中國人民大學

RENMIN UNIVERSITY OF CHINA

基于逻辑回归和 XGBoost 模型的信贷平台的贷款违约率影响因素研究

姓　　名：　罗则鸣

学　　号：2019200884

学　　院：　商学院

专　　业：工商管理类

2021 年 6 月

摘 要

本文使用了某信贷平台提供的贷款顾客数据，根据用户的社会人口学变量、交易信息以及 n 系列匿名行为特征，对顾客贷款违约的可能性进行预测，并希望通过分析筛选出的显著影响因素，对信贷平台审慎发放贷款，预防违约风险提出具体建议。本文首先介绍防范贷款违约风险，对提升银行竞争力，增强银行盈利能力的重要性；其次以收集到的数据集为案例，对样本数据进行了一系列的统计描述，展现变量弱相关和分布差异大的特点；之后基于数据的特点，先后建立逻辑回归和 XGBoost 模型，最终实现了对目标变量的准确预测，并找到影响顾客是否流失的关键因素。

关键词：贷款违约率，交易记录，逻辑回归，集成学习

目录

1 引言

1.1 研究背景

贷款是商业银行的一项重要业务，是其主要盈利来源之一，对于商业银行的发展有着十分重要的意义，因此贷款风险也一直是银行业关注的重点风险。不良贷款率作为评价金融机构信贷资产安全状况的一项重要指标，是贷款风险的重要反映。在 2012 年到 2013 年上半年期间，我国商业银行不良贷款率水平较低，维持在 0.9%附近。但从 2013 年下半年开始，不良贷款率从 0.9%迅速攀升

到 1.7%附近，然后稳定在了较高水平，2018 年期间不良贷款率达到了近年来的最高水平 1.87%。不良贷款的产生和积累会造成银行资本充足率的下降，给银行发展带来瓶颈威胁，因此建立合适的模型控制贷款违约风险显得至关重要。

银保监会曾下发文件，提出要积极发展消费金融，增强消费对经济的拉动作用，支持发展消费信贷，满足人民群众日益增长的美好生活需要。在互联网金融监管收紧前，以消费金融为主的个人短期贷款一度发展迅速。个人短期贷款快速发展一方面促进消费，缓解了宏观经济压力，但另一方面贷款增速过快将增加贷款违约风险，尤其是在储蓄结构与消费结构年龄差距越来越大的背景下。

据蚂蚁金服和福达国际《中国养老前景调查报告》显示，中国 35 岁以下年轻人中 56%没有储蓄，有储蓄的年轻人中，每月储蓄仅 1389 元;而在消费贷款中，80 后和 90 后占比高达 80.77%。消费贷款进一步趋向年轻化，储蓄趋向年老化，年轻人借贷过多，一旦工资增长停滞，个人短期贷款违约不可避免。因此年龄、学历等人口特征对短期贷款违约风险的影响及基于人口特征的短期贷款违约风险预测研究，具有重要的现实意义。

1.2 文献综述

1.2.1 贷款违约

关于贷款违约问题，国外的研究相对较早。早在 20 世纪中晚期，David 和 Jon 通过使用贷款发放时提供的信息来确定贷款违约类与不违约类之间是否存在差异；Foster 和 Order 基于期权的方法对信用违约进行分析，研究发现期权违约模型能够很好地分析信用违约问题。之后，随着计算机水平的发展，出现了一些新兴的贷款违约评估方法，最常用的是机器学习方法和统计方法。机器学习方法中有神经网络、随机森林、支持向量机等。比如 Ma 和 Yang 建立了短期贷款人工神经网络模型；杨保安和季海基于神经网络方法对贷款违约风险进行预测研究；方匡南等基于随机森林分析影响住房贷款违约的风险因素；Danenas 和 Garsva 在粒子群优化技术上构建支持向量机进行违约风险研究。统计方法中最常用的是 Logistic 回归。比如 Lobna Abid 等对突尼斯某商业银行的违约数据采用 Logistic 回归来区分“好”和“坏”的借款人；胡毅等在 Logistic 回归下分析影响住房贷款违约的风险因素；舒扬和杨秋怡发现 Logistic 回归在分析汽车贷款违约数据时取得了良好效果。也有不少学者基于 Logistic 回归构建组合模型，如石庆焱组合 Logistic 回归与神经网络，建立了一个比 Logistic 回归精度更高、稳健性更好的两阶段混合模型；张奇等建立 Logistic 和支持向量机组合型贷款违约风险预警模型。近年来，考虑文本数据的违约建模也越来越多，比如陈林等在 P2P 借款陈述文本中提取了三类信息：文字特征信息、反映还款能力和还款意愿的信息以及对资金需求的情感特征信息。王小燕等在 Logistic 回归中通过文本挖掘技术引入了文本先验信息以体现信用风险指标的重要性。

1.2.2 个人住房按揭抵押贷款违约

宏观层面来讲，就业情况与住房贷款违约正相关 (Burrows, 1998)，利率与住房贷款违约负相关 (何晓晴等, 2005)，宏观经济越好，就业机会越多，收入增长带动住户还款能力增强，利率越低，住户还款负担越小，违约风险越低。

宏观因素只能解释群体性违约风险，但无法解释不同个体间的违约风险差异。因此需要从借款人特征、贷款特征等微观因素层面研究住房抵押贷款违约。Davidoff 和 Welke (2017) 发现离婚、怀孕等情形导致的个人破产、家庭关系不和等因素增加了借贷违约风险。高广春 (2017) 认为个人贷款金额增加、贷款期限延长增加了违约风险，进而限制了个人住房抵押贷款的增长空间。

1.2.3 个人短期贷款

目前学者对个人贷款逾期违约研究，却主要集中于个人住房按揭抵押贷款 (Mortgage) 方面，针对个人短期贷款违约风险的研究鲜有涉及。

而且，上述研究均基于按揭贷款经验数据进行分析，个人按揭贷款期限长，且有住房作为担保物，容易构建违约风险模型，并对其进行预测。

此外，个人短期贷款期限短，具有以下三点差异化特征:第一、期限在一年以内，与商业按揭贷款存在明显的区别，商业按揭贷款期限一般在五年以上。第二、贷款目的性的不同。个人短期贷款主要用于个体经营或消费需求，经营投入对象与消费内容具有较大的分散性。而个人住房按揭贷款则主要为满足住房需求而申请的房产贷款，投资标的较为统一。第三、贷款申请主体特征差异明显。个人短期贷款的主体以经商企业主、个体工商户、具有较高消费性需求的人群为主，而个人按揭贷款的主体以具有购房意向的住户群体为主。

个人短期贷款具有的期限短、资金流向复杂等特征，加大了违约风险研究和预测难度。再加上，个人短期借款数据属于银行商业秘密，一般不对外披露，造成关于个人短期贷款违约风险的研究著述较少。

本文创新点主要有三个方面：第一，填补个人短期贷款违约风险研究的空缺。个人短期借款数据属于传统银行的商业秘密，一般不对外披露，造成关于个人短期贷款违约风险的研究著述较少。本文则乘互联网金融发展的东风，利用某信贷平台的公开数据集，抛砖引玉，试水个人短期贷款违约风险的研究；第二，数据集的样本质量更优。已有的研究样本大多是来自传统的商业银行，受限于业务模式和处理能力，已有的研究获取的样本量一般都不超过几万，样本特征一般不超过三十个。得益于大数据技术，本文的数据集则源自某信贷平台的贷款记录，总样本量为一百万条，变量达 47 个，不仅丰富了了传统的人口社会学因素和贷款因素下的变量范畴，还创新增加了 15 个贷款人行为计数特征；第三，尝试为个人贷款违约引入新研究模型。随着个人短期贷款规模迅速扩大，其

