欧亚学院 eurasia UNIV.

19-20（1）

《金融科技实践》

课程期末报告

学　　院　金融学院

班　　级　统本大数据 1802

题　　目　征信数据分析实践

姓　　名　刘嘉玲

时　　间　2020.6.18

目录

征信数据分析实践.....1
一、背景介绍.....1
二、数据获取方式及说明.....2
三、描述性分析.....2
（一）因变量：是否按期还款.....3
（二）自变量：基于连续变量的描述分析.....3
（三）自变量：基于连续变量的描述分析.....5
（四）自变量：基于所有自变量之间的相关性分析.....6
四、变量筛选.....7
（一）分箱.....7
（二）对定量变量分箱处理.....8
（三）图形化展示定性变量.....10
（五）对变量进行筛选.....12
五、模型建立.....12
（一）逻辑回归模型.....13
（二）模型预测.....15
（四）建立打分卡.....16
六、结论与建议.....17
（一）商业应用.....17
（二）建议.....18
（三）未来研究方向.....19

七、参考文献.....19

征信数据分析实践

摘要：面向小微商户以及个人消费的小微信贷是当前互联网金融的重要发展方向，并且正在经历爆发式增长。在这个增长过程中，如何在没有实物抵押的情况下，通过互联网大数据分析实现快速准确征信是一个非常重要的问题。为此，不同的数据来源将各显神通地为信用评估提供依据。本文将通过一个真实的案例出发，进行分析和探讨，针对用户历史行为数据建立信用评分模型，并通过该模型改进信用评估的预测效果。

本报告主题以征信公司客户数据为主要对象，对客户的相关数据使用 0-1 变量回归分析，研究客户性别、已婚_未婚、已育_未育，收入等多项数据对客户留存情况影响，建立回归模型，能够刻画移动通信客户的行为特征，并以此预测客户是否按期还款来更好地为公司提供决策，规避风险。模型检验证明，该方法实用、可操作性强对支持企业客户关系管理产生了积极的影响。

关键词：互联网征信；逻辑回归；打分卡。

一、背景介绍

征信是指对企业组织和个人的信用信息进行采集、整理、保存和加工，并向信息使用者提供的活动，其本质在于利用信用信息对金融主体进行数据刻画。征信是现代金融体系的基础设施。征信本身不创造信用，却是信用活动乃至整个经济金融体系征信的基石。现代金融体系中，征信的作用在于利用数据对每个金融主体进行刻画和信用评估，进而激发金融主体间的潜在融资需求，并支撑起总体融资规模的扩大。因为征信机构承担了部分信用风险管理的职能，金融机构的中介属性将会弱化，整个金融体系的交易成本有望降低。征信行业的竞争越来越激烈，如何把控金融流通，降低信贷风险，成为征信公司保证稳定收入的一个重要因素。

二、数据获取方式及说明

本案例共 1000 条数据，数据共包括 9 个变量，1 个因变量——是否按期还款，8 个自变量——基本信息方面：性别、已婚_未婚、已育_未育，收入，学习能力方面：教育水平、英语水平，人脉方面的微博好友数和消费理念。该数据来自某征信网的用户基本信息，网上简历数据和用户行为数据。

表 2-1 数据变量说明表

变量类型		变量名	详细说明	取值范围	备注
因变量		是否按期还款	定性变量（2 水平）	0 代表未按期还款；1 代表按期还款。	按期还款率占比 66.7%
自变量	基本信息	性别	定性变量（2 水平）	0 代表女性；1 代表男性。	女性占比 50.3%
		已婚_未婚	定性变量（2 水平）	0 代表单身；1 代表已婚。	单身占比 50.3%
		已育_未育	定性变量（2 水平）	0 代表未育；1 代表已育。	未育占比 51.2%
	收入水平	收入	连续变量呈左偏分布	426~120940	单位：元
	学习能力	教育水平	定性变量（4 水平）	1 代表高中及以下；2 代表大专或本科；3 代表硕士研究生；4 代表博士研究生及以上.	大专或本科占比 33.5%
		英语水平	定性变量（4 水平）	1 代表四级以下；2 代表四级；3 代表六级；4 代表六级以上。	四级占比 32.9%
	社交人脉	微博好友数	连续变量呈左偏分布	6~114	单位：人
	消费理念	消费理念	连续变量呈右偏分布	0~1	消费理念=信用卡消费/总消费

三、描述性分析

在对个人征信数据的是否按期还款进行模型探究之前，首先对各变量进行描述性分析，通过 RStudio 以初步判断是否按期还款的影响因素，为后续建模做铺垫。

（一）因变量：是否按期还款

在本案例中，一位已育的单身女性，学历为高中以下，英语水平六级，微博好友数为36位，消费理念行用卡的使用率为0.1，收入为19823元，是按期还款的。其中1代表按期还款，0代表未按期还款。

通过图3-1按期还款的情况饼图，可以看出：蓝色部分为按期还款的约占被记录人数的67%，还有33%红色部分的用户未能按期还款，为了公司的稳定利益，研究评估影响客户是否按期还款的行为预测尤为重要。

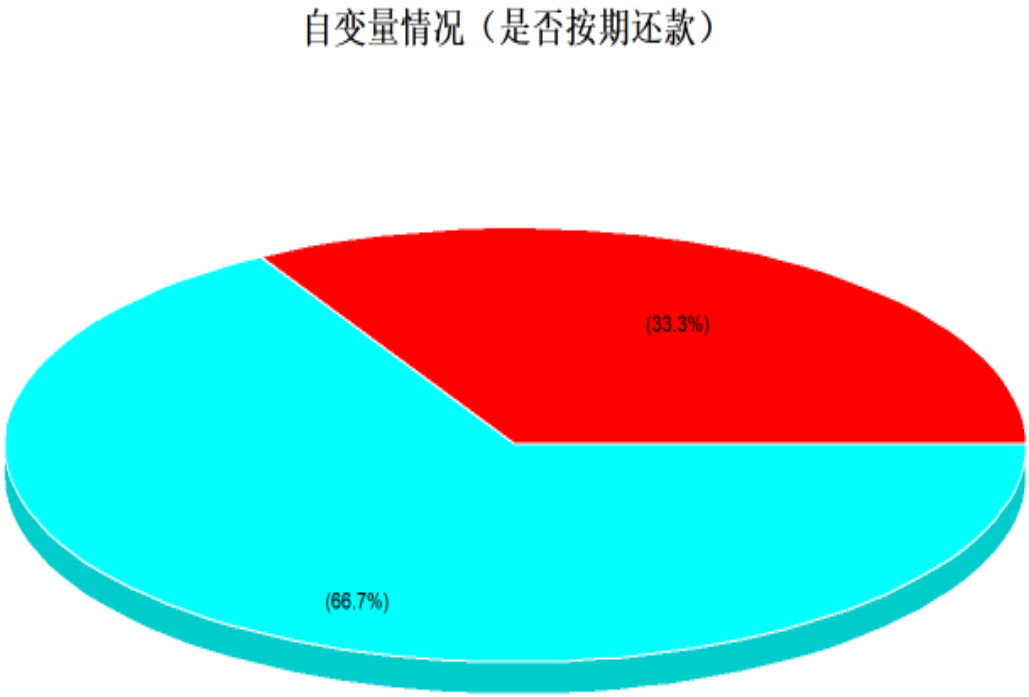

图3-1 按期还款情况的饼图

（二）自变量：基于连续变量的描述分析

通过图3-2收入水平、社交人脉和消费理念的直方图，可以看到：用户收入和微博好友数量都是呈左偏正态分布的；这一现象说明社会阶层和圈子的不同与收入是相互影响的。关于消费理念，基于对信用卡的使用率来看，是右偏的，呈递减的趋势，这一现象说明只有少数人是忠于信用卡的。

