马上 AI 全球挑战者大赛——违约用户风险预测

一、方案概述

（一）赛题背景

在金融领域，无论是投资理财还是借贷放款，风险控制永远是业务的核心基础。对于消费金融来说，其主要服务对象的特点是：额度小、人群大、周期短，这个特性导致其被公认为是风险最高的细分领域。随着人工智能和大数据等技术不断渗透，依靠金融科技主动收集、分析、整理各类金融数据，为细分人群提供更为精准的风控服务，成为解决消费金融风控问题的有效途径。

这是个典型二分类问题，或者说就是预测用户是否违约，在金融风险领域二分类挑战是正负样本极不平衡。本次的比赛的目标是要求参赛者预测样本违约的百分比概率是多少。所以，本次比赛的目的就是让参赛者能设计出色的分类模型，能够对样本进行精准预测以实现风险最小化。

（二）处理流程

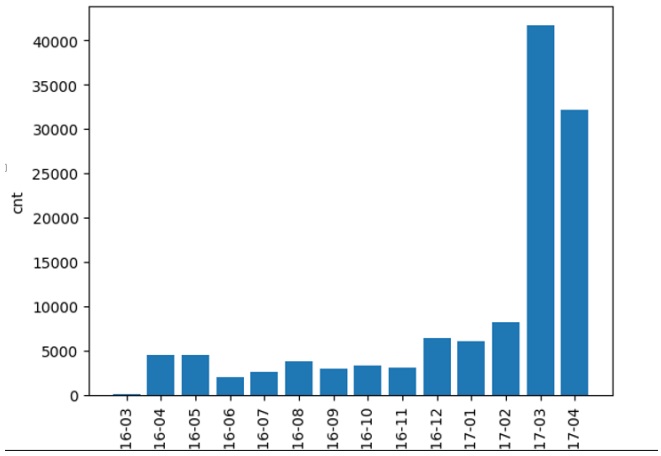
我们的实现方案是前期各自独立完成特征工程和构建模型，最终进行一个汇总交流以及模型融合。我们一共构建了 3 个模型，包括两个 XGBoost 模型，一个 LogisticRegression 回归模型。最终采用了线性加权融合得出了最优的结果。

（三）创新点

本小组认为我们本次的创新点在于模型异构方面做的比较好，最后通过特征比对分析，三个模型所用的特征差异性比较大，采用了两种不同类型的模型，正是由于特征和模型的较大差异性，所以最终的方案具有很好的泛化能力，这通过 AB 轮的结果可以分析得出（A 榜排名第 14，B 榜排名第 4)。

二、数据洞察

拿到数据，我们做了一些数据可视化：



图一　训练集时间分布

从图一可以看出训练集的时间维度是：2016 年 3 月 - 2017 年 4 月，计算得出样本个数 120930；训练数据主要集中在 2017 年的 3、4 月。

其中违约用户 3388 条，正负样本比例为 1：35。

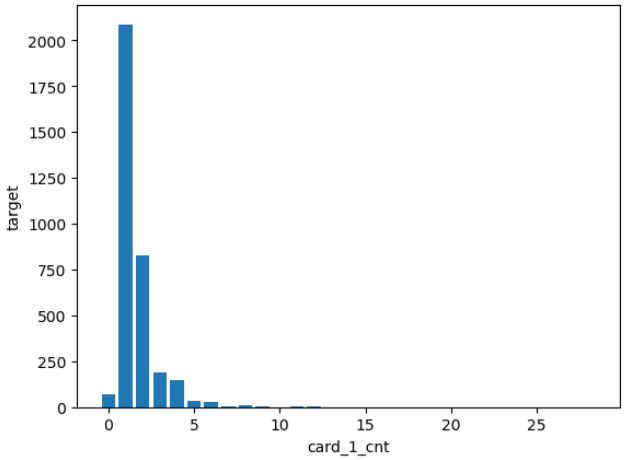
A 轮测试集的时间维度是：2017 年 5 月，样本个数为 47032；B 轮测试的时间维度是：2017 年 6 月，样本个数为 52017。本赛题的数据样本都有一个唯一的 id，所以不能使用常规的数据集划分方法来扩充数据集。

