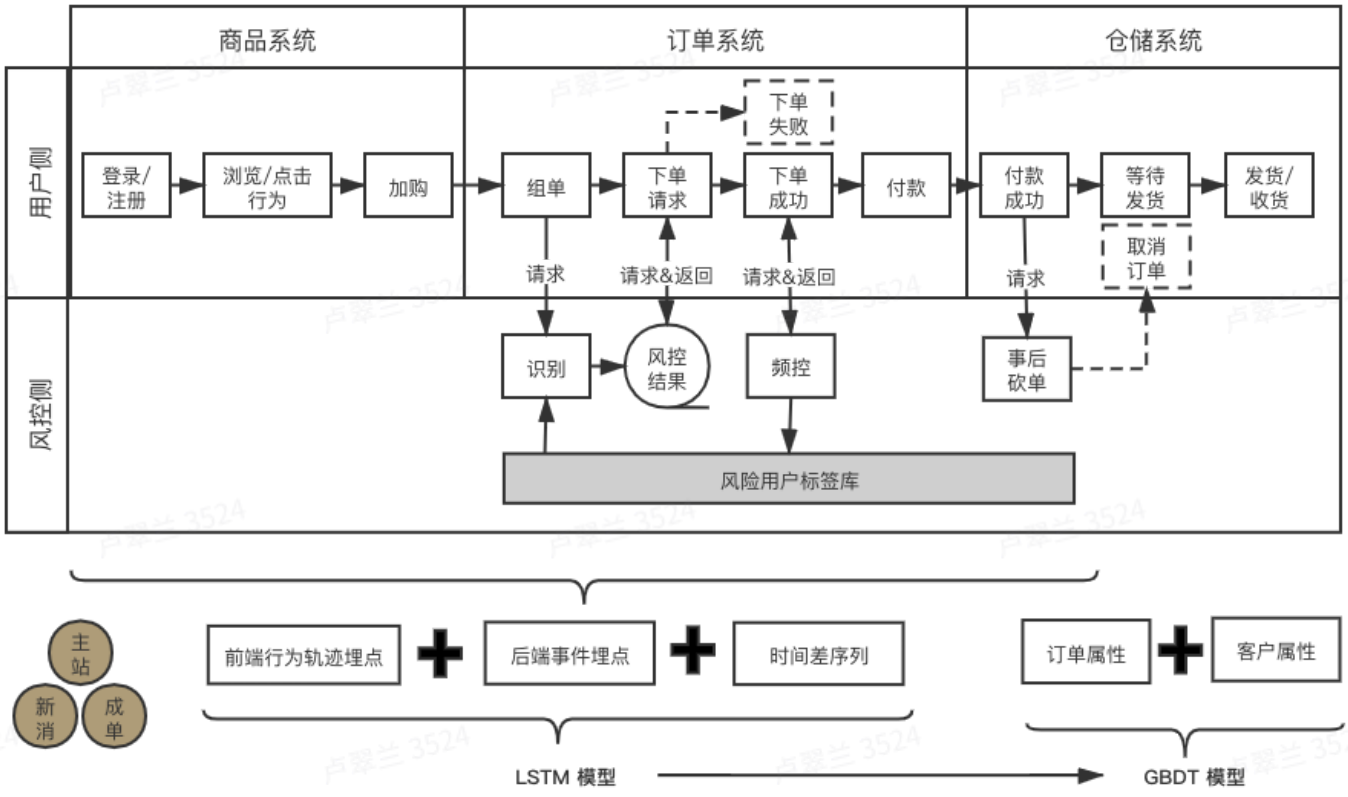
**截至 2020 年 8 月 25 日，模型准确率 90.88% , 召回率 74.96% ，每日风险水位为 3-8%。**

## 一. 项目背景

风险用户定义为**黑产团体**，包含代购，不考虑个体薅羊毛、账户盗用等情况。黑产为了最小化成本，常见行为有：重复使用设备/支付ID下单、在第三方平台(如微信)指定操作教程指导他人代下单等，可以认为该类风险订单的**行为链路较固定**。

目前策略侧已针对新人营销活动、礼品卡、积分等各业务场景完成黑白名单、频控等策略部署，并通过图关联进行“同人”识别，对于在大团伙组单行为进行持续的拦截。

风险水位项目可为策略方提供①**当前风险水位概览、②风险水位变化监控**。具体地，风险水位识别模型通过前端行为轨迹埋点、后端事件埋点、行为时间差序列、订单/客户属性学习黑产固定的行为模式，对当日**主站支付成功的新消订单**(不包含线下店订单)进行模型预测，返回每个订单的风险评分标签。通过评分分布来判断当日高风险用户占比是否异常，从而提升对未知、新增的风险行为的响应时效，协助风控策略的及时优化。