违约风险的研究价值愈发重要。因此本文在借鉴前人研究人口特征对住房按揭抵押贷款违约影响的经验基础上,利用 逻辑回归和 XGBoost 模型进行数据挖掘分析。

2. 数据介绍

2.1 数据来源

本文的数据来自数据科学网站和鲸社区 https://www.heywhale.com/home/dataset。

该数据来自某信贷平台的贷款记录，为两个 csv 文件，总数据量为 100 万条，80 万条为训练集，20 万条为测试集 A，包含 47 列变量信息，其中 15 列为匿名变量。employmentTitle、purpose、postCode 和 title 等信息进行过脱敏处理。

特征分为三大类，分别为社会人口学特征、贷款信息特征以及用户行为特征。此外，对于数值型变量中连续性和离散型的划分，本文是按数值不重复个数是否大于 15 的标准划分地。变量详细信息见表 1：

表 1 变量名称、含义及类型

变量名称	变量含义	变量类型
id	为贷款清单分配的唯一信用证标识	连续型变量
loanAmnt	贷款金额	连续型变量
term	贷款期限（year）	离散型变量
interestRate	贷款利率	连续型变量
installment	分期付款金额	连续型变量
grade	贷款等级	字符串型变量
subGrade	贷款等级之子级	字符串型变量
employmentTitle	就业职称	连续型变量
employmentLength	就业年限（年）	字符串型变量
homeOwnership	借款人在登记时提供的房屋所有权状况	离散型变量
annualIncome	年收入	连续型变量
verificationStatus	验证状态	离散型变量
issueDate	贷款发放的月份	字符串型变量
purpose	借款人在贷款申请时的贷款用途类别	离散型变量
postCode	借款人在贷款申请中提供的邮政编码的前 3 位数字	连续型变量

regionCode	地区编码	连续型变量
dti	债务收入比	连续型变量
delinquency_2years	借款人过去 2 年信用档案中逾期 30 天以上的违约事件数	连续型变量
ficoRangeLow	借款人在贷款发放时的 fico 所属的下限范围	连续型变量
ficoRangeHigh	借款人在贷款发放时的 fico 所属的上限范围	连续型变量
openAcc	借款人信用档案中未结信用额度的数量	连续型变量
pubRec	贬损公共记录的数量	连续型变量
pubRecBankruptcies	公开记录清除的数量	离散型变量
revolBal	信贷周转余额合计	连续型变量
revolUtil	循环额度利用率，或借款人使用的相对于所有可用循环信贷的信贷金额	连续型变量
totalAcc	借款人信用档案中当前的信用额度总数	连续型变量
initialListStatus	贷款的初始列表状态	离散型变量
applicationType	表明贷款是个人申请还是与两个共同借款人的联合申请	离散型变量
earliesCreditLine	借款人最早报告的信用额度开立的月份	字符串型变量
title	借款人提供的贷款名称	连续型变量
policyCode	公开可用的策略_代码=1 新产品不公开可用的策略_代码=2	单值变量
n 系列匿名特征	匿名特征 n0-n14，为一些贷款人行为计数特征的处理	连续型变量

2.2 数据预处理

2.2.1 缺失值

本文数据无重复值，但存在一定的缺失值情况。图 1 展示了存在缺失的特征及其缺失率。可以发现，n 系列匿名特征及 employmentLength 存在较为严重的缺失情况，缺失率都超过了 4%。本文采用缺失值填充的办法处理缺失值。

对于数值型变量，本文使用变量的中位数进行填充；对于分类型变量，本文使用变量的众数进

行填充；对于 employmentLength（就业年限），本文采取随机森林的方法进行拟合填充。

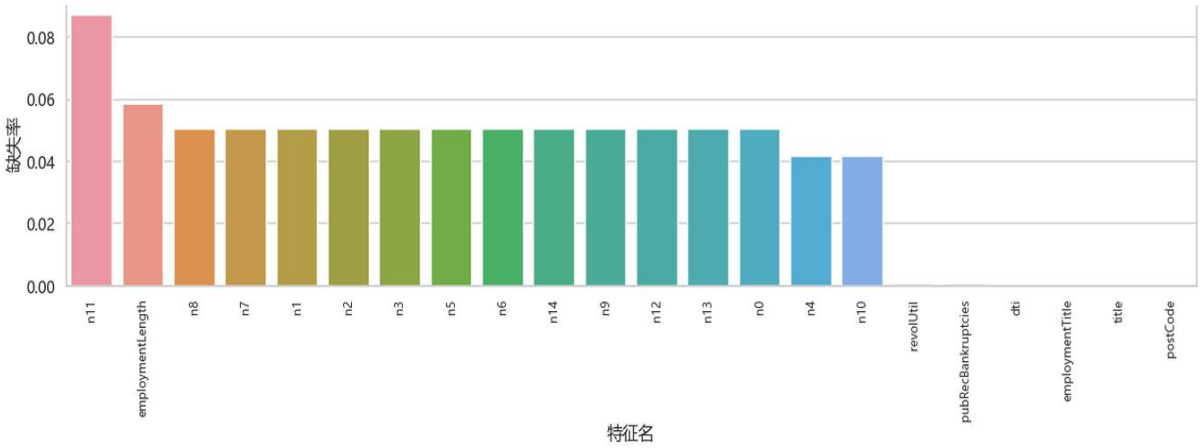


图 1：缺失特征的分布图

2.2.2 编码

在 47 个变量中，“grade”, “subGrade”,“employmentLength”，“issueDate”和 “earliesCreditLine”等 5 个变量是字符串类型，需要进行处理。

对于字符串型变量“grade”, “subGrade”和“employmentLength”，可以直接转化成数值型变量；对于“issueDate”（贷款发放的月份）和 “earliesCreditLine”（借款人最早报告的信用额度开立的月份），为了便于后续数据分析以及出于可解释性的考虑，本文对这两个变量合并，即做了两者年份之差，从而形成变量“最早报告信用额度开立的到贷款发放时的时长（年）”，并以其代替之。

在数值型变量中，“homeOwnership”，“verificationStatus”，“purpose”为分类数据，对事物进行分类但不具有内在的逻辑顺序，因此采用 One-Hoe 编码，给每一个类型分配一个独立的数值向量。

表 2 字符型变量处理（部分）

变量	字符型数据（部分）	数值型数据（部分）
grade（贷款等级）	B	2
	C	3
	A	1
	D	4
	E	5
	F	6
	G	7

subGrade（贷款等级之子级）	C1	31
	B4	24
	B5	25
	B3	23
	C2	32
	C3	33
	C4	34
employmentLength（就业年限（年））	10+ years	0
	2 years	2
	< 1 year	5
	3 years	8
	1 year	10
	5 years	7
	4 years	9
	6 years	1
	8 years	3
	7 years	4
	9 years	6

表 3：对“issueDate”和“earliesCreditLine”合并为“CreditLine”（最早报告信用额度开立的到贷款发放时的时长（年））

issueDate（部分）	earliesCreditLine（部分）	CreditLine（年）
2014/7/1	Aug-01	13
2012/8/1	May-02	10
2015/10/1	May-06	9
2015/8/1	May-99	16
2016/3/1	Aug-77	39

表 4： 对部分变量的 One-hot 编码

变量	水平	One-hot 编码
homeOwnership	x0_0	0
	x0_1	1
	x0_2	2
	x0_3	3
	x0_4	4
	x0_5	5
verificationStatus	x1_0	6
	x1_1	7
	x1_2	8
purpose	x2_0	9
	x2_1	10
	x2_2	11
	x2_3	12
	x2_4	13
	x2_5	14
	x2_6	15
	x2_7	16
	x2_8	17
	x2_9	18
	x2_10	19
	x2_11	20
	x2_12	21
	x2_13	22