本赛题数据的最鲜明的特点是空缺值太多，存在较多中文属性值，并且某些属性值含义不清晰就像用户的爱好以及某些属性值表示不规范，如银行卡名称中英文混搭。对于空缺值过多问题，在本赛题的应用场景中个人相关信息对于借贷业务来说应该是一个比较重要的因素，往往个人信息越完善违约风险可能会越小，所以我们在后期做特征工程的时候对部分属性是否为空提取了特征。

order_info 表中的交易额有很多空值，我们把空值用均值填充然后提取特征，对于标称类型数据的空值，进行 one-hot 编码。

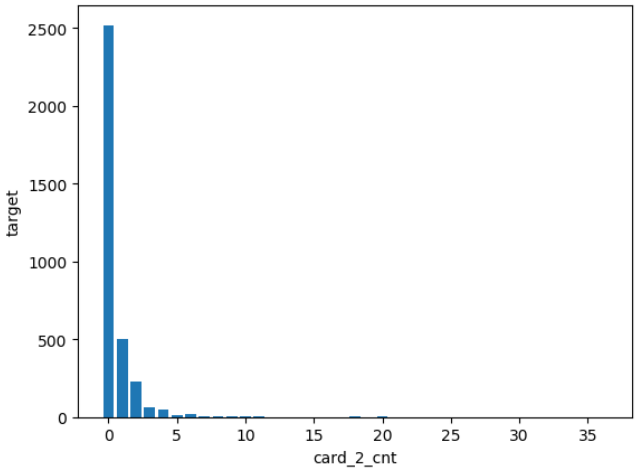
图二为储蓄卡的数量对应违约用户的数量的直方图，发现用户有储蓄卡张数为 1 或 2

张的时候违约可能性较大。



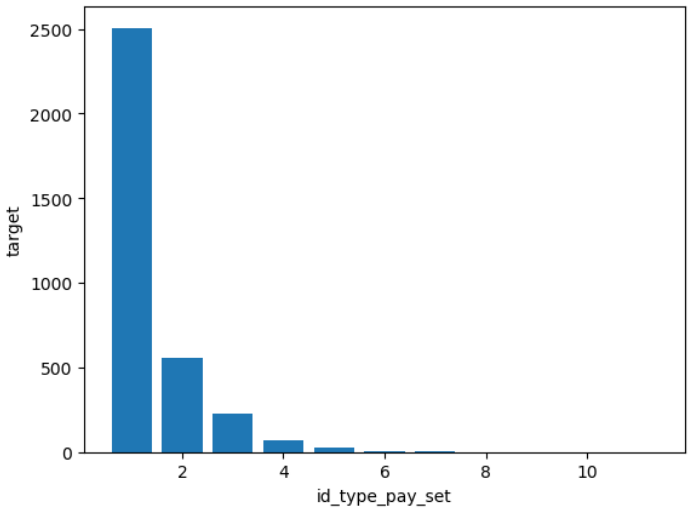
图二　储蓄卡的数量对应违约用户的数量

图三为信用卡的数量对应违约用户的数量的直方图，发现用户有信用卡张数为 0 张的时候违约可能性较大。



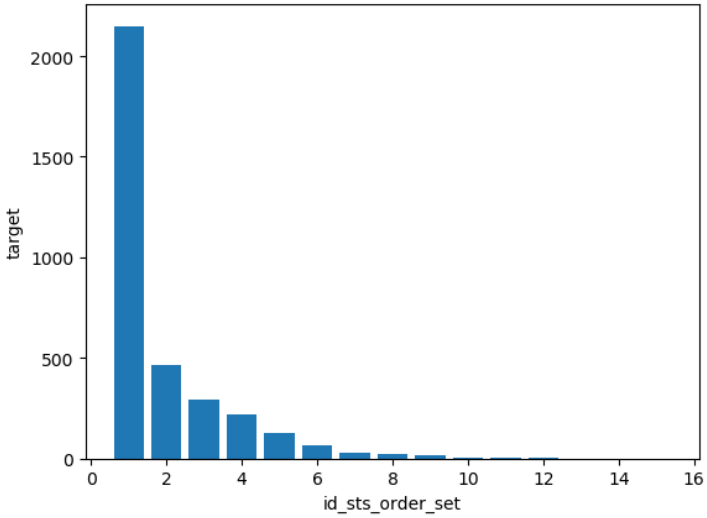
图三　信用卡的数量对应违约用户的数量

图四为支付方式的种类数和违约用户数量的直方图，发现支付方式比较单一违约用户较多。



图四　支付方式的种类数对用应的违约用户数量

从图五可以看出订单状态的种类数量为 1 的违约用户较多。



图五　用户订单状态的种数对应违约用户的数量

三、特征工程

特征提取以特征群的形式进行的，主办方一共给了 7 个数据表，其中训练集中的 train_target 和测试集中的 test_list 表单表没有提取相关特征，但在联表进行特征提取的时候用到过。