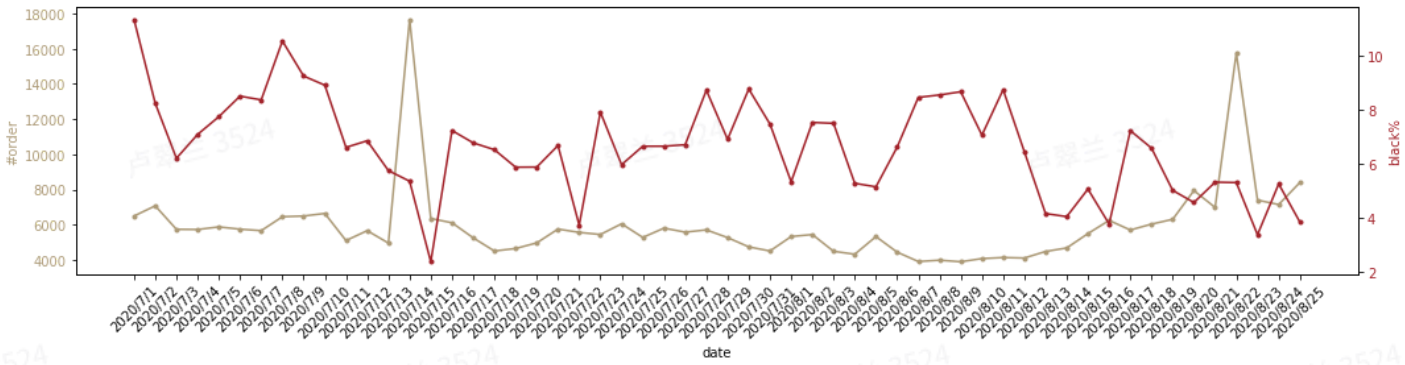
# 二. 方案介绍

下面从数据和模型两个角度分析风险水位整体的模型方案。

## 2.1 输入数据

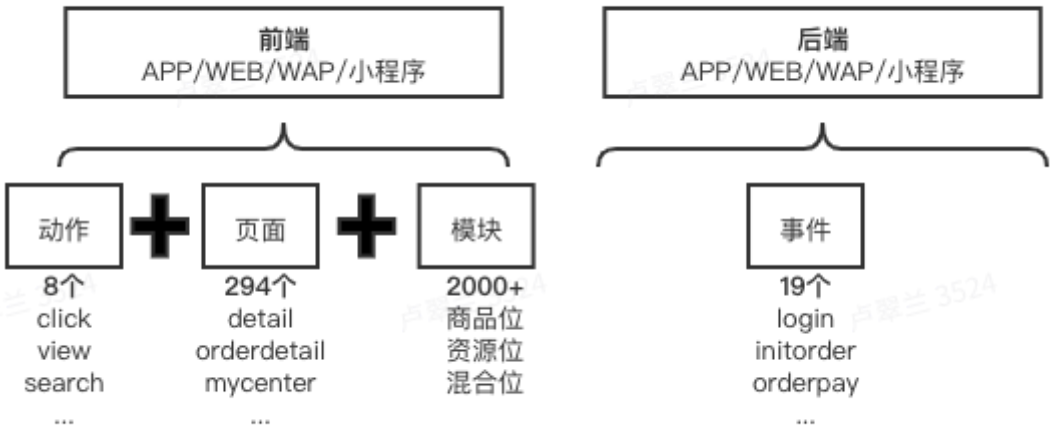
### 2.1.1 黑白标签

策略方提供了黑白标签的数据表，打标逻辑为：**事后(T+1)**根据下单主体(用户 id、收货手机等)信息为新消订单打上风险标签。 2020 年 7 月 1 日-2020 年 8 月 25 日的新消订单及风险订单占比如下图所示(日均订单数近 6,000 单，黑样本平均占比 6.61%)：

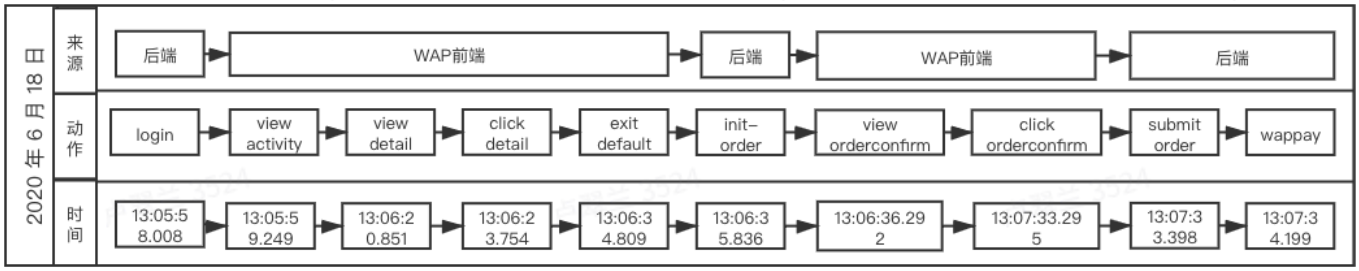



### 2.1.2 埋点数据

前端埋点由 **动作、页面、模块** 三个部分组成，后端日志主要由 **登录、支付**等事件触发。为了避免行为类别太多导致的数据稀疏、新增营销活动导致模块的变动频繁等情况，前端的“行为”定义为动作+页面组合。




