风险水位模型优化项目

一. 项目背景

原有风险水位识别模型通过用户的前端动作轨迹、后端事件埋点、行为时间差序列、以及订单/用户属性学习黑产团体（包含代购，重复使用设备、支付ID下单，以及第三方平台指导他人代下单等团体）的固定行为模式。之后利用训练模型对(T+1)昨日主站支付成功的新消订单（不包含线下店订单）进行预测，返回每个订单的风险评分标签。通过当日的评分分布来判断高风险用户占比是否异常，以此来预警未知的风险行为的出现并提高策略方响应时效。

原有训练模型采用了1:1的黑白样本集。输入的前端动作轨迹由**动作**、**页面**两部分组成，可能的组合方式共有2352种。为了避免在上线预测时遇见未经过表征化的动作组合，在预处理中将未表征化的动作组合由该组合中的动作部分替代（比如click_xxxx 被click替代）。利用LSTM对表征化后的行为轨迹（包含前端动作组合和后端事件）和时间差序列进行学习和预测，用GBDT模型把LSTM输出的风险概率和用户的统计类特征作为特征，完成最终的风险概率预测。

虽然训练结果的准确率和召回率较为可观，但在抽样后得出模型鲁棒性不足，线上指标效果与离线相差较大。并且随着前后端埋点的页面、动作节奏变化，线上结果存在黑样本（风险样本）召回率不足的问题。模型对于真实业务环境中的黑样本识别准确率下降到了77%，召回率下降到了37%。因此提高模型的召回率又同时保证准确率，是此次优化项目的最终目标。

二. 优化方案

优化方案将从**数据**和**模型**两个角度进行介绍

2.1 输入数据

2.1.1 黑白标签

在训练优化模型时，采用了靠近真实黑白样本分布的数据集（黑样本占比约8.44%）。希望缩减离线与线上的模型指标距离并提高模型上线后对当日中占较小比例的风险用户的识别能力。在模型上线后，其时效性从原本**T+1**提升到了**准实时**。有效提高了风控策略的及时响应和优化。

2.1.2 埋点数据

优化后的训练模型采用了2020年9月到2021年3月分区表的数据，总计约900000条记录。完善了之前训练样本中个别用户前端动作轨迹缺失的问题。同时分区表中记录的前后端行为更加贴近用户当下在严选主站消费的行为轨迹，提高了预测结果的准确性。

2.2 模型算法

新的风险水位模型采用bidirectional LSTM网络和LGB框架运行GBDT，在预处理过程中对文本进行半遮盖处理，添加attention层，并利用了交叉验证完成了超参优化。又因为输入数据中黑白样本失衡，在建立bidirectional LSTM和LGB模型时手动初始化了偏置向量bias以提高模型训练效果。

2.2.1 Embedding

优化后的风险水位模型仍考虑到前端动作组合（动作+页面），利用fastText进行学习。同时为了增强模型的鲁棒性，**通过对动作组合采取半遮盖处理的方式在行为序列中学习前端动作本身的数值表征**，用于后续埋点更新的OOV表征。具体来说，优化后的风险水位模型在训练样本中随机选择约30000条样本（占比3.1%），并在所选样本中随机挑选3个动作组合，遮蔽其中的页面部分，利用fasttext算法对所有样本进行embedding，最终得到所有已知的单个前端动作本身、前端动作组合、后端事件的数值表征。

2.2.2 LSTM

LSTM模型部分的主要优化是将LSTM模型升级为Bidirectional LSTM模型和增加了attention（注意力）层

Attention 层

风险用户群体存在明显的异常动作及相关轨迹，比如说会频繁更换手机以及登入登出。如果在训练LSTM时将有限的注意力集中在该类重点信息上，就能获得更多所需要关注的风险用户的细节信息，并抑制其它无用信息。下面是将attention层置于embedding层后、LSTM网络前的attention输出结果可视图。（颜色越红，值越高；反之，颜色越绿，值越接近0）

