基于 XGBoost-logistic 串行组合模型的个人信用评分体系

所在学校：上海财经大学

参赛编号：0137

团队负责人：王东岳

团队成员：龙宇航、闫凌曌、吴奕琳、段怡君

指导老师：王黎明

目录

表格和插图清单

摘要..... 1

论文主体

1. 问题描述..... 2

2. 特征选择

2.1 构造特征

2.1.1 个人信息汇总类..... 3

2.1.2 贷款信息类..... 4

2.1.3 贷记卡信息类..... 6

2.1.4 准贷记卡信息类..... 7

2.2 指标处理

2.2.1 缺失值填补..... 8

2.2.2 编码..... 8

2.2.3 非线性变换..... 9

2.3 指标筛选

2.3.1 IV 信息量..... 10

2.3.2 Spearman 和 Hoeffding's D 统计量..... 10

2.3.3 Binning..... 11

2.3.4 变量聚类..... 13

2.3.5 总结..... 13

3. 数据描述

3.1 正态性检验..... 14

3.2 描述性分析

3.2.1 信用卡总授信额度..... 14

3.2.2 未结清贷款总授信额度..... 17
4. 模型建立
4.1 XGBoost 模型..... 21
4.2 逻辑回归模型..... 22
4.3 XGBoost-logistic 串行组合模型..... 22
4.4 信用规则建立..... 23
5. 求解和检验
5.1 XGBoost 模型
5.1.1 预实验..... 25
5.1.2 确定 eta 和 nround..... 26
5.1.3 确定 max_depth 和 min_child_weight..... 26
5.1.4 确定 gamma..... 27
5.1.5 确定 subsample 和 colsample_bytree..... 27
5.1.6 确定 alpha 和 lambda..... 27
5.1.7 降低学习率..... 28
5.1.8 总结..... 28
5.2 逻辑回归模型
5.2.1 因子分析..... 29
5.2.2 逻辑回归..... 30
5.3 XGBoost-logistic 串行组合模型..... 32
5.4 个人信用评分体系..... 34
6. 模型结果分析
6.1 模型分析
6.1.1 XGBoost-logistic 模型..... 35
6.1.2 个人信用评分体系..... 37

6.2 模型预测..... 38
结论和建议..... 39
参考文献..... 41
附录

表格和插图清单

表格 1. 个人信用汇总类.....3
表格 2. 贷款信息类.....4
表格 3. 贷记卡信息类.....6
表格 4. 准贷记卡信息类.....7
表格 5. IV 信息量.....10
表格 6. 指标列表.....13
表格 7. 信用卡授信额度参考因素.....15
表格 8. 贷款授信额度参考因素.....18
表格 9. 贷款最长逾期月数非参数检验.....19
表格 10. 关联性统计量列表.....19
表格 11. 个人信用评分体系.....23
表格 12. XGBoost 参数列表..... 28
表格 13. XGBoost 分类精度..... 29
表格 14. 通过检验的指标 1..... 30
表格 15. 预测概率和观测响应的关联 1..... 30
表格 16. 混淆矩阵 1..... 31
表格 17. 逻辑回归分类精度.....31
表格 18. 通过检验的指标 2..... 32
表格 19. 预测概率和观测响应的关联 2..... 32
表格 20. 混淆矩阵 2..... 33
表格 21. XGBoost-logistic 分类精度.....33
表格 22. 预测概率分位数.....34
表格 23. 个人信用评分体系.....34
表格 24. 特征因子.....35

图 1. Spearman 和 Hoeffding 秩散点图..... 11
图 2. 非线性变换..... 12
图 3. 信用卡总授信额度箱线图..... 14
图 4. 收入水平箱线图..... 15
图 5. 负债水平箱线图..... 16
图 6. 总逾期笔数箱线图..... 16
图 7. 流程风险箱线图..... 17
图 8. 贷款总授信额度箱线图..... 18
图 9. 最近 6 个月平均应还款箱线图..... 19
图 10. 贷款通过率箱线图..... 20
图 11. 短期贷款笔数占比箱线图..... 21
图 12. 验证集 AUC - 迭代轮数 1..... 25
图 13. 验证集 AUC - 迭代轮数 2..... 26
图 14. 验证集 AUC - 迭代轮数 3..... 28
图 15. ROC 曲线 1..... 31
图 16. ROC 曲线 2..... 33
图 17. XGBoost-重要性度量条形图..... 37
图 18. 模型预测流程图..... 38

摘要

互联网金融时代下，精准的个人信用评分体系能够提高金融机构的操作效率，降低企业的授信成本，控制消费信贷的风险，对完善金融市场有着重要意义。鉴于此，我们通过对原始数据进行特征工程，先变量聚类筛选指标，再因子分析提取出 28 个有经济意义的特征因子，计算因子得分。在构造的特征集合上，我们运用机器学习算法 XGBoost 和统计学方法逻辑回归建立个人违约预测模型，并创造性地提出了 XGBoost-logistic 串行组合模型。我们发现该模型兼具 XGBoost 分类精度高以及逻辑回归稳健性好、解释性强的优点，模型指出影响违约的 3 个主要因子分别是：借贷通过因子、查询风险因子和流程风险因子。据此我们给出有针对性的结论和建议。最后，借鉴 FICO 评分系统，我们将模型的输出概率转化成信用得分，建立了个人信用评分体系，将用户的信用划分成优秀、良好、一般和较差四个等级，并给出每个等级下的得分区间和实际违约率。

关键词：特征工程　变量聚类　　因子分析　　组合模型　　FICO 评分

基于 XGBoost-logistic 串行组合模型的个人信用评分体系

1. 问题描述

互联网金融是在互联网逐渐融入人们生活的进程中发展起来的，作为其载体的互联网所具有的开放性和虚拟性也为这种新兴的金融形式带来了更多机遇与挑战。当网络背后真实存在的人被虚拟化，信用就显得尤为重要。如何能够将一个人的信息与他的行为历史数据整合评估，得出有效的信用评价，是互联网金融目前所面临的重要问题之一。

传统的个人信用评分体系属美国的 FICO 评分系统使用最为广泛，FICO 信用分模型利用高达 100 万的大样本的数据，首先确定刻画消费者的信用、品德，以及支付能力的指标，再把各个指标分成若干个档次以及各个档次的得分，然后计算每个指标的加权，最后得到消费者的总得分。

目前数据挖掘与机器学习算法的飞速发展使得我们能够对这一经典问题产生新的视角，而互联网上的繁杂信息也使得原始的指标体系得到大大扩充。一个合理、有效的信用评分体系能够为银行和金融机构省下大量的人力成本与判别时间，对于评价对象也更能管理和约束他们的行为。

本题基于主办方所发放的某贷款机构的历史业务数据展开，主要分为三个部分：

1）从原始数据中构造特征，进行数据挖掘，建立个人违约预测模型，通过训练与测试调整模型，利用所建立的模型预测业务数据中所包含的一万条测试数据的对应结果。

2）基于个人违约预测模型建立信用评分体系，对用户划分信用等级，并运用于一万条测试数据。

3）对个人违约预测模型和个人信用评分体系进行评价。

2. 特征选择

2.1 构造特征

根据所提供 40000 条观测 ID 的基本信息及其部分详细的征信报告板块，我们查阅了大量相关资料并咨询了业内专业人士，运用提取、转换和计算的数据预

处理方法，一共构造了 7 个名义变量和 99 个定量变量作为我们建立模型的候选指标。有关各指标的类型、名称与含义如下。（详细的计算方式请见附录 1）

2.1.1 个人信息汇总类

表格 1. 个人信用汇总类

指标类型	指标名称	含义
基本信息	SEX	性别
基本信息	EDU_LEVEL	教育水平
基本信息	MARRY_STATUS	婚姻状况
基本信息	IS_LOCAL	是否为本地籍
基本信息	AGENT	客户渠道
还款能力	HAS_FUND	是否有公积金
还款能力	SALARY	收入水平
账户时长	LOAN_AGE	账龄
新开立账户	ALL_RECENTLOAN_COUNT	新开立账户数
借贷通过率	LD_PASS_PERCENT	借贷通过率
借贷偏好	PREFERENCE1	贷款偏好
借贷偏好	PREFERENCE2	贷记卡偏好
系统风险	QUERY_RECENT	近期查询次数
系统风险	LOAN_TIME	信息滞后风险
负债水平	ALL_6_AMOUNT	最近 6 个月的平均负债
授信额度	ALL_DS_CREDIT_LIMIT	信用卡总授信额度
授信额度	MAX_DS_CREDIT_LIMIT	信用卡最大授信额度
授信额度	MIN_DS_CREDIT_LIMIT	信用卡最小授信额度
逾期频度	ALL_OVD_COUNT	逾期总笔数和账户数

结合上表，我们认为用户是否有公积金及其收入水平，是最能直接作为衡量其还款能力的标准，对于拥有公积金和较高收入水平的用户，在客观上这类人群具有比较强的还款能力，从而贷款违约的风险也就相对较低。

用户的账龄能够反映与贷款机构维持借贷关系的时间长短，对账龄越长的用户，贷款机构往往掌握着更充足的历史信息，这对判断贷款是否违约将起到重要作用。此外，用户的新开立账户数和近期查询次数越多，说明其经济能力很可能在短期内出现了问题，有较多的资金需求，这对判断是否违约会有一定影响。借贷通过率反映的是用户能否成功借贷的历史信息，较高的通过率意味着贷款机构普遍对其信用有比较高的评价，认可度较高，那么违约的可能性也就自然较低。

我们认为用户偏好贷款或者信用卡的习惯也可能会对偿还贷款起到影响。同时，当信息滞后的时间越长，贷款机构对用户在报告生成时间到放款时间这段时间内的借贷行为就越不清楚，会因信息不对称而对是否违约产生错误判断。

最近 6 个月的平均负债体现用户短期内的负债水平，贷款机构可据此对用户剩余的还款能力做出评估，进而对是否发生违约事件做出判断。相类似地，总授信额度(合同金额)反映的是全体贷款机构对用户总还款能力的评估，最大和最小授信额度分别是单个机构给出的最高和最低的评估，这些对是否放款都有着比较重要的借鉴意义。而逾期频度则是从用户风险的角度进行评估，当用户的逾期笔数和账户数越多，则说明其存在拖欠贷款的习惯或者经济能力较弱，这些都使用户会产生违约的倾向。

2.1.2 贷款信息类

表格 2. 贷款信息类

指标类型	指标名称	含义
账户时长	L_LOAN_AGE	贷款账龄
账户时长	L_LASTLOAN_AGE	贷款新开立账户时长
新开立账户	L_RECENTLOAN_COUNT	新开立贷款笔数
贷款通过率	L_PASS_PERCENT	贷款通过率
贷款来源	L_FINANCE_ORG_COUNT	未结清贷款机构数
负债水平	L_USING_COUNT	当前未结清贷款数
负债水平	L_BALANCE	当前未结清贷款总余额
负债水平	L_SCH_PAY_AMOUNT	本月应还款总额
负债水平	L_6_AMOUNT	最近 6 个月的平均应还款
负债承受能力	L_PERCENT	当前已使用贷款额度比
授信额度	L_CREDIT_LIMIT	未结清贷款总授信额度
逾期频度	L_OVD_COUNT	贷款总逾期笔数
逾期频度	L_OVD_MONTH	贷款总逾期月份数
逾期现状	L_OVD_MONTH_AVG	当前平均逾期期数
逾期现状	L_OVD_MONTH_MAX	当前最大逾期期数
逾期现状	L_OVD_AMOUNT_AVG	当前平均逾期金额
逾期现状	L_OVD_AMOUNT_MAX	当前最大逾期金额
逾期波动	L_OVD_MONTH_STD	贷款逾期月数的标准差
逾期波动	L_OVD_MONTH_CV	贷款逾期月数的变异系数
逾期严重程度	L_HIGHEST_OA_PER_MON	贷款单月最高逾期总额
逾期严重程度	L_DURATION_MAX	贷款最长逾期月数
还款缺陷	L_YJ1_COUNT	贷款异常结清(展期/代还)占比
还款污点	L_YJ2_COUNT	贷款异常结清(提前)占比
波动性	L_CREDITLIMIT_STD	贷款金额标准差
波动性	L_CREDITLIMIT_CV	贷款金额变异系数
贷款结构	L_ZC_PERCENT	贷款账户状态(正常)占比

贷款结构	L_JQ_PERCENT	贷款账户状态(结清)占比
贷款结构	L_QT_PERCENT	贷款账户状态(其他)占比
贷款结构	L_ZFZ_PERCENT	住房贷款应还款占比
贷款结构	L_LONG_PERCENT	长期贷款笔数占比
贷款结构	L_MED_PERCENT	中期贷款笔数占比
贷款结构	L_SHORT_PERCENT	短期贷款笔数占比
贷款结构	L_CREDIT_PERCENT	信用担保笔数占比
贷款结构	L_NCREDIT_PERCENT	非信用担保笔数占比
贷款结构	HOUSE_LOAN_COUNT	个人住房贷款笔数
贷款结构	COMMERCIAL_LOAN_COUNT	个人商用房贷款笔数
贷款结构	OTHER_LOAN_COUNT	其他贷款笔数
贷款结构	ALL_LOAN_COUNT	总贷款笔数
还款能力	L_YQ_PERCENT	当前逾期贷款占比
还款能力	L_Z_PERCENT	过去转出贷款占比
损失风险	L_LOSTBALANCE_MAX	最大可能损失本金

结合上表，我们认为机构数会对发放贷款产生双向影响，当未结清贷款的机构数越多时，说明有越多的贷款机构肯定用户的信用水平，说明其违约的可能性较低，然而一旦发生了违约，该贷款机构与其他贷款机构便会产生竞争倾向，从而造成比较高的损失风险。

当前未结清贷款数、贷款总余额和本月应还款额都是从不同的角度对用户当前的负债水平进行刻画，当用户的负债水平过高，甚至超出了其还款能力，便会发生违约。同时，已使用贷款额度比便是对用户负债承受能力的刻画，当已使用额度越接近授信额度时，我们认为贷款机构需要谨慎放款，此时的违约风险往往较高。

在刻画用户风险时，相比其总体的逾期频度，用户的逾期现状往往更能说明问题。当前最大逾期期数和金额，反映出近期用户逾期造成的最大损失风险，这可以与逾期严重程度对历史信息的刻画相互参照。而当前平均逾期期数和金额，则反映的是平均损失风险。此外，逾期月数的标准差和变异系数能够对用户的逾期波动进行刻画，这对判断是否违约会有一定影响。

虽然还款缺陷和污点相比逾期反映的风险程度较低，但前者借助担保人代还和展期的方式偿还，还是暴露出用户自身还款能力弱的问题，而后者提前还款虽然是对用户还款能力的正向体现，但由于提前偿还会直接造成贷款机构利息收入的减少，固也会影响对贷款的发放。

用户每笔贷款金额的波动其实也反映出其对风险的把控，当波动越大时，我们认为该用户越倾向于缺乏理财能力，进而在还款计划上也相对缺乏科学性，可

能会产生违约风险。

此外，我们认为用户的贷款结构关系到其现金流的分配，这也可能会对违约造成影响。

逾期贷款的占比是对用户还款能力的刻画，该比重越大，能比较直观地说明当前用户的逾期风险越高。而损失风险则是对还款能力的另一种刻画，我们根据不同状态下贷款损失本金进行区间估计，取其置信上限作为最大损失本金，当损失风险越大，用户越有可能会发生违约。

2.1.3 贷记卡信息类

表格 3. 贷记卡信息类

指标类型	指标名称	含义
账户时长	D_LOAN_AGE	贷记卡账龄
账户时长	D_LASTLOAN_AGE	贷记卡新开立账户时长
账户数	LOANCARD_COUNT	贷记卡账户数
新开立账户	D_RECENTLOAN_COUNT	新开立贷记卡账户数
贷记卡通过率	D_PASS_PERCENT	贷记卡通过率
贷记卡来源	D_FINANCE_ORG_COUNT	未销贷记卡机构数
负债水平	D_USING_COUNT	当前未销贷记卡数
负债水平	D_USED_CREDIT_LIMIT	当前未销户贷记卡总已用额度
负债水平	D_SCH_PAY_AMOUNT	本月应还款总额
负债水平	D_6_AMOUNT	最近 6 个月的平均使用额度
负债承受能力	D_PERCENT	当前已使用贷记卡额度比
授信额度	D_CREDIT_LIMIT	未销贷记卡总授信额度
授信额度	D_MAX_CREDIT_LIMIT_PER_ORG	未销贷记卡最大授信额度
授信额度	D_MIN_CREDIT_LIMIT_PER_ORG	未销贷记卡最小授信额度
逾期频度	D_OVD_COUNT	贷记卡总逾期账户数
逾期频度	D_OVD_MONTH	贷记卡总逾期月份数
逾期现状	D_OVD_MONTH_AVG	当前平均逾期期数
逾期现状	D_OVD_MONTH_MAX	当前最大逾期期数
逾期现状	D_OVD_AMOUNT_AVG	当前平均逾期金额
逾期现状	D_OVD_AMOUNT_MAX	当前最大逾期金额
逾期波动	D_OVD_MONTH_STD	贷记卡逾期月数标准差
逾期波动	D_OVD_MONTH_CV	贷记卡逾期月数变异系数
逾期严重程度	D_HIGHEST_OA_PER_MON	贷记卡单月最高逾期总额
逾期严重程度	D_MAX_DURATION	贷记卡最长逾期月数
波动性	D_CREDITLIMIT_STD	贷记卡信用额度标准差
波动性	D_CREDITLIMIT_CV	贷记卡信用额度变异系数

贷记卡结构	D_STATE1	贷记卡正常账户占比
贷记卡结构	D_STATE2	贷记卡销户(未激活)账户占比
贷记卡结构	D_STATE3	贷记卡其他状态占比
贷记卡结构	D_CREDIT_PERCENT	信用担保账户数占比
贷记卡结构	D_NCREDIT_PERCENT	非信用担保账户数占比
还款能力	D_YQ_PERCENT	当前逾期贷记卡占比
还款能力	D_ZC_SCH_PERCENT	正常贷记卡当月应还款占比

结合上表，我们可以看到指标对贷记卡和贷款刻画的角度极为相似，以负债水平为例，由于贷记卡的已用额度和贷款余额本质相同，反映的都是用户需要偿还的金额，所以在这里用已用额度去替代刻画负债水平。

2.1.4 准贷记卡信息类

表格 4. 准贷记卡信息类

指标类型	指标名称	含义
账户时长	S_LOAN_AGE	准贷记卡账龄
准贷记卡来源	S_FINANCE_ORG_COUNT	未销准贷记卡机构数
负债水平	S_USING_COUNT	当前未销准贷记卡数
负债水平	S_USED_CREDIT_LIMIT	当前未销户准贷记卡总已用额度
负债水平	S_6_AMOUNT	最近 6 个月的平均使用额度
负债承受能力	S_PERCENT	当前准贷记卡已用额度比
授信额度	S_CREDIT_LIMIT	准贷记卡总授信额度
授信额度	S_MAX_CREDIT_LIMIT_PER_ORG	准贷记卡最大授信额度
授信额度	S_MIN_CREDIT_LIMIT_PER_ORG	准贷记卡最小授信额度
逾期频度	S_OVD_COUNT	准贷记卡总逾期账户数
逾期频度	S_OVD_MONTH	准贷记卡总逾期月份数
逾期严重程度	S_HIGHEST_OA_PER_MON	准贷记卡单月最高逾期总额
逾期严重程度	S_DURATION_MAX	准贷记卡最长逾期月数