通过埋点日志的记录可收集当天新消订单在支付成功之前的所有行为轨迹，例如，2020 年 6 月 18 日的某个新消订单行为序列如下图所示：




**预处理：**

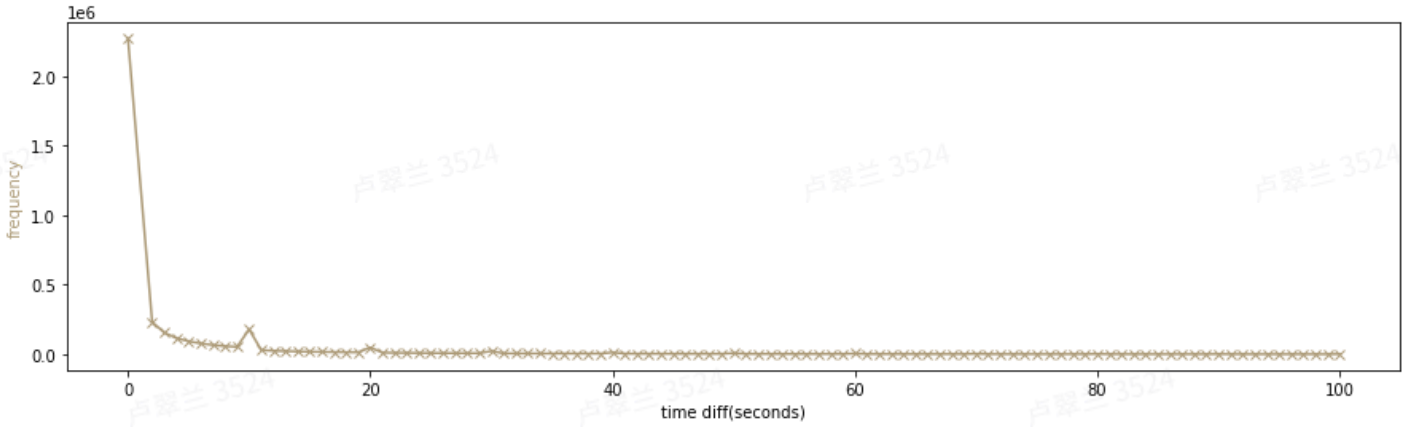
- **动作：**

对训练数据(2020年 3 月 1 日-2020 年 6 月 30 日）的行为轨迹进行分析，为保证行为序列表达的丰富性，取支付成功前99个行为组成序列。

| 黑白标签 | 订单数 | 路径长度 | | | | |
|---|---|---|---|---|---|---|
| | | 最小 | 最大 | 平均 | 标准差 | 中位数 |
| 0 | 271,805 | 1 | 10,226 | 95.63 | 136.23 | **48.84** |
| 1 | 19,841 | 1 | 3,394 | 67.95 | 103.88 | **41.43** |

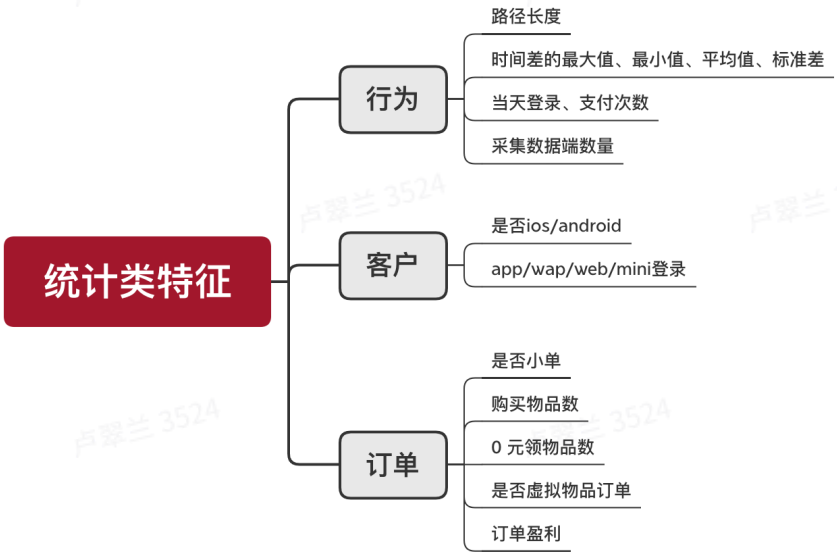
- **时间：**

经分析，时间维度(毫秒级)除了可以对动作进行排序，构成有序行为序列外，相邻动作的时间间隔也具有指向性（固定/重复的行为操作时间间隔更短）。但由于时间差分布具有典型的长尾效应，对于超过 3 小时以上的时间取 3 小时为上限，并完成归一化。




### 2.1.3 订单及客户属性

除了行为轨迹外，增加客户属性、订单属性等统计类特征，如下图：



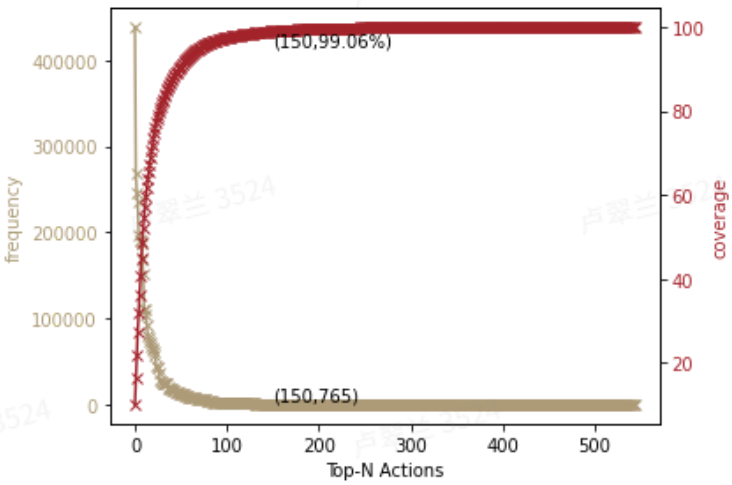

## 2.2 模型算法

风险水位项目采用了 2 个模型：LSTM 模型学习、预测行为序列和时间差序列，GBDT 模型把 LSTM 预测的风险概率作为输入，融合统计类特征完成最终的风险评分预测。在此之前，行为序列需要使用 embedding(词嵌入)的形式完成数值表征。

