信也科技杯 Data Science Competition | 信也 FINVOLUTION

FRAUD USER DETECTION

聚焦数创·连接未来

第七届信也科技杯算法大赛

7th Finvolution Data Science Competition

欺诈用户风险识别

2022 第七届信也科技杯图算法大赛

团队名称：ppd小试牛刀

日期：2022.09.17

目录

ppd小试牛刀

中原消费金融 ZHONGYUAN CONSUMER FINANCE

1.金融风控领域的从业者

2.金融风控领域算法探索践行者

- 蚂蚁金服--支付风险识别 TOP1
- 中国银联--信贷逾期预测 TOP2
- 融360--拒绝推断 TOP3
- 信也科技--现金流预测 TOP4
- 也有其他竞赛多次top3经历

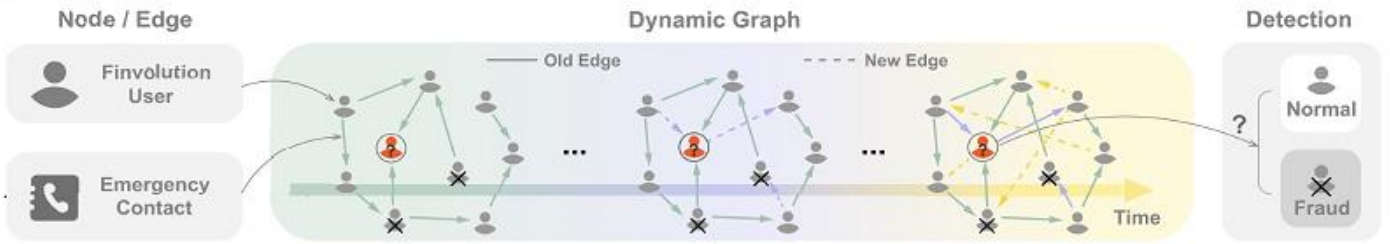
最终得分：0.81395

排名：3

与第一名差距：0.00073

与第二名差距：0.00015

任务目标：根据图节点基础特征群X，以及节点有向边edge，预测用户是否欺诈



数据表：

a.节点特征表：X

	id	feat_0	feat_1	feat_2	...	feat_15	feat_16	label
0	0	0.0	2.0	-1.00	...	-1.000000	-1.000000	2
1	1	0.0	2.0	0.68	...	0.007353	0.142857	0
2	2	0.0	2.0	-1.00	...	-1.000000	-1.000000	2
3	3	0.0	3.0	-1.00	...	-1.000000	-1.000000	2
4	4	0.0	-1.0	-1.00	...	-1.000000	-1.000000	2

b.节点之间的关联关系：edge

from_id	to_id	type	time
A	B	夫妻	2021/8/1
C	D	朋友	2021/4/2
B	C	同事	2021/5/3
D	C	姐妹	2021/8/4
E	F	亲戚	2021/8/5

主办方提供了风控场景的非常经典的图数据，这也是我们日常风控工作中需要处理面对的一类数据。关于此类数据的处理，我们日常工作中有两种思路：

基于图数据，人工衍生传统的关联类特征，用聚合的方法将关联类用户的信息表示出来，再用传统模型建模

- 衍生出来的特征更加有业务含义，模型学习到什么信息可以心中有数
- 可以让模型同时学习出不同边类型、不同时间窗的关联用户的关联特征，基础统计特征就可以表示出关联用户的风险信息。
- 关于特征的想象力有多远我们就能走多远