2.2.3 高相关变量的处理

对数值型变量的相关系数进行初步的计算。相关系数是统计学分析变量之间相关性的最基础的手段。相关系数越接近 1，则变量之间的相关性越强。

由图 2 可知，少数变量之间的相关系数达到 0.75 以上，具有较为显著的线性相关。表 4 则展示了相关性系数大于 0.75 的相关关系。为了避免因相关性过高对模型表现性能的制约，本文从变量的解释性出发，剔除“installment”，“interestRate”，“grade”和“ficoRangeLow”等四个解

释性与其他变量重复的变量。对于其余的高相关变量，本文则通过主成分分析法进行降维处理。

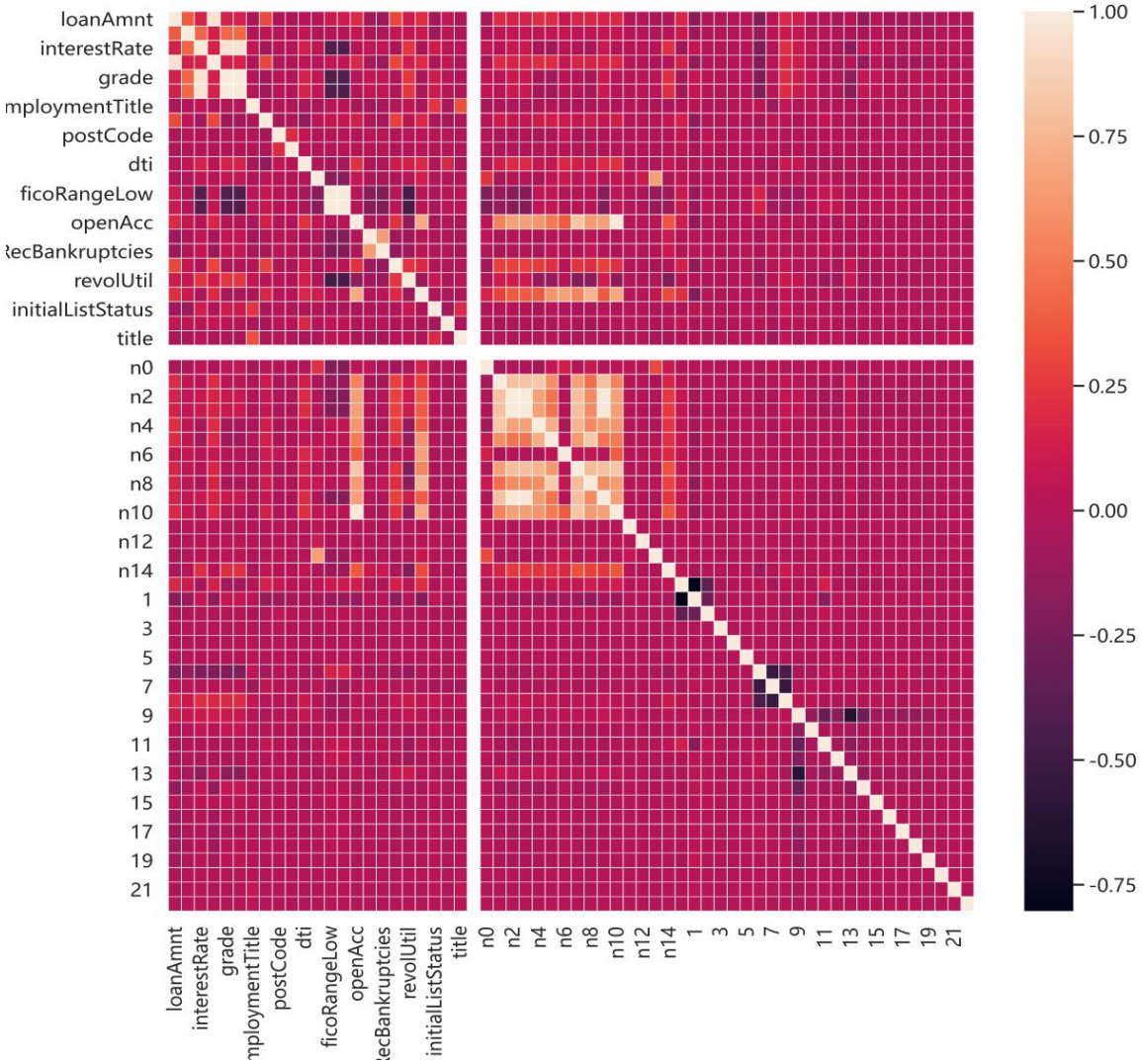


图 2：相关性热力图

表 5：相关性大于 0.75 的变量

feature_x	feature_y	correlation
n2	n3	1
ficoRangeLow	ficoRangeHigh	0.9999999
grade	subGrade	0.9939067
openAcc	n10	0.9838900
n2	n9	0.9820449
n3	n9	0.9820449
interestRate	subGrade	0.9708469
loanAmnt	installment	0.9533693
interestRate	grade	0.9532686

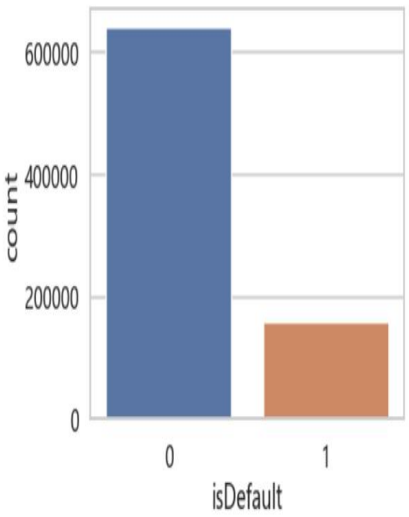
n5	n8	0.8384203
n7	n10	0.8268071
n1	n4	0.8266510
openAcc	n7	0.8175235
n1	n2	0.8081573
n1	n3	0.8081573
n1	n9	0.8012436
n7	n9	0.7948123
n2	n7	0.7907639
n3	n7	0.7907639
n7	n8	0.7754596

2.3 数据描述

2.3.1 字符型变量

首先对因变量和 5 个字符型变量进行基础的描述性统计，初步了解数据的分布情况，并将统计结果绘图如下。从图上可以看出，数据集并非关于目标变量均匀分布，样本中的大部分顾客为非违约客户；在研究的样本中，大部分的贷款等级为 A、B、C 等三个较高等级；本文注意到，工作年限大于十年的客户，其贷款违约率明显低于其他客户，这可能与其稳定的收入，更为充裕的资产以及成熟带来的自律有关；本文还注意到，不同的贷款等级以及贷款子等级下违约率也有差异。

此外，对于“issueDate”（贷款发放的月份）和 “earliesCreditLine”（借款人最早报告的信用额度开立的月份）合并后的新变量“CreditLine”（最早报告信用额度开立的到贷款发放时的时长（年）），其为数值型变量，由图 4 的分布图可以看出，大多数客户的已有报告信用额度为二十年左右。



