实 验 报 告

实验名称 交易反欺诈的风险识别

课程名称 网络数据风控技术

专业班级: 2019 级数据科学与大数据技术专业

学生姓名: 谭芷妍 学 号: 1851150

同组学生: 学 号:

指导教师: 成 绩:

实验日期: 2021 年 12 月 9 日

同济大学

一、 实验目的

近年来，银行线上业务随着移动互联网的兴起而快速发展，互联网金融业务涉及多种欺诈风险，国家也密切关注账户资金流动的合规性风险，监管趋严，监管文件频出。如银发【2016】261 号文、银发【2019】85 号文、银保监办发【2021】49 号文等。同时，2020 年以来涉赌涉诈案件频发，公安部加大打击力度，要求银行对其账户做内部排查，及时关停可疑账户。为保障银行线上金融业务的平稳运行、风险可控、账户安全规范，提出了基于线上交易反欺诈的风险识别的需求。

本实验提供风险账户的基础信息、操作行为信息和交易行为信息，通过数据分析，从账户开户特征、设备操作特征、交易频次、交易时间、交易金额、地域分布等维度进行特征挖掘，建立有监督的机器学习模型，有效甄别存在高风险交易的账户。

二、 问题分析与数据质量分析

该问题的训练集和测试集分别提供了三张数据表：_base、_op 和 _trans。表中的特征可分成三类：有大小关系的类别编码、无大小关系的类别编与连续型编码。该问题是一个以是否发生欺诈为因变量的二分类问题，为了求解该问题首先需要根据这些数据进行特征衍生，对特征进行处理后，分别使用 catboost、xgboost、lightgbm 等模型进行训练。

首先分别对 train_base、train_op 和 train_trans 作缺失值分析，得到如下结果：

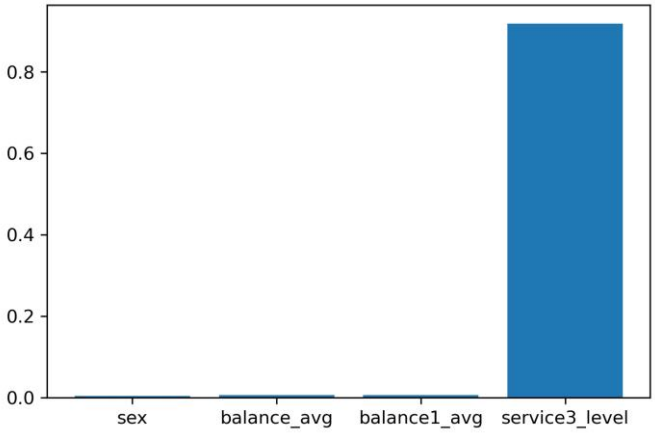

图 1 train_base 表的缺失值统计

在 train_base 中有 4 个属性有缺失值，其中’service3_level’缺失率超过 90%，将该特征剔除。’balance_avg’和’balance1_avg’都是有序类别特征，且缺失率在 0.5%左右，对这些样本进行补 0 的操作。’sex’特征缺失的样本非常少，如使用 lightgbm 模型可以不处理，但在使用其他模型时需要处理，考虑通过分箱实现将缺失值化为同一分箱。

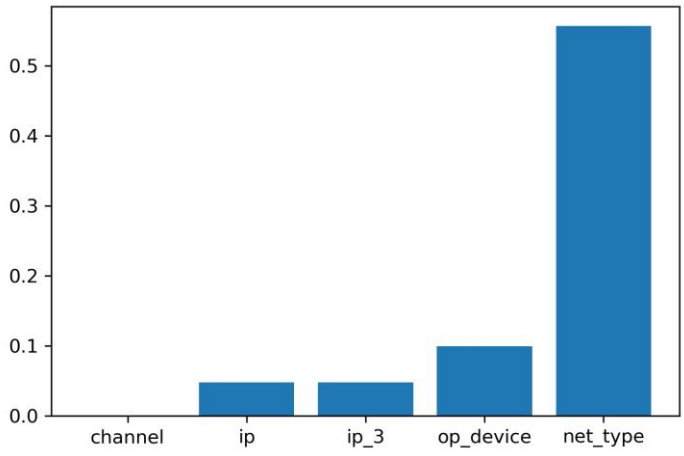

图 2 train_op 表的缺失值统计

在 train_op.csv 中’net_type’属性缺失过多，后续不考虑基于该属性进行特征衍生计算。对于 ip、ip_3 和 channel 由于缺失非常少，可以利用这些属性进行特征衍生。