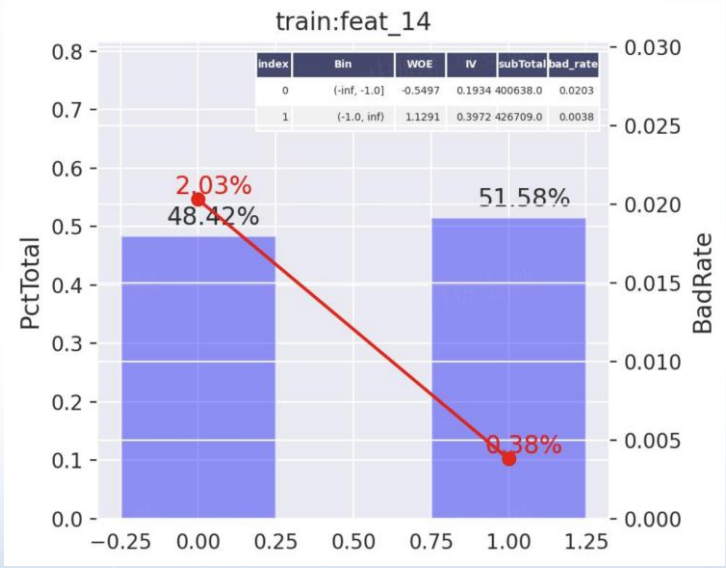
VS

基于图数据，利用主流图深度学习算法，例如GCN等传统算法，以及可以学习不同边类型权重的异构图和可以学习不同时间段节点信息的动态图等

- 无需过多的衍生特征，对关联用户特征学习出不同的特征加权权重，可以得到比基础统计特征更高的模型精度
- 不同的网络结构可以学习出不同边类型，不同时间窗下的关联用户信息。
- AI可以打败人工，AI的创造力无法想象

EDA: 到底什么对预测用户欺诈重要?

对原始特征群X的分析:



EDA: 到底什么对预测用户欺诈重要?

对关联关系:

作为联系人类型:

index	Bin	WOE	IV	subTotal	bad_rate
0	(-inf, 2.0]	0.6998	0.0398	112591	0.0062
1	(2.0, 4.0]	0.1015	0.0022	220150	0.0113
2	(4.0, 7.0]	0.0496	0.0013	522333	0.0119
3	(7.0, 9.0]	-0.3066	0.009	82453	0.0169
4	(9.0, inf)	-0.7654	0.0566	65398	0.0264

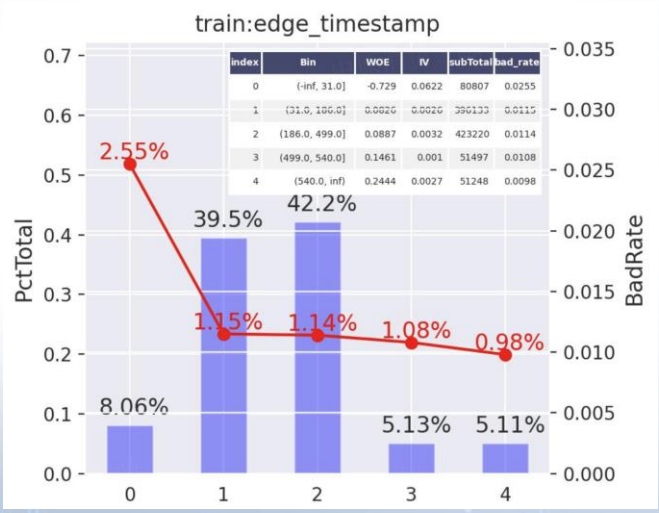
自填联系人类型:

index	Bin	WOE	IV	subTotal	bad_rate
0	(-inf, 1.0]	0.7529	0.0355	140527	0.0019
1	(1.0, 5.0]	0.5963	0.1618	952844	0.0022
2	(5.0, 8.0]	0.4509	0.0309	298052	0.0025
3	(8.0, inf)	-1.4544	0.5629	188773	0.0168

EDA: 到底什么对预测用户欺诈重要?

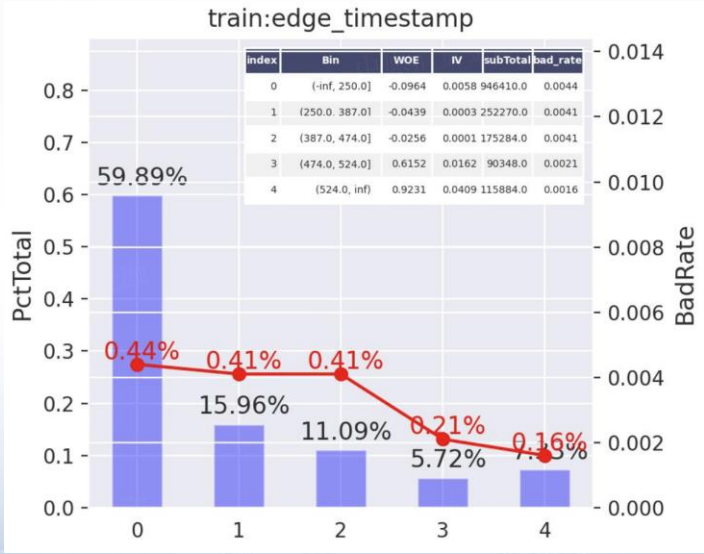
对关联关系:

作为联系人时间:



index	Bin	WOE	IV	subTotal	bad_rate
0	(-inf, 31.0]	-0.729	0.0622	80807	0.0255
1	(31.0, 186.0]	0.0026	0.0026	396133	0.0115
2	(186.0, 499.0]	0.0887	0.0032	423220	0.0114
3	(499.0, 540.0]	0.1461	0.001	51497	0.0108
4	(540.0, inf)	0.2444	0.0027	51248	0.0098

自填联系人时间:



index	Bin	WOE	IV	subTotal	bad_rate
0	(-inf, 250.0]	-0.0964	0.0058	946410.0	0.0044
1	(250.0, 387.0]	-0.0439	0.0003	252270.0	0.0041
2	(387.0, 474.0]	-0.0256	0.0001	175284.0	0.0041
3	(474.0, 524.0]	0.6152	0.0162	90348.0	0.0021
4	(524.0, inf)	0.9231	0.0409	115884.0	0.0016

- 依托于高质量图数据，构造了线下cv 评估方式，不管是初赛还是复赛，线上评分和线下同增减。
- 全面的特征体系框架

A-->B，对A聚合B信息

B-->A，对B聚合A信息

A-->B--->C,对A聚合C信息

A-->B<---C,对A聚合C信息

B<---A--->C,对B聚合C信息

B<---A<---C,对B聚合C信息

A对B信息聚合：

- 关联类型，关联时间，B的X信息，B的背景节点信息
- 信息交叉，如关联类型与关联时间交叉

聚合方式：
max,min,mean,nunique,std,count,sum等

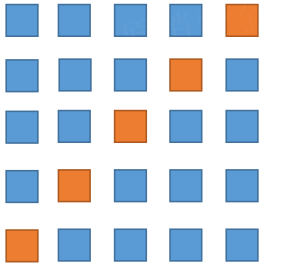
基于5fold lightgbm，我们评估了所构造特征的线下cv评分

特征群	线下得分
基础特征群X	0.7366
A-->B，对A聚合B信息	0.8293
B-->A，对B聚合A信息	0.8467
A-->B--->C,对A聚合C信息	0.8474
A-->B<---C,对A聚合C信息	0.8483
B<---A--->C,对B聚合C信息	0.8488
B<---A<---C,对B聚合C信息	0.8494

变量英文名	变量中文示意	imp
cnt_to_type_l8	作为别人联系类型小于8次数	49800.75
min_type	联系人类型最小值	19975.55
feat_2	初始特征2	9886.963
cnt_to_type_l8_divi_feat_2	作为别人联系类型小于8feat2均值	8516.859
per_from_id_y_1	联系人背景节点占比	4543.87
min_etime	最早联系人关联时间	4510.487
feat_6	初始特征6	4421.145
feat_3	初始特征3	3785.302
feat_12	初始特征12	3579.781
mean_from_feat16	联系人初始特征16均值	3469.81
median_type	联系人类别中位数	3049.551
feat_11	初始特征11	2403.408
feat_1	初始特征1	2203.083
mean_from_feat15	联系人初始特征15均值	2159.19
cnt_from_id_y_1	联系人背景节点个数	2096.883
feat_8	初始特征8	1971.786
max_etime	最晚联系人关联时间	1616.485
mean_from_type	联系人类别均值	1481.195
feat_0	初始特征0	1465.374

思路1：基于传统特征，构建5-folds lightgbm模型

交叉验证思路进行特征选择



robust特征交叉

在筛选出的最稳健的特征集合里，两两交叉，选取iv最高的topN放到原始特征集合里，通过线下验证集效果，确定最终入选的N组交叉特征。

kfold--avgeage

单模型做5个fold，4个fold训练，另外一个fold测试，共5个模型，在最终线上预测时，取5个模型的均值，这可以有效的降低模型的Variance，增加整体模型的鲁棒性。