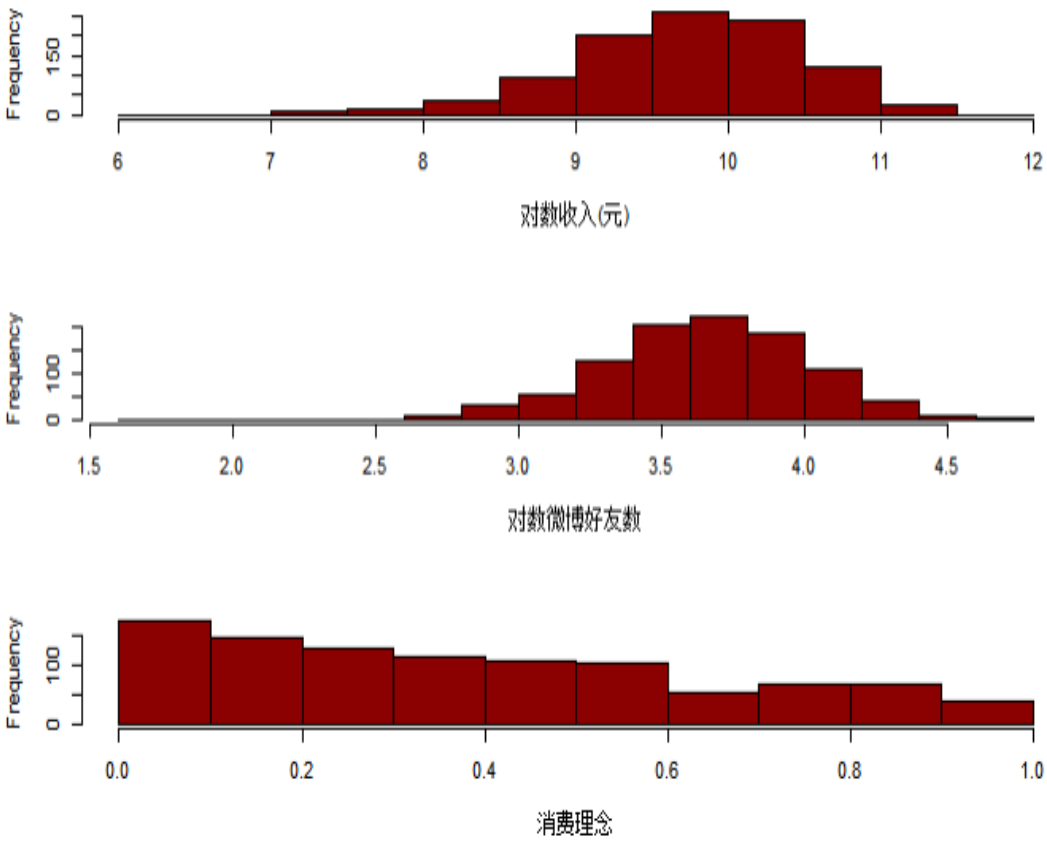

图 3-2 收入水平、社交人脉和消费理念的直方图

通过图 3-3 用户的收入水平、社交人脉和消费理念与是否按时还款的箱线图，可以看到：能按期还款用户的平均收入水平要高于未按期还款的，微博平均好友数多的用户和平均消费理念高的用户无法按期还款可能性会更高一些，可能是社交能手对时间观念不是特别强烈的缘故吧，经常大概率使用信用卡的用户出现未按期还款的情况明显高于不是很依赖信用卡的用户。

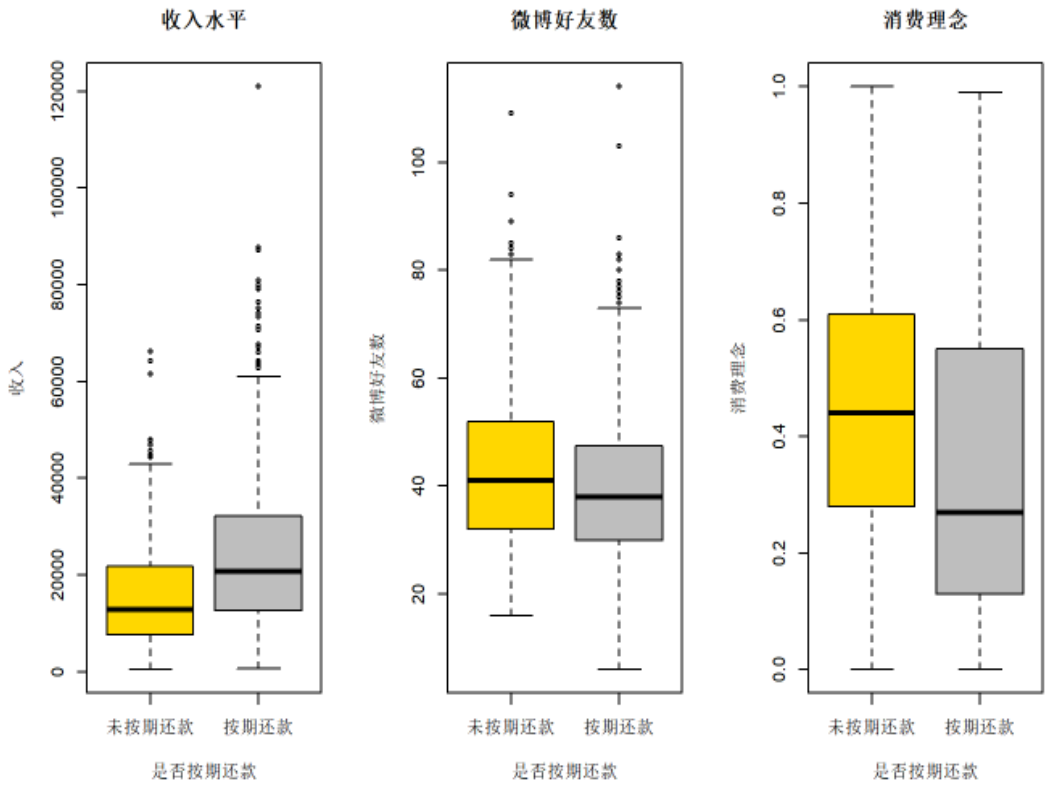

图 3-3 用户的收入水平、社交人脉和消费理念与是否按时还款的箱线图

（三）自变量：基于连续变量的描述分析

图 3-4 表示用户的基本信息、学习能力与是否按时还款的 5 个荆棘图。由图 1 得出男性未按期还款的比女性多；由图 2 得单身的未按时归坏率高于已婚的；由图 3 得未育的年轻人往往会比已育的人群未按期还款的情况多。图 4 和图 5 的教育水平和英语水平影响因素相似，知识越多，逾期的情况就越少，同人们的素质有很大的关系。其中教育水平为高中及以下和大专或本科对是否守时影响一致，没有明显区别，同时也提醒社会教育引起反思，高等教育的价值和意义该如何提升等系列问题。