由于原始数据中没有提供准贷记卡的明细信息，所以我们构造的特征数量相对较少，考虑到准贷记卡与贷记卡极为类似，对刻画角度的解读可以参照之前内容。

2.2 指标处理

2.2.1 缺失值填补

在构造的特征集合上，我们认为指标的缺失值存在“有意义”和“无意义”两种。

前者出现在定量变量中，这是因为在原始数据中，并非所有用户都有贷款、贷记卡和准贷记卡的信息，固当我们从这三个角度构造指标时，会因缺少局部借贷信息而造成部分用户在相应指标下出现缺失的情况。显然，这种缺失有一定的解释性。

对于这种“有意义”的缺失值，我们主要的处理方法是将缺失的地方赋“0”，这是因为零值能够比较客观地体现出用户因缺少相应业务而造成该指标缺失的现实意义。

后者则出现在名义变量中，当用户出于对自身信息的保护或者贷款机构出于保护用户的隐私时，会造成部分信息不公开。因为这种缺失没有明显的指向性，所以我们称之为“无意义”的缺失。

对于这种缺失，传统的填补方法像单值填补和多重填补可以在一定程度上保留特征的完整性，提高违约预测的准确性。向辉[1]曾指出当指标缺失在 10%以内时，宜采用删除法；缺失在 20%~40%，宜采用多重填补法；缺失在 50%以上时，各种填补方法均不再有效，最好删除该指标。

SALARY 的缺失程度在 90%以上，对于这种变量原则上应直接删除。但我们经过预实验发现，该指标可以提供比较大的信息量，对预测违约影响较大，固直接删除并不合适。于是我们将缺失的部分单独作为一类，将定序变量作为名义变量处理。

EDU_LEVEL 的缺失程度在 10%~20%之间，考虑到“专科及以下”、“硕士研究生”和“其他”会对原本受教育程度的定序性质产生干扰，所以我们没有将该指标视为定序变量，而是将其视为名义变量。自然的，对于缺失的部分我们也将其单独作为一类处理。

MARRY_STATUS 的缺失程度在 10%以内，考虑到各婚姻状态并没有显著的定序特征，固也将其作为名义变量处理。

2.2.2 编码

我们对没有显著定序特征的名义变量或者存在大量缺失值的定序变量重新编码，主要采用 Dummy 和 WOE 这两种方式。

如果追求该变量的每一类别对响应变量的解释性，我们往往会对其进行 Dummy 化，生成相应的哑变量，将每一类别作为一种状态，使其具有更直观的意义。具体来说，指标所取的每一类别转化成相应的哑变量后，每个哑变量反映是否满足相应状态，不满足取 0，满足取 1。以 EDU_LEVEL 为例，该指标取值为 1-10，对应受教育程度的 10 个类别，故产生相应的 EDU_LEVEL1-EDU_LEVEL10 共 10 个哑变量。但需要注意的是，如果将这些哑变量全部作为解释变量放入模型中，会产生多重共线性，使设计矩阵不满秩，轻则影响参数估计的方差，重则使模型无法求解，所以我们往往会选择一个哑变量作为参照组，不将其投入模型中。

如果不追求变量的每一类别对响应变量的解释性，我们往往会对其进行 WOE 编码，计算公式如下：

$$\text{WOE} = \left[\ln\left(\frac{Distr\ Good}{Distr\ Bad}\right)\right] \times 100$$

其中 Distr Good 表示该类别不违约观测数占不违约观测总数的比例，而 Distr Bad 表示的则是该类别违约观测数占违约观测总数的比例。

通过计算变量每一类别的 WOE，我们将得到的 WOE 值去替换原始值，并将编码后的变量视为定量变量处理。经验告诉我们，这样的处理往往会提升模型的整体效果。

2.2.3 非线性变换

如果定量解释变量跟响应变量之间呈现某种非线性关系(此时的响应变量会先被转换成 empirical logits)，那么我们对该变量进行分箱(binning)处理，通过用离散化的变量去替代原始变量，可以使变量处理后跟响应变量间大致呈现线性关系，提升模型的整体效果，同时保留变量的解释性。

二元响应变量为了更好地揭示与连续型变量间的关系需要被平滑，一种简单而稳健的平滑方法是对绘制出解释变量分位数的 empirical logits，这些 logits 对每一个 bin 下违约事件的比率进行 Minimax 估计，其计算公式如下：

$$\text{Empirical Logits} = \ln\left(\frac{m_i + \frac{\sqrt{M_i}}{2}}{M_i - m_i + \frac{\sqrt{M_i}}{2}}\right)$$

其中在每个 bin 下，m_i是违约发生数，M_i是观测实例数。

2.3 指标筛选

2.3.1 IV 信息量

我们在有标签样本集上对名义变量进行 WOE 编码并计算 IV信息量, 判断各指标对违约情况的解释能力，计算公式如下：

$$\mathrm{IV} = \sum_{i=1}^{n} (Distr\ Good - Distr\ Bad) \times \ln\left(\frac{Distr\ Good}{Distr\ Bad}\right)$$

各指标结果见下表:

表格 5. IV 信息量

指标名称	IV
SEX	0.0227757
EDU_LEVEL	0.2242483
MARRY_STATUS	0.0258425
IS_LOCAL	0.0257412
AGENT	0.2607333
HAS_FUND	0.0009402
SALARY	0.2130032

根据上表，我们看到除 HAS_FUND 外，其余名义变量对违约情况具有比较强的解释能力，以 EDU_LEVEL、AGENT 和 SALARY 更为显著，这说明我们提取的指标比较合理，同时在建模前可以删去对响应变量影响不大的 HAS_FUND。

2.3.2 Spearman 和 Hoeffding's D 统计量

Spearman 相关系数通过对解释变量进行排秩，计算与响应变量之间的相关性，相比 Pearson 相关系数，前者对变量间的非线性以及异常值更不敏感。但是当变量间不再是单调相关(monotonically related)时，使用 Spearman 相关系数可能会使我们错过变量间的一些重要关联。此时一种常用而稳健的统计测度 Hoeffding's D 可能就更加有效，帮助我们捕捉到比较广泛的变量关联性。

故我们分别计算定量变量与响应变量之间的 Spearman 和 Hoeffding's D，并根据变量在这两个统计量上表现出来的相关性由强到弱进行排秩，最后对这两组排秩结果绘制散点图如下：

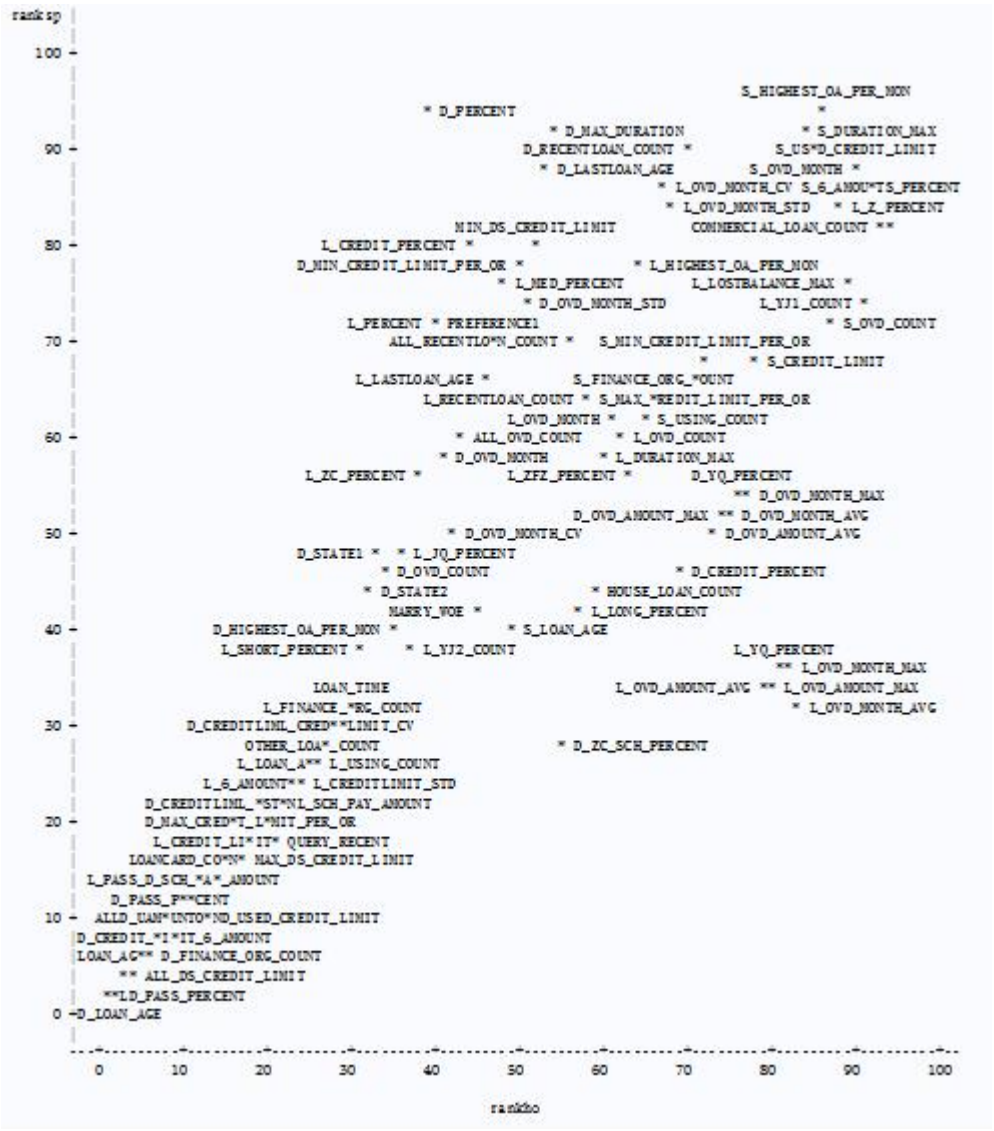

图 1. Spearman 和 Hoeffding 秩散点图

根据上图，我们可以看到位于右上角的变量 S_HIGHEST_OA_PER_MON、S_DURATION_MAX、S_OVD_MONTH、S_USED_CREDIT_LIMIT 和 S_PERCENT，他们的 Spearman 和 Hoeffding's D 排得都比较靠后，说明这些变量跟响应变量间的相关性都比较小，通常情况下我们可以删除这些变量。

对于 Spearman 排秩靠后，而 Hoeffding's D 排秩靠前的变量，它们位于左上方区域，比如 D_PERCENT、D_LASTLOAN_AGE 和 L_PERCENT，很可能跟响应变量间存在非线性关系，我们需要对其进行变换。

2.3.3 Binning

根据 Spearman 和 Hoeffding's D 统计量，我们发现有些变量与响应变量之间

存在非线性关系，故对其进行变换。常用的非线性变换像加入交叉项、高次项以及 Xlog(X)形式的确可以提升模型的拟合效果，但缺点是变换后的变量比较难找到对应的科学解释。

我们这里对连续型变量分箱(binning)，使变换后的变量与响应变量间呈现比较好的线性关系。这里以 L_PERCENT 进行说明：

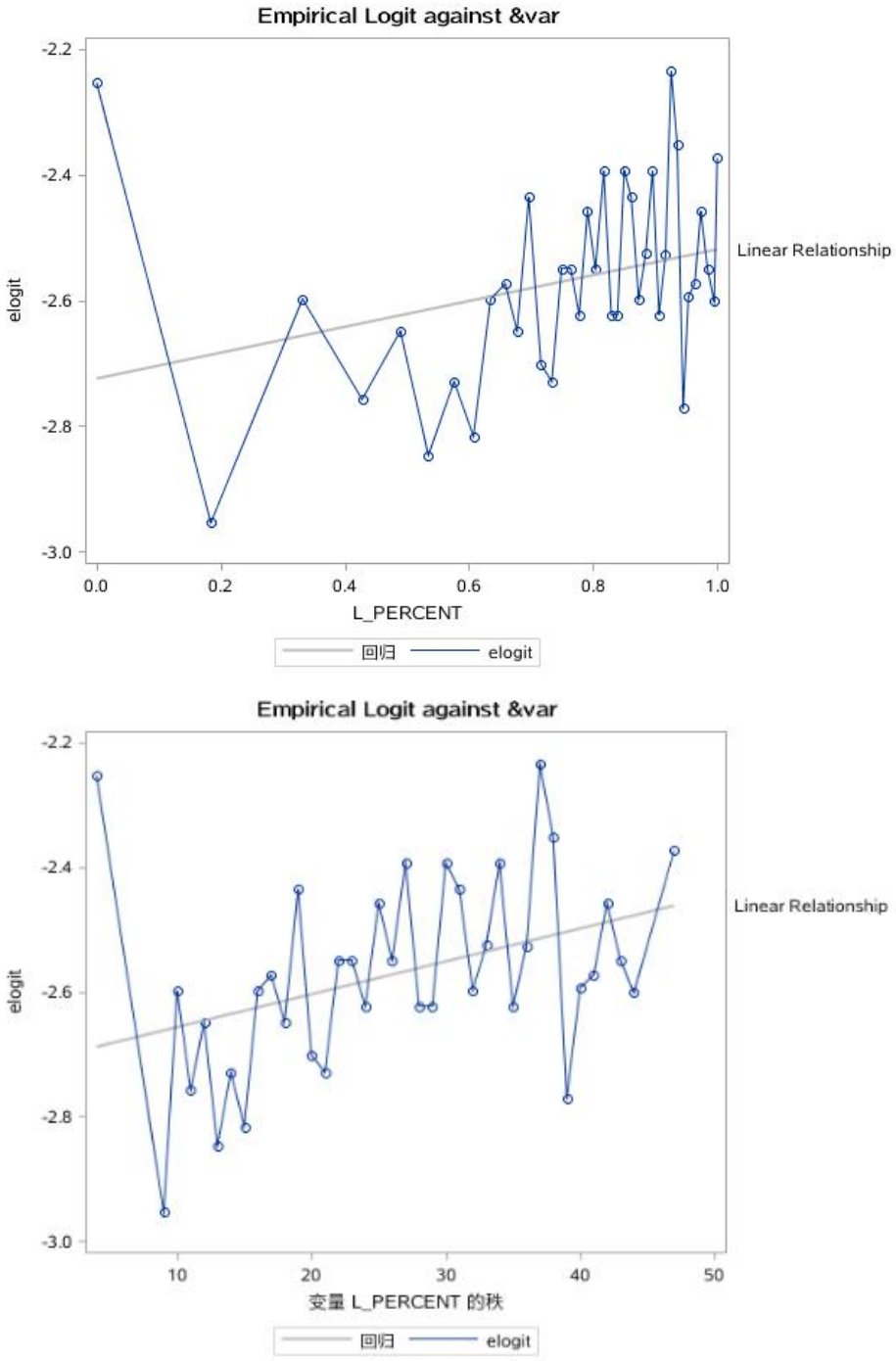

图 2. 非线性变换

根据上图我们可以看到，在进行变换前，L_PERCENT 与 empirical logits 并不是完全线性的，而在变换后可以看到 L_PERCENT 基本在 logits 的线性趋势周围波动，说明我们的变换是有效的。

此外，进行变换的解释变量还有 D_PERCENT 和 D_LASTLOAN_AGE。

2.3.4 变量聚类

通过将相关性强的变量聚到相同类，而相关性弱的变量聚到不同类，我们对解释变量进行分组，再选择每个组内 RSquareRatio 最小的作为代表变量，从而实现特征降维。(聚类结果请见附录 2)

2.3.5 总结

通过对特征进行筛选，我们最终得到了 64 项指标，可以认为这些指标对响应变量间有着比较强的解释能力，具体指标请见下表：

表格 6.指标列表

个人信用汇总类		
IS_LOCAL	SEX	SALARY1
SALARY2	SALARY3	SALARY4
SALARY5	SALARY6	SALARY7
EDU_LEVEL1	EDU_LEVEL2	EDU_LEVEL3
EDU_LEVEL5	EDU_LEVEL6	EDU_LEVEL7
EDU_LEVEL8	ALL_6_AMOUNT	MARRY_WOE
QUERY_RECENT	LD_PASS_PERCENT	LOAN_TIME
贷款信息类		
L_FINANCE_ORG_COUNT	L_SHORT_PERCENT	L_YJ2_COUNT
L_OVD_MONTH	L_LASTLOAN_AGE	L_LOAN_AGE
L_JQ_PERCENT	L_OVD_MONTH_CV	L_OVD_COUNT
L_HIGHEST_OA_PER_MON	L_PERCENT_BIN	L_YQ_PERCENT
L_ZFZ_PERCENT	L_YJ1_COUNT	L_Z_PERCENT
L_MED_PERCENT	L_OVD_MONTH_STD	L_LOSTBALANCE_MAX
COMMERCIAL_LOAN_COUNT	L_RECENTLOAN_COUNT	L_6_AMOUNT
L_BALANCE	L_OVD_AMOUNT_AVG	
贷记卡信息类		
D_OVD_MONTH_CV	D_CREDIT_PERCENT	D_OVD_AMOUNT_MAX
D_PERCENT_BIN	D_LASTLOAN_AGE_BIN	D_LOAN_AGE
D_USING_COUNT	D_OVD_COUNT	D_STATE1

D_YQ_PERCENT	D_OVD_MONTH	D_OVD_MONTH_MAX
D_MIN_CREDIT_LIMIT_PER_ORG	D_RECENTLOAN_COUNT	D_MAX_CREDIT_LIMIT_PER_ORG
D_CREDITLIMIT_CV	D_CREDITLIMIT_CV	
准贷记卡信息类		
S_OVD_COUNT	S_FINANCE_ORG_COUNT	S_MAX_CREDIT_LIMIT_PER_ORG

3. 数据描述

我们将定量变量视为自变量，用户的违约情况视为因变量，从而对各定量变量与是否违约间的内在关联进行统计分析。

3.1 正态性检验

我们对定量变量进行正态性检验，发现均没有通过检验，所以在进行连续性自变量与因变量的描述性分析时应该采用非参数检验的方法，由于因变量是二分类变量，故我们采用 Wilcoxon 秩和检验，结果显示两类人群在不同指标下均存在显著差异。(结果请见附录 3)

3.2 描述性分析

我们分别从信用卡和贷款的角度对用户特征进行分析。

3.2.1 信用卡总授信额度

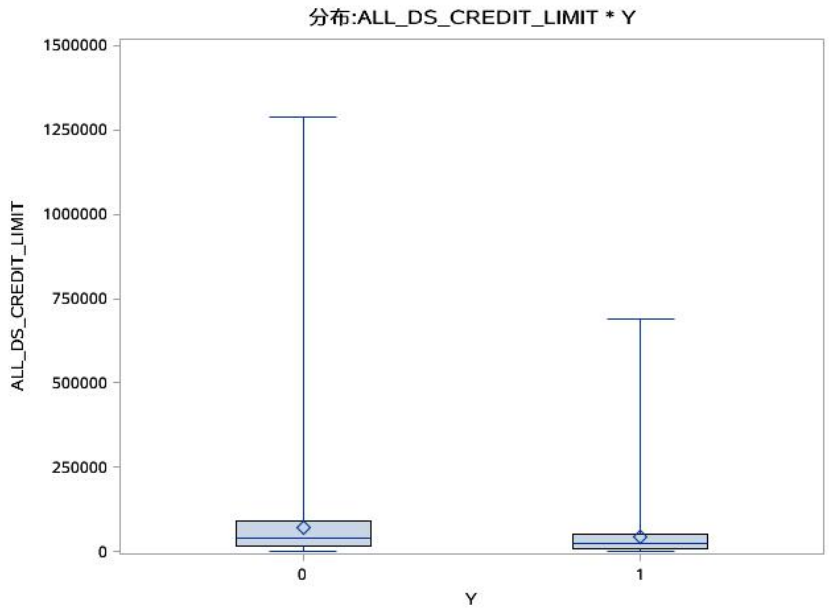

图 3. 信用卡总授信额度箱线图

由上图可以看出，违约客户的信用卡总授信额度显著低于未违约客户，也就是说对违约客户的授信额度要低于未违约客户。这说明金融机构的风险控制是有效的。在对潜在违约客户进行信用评价时，有效的评价应该给予一个消极的评价，即使决定授信，也应该控制授信额度，使得对于潜在违约客户的授信额度低于那些预期不会违约的客户。金融机构在确定客户的授信额度时应该综合考虑以下四点：

