风控评分模型报告

一、背景描述

现如今，金融信用消费产品的数量，种类越来越多，信用消费的规模也越来越大，种类由过去单一、大额、手续繁杂的房贷、车贷产品，发展出现如今小额，手续便捷，响应速度更快的花呗、京东白条、借呗等全新的信用消费产品。随着数量、种类的快速增多，信用消费便捷快速的同时，需要的是更严格的风险控制。

为了更高效，便捷的管控金融信用风险，在央行的监管下两家个人征信机构百行征信，朴道征信相继成立，其主要产品按服务对象划分为：个人征信产品，企业征信产品，政务产品，可用于查询个人信用报告，帮助中小微企业提供征信报告，有助于企业融资，银行进行优质贷款，为政府提供便民服务。所提供的征信包括了个人或企业的基本情况以及信用评估报告，可用于信用证明，信用消费等与信用相关的业务。不仅可以帮助个人或企业分辨优质的用户与合作伙伴，国家也可依据征信报告快捷办理便民业务，也可对失信人员进行约束，如限制购机票、火车票等举措，督促失信人员尽快弥补失信行为，高效的降低了资金流动中的信用风险。

风控评分能够有效的量化风险控制，有效的降低了金融放贷机构的信用风险。风控评分对于银行的重要性如下： 1. 风控评分可对贷款人进行信用分级（未来考虑对不同等级人群给予不用的贷款利率） 2. 风控评分降低贷款人还贷逾期风险（提高银行现金流的稳定性，为银行带来高确定性的收益，降低银行坏账风险） 3. 风控评分增加了信用评价量化标准，未来可为低抵押甚至无抵押的用户进行贷款，提高银行收益，增加银行资产流动性。

本文选取具有市场代表性的信用消费产品——银行信用卡作为研究对象，对 1000 名用户的数据进行逻辑回归分析，量化评价不同用户的按时还款能力。

二、数据概述

表 1:数据描述表

数据类型		数据名称	详细说明	取值范围	数据分布	备注
因变量		还款状况	定性变量	0：坏人；1:好人	0:333（33.3%），1:667（66.7%）	好人占比 66.7%
自变量	基本信息	性别	定性变量	0：女性；1:男性	0:503（50.3%），1:497（49.7%）	男性占比 49.7%
		单身/已婚	定性变量	0：单身；1:已婚	0:503（50.3%），1:497（49.7%）	已婚占比 49.7%
		未育/已育	定性变量	0：未育；1:已育	0:512（51.2%），1:488（48.8%）	已育占比 48.8%
	收入水平	收入	连续变量 呈右偏分布	426-120940	平均值：21549，中位数：18080，数据右偏	单位：元
	学习能力	教育水平	定性变量	1：高中及以下，2：大专或本科，3：硕士研究生，4:博士研究生及以上	1:212（21.2%），2:335（33.5%），3:354（35.4%），4:99（9.9%）	
		英语水平	定性变量 共 4 个水平	1：四级以下，2：四级，3：六级，4:六级以上	1:224（22.4%），2:329（32.9%），3:305（30.5%），4:142（14.2%）	
	社交人脉	微博好友数	连续变量	6-114（整数）	平均值 40.99，中位数：39，数据右偏	单位：个
	消费理念	消费理念	连续变量	0-1	平均值：0.3903，中位数：0.35，数据右偏	消费理念=信用卡消费/总消费