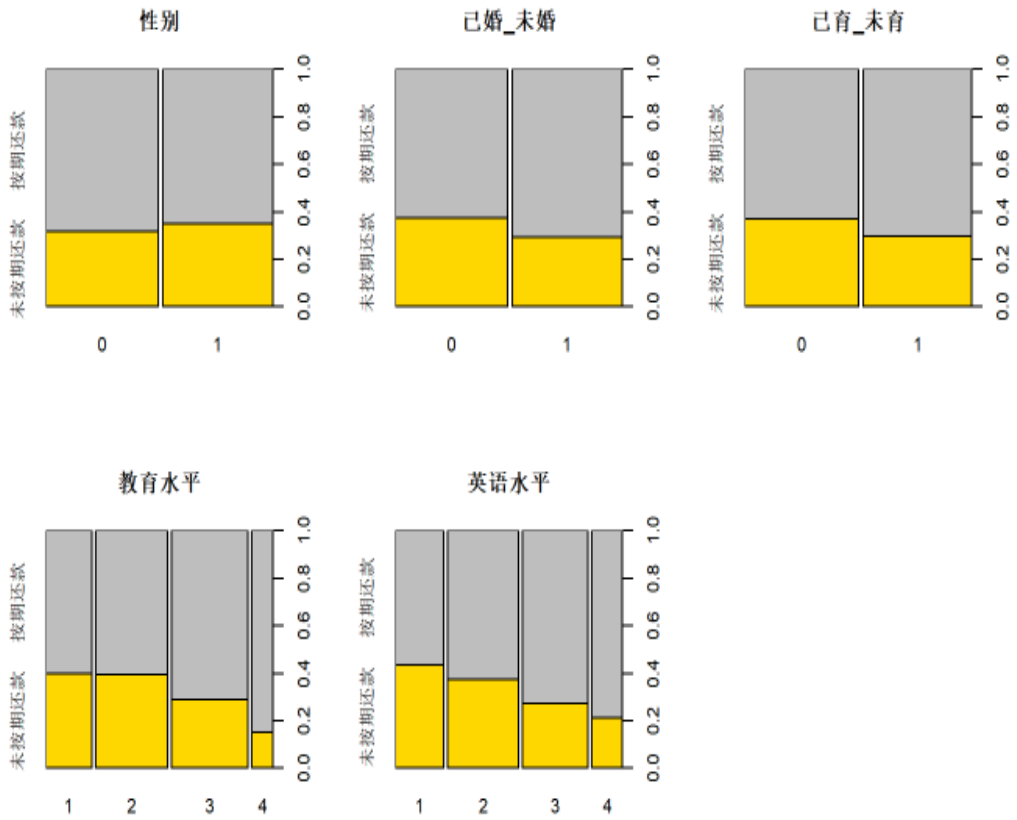

图 3-4 用户的基本信息、学习能力与是否按时还款的荆棘图

（四）自变量：基于所有自变量之间的相关性分析

图 3-5 所有解释型变量盒状图可以看出，为了综合评价用户行为影响因素的整体表现，本案例对数据中的 8 个自变量，进行因子分析。可以明显看到，部分变量的线性相关性较强，例如，收入与消费理念有较强的相关性。为我们建立逻辑回归模型和变量选择提前做预测。我们还发现收入水平和学习能力对按期还款有着积极正向的作用，而社交能力和消费观念对按期还款是负相关。

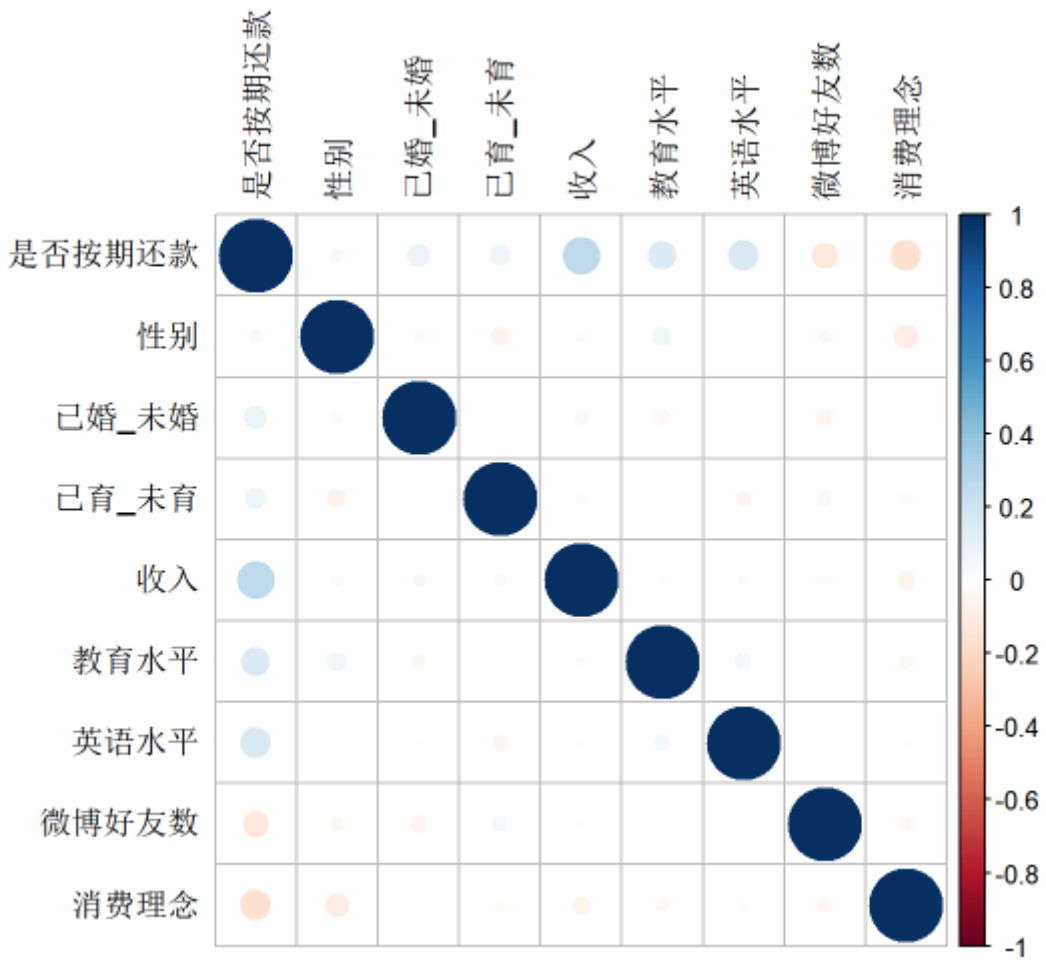

图 3-5 所有解释型变量盒状图

四、变量筛选

为了更准确预测各因素对是否按期还款行为的影响，我们通过使用 WOE 证据权重对原始自变量重新分组并编码。再计算其 IV 值，完成变量筛选。

（一）分箱

由图 4-1 分箱概览图我们得出收入这个自变量的响应比例最大，消费理念和微博好友数中等。

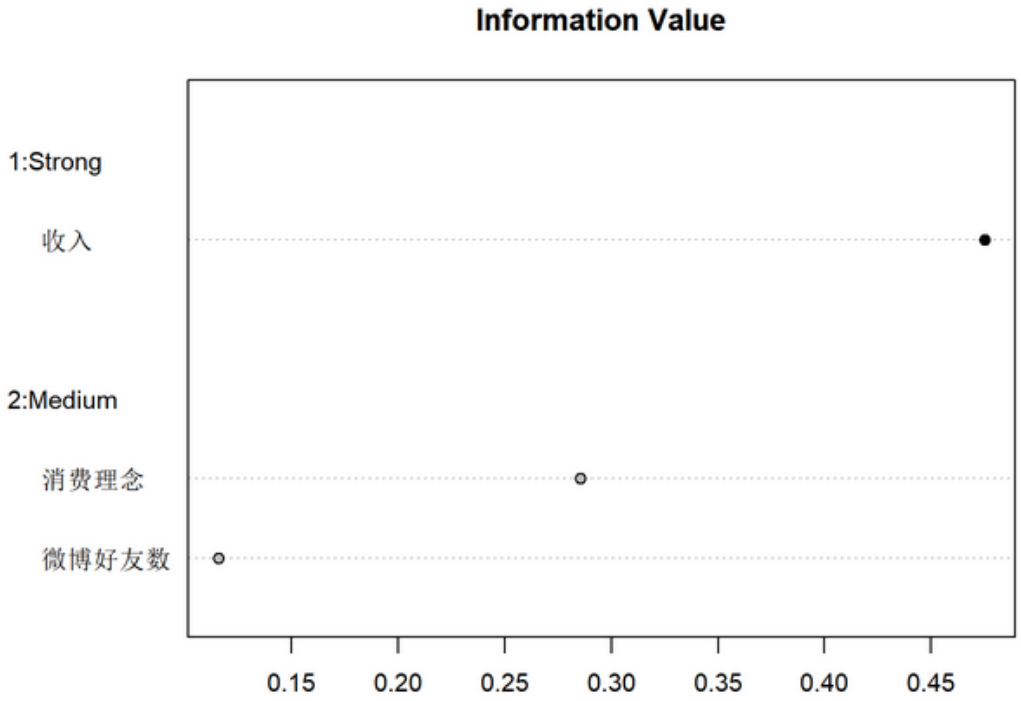

图 4-1 分箱概览图

（二）对定量变量分箱处理

由图 4-2 分箱后收入图，我们得出收入对按期还款的响应显著。收入高低俩端和中间响应的比例相差较大，IV 值会很大吧，其中正向影响跟多一些，收入越高的群体正影响越大。