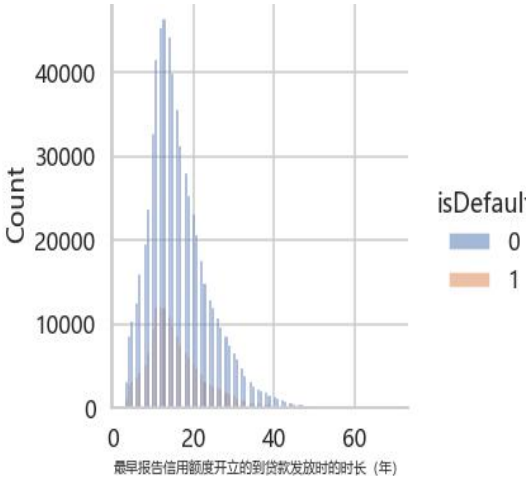


图 3：因变量的分布情况　　图 4：CreditLine 的违约率分布

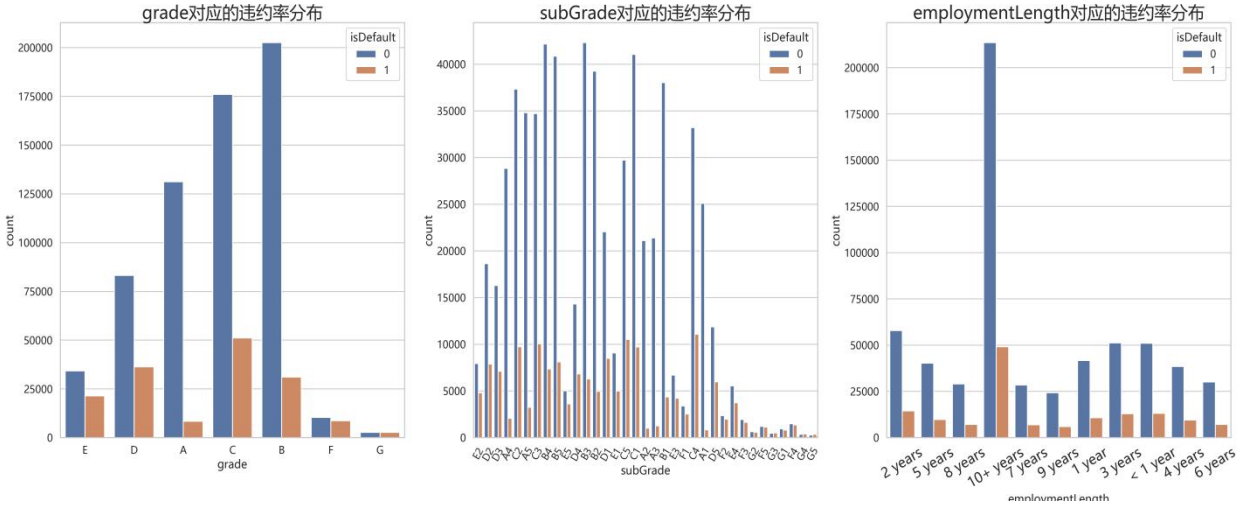


图 5：部分字符型变量的违约率分布

2.3.2 连续性数值变量

绘制 30 个数值型变量的分布图，从图 6 可以看出，样本大多服从偏态分布，且大都右偏；从集中度上看，部分特征的分布高度集中在某一狭窄区间内，如“annualIncome”以及 n 系列匿名特征；

样本数据的特征几乎都不服从正太分布，既有多峰的分布，如：总交易数和交易变化数；也有偏态的分布，如：信用额度和平均使用率的分布；从分布的区间上看，不同特征分布的区间跨度也很大。信用卡的平均使用率分布区间最短，在[0,1]之间；信贷上限的分布区间则最宽，在[0,35000]之间。

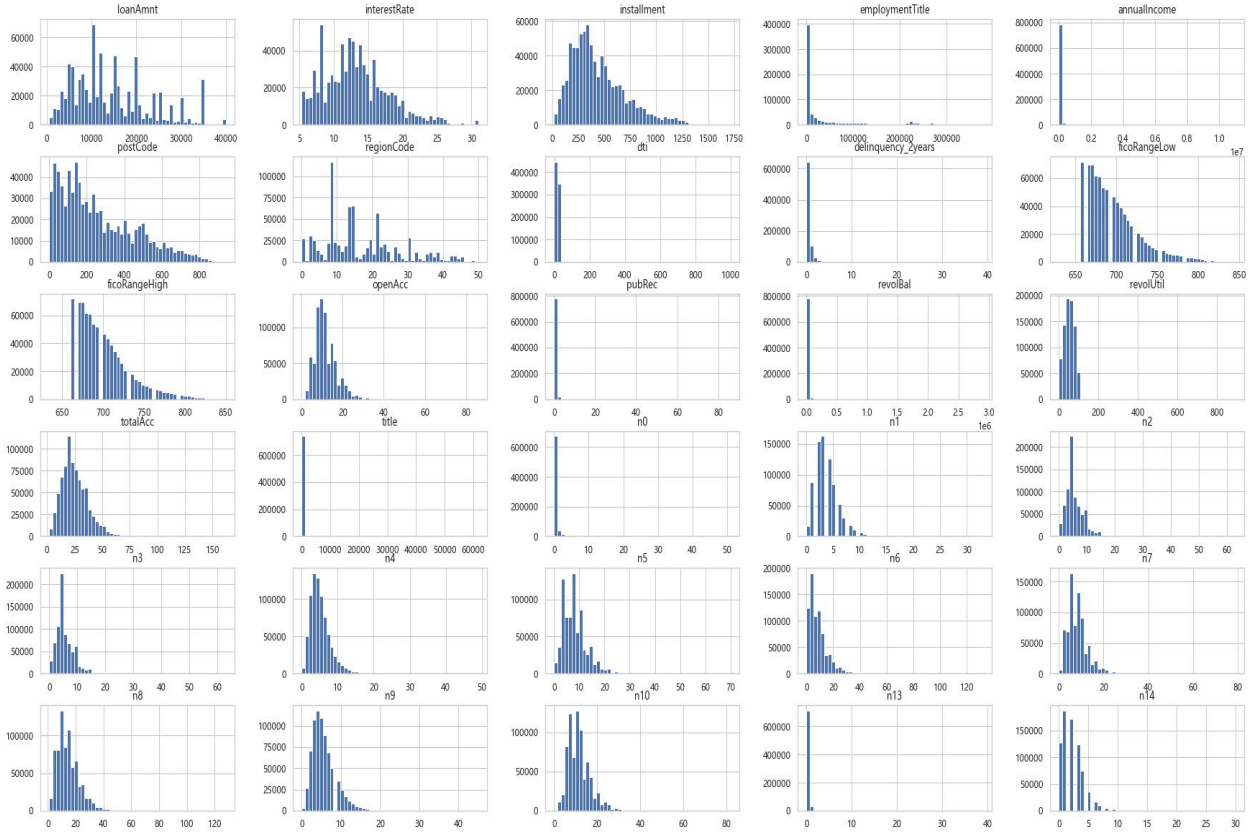


图 6：连续性变量的分布情况

2.3.3 离散型数值变量

本文将不同数值个数小于 15 的数值变量归为离散型数值变量。图 7 展示了离散型数值变量的分布情况，可以发现贷款期限以三年为主；借贷者持有的房产大多不超过 1 套，且无房产者居多；贷款的验证状态和“initialListStatus ”（贷款的初始列表状态）的分布较为均匀；“purpose ”（借款人在贷款申请时的贷款用途类别）和“pubRecBankruptcies ”（公开记录清除的数量）多为 0；“applicationType ”（表明贷款是个人申请还是与两个共同借款人的联合申请）和 n 系列匿名特征同值化严重，数值高度集中。