表格 7. 信用卡授信额度参考因素

因素	代表变量
还款能力	收入
当前负债水平	负债水平(最近)
信用历史	总逾期笔数(账户数)
流程风险	信息滞后风险

3.2.1.1 还款能力

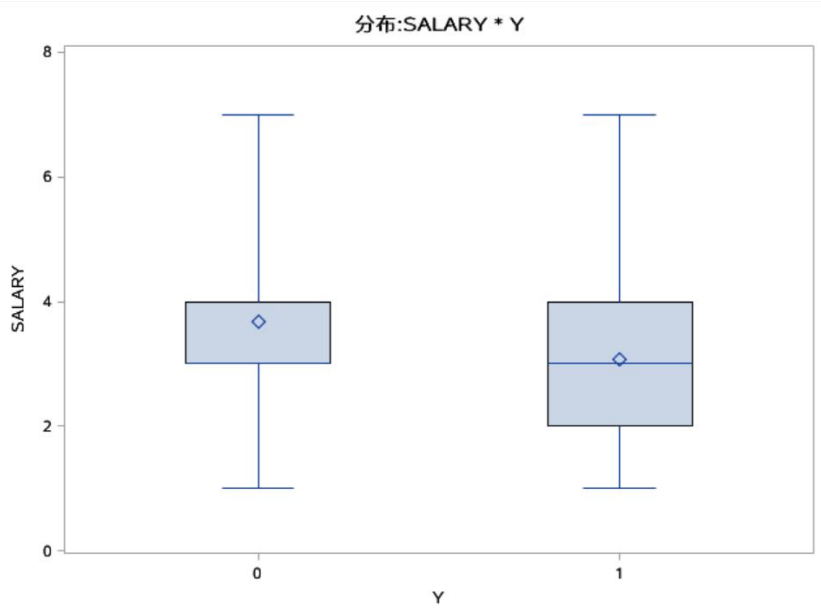

图 4. 收入水平箱线图

由上图可以看出，违约客户的收入显著低于未违约客户，也就是说违约客户的还款能力低于未违约客户。收入低的客户还款能力弱，如果出现预期外的开支很容易就发生违约。金融机构在授信时收入往往是考虑还款能力的一项重要指标。

3.2.1.2 当前负债水平

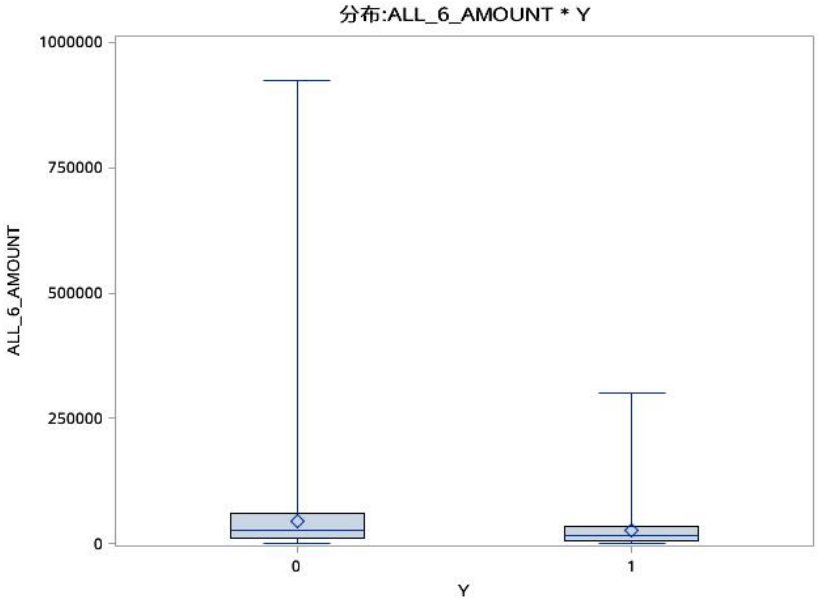

图 5. 负债水平箱线图

由上图可以看出，违约客户的当前负债水平显著低于未违约客户，也就是说违约客户的当前负债水平要低于未违约客户。由于对于预期不会违约的客户的授信额度较高，因此预期不会违约的客户的负债水平应该显著高于潜在违约客户，正是通过对于预期不会违约的客户的授信额度的增加，这一类客户的的负债水平较高，金融机构才能实现在控制风险的前提下实现利益最大化。

3.2.1.3 信用历史

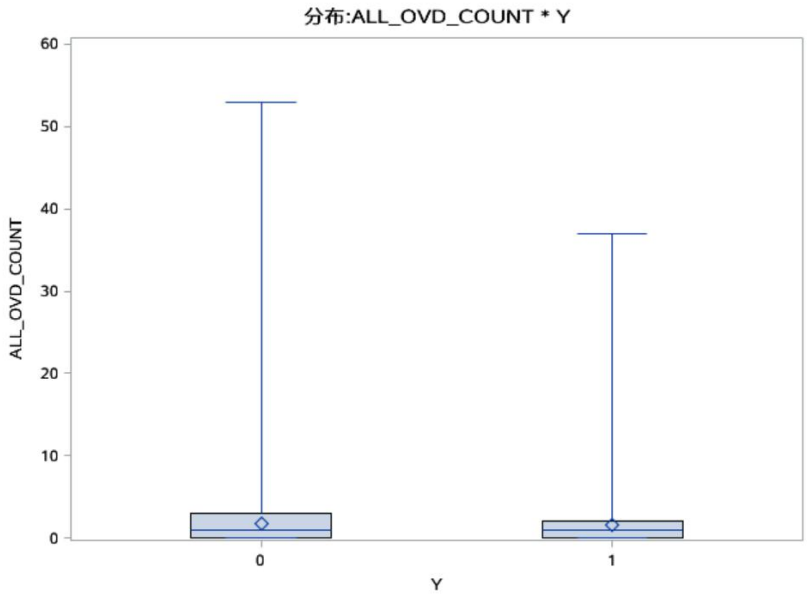

图 6. 总逾期笔数箱线图

由上图可以看出，违约客户的总逾期笔数(账户数)显著低于未违约客户，也就是说违约客户的信用历史优于未违约客户。违约客户的总逾期笔数(账户数)显著低于未违约客户是因为逾期与违约之间存在差别，逾期表示延迟还款，而违约是拒绝还款。未违约客户的总逾期笔数(账户数)较高在给金融机构的流动性带来不利影响的同时需要给金融机构以各种形式的补偿，比如罚款等。未违约客户的总逾期笔数(账户数)显著高于未违约客户可以为金融机构在不至于发生违约风险的前提下增加收益。

3.2.1.4 流程风险

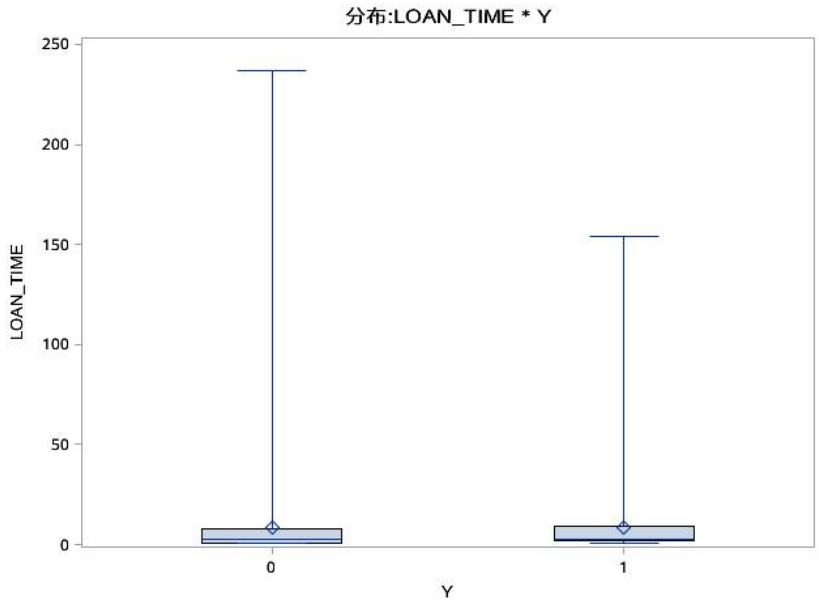

图 7. 流程风险箱线图

由上图可以看出，违约客户的信息滞后风险显著高于未违约客户，也就是说对违约客户的系统性风险要高于未违约客户。这体现了金融机构的风险控制过程。对于潜在的违约客户，从报告生成到发放贷款的时间较长，通过对于潜在违约客户的进一步考察从而减少对违约客户的授信。

3.2.2 未结清贷款总授信额度

因为 Wilcoxon 统计量的 p 值小于 0.05，在显著性水平 0.05 条件下认为违约客户与不违约客户的未结清贷款总授信额度存在显著差异。

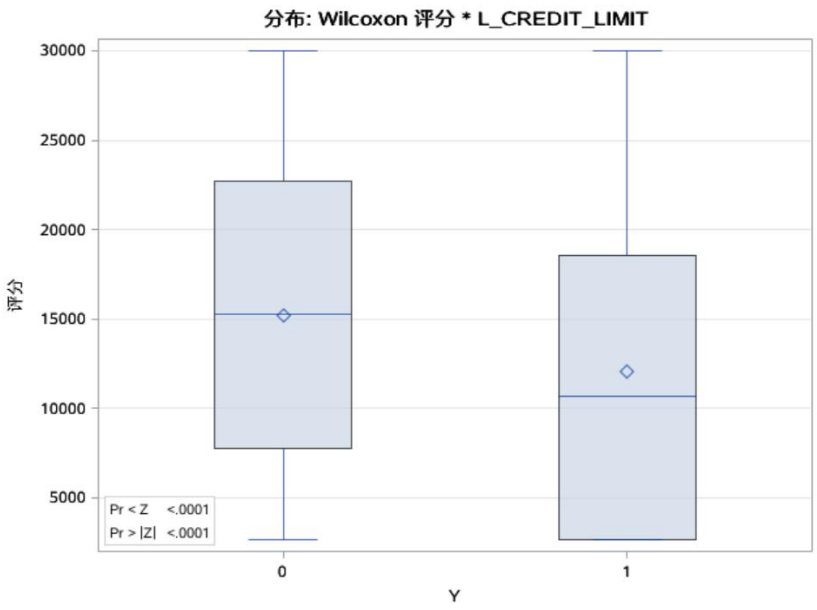

图 8. 贷款总授信额度箱线图

由上图可以看出，违约客户的未结清贷款总授信额度显著低于未违约客户，也就是说对违约客户的授信额度要低于未违约客户。这说明金融机构的风险控制是有效的。类似地，金融机构在确定客户的贷款授信额度时应该综合考虑以下四点：

表格 8. 贷款授信额度参考因素

因素	代表变量
当前负债水平	最近 6 个月的平均应还款
信用历史	贷款最长逾期月数
系统性风险	贷款通过率
贷款结构	短期贷款笔数占比

3.2.2.1 当前负债水平

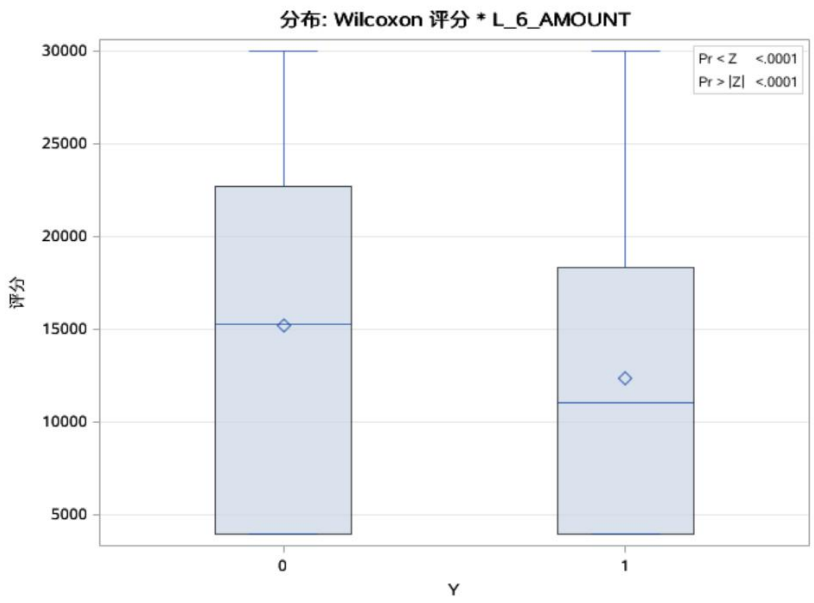

图 9. 最近 6 个月平均应还款箱线图

由上图可以看出，违约客户的最近 6 个月的平均应还款显著低于未违约客户，也就是说对违约客户的当前负债水平要低于未违约客户。正是通过使得对于预期不会违约的客户的授信额度较高，这一类客户的负债水平较高，金融机构才能实现在控制风险的前提下实现利益最大化。

3.2.2.2 信用历史

表格 9. 贷款最长逾期月数非参数检验

统计量	自由度	值	概率
卡方	7	41.6647	<.0001
似然比卡方检验	7	31.621	<.0001
Mantel-Haenszel 卡方	1	23.3038	<.0001
Phi 系数		0.0373	
列联系数		0.0372	
Cramer V		0.0373	

卡方统计量的 p 值以及 Mantel-Haenszel 卡方统计量的 p 值小于 0.05，在显著性水平 0.05 条件下认为贷款最长逾期月数与是否违约之间存在显著有序相关。

表格 10. 关联性统计量列表

统计量	值	ASE
Gamma	0.1109	0.0293
Kendall's Tau-b	0.0215	0.0061
Stuart's Tau-c	0.0078	0.0022

Somers' D C|R	0.0332	0.0094
Somers' D R|C	0.014	0.004
Pearson 相关	0.0279	0.007
Spearman 相关	0.0218	0.0062
Lambda 非对称 C|R	0	0
Lambda 非对称 R|C	0	0
Lambda 对称	0	0
不确定系数 C|R	0.0009	0.0004
不确定系数 R|C	0.0023	0.0009
不确定系数对称	0.0013	0.0005

由以上相关系数表可以看出，这种有序相关是有序正相关。贷款最长逾期月数越长的客户越有可能是违约客户。贷款最长逾期期数越长说明客户的信用历史较差，随着逾期期数的增加，坏账的可能性也随之增加。

3.2.2.3 系统性风险

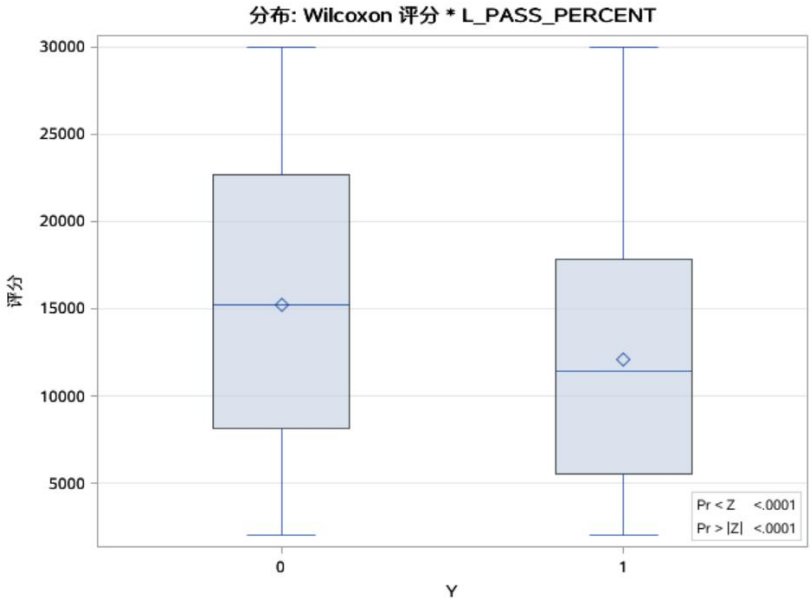

图 10. 贷款通过率箱线图

由上图可以看出，违约客户的贷款通过率显著低于未违约客户，也就是说违约客户的系统性风险要低于未违约客户。通过降低对于潜在违约客户的贷款通过率，金融机构可以实现风险控制。

3.2.2.4 贷款结构

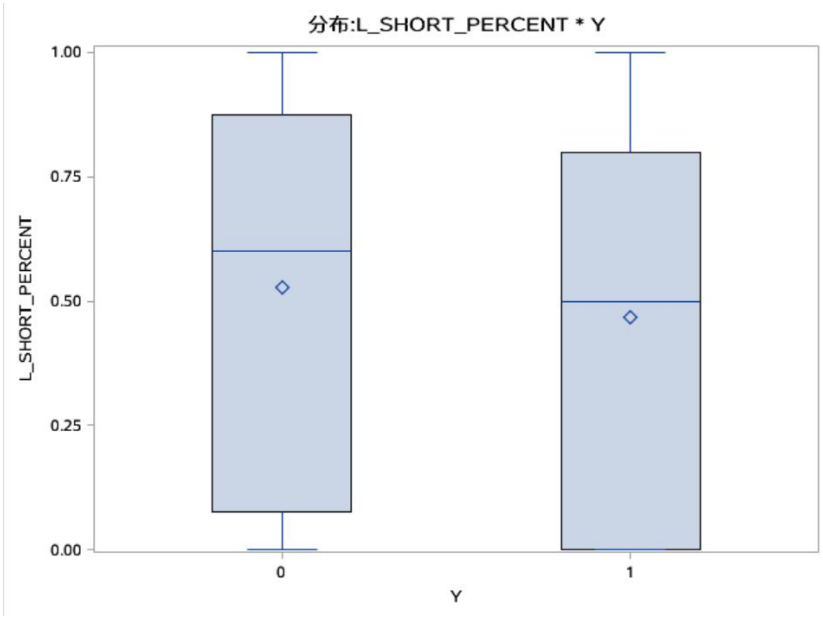

图 11. 短期贷款笔数占比箱线图

由上图可以看出，违约客户的短期贷款笔数占比显著低于未违约客户，也就是说，短期贷款笔数占比越高的客户越有可能违约。

4. 模型建立

4.1 XGBoost 模型

我们运用时下热门的机器学习算法 XGBoost，对用户的违约情况进行预测。

众所周知，XGBoost 以“正则化提升”技术而闻名，对传统的 Boosting 算法进行了全面的优化，它允许用户自定义优化目标和评价标准，使得模型有高度的灵活性，与此同时，该算法还实现了并行处理，运行效率高，时间开销小。这些优点都非常适合对借贷违约进行准确预测。

相比于传统的统计模型，XGBoost 可以捕捉变量间更加复杂的非线性关系，提高模型的预测精度。但跟其他机器学习算法一样，它属于黑箱模型，并不能直观地反映各指标对违约情况的解释性，因此比较难给出经济学解释。此外，XGBoost 中参数较多，调整起来比较复杂，所设置的参数会对预测结果产生显著影响，对此我们在查阅相关资料后，给出的调参方法如下:

STEP1. 设置初始值；

STEP*2*. 确定学习速率(eta)和相应的树的数目(nround)；

STEP*3*. 栅格搜索（grid search）确定每棵树的最大深度（max_depth）和最小节点权重（min_child_weight）；

STEP*4*. 调节 gamma 参数；

STEP*5*. 栅格搜索确定每棵树训练的样本子集（subsample）和特征子集（colsample_bytree）的大小；

STEP*6*. 栅格搜索确定每棵树的正则化参数 alpha 和 lambda；

STEP*7*. 重新设置更小的学习率，确定此时的 nround。

4.2 逻辑回归模型

逻辑回归作为广义线性回归中的一种，其本身就具备可靠的统计理论基础，稳健性较高。由于其属于白箱模型，从而能提供各指标与违约情况的显示表达式，可以直观地反映各指标对违约的解释性，进而给出相应的经济学解释。