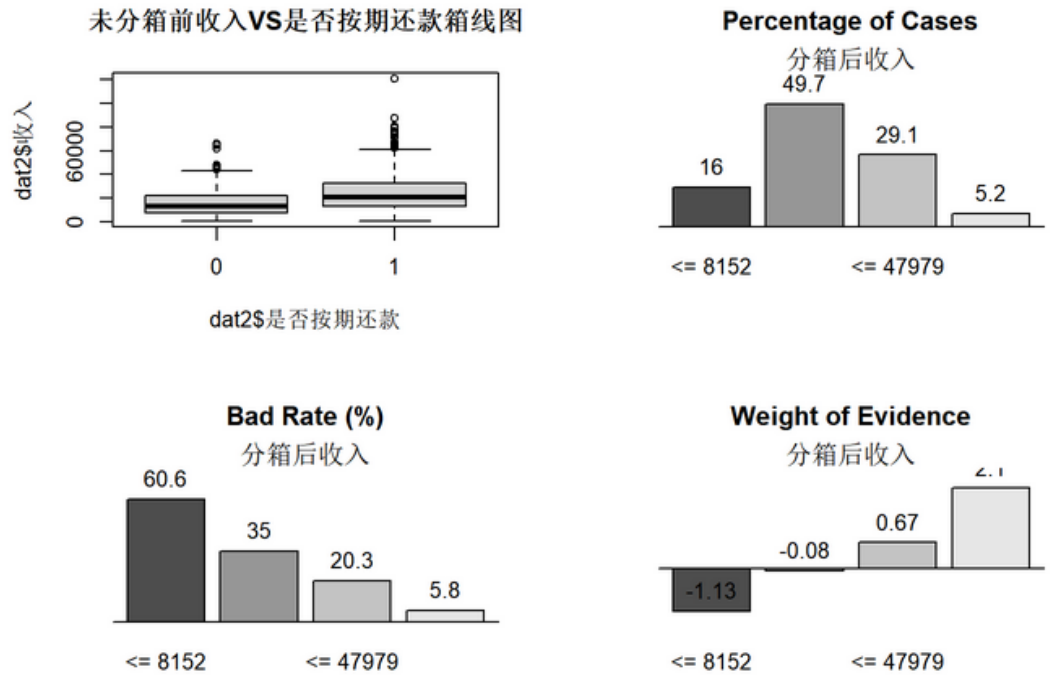

图 4-2 分箱后收入

由图 4-3 分箱后微博好友数，我们得出社交能力对按期还款的响应显著。微博好友数响应的比例相差较大，IV 值不会很低吧。微博好友少于 25 个人数呈现正影响。

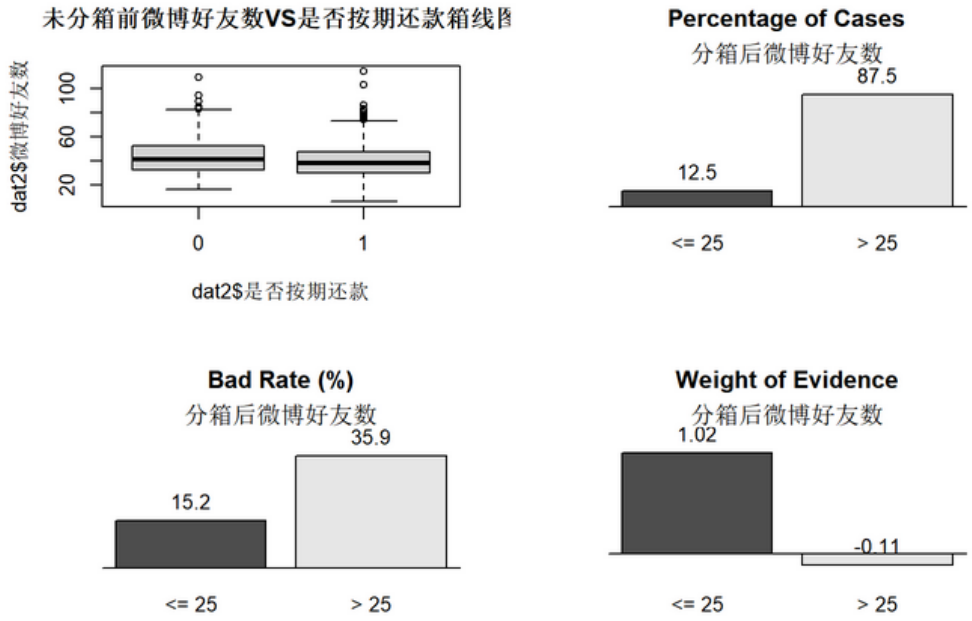

图 4-3 分箱后微博好友数

由图 4-4 分箱后消费理念，我们得出消费理念对按期还款的响应显著。消费理念低于 0.27 的呈现正影响。

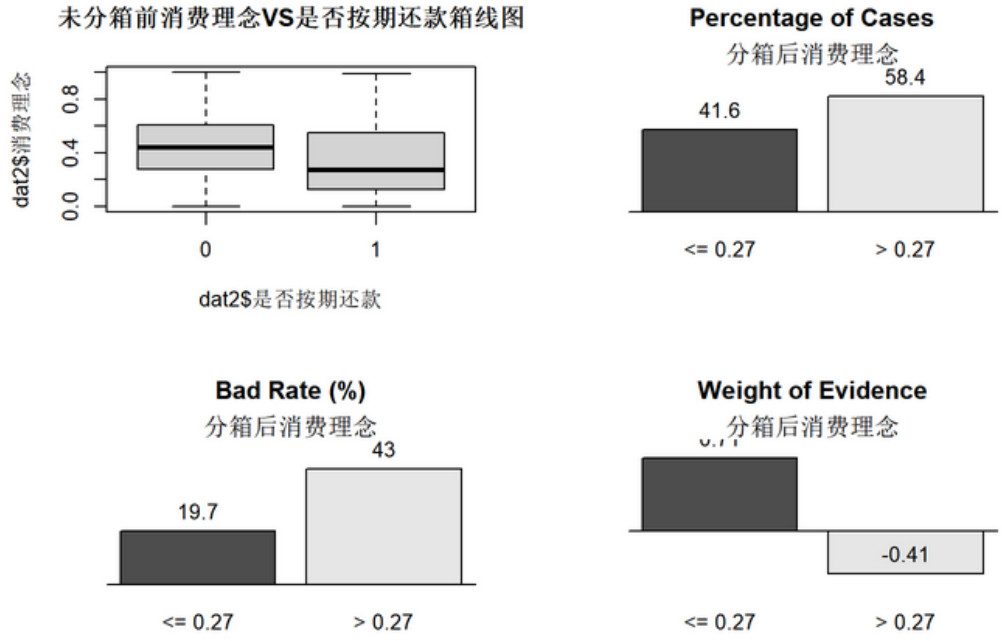

图 4-4 分箱后消费理念

（三）图形化展示定性变量

由图 4-5 分箱后性别，我们得出性别对是否按期还款没有明显响应。女性呈现正影响。

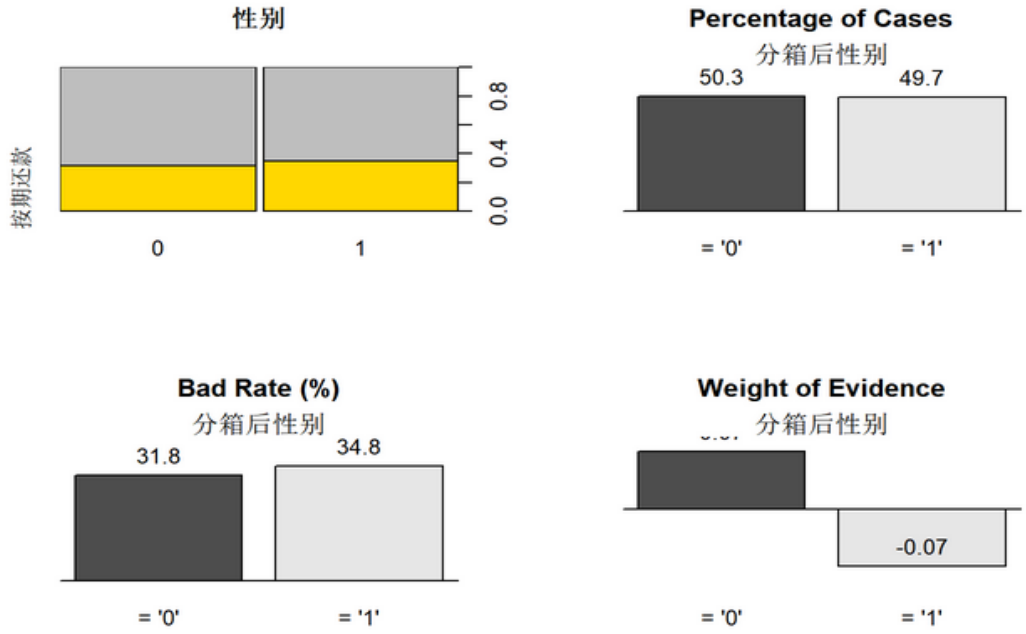

图 4-5 分箱后性别

由图 3-6 分箱后已婚-未婚，我们得出已婚-未婚对是否按期还款没有明显响应。已婚未婚方面已婚呈现正影响。

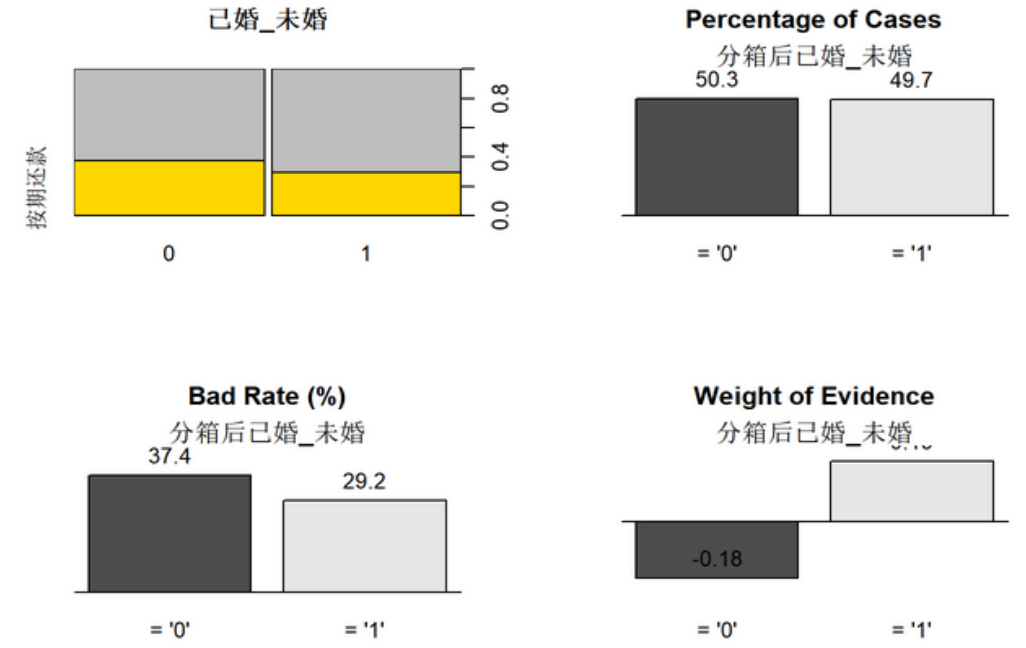

图 4-6 分箱后婚配情况

由图 4-7 分箱后已育-未育，我们得出已育-未育对是否按期还款没有明显响应。已育和未育方面已育呈现正影响。

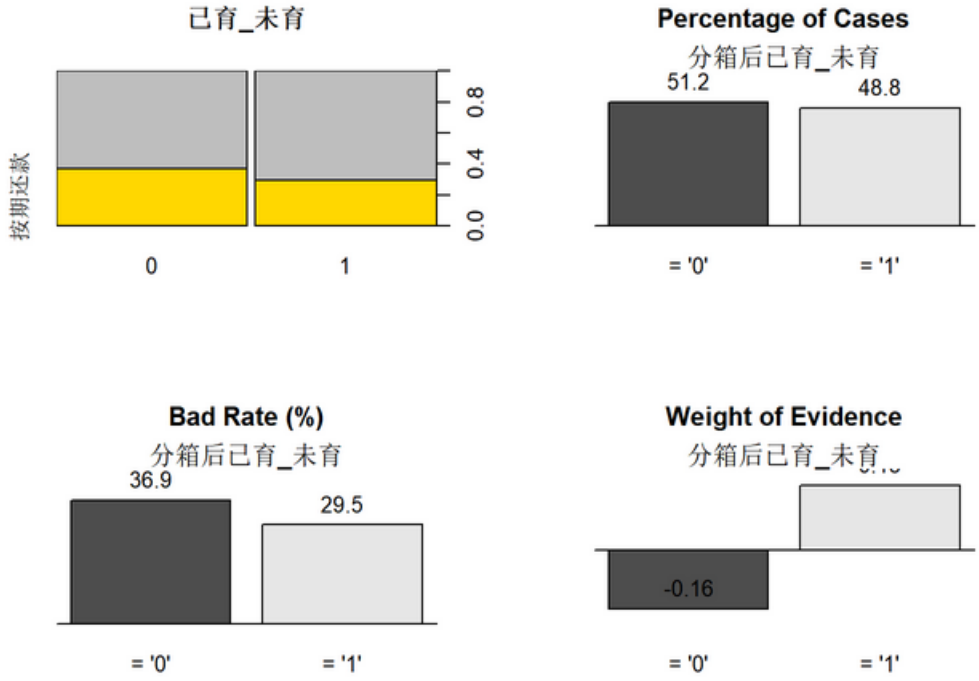

图 4-7 分箱后生育情况

由图 4-8 分箱后教育水平，我们得出教育水平对是否按期还款有明显的响应。随着其水平的提高呈现正影响。

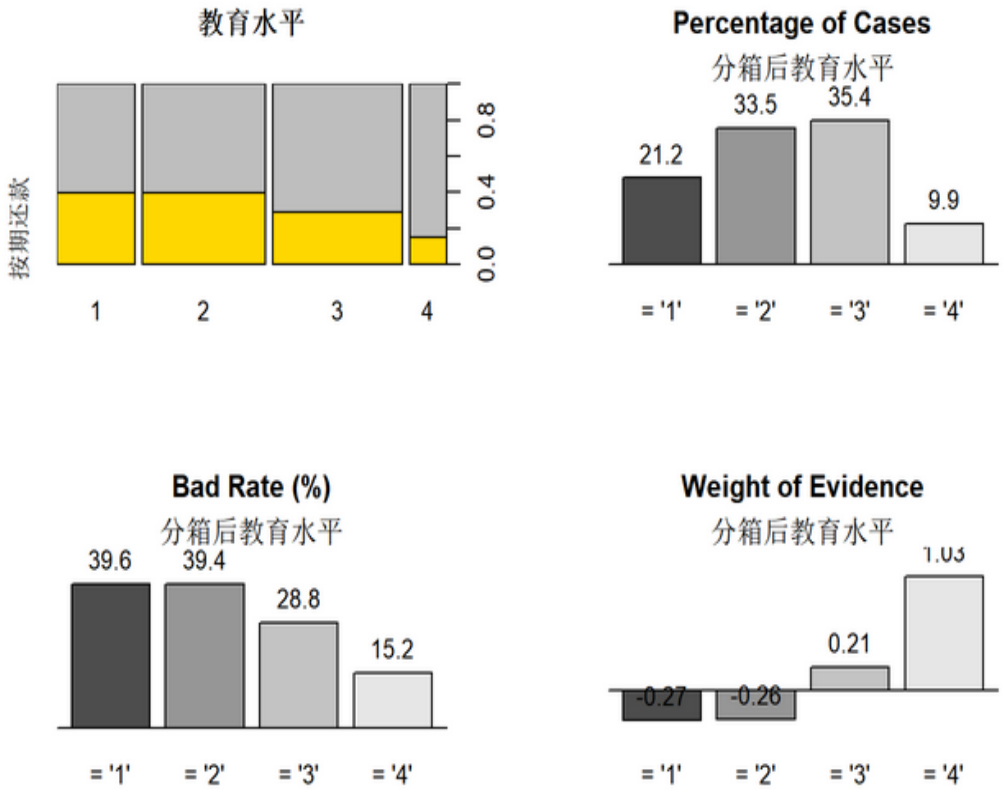