login	login_login	view_index	login	view_mypage	click_mypage	view_index	exit_index	login	view_cart	view_index	search_searchre
0.027416993	0.007813412	0.002566242	0.01233249	0.008097552	0.0481114	0.01305519	0.013936457	0.006724651	0.009262459	0.0224556	0.005251174
view_activity	login	click_activity	view_cart	view_index	click_index	view_index	click_index	click_index	view_catelev1	search_searchre	click_searchkw
0.017930476	0.004170072	0.001194157	0.005033554	0.003859295	0.001815706	0.00548698	0.024116615	0.013937777	0.008826544	0.001186289	0.005144992
view_activity	autobindmobile	login	view_activity	login	login	view_index	click_index	view_signin	login	click_signin	view_index
0.000273038	0.20001686	0.003090137	0.012130106	0.000857935	0.001118871	0.010402794	0.024424545	0.003308571	0.00169513	0.002267405	0.001944493
login_login	view_mypage	view_index	click_default	view_index	click_default	view_msgcenter	click_msgcenter	view_customers	login	view_activity	click_default
0.012243037	0.000553458	0.010959891	0.10653718	0.019256845	0.001812562	0.001967354	0.007788671	0.02527551	0.039051086	0.002605451	0.002098852
login	login	login	login	login	login	view_orderlist	login	searchdefaultdis	view_detail	login	visit
0.084083974	0.006036541	0.001333554	0.04768697	0.003550236	0.021415083	0.001912016	0.002412507	0.000572804	0.01086655	0.003335837	0.001135855
login	view_detail	view_detail	view_detail	click_detail	click_detail	click_detail	click_detail	click_detail	view_detail	view_detail	click_detail
0.004075048	0.005382965	0.01285709	0.002697695	0.003061298	0.022972329	0.002496582	0.009326345	0.00146505	0.010770769	0.004529591	0.002504207
login	login	login	login	login	login	login	login	login	login	login	login
0.007277309	0.010606231	0.007970678	0.002377391	0.034463055	0.001864819	0.005969382	0.005577401	0.00267864	0.003693933	0.007989868	0.002323375
view_launchpag	view_index	login	login	click_launchpage	click_launchpage	click_launchpage	click_launchpage	click_launchpage	click_launchpage	click_launchpage	click_launchpage
0.043837998	0.000330266	0.09347111	0.011579483	0.002402097	0.000297392	0.006980444	0.048423607	0.000209504	0.021377016	0.015990445	0.022641998
login	login	login	login	view_orderlist	login	view_detail	login	visit	view_activity	view_detail	click_activity
0.010059512	0.002029091	0.005829494	0.022839628	0.005412174	0.006039334	0.00866188	0.003647131	0.01528875	0.001647293	0.002324809	0.001320727
login	view_orderlist	login	searchdefaultdis	login	view_detail	visit	view_detail	view_activity	click_activity	add_detail	searchdefaultdis
0.011222787	0.004416269	0.014312978	0.001059296	0.001132703	0.003191679	0.005412462	0.019088808	0.021176953	0.002724302	0.001992296	0.012173297
login	login	login	login	view_detail	click_detail	click_orderconfi	click_orderconfi	view_payselect	login	login	visit
0.00625784	0.014989981	0.003076582	0.001563551	0.014265856	0.022865823	0.07068117	0.008847455	0.002116371	0.002844584	0.06260129	0.001278632
view_launchpag	view_index	exit_default	open_default	view_login	click_login	view_mypage	view_index	login_login	login	view_index	click_default
0.002723879	0.001261717	0.012074396	0.001081784	0.001275435	0.004256911	0.006439208	0.017731048	0.019023988	0.002796473	0.001512205	0.003930523
login	login	login	login	login	login	login	searchdefaultdis	click_default	view_index	view_orderlist	view_default
0.01129313	0.002806578	0.015917884	0.002921325	0.008233722	0.044541128	0.001125328	0.0359478	0.027635932	0.013118355	0.000811567	0.003201851
login	login	login	login	coupon	login	login	login	view_orderlist	login	view_detail	visit
0.001668489	0.003143479	0.029048648	0.002687987	0.003032971	0.006308933	0.02747772	0.004990418	0.007949534	0.000855114	0.000913398	0.035324253
click_login	login_login	view_index	view_mypage	view_index	click_default	view_msgcenter	click_default	click_msgcenter	view_customers	login	view_activity
0.001009441	0.002428091	0.01462202	0.008141511	0.004194539	0.000570825	0.003379683	0.011771445	0.001714167	0.002579391	0.000989687	0.00229032
login	searchdefaultdis	view_default	searchdefaultdis	login	view_detail	visit	view_activity	searchdefaultdis	view_detail	click_activity	add_detail

如上图所示，attention层学习到了文本本身的内容。attention层认为在学习风险用户行为时autobindmobile, login, click-default等动作是值得注意的。之后模型会将attention层输出结果与行为轨迹的表征值相结合最终传递给LSTM网络。