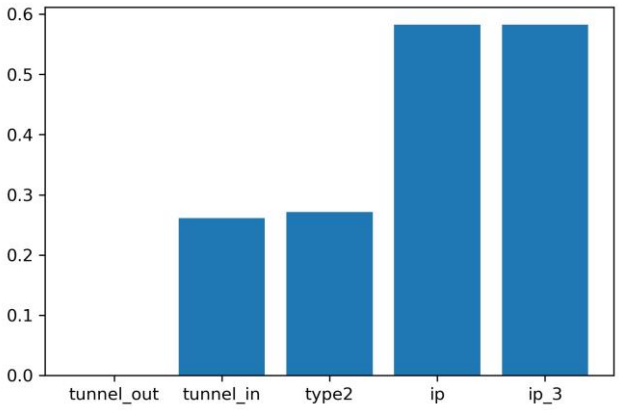

图 3 train_trans 表缺失值统计

原始训练数据正负样本比例统计

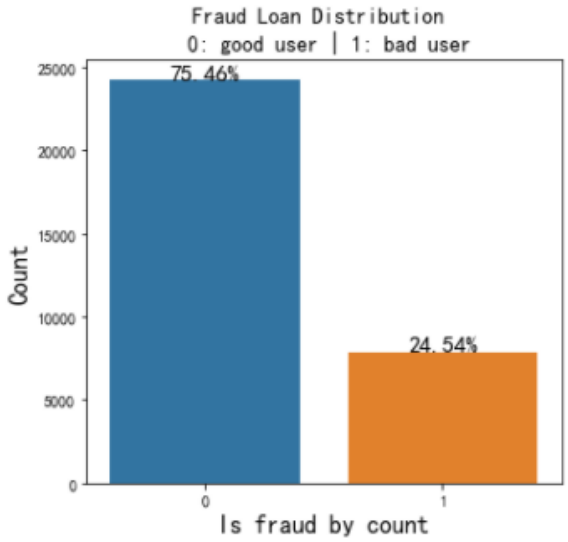

图 4 正负样本分布统计（1 为欺诈样本 0 为非欺诈样本）

整体来看，正负样本的比例是较为均衡的，因此不考虑对样本进行采样的操作。此外也进行了，重复值检验，不存在单值特征。

三、 特征工程

特征衍生主要基于 opt 表和 trans 表进行的。这两张表都有时间特征（距离某观测时间点的时间间隔），因此考虑对时间进行切片。对 opt 表和 trans 表中记录按时间画出分布直方图可得到如下两张图：

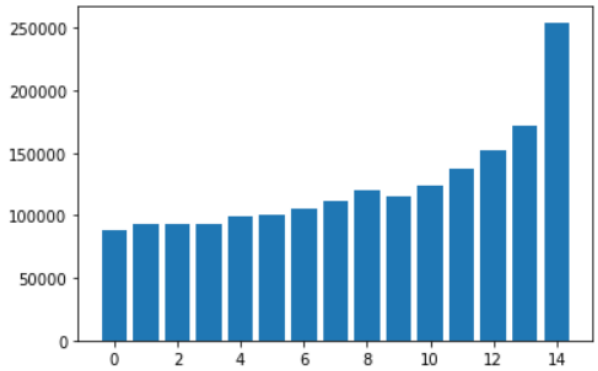

图 5 opt 表时间跨度

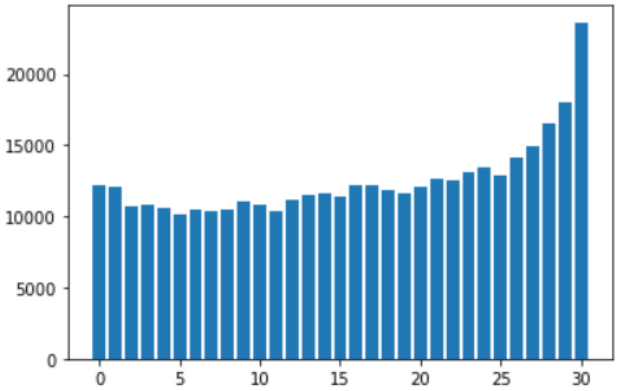

图 6 trans 表时间跨度

对于 opt 表，时间切片为 1/7/14 天，对于 trans 表，时间切片为 1/7/15/30 天。

① 基于 opt 表的特征衍生包括：

在某个观测时间段 i 内，用户的总操作次数、平均操作次数、用户使用的 ip 类型个数、用户使用的操作渠道个数。

② 基于 trans 表的特征衍生包括：

在某个观测时间段 i 内，用户所完成的交易金额总额、平均值、最大值、最小值和标准差、用户进行交易的总次数和均值；过去 30 天内用户进行交易的平台类型统计、过去 1 天交易超过 1 万元次数。[1]

完成对 opt 表和 trans 表的特征衍生后加入 base 表可以得到 85 个特征的总表。进一步观察总表的单个变量的分布。

（一） 特征分布分析与特征初筛

首先是去除缺失率过高的特征：service3_level（缺失率为 91.56%）；分别对离散特征和连续特征画出频率分布直方图与密度分布曲线。分析离散型特征找出严重偏向某一取值的特征：service1_cnt、service2_cnt、agreement1，后续训练时进行剔除，筛除后的全部特征如下：

表 1 用于训练的所有特征

level：用户等级	verified：是否实名	using_time：使用时长	regist_type：注册类型
op1_cnt：某类型 1 操作数量	op2_cnt：某类型 2 操作数量	card_d_cnt：d 卡数量	agreement_total：开通协议数量
province：省份	city：城市	balance：余额等级	balance_avg：近某段时期余额均值等级
service3：是否服务 3 用户	login_cnt_period1：某段时期 1 的登录次数	product1_amount：产品 1 金额等级	product2_amount：产品 2 金额等级
product7_cnt：产品 7 申请次数	product7_fail_cnt：产品 7 申请失败次数	p_oneDayOpt：过去 1 天操作次数	p_7DayOpt：过去 7 天操作次数
p_oneDayOptChanType：过去 1 天使用的操作渠道个数	p_7DayOptChanType：过去 7 天使用的操作渠道个数	p_15DayOptChanType：过去 15 天使用的操作渠道个数	AllMon_1：过去 1 天总交易金额
过去 1 天累计交易金额方差	过去 1 天最大交易金额	过去 1 天最小交易金额	过去 30 天平均累计交易金额
过去 15 天累计交易金额方差	过去 15 天最大交易金额	过去 15 天最小交易金额	过去 7 天平均累计交易金额
过去 1 天累计交易次数	过去 7 天累计交易次数	过去 15 天累计交易次数	过去 7 天平均交易次数
user	sex：性别	age：年龄	provider：运营商类型
card_a_cnt：a 卡数量	card_b_cnt：b 卡数量	card_c_cnt：c 卡数量	acc_count：账户数量
agreement2：是否开通协议 2	service1_amt：某业务 1 产生金额	agreement3：是否开通协议 3	agreement4：是否开通协议 4
login_cnt_period2：某段时期 2 的登录次数	ip_cnt：某段时期登录 ip 个数	login_cnt_avg：某段时期登录次数均值	login_days_cnt：某段时期登录天数
balance1：类型 1 余额等级	balance1_avg：近某段时期类型 1 余额均值等级	balance2：类型 2 余额等级	balance2_avg：近某段时期类型 2 余额均值等级
product3_amount：产品 3 金额等级	product4_amount：产品 4 金额等级	product5_amount：产品 5 金额等级	product6_amount：产品 6 金额等级
p_15DayOpt：过去 15 天操作次数	p_oneDayOptIpType：过去 1 天使用的 ip 地址个数	p_7DayOptIpType：过去 7 天使用的 ip 地址个数	p_15DayOptIpType：过去 15 天使用的 ip 地址个数
AllMon_7：过去 7 天总交易金额	AllMon_15：过去 15 天总交易金额	AllMon_30：过去 30 天总交易金额	过去 1 天平均累计交易金额
过去 30 天累计交易金额方差	过去 30 天最大交易金额	过去 30 天最小交易金额	过去 15 天平均累计交易金额
过去 30 天累计交易次数	过去 7 天累计交易金额方差	过去 7 天最大交易金额	过去 7 天最小交易金额
过去 15 天平均交易次数	过去 30 天平均交易次数	过去 30 天交易平台类型	过去 1 天交易超过 1 万元次数