三、描述分析

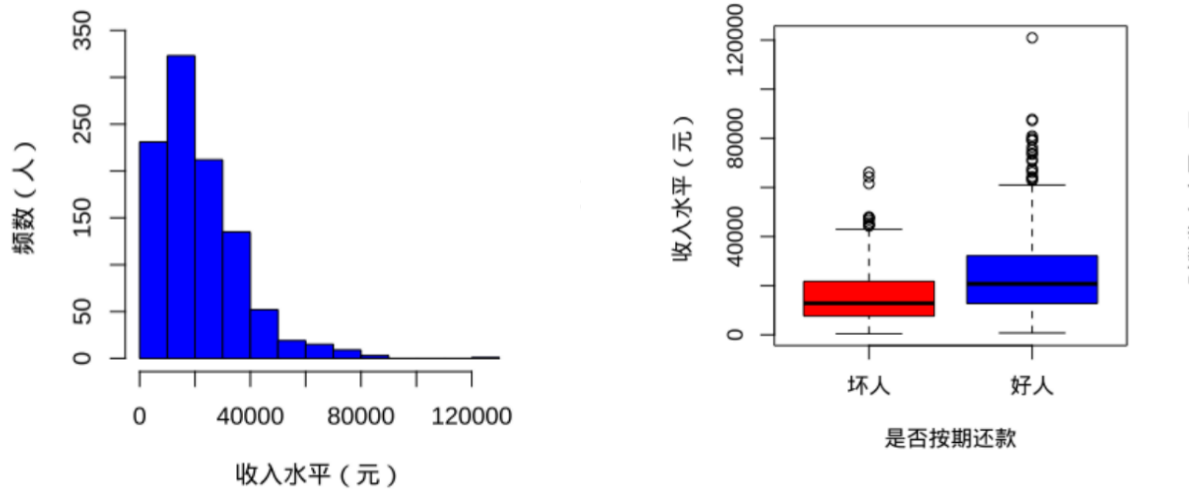


图 1: 收入水平直方图，箱线图

收入水平呈右偏分布，收入平均值：21549 元，收入中位数：18080 元。好人的收入中位数显著高于坏人，初步判断收入水平与按时还款正相关。

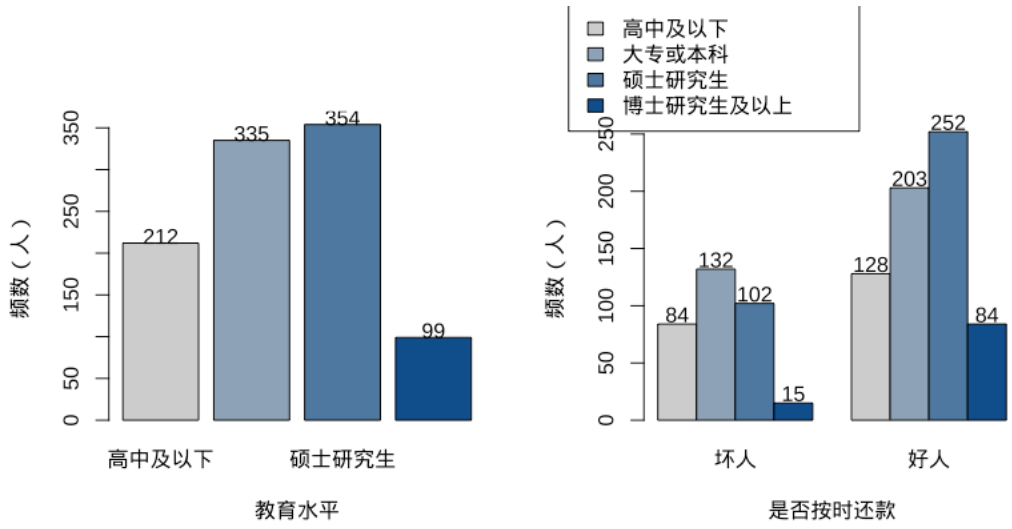


图 2: 教育水平直方图

高中及以下学历人群违约率：84/212=39.62%，大专或本科学历人群违约率：132/335=39.40%，硕士研究生学历人群违约率：102/354=28.81%，博士研究生及以上学历