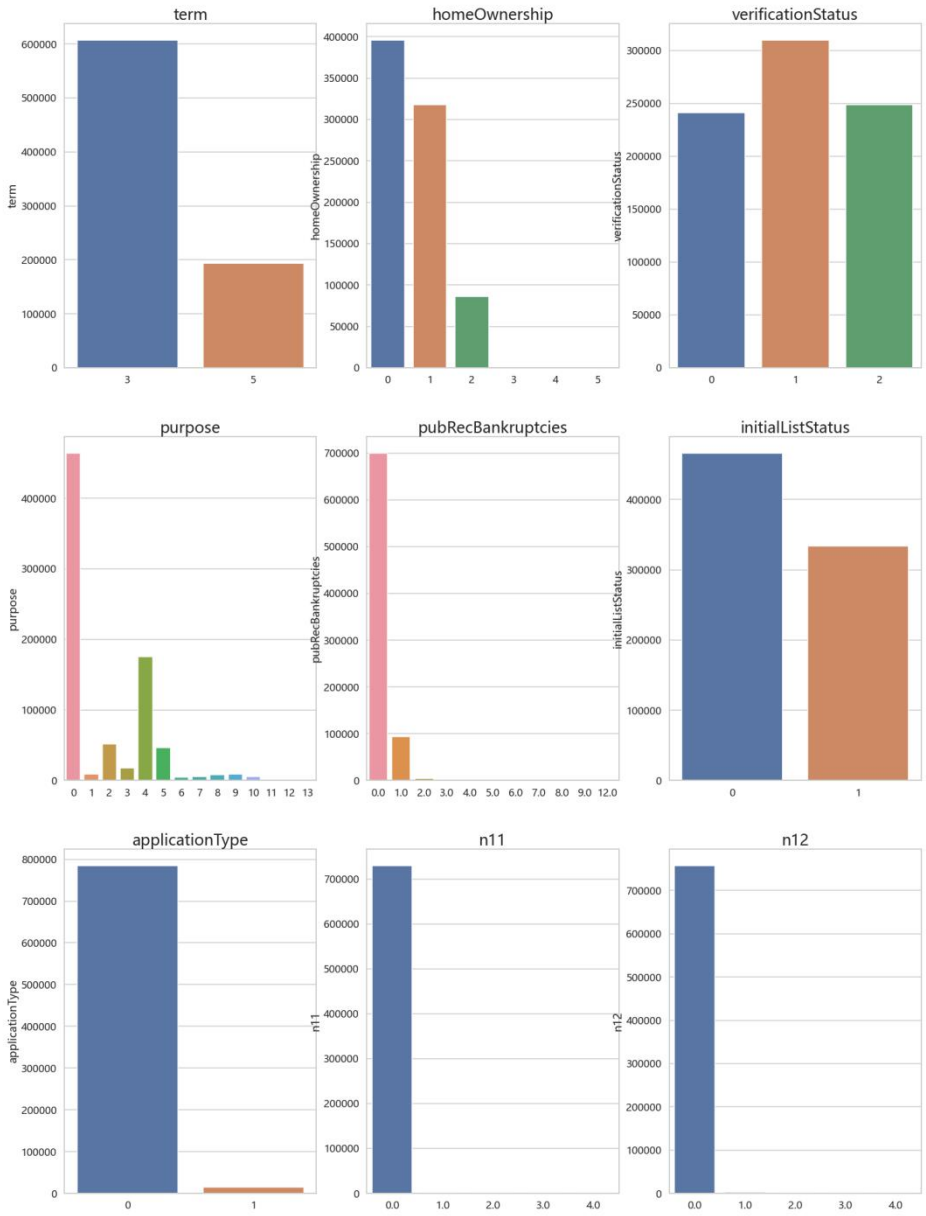


图 7：离散型数值变量的分布情况

3.模型选取

3.1 逻辑回归

逻辑回归，也称 logistic 回归，是一种广义线性回归（generalized linear model），因此与多重线性回归分析有很多相同之处。它们的模型形式基本上相同，都具有 wx+b，其中 w 和 b 是待求参数，其区别在于他们的因变量不同，多重线性回归直接将 wx+b 作为因变量，即 y =wx+b，而 logistic 回归则通过函数 L 将 wx+b 对应一个隐状态 p，p =L(wx+b),然后根据 p 与 1-p 的大小决定因变量的值。如果 L 是 logistic 函数，就是 logistic 回归，如果 L 是多项式函数就是多项式回归。

logistic 回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加

容易解释，多类可以使用 softmax 方法进行处理。实际中最为常用的就是二分类的 logistic 回归。

Logistic 回归模型的适用条件有四个。

第一，因变量为二分类的分类变量或某事件的发生率，并且是数值型变量。但是需要注意，重复计数现象指标不适用于 Logistic 回归。

第二，残差和因变量都要服从二项分布。二项分布对应的是分类变量，所以不是正态分布，进而不是用最小二乘法，而是最大似然法来解决方程估计和检验问题。

第三，自变量和 Logistic 概率是线性关系

第四，各观测对象间相互独立。

Logistic 回归的实质是发生概率除以没有发生概率再取对数。就是这个不太繁琐的变换改变了取值区间的矛盾和因变量自变量间的曲线关系。究其原因，是发生和未发生的概率成为了比值 ，这个比值就是一个缓冲，将取值范围扩大，再进行对数变换，整个因变量改变。不仅如此，这种变换往往使得因变量和自变量之间呈线性关系，这是根据大量实践而总结。所以，Logistic 回归从根本上解决因变量要不是连续变量怎么办的问题。还有，Logistic 应用广泛的原因是许多现实问题跟它的模型吻合。例如一件事情是否发生跟其他数值型自变量的关系。

逻辑回归的常规步骤有三步：

首先，构造预测函数 h(x)。

Logistic 函数（或称为 Sigmoid 函数），函数形式为 $g(z)=\frac{1}{1+e^{-z}}$。对于线性边界的情况，边界形式如下：

$$z=\theta^T x=\theta_0 x_0+\theta_1 x_1+\cdots+\theta_n x_n=\sum_{i=0}^{n}\theta_i x_i$$

其中，训练数据为向量 x，最佳参数 为 θ 。

于是，构造预测函数为：

$$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}}$$

函数 h(x)的值有特殊的含义，它表示结果取 1 的概率，因此对于输入 x 分类结果为类别 1 和类别 0 的概率分别为：

$$P(y=1 \mid x;\theta)=h_\theta(x)$$

$$P(y=0 \mid x;\theta)=1-h_\theta(x)$$

其次，构造损失函数 J（m 个样本，每个样本具有 n 个特征）。

Cost 函数和 J 函数如下，它们是基于最大似然估计推导得到的。

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x_i), y_i) = -\frac{1}{m}\left[\sum_{i=1}^{m}(y_i log h_\theta(x_i) + (1 - y_i)\log(1 - h_\theta(x_i)))\right]$$

最后，想办法使得 J 函数最小并求得回归参数（θ）。一般使用梯度下降法，来不断更新参数 θ，使得损失函数 J 达到最小。

θ更新过程：

$$\theta_j := \theta_j - \alpha\frac{\delta}{\delta_{\theta_j}}J(\theta)$$

$$\frac{\delta}{\delta_{\theta_j}}J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i\frac{1}{h_\theta(x_i)}\frac{\delta}{\delta_{\theta_j}}h_\theta(x_i) - (1-y_i)\frac{1}{1-h_\theta(x_i)}\frac{\delta}{\delta_{\theta_j}}h_\theta(x_i)\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y_i\frac{1}{g(\theta^T x_i)} - (1-y_i)\frac{1}{1-g(\theta^T x_i)}\right)\frac{\delta}{\delta_{\theta_j}}g(\theta^T x_i)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y_i\frac{1}{g(\theta^T x_i)} - (1-y_i)\frac{1}{1-g(\theta^T x_i)}\right)g(\theta^T x_i)(1-g(\theta^T x_i))\frac{\delta}{\delta_{\theta_j}}\theta^T x_i$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y_i(1-g(\theta^T x_i)) - (1-y_i)g(\theta^T x_i)\right)x_i^j$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y_i - g(\theta^T x_i)\right)x_i^j$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x_i) - y_i\right)x_i^j$$