### 2.2.1 Embedding

常见动作及动作频率统计如下：

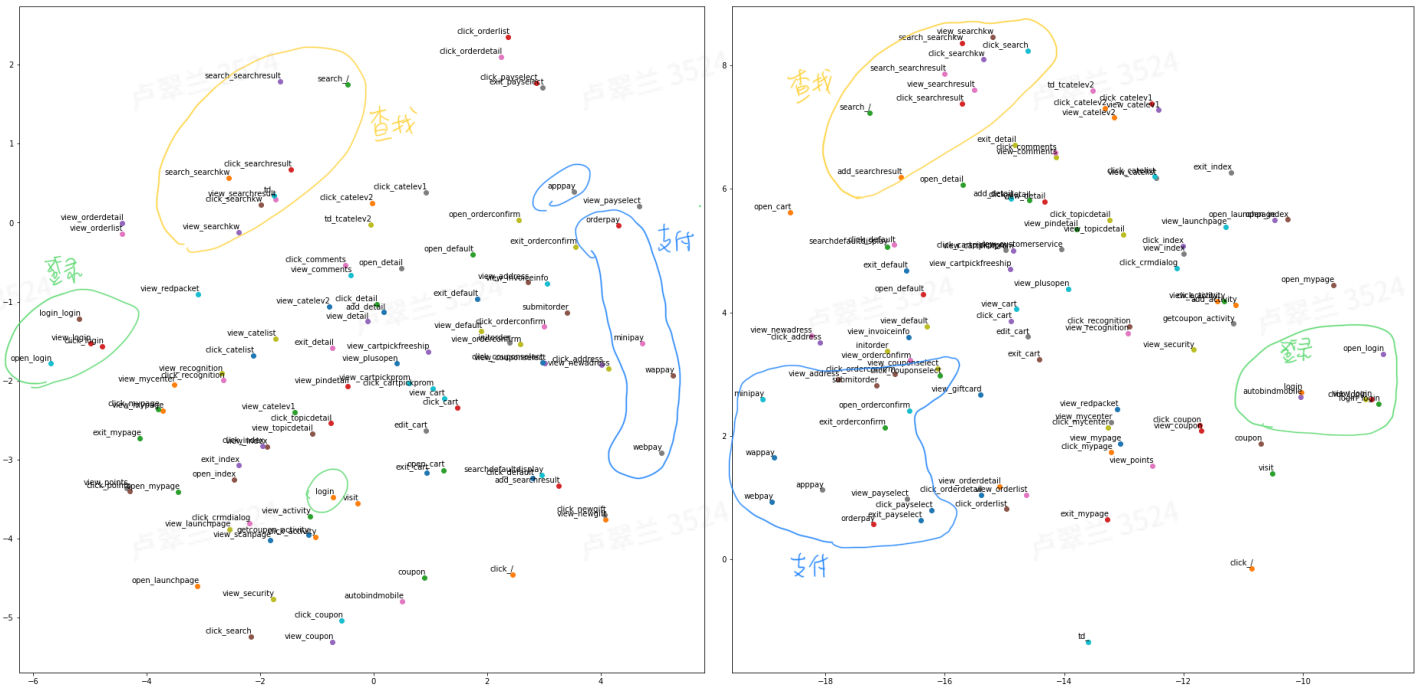
| 排序 | 动作 | 排序 | 动作 |
|---|---|---|---|
| 0 | login | 15 | view_searchresult |
| 1 | click_detail | 16 | apppay |
| 2 | view_detail | 17 | view_default |
| 3 | click_cart | 18 | add_detail |
| 4 | click_orderconfirm | 19 | view_mypage |
| 5 | view_orderconfirm | 20 | click_mypage |
| 6 | view_index | 21 | searchdefaultdisplay |
| 7 | initorder | 22 | click_searchresult |
| 8 | view_cart | 23 | click_searchkw |
| 9 | exit_default | 24 | view_searchkw |
| 10 | view_activity | 25 | view_orderlist |
| 11 | click_activity | 26 | view_payselect |
| 12 | orderpay | 27 | view_comments |
| 13 | click_index | 28 | view_launchpage |
| 14 | submitorder | 29 | click_payselect |




与时间差序列一样，行为序列也存在长尾现象，baseline 模型使用one-hot离散化表示最高频的 150 个动作，取得了不错的效果。但考虑到 150 维训练较慢，可表达的动作有限，使用词嵌入的表示方法对模型进行优化。