train_auth_info 表：

身份证账号是否为空
认证时间是否为空
电话号码是否为空
是否所有信息都为空，除了 id
是否所有信息都不为空
认证时间和身份证是否同时为空
认证时间和手机号码是否同时为空

train_bankcard_info 表：

用户有多少条记录
用户有多少个不同的手机号码
用户的储蓄卡的数量
用户的信用卡的数量
用户是否有信用卡
用户有几种不同类型的银行卡
用户银行卡数量是否大于 6
用户是否只有一张银行卡

train_credit_info 表：

用户信用积分
用户额度是否为 0
用户已使用的额度
用户的信用额度

用户剩余的额度
用户额度使用率
用户额度排名
是否所有信息都为空，除了 id
是否所有信息都为 0，除了 id
信用额度是否为 0
信用积分是否为 0
用户是否还有剩余的额度

train_order_info 表:

用户记录是否有除 id 外都为空
用户关于商品单价的统计特征
用户关于订单金额的统计特征
支付方式的离散特征
订单状态的离散特征

train_recieve_addr_info 表:

用户记录中是否有除 id 外都为空
'addr_id', 'region', 'phone', 'fix_phone', 'receiver_md5'是否同时为空
用户的记录数
用户收获地址中的省份离散特征
用户收获地址中有多少不同的省份

train_user_info 表：

用户生日是否是“0000-00-00”
用户性别的 one-hot 编码
用户婚姻状况的 one-hot 编码
用户会员等级的 one-hot 编码
用户是否绑定 QQ
用户是否绑定微信号
用户学历是否是“硕士、其它、博士”
用户身份证号是否为空
用户会员收入的 one-hot 编码