图 4-8 分箱后教育水平

由图 4-9 分箱后英语水平，我们得出英语水平对是否按期还款有明显的响应。随着其水平的提高呈现正影响。

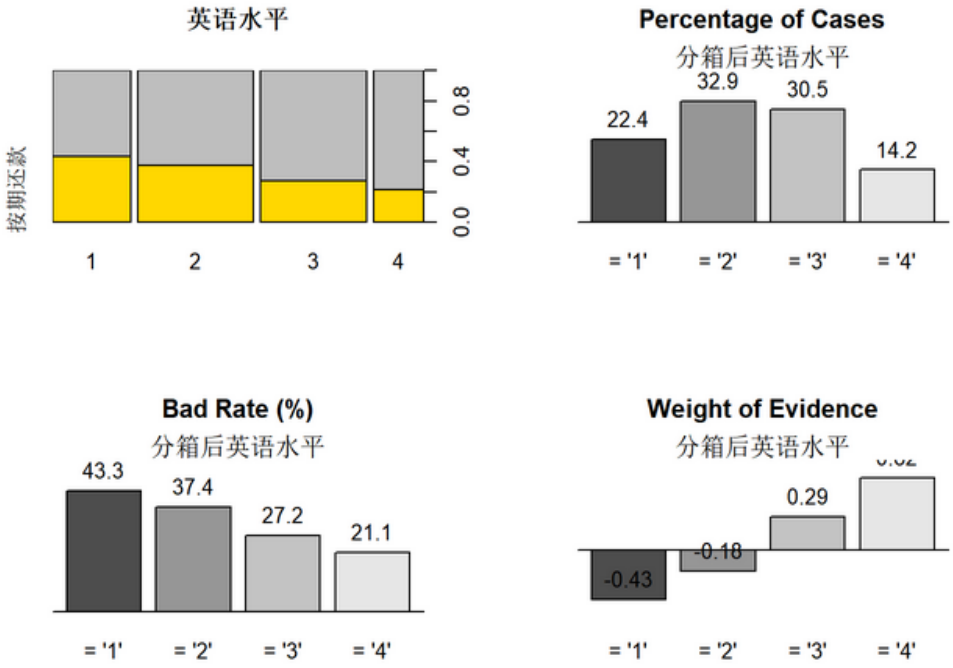

图 4-9 分箱后英语水平

（五）对变量进行筛选

我们通过使用 IV 的方法，来衡量变量的预测能力，值越大，表示此变量的预测能力越强。

表 4-10 变量 IV 值

变量名	性别	已婚-未婚	已育-未育	收入	教育水平	英语水平	微博好友数	消费理念
IV 值	0.046	0.0342	0.0279	0.4750	0.139	0.139	0.1159	0.2855
判断	无预测能力	低无预测能力	无预测能力	高	中	中	中	中

由上表 4-10 变量 IV 值我们对变量进行筛选，最后决定留下收入，教育水平，英语水平，微博好友数和消费理念这 5 个变量。

五、模型建立

为了更准确预测各因素对是否按期还款行为的影响，本次分析报告将建立是否按期还款预测关于自变量和因变量的逻辑回归分析模型，使用 ROC 曲线的方式确定最佳阈值，

并绘制混淆矩阵，试图通过该模型来预测用户是否按期还款情况，为公司提供决策参考。

（一）逻辑回归模型

首先，利用征信数据建立简单的自变量逻辑回归全模型，结果如表 5-1 所示。

表 5-1：自变量建模和 AIC 准则变量选择

变量	逻辑回归模型		AIC 准则		备注
	回归系数	显著性	回归系数	显著性	
截距项	0.343	0.404	0.240	0.551	
性别男	-0.180				基准组：女性
已婚_未婚.已婚	0.418	**	0.420	**	基准组：未婚