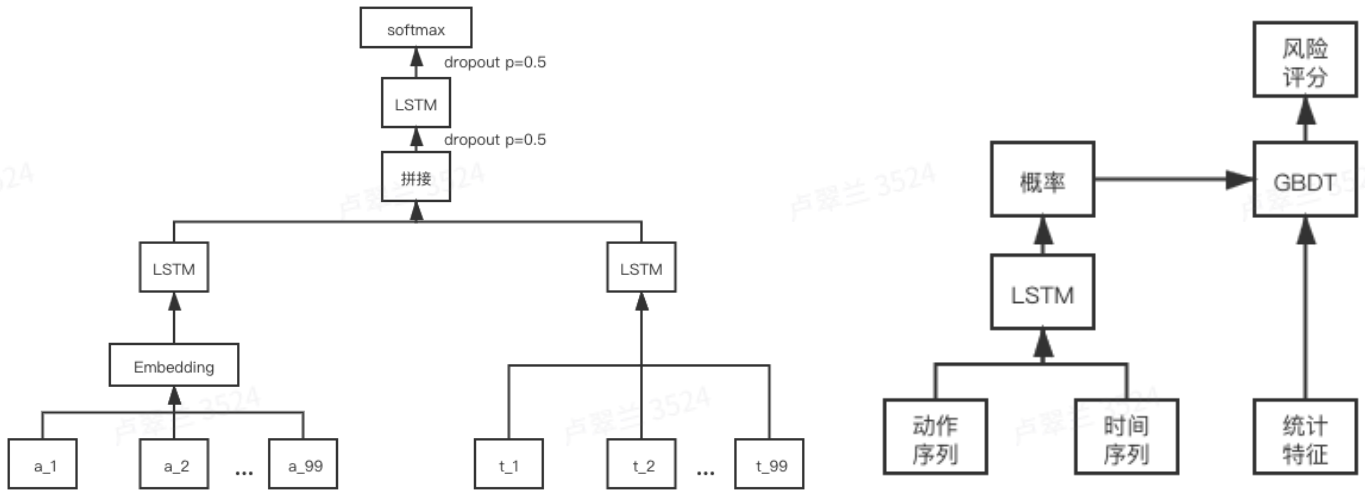
下图为 word2vec(左) 与 fasttext(右) 的embedding 表示可视化后的效果。由于fasttext除了动作前后组合以外，还会考虑动作的内部表达(比如 click_xxx、xxx_searchkw)，相似动作之间在数值表示

上更相近，因此最终采用的是 fasttext 算法，以20 维数值向量表示动作的预训练结果。




### 2.2.2 LSTM

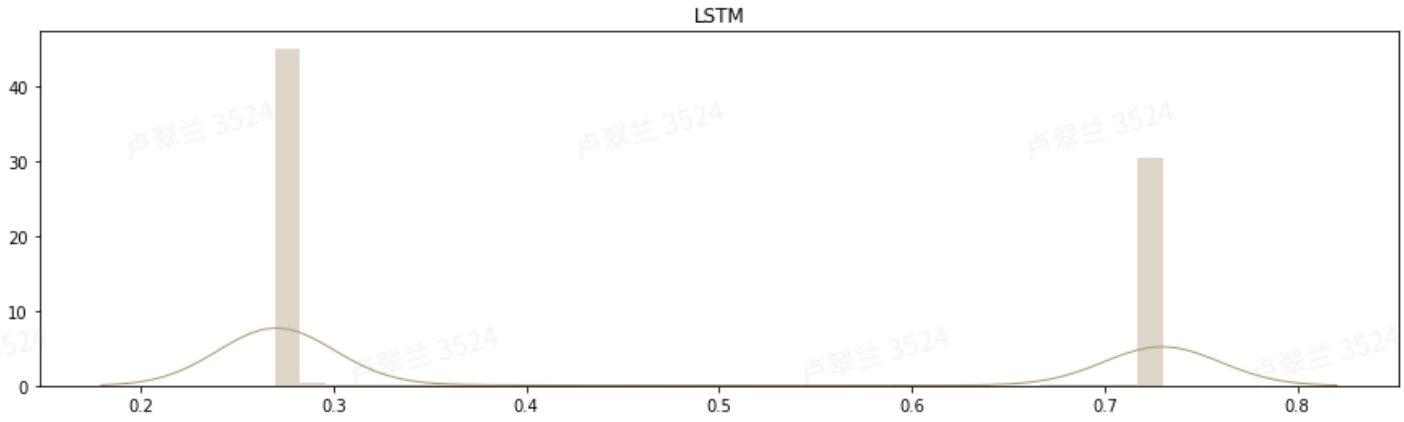
LSTM 适合对时序特征进行处理。经试验，时间差序列先过一层 LSTM 再拼接，会比直接拼接在动作后面效果会稍好一些。风险水位 LSTM 部分的架构如下左图：