3.2 集成学习

Xgboost 的全称是 eXtreme Gradient Boosting，即极端梯度提升树，是梯度提升机器算法 (Gradient Boosting Machine) 的扩展。Boosting 分类器属于集成学习模型，其基本思想是把成百上千个分类准确率较低的树模型组合成一个准确率较高的模型。该模型不断迭代，每次迭代生成一棵新的树，如何在每一步生成合理的树是 Boosting 分类器的核心。Gradient Boosting Machine 算法在生成每一棵树的时候采用梯度下降的思想，以上一步生成的所有树为基础，向着最小化给定目标函数的方向前进。在合理的参数设置下，需要生成一定数量的树才能达到预期准确率，在数据集较大较复杂的时候，Gradient Boosting Machine 算法的计算量巨大。Xgboost 是 Gradient Boosting Machine 的实现，能自动利用 CPU 的多线程进行并行，并对算法加以改进以提高精度。Xgboost 的基学习器既有树 (gbtree) 又有线性分类器 (gblinear)，从而得到带 L1+L2 惩罚的线性回归或逻辑回归，其损失函数采用二阶泰勒展开，具有高准确度、不易过拟合、可扩展性等特点，能分布式处理高维稀疏特征，因此在同等情况下，Xgboost 算法比同类算法快 10 倍以上。

Boosting 是一种非常有效的集成学习算法，采用 Boosting 方法可以将弱分类器转化为强分类器，从而达到准确的分类效果。其步骤如下所示:

第一步， 将所有训练集样本赋予相同权重;

第二步,进行第 m 次迭代, 每次迭代采用分类算法进行分类, 采用公式 (1) 计算分类的错误率:

$$\mathrm{err}_m = \frac{\sum \omega_i I(y_i \neq G_m x_i)}{\sum \omega_i} \qquad (1)$$

式中 ωi 代表第 i 个样本的权重, Gm 代表第 m 个分类器;

第三步，计算下式

$$\alpha_{\mathrm{m}}=\log\left((1-\mathrm{err}_{\mathrm{m}})\ /\mathrm{err}_{\mathrm{m}}\right)$$

第四步，对于第 m+1 次迭代，将第 i 个样本的权重ω_i重置为

$$\omega_i \times \mathrm{e}^{\alpha_m \times I(y_i \neq G_m x_i)}$$

第五步，完成迭代后得到全部的分类器, 采用投票方式得到每个样本的分类结果。其核心在于每次迭代后, 分类错误的样本都会被赋予更高的权重, 从而改善下一次分类的效果。

Gradient Boosting 是 Boosting 的一个改进版本, 经证明, Boosting 的损失函数是指数形式, 而 Gradient Boosting 则是令算法的损失函数在迭代过程中沿其梯度方向下降, 从而提升稳健性。其算法流程如下所示:

（1）初始化

$$f_0(x) = \mathrm{argmin}_p \sum_{i=1}^{N} L(y_i, \gamma)$$

（2）对于 1-m 次迭代:

$$① F_0(x) = \mathrm{argmin}_\rho \sum_{i=1}^{N} L(y_i, \rho)$$

（3）对于 m=1 到 M

$$\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \ldots, N$$
$$a_m = \mathrm{argmin}_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(X_i; a)]^2$$
$$\rho_m = \mathrm{argmin}_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(X_i) + \rho h(X_i; a_m))$$
$$F_m(X) = F_{m-1}(X) + \rho_m h(X; a_m)$$

(5) 完成迭代后得到全部的分类器, 采用投票方式得到每个样本的分类结果。其核心在于每次迭代后, 分类错误的样本都会被赋予更高的权重, 从而改善下一次分类的效果。

4 实证分析

针对金融风控领域，大部分选取的是机器学习模型中的 Random Forests, lightgbm，xgboost，catboost 等树模型，而不是深度学习模型，原因主要包括：1.样本数量小 2.样本不均衡 3.深度

学习模型对于特定结构的特征学习效果较好（比如文本和图像），而针对具有实际意义的金融领域特征来说，传统树模型构造的可解释性特征效果显著。

机器学习方法能够处理海量数据，一般来说抗噪能力良好，稳健性较强，而 Logistic 回归结构简单，稳健性良好，且回归系数解释能力强，所以本文又采用了逻辑回归的方法，与 XGBoost 对比分析，共同寻找合适的风险指标体系，来预测借款人的违约风险。

回顾本文研究的因变量和解释变量：

因变量：贷款是否违约

解释变量：源数据下清洗后（包括剔除和编码新增的）后保留的变量 48 个，手工合并生成变量 1 个，以及主成分分析降维合并生成的变量 6 个，共计 55 个变量。

4.1 逻辑回归建模

在逻辑回归建模中，本文对样本进行了正则化处理。在阈值为 0.5 的条件下，筛选出权重前 20 的变量，如表 6 所示；图 8 则展示了阈值为 0.5 下的混淆矩阵与 ROC 曲线，模型的正确率 0.5705，精准率为 0.269，召回率 0.6677，调和平均值 F1 为 0.383769。

表 6：逻辑回归的权重前二十的变量

变量	虚拟变量对应的真实变量名	系数	权重
10	purpose	0.27457	0.27457
term		0.25228	0.25228
0	homeOwnership	-0.21168	0.21168
6	verificationStatus	-0.12028	0.12028
1	homeOwnership	0.10632	0.10632
13	purpose	-0.10327	0.10327
initialListStatus		-0.07206	0.07206
n14		0.06132	0.06132
9	purpose	-0.05386	0.05386
delinquency_2years		0.04663	0.04663
18	purpose	0.04181	0.04181
subGrade		0.03125	0.03125
pca2		-0.02767	0.02767
17	purpose	-0.02338	0.02338

n13	-0.01676	0.01676
n6	0.01410	0.01410
pubRec	0.01282	0.01282
dti	0.01109	0.01109
pca4	-0.01085	0.01085
pca5	-0.01071	0.01071

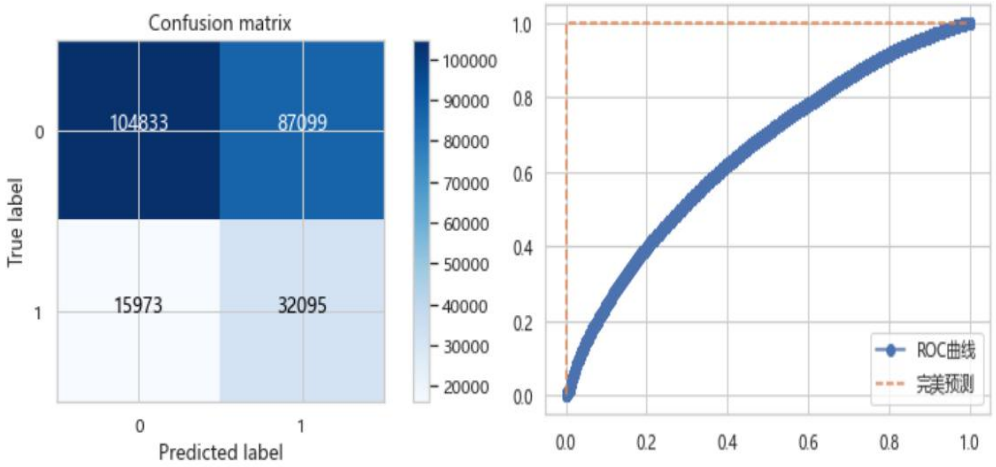


图 8：阈值为 0.5 下的混淆矩阵与 ROC 曲线

对于逻辑回归算法来说，我们还可以指定一个阈值，也就是说最终结果的概率是大于多少我们把它当成是正或者负样本。这里通过 Sigmoid 函数将得分值转换成概率值，默认情况下，模型都是以 0.5 为界限来划分类别：p>0.5 为正例，p<0.5 为负例。0.5 是一个经验值，可以根据实际经验进行调整。我们指定 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9，9 个不同阈值，分别查看模型效果。图 9 展示了 9 个不同阈值下的混淆矩阵。图 10 则展示了正确率、精准率、召回率以及调和平均值 F1 随阈值的变动情况。