已育_未育.已育	0.371	*	0.380	*	基准组：未育
收入.中	0.905	***	0.924	***	基准组:收入较低
收入.较高	1.600	***	1.612	***	
收入.高	3.139	***	3.149	***	
教育水平.大专或本科	-0.093		-0.102		基准组:高中及以下
教育水平.硕士研究生	0.451	*	0.430	*	
教育水平.博士研究生及以上	1.224	***	1.214	***	
英语水平.四级	0.390	.	0.379	.	基准组:四级以下
英语水平.六级	0.841	***	0.840	***	
英语水平.六级以上	1.075	***	1.070	***	
微博好友数	-1.098	***	-1.090	***	社交人脉
消费理念	-1.083	***	-1.075	***	消费理念
模型显著性检验	P<0.001				

注：***0.001 显著；**0.01 显著；*0.05 显著；.0.1 显著

模型解读：

在控制其他因素不变时，可以得到如下结论：

- 对于用户基本信息的相关变量，性别对是否按期还款并没有显著性；是否已婚和是否

生育与因变量呈现正相关，自变量增加，按时归还可能性会更大。相比之下，婚否对征信的影响比生育更显著。

- 用户的收入水平对是否按期还款都是正的强相关。收入水平越高，按时归还可能性会更大。教育水平只有硕士研究生及以上比较显著，本科及以下的并没有多少影响。英语水平四级及以上的都比较显著。
- 用户社交人脉管理的微博好友数以及用户消费理念对是否按期还款是负的强相关。好友数增加，消费理念增加都会导致为按时归还的可能性增加。