进一步分析离散特征的分布发现大量的特征存在某一个取值的欺诈样本数远高于其他取值的情况，如直接对这种类型的特征直接进行 LabelEncode 编码转化并使用 SVM 等线性模型进行训练可能会对分类结果有比较偏向性的影响，因此考虑使用树模型进行后续训练。

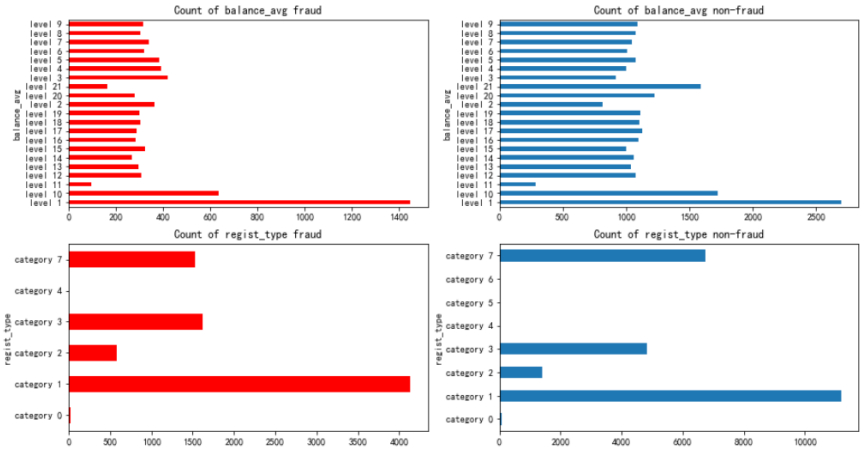

图 7 两个离散特征的分布直方图（左图为欺诈样本右图为非欺诈样本）

同时，部分离散特征存在多值偏向性。如 province、city 和 age 这三个特征，这三个特征有较多的取值，并且某一些取值的欺诈样本数极高。除了考虑使用树模型，也考虑对这些特征进行分箱处理。

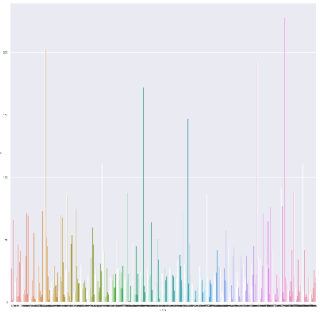

图 8 欺诈样本的 city 分布

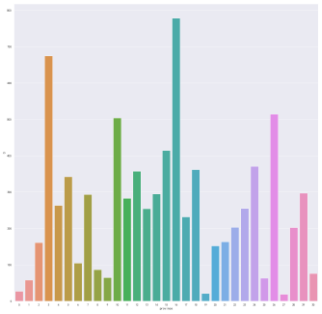

图 9 欺诈样本的 province 分布

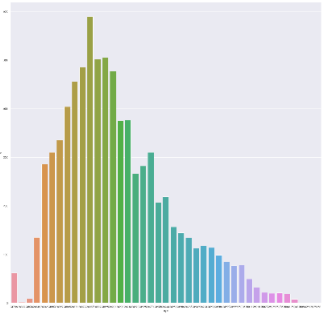

图 10 欺诈样本的 age 分布

对连续型特征画出频率分布直方图与密度分布曲线分析发现连续型特征也存在类似的问题。超过 20 个特征倾斜严重，在进行了对数运算后特征的倾斜程度改善不明显，考虑对这些特征进行分箱以减轻倾斜情况。