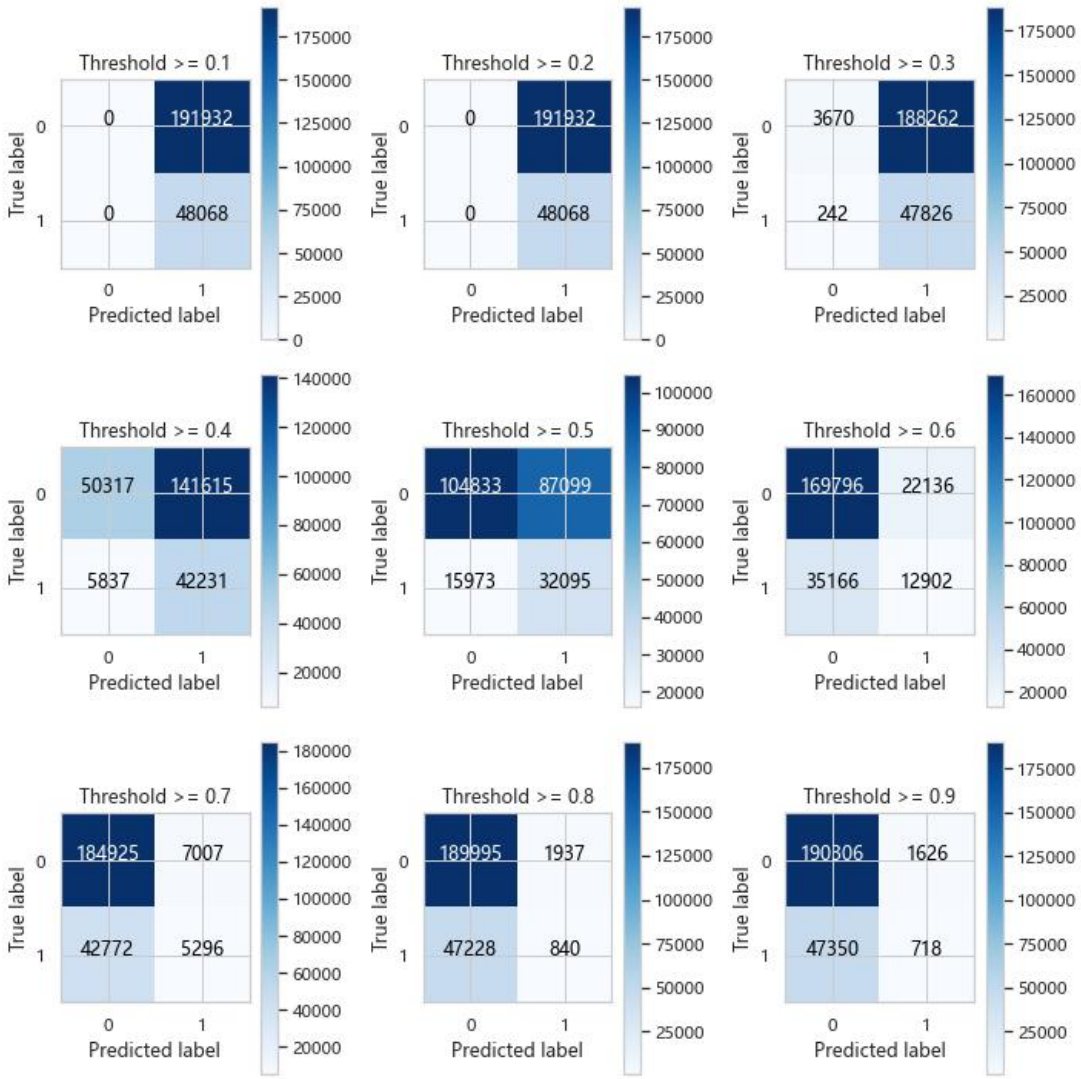


图 9：不同阈值下的混淆矩阵

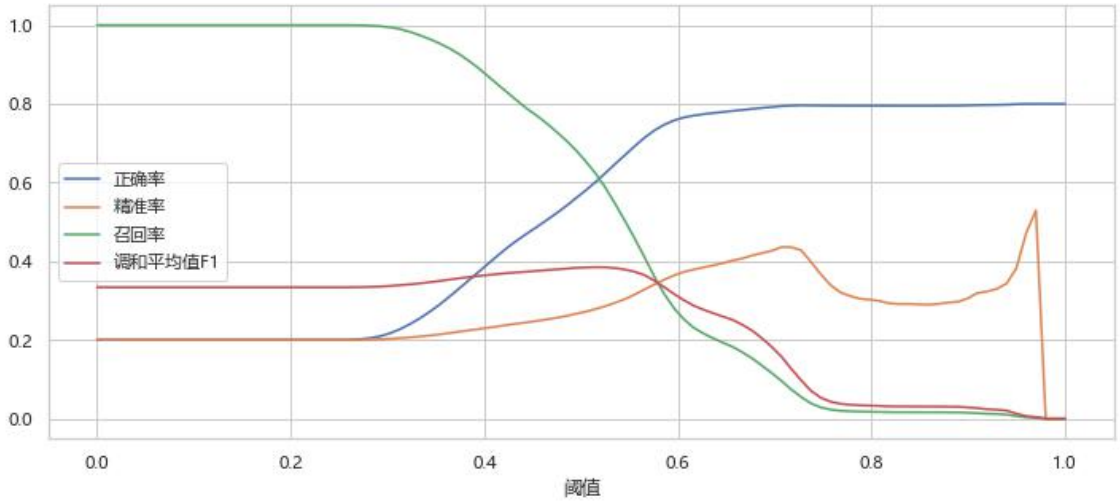


图 10：不同阈值下的得分折线图

4.2 XGBoost 建模

由表 7 可以看出，XGBoost 的正确率显著高于阈值为 0.5 下的逻辑回归。图 11 则展示了 XGBoost

的混淆矩阵。可见，在此情景下，XGBoost 的性能要显著优于逻辑回归。

表 7：XGBoost 和阈值为 0.5 下的逻辑回归得分对比

指标	XGBoost	逻辑回归
正确率	0.8042	0.5880
精准率	0.5579	0.2864
召回率	0.1077	0.7087
调和平均值 F1	0.1805	0.4080