人群违约率：15/99=15.15%。高学历人群违约率显著低于低学历人群。学历水平与按时还款正相关。

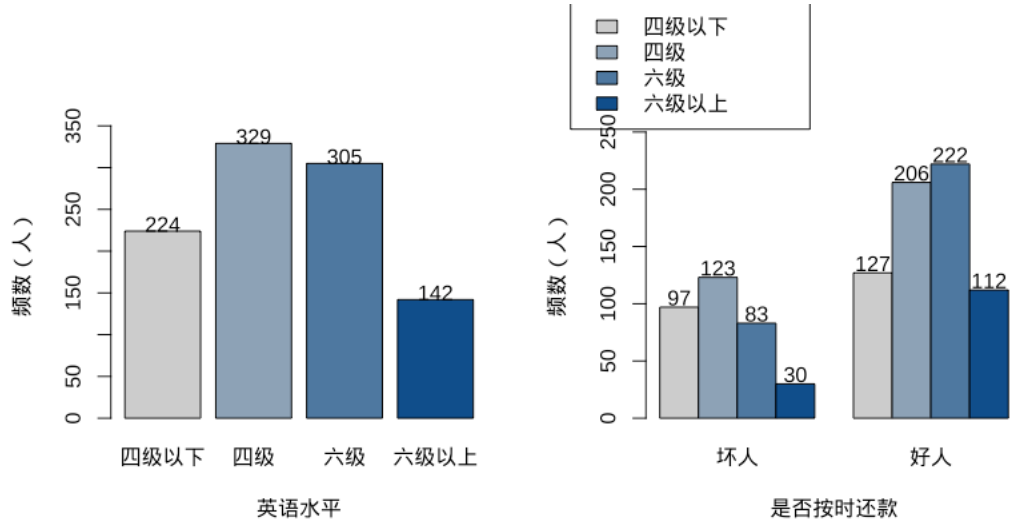


图 3: 英语水平直方图

英语水平四级以下人群违约率：97/224=43.30%，四级人群违约率：123/329=37.39%，六级人群违约率：83/305=27.21%，六级以上人群违约率：30/142=21.13%。高英语水平人群违约率显著低于低英语水平人群。英语水平与按时还款正相关。

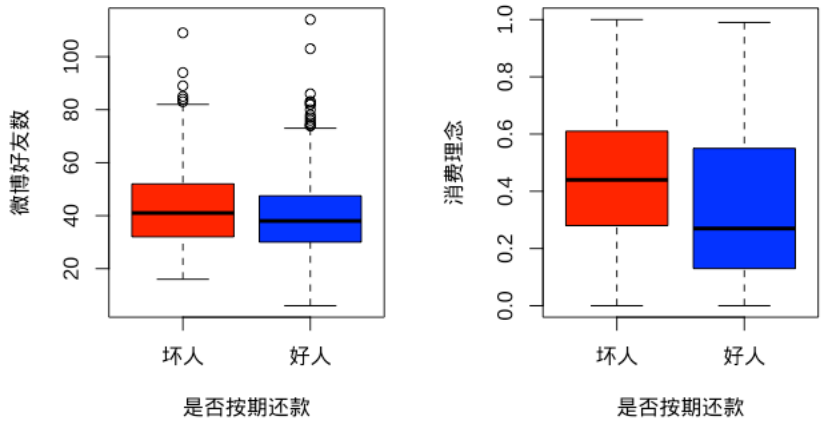


图 4: 微博好友数，消费理念箱线图

好人的微博好友，消费理念中位数显著低于坏人，初步判断微博好友数，消费理念与按时还款负相关。

四、模型建立

本逻辑模型包含一个自变量：是否按期还款，8 个自变量性别，单身/已婚，未育/已育，收入，受教育程度，英语水平，微博好友数，消费理念。将案例数据的连续变量标准化后，对自变量进行逻辑回归，回归结果如下表。

表 2：标准化后回归系数表

变量名	回归系数	显著性	备注
收入	0.748	<0.001	不显著
微博好友数	-0.309	<0.001	
消费理念	-0.382	<0.001	
性别：男性	-0.266	0.077	基准组：女性
单身/已婚：已婚	0.398	0.008	基准组：单身
未育/已育：已育	0.364	0.015	基准组：未育
教育水平：大专或本科	-0.004	0.982	基准组：高中及以下
教育水平：硕士研究生	0.541	0.007	
教育水平：博士研究生及以上	1.374	<0.001	
英语水平：四级	0.348	0.076	基准组：四级以下
英语水平：六级	0.784	<0.001	
英语水平：六级以上	1.183	<0.001	
全样本 AUC 值	0.759	**AIC**	1104.5

由上表得出，控制其他变量不变的情况下：

- 收入回归系数为正且显著，收入高的用户更倾向于按时还款
- 微博好友数，消费理念回归系数为负，微博好友数越多，信用卡消费比例越大的用户按时还款可能性越小
- 硕士，博士及以上学历用户比高中及以下学历用户更有可能按时还款，但大专或本科用户与高中及以下学历用户按时还款无明显差异
- 英语水平六级及以上用户比四级以下用户更有可能按时还款，但四级用户与四级以下用户按时还款无明显差异
- 已婚，已育用户比单身，未育用户更有可能按时还款