模型的基本公式如下：

$$logit(p) = X\beta + \epsilon, \epsilon \sim N(0, I\sigma^2)$$

$$\hat{p} = \frac{exp\left(X\hat{\beta}\right)}{1 + exp\left(X\hat{\beta}\right)}, \hat{p}\text{为违约概率}$$

但是在过抽样的样本上训练模型时，预测概率会产生偏置$\ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right)$，此时模型为：

$$logit(p^*) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + X\beta + \epsilon, \epsilon \sim N(0, I\sigma^2)$$

其中，p^*是有偏样本的后验概率，并且有$\pi_0 > \rho_0$和$\pi_1 < \rho_1$，故调整后的后验概率为$\bar{p} = \frac{p^*\rho_0\pi_1}{(1-p^*)\rho_1\pi_0 + p^*\rho_0\pi_1}$。

4.3 XGBoost-logistic 串行组合模型

在串行结构的组合模型中，各分类器的学习是按照顺序进行的，前一个分类器的输出作为后一个的输入。如果能选择合适的分类器进行串行组合，那么组合分类器可能比单一分类器具有更加优秀的性质。

机器学习 XGBoost 模型分类精度高，但结果缺乏稳健性和解释性，传统的逻辑回归模型尽管在分类精度上比机器学习模型稍显逊色，但其稳健性好，可解释性高。所以，一种自然的解决思路就是将二者结合起来，形成优势互补，最终构建一个分类精度、稳健性俱佳的组合模型。

根据这个思路，我们构建了 XGBoost-logistic 串行组合模型：

STEP1. 在训练集上训练 XGBoost 模型；

STEP*2*. 对训练集和测试集中样本的违约情况进行预测并输出预测标签；

STEP*3*. 将训练集上的预测标签和其他特征指标一起建立逻辑回归模型；

STEP*4*. 用 STEP*3* 得到的逻辑回归模型计算测试集中样本的违约概率，得到相应的预测标签。

4.4 信用规则建立

经典的 FICO 信用评分模型将用户的信用水平通过 300 至 800 的分数进行体现，得分越高的用户，信用评价越好，出现违约的可能性也就越低。借鉴 FICO 的思路，我们也将用户的信用评价用 300~800 的得分区间表示，其中 300 分为基本得分，剩余 500 分根据模型预测的违约概率确定。

通过之前的 XGBoost-logistic 模型，我们可以得到每个用户的预测概率 p。

由此，相应的信用评分公式如下：

$$Score = 300 + 500 \cdot (1 - p)$$

考虑到模型的分类阈值为 0.0625，我们分别对预测为违约与不违约这两类人群的概率 p 进行分位数统计。我们将不违约用户 75%分位数视为分类下限，违约用户 25%分位数视为分类上限。对违约概率小于分类下限的用户，我们认为用户的信用比较好，可以优先放款；对违约概率大于分类下限，而小于分类阈值的用户，他们的信用水平相对较好，不过有待进一步考察；对违约概率大于分类阈值，而小于分类上限的用户，他们的信用水平相对较差，不过有待进一步考察；对违约概率大于分类上限的用户，我们认为用户的信用水平较差，应不给予放款以控制风险。

最终我们得到的信用评价体系见下表：

表格 11. 个人信用评分体系

得分区间	信用评价
$300+500\cdot(1-p_{0,0.75})\sim 800$	优秀
$300+500\cdot(1-cutoff)\sim 300+500\cdot(1-p_{0,0.75})$	良好
$300+500\cdot(1-p_{1,0.25})\sim 300+500\cdot(1-cutoff)$	一般
$300\sim 300+500\cdot(1-p_{1,0.25})$	较差

5. 求解和检验

我们将有标签的 30000 条观测进行三七分。70%作为训练集以训练模型，剩下 30%的样本作为测试集以检验模型的最终效果。

在训练 XGBoost 模型时，我们根据在训练集上交叉验证的结果选择最优的模型参数。具体来说在每轮调参中，我们在训练集上进行 5 折-交叉验证，并选择在交叉验证中验证集平均 AUC 得分最高的作为最优模型，而其对应的参数便是我们该轮的最优参数。

考虑到训练集正类和反类是不平衡的，我们在设置分类阈值(cutoff)时并不使用 0.5 作为划分标准，而是从几率(odds)的角度考虑，模型对样本的预测几率为$\frac{\hat{y}}{1-\hat{y}}$，设θ等于正类样本数除以全部样本数，那么样本的真实几率为$\frac{\theta}{1-\theta}$，当观测几率大于真实几率时，也就是$\hat{y}>\theta$时，那么就判断这个样本属于正类。

虽然我们无法获悉真实样本总体，但对于样本集，存在这样一个假设：样本集是真实样本总体的无偏采样。正因为这个假设，所以我们认为有标签样本集的观测几率$\frac{\hat{\theta}}{1-\hat{\theta}}$，就代表了真实几率$\frac{\theta}{1-\theta}$。

这里的$\hat{\theta}$ =1875/30000=0.0625，此为分类阈值(cutoff)。

值得注意的是，实际中的 cutoff 会随总体分布的变化而变化，而总体分布又时常会发生改变，故我们在选择最优模型时不应该将跟 cutoff 有关的统计量纳为检验标准(如查全率，查准率等)，于是我们选择 AUC，这一业内比较能接受的统计量作为我们比较模型优劣的标尺。

此外根据经验，模型在平衡的数据集上训练，相比非平衡数据，更容易捕捉到少类的用户特征，而常用的平衡方法便是过抽样。于是，我们这里将训练集中的多类和少类过抽样至 2:1 进行训练。

5.1 XGBoost 模型

5.1.1 预实验

我们使用 R 软件中 xgboost 程序包建立模型，模型的输入为 16 个哑变量和 48 个定量变量。

train_auc_mean 和 test_auc_mean 是模型在进行 5 折-交叉验证时，分别在训练样本上的平均 AUC 和验证样本上的平均 AUC；而 train_auc_std 和 test_auc_std 表示的则是模型在 5 轮交叉训练中 AUC 的标准差，反映出波动大小。

我们先在过抽样训练集上训练 XGBoost 模型，给定合适的初始参数后训练 1500 轮，验证集上的 AUC 结果如下：

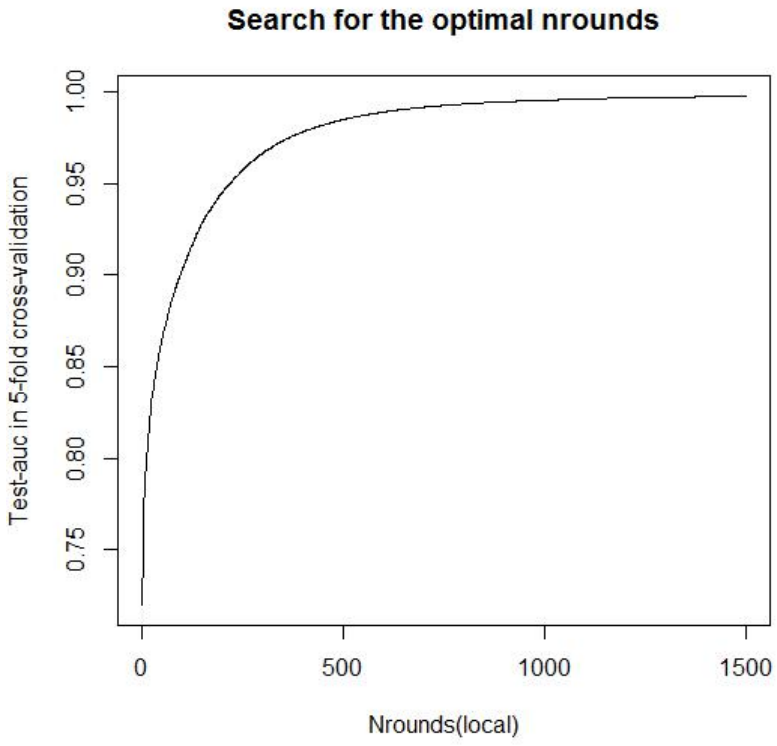

图 12. 验证集 AUC－迭代轮数 1

结合上图，我们发现模型在训练集和验证集上的 AUC 随着迭代次数的增加，都单调递增趋于 1，经计算，模型在测试集上的 AUC 却只处在 0.7 的水平，初步判断应该是出现了过拟合现象。由于模型在验证集上无法收敛到最优情况，故我们也无法对参数进行调节。

所以我们选择在非平衡的样本集上训练 XGBoost 模型，事后表明这样的选择是正确的。我们认为这很可能是因为该模型本身可以通过迭代的方式，对少类进

行很好地识别，而过抽样反而弄巧成拙造成过拟合。

5.1.2 确定 eta 和 nround

Eta 表示学习率，通过减少每一步的权重，可以提高模型的鲁棒性。Nround 表示迭代轮数，也是模型中树的棵树。

我们在非平衡数据集上设置一组合理的初始参数，并将学习率(eta)设为 0.1，基于该参数训练模型。通过计算在不同训练轮数(nround)上验证集的 AUC，我们选择 AUC 最高 nround 作为我们之后训练的参数。

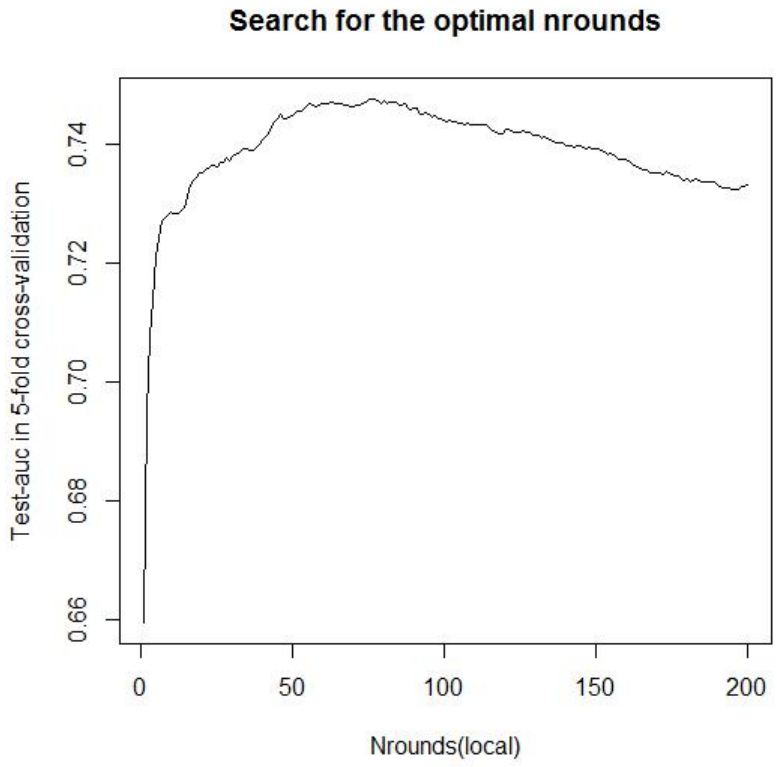

图 13. 验证集 AUC－迭代轮数 2

根据上图，我们可以看到在 eta=0.1 时，模型在验证集上的 AUC 先迅速提高，在达到 71 轮之后，AUC 便逐步降低，这充分反映出模型由欠拟合向过拟合的变化过程，故我们将 nround 设为 71，此时模型的泛化能力最好。

5.1.3 确定 max_depth 和 min_child_weight

Max_depth 表示树的最大深度，取值越大，模型会学到更具体更局部的样本。

Min_child_weight 表示最小叶子节点样本权重和，取值越大，可以避免模型学习到局部的特殊样本。我们经过栅格搜索，将 max_depth 的范围设为[3 , 7]，将 min_child_weight 的范围设为[1 , 50]。

实验结果显示，当 max_depth=3，min_child_weight=10 时，此时模型在验证集上的 AUC 最高，均值为 0.751762，标准差为 0.013742。（结果参见附录 4）

5.1.4 确定 gamma

Gamma 表示节点分裂所需的最小损失函数下降值，取值越大，算法越保守，避免过拟合。我们将 gamma 分别取为 0, 0.01, 0.05, 0.1, 0.5, 1 和 5。

实验结果显示，当 gamma=1 时，此时模型在验证集上的 AUC 最高，均值为 0.750222，标准差为 0.010381。我们注意到相比调参前，AUC 略显下降，这主要是由之前的标准差较大造成的。（结果参见附录 4）

5.1.5 确定 subsample 和 colsample_bytree

Subsample 表示每棵树对样本随机采样的比例，取值越小，算法越保守，避免过拟合。colsample_bytree 表示每棵树在特征集合上的采样比例，跟 subsample 的作用相近。我们经过栅格搜索，将 subsample 和 colsample_bytree 的范围均设为[0.5 , 1]。

实验结果显示，当 subsample=0.9，colsample_bytree=0.8 时，此时模型在验证集上的 AUC 最高，均值为 0.751259，标准差为 0.006942。我们注意到相比调参前，AUC 有所提升，标准差也变得更小了。（结果参见附录 4）

5.1.6 确定 alpha 和 lambda

Alpha 表示权重的 L1 正则化项，lambda 则表示权重的 L2 正则化项，两个参数均用来控制 XGBoost 的正则化部分。我们经过栅格搜索，将 alpha 和 lambda 的范围均设为[0 , 0.5]。

实验结果显示，当 alpha=0.1，lambda=0.01 时，此时模型在验证集上的 AUC 最高，均值为 0.750163，标准差为 0.009102。（结果参见附录 4）

5.1.7 降低学习率

我们将 eta 重设为 0.01，让模型更加仔细地学习，重新选择相应 nround。

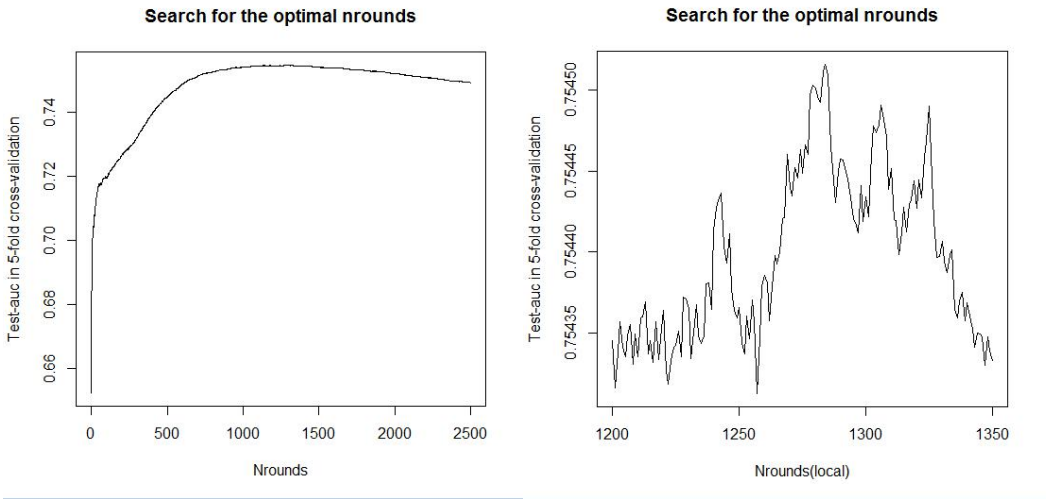

图 14. 验证集 AUC－迭代轮数 3

根据上图，我们可以看到当 eta=0.01 时，模型验证集上的 AUC 在 1284 轮时达到最高点，此时模型的泛化能力最好。

到这里，我们整个调参过程结束，模型最终在训练集上的 AUC 为 0.8304384，验证集上的 AUC 为 0.754516，测试集上的 AUC 为 0.7456046。

5.1.8 总结

最终模型训练的参数如下：

表格 12. XGBoost 参数列表

参数名	取值	参数名	取值
eta	0.01	subsample	0.9
nround	1284	colsample_bytree	0.8
max_depth	3	alpha	0.1
min_child_weight	10	lambda	0.01
gamma	1		

我们将 0.0625 作为 cutoff，得到在训练集和测试集上相应的查全率 $(\frac{TP}{TP+FN})$ 以及整体错分率如下：

表格 13. XGboost 分类精度

指标	值
Accurary0_train	67.9262%
Accurary1_train	81.0606%
Accurary_all_train	68.7524%
Accurary0_test	67.7224%
Accurary1_test	69.3694%
Accurary_all_test	67.8239%

根据查全率，我们可以看到模型对违约用户的识别率要高于对不违约用户的识别率。考虑到现实中贷款机构往往是风险厌恶者(即宁可将不违约用户错判成违约用户，也不将违约用户错判成不违约)，所以我们的模型对贷款机构降低风险会有显著的帮助，是有效的。

5.2 逻辑回归模型

5.2.1 因子分析

为了使最终模型的变量可以更加直观地刻画用户特征，我们对前期筛选得到的 48 个定量变量进行因子分析提取公因子。一方面,我们发现不同指标在含义上仍存在交叉，经过变换，将指标信息压缩到少数相互独立的因子上，这可以帮助我们重整信息，提高模型计算的效率，也可以避免多重共线性对参数估计造成的不准确。另一方面，构造具有实际经济意义的因子，计算因子得分，比较不同用户在相同因子下的得分，可以使我们更好地描述用户特征。

我们使用 SAS 9.4 中的 FACTOR 过程步，运用主成分分析法计算因子载荷，提取公因子，并通过方差最大正交旋转，增加因子间的独立性，最后计算出各因子得分。

最终，我们共提取了 28 个因子，保留了原始特征集合上 80%的信息量，并将其命名为 Factor1-Factor28，替换原先的定量变量，构造新的特征集合。(因子载荷矩阵请见附录 5)

5.2.2 逻辑回归

我们先用 SAS 9.4 中 REG 过程步检查解释变量与响应变量间的共线性关系，结果显示并不存在显著的多重共线关系，这说明我们前期的变量筛选是有效的。于是我们将 16 个哑变量和 28 个特征因子作为模型的输入变量，再用 LOGISTIC 过程步建立模型，经过抽样训练样本集中的正类和反类被平衡为 2:1，并将显著性水平设为 0.05，运用逐步回归(stepwise 方法)对变量进行选择，最终进入模型的变量见下表：

表格 14. 通过检验的指标 1

变量名	系数	标准化估计	变量名	系数	标准化估计
Intercept	-1.7548		Factor4	0.0733	0.0537
IS_LOCAL	0.2732	0.0735	Factor5	-0.1688	-0.0971
SEX	0.3881	0.0959	Factor6	0.0502	0.0268
SALARY1	1.4523	0.061	Factor7	-0.1366	-0.0718
SALARY2	1.137	0.1743	Factor8	0.118	0.0773
SALARY3	0.9931	0.1863	Factor11	-0.1398	-0.0728
SALARY4	0.325	0.0458	Factor14	0.0511	0.028
EDU_LEVEL1	0.293	0.0332	Factor15	-0.6366	-0.3297
EDU_LEVEL2	0.3244	0.0719	Factor20	0.0622	0.0309
Factor1	-0.3409	-0.1766	Factor21	0.3203	0.1823
Factor2	-0.0547	-0.0299	Factor24	-0.1248	-0.0686
Factor3	-0.2129	-0.1032	Factor28	0.2422	0.1024

SAS 输出的模型拟合效果如下：

表格 15. 预测概率和观测响应的关联 1

一致对占比	73.9	Somers D	0.481
不一致对占比	25.8	Gamma	0.482
结值百分比	0.3	Tau-a	0.215
对	196640000	c	0.74