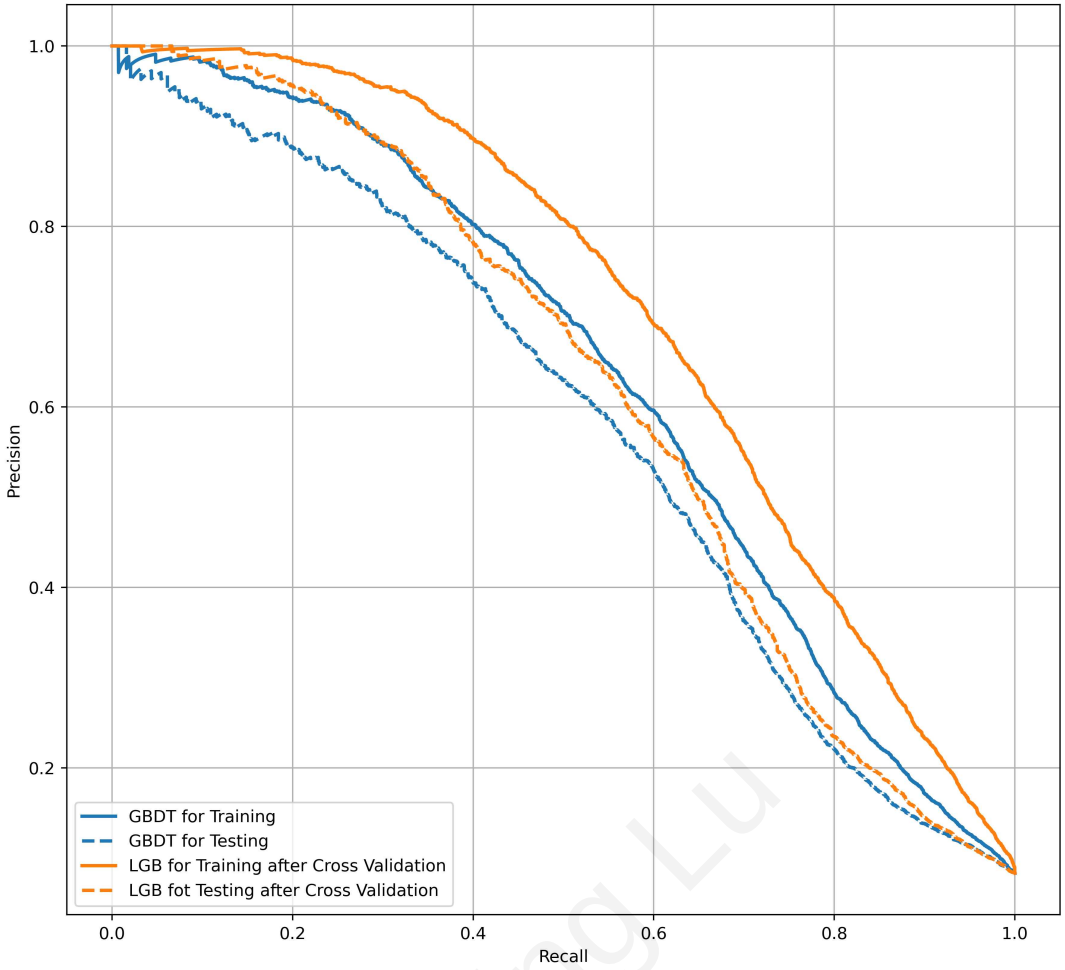
Bidirectional LSTM

标准的循环神经网络（RNN）能够存取的上下文信息范围很有限。这个问题就使得隐含层的输入对于网络输出的影响随着网络环路的不断递归而衰退。在真实业务场景中，风险用户虽然有固定的行为链路，但是如果行为链路较长，单向学习该行为链路难免会造成对序列后端的记忆不深从而导致信息利用不足。并且相关风险行为可能也会出现在正常用户的行为轨迹中，以单向路径判断很可能会将该用户误判为风险用户从而导致低准确率。因此使用bidirectional（双向） LSTM网络分别从”上下文“两个方向学习行为路径能避免以上问题所带来的负面影响。实验结果也表明，双向LSTM网络确实对最终模型的PRC曲线有优化效果。下图是针对黑样本的最终识别效果图