该 0-1 回归模型总的来说，用户婚育情况、收入水平和学习能力的值越大，按时还款率越高。

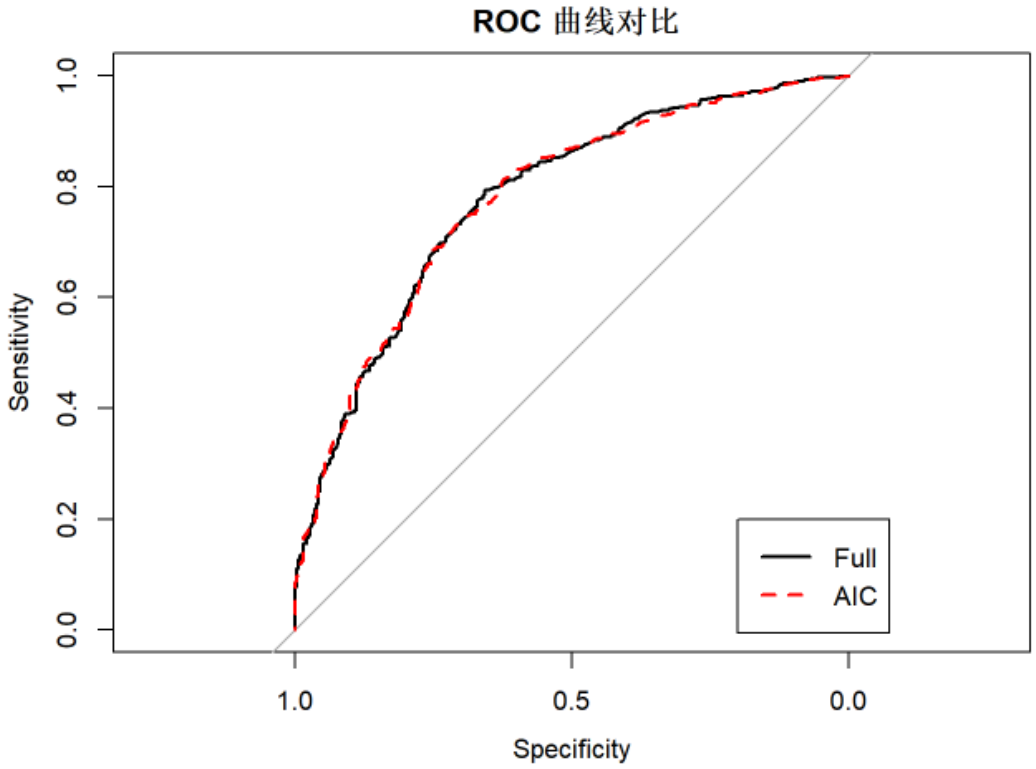

图 5-2：模型的 ROC 曲线对比

基于模型对变量选取的诊断，我们根据表中极大似然比检验，因为得出的结论一致，我们选取 AIC 检验的模型，故得绘制了 AIC 模型的 ROC 曲线并计算其 AUC 的值。

故我们得出以下模型式子：

是否按时还款 = 0.240 + 0.420 * 婚否 + 0.380 * 生育情况 + 0.924|(0.430)|(3.149) * 收入水平 + (-0.102)|(0.430)|(1.214) * 教育水平 + (0.379)|(0.840)|(1.070) * 英语水平 - 1.090 * 微博好友数 - 1.075 * 消费观念

（二）模型预测

图 4-6 ROC 曲线展示了该模型下的最佳阈值。

- 预测的最佳阈值：P=0.646

  预测概率>=P，预测为按时还款

  预测概率<P，预测为未按时还款

最佳阈值的选择标准：平衡 TPR 和 FPR

在实际数据分析中，也可考虑使用样本流失率作为阈值。

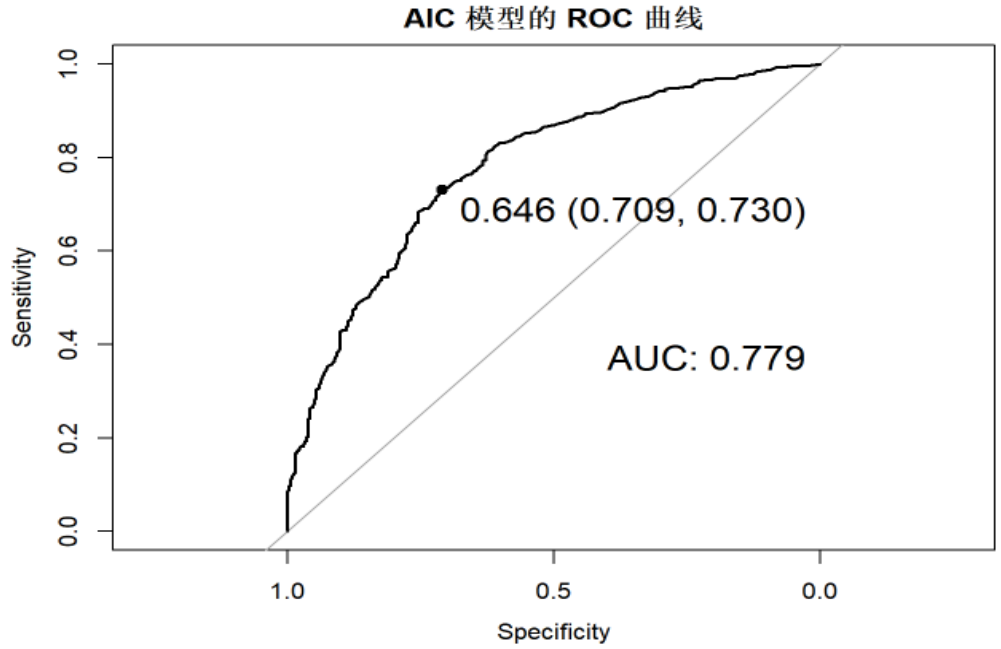

图 5-3：AIC 模型的 ROC 曲线

AUC 的取值为 0.779。说明该模型的预测效果不错，正确率较高，有很强的区分度。

表 5-4 混淆矩阵展示了在该模型预测下的情况。

表 5-4：混淆矩阵

混淆矩阵		预测值		总计
		0	1	
真实值	0	236	97	333
	1	180	487	667
总计		416	584	1000

注：0 为未按期还款，1 为按期还款。

根据 ROC 曲线选取阈值为 0.646。

整体错判率：27.7%

TPR：70.9 % FPR：30.3%

灵敏度：Sensitivity=TPR=70.9% 特异度：Specificity=1-FPR=69.7%

（四）建立打分卡

表 5-5 打分卡

变量	取值	评分
收入	<= 8152	75
	<= 23978	101
	<= 47979	121
	> 47979	165
好友	<= 25	75
	> 25	43
消费观念	<= 0.27	75
	> 0.27	43
性别	’0’	75
	’1’	69
婚姻	’0’	75
	’1’	87
生育	’0’	75
	’1’	85
教育水平	’1’	75
	’2’	72
	’3’	88
	’4’	110
英语水平	’1’	75
	’2’	86
	’3’	99
	’4’	106