将训练出的模型分别在未平衡训练集和测试集上进行相应预测，我们计算出的 AUC 分别为 0.7413 和 0.7214，下图为测试集上的 ROC 曲线：

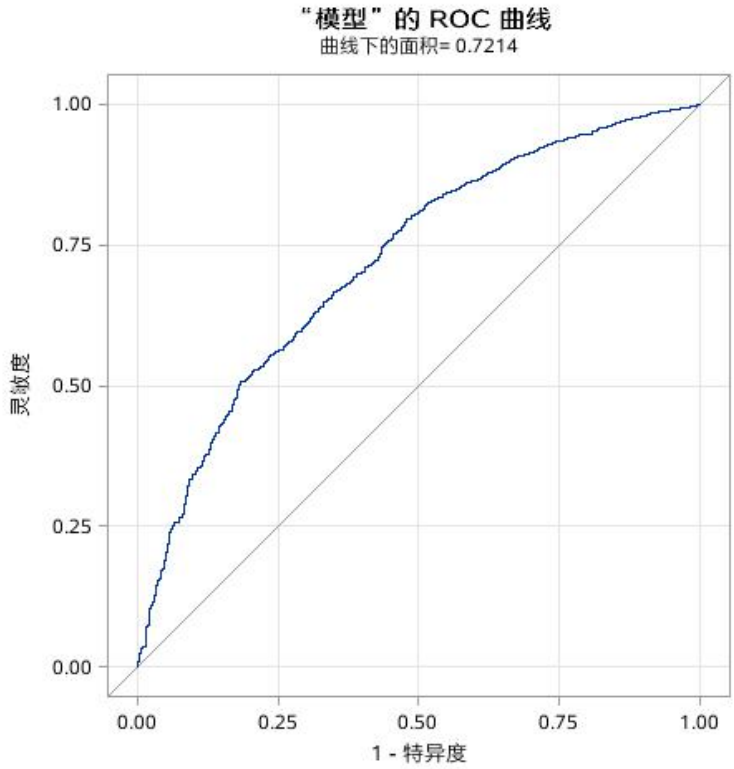

图 15. ROC 曲线 1

我们可以看出跟 XGBoost 模型在测试集上的 AUC(0.7456)相比，逻辑回归模型还是稍显逊色的。

我们最后将 0.0625 作为 cutoff，得到在训练集和测试集上相应的混淆矩阵、查全率($\frac{TP}{TP+FN}$)以及整体错分率如下：

表格 16. 混淆矩阵 1

训练集	预测＼实际	0	1	测试集	预测＼实际	0	1
	0	13264	414		0	5780	210
	1	6400	906		1	2681	345

表格 17. 逻辑回归分类精度

指标	值
Accurary0_train	67.45%
Accurary1_train	68.64%
Accurary_all_train	67.53%
Accurary0_test	68.31%
Accurary1_test	62.16%
Accurary_all_test	67.93%

根据上表我们看到 Logistic 模型在测试集上对违约用户的识别没有在训练集上好。通过比较也不难发现，XGBoost 预测精度要整体优于逻辑回归，这也再次验证了 XGBoost 作为时下热门的机器学习算法，确实可以实现比较高的分类精度，在这点上要优于传统的统计学模型。

5.3 XGBoost-logistic 串行组合模型

我们用 SAS 9.4 中 LOGISTIC 过程步在平衡后的训练集上建立模型，模型的输入为 XGBoost 预测的样本标签(LABEL_XG)、16 个哑变量和 28 个特征因子，并通过逐步回归对变量进行筛选，最终进入模型的变量见下表：

表格 18. 通过检验的指标 2

变量名	系数	标准化估计	变量名	系数	标准化估计
Intercept	-2.1061		Factor2	0.0426	0.0233
LABEL_XG	1.8933	0.5217	Factor4	0.0355	0.026
IS_LOCAL	0.1612	0.0434	Factor8	0.0487	0.0319
SEX	0.1655	0.0409	Factor11	-0.0472	-0.0246
SALARY1	1.1508	0.0483	Factor15	-0.1823	-0.0944
SALARY2	0.4956	0.076	Factor21	0.0888	0.0506
SALARY3	0.3282	0.0616	Factor24	-0.0567	-0.0312
SALARY6	-0.4248	-0.0276	Factor28	0.1221	0.0516
EDU_LEVEL5	-0.1481	-0.0326			

SAS 输出的模型拟合效果如下：

表格 19. 预测概率和观测响应的关联 2

一致对占比	77.6	Somers D	0.556
不一致对占比	21.9	Gamma	0.559
结值百分比	0.5	Tau-a	0.249
对	196640000	c	0.778

将训练出的模型分别在未平衡训练集和测试集上进行相应预测，我们计算出的 AUC 分别为 0.7773 和 0.7339，不难发现串行后的 Logistic 模型无论是在训练

集还是测试集，AUC 都得到了明显提高，下图为测试集上的 ROC 曲线：

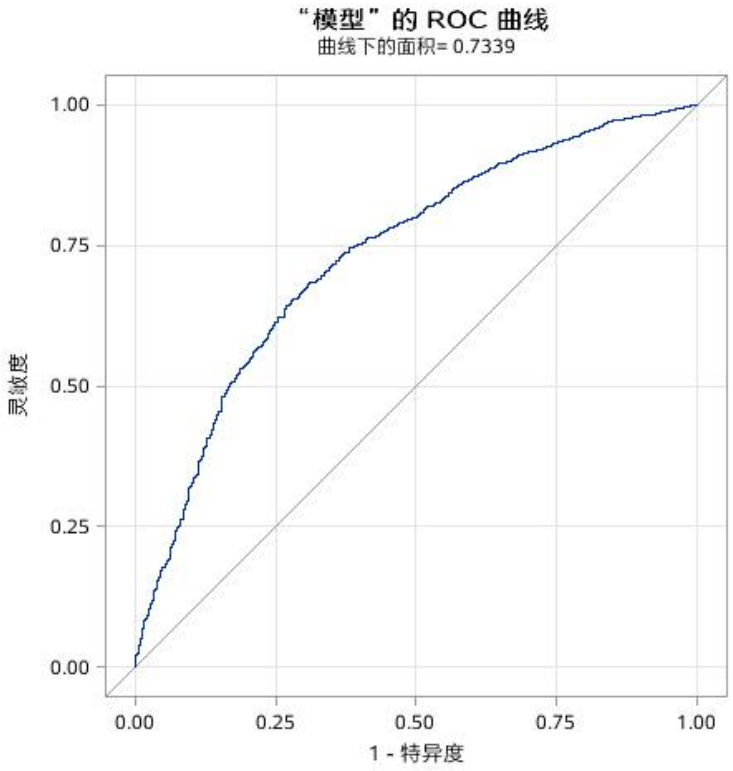

图 16. ROC 曲线 2

我们可以看出串行后的模型，测试集上 AUC 的表现要介于 XGBoost 和逻辑回归之间。

我们最后将 0.0625 作为 cutoff，得到在训练集和测试集上相应的混淆矩阵、查全率($\frac{TP}{TP+FN}$)以及整体错分率如下：

表格 20. 混淆矩阵 2

训练集	预测＼实际	0	1	测试集	预测＼实际	0	1
	0	13357	251		0	5737	170
	1	6307	1069		1	2724	385

表格 21. XGBoost-logistic 分类精度

指标	值
Accurary0_train	67.93%
Accurary1_train	80.98%
Accurary_all_train	68.75%
Accurary0_test	67.81%
Accurary1_test	69.37%

Accurary_all_test	67.90%

根据上表我们可以看到， 无论是在训练集还是测试集，XGBoost-logistic 的分类精度都非常接近单个 XGBoost 的分类精度，甚至在测试集上，组合模型的两类查全率都超过了单个 XGBoost 的查全率，充分继承了机器学习模型高分类精度的特点。跟逻辑回归模型相比，违约用户的查全率在测试集上提升了 7%，这对贷款机构降低风险会有很好的帮助，模型同时也继承了对变量的解释性。

总得来说，XGBoost-logistic 串行模型的确结合了机器学习较高的分类精度与逻辑回归解释性的优点，表现是很不错的。

5.4 个人信用评分体系

我们分别计算出违约与不违约这两类人群预测概率 p 的分位数如下表：

表格 22. 预测概率分位数

百分位数	不违约用户	违约用户
25	0.0152727	0.1182180
50	0.0179098	0.1365289
75	0.0208358	0.1633237

由此我们得到的信用评分体系如下：

表格 23. 个人信用评分体系

得分区间	信用等级	违约率
790~800	优秀	1.86%
769~789	良好	3.05%
741~768	一般	10.30%
300~740	较差	15.06%

6. 模型结果分析

6.1 模型分析

6.1.1 XGBoost-logistic 模型

由 5.3 我们得到了模型显著的指标以及相应的回归系数，据此我们分析各指标对违约的影响如下：

LABEL_XG 的回归系数为 1.8933，说明被 XGBoost 模型预测为违约的用户，在 XGBoost-logistic 中被判为违约的概率是不违约概率的 6.6 倍。

SEX 的回归系数为 0.1655，这说明相比女性而言，男性发生贷款违约的概率是不发生的 1.17 倍，所以男性比女性的违约可能性要大一些。

IS_LOCAL 的系数为 0.1612，说明本地借款人相比外地借款人，发生违约的倾向为 1.17 倍，本地略高于外地。

SALARY 只有状态 1、2、3 和 6 显著，在前 3 种低收入状态中，最大的系数为 1.1508，这说明收入状态为 1 的用户相比其他用户，他们发生违约的可能性是不发生的 3.16 倍。对于收入状态为 6 的用户来说，他们发生违约的可能性是不发生的 0.65 倍。总体上看，随着收入水平的提高，系数递减，表明发生违约的可能性逐渐减少，这从直观意义上也很容易理解。从收入水平为 4 开始，更高的收入状态多数不再对预测违约有显著的影响，这说明对中高收入人群，收入不再是影响是否违约的主要因素。

在分析各因子对违约的影响前，我们先通过因子载荷矩阵，挑选出载荷大的指标，并根据这些指标赋予显著的因子以经济意义如下：

表格 24. 特征因子

因子名	载荷指标	经济意义
Factor2	D_OVD_MONTH, D_OVD_COUNT, D_OVD_MONTH_CV	贷记卡逾期频度与波动因子
Factor4	L_YQ_PERCENT, L_OVD_AMOUNT_AVG	未结清贷款逾期因子
Factor8	D_YQ_PERCENT, D_OVD_MONTH_MAX	未销户贷记卡逾期因子
Factor11	L_ZFZ_PERCENT, L_BALANCE	贷款结构与负债因子
Factor15	D_PASS_PERCENT, LD_PASS_PERCENT	借贷通过因子
Factor21	QUERY_RECENT	查询风险因子
Factor24	MARRY_WOE	婚姻状态因子
Factor28	LOAN_TIME	流程风险因子

贷记卡逾期频度与波动因子的系数为0.0426，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的1.04倍。具体来说，贷记卡历史违约期数和账户数越多的用户，该笔发放贷款往往有越高的违约风险；用户所持所有未销贷记卡违约月份数的波动越大，该用户的违约风险也较高。

未结清贷款逾期因子的系数为0.0355，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的1.04倍。具体来说，用户当前有逾期记录的贷款笔数在所有未结清贷款中占比越高，该用户往往具有较高的违约风险；用户当前未结清贷款的平均逾期期数越高，该用户的违约风险也较高。

未销户贷记卡逾期因子的系数为0.0487，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的1.05倍。具体来说，用户当前有逾期记录的贷记卡账户在所有未销户贷记卡中占比越高，该用户往往具有较高的违约风险；未销户贷记卡中违约期数越大的用户，该用户的违约风险也较高。

贷款结构与负债因子的系数为-0.0472，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的0.95倍。具体来说，用户目前个人住房和商用房的月应还款额在总的月应还款中占比越高，该用户往往具有较低的违约风险；用户未结清贷款总余额越高，往往也具有较低的风险。

借贷通过因子的系数为-0.1823，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的0.83倍。这表明用户申请的贷记卡账户通过率或者信用卡(贷记卡和准贷记卡)账户通过率越高，他们越倾向与有较低的违约可能。

查询风险因子的系数为0.0888，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的1.09倍。这表明用户的征信报告在最近一个月内被查询的次数越多，他们越倾向与有较高的违约可能。

婚姻状态因子的系数为-0.0567，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的0.945倍。

流程风险因子的系数为0.1221，每当该因子得分增加1个单位，相比现在，用户发生违约的可能性将变为不发生的1.13倍。这表明从报告生成时间到放款时间这段时间越长，出于信息滞后的原因，客观上使得用户增加了违约的可能性。

此外，根据模型的标准化估计，除了LABEL_XG外，对识别违约用户最重要的前四个因子分别为：借贷通过因子、SALARY3、流程风险因子和查询风险因子。

在基本信息中，我们可以看到SALARY2和SALARY3对用户违约情况影响最大，其次影响较大的指标为SALARY1、IS_LOCAL和SEX，它们对违约的影响程度

相近，而影响程度最低的指标是 SALARY6.

而在 8 项因子中，对违约判断最重要的是借贷通过因子，其次是查询风险因子和流程风险因子，它们对违约的影响程度相近，影响较弱的因子则是贷记卡逾期频度与波动因子、未结清贷款逾期因子和贷款结构与负债因子。

考虑到LABEL_XG作为XGBoost模型的输出，其本身也带有比较高的信息量，故我们提取该模型中重要度排在前 50%的指标，结果如下：

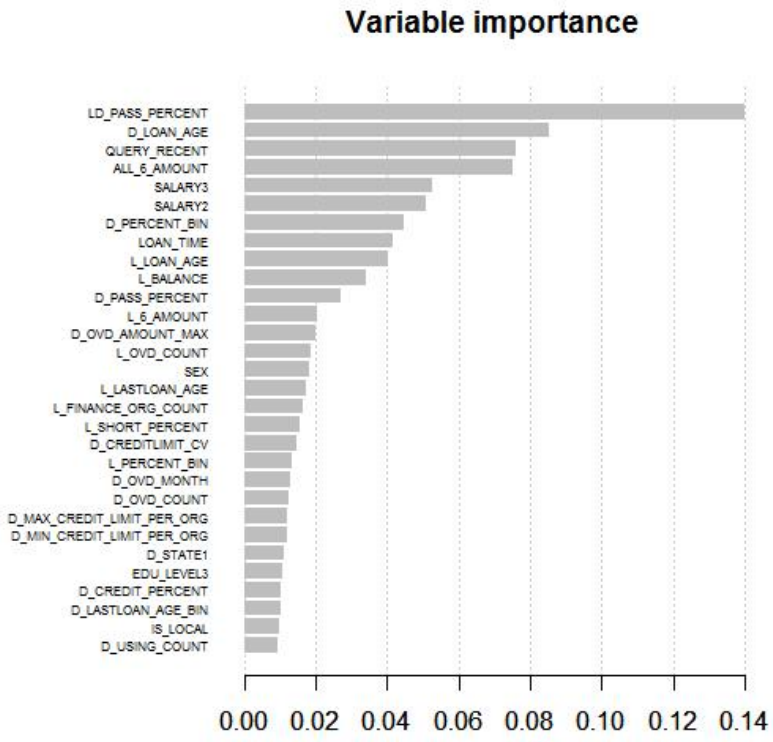

图 17. XGBoost-重要性度量条形图

根据上图，可以看出XGBoost模型认为最重要的指标为LD_PASS_PERCENT，这在借贷通过因子中也同样有所体现。其次比较重要的有 D_LOAN_AGE、QUERY_RECENT 和 ALL_6_AMOUNT,其中除了 QUERY_RECENT 在查询风险因子中有所体现外，反映账龄信息的 D_LOAN_AGE 和反映负债水平的 ALL_6_AMOUNT 都没有在其余因子中体现，但我们在进行经济意义解释时，却有必要将它们纳入考虑范围内。

故我们将 XGBoost 模型识别出的这 30 项重要指标作为结论和建议部分的额外参考信息。

6.1.2 个人信用评分体系

由 5.4 我们可以看到信用优秀的用户中，实际违约率为 1.86%，信用良好的用户中，实际违约率为 3.05%，信用一般的用户违约率为 10.30%，而信用较差的用户违约率为 15.06%。

我们将信用水平处于良好及以上的用户群体，违约率控制在了 1.8%-3.1%之间，并且总体上呈现用户的信用等级越低，违约率越高的特点，这说明我们的信用评分体系是合理有效的，可以为贷款机构提供一定程度的借鉴意义。

6.2 模型预测

我们先用 5.1.6 设置的参数在样本容量为 10000 的无标签测试集上训练 XGBoost 模型，并对测试集预测标签。然后将 XGBoost 预测出来的标签连同 5.3 中通过显著性检验的哑变量和因子作为逻辑回归的输入指标，并结合 5.3 中相应的回归系数在测试集上预测标签，此处分类阈值仍取 0.0625。

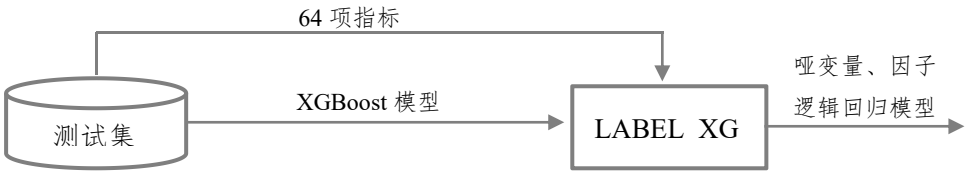

图 18. 模型预测流程图

XGBoost-logistic 的结果显示，在一万条观测中有 3895 位用户被预测为违约用户，剩余 6105 位被预测为不违约用户。REPORT_ID 对应的信用评分、信用等级和预测标签请见附件 1。

值得指出的是，模型预测的违约率要远高于实际的 6.25%，显然这些“违约”用户中实际将发生违约的只是其中一小部分，但我们的模型却能够比较好得将他们识别出来，为贷款机构有效地降低了放款风险。

结论和建议

我们主要从信用习惯、偿付能力、负债水平角度出发，分如下七方面来做出是否授信决策：个人特质、历史授信记录、历史偿还记录、偿付能力、目前负债水平、目前负债结构、风险。

在个人特质上，我们发现教育程度对违约可能性有显著的影响。随着学历的升高，违约可能性逐渐降低。但是从本科学历开始，教育水平不再是影响违约可能性的因素。所以对较低学历水平用户，应考虑学历水平限制对违约的影响，而对本科以上用户在放款时可考虑其他因素的影响。

同样在个人特质上，我们发现性别对于违约可能性有一定程度的影响，男性比女性的违约可能性要大一些，相比女性而言，男性发生贷款违约的概率是不发生的 1.17 倍，建议贷款机构可以考虑提高对女性放款的比例；就地域而言，本地借款人发生违约概率略高于外地，为 1.17 倍，建议贷款机构可以考虑提高对外地放款的比例。

从偿付能力来看，包括三部分：流量、存量、潜量。**从流量上说，**收入越高意味着资金流动能力越高，因此对于利息或分期应付款等短期负债而言，还款能力越强。我们发现薪酬水平对违约可能性有显著的影响，对于状态为 1,2,3 的低收入群体而言，收入的高低会影响违约可能性，而对于中高收入群体，收入不再是影响违约可能性大小的因素。因此，我们建议贷款用户对收入水平为 1,2,3 的用户要谨慎放款，而对中高收入水平的用户在放款时需要参考其他更多的因素。**从存量上说，**有更高的净资产或者更多的担保，意味着对于本金等数额大的长期负债，偿付能力越强，会提高履约的可能性，我们从贷记卡的信用担保比率可以看出，有担保的用户要比无担保的用户，履约可能性更高。因此，我们建议对于信用贷款申请要审慎放款。**从潜量上说，**如果存在抚养责任等压力，偿付能力将会变低，从婚姻状态因子来看，单亲家庭的家长有更低的偿付能力，离异或丧偶状态会加大违约可能性。因此我们建议对于“离异”和“丧偶”状态的申请者谨慎放款。