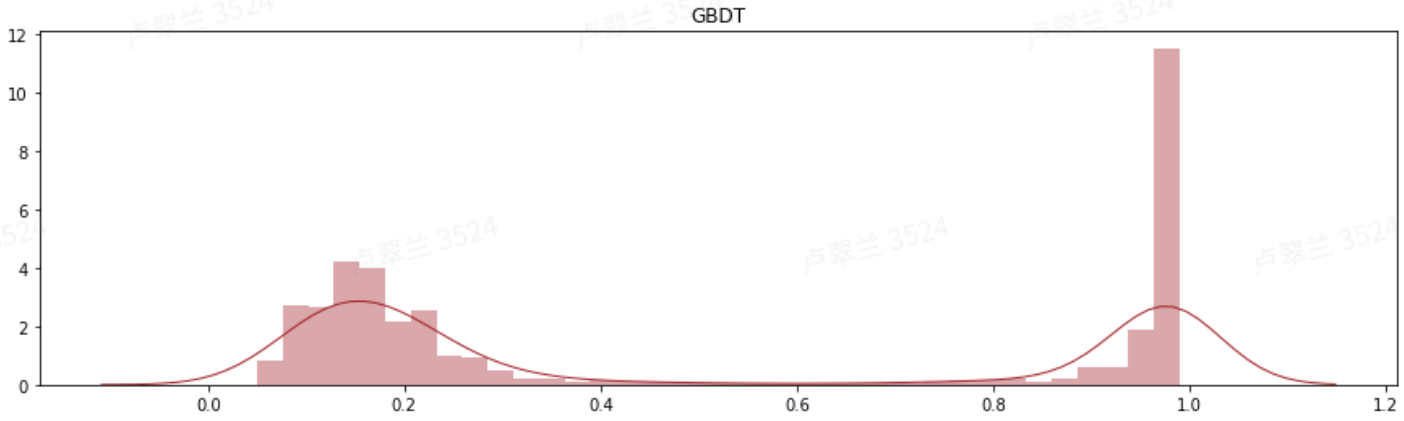



### 2.2.3 GBDT

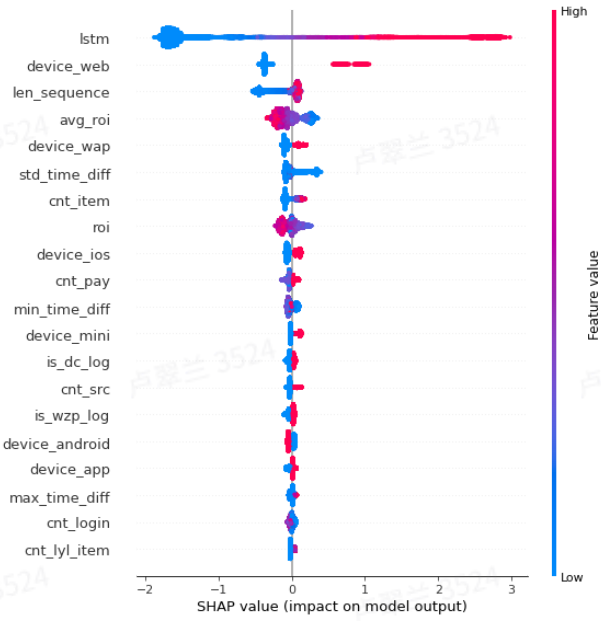
由于订单属性、客户属性这类统计类特征缺乏时序特性，而GBDT 适用于非时序、稠密型变量，因此使用 GBDT 模型在 LSTM 模型的基础上做一层 stacking。从LSTM、GBDT得到的概率分布可以看

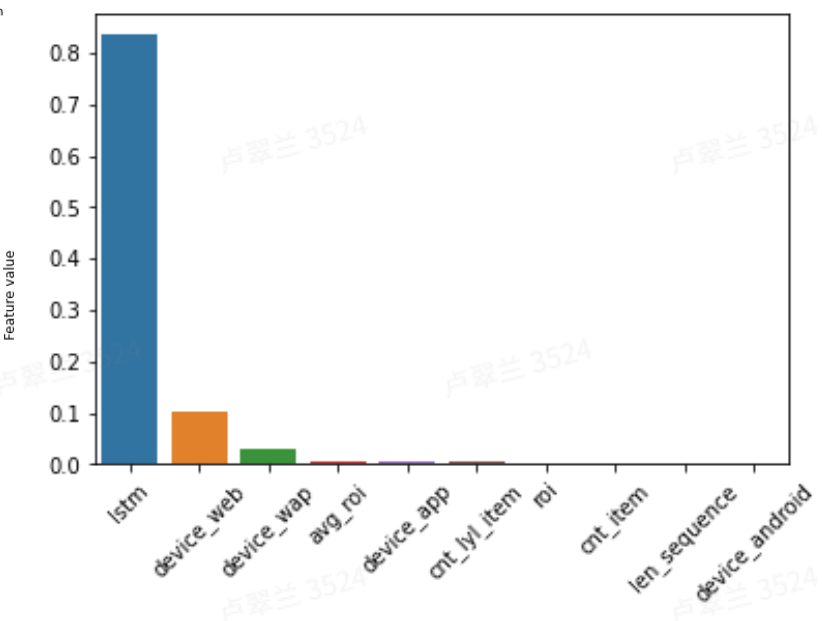
出，GBDT 模型对 LSTM 的预测结果有进一步细化的作用：




从其特征重要性分析，GBDT的预测结果强依赖于 LSTM 的预测概率：



# 三. 效果分析

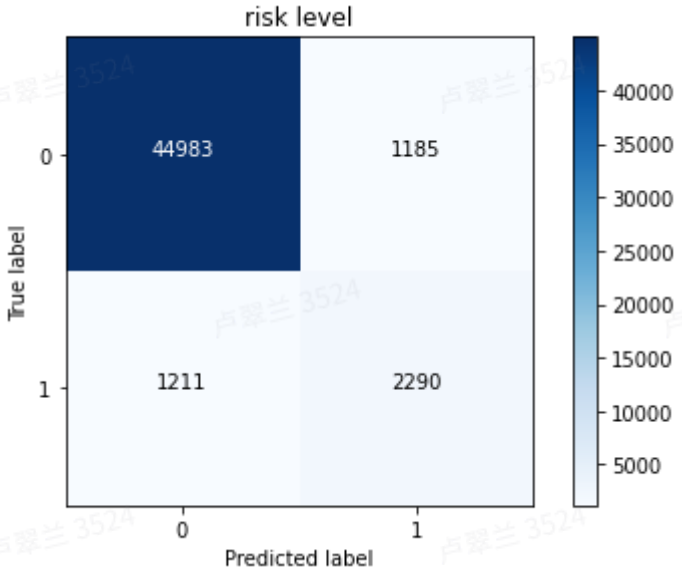
## 3.1 模型指标

使用 2020年 3 月 1 日-2020 年 6 月 30 日共计3 个月的新消订单作为训练数据（去除包含预约、抢购、特供SKU的订单），对训练数据的黑白样本进行 1:1 抽样，并按 7:3 随机分为训练集和测试集。使用 2020 年 7 月 20 日-2020 年 7 月 29 日共计 10 天的新消订单作为跨时区验证，整体模型效果如下：