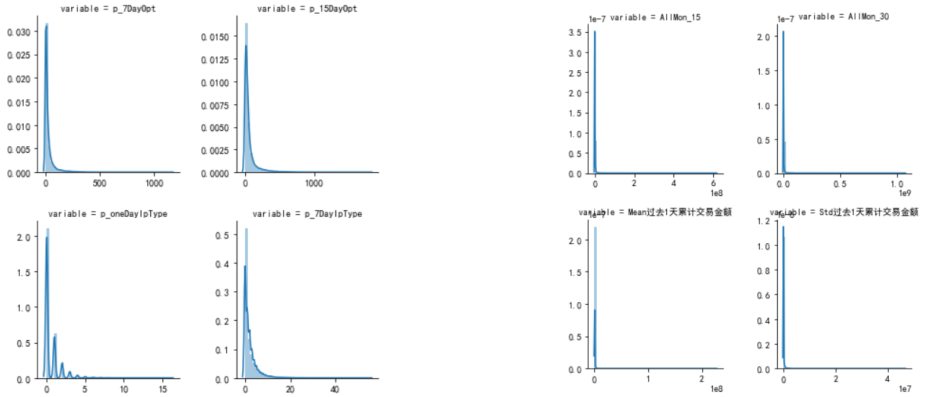

图 11 部分连续型特征的分布图

（二） 缺失值处理与离散型特征编码

1、 缺失值处理

连续型特征与离散型特征均有特征存在缺失的情况。衍生得到的特征中存在较多的缺失情况，这主要是因为在 opt 表中和 trans 表中并不包含全体样本用户的操作记录。对于连续型特征一般有填充 0、填充中位数和填充众数三种方法，在实验中，这三类方法对训练效果影响不大，最终选择了填充中位数的方法。对于离散型特征，没有大小关系的特征进行了类别编码；分别尝试了分箱的方法和填充众数的方法，相比而言分箱的方法效果更好可解释性也更强，因此选择了分箱的方法。

2、 离散型特征编码

（三） 相关性分析与异常值处理

1、 异常值分析

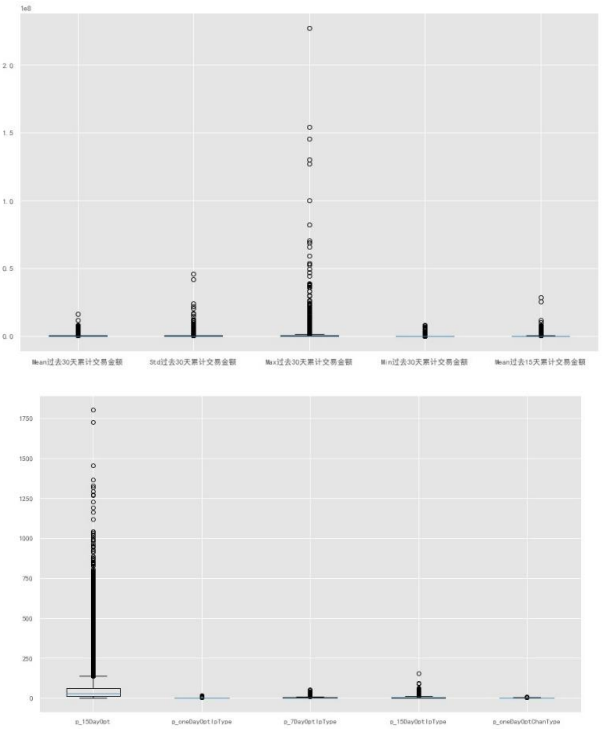

图 12 部分箱型图

基于前文对连续型特征的分布分析结合，箱型图中大量特征出现较

多的异常值与前面对变量分布的分析是相吻合的，考虑到如果直接剔除异常值可能会删去大量与分类有关的信息，因此对异常值不进行处理。

2、 相关性分析

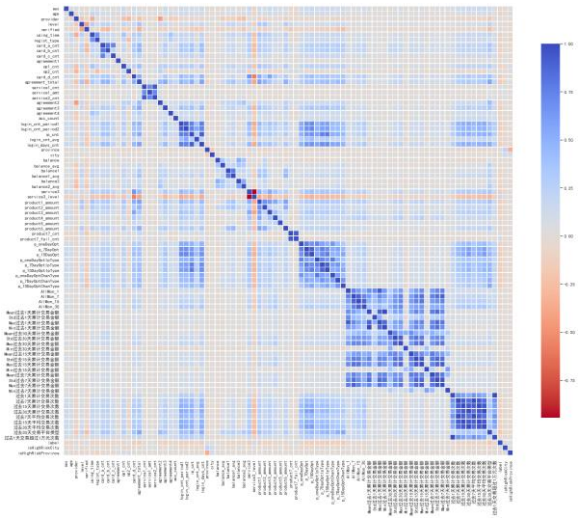

图 13 原始特征的相关性分析

从原始特征来看，除了部分衍生特征以外，其余特征之间的基本不相关，同时各特征与因变量 label 之间基本不相关。

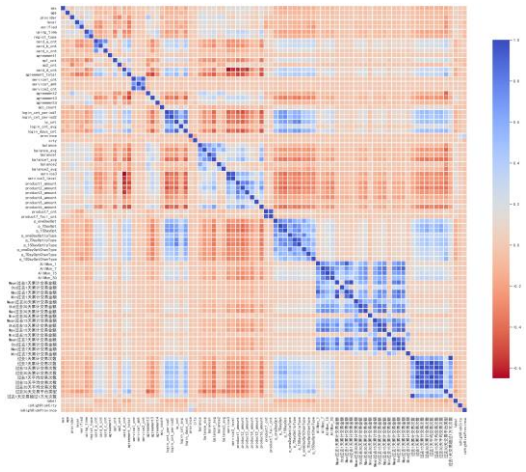

图 14 决策树分箱后特征相关性分析

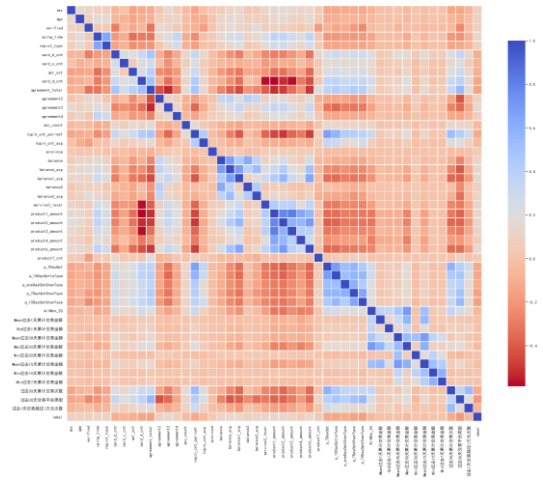

图 15 单特征 iv 值筛选后特征相关性分析

基于前文的分析，连续型特征倾斜严重，需要进行分箱，同时对于有缺失值的离散特征 sex 也通过分箱将缺失值单独分成一类，其余离散型特征进行 LabelEncoder 类别编码。