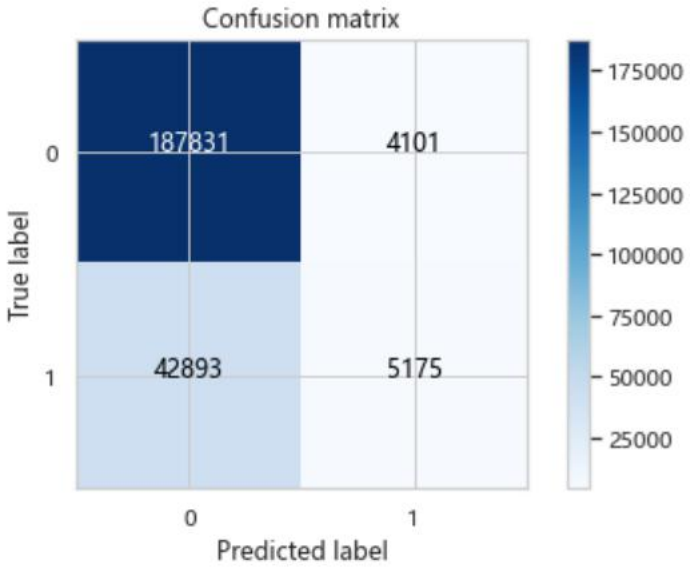


图 11：XGBoost 的混淆矩阵

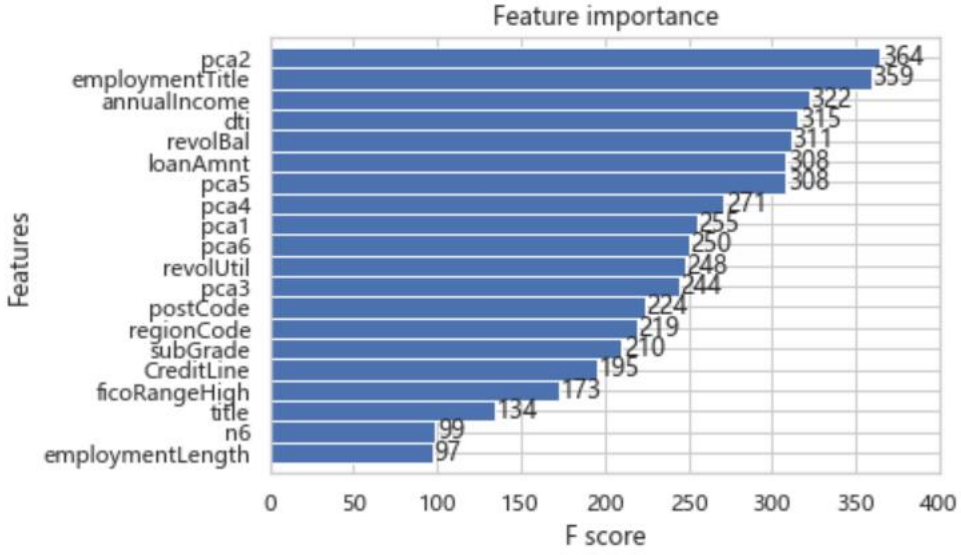


图 12：XGBoost 的权重前 20 的变量

图 12 展示了 XGBoost 筛选出的 20 个最重要的变量。可见工作职称、年收入、债务收入比、信

贷周转余额合计、借款人在贷款申请中提供的邮政编码的前 3 位数字、地区编码等变量以及我们经过主成分分析法合并产生的特征对贷款违约率的影响最大。

5.总结与反思

5.1 论文总结

本文使用了某信贷平台提供的贷款顾客数据，根据用户的社会人口学变量、交易信息以及 n 系列匿名行为特征，对顾客贷款违约的可能性进行预测，并希望通过分析筛选出的显著影响因素，对信贷平台审慎发放贷款，预防违约风险提出具体建议。

本文首先介绍防范贷款违约风险，对提升银行竞争力，增强银行盈利能力的重要性；其次以收集到的数据集为案例，对样本数据进行了一系列的统计描述，展现变量弱相关和分布差异大的特点；之后基于数据的特点，先后建立逻辑回归和 XGBoost 模型，最终实现了对目标变量的准确预测，并找到影响顾客是否流失的关键因素。

相比之下，用户的社会人口学特征比用户的交易信息对准确预测客户的动向更有帮助。在社会人口学变量方面，工作职称、年收入、债务收入比、地区等因素影响显著；交易信息方面，信贷周转余额合计以及借贷金额影响较大。而 n 系列匿名行为特征的预测影响不大。

基于上述分析结果，本文提出以下建议与意见：

关注客户的收入实力。本文注意到高工作职称和年收入的顾客，贷款违约率更低，背后一大原因在于其收入的稳定和丰厚降低了贷款违约的可能。

关注客户的年龄。年轻人的消费欲望更强，资金周转可能更为频繁；中年人逐渐步入事业的顶峰，账户的存款余额比较稳定。

关注客户的资产结构。收入只是一方面，客户资产结构对于其经济状况得分波动情况有很大的影响。要密切注意客户的债务收入比，客户过高的杠杆意味着更大违约的风险。

关注交易信息。比如说，更长期的贷款，顾客违约的风险会更大，因为更长的时间会带来更大的不确定风险，对于长期贷款，借贷平台应该审慎发放。

5.2 反思与不足

尽管本研究取得了一定的成果，本报告得出的结论在许多方面仍值得进一步的推敲和探究。具体如下：

受限于硬件条件限制，未能对逻辑回归的正则化惩罚力度参数进行很好的调参，所以制约了文中逻辑回归模型的性能。

筛选出的特征对因变量的具体影响方式未知。考虑到用户的交易行为具有复杂性，解释变量和预测变量并不具备线性可分的性质，采用 XGBoost 模型对目标变量进行预测具有很好的性能。但是，这样却导致筛选出的特征虽然和目标变量显著相关，但是具体的作用机制未知，例如，不能准确判断特征对因变量是正向影响还是负向影响，因此对具体影响方式的分析存在一定的主观性，可以做进一步的探究。

参考文献

[1]丁正斌、施建军，2017，《住房按揭贷款逾期风险及其管理探析》，《审计与经济研究》第 1 期，105-111。

[2]高广春，2017，《个人住房抵押贷款还有多大增长空间》，《银行家》第 12 期，49-52。

[3]何晓晴、谢赤、吴晓，2005，《住房按揭贷款违约风险及其防范机制》，《社会科学家》第 6 期，65-67。

[4]胡金焱、宋唯实，2018，《借贷意愿、融资效率与违约风险-网络借贷市场参与者的性别差异研究》，《东岳论丛》第 3 期，27-34。

[5]胡颖、谢芳，2009，《商业银行个人住房抵押贷款违约风险研究》，《经济前沿研究》第 8 期，49-55。

[6]胡晏，2017，《信用等级、借款成功率与违约风险-基于“拍拍贷”数据的经验证据》，《投资研究》第 8 期，143-158。

[7]况伟大，2014，《中国住房抵押贷款拖欠风险研究》，《经济研究》第 1 期，156-168。

[8]马宇，2009，《我国个人住房抵押贷款违约风险影响因素的实证研究》，《统计研究》第 5 期，101-107。

[9]欧阳远芬，2014，《银行住房贷款违约率的宏观风险分析-基于 MVAR 模型的实证研究》，《投资研究》第 33 期，4-11。

[10]平新乔、杨慕云，2009，《消费信贷违约影响因素的实证研究》，《财贸经济》第 7 期，32-39。

[11]任兆璋、杨绍基，2006，《商业银行信贷违约风险测度的 SBP 模型研究》，《金融研究》第 11 期，130-137。

[12]舒扬、杨秋怡，2017，《基于大样本数据模型的汽车贷款违约预测研究》，《管理评论》第 7 期，60-72。

[13]王福林、贾生华、邵海华, 2005, 《个人住房抵押贷款违约风险影响因素实证研究-以杭州市为例》, 《经济学》季刊第 5 期, 739-750。

[14] 王吉恒、王思祺, 2017, 《村镇银行的信贷风险及防范措施研究》, 《中国经贸》第 5 期, 9-16。

[15]吴姗姗, 2018, 《个人住房抵押贷款违约风险研究评述和展望》, 《经济师》第 3 期, 19-22。

[16]Arminger G, Enache D, Bonne T, 1997, "Analyzing Credit Risk Data:A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feed forward Networks", Social Science Electronic Publishing, 12 (2) , pp.293-310.

[17]Arya S, Eckel C, Wichman C, 2013, "Anatomy of the credit score", Journal of Economic behavior and organization, 95, pp.175-185.