联表提取特征：

用户年龄
用户注册天数
用户借贷日期是否早于注册日期
下订单时间与注册时间的天数差的最大、最小、平均

以上特征经过特征选择再构建模型的，特征选择方案如下一章所述。

四、特征预筛选

方案一：

采取的特征选择方案是根据线上线下的特征表现结果决定是否保留该特征。

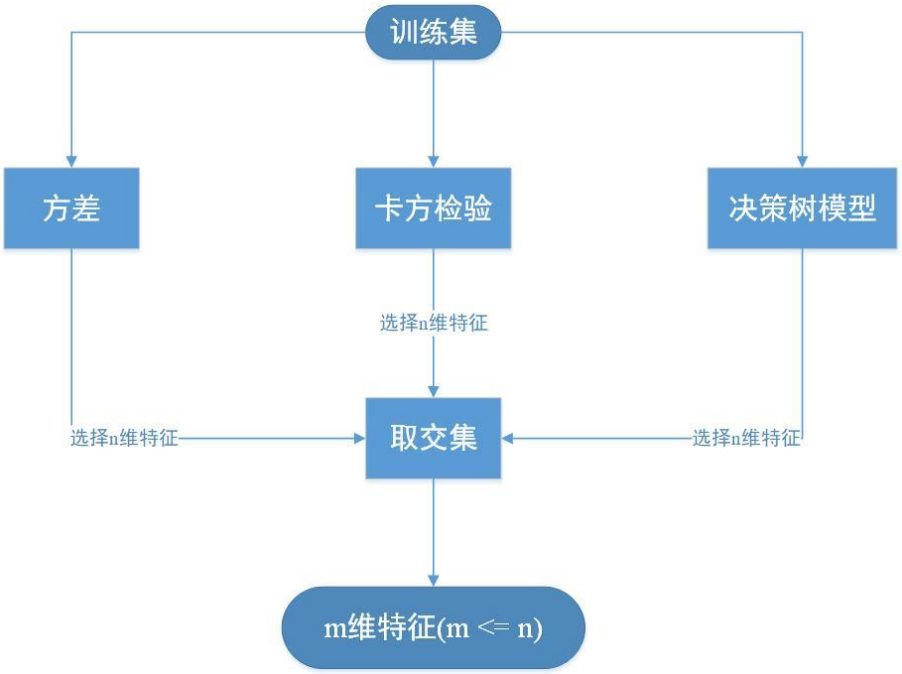
方案二：

采用的特征选择方案是融合三种基本的特征选择方法，三种常用的特征选择方法分别是：

① 基于特征方差的特征选择:将每个特征的方差计算出来,最终选取方差较高的 n 维特征,其中参数 n 是由我们给定的。

② 基于卡方检验的特征选择：使用卡方检验的方法选择出 n 维特征，参数 n 由我们指定。

③ 基于决策树模型的特征选择：使用决策树用训练集训练得出一个模型，然后输出每个特征的重要性，最后选出得分较高的前 n 维特征，参数 n 由我们指定。



图六　特征选择流程图

特征选择的流程如图六所示，特征评估方法是采用 2017 年 4 月作为验证集，4 月前面所有样本作为训练集，最开始我们采用验证集是 2016 年 3 月到 8 月，一共 2 万多个样本，但是发现这种方法验证效果误差太大，线上线下差距过大，所以最后选择的是 2017 年 4 月用来验证。

方案二一共提取了 194 维特征，使用了一个循环 n 的初始值维 50，每执行一次 n 加 5，循环终止条件为 n 小于等于 194。本方案共得到了两个模型一个 XGBoost 一个 LogisticRegression，所以评价方法里面使用的模型就与之相对应。最终得出 m 维特征是根据

每次循环迭代得出的验证结果最好的一个特征组合。

最终选出的用于 XGBoost 模型的特征有 73 维，用于 LogisticRegression 模型的特征有 96 维。

五、模型训练

5.1 模型的算法选取

本小组使用了 Xgboost，Lightgbm 和 LogisticRegression 模型进行验证，最终选择了 Xgboost 和 LogisticRegression 作为最终的两个模型。

5.2 设计框架

5.2.1 Xgboost

XGBoost（eXtreme Gradient Boosting）是工业界逐渐风靡的基于 GradientBoosting 算法的一个优化的版本，可以给预测模型带来能力的提升，xgboost 具有以下优点：

(1)正则化

标准 GBM 的实现没有像 XGBoost 这样的正则化步骤。正则化对减少过拟合也是有帮助的。

(2)并行处理

XGBoost 可以实现并行处理，相比 GBM 有了速度的飞跃，LightGBM 也是微软最新推出的一个速度提升的算法。

(3)高度的灵活性

XGBoost 允许用户定义自定义优化目标和评价标准 。

(4)缺失值处理

XGBoost 内置处理缺失值的规则。用户需要提供一个和其它样本不同的值，然后把它作为一个参数传进去，以此来作为缺失值的取值。XGBoost 在不同节点遇到缺失值时采用不同的处理方法，并且会学习未来遇到缺失值时的处理方法。

(5)剪枝

当分裂时遇到一个负损失时，GBM 会停止分裂。因此 GBM 实际上是一个贪心算法。XGBoost 会一直分裂到指定的最大深度(max_depth)，然后回过头来剪枝。如果某个节点之后不再有正值，它会去除这个分裂。

(6)内置交叉验证

XGBoost 允许在每一轮 boosting 迭代中使用交叉验证。因此，可以方便地获得最优 boosting 迭代次数。

5.3 建模流程

数据分析 —> 数据预处理 —> 划分训练集和验证集 —> 特征提取 —> 模型选择 —> 模型融合。

5.4 模型评估效果

本小组采用的线下验证统一用 2017 年 4 月的数据作为验证集，2017 年 4 月之前所有数据作为线下训练集，所有数据用作最终模型的构建。

本小组一共使用了 3 个模型，方案一的 Xgboost（XGB1）和方案二的 Xgboost(XGB2)和 Logistic Regression(LR)模型。

XGB1 模型特征 64 维，线下 auc=0.809，线上 a 榜 auc=0.82284，具体参数见下图:

```
params = {
    'booster': 'gbtree',
    'objective': 'binary:logistic',
    'gamma': 0.1,  # 用于控制是否后剪枝的参数,越大越保守，一般0.1、0.2这样子。
    'max_depth': 5,  # 构建树的深度，越大越容易过拟合
    'lambda': 2,  # 控制模型复杂度的权重值的L2正则化项参数，参数越大，模型越不容易过拟合。
    'subsample': 0.8,  # 随机采样训练样本
    'colsample_bytree': 0.8,  # 生成树时进行的列采样
    'min_child_weight': 18,
    'silent': 0,  # 设置成1则没有运行信息输出，最好是设置为0.
    'eta': 0.03,  # 如同学习率
    'eval_metric': 'logloss'
}
```