五、模型预测

使用上述逻辑回归模型对内样本进行预测，得到 ROC 曲线图：

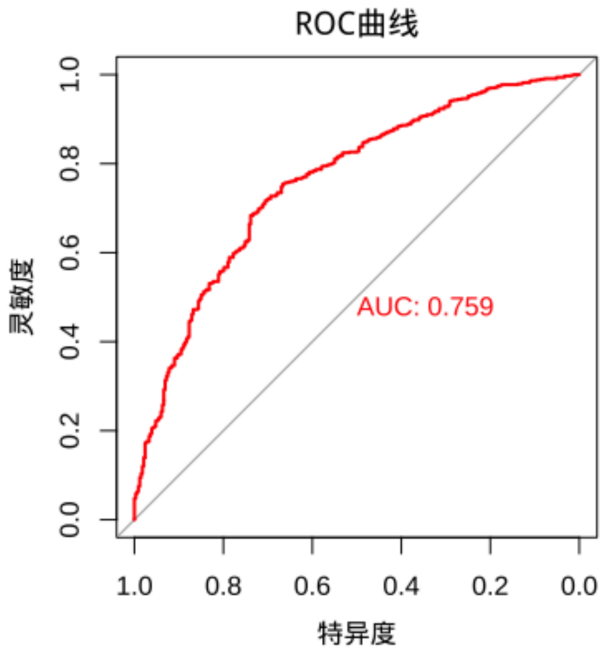


图 5: ROC 曲线图

由图 5 可知，全模型的内样本 AUC 值为 0.759，即在 ROC 曲线下方的面积为 0.759

为了验证模型在真实测试数据的准确性，随机拆分 80%的数据作为训练集，20%的数据作为测试集，模拟 100 次，得到外样本的 AUC 取值为 0.747

六、商业化结果展示

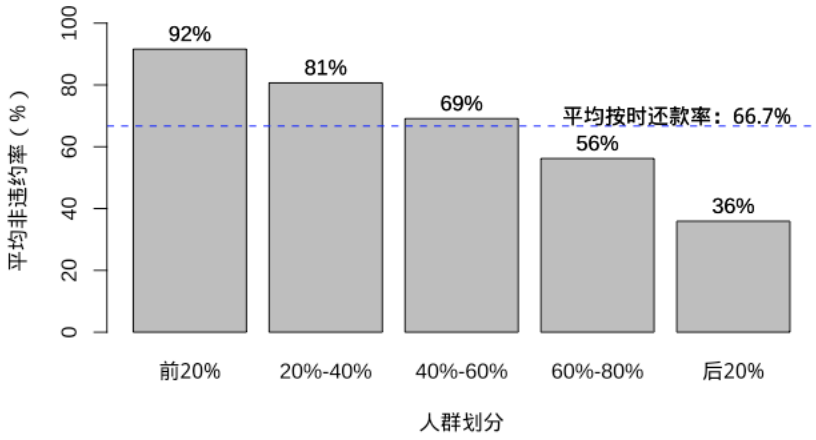
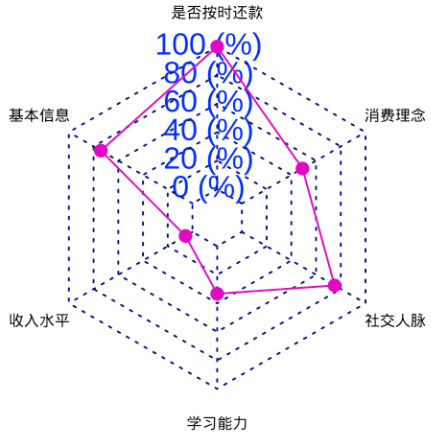
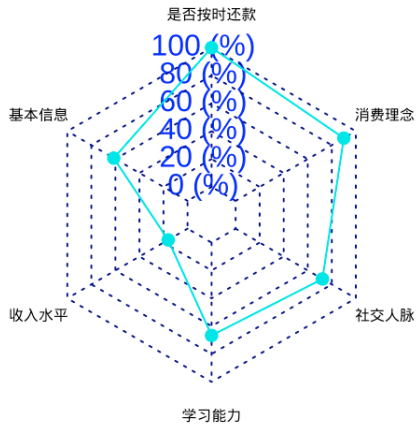