在多次试验后，选择了决策树分箱作为分箱方式，分箱后对分箱结果进行微调并进行微调后得到图 15 的特征相关性热力图，各特征与 label 变量相关性依然比较低。进行 WOE 编码转换后进行单特征筛选。单特征筛选的标准是 iv 值小于 0.02 和相关系数大于 0.8 的特征筛去，最终剩余 46 个特征。

四、 模型训练与优化

在前文的分析中，各类特征均不同程度的倾斜。因此在后续训练时首先考虑使用对离散特征各取值出现频率进行统计并生成新的数值型特征的 CatBoost 模型和使用直方图算法对特征进行划分的 LightGBM 模型进行训练。

(一) CatBoost

首先选用 CatBoost 官网给出的案例初步设置参数进行训练。

```
params = {'depth': 8, 'eval_metric':'AUC','l2_leaf_reg': 10,
'random_strength':12,'one_hot_max_size':32,
'od_type': 'Iter', 'od_wait': 50, 'random_seed': 11, 'allow_writing_files': False}
```

图 16 CatBoost 参数

结合网格法对部分超参数进行调整

```
grid = {'learning_rate': [0.03,0.05, 0.1],
        'depth': [4, 6, 10],
        'l2_leaf_reg': [1, 3, 5, 7, 9]}
```

图 17

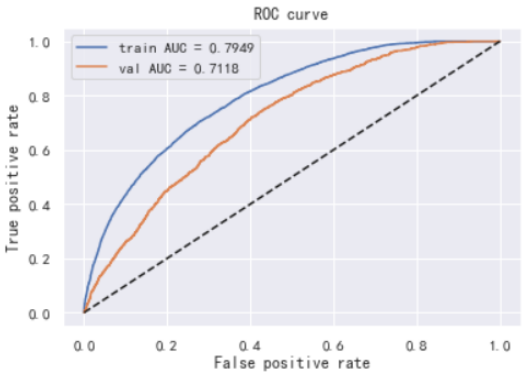

图 18 CatBoost 训练结果 ROC 曲线

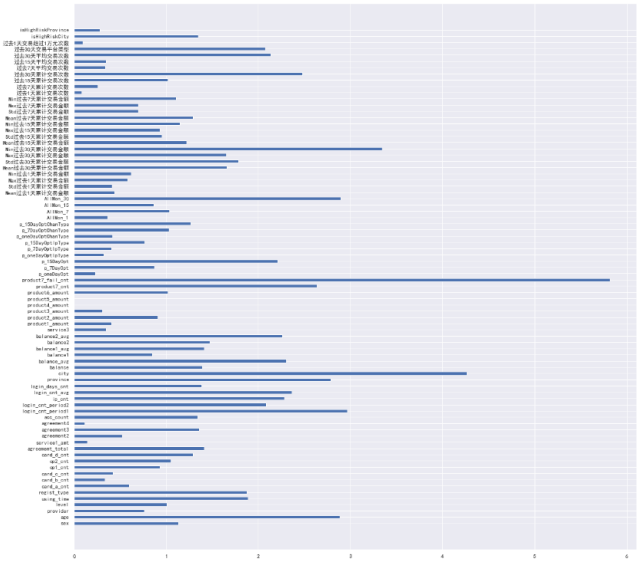

图 19 CatBoost 模型特征重要性分析

打印特征重要性，前 5 位的特征是 product7_fail_cnt、city、login_cnt_period1、过去 30 天累计交易金额和 province。

（二） LightGBM

首先参考了大作业平台中 lgb 模型的参数设置进行了第一次训练。此次训练分别使用了原始数据与进行了 WOE 编码转换并进行了单特征筛选后的数据进行训练。

```
'boosting_type': 'gbdt',
'objective': 'binary',
'metric': 'auc',
'learning_rate': 0.01,
'num_leaves': 95,
'max_depth': 3,
'min_data_in_leaf': 43,
'min_child_weight':9.5,
'bagging_fraction': 0.98,
'feature_fraction': 0.96,
'bagging_freq': 42,
'reg_lambda': 8,
'reg_alpha': 4,
'min_split_gain': 0.25,
'nthread': 8,
'seed': 2020,
'silent': True,
'verbose': -1,
```