从历史授信记录来看，愿意授信的机构数越多，审批通过率越高，且额度越高，意味着该申请人经评估后历史信用程度越高，相应地我们可以推断该申请人目前有更高的资信可以进行授信。另一方面，授信的机构数与账户数越多，卡龄越久，表明用户更有可能在使用中培养了良好的信用习惯。需要指出的是，卡龄是历史授信记录中相当重要的判断依据，使用信用的年限越长，该授信机构对于用户的使用习惯等信息掌握越全，风险会随之下降；同时卡龄越长的用户也更有可能培养了良好的信用习惯。因此，建议放款时可以偏向账户数多，借贷通过率

高，授信额度大，使用信用年限长的用户。

从历史偿还记录来看，历史的偿还与否，逾期与否，不仅可以显示着用户的历史偿还能力，而且可以显示用户的信用习惯好坏。如果历史上出现止付、冻结和呆账比重较大，其目前违约风险就会加大。我们这里建议贷款机构对于历史上出现止付、冻结和呆账记录的申请者放贷请求一概不予通过。此外，对历史逾期期数和账户数较多的用户，他们的信用习惯也不是特别好，贷款机构在放款时也应该谨慎考虑。

从目前的负债水平来看，当前逾期时间长，逾期金额较高且占总负债比重较高的用户，更有可能在目前面临资金紧缺、还款能力不足的问题。相应地从未结清贷款逾期因子和未销户贷记卡逾期因子可以看出，如未销户贷记卡逾期因子每增加 1 单位，发生违约的可能性就将变为不发生的 1.05 倍。因此，我们建议贷款机构关注用户的近期逾期状况，包括逾期时间与逾期金额绝对数、相对数，如果逾期情况严重，则谨慎放贷。

从目前的负债结构来看，个人住房和商用房属于贷款结构中的刚需部分，其月应还款额在总的月应还款中占比越高，该类用户为满足生存需要通常越会积极还款，对应的违约风险也就越低。此外，房产作为一种价值稳定的固定资产，可以降低用户违约发生时对贷款机构造成的损失。因此，我们建议对于目前个人住房和商用房月应还款占比较高的申请者可以提高放贷额度。此外，现有贷款期限的长短意味着偿债压力的大小。短期利率低，但需要流动资金偿还；长期利率高，但可以通过长期资产逐渐变现，用持续一段时间的现金流偿还。贷款期限在 1 到 5 年的中期贷款的偿还压力最大，因为在利率偏高、流动资金量要求较高的情况下，申请者越难履约。因此，我们建议对于目前中期贷款占比较高的申请者谨慎放贷。

从风险角度来看，在最近一个月内，对于征信报告被多次查询的用户来说，这类人群很可能出现了比较大的短期资金缺口，其现金流的不稳定以及短期内的过剩需求增加了他们无法偿付的风险。查询风险因子每增加 1 单位，相比现在，发生违约可能性将变为不发生的 1.09 倍，故我们建议放款机构对此需要谨慎放款。此外，由于目前报告生成时间与放款时间之间存在较长时间间隔，时滞性会导致信息的滞后，这种信息的不确定性导致准确判断用户违约可能性的难度上升，放款风险加大。流程风险因子每增加 1 单位，相比现在，发生违约可能性将变为不发生的 1.13 倍。因此，我们建议放款机构尽早做出是否放贷决策，并且尽可能考虑报告生成时间较接近当前的用户。

参考文献

[1] 向晖. 个人信用评分组合模型研究与应用[D]. 湖南大学, 2011.

[2] 曾辉. 基于数据挖掘的银行个人客户信用评分模型的研究[D]. 对外经济贸易大学, 2007.

[3] 薛薇. R 语言数据挖掘[M]. 中国人民大学出版社, 2016.

[4] 周志华.基于分歧的半监督学习[J].自动化学报, 2013, 39(11):1871-1878.

[5] 周志华.半监督学习中的协同训练算法[M]//周志华，王钰.机器学习及其应用.北京：清华大学出版社，2007:259-275

[6] Goldman S A, Zhou Y. Enhancing Supervised Learning with Unlabeled Data[C]// Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2000:327-334.

[7] Zhou Z H, Chen K J, Jiang Y. Exploiting Unlabeled Data in Content-Based Image Retrieval[J]. Lecture Notes in Computer Science, 2004, 3201:525--536.

[8] Zhou Z H. Enhancing relevance feedback in image retrieval using unlabeled data[J]. Acm Transactions on Information Systems, 2006, 24(2):219-244.

附录

附录 1. 指标计算方式：

个人信息汇总类

指标名称	计算方式	来源
SEX	女=0，男=1	基础表
EDU_LEVEL	初中=1，高中=2，专科=3，专科及以下=4，本科=5，硕士研究生=6，硕士及以上=7，博士研究生=8，其他=9，缺失为 10	基础表
MARRY_STATUS	未婚=1，已婚=2，离婚=3，离异=4，丧偶=5，其他=6，缺失为 7	基础表
IS_LOCAL	非本地籍=0，本地籍=1	基础表
AGENT	未公开收入水平的为 0，公开收入水平的为 1	基础表
HAS_FUND	无=0，有=1	基础表
SALARY	1=1,2=2,3=3,4=4,5=5,6=6,7=7，缺失为 8	基础表
LOAN_AGE	*Max*{贷款账龄，贷记卡账龄，准贷记卡账龄}	报告主表、信用提示
ALL_RECENTLOAN_COUNT	贷款新开立账户数+贷记卡新开立账户数	报告主表
LD_PASS_PERCENT	(贷款总笔数+贷记卡总卡数)/总查询次数	贷款明细、贷记卡明细、信贷审批查询记录明细
PREFERENCE1	贷款总笔数/(贷款总笔数+贷记卡总账户数)	贷款明细、贷记卡明细
PREFERENCE2	贷记卡总账户数/(贷款总笔数+贷记卡总账户数)	贷款明细、贷记卡明细
QUERY_RECENT	报告生成时间 1 个月内的查询次数	报告主表、信贷审批查询记录明细
LOAN_TIME	放款时间−报告生成时间	基础表、报告主表
ALL_6_AMOUNT	最近 6 个月的平均应还款额+贷记卡平均使用额度+准贷记卡平均使用额度	未销贷记卡或未结清贷款
ALL_DS_CREDIT_LIMIT	贷记卡合同金额+准贷记卡合同金额	未销贷记卡或未结清贷款
MAX_DS_CREDIT_LIMIT	贷记卡平均单个机构最大合同金额+准贷记卡平均单个机构最大合同金额	未销贷记卡或未结清贷款
MIN_DS_CREDIT_LIMIT	贷记卡平均单个机构最小合同金额+准贷记卡平均单个机构最小合同金额	未销贷记卡或未结清贷款
ALL_OVD_COUNT	贷款总逾期笔数+贷记卡总逾期账户数+准贷记卡总逾期账户数	逾期(透支)信息汇总

贷款信息类

指标名称	计算方式	来源
L_LOAN_AGE	报告生成时间—首笔贷款发放月份	报告主表、信用提示
L_LASTLOAN_AGE	报告生成时间—最后一笔贷款发放时间	报告主表、贷款明细
L_RECENTLOAN_COUNT	报告生成 1 个月内的贷款笔数	报告主表、贷款明细
L_PASS_PERCENT	贷款总笔数/(贷款审批记录数+担保审批记录数)	贷款明细、信贷审批查询记录明细
L_FINANCE_ORG_COUNT	未结清贷款机构数	未销贷记卡或未结清贷款
L_USING_COUNT	未结清贷款数	未销贷记卡或未结清贷款
L_BALANCE	贷款余额	未销贷记卡或未结清贷款
L_SCH_PAY_AMOUNT	每笔未结清贷款的本月应还款总额	贷款明细
L_6_AMOUNT	最近 6 个月的平均应还款额(平均使用额度)	未销贷记卡或未结清贷款
L_PERCENT	贷款余额/合同金额	未销贷记卡或未结清贷款
L_CREDIT_LIMIT	合同金额	未销贷记卡或未结清贷款
L_OVD_COUNT	逾期笔数	逾期(透支)信息汇总
L_OVD_MONTH	逾期月份数	逾期(透支)信息汇总
L_OVD_MONTH_AVG	未结清贷款当前逾期期数的平均值	贷款明细
L_OVD_MONTH_MAX	*Max*{每笔未结清贷款当前逾期期数}	贷款明细
L_OVD_AMOUNT_AVG	未结清贷款当前逾期金额的平均值	贷款明细
L_OVD_AMOUNT_MAX	*Max*{每笔未结清贷款当前逾期金额}	贷款明细
L_OVD_MONTH_STD	根据 24 个月还款状态计算出历史逾期期数，并求出所有未结清贷款逾期期数的标准差	贷款明细
L_OVD_MONTH_CV	根据 24 个月还款状态计算出历史逾期期数，并求出所有未结清贷款逾期期数的变异系数	贷款明细
L_HIGHEST_OA_PER_MON	单月最高逾期总额	逾期(透支)信息汇总
L_DURATION_MAX	最大贷款时长	逾期(透支)信息汇总

L_YJ1_COUNT	担保人还款笔数+展期还款笔数	贷款特殊交易
L_YJ2_COUNT	提前还款笔数	贷款特殊交易
L_CREDITLIMIT_STD	全部贷款合同金额的标准差	贷款明细
L_CREDITLIMIT_CV	全部贷款合同金额的变异系数	贷款明细
L_ZC_PERCENT	正常状态贷款笔数/贷款总笔数	贷款明细
L_JQ_PERCENT	结清状态贷款笔数/贷款总笔数	贷款明细
L_QT_PERCENT	1−L_ZC_PERCENT−L_JQ_PERCENT	贷款明细
L_ZFZ_PERCENT	(个人住房月应还款+商用房月应还款)/总应还款	贷款明细
L_LONG_PERCENT	贷款时间在 5 年以上的贷款数/贷款总笔数	贷款明细
L_MED_PERCENT	贷款时间在 1 至 5 年的贷款数/贷款总笔数	贷款明细
L_SHORT_PERCENT	贷款时间在 1 年以下的贷款数/贷款总笔数	贷款明细
L_CREDIT_PERCENT	信用方式担保贷款笔数/贷款总笔数	贷款明细
L_NCREDIT_PERCENT	1−L_CREDIT_PERCENT	贷款明细
HOUSE_LOAN_COUNT	个人住房贷款笔数	信用提示
COMMERCIAL_LOAN_COUNT	商用房贷款笔数	信用提示
OTHER_LOAN_COUNT	其他贷款笔数	信用提示
ALL_LOAN_COUNT	个人住房贷款笔数+商用房贷款笔数+其他贷款笔数	信用提示
L_YQ_PERCENT	当前有逾期情况的贷款数/未结清贷款总数	贷款明细、未销贷记卡或未结清贷款
L_Z_PERCENT	转出贷款数/(转出贷款数+结清贷款数)	贷款明细
L_LOSTBALANCE_MAX	所有未结清贷款按5级分类的标准计算最大损失金额在本金余额的占比	贷款明细

贷记卡信息类

指标名称	计算方式	来源
D_LOAN_AGE	报告生成时间−首笔贷记卡发放月份	报告主表、信用提示
D_LASTLOAN_AGE	报告生成时间−最后一张贷记卡账户生成时间	报告主表、贷记卡明细
LOANCARD_COUNT	贷记卡账户数	贷记卡明细
D_RECENTLOAN	报告生成 1 个月内的贷记卡账户数	报告主表、贷记卡

_COUNT		明细
D_PASS_PERCENT	贷记卡总数/贷记卡审批记录数	贷记卡明细、信贷审批查询记录明细
D_FINANCE_ORG_COUNT	未销贷记卡机构数	未销贷记卡或未结清贷款
D_USING_COUNT	未销贷记卡账户数	未销贷记卡或未结清贷款
D_USED_CREDIT_LIMIT	未销贷记卡的已用额度	未销贷记卡或未结清贷款
D_SCH_PAY_AMOUNT	未销户贷记卡的本月应还款总额和	贷记卡明细
D_6_AMOUNT	最近 6 个月的平均使用额度	未销贷记卡或未结清贷款
D_PERCENT	贷记卡已用额度/合同金额	未销贷记卡或未结清贷款
D_CREDIT_LIMIT	合同金额	未销贷记卡或未结清贷款
D_MAX_CREDIT_LIMIT_PER_ORG	平均单个机构最大合同金额	未销贷记卡或未结清贷款
D_MIN_CREDIT_LIMIT_PER_ORG	平均单个机构最小合同金额	未销贷记卡或未结清贷款
D_OVD_COUNT	逾期笔数	逾期(透支)信息汇总
D_OVD_MONTH	逾期月份数	逾期(透支)信息汇总
D_OVD_MONTH_AVG	未销户贷记卡当前逾期期数的平均值	贷记卡明细
D_OVD_MONTH_MAX	*Max*{每笔未销户贷记卡当前款逾期期数}	贷记卡明细
D_OVD_AMOUNT_AVG	未销户贷记卡当前逾期金额的平均值	贷记卡明细
D_OVD_AMOUNT_MAX	*Max*{每笔未销户贷记卡当前款逾期金额}	贷记卡明细
D_OVD_MONTH_STD	根据 24 个月还款状态计算出历史逾期期数，并求出所有未销户贷记卡逾期期数的标准差	贷记卡明细
D_OVD_MONTH_CV	根据 24 个月还款状态计算出历史逾期期数，并求出所有未销户贷记卡逾期期数的变异系数	贷记卡明细
D_HIGHEST_OA_PER_MON	单月最高逾期总额	逾期(透支)信息汇总
D_MAX_DURATION	最大贷款时长	逾期(透支)信息汇总
D_CREDITLIMIT_STD	所有贷记卡账户信用额度(贷款金额)的标准差	贷记卡明细

D_CREDITLIMIT_CV	所有贷记卡账户信用额度(贷款金额)的变异系数	贷记卡明细
D_STATE1	正常账户数/总账户数	贷记卡明细
D_STATE2	(销户＋未激活账户数)/总账户数	贷记卡明细
D_STATE3	1－D_STATE1－D_STATE2	贷记卡明细
D_CREDIT_PERCENT	信用担保账户数占比	贷记卡明细
D_NCREDIT_PERCENT	1－D_CREDIT_PERCENT	贷记卡明细
D_YQ_PERCENT	当前有逾期的贷记卡账户数/未销户贷记卡账户总数	贷记卡明细、未销贷记卡或未结清贷款
D_ZC_SCH_PERCENT	正常状态的本月应还款和/未销户状态的本月应还款和	贷记卡明细

准贷记卡信息类

指标名称	计算方式	来源
S_LOAN_AGE	报告生成时间－首笔准贷记卡发放月份	报告主表、信用提示
S_FINANCE_ORG_COUNT	未销准贷记卡机构数	未销贷记卡或未结清贷款
S_USING_COUNT	未销准贷记卡账户数	未销贷记卡或未结清贷款
S_USED_CREDIT_LIMIT	未销准贷记卡已用额度	未销贷记卡或未结清贷款
S_6_AMOUNT	最近 6 个月的平均使用额度	未销贷记卡或未结清贷款
S_PERCENT	准贷记卡已用额度/合同金额	未销贷记卡或未结清贷款
S_CREDIT_LIMIT	合同金额	未销贷记卡或未结清贷款
S_MAX_CREDIT_LIMIT_PER_ORG	平均单个机构最大合同金额	未销贷记卡或未结清贷款
S_MIN_CREDIT_LIMIT_PER_ORG	平均单个机构最小合同金额	未销贷记卡或未结清贷款
S_OVD_COUNT	逾期笔数	逾期(透支)信息汇总
S_OVD_MONTH	逾期月份数	逾期(透支)信息汇总
S_HIGHEST_OA_PER_MON	单月最高逾期总额	逾期(透支)信息汇总
S_DURATION_MAX	最大贷款时长	逾期(透支)信息汇总

附录 2. 变量聚类结果

Cluster	Variable	RSquareRatio
Cluster 1	ALL_6_AMOUNT	0.1011
	ALL_DS_CREDIT_LIMIT	0.2674
	D_USED_CREDIT_LIMIT	0.1176
	D_SCH_PAY_AMOUNT	0.3807
	D_6_AMOUNT	0.1032
	D_CREDIT_LIMIT	0.2701
Cluster 2	PREFERENCE1	0.5305
	L_FINANCE_ORG_COUNT	0.3966
	L_USING_COUNT	0.3978
	L_CREDITLIMIT_CV	0.5591
Cluster 3	S_6_AMOUNT	0.3704
	S_CREDIT_LIMIT	0.1606
	S_MAX_CREDIT_LIMIT_PER_ORG	0.1145
	S_MIN_CREDIT_LIMIT_PER_ORG	0.1301
Cluster 4	LOAN_AGE	0.11
	D_LOAN_AGE	0.0925
Cluster 5	L_BALANCE	0.0872
	L_CREDIT_LIMIT	0.0917
	L_CREDITLIMIT_STD	0.4638
Cluster 6	L_OVD_MONTH_MAX	0.5249
	L_OVD_AMOUNT_MAX	0.3594
	L_OVD_AMOUNT_AVG	0.355
Cluster 7	LD_PASS_PERCENT	0.1673
	L_PASS_PERCENT	0.206
Cluster 8	D_OVD_MONTH	0.2484
	D_OVD_MONTH_STD	0.469
	D_MAX_DURATION	0.4682
Cluster 9	D_OVD_AMOUNT_AVG	0.0153
	D_OVD_AMOUNT_MAX	0.0148
Cluster 10	S_LOAN_AGE	0.5489
	S_FINANCE_ORG_COUNT	0.1146
	S_USING_COUNT	0.1854
Cluster 11	L_OVD_MONTH_AVG	0.0198
	L_YQ_PERCENT	0.018
Cluster 12	LOANCARD_COUNT	0.1564
	D_FINANCE_ORG_COUNT	0.3138
	D_USING_COUNT	0.0968
Cluster 13	ALL_OVD_COUNT	0.2294
	D_OVD_COUNT	0.2097
	D_HIGHEST_OA_PER_MON	0.6658

Cluster 14	MIN_DS_CREDIT_LIMIT	0.0498
	D_MIN_CREDIT_LIMIT_PER_ORG	0.0448
Cluster 15	D_OVD_MONTH_AVG	0.18
	D_OVD_MONTH_MAX	0.1643
Cluster 16	ALL_RECENTLOAN_COUNT	0.1273
	L_RECENTLOAN_COUNT	0.1006
Cluster 17	L_SHORT_PERCENT	0.2124
	L_CREDIT_PERCENT	0.2225
Cluster 18	AGENT1	0.5423
	SALARY4	0.3029
Cluster 19	L_OVD_MONTH	0.1841
	L_DURATION_MAX	0.188
Cluster 20	D_PERCENT_BIN	0.3555
	D_ZC_SCH_PERCENT	0.3686
Cluster 21	L_ZFZ_PERCENT	0.2932
	L_LONG_PERCENT	0.3064
	HOUSE_LOAN_COUNT	0.3427
Cluster 22	D_STATE1	0.0332
	D_STATE2	0.0337
Cluster 23	EDU_LEVEL4	0.4508
	SALARY3	0.3669
Cluster 24	L_OVD_MONTH_STD	0
Cluster 25	SALARY2	0
Cluster 26	EDU_LEVEL2	0
Cluster 27	L_PERCENT_BIN	0.3261
	L_ZC_PERCENT	0.4406
Cluster 28	EDU_LEVEL5	0
Cluster 29	SALARY6	0
Cluster 30	EDU_LEVEL1	0
Cluster 31	IS_LOCAL	0
Cluster 32	EDU_LEVEL8	0
Cluster 33	L_YJ1_COUNT	0
Cluster 34	L_Z_PERCENT	0
Cluster 35	SEX	0
Cluster 36	LOAN_TIME	0
Cluster 37	L_MED_PERCENT	0
Cluster 38	D_RECENTLOAN_COUNT	0
Cluster 39	SALARY1	0
Cluster 40	COMMERCIAL_LOAN_COUNT	0
Cluster 41	MARRY_WOE	0
Cluster 42	D_LASTLOAN_AGE_BIN	0
Cluster 43	SALARY7	0
Cluster 44	QUERY_RECENT	0