思路2：基于深度图算法，构建预测模型

基于传统特征构造方式，我们训练了lgb模型，对于数据有了较深的理解和认识，我们认为深度图算法，需要在以下几个环节设计方案：

- 原始特征X缺失值（-1）处理
- 不同边类型、不同时间窗口信息差异化
- 关联到的背景节点如何使用

GraphSAGE (NeighborSampler)

GraphSAGE (NeighborSampler)模型

处理方法add	线下得分
初始处理	0.7885
缺失值处理	0.7913
添加边类型属性	0.8281
添加边时间信息	0.8294
添加背景节点信息	0.8321

思路2：基于深度图算法，构建预测模型

GraphSAGE (NeighborSampler)处理方法：

a.缺失值处理：

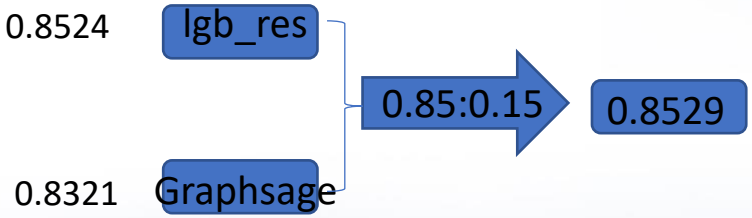
```
x = data['x']
r1 = (x[:, 1] == -1) * 1.0
r2 = (x[:, 2] == -1) * 1.0
r3 = (x[:, 15] == -1) * 1.0
x = np.column_stack((x, r1))
x = np.column_stack((x, r2))
x = np.column_stack((x, r3))
x[x == -1] = 0
```

b.边属性、时间信息，与背景节点信息处理：

```
a = edge_index[edge_index[:, 0].argsort()]
b = np.unique(a[:, 0], return_index=True)[0]
a = np.split(a, np.unique(a[:, 0], return_index=True)[1][1:])

c1 = np.array([np.sum(tmp[:, 2] > 8) for tmp in a])
c2 = np.array([np.sum(tmp[:, 2] <= 8) for tmp in a])

c3 = np.array([np.sum(tmp[:, 3] <= 31) for tmp in a])
c4 = np.array([np.sum(tmp[:, 3] > 480) for tmp in a])

c5 = np.array([np.sum(y[tmp[:, 1]] != 1) for tmp in a])
```

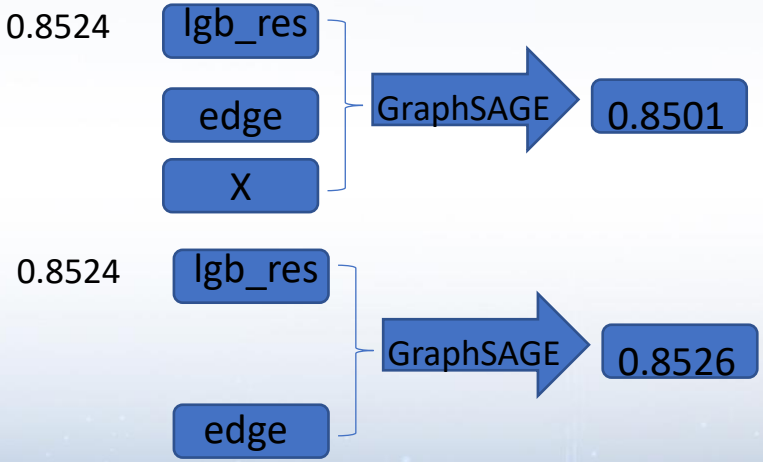
模型融合

方法1：GraphSAGE最优模型和lgb最优模型融加权融合单折结果：



Hard for me online

方法2：lgb评分当特征加入X里，再GraphSAGE单折结果：



优点：

- 最终线上方案使用了LGB单模型，模型应用起来方便简单，且特征构造的可扩展性极强
- 对整体数据有较深的理解，构造了全面的特征体系框架。
- 依托于高质量的图数据，构造出线上线下同增减评估方法

缺点：

- 最对深度网络设计缺乏创造力，导致最终结果还未到最优层面

信也科技杯 FINVOLUTION Data Science Competition | 信也 FINVOLUTION

TNANK YOU