图 20 lgb 初设参数

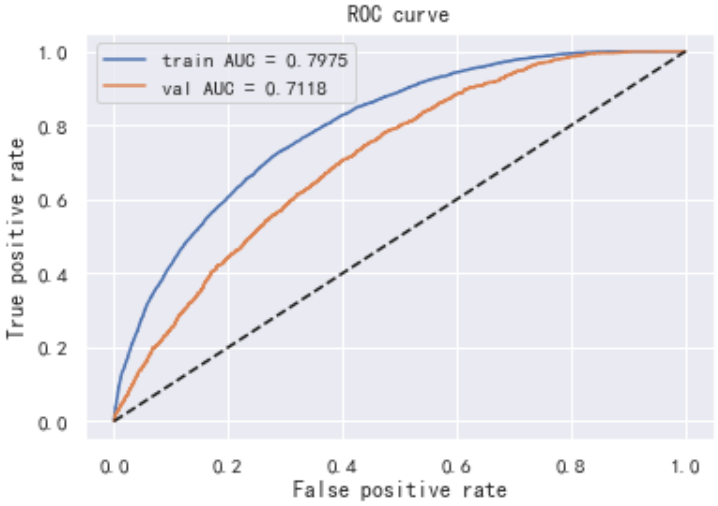

图 21 使用原始数据进行 lgb 模型训练

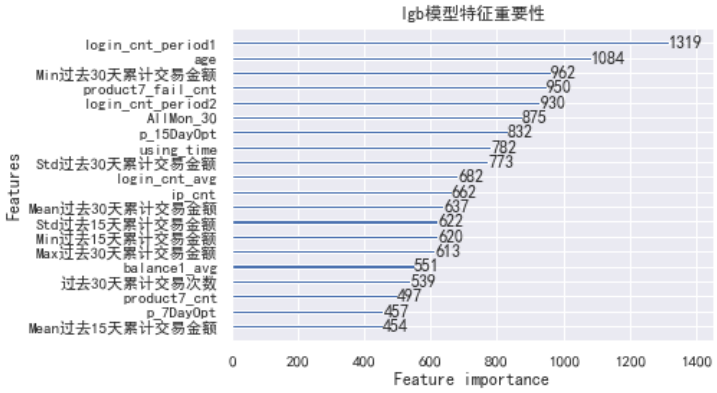

图 22 图 21 对应特征重要性分析

在使用原始数据进行训练时其结果与 CatBoost 结果相近。

使用 WOE 编码并进行单特征筛选后的数据进行训练的结果如下：

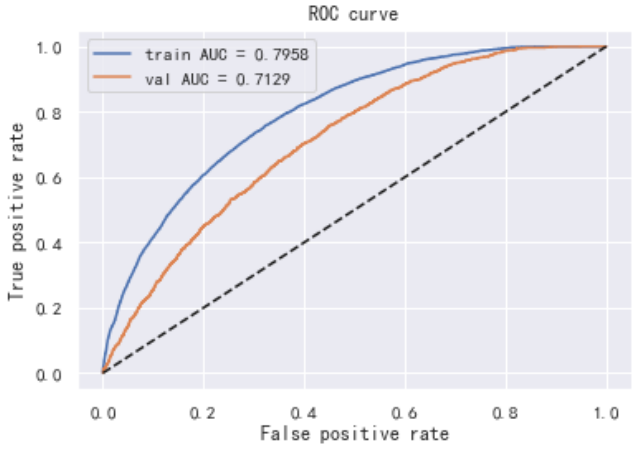

图 23 采用进行特征工程后数据进行 lgb 训练

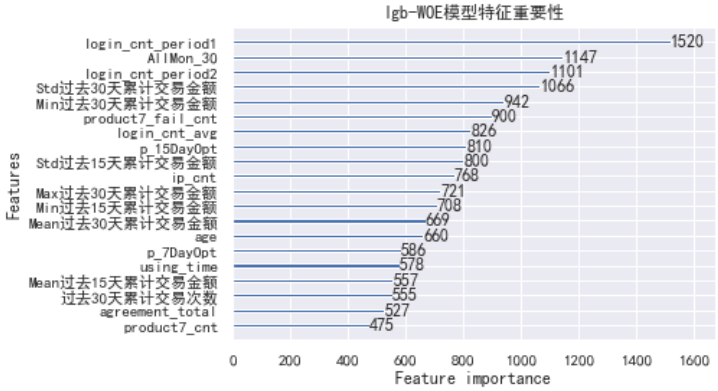

图 24 图 23 对应特征重要性分析

其结果比原始数据的结果要稍好一些，但也存在一定程度的过拟合，考虑通过调参进行进一步尝试。由于网格搜索对于 lgb 的超参数调整来说时间太长，这里采用了贝叶斯调参的方法进行尝试。

```
from bayes_opt import BayesianOptimization

bayes_lgb = BayesianOptimization(
    rf_cv_lgb,
    {
        'num_leaves':(10, 200),
        'max_depth':(3, 20),
        'bagging_fraction':(0.5, 1.0),
        'feature_fraction':(0.5, 1.0),
        'bagging_freq':(0, 100),
        'min_data_in_leaf':(10,100),
        'min_child_weight':(0, 10),
        'min_split_gain':(0.0, 1.0),
        'reg_alpha':(0.0, 10),
        'reg_lambda':(0.0, 10),
    }
)
```

图 25 贝叶斯调参参数范围设置

```
'boosting_type': 'gbdt',
'objective': 'binary',
'metric': 'auc',
'learning_rate': 0.01,
'num_leaves': 132,
'max_depth': 5,
'min_data_in_leaf': 16,
'min_child_weight':9.4,
'bagging_fraction': 0.73,
'feature_fraction': 0.52,
'bagging_freq': 25,
'reg_lambda': 9,
'reg_alpha': 8,
'min_split_gain': 0.25,
'nthread': 8,
'seed': 2020,
'silent': True,
'verbose': -1,
```