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | 模型 | LSTM(阈值 0.70) | | | GBDT(阈值 0.95) | | |
| 2 | 数据集 | 训练集 | 测试集 | 跨时区验证集 | 训练集 | 测试集 | 跨时区验 |
| 3 | 样本量 | 52,363 | 22,442 | 49,669 | 52,363 | 22,442 | |
| 4 | 黑样本占比 | 49.61% | 49.34% | 7.05% | 49.61% | 49.34% | |
| 5 | 正确率 | 88.53% | 85.67% | 92.14% | 77.51% | 77.62% | 9 |
| 6 | 准确率 | 97.35% | 94.90% | 46.54% | 99.02% | 98.77% | **6** |
| 7 | 召回率 | 78.98% | 75.10% | 77.12% | 55.21% | 55.34% | **6** |
| 8 | AUC | 90.99% | 89.00% | 89.60% | 91.92% | 91.91% | 8 |
| 9 | KS | 78.09% | 71.20% | 70.33% | 75.96% | 75.93% | 7 |

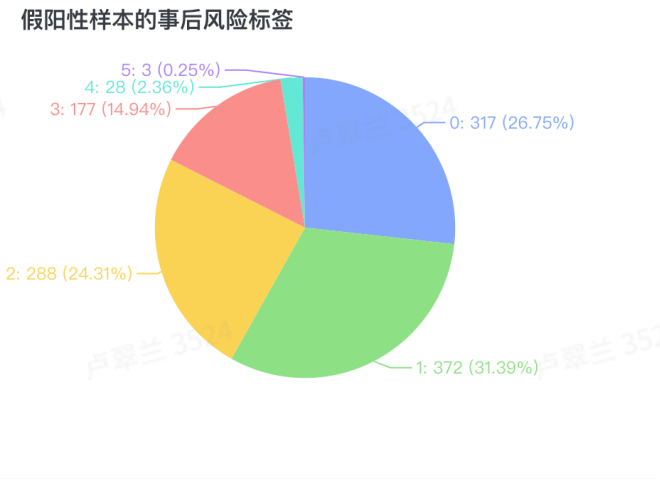
## 3.2 案例分析

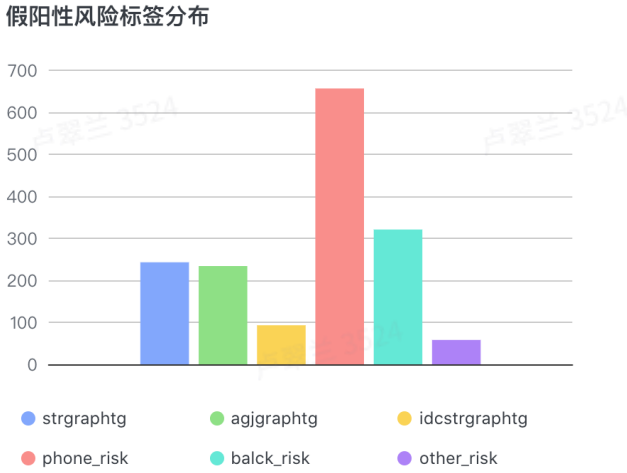
跨时区验证集的黑白样本占比为全量还原情况，GBDT 模型预测结果(概率≥0.95的为黑样本)与原始黑白标签的混淆矩阵如下图所示。在黑白标签为 100%准确的前提下，有1,185个白样本被模型判定为黑样本(假阳性False Positive)，有 1,211个黑样本被模型判定为白样本(假阴性False Negative)：



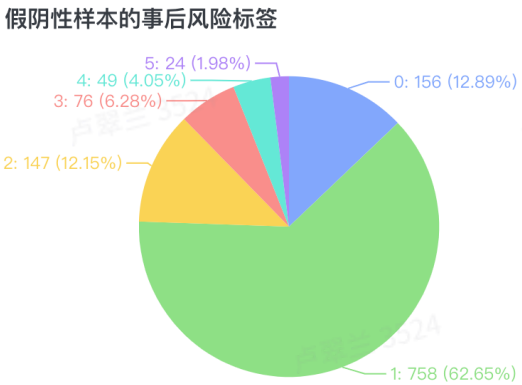
值得注意的是，如果没有及时拦截黑产用户，那么随着时间推移，其风险表现将逐渐暴露。黑白标签使用 T+1 的信息界定风险标签，下面分别对错分的样本(假阳性、假阴性)在 2020 年 8 月 27 日(约 1 个月)的事后风险标签进行分析。

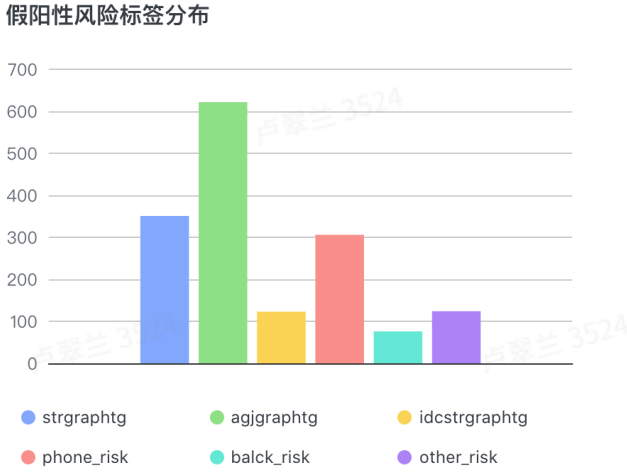
- **假阳性：**73.25% 的样本在一个月后打上了风险标签（图标签判定阈值为≥20），说明模型具备发现未知风险的能力。经此调整后的准确率为**90.88%**。