Cluster 45	EDU_LEVEL7	0
Cluster 46	D_YQ_PERCENT	0
Cluster 47	EDU_LEVEL6	0
Cluster 48	L_HIGHEST_OA_PER_MON	0
Cluster 49	MAX_DS_CREDIT_LIMIT	0.0727
	D_MAX_CREDIT_LIMIT_PER_ORG	0.0596
	D_CREDITLIMIT_STD	0.1726
Cluster 50	S_OVD_COUNT	0
Cluster 51	D_CREDIT_PERCENT	0
Cluster 52	SALARY5	0
Cluster 53	L_LASTLOAN_AGE	0
Cluster 54	L_LOAN_AGE	0
Cluster 55	L_JQ_PERCENT	0
Cluster 56	L_OVD_MONTH_CV	0
Cluster 57	EDU_LEVEL3	0
Cluster 58	D_CREDITLIMIT_CV	0
Cluster 59	L_SCH_PAY_AMOUNT	0.3567
	L_6_AMOUNT	0.3155
Cluster 60	D_PASS_PERCENT	0
Cluster 61	L_OVD_COUNT	0
Cluster 62	L_LOSTBALANCE_MAX	0
Cluster 63	L_YJ2_COUNT	0.1921
	OTHER_LOAN_COUNT	0.2975
Cluster 64	D_OVD_MONTH_CV	0

附录 3. 正态性检验和秩和检验

变量名	Kolmogorov-Smirnov D 统计量 p 值	Cramer-von Mises W-Sq 统计量 p 值	Anderson-Darling A-Sq 统计量 p 值
D_YQ_PERCENT	<0.010	<0.005	<0.005
D_OVD_MONTH_MAX	<0.010	<0.005	<0.005
D_OVD_AMOUNT_MAX	<0.010	<0.005	<0.005
D_MAX_CREDIT_LIMIT_PER_ORG	<0.010	<0.005	<0.005
L_BALANCE	<0.010	<0.005	<0.005
L_OVD_MONTH_CV	<0.010	<0.005	<0.005
D_PASS_PERCENT	<0.010	<0.005	<0.005
L_FINANCE_ORG_COUNT	<0.010	<0.005	<0.005
D_STATE1	<0.010	<0.005	<0.005
ALL_6_AMOUNT	<0.010	<0.005	<0.005
QUERY_RECENT	<0.010	<0.005	<0.005
L_YJ2_COUNT	<0.010	<0.005	<0.005
L_MED_PERCENT	<0.010	<0.005	<0.005
L_OVD_COUNT	<0.010	<0.005	<0.005
L_SHORT_PERCENT	<0.010	<0.005	<0.005
D_RECENTLOAN_COUNT	<0.010	<0.005	<0.005
L_JQ_PERCENT	<0.010	<0.005	<0.005
L_HIGHEST_OA_PER_MON	<0.010	<0.005	<0.005
S_OVD_COUNT	<0.010	<0.005	<0.005
L_LOAN_AGE	<0.010	<0.005	<0.005
LD_PASS_PERCENT	<0.010	<0.005	<0.005
S_MAX_CREDIT_LIMIT_PER_ORG	<0.010	<0.005	<0.005
MARRY_WOE	<0.010	<0.005	<0.005
D_CREDITLIMIT_CV	<0.010	<0.005	<0.005
L_Z_PERCENT	<0.010	<0.005	<0.005
L_YQ_PERCENT	<0.010	<0.005	<0.005
L_YJ1_COUNT	<0.010	<0.005	<0.005
L_PERCENT_BIN	<0.010	<0.005	<0.005
D_OVD_MONTH	<0.010	<0.005	<0.005
S_FINANCE_ORG_COUNT	<0.010	<0.005	<0.005
D_CREDIT_PERCENT	<0.010	<0.005	<0.005
COMMERCIAL_LOAN_COUNT	<0.010	<0.005	<0.005
L_LASTLOAN_AGE	<0.010	<0.005	<0.005
L_OVD_AMOUNT_AVG	<0.010	<0.005	<0.005
L_OVD_MONTH_STD	<0.010	<0.005	<0.005
D_LASTLOAN_AGE_BIN	<0.010	<0.005	<0.005
D_PERCENT_BIN	<0.010	<0.005	<0.005

LOAN_TIME	<0.010	<0.005	<0.005
L_6_AMOUNT	<0.010	<0.005	<0.005
D_MIN_CREDIT_LIMIT_PER_ORG	<0.010	<0.005	<0.005
L_ZFZ_PERCENT	<0.010	<0.005	<0.005
D_USING_COUNT	<0.010	<0.005	<0.005
L_OVD_MONTH	<0.010	<0.005	<0.005
L_RECENTLOAN_COUNT	<0.010	<0.005	<0.005
D_OVD_COUNT	<0.010	<0.005	<0.005
D_LOAN_AGE	<0.010	<0.005	<0.005
D_OVD_MONTH_CV	<0.010	<0.005	<0.005
L_LOSTBALANCE_MAX	<0.010	<0.005	<0.005

信用卡总授信额度秩和检验

Wilcoxon 双样本检验	
统计量	21650130.5
近似正态分布	
Z	-17.8355
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

收入水平秩和检验

Wilcoxon 双样本检验	
统计量	3159228
近似正态分布	
Z	-13.4792
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

负债水平秩和检验

Wilcoxon 双样本检验

统计量	21730867

近似正态分布	
Z	-17.6125
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

总逾期笔数秩和检验

Wilcoxon 双样本检验	
统计量	**26862926**

近似正态分布	
Z	-3.5864
单侧 Pr < Z	**0.0002**
双侧 Pr > |Z|	0.0003
t 近似值	
单侧 Pr < Z	**0.0002**
双侧 Pr > |Z|	0.0003
Z 包括 0.5 的连续性校正。	

系统风险秩和检验

Wilcoxon 双样本检验	
统计量	30991220

近似正态分布	
Z	8.0483
单侧 Pr > Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr > Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

贷款总授信额度秩和检验

Wilcoxon 双样本检验	
统计量	22689330

近似正态分布	
Z	-15.0146
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

最近 6 个月平均应还款秩和检验

Wilcoxon 双样本检验	
统计量	23130366

近似正态分布	
Z	-13.8842
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

贷款通过率秩和检验

Wilcoxon 双样本检验	
统计量	22615844

近似正态分布	
Z	-15.1996
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
Z 包括 0.5 的连续性校正。	

短期贷款笔数占比秩和检验

Wilcoxon 双样本检验	
统计量	25774954

近似正态分布	
Z	-6.5453

单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001

Z 包括 0.5 的连续性校正。

附录 4. XGBoost 调参结果

表格 21. XGBoost 调参 1

max_depth	min_child_weight	train_auc_mean	train_auc_std	test_auc_mean	test_auc_std
3	1	0.812427	0.003162	0.746108	0.011873
3	5	0.810444	0.001922	0.744219	0.012125
3	10	0.806918	0.002911	0.751762	0.013742
3	25	0.798457	0.002927	0.744817	0.010903
3	50	0.787704	0.004706	0.743802	0.019732
5	1	0.906326	0.003576	0.746491	0.016074
5	5	0.886751	0.003214	0.748807	0.016170
5	10	0.871094	0.004662	0.751451	0.019368
5	25	0.844502	0.002298	0.746200	0.013585
5	50	0.814926	0.002636	0.747207	0.015546
7	1	0.979687	0.001415	0.740223	0.012314
7	5	0.947921	0.001332	0.740687	0.008269
7	10	0.922621	0.001672	0.748551	0.013399
7	25	0.872344	0.001542	0.743937	0.016026
7	50	0.830102	0.002844	0.747576	0.012376

表格 22. XGBoost 调参 2

gamma	train_auc_mean	train_auc_std	test_auc_mean	test_auc_std
0	0.806819	0.003185	0.746505	0.014616
0.01	0.805479	0.004075	0.749742	0.018791
0.05	0.805330	0.002434	0.746421	0.014993
0.1	0.806629	0.002351	0.749381	0.009935
0.5	0.806151	0.002961	0.745357	0.017099
1	0.805704	0.002863	0.750222	0.010381
5	0.803785	0.001524	0.746621	0.013609

表格 23. XGBoost 调参 3

subsample	colsample_bytree	train_auc_mean	train_auc_std	test_auc_mean	test_auc_std
0.5	0.5	0.795807	0.003080	0.746349	0.010841
0.5	0.6	0.799537	0.001585	0.743019	0.008723
0.5	0.7	0.799822	0.001517	0.746622	0.006620
0.5	0.8	0.800335	0.001894	0.747168	0.004271
0.5	0.9	0.801024	0.002852	0.747552	0.018901
0.5	1	0.801786	0.003778	0.743620	0.016079
0.6	0.5	0.799106	0.004238	0.748424	0.018548
0.6	0.6	0.799107	0.002357	0.746917	0.014900
0.6	0.7	0.803291	0.001776	0.744222	0.010433

0.6	0.8	0.803578	0.000746	0.747632	0.010090
0.6	0.9	0.804815	0.002436	0.749456	0.015881
0.6	1	0.804849	0.001396	0.747396	0.008767
0.7	0.5	0.802373	0.003709	0.750608	0.009759
0.7	0.6	0.802447	0.003891	0.748001	0.017985
0.7	0.7	0.803993	0.004875	0.747643	0.015688
0.7	0.8	0.805153	0.001617	0.748977	0.013507
0.7	0.9	0.805639	0.002997	0.747888	0.015123
0.7	1	0.806807	0.004218	0.744888	0.021036
0.8	0.5	0.802969	0.002724	0.747695	0.014310
0.8	0.6	0.805105	0.003742	0.747266	0.013220
0.8	0.7	0.805840	0.003890	0.745969	0.008164
0.8	0.8	0.807094	0.001663	0.747768	0.011162
0.8	0.9	0.806432	0.003332	0.749306	0.016396
0.8	1	0.806479	0.002980	0.748768	0.007422
0.9	0.5	0.803951	0.003156	0.745788	0.013179
0.9	0.6	0.804635	0.002812	0.748881	0.022507
0.9	0.7	0.806202	0.001396	0.746200	0.010693
0.9	0.8	0.806845	0.003170	0.751259	0.006942
0.9	0.9	0.807411	0.003068	0.747421	0.014641
0.9	1	0.808708	0.003478	0.748368	0.012638
1	0.5	0.803857	0.002400	0.748944	0.013574
1	0.6	0.804961	0.001958	0.749433	0.013370
1	0.7	0.806701	0.003184	0.744398	0.015162
1	0.8	0.806670	0.003412	0.747756	0.011647
1	0.9	0.808118	0.003175	0.746014	0.013726
1	1	0.807657	0.002543	0.745648	0.013433

表格 24. XGBoost 调参 4

alpha	lambda	train_auc_mean	train_auc_std	test_auc_mean	test_auc_std
0	0	0.807569	0.003489	0.748032	0.012041
0	0.001	0.807158	0.002844	0.747202	0.006677
0	0.01	0.806888	0.003008	0.745688	0.024455
0	0.1	0.807001	0.003902	0.748912	0.011874
0	0.5	0.805918	0.002214	0.745990	0.018387
0.001	0	0.807144	0.003576	0.749531	0.015097
0.001	0.001	0.805502	0.001389	0.749974	0.008746
0.001	0.01	0.807199	0.002244	0.748659	0.024330
0.001	0.1	0.807488	0.002855	0.742809	0.008732
0.001	0.5	0.808301	0.002550	0.743613	0.005176
0.01	0	0.808497	0.001986	0.746327	0.013832
0.01	0.001	0.807882	0.001140	0.745011	0.009703

0.01	0.01	0.807830	0.002581	0.745956	0.019772
0.01	0.1	0.807490	0.001544	0.742809	0.007443
0.01	0.5	0.806455	0.003917	0.746672	0.016765
0.1	0	0.808043	0.003627	0.748338	0.019278
0.1	0.001	0.806647	0.002709	0.749709	0.011802
0.1	0.01	0.806919	0.001940	0.750163	0.009102
0.1	0.1	0.806353	0.002493	0.749281	0.012843
0.1	0.5	0.806662	0.003859	0.748284	0.015316
0.5	0	0.807263	0.005505	0.748579	0.029655
0.5	0.001	0.806775	0.002055	0.747499	0.007840
0.5	0.01	0.805568	0.003556	0.749670	0.008562
0.5	0.1	0.806377	0.001835	0.750155	0.010404
0.5	0.5	0.806410	0.002748	0.749601	0.013987

附录 5. 因子载荷矩阵：

指标名	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
LOAN_TIME	-0.00835	0.01728	-0.01527	0.00149	-0.00155	-0.01504
L_OVD_MONTH	0.02653	-0.00919	-0.13081	-0.04727	0.03785	0.11322
D_OVD_AMOUNT_MAX	0.04774	-0.00434	-0.1142	-0.11829	-0.01939	0.01458
D_OVD_MONTH_CV	0.14898	0.3283	-0.09521	0.01233	-0.00955	0.02704
L_MED_PERCENT	-0.02598	-0.01528	0.11263	-0.00044	0.05231	0.00137
L_ZFZ_PERCENT	-0.0223	0.02307	-0.15051	0.00141	0.00682	0.04025
D_USING_COUNT	0.22454	0.06497	-0.00863	0.00609	-0.00998	0.00736
L_LASTLOAN_AGE	-0.03441	-0.0076	-0.11073	0.02717	0.49166	0.03858
D_CREDIT_PERCENT	-0.00985	-0.01158	0.00445	-0.00068	-0.00737	-0.02058
MARRY_WOE	-0.02757	-0.01787	-0.01336	0.00321	-0.01774	0.00466
D_OVD_COUNT	-0.01084	0.42285	0.00795	-0.0048	-0.03426	-0.01638
D_OVD_MONTH	-0.11694	0.42792	0.0427	-0.00602	-0.01865	-0.03686
D_MIN_CREDIT_LIMIT_PER_ORG	0.11481	0.05212	-0.04326	0.00122	-0.00887	0.01101
LD_PASS_PERCENT	-0.04471	-0.06714	-0.07738	0.00171	-0.00374	-0.02374
L_FINANCE_ORG_COUNT	-0.07338	0.02384	0.42399	-0.00514	0.02741	0.0836
L_RECENTLOAN_COUNT	0.02285	0.05449	-0.1927	0.01344	0.00491	0.17845
D_PASS_PERCENT	-0.00386	-0.07523	-0.05323	0.00356	-0.02962	0.07925
L_SHORT_PERCENT	-0.01104	0.00833	0.06002	0.01743	0.09837	0.04154
L_OVD_MONTH_STD	0.01063	-0.02412	-0.13863	-0.01994	-0.02012	0.55034
S_FINANCE_ORG_COUNT	-0.05175	0.00447	-0.02003	-0.00906	0.00024	0.02534
L_OVD_AMOUNT_AVG	0.00644	-0.0013	0.00336	0.51067	0.00256	-0.01248
S_OVD_COUNT	0.00131	-0.0253	0.01541	0.0036	-0.01454	-0.02587
L_HIGHEST_OA_PER_MON	-0.01304	-0.0067	-0.02693	0.00153	-0.01831	0.0116
L_YQ_PERCENT	0.01527	-0.00532	-0.01822	0.58413	-0.00101	-0.02075
L_Z_PERCENT	-0.00355	-0.00429	0.00905	-0.00405	-0.00277	-0.01042
D_PERCENT_BIN	0.01097	-0.02547	-0.03609	0.00187	-0.00108	0.01545
D_LASTLOAN_AGE_BIN	-0.13519	0.1218	0.08231	0.00208	-0.00052	-0.00671
COMMERCIAL_LOAN_COUNT	-0.01854	-0.001	-0.01215	0.00461	-0.01627	-0.01218
D_YQ_PERCENT	-0.00808	-0.03939	0.03879	0.01939	0.0044	-0.00725
D_LOAN_AGE	0.02607	0.14811	0.06838	-0.00082	0.03112	-0.00185
L_YJ1_COUNT	0.00073	-0.00905	-0.00044	0.00589	-0.0229	-0.00697
ALL_6_AMOUNT	0.31441	-0.03579	0.01305	0.00584	-0.01655	-0.02005
L_LOAN_AGE	0.00472	-0.02953	0.05808	-0.01469	0.42118	0.02611
L_OVD_COUNT	0.0172	-0.02255	0.01684	-0.00815	-0.11772	-0.10493
QUERY_RECENT	-0.01874	-0.00263	0.0117	-0.00258	0.00655	-0.01042
L_PERCENT_BIN	-0.02442	-0.00265	-0.08405	0.0024	-0.00341	-0.04961
S_MAX_CREDIT_LIMIT_PER_ORG	-0.02751	-0.02002	-0.01696	-0.00106	-0.01877	-0.01598
D_OVD_MONTH_MAX	0.03238	-0.02197	-0.01729	-0.02557	0.0029	0.01028
D_RECENTLOAN_COUNT	-0.0781	0.04047	0.01552	0.0022	-0.00022	0.00681
L_6_AMOUNT	-0.05244	0.00985	0.53184	-0.00671	-0.0075	-0.0577

L_JQ_PERCENT	-0.04154	-0.01041	0.03797	-0.0052	0.46018	-0.07105
D_CREDITLIMIT_CV	0.26912	-0.08966	-0.1594	0.00264	-0.03864	0.00475
D_MAX_CREDIT_LIMIT_PER_ORG	0.42374	-0.09752	-0.12552	0.00662	-0.03319	-0.01569
D_STATE1	0.06976	-0.05212	-0.03141	-0.00417	0.01204	-0.0014
L_YJ2_COUNT	-0.04223	-0.0275	0.11321	-0.00743	-0.06165	-0.21914
L_BALANCE	-0.04958	0.00697	0.23675	0.0048	-0.03241	-0.13283
L_OVD_MONTH_CV	-0.04167	-0.0105	0.09683	-0.01143	-0.001	0.52274
L_LOSTBALANCE_MAX	-0.06427	0.00698	0.09365	-0.01197	0.01169	-0.03121