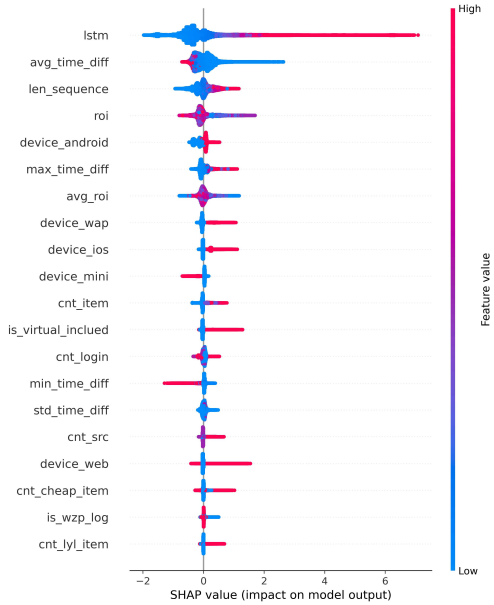
		准确率	召回率	F1
训练集	LSTM	0.66	0.52	0.58
	BILSTM	0.71	0.57	0.64
验证集	LSTM	0.65	0.52	0.58
	BILSTM	0.70	0.57	0.63
测试集	LSTM	0.62	0.50	0.56
	BILSTM	0.69	0.56	0.62

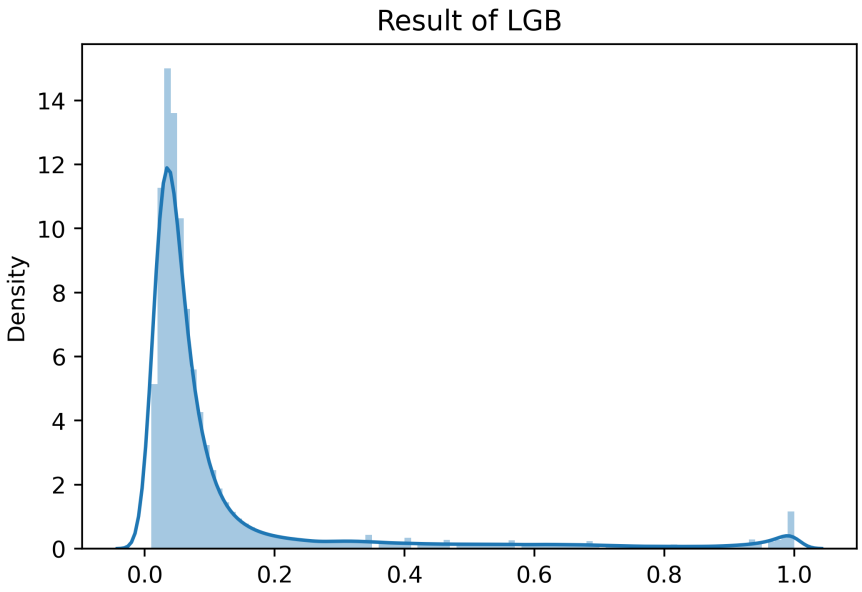
LGB和超参优化

采用了LightGBM框架运行GBDT。与原本的sklearn相比，LGB有更好的分类表现。同时使用了网格搜索和随机搜索两种方式对超参进行筛选，设置召回率为评价指标。下图显示了LGB在超参优化后在训练集和测试集上的表现都优于GBDT



从特征重要性角度分析，LGB的预测结果仍强依赖于LSTM的预测结果, 但其他特征的重要性有所变化（左图）。右图为LGB输出的概率分布，可以看出概率分布细化并符合黑白样本的不均衡分布。





三. 效果分析

3.1 训练效果

使用2020年9月到2021年3月共计6个月的新消订单作为训练数据。对训练数据按照7 : 3随机分为训练组和测试组对LSTM进行训练和测试，并按照6.4 : 1.6 : 2的比例将用户统计类特征和LSTM输出数据随机分为训练组，验证组和测试组进行训练，超参优化和测试。 整体模型效果如下：

模型	BILSTM（阈值0.5）		LGB（阈值0.5）		
数据集	训练集	测试集	训练集	测试集	验证集
样本量	677211	290234	619164	154792	193489
黑样本占比	8.46%	8.38%	8.46%	8.41%	8.40%
准确率	0.72	0.69	0.71	0.70	0.69
召回率	0.42	0.41	0.57	0.57	0.56
F1值	0.53	0.51	0.64	0.63	0.62