图 6: 人群划分图

由图 6 可知，前三个分位数样本按时还款率均高于平均水平，为了降低坏账风险，建议向按时还款率靠前的分组样本放贷。





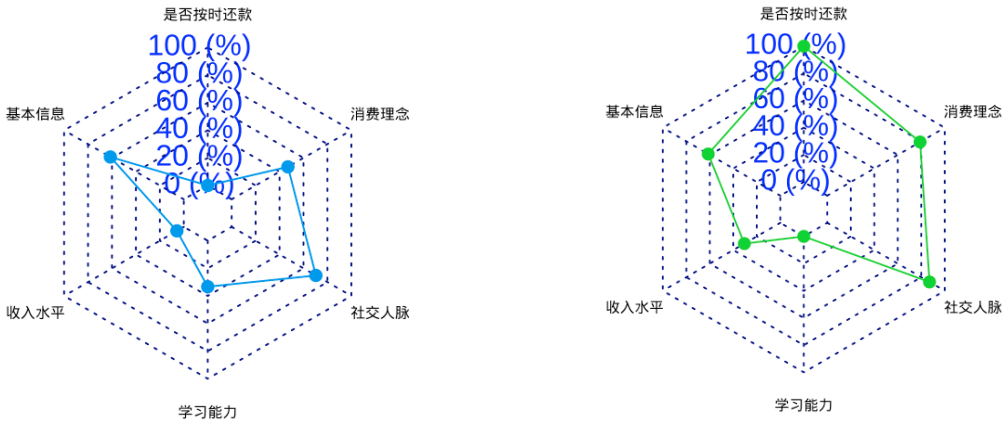


图 7：用户画像雷达图

图 7 中选取了 4 个样本的数据，其中 3 位用户按时还款，这 3 位用户在收入，学习能力两个正相关指标表现较好，在消费理念，社交人脉两个负相关指标表现较差，因此按时还款，被归类为优质客户。

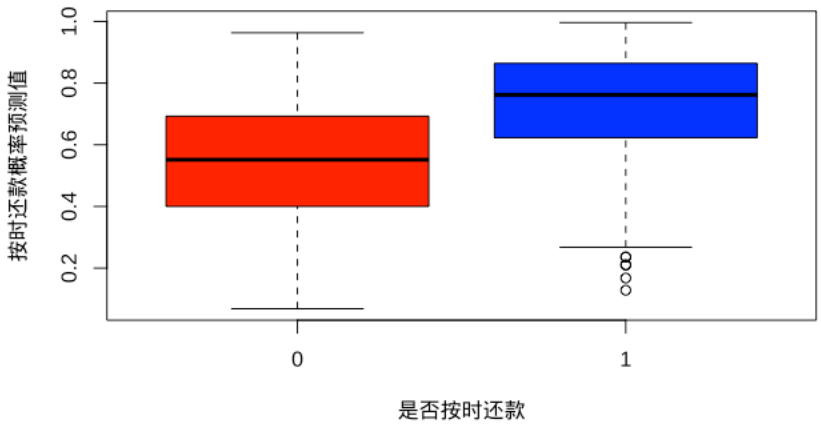


图 8: 按时还款概率预测值的分组箱线图

由图 8 可知，按时还款与非按时还款用户的中位数差距显著，因此本模型具有较强的区分度。

七、结论与建议

根据风控打分模型，控制其他变量不变的情况下，我们可以得出如下结论：

- 收入，学历水平，英语水平高的用户更倾向于按时还款
- 微博好友数，消费理念高的用户更不容易按时还款
- 已婚，已育用户比单身，未育用户更有可能按时还款

因此，针对信用卡还款给出如下建议：

- 为降低坏账风险，在其他资质相同的情况下，尽可能选择收入高，学历水平，英语水平高的已婚已育用户，且微博好友数少，消费理念低的用户开放高额的信用消费