- **假阴性：**87.11% 的样本在一个月后仍存在风险标志，说明策略方标签准确率非常高，但本模型的召回率不高，为 **74.96%** 。



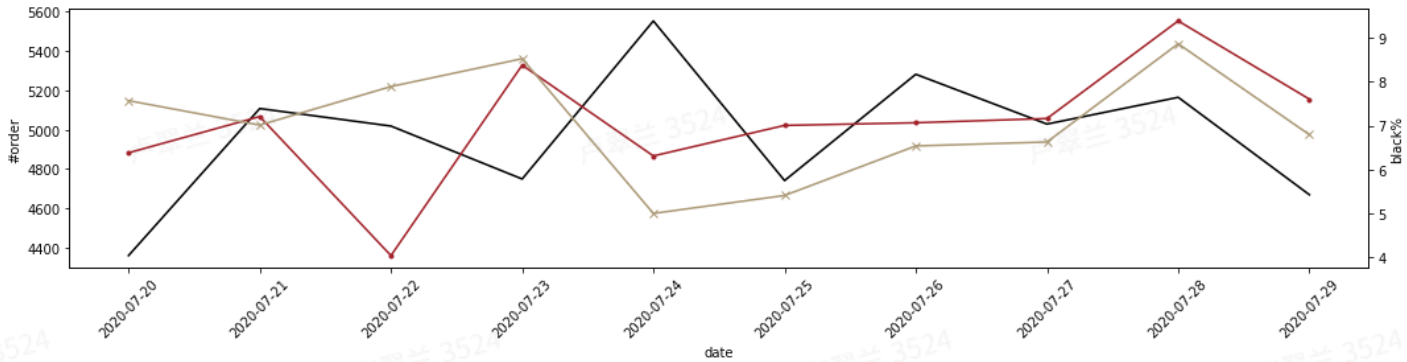



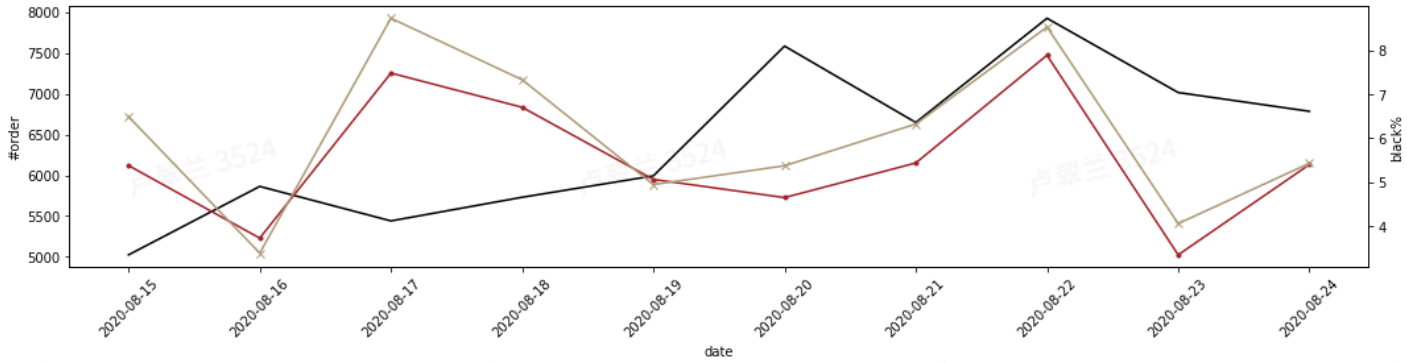

## 3.3 水位分析

从下图可以看出，模型拟合的效果（金色）与黑白标签（红色）相接近，但因为召回率不高，整体水平偏低。7 月 20 日-7 月 29 日、8 月 15 日-8 月 24 日两个时间段的**风险水位大约为 3-8%**。

值得注意的是，在 7 月 22 日黑白标签占比为4.03%，模型预测风险占比为 7.89%，占比提升 96.02%，即可**通过行为轨迹及时发现黑产团伙的异常风险**（7 月 23 日策略方发现首单全额返活动的

新风险点）。




# 四. 应用说明

通过 T+1 运行模型，并将每日新消风险以监控报表的形式产出，建议针对预测为高风险的新消订单购买的商品、使用的红包/礼券进行**案例分析**，特别是策略方未标注为黑样本的部分订单值得重点关注。

另外，建议在出现以下情况（后续待更新）则进行**风险提示**：

1. 与策略提供的黑样本占比相比，提升 50%时；
2. 与模型历史评分分布占比相比，PSI 变化高达 10%时。

# 五. 未来工作

### 5.1 监控报表制作

### 5.2 模型优化：

1. 提高未召回样本的识别程度；
2. 与前后端沟通，随着埋点的页面/动作变化节奏，更新模型，保证模型的鲁棒性；
3. 随着营销节奏的变化，放开到对所有订单进行监控。