图 26 贝叶斯调参结果

参数调整后模型训练结果有所提升：

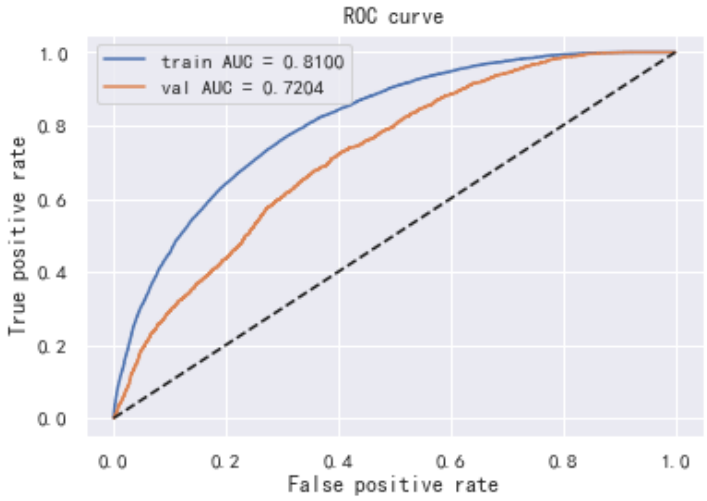

图 27 调参后 lgb 模型训练结果 ROC 曲线

（三） XGBoost

考虑到 lgb 模型和 CatBoost 模型内部对离散特征均有所处理，因此进一步使用了 xgboost 模型进行了训练。

```
params = {'booster': 'gbtree',
        'objective': 'binary:logistic',
        'eval_metric': 'auc',
        'gamma': 1,
        'min_child_weight': 1.5,
        'max_depth': 5,
        'subsample': 0.7,
        'colsample_bytree': 0.7,
        'colsample_bylevel': 0.7,
        'eta': 0.04,
        'tree_method': 'exact',
        'seed': 2020,
        'nthread': 36,
        "silent": True,
        'num_boost_round':50000,
        'verbose_eval':3000, 'early_stopping_rounds':200
        }
```

图 28

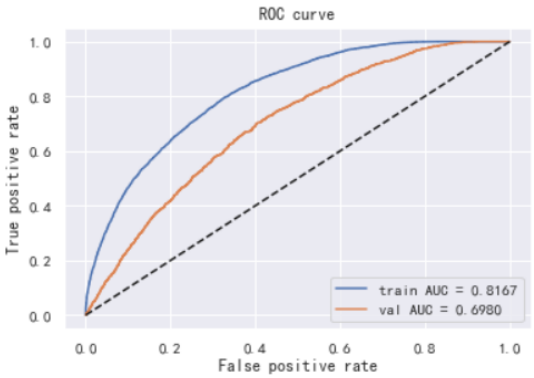

图 29

从结果上看，相同情况下 xgboost 模型确实比其余两个稍差一些。

对比来看，在结果上 lgb 模型要比 catboost 模型的结果稍好一些；在训练时间上两者相当，但 lgb 涉及的超参数较多，调参比较麻烦，最后打印两者的树结点的个数分别是 lgb 的 791 与 catboost 的 585，catboost 分裂的树节点更少。

五、 模型效果评估

分别保留以上三个单模型的最优参数进行 5 折交叉验证，取 auc 的平均得分作为最终得分。在分类阈值的选择上，由于在这个任务中，希望欺诈样本尽可能地被选出来，即更多地考虑查全率，因此考虑适当降

低分类阈值。选择分类阈值 0.3 计算查准率与查全率得到如下的结果：

表 2

模型	AUC	precision	recall	F1-score
CatBoost	0.712	0.378	0.531	0.442
LightGBM	0.719	0.402	0.558	0.467
XGBoost	0.70	0.400	0.527	0.455

混淆矩阵：

CatBoost

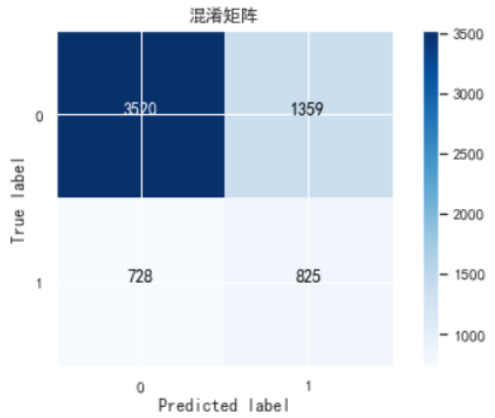

图 30

LightGBM

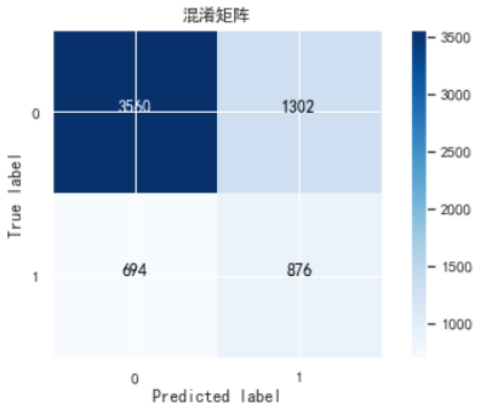

图 31

XGBoost

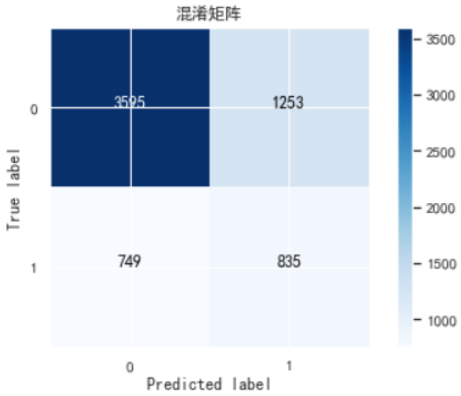

图 32

横向对比来看，lightgbm 模型的训练效果是最好的，catBoost 次之，xgboost 效果最不好，但三者间的差距并不大；在同一分类阈值划分下，lightgbm 模型的 F1-score 也是最高的，综合来看在这个任务背景下 lightgbm 模型是更合适的。

六、 结论和心得体会

1、在本次实验的过程中出现的主要问题是模型的训练效果无法通过特征工程实现有效的提升。比如在多特征筛选上尝试了不同的方法，但对同一模型、同一数据集来说训练结果并没有明显的改变。