给每个用户打分之后，我们得出最高分为 778，最低分为 527；由图 3-2 打分频率分布图可以得出所有用户得分整体呈正态分布，众数在 600~620 之间。

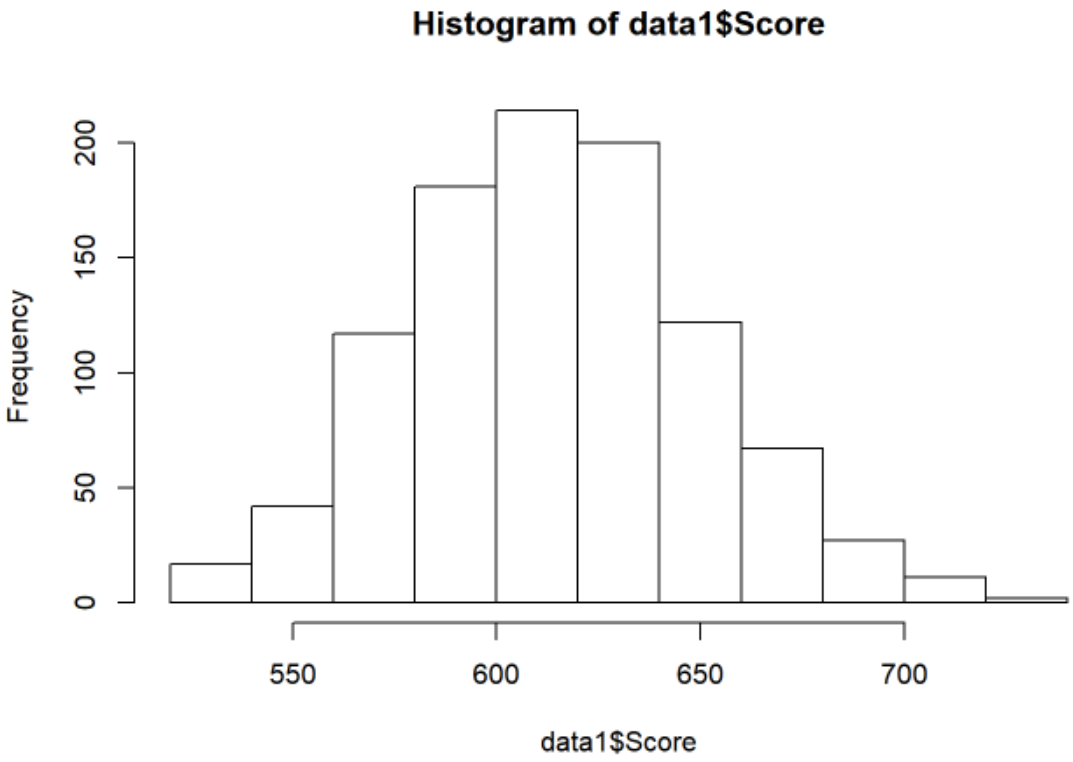

图 5-6 打分频率分布图

六、结论与建议

本案例基于某征信公司客户数据对是否按时还款进行了分析和预测。在数据预处理中，根据数据中数字代表的含义，转化为文字，进行了可视化分析。在数据的描述性分析中，通过绘制箱线图和荆棘图展示了各个自变量对因变量的不同影响。在构建模型的过程中，采用了 R 分析建立 0-1 回归模型。在征信公司客户数据统计中，征信客户数据样本数量相对较小，所以会存在样本类别不均衡的问题，本案例的处理方法能够在一定程度上对征信公司客户是否按时还款预测的问题上提供参考。

通过数据分析得出：影响征信公司客户是否按时还款的因素有性别、已婚_未婚、已育_未育，收入，学习能力方面：教育水平、英语水平，人脉方面的微博好友数和消费理念等。

（一）商业应用

1、个性化产品

- 根据客户按时还款率的因素模型，制定个性化信贷方案。
- 进一步结合客户使用行为数据，制定基于客户使用行为的信贷产品。

2、人群细分

- 按照 AIC 模型的预测客户按时还款的概率从高到低排序。
- 将排序后的客户等分成 4 份，高按时还款率，正常按时还款率，低按时还款率和超低按时还款率，代表不同风险人群。
- 考察这 4 种人群的实际流失概率。
- 模型识别的高风险人群：

占总人数的 34%

实际流失率只有 63%

VS.样本的整体流失率为 66.7%

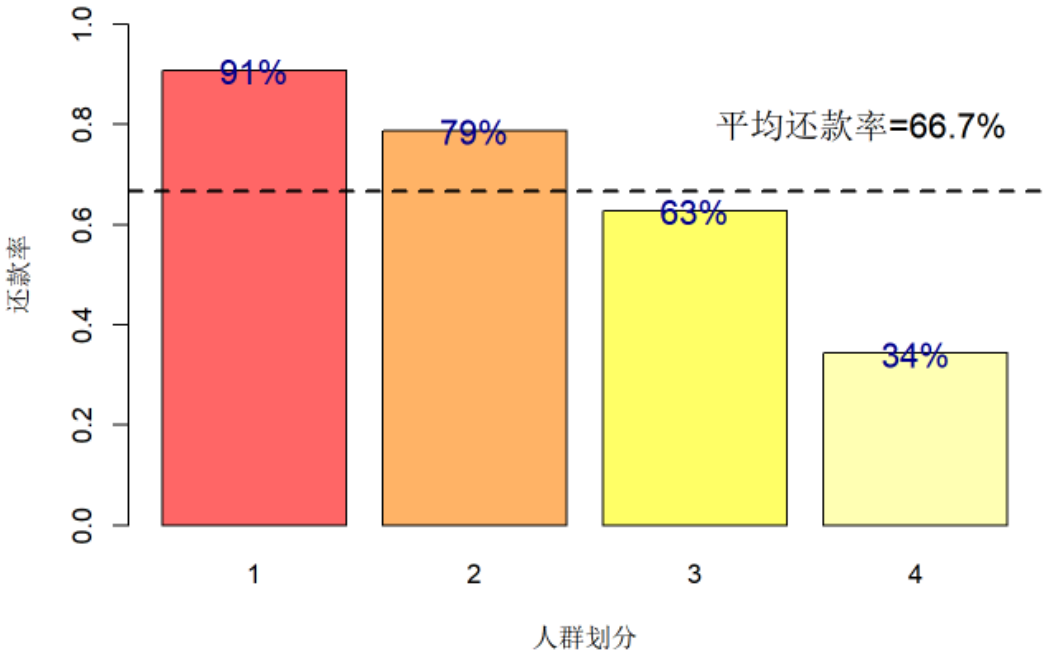

图 5-1 人群风险等级条形图

（二）建议

通过改变体制向客户需求靠拢，给人群划分，提供不同产品套餐，尽量给予客户科学实惠的服务。

（三）未来研究方向

- 案例的结论基于内样本，会高估模型预测精度。
- 考虑客户行为数据，定制基于客户行为的征信方案。
- 结合客户按时还款率等其他因素数据，预测出合适的信贷利率。

七、参考文献

[1]王梓骏.基于大数据的华振金融个人信贷信用评价研究[D].哈尔滨理工大学,2019.

王梓骏以华振金融个人信贷信用为研究对象,提出基于大数据的个人信贷信用评价方法,为华振金融解决依靠传统的信用评价方法无法准确评价其个人信贷信用的难题。论文最后提出了提高准确性的对策,主要包括完善信用数据库与提高办理个人信贷业务人员素质等内容。（字数：120）

[2]蔡金鑫,王一卓,郭文.大数据背景下的个人征信体系建设研究[J].技术经济与管理研究,2018(03):3-8.

蔡金鑫等人用"大数据+征信"的全新理念为我国的社会信用体系建设提供了新思路。文章通过改进传统的"5C 信用评估法",构建个人信用评估指标体系,计算相关权重;并针对我国大数据征信面临的问题,从多个角度提出具备一定可操作性和实际应用价值的对策建议。（字数：118）

[3]叶文辉.大数据征信机构的运作模式及监管对策——以阿里巴巴芝麻信用为例[J].国际金融,2015(08):18-22.

叶文辉等人将对芝麻信用的运作模式及特点进行分析，并在此基础上，提出促进大数据征信机构规范发展的相关建议。(字数：51)

[4]韩嵩,李晓俊.大数据背景下我国企业信用研究综述——基于 CSSCI 检索论文的分析[J].金融理论与实践,2018(10):107-113.

韩嵩等人基于 CSSCI 关于企业信用的学术成果,对国内企业信用研究的总体趋势以及具体的研究内容进行详细的梳理与总结,针对我国企业信用现状研究、企业信用指标研究、企业信用评价模型研究、企业信用管理研究四个方面对企业信用研究文献展开了述评。(字数：113)

[5] Nir Kshetri. Big data’s role in expanding access to financial services in China[J]. International Journal of Information Management,2016,36(3).

Nir Kshetri's first major goal is to study the role of big data in facilitating access to financial products by economically active low-income households and micro-enterprises in China.The second goal is to investigate how formal and informal institutions promote and limit the use of big data in China's financial sector and markets.The analysis shows that the main reason for low-income households and enterprises' lack of financial services in China and other emerging economies is not because they lack credibility, but simply because Banks and financial institutions lack data, information and the ability to effectively access the credibility and provide financial services to this financially vulnerable group. (word：106)