指标名	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12
LOAN_TIME	0.007640	-0.017770	-0.000330	0.005750	-0.003160	0.008110
L_OVD_MONTH	0.006710	-0.016850	0.000310	0.005420	-0.002820	0.009590
D_OVD_AMOUNT_MAX	-0.010220	-0.020920	-0.009170	0.038540	0.012270	0.007470
D_OVD_MONTH_CV	0.007670	-0.002770	0.006060	-0.014440	-0.004230	0.001560
L_MED_PERCENT	0.008730	-0.007970	0.005760	-0.009560	-0.002480	0.000270
L_ZFZ_PERCENT	0.004220	0.007420	0.010720	-0.011790	-0.000890	-0.010890
D_USING_COUNT	-0.017480	-0.002540	0.035090	-0.022540	0.007410	-0.019190
L_LASTLOAN_AGE	0.037340	-0.047260	0.043080	0.029950	0.006010	0.010060
D_CREDIT_PERCENT	0.008840	-0.001040	-0.010110	0.005590	-0.003270	0.008610
MARRY_WOE	0.010000	-0.000890	-0.010750	0.006460	-0.003600	0.007160
D_OVD_COUNT	0.018360	0.024450	0.050250	0.085180	-0.009520	0.001970
D_OVD_MONTH	0.063860	-0.082780	0.016600	0.098690	0.028870	0.023390
D_MIN_CREDIT_LIMIT_PER_ORG	-0.004000	-0.021550	0.014660	0.017550	0.080850	0.064140
LD_PASS_PERCENT	-0.052640	0.242050	0.072130	-0.138620	0.000110	0.019100
L_FINANCE_ORG_COUNT	0.041450	-0.051010	0.009550	0.157560	0.019390	0.026340
L_RECENTLOAN_COUNT	-0.022730	0.011270	0.016970	-0.026660	0.011130	-0.045310
D_PASS_PERCENT	0.768490	0.259960	-0.070080	-0.254570	-0.009060	0.065770
L_SHORT_PERCENT	0.002400	0.002650	0.001820	-0.004790	-0.008990	-0.020500
L_OVD_MONTH_STD	0.000640	0.016150	0.057070	0.004070	-0.005040	-0.014500
S_FINANCE_ORG_COUNT	-0.003750	0.005350	0.014010	-0.018720	-0.003530	-0.025910
L_OVD_AMOUNT_AVG	-0.104600	0.087440	0.095320	0.862390	0.002940	0.021250
S_OVD_COUNT	-0.021250	0.123710	0.344250	0.746700	-0.004110	0.025020
L_HIGHEST_OA_PER_MON	-0.037540	-0.015950	0.021320	0.015240	0.013480	0.918200
L_YQ_PERCENT	0.569320	0.493100	0.168500	0.091130	-0.002500	0.009130
L_Z_PERCENT	-0.005370	-0.006270	-0.002430	-0.000050	0.000950	0.019300
D_PERCENT_BIN	-0.061410	-0.009320	-0.021380	-0.005990	0.052570	0.169670
D_LASTLOAN_AGE_BIN	0.001400	-0.011330	0.025790	-0.017260	0.103660	-0.011460
COMMERCIAL_LOAN_COUNT	-0.004440	0.000640	-0.007650	0.008070	0.001670	0.005800
D_YQ_PERCENT	-0.001180	-0.000620	0.010880	0.012420	-0.006340	0.016690
D_LOAN_AGE	0.094170	0.035730	0.000320	-0.033980	0.004280	-0.129440
L_YJ1_COUNT	-0.037620	-0.213150	-0.177270	0.440640	0.007940	-0.000360
ALL_6_AMOUNT	0.024530	-0.008950	0.373650	0.702250	0.008920	0.035110
L_LOAN_AGE	-0.018330	-0.020620	0.060450	0.029880	0.008120	-0.007070

L_OVD_COUNT	-0.019090	-0.019740	-0.025310	0.023330	0.021070	0.177250
QUERY_RECENT	0.683400	-0.057720	0.065140	-0.015650	-0.003890	-0.153040
L_PERCENT_BIN	-0.02497	-0.00019	0.02093	0.06021	0.06045	0.01424
S_MAX_CREDIT_LIMIT_PER_ORG	-0.00717	0.00488	0.03103	-0.01124	0.00328	0.00112
D_OVD_MONTH_MAX	0.01454	0.57499	-0.00675	-0.00077	0.00322	0.00652
D_RECENTLOAN_COUNT	-0.03603	0.00845	-0.00374	0.00499	-0.00082	-0.00078
L_6_AMOUNT	-0.10712	-0.00756	-0.01822	0.03998	0.038	0.03826
L_JQ_PERCENT	0.02368	0.0083	0.06052	-0.14469	-0.13149	-0.00684
D_CREDITLIMIT_CV	0.08112	0.0047	0.01582	-0.00191	0.02147	0.00799
D_MAX_CREDIT_LIMIT_PER_ORG	0.00962	0.02529	0.03326	-0.0117	-0.0153	-0.00605
D_STATE1	0.00046	0.01351	0.02136	0.00487	0.00154	-0.00308
L_YJ2_COUNT	0.43615	0.0097	0.26137	0.04058	0.05746	-0.0081
L_BALANCE	-0.05197	0.01766	0.02892	0.01196	0.42198	-0.03335
L_OVD_MONTH_CV	0.03298	0.01523	-0.11046	0.01764	-0.06585	-0.03148
L_LOSTBALANCE_MAX	-0.06842	-0.11731	0.00937	-0.00164	-0.02957	0.58221

指标名	Factor13	Factor14	Factor15	Factor16	Factor17	Factor18
LOAN_TIME	-0.00376	-0.01162	0.01147	0.00837	-0.01097	-0.01222
L_OVD_MONTH	-0.00954	-0.0187	-0.047	0.02202	-0.00926	0.07908
D_OVD_AMOUNT_MAX	0.00722	-0.02595	-0.04833	0.01186	0.00353	0.01547
D_OVD_MONTH_CV	-0.02819	0.42296	-0.09954	-0.10288	0.03898	0.01327
L_MED_PERCENT	-0.00553	0.00345	-0.01367	-0.00681	0.01607	0.2016
L_ZFZ_PERCENT	0.01042	0.01104	-0.03146	-0.03785	0.01804	0.01602
D_USING_COUNT	-0.0051	0.20671	0.07023	-0.1491	0.17107	-0.02133
L_LASTLOAN_AGE	-0.00872	-0.0112	-0.05645	-0.01141	-0.00039	-0.21999
D_CREDIT_PERCENT	0.00885	0.01858	-0.00911	0.0123	0.06273	-0.01111
MARRY_WOE	-0.00344	0.01606	-0.01052	-0.00442	0.00382	-0.00242
D_OVD_COUNT	-0.0046	0.07478	-0.05054	0.03434	0.03432	0.00217
D_OVD_MONTH	-0.00073	-0.07722	-0.13223	0.08291	0.02812	-0.01628
D_MIN_CREDIT_LIMIT_PER_ORG	-0.02882	0.03583	-0.08395	0.75793	0.10602	0.02013
LD_PASS_PERCENT	-0.01757	0.06425	0.47725	-0.00575	0.08738	-0.06664
L_FINANCE_ORG_COUNT	-0.00762	-0.02779	-0.08997	-0.02575	0.00549	0.09016
L_RECENTLOAN_COUNT	-0.00281	-0.02794	-0.15232	0.06398	-0.06988	0.10149
D_PASS_PERCENT	-0.0273	0.06188	0.72987	-0.06221	-0.03944	0.09568
L_SHORT_PERCENT	-0.0007	-0.00347	-0.00846	0.02211	0.00237	0.18365
L_OVD_MONTH_STD	0.00936	-0.00022	0.04058	0.00348	0.00636	0.01369
S_FINANCE_ORG_COUNT	0.55359	0.01898	-0.0055	-0.04035	0.00602	0.01267
L_OVD_AMOUNT_AVG	0.00031	0.00222	0.00355	0.00148	0.00226	-0.00651
S_OVD_COUNT	-0.07553	0.02523	0.00959	-0.00051	0.01611	-0.02466
L_HIGHEST_OA_PER_MON	0.00136	0.00411	0.00225	-0.02146	0.01762	-0.00591
L_YQ_PERCENT	-0.00981	0.00227	0.00038	-0.00351	0.00245	0.00882
L_Z_PERCENT	-0.00003	0.00529	-0.0064	-0.00477	-0.00017	0.00926
D_PERCENT_BIN	-0.01392	-0.05197	-0.00656	-0.03544	0.07553	-0.0049

D_LASTLOAN_AGE_BIN	0.02462	-0.08795	0.12189	0.22321	-0.31252	-0.01014
COMMERCIAL_LOAN_COUNT	-0.0066	0.00539	0.00141	-0.0188	0.0043	-0.00708
D_YQ_PERCENT	0.00066	0.00052	0.03342	-0.00042	0.00887	-0.00145
D_LOAN_AGE	-0.00613	-0.24224	0.07647	0.01956	-0.052	0.01527
L_YJ1_COUNT	-0.00376	0.00125	0.00294	-0.00904	0.00064	-0.01825
ALL_6_AMOUNT	0.02471	0.05213	0.04234	0.10165	-0.01604	-0.03576
L_LOAN_AGE	-0.00706	0.02184	0.03345	0.01272	-0.01293	0.09554
L_OVD_COUNT	-0.00379	0.03747	-0.01486	-0.02889	0.00544	-0.01763
QUERY_RECENT	-0.00495	0.0259	0.06225	-0.01728	-0.03916	-0.00325
L_PERCENT_BIN	0.00884	0.01185	0.04695	0.00871	0.00855	0.79812
S_MAX_CREDIT_LIMIT_PER_ORG	0.67443	-0.01831	-0.0498	0.01181	-0.00491	-0.00288
D_OVD_MONTH_MAX	0.00141	0.00744	-0.00464	-0.00623	0.00336	0.00753
D_RECENTLOAN_COUNT	0.0059	-0.09947	0.03791	0.16639	0.80402	0.00212
L_6_AMOUNT	-0.01684	-0.01944	-0.01577	0.02579	-0.02551	-0.18842
L_JQ_PERCENT	-0.00426	0.01667	-0.04586	-0.01995	0.00828	0.07954
D_CREDITLIMIT_CV	-0.0646	-0.43862	-0.19572	-0.22204	-0.1463	0.00004
D_MAX_CREDIT_LIMIT_PER_ORG	-0.04723	-0.10031	-0.08363	0.24573	-0.09847	-0.00499
D_STATE1	-0.02048	0.5423	0.01873	-0.01613	-0.14576	0.01581
L_YJ2_COUNT	0.01108	0.01766	0.01443	-0.04414	0.01484	-0.20309
L_BALANCE	-0.01658	0.01245	0.00497	-0.00887	-0.02016	0.01397
L_OVD_MONTH_CV	-0.00094	0.00307	0.02285	-0.00731	0.00806	-0.08668
L_LOSTBALANCE_MAX	-0.00228	0.02638	0.05481	-0.01785	0.00078	-0.00253

指标名	Factor19	Factor20	Factor21	Factor22	Factor23	Factor24
LOAN_TIME	-0.00035	0.02212	-0.00626	-0.0002	0.00415	0.01709
L_OVD_MONTH	-0.04361	0.04636	-0.00283	0.04977	0.01952	-0.00308
D_OVD_AMOUNT_MAX	0.01394	0.01088	-0.04857	-0.04758	-0.01027	-0.01012
D_OVD_MONTH_CV	-0.14507	-0.13231	-0.04387	0.02174	-0.0143	-0.00581
L_MED_PERCENT	-0.01081	0.00431	-0.05321	-0.04297	-0.01963	-0.04912
L_ZFZ_PERCENT	0.01556	0.01615	-0.05055	-0.10652	-0.01716	-0.01136
D_USING_COUNT	-0.06379	-0.00397	-0.0184	-0.0127	-0.0058	-0.00428
L_LASTLOAN_AGE	0.00892	0.05208	-0.01916	0.00505	-0.02259	-0.05761
D_CREDIT_PERCENT	-0.05578	0.92121	-0.02238	-0.00424	-0.00988	-0.00452
MARRY_WOE	-0.00313	-0.00799	-0.01976	-0.01025	-0.00933	0.99434
D_OVD_COUNT	-0.00116	-0.04404	0.00166	-0.01145	-0.02062	-0.03556
D_OVD_MONTH	-0.00705	0.01953	-0.00626	-0.0146	-0.02531	-0.01155
D_MIN_CREDIT_LIMIT_PER_ORG	-0.06107	-0.00038	-0.03694	-0.01732	0.00777	-0.00253
LD_PASS_PERCENT	0.02244	-0.03532	-0.16317	-0.00973	-0.01951	0.00066
L_FINANCE_ORG_COUNT	0.00092	0.01182	-0.03843	-0.04835	-0.00686	-0.01549
L_RECENTLOAN_COUNT	-0.02251	0.04566	0.08875	0.11083	0.04678	0.01304
D_PASS_PERCENT	-0.02455	-0.0048	0.14403	0.02218	0.02497	-0.01352
L_SHORT_PERCENT	-0.02276	0.03115	-0.02911	-0.02936	-0.00803	-0.04512

L_OVD_MONTH_STD	0.00432	-0.00533	-0.01255	-0.00942	-0.0261	0.00584
S_FINANCE_ORG_COUNT	-0.02298	0.00376	-0.01296	0.02209	0.10349	-0.01776
L_OVD_AMOUNT_AVG	0.00219	-0.00335	0.00457	0.00183	0.00054	0.0049
S_OVD_COUNT	0.01479	-0.01162	-0.00852	-0.01756	0.97318	-0.00733
L_HIGHEST_OA_PER_MON	0.01051	-0.0018	-0.02109	0.97213	-0.01347	-0.00947
L_YQ_PERCENT	0.00096	0.00172	-0.00936	0.00079	0.00672	0.00127
L_Z_PERCENT	0.0026	-0.00178	-0.00348	-0.00162	-0.01279	-0.00448
D_PERCENT_BIN	0.94667	-0.05334	-0.00166	0.01345	0.01692	-0.00171
D_LASTLOAN_AGE_BIN	0.12626	0.1219	0.01894	-0.00979	-0.03411	-0.03382
COMMERCIAL_LOAN_COUNT	0.00268	-0.00228	-0.00788	-0.01797	0.0004	0.0017
D_YQ_PERCENT	0.02161	0.00655	0.02221	0.01079	0.00686	0.0106
D_LOAN_AGE	0.00048	0.092	0.04581	-0.00011	0.01834	0.05465
L_YJ1_COUNT	0.00042	-0.00179	-0.00293	-0.01601	-0.01177	-0.00334
ALL_6_AMOUNT	0.17229	-0.0628	0.03344	-0.02027	-0.01818	-0.01643
L_LOAN_AGE	-0.01023	-0.02434	0.03085	-0.00801	0.02733	0.04176
L_OVD_COUNT	-0.00702	-0.01316	0.03142	-0.07827	0.01317	-0.0053
QUERY_RECENT	0.00706	-0.02136	0.95943	-0.02336	-0.00969	-0.01706
L_PERCENT_BIN	0.00534	-0.00541	-0.00261	-0.01049	-0.02238	-0.00037
S_MAX_CREDIT_LIMIT_PER_ORG	0.00709	0.00918	-0.00117	-0.02513	-0.24151	0.00761
D_OVD_MONTH_MAX	-0.03073	-0.0106	-0.00319	0.00043	0.00427	-0.00377
D_RECENTLOAN_COUNT	0.1292	0.1069	-0.03684	0.01481	0.00424	-0.0111
L_6_AMOUNT	-0.03536	0.01593	0.02425	0.00091	0.01698	-0.00585
L_JQ_PERCENT	0.00049	-0.03363	0.00585	-0.03895	-0.02742	-0.02883
D_CREDITLIMIT_CV	0.03662	0.14369	-0.07974	0.00806	0.01101	-0.03514
D_MAX_CREDIT_LIMIT_PER_ORG	-0.06283	0.02654	-0.01767	-0.01546	0.00976	-0.0337
D_STATE1	0.01152	0.16357	0.0169	0.00385	0.03557	0.01381
L_YJ2_COUNT	0.03493	-0.02843	-0.0027	-0.09568	-0.06024	0.01054
L_BALANCE	0.00312	-0.00831	0.02792	0.1051	0.02404	-0.00923
L_OVD_MONTH_CV	0.01913	-0.02108	0.01442	0.01129	0.00142	0.00091
L_LOSTBALANCE_MAX	-0.01956	-0.00368	0.043	0.02471	0.00869	0.03006

指标名	Factor25	Factor26	Factor27	Factor28
LOAN_TIME	-0.00467	0.00007	-0.00443	0.99987
L_OVD_MONTH	0.01147	0.00252	0.01584	0.03539
D_OVD_AMOUNT_MAX	0.00279	-0.00028	0.01677	0.02756
D_OVD_MONTH_CV	0.01331	-0.00129	-0.00794	0.02468
L_MED_PERCENT	-0.0483	-0.02409	-0.0035	0.02274
L_ZFZ_PERCENT	-0.0019	-0.01727	-0.02205	0.02191
D_USING_COUNT	0.00233	0.00043	-0.02485	0.02105
L_LASTLOAN_AGE	-0.06529	0.00598	-0.04851	0.02
D_CREDIT_PERCENT	-0.00153	0.0002	-0.00244	0.01958
MARRY_WOE	-0.00382	-0.00455	0.00241	0.01754
D_OVD_COUNT	-0.00976	-0.00972	-0.00871	0.01373

D_OVD_MONTH	-0.01198	-0.00067	0.01301	0.01308
D_MIN_CREDIT_LIMIT_PER_ORG	-0.00496	-0.00385	-0.01869	0.01141
LD_PASS_PERCENT	-0.00146	0.00159	-0.00384	0.01039
L_FINANCE_ORG_COUNT	0.00979	0.0037	-0.00684	0.00977
L_RECENTLOAN_COUNT	0.03104	0.00282	0.06079	0.00965
D_PASS_PERCENT	0.00906	-0.00837	0.01247	0.00876
L_SHORT_PERCENT	-0.04418	-0.01036	-0.00067	0.00795
L_OVD_MONTH_STD	-0.01595	0.01163	-0.00456	0.00612
S_FINANCE_ORG_COUNT	-0.00933	-0.00146	-0.01607	0.00503
L_OVD_AMOUNT_AVG	0.00721	-0.00852	0.00377	0.00348
S_OVD_COUNT	-0.01209	-0.01127	-0.00079	0.00274
L_HIGHEST_OA_PER_MON	-0.01557	-0.00063	-0.01441	0.00138
L_YQ_PERCENT	0.0028	0.00014	0.004	-0.00056
L_Z_PERCENT	0	0.99711	0.00074	-0.00067
D_PERCENT_BIN	0.00106	0.00332	0.00313	-0.00083
D_LASTLOAN_AGE_BIN	0.00155	-0.01487	-0.00503	-0.00313
COMMERCIAL_LOAN_COUNT	-0.0012	0.00047	0.99929	-0.00359
D_YQ_PERCENT	-0.00172	0.00178	-0.005	-0.00363
D_LOAN_AGE	-0.00698	0.0121	-0.01172	-0.0045
L_YJ1_COUNT	0.99598	0.00044	-0.00101	-0.00474
ALL_6_AMOUNT	-0.00656	0.00193	-0.01009	-0.00495
L_LOAN_AGE	0.04216	0.02377	0.03068	-0.00527
L_OVD_COUNT	-0.03516	-0.00511	-0.01078	-0.00539
QUERY_RECENT	-0.00393	-0.00316	-0.00762	-0.00718
L_PERCENT_BIN	-0.01621	0.01063	-0.01494	-0.00771
S_MAX_CREDIT_LIMIT_PER_ORG	-0.00008	0.00028	0.00217	-0.011
D_OVD_MONTH_MAX	0.00392	0.00357	0.00084	-0.01151
D_RECENTLOAN_COUNT	-0.0002	-0.0061	0.00372	-0.01386
L_6_AMOUNT	-0.00484	0.0073	-0.00071	-0.01403
L_JQ_PERCENT	-0.02833	-0.03701	-0.01866	-0.01438
D_CREDITLIMIT_CV	0.00699	-0.01382	-0.01056	-0.01612
D_MAX_CREDIT_LIMIT_PER_ORG	-0.00472	-0.00664	-0.00095	-0.01711
D_STATE1	-0.00158	0.00157	0.00377	-0.02348
L_YJ2_COUNT	-0.02316	0.01495	-0.05824	-0.02369
L_BALANCE	-0.00678	-0.00369	-0.03216	-0.02568
L_OVD_MONTH_CV	0.00378	-0.02879	-0.01767	-0.03216
L_LOSTBALANCE_MAX	-0.00666	0.00118	-0.01586	-0.04352