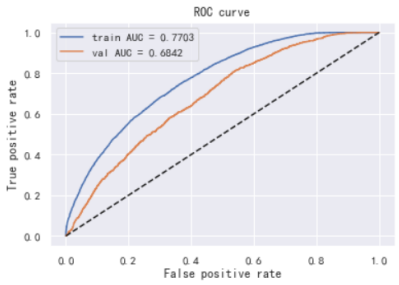

图 33 相关系数筛选

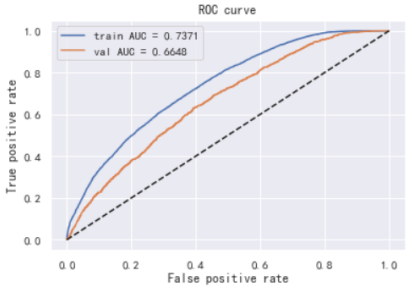

图 34 带惩罚项的特征筛选

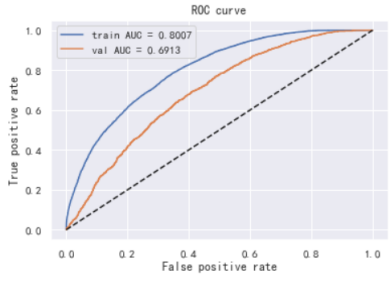

图 35 基于 GBDT 的特征筛选

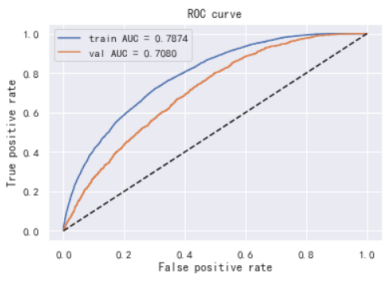

图 36 基于特征重要性进行筛选

前期有比较多的时间都卡在了这一步上所以耽误了时间，通过与同学的交流发现，可能主要问题是出现在特征衍生的不够充分这一步上。分析后面得到的特征重要性图与相关性图，衍生出的特征存在部分特征相关性比较高，在特征重要性图中 base 表内原有的特征也占据了一半。这说明在特征衍生这一步并没有把能体现更多分类信息的特征挖掘出来，或者可以考虑特征组合对特征进一步细化，这一点需要进行改进。

2、在实验时尝试了模型融合的方法（VotingClassifier 和 blending）融合的第一层使用完成调参的 catboost 和 lgb 模型，第二层采用 logistic 回归模型但最终实现效果都非常差。

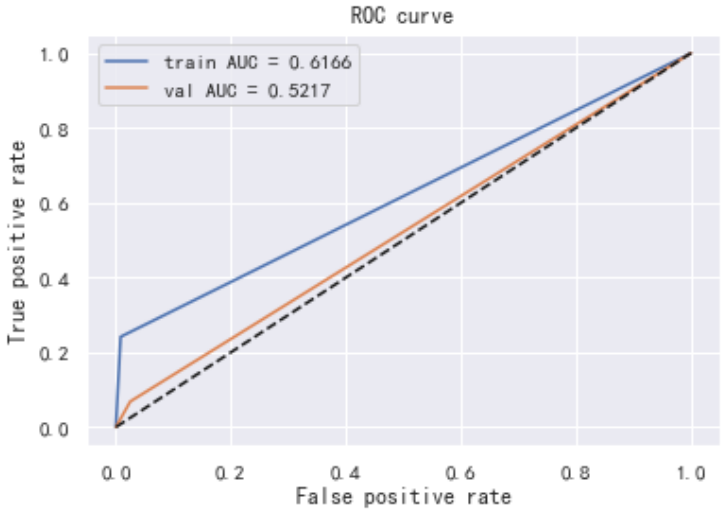

图 37

考虑到如果个体学习器的准确性越高、多样性越大，融合的效果会更好，这里使用的 lgb 模型和 catboost 模型性能差距并不大但可能是由于单个模型本身的训练效果不是太高而且对数据多样性没有做足够的处理，所以初步实现的效果并不好。

3、数据采样：尽管我原本认为本次提供的数据正负样本比例是合适的但由于想尝试不同的方法我还是尝试了数据采样。我尝试的方法是 SMOTE 过采样，但出现了非常严重的过拟合现象。

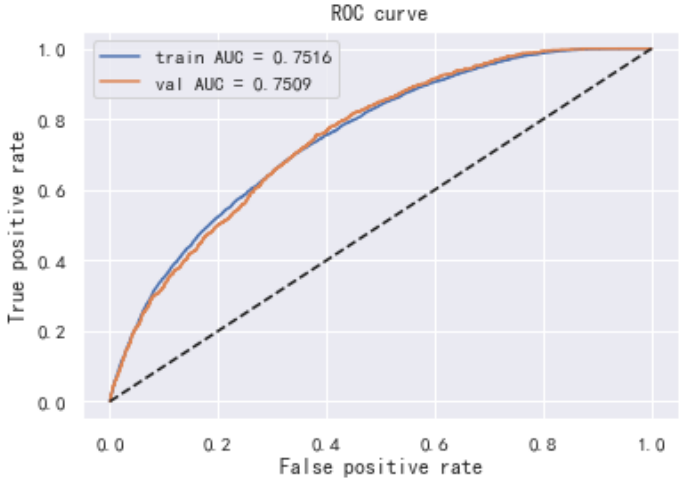

图 38

可能是因为 SMOTE 方法进行插值每次构造的是与样本邻近的数据，选取新样本的标准是为了提高模型的训练效果，因此很容易产生过拟合。下一步的改进可以尝试进行随机欠采样实验，考察是否会得到一定的数据增强的效果。

七、 参考文献：

[1]曹汉平,张晓晶,祝睿杰,等. 数字金融时代机器学习模型在实时反欺诈中的应用与实践[J]. 智能科学与技术学报,2019,1(4):342-351. DOI:10.11959/j.issn.2096-6652.201939.

[2]刘思茹. 基于特征工程的信用卡欺诈识别研究[D].兰州大学,2021.DOI:10.27204/d.cnki.glzhu.2021.001715.

[3]CatBoost:https://catboost.ai/en/docs/concepts/python-reference_catboostclassifier_grid_search

[4]LightGBM:https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html