XGB2 模型特征 73 维，线下 auc=0.817653372567 ，线上 a 榜 auc=0.82213，具体参数见下图:

```
num_round = 500     # 迭代次数 #
params = {
    'booster': 'gbtree',
    'max_depth': 4,
    'colsample_bytree': 0.8,
    'subsample': 0.8,
    'eta': 0.03,
    'silent': 1,
    'objective': 'binary:logistic',
    'eval_metric': 'auc',
    'min_child_weight': 1,
    'scale_pos_weight': 1,
    'seed': 27,
    'reg_alpha': 0.01
}
```

LR 模型特征 96 维，线下验证 auc=0.812781111086 ，线上 a 榜 auc=81328，具体参数见下图:

```
module_two = LogisticRegression(
    penalty='l2',
    solver='sag',
    max_iter=500,
    random_state=42,
    n_jobs=4
)
```

5.5 算法创新点

（1）考虑到离散特征比较多，采用了 LR 进行模型的训练，因为 LR 对离散的特征有更好的处理机制，而 Xgboost 对连续值处理更优秀。

（2）采用 Xgboost，没有用 GBDT

因为传统 GBDT 以 CART 作为基分类器，xgboost 还支持线性分类器，这个时候 xgboost

相当于带 L1 和 L2 正则化项的逻辑斯蒂回归（分类问题）或者线性回归（回归问题）。传统 GBDT 在优化时只用到一阶导数信息，xgboost 则对代价函数进行了二阶泰勒展开，同时用到了一阶和二阶导数。

（3）xgboost 在代价函数里加入了正则项，用于控制模型的复杂度

正则项里包含了树的叶子节点个数、每个叶子节点上输出的 score 的 L2 模的平方和，使学习出来的模型更加简单，防止过拟合。

（4）列抽样（column subsampling）。

xgboost 借鉴了随机森林的做法，支持列抽样，不仅能降低过拟合，还能减少计算。

六、重要特征

列出模型所选的重要特征的前 20 个：表格样式如下：

方案一：

特征名称	特征释义	特征重要性排名
days	用户注册天数	1
credit_score	网购平台信用评分	2
time_days_mean	认证时间和交易时间天数的均值	3
qu_xiao_cnt	用户取消订单的次数	4
q_o	信用额度-额度使用值	5
way1_cnt	用户在线支付的次数	6
overdraft	网购平台信用额度使用值	7
wan_cheng_cnt	用户完成订单的次数	8
0000-00-00	生日为 ’0000-00-00’	9

card_2_cnt	用户有几张信用卡	10
quota	网购平台信用额度	11
qq_bound_one	已绑定 qq	12
sex2	性别为男	13
way2_cnt	用户使用货到付款的次数	14
account_grade_one	用户是否是注册会员	15
card_1_cnt	用户有几张储蓄卡	16
id_bank_set	用户去过几种银行	17
q/o	信用额度/额度使用值	18
account_grade_two	用户是否是铜牌会员	19
id_bank_cnt	用户去过几次银行	20

方案二:

特征名称	特征释义	特征重要性排名
credit_score_rank	用户信用积分排名	1
register_days	用户注册天数	2
quota_surplus	用户已使用的信用额度	3
quota_rate	用户额度使用率	4
account_grade_is_null	用户会员等级是否为空	5
quota	用户信用额度	6
record_count	用户有多少条收货记录	7
store_card_count	用户储蓄卡数量	8

credit_score	用户信用积分	9
credit_count	用户信用卡数量	10
card_record_count	用户在银行卡信息表中的记录数	11
qq_bound_is_null	是否绑定 QQ 号	12
sts_order_len	用户有几种不同的订单状态	13
income1	用户收入等级为“4000-5999 元”	14
货到付款	支付方式中是否包含货到付款	15
birthday_is_zero	用户生日是否是“0000-00-00”	16
订单取消	用户订单状态中是否包含订单取消	17
四川	用户收货地址是否包含四川省	18
type_pay_len	用户有几种不同的支付方式	19
age_two	用户年龄是否在 18-30 这个区间	20

七、创新点

本小组认为我们本次的创新点在于模型异构方面做的比较好，最后通过特征比对分析，三个模型所用的特征差异性比较大，采用了两种不同类型的模型，正是由于特征和模型的较大差异性，所以最终的方案具有很好的泛化能力。

八、赛题思考

1. 对用户的个人信息统计地越详细，违约的可能性越低。
2. 可以定期与用户联系，保证其信用良好。