银行个人客户信用与价值管理

项目简介：本次项目通过建立逻辑回归模型对客户的违约和营销响应进行预测和评分，对客户进行细分，针对不同单元客户制定营销策略。

一、 数据的诊断

1.1 单变量诊断

删除了申请表中重复的 ID 信息，共 47 个样本。

分类变量的诊断情况如以下图表所示。

变量	分类取值	总样本数	分类样本数	分类占比
性别	女	438510	294406	67.14%
性别	男	438510	144104	32.86%

图表 1 性别变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
是否有车	否	438510	275427	62.81%
是否有车	是	438510	163083	37.19%

图表 2 是否有车变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
是否有房产	是	438510	304040	69.33%
是否有房产	否	438510	134470	30.67%

图表 3 是否有房产变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
收入类型	工作	438510	226076	51.56%
收入类型	商业合伙人	438510	100744	22.97%
收入类型	退休金	438510	75488	17.21%
收入类型	公务员	438510	36185	8.25%
收入类型	学生	438510	17	0.00%

图表 4 收入类型变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
教育程度	高中	438510	301788	68.82%
教育程度	大学	438510	117512	26.80%
教育程度	不完全大学	438510	14847	3.39%
教育程度	初中	438510	4051	0.92%
教育程度	学位等级	438510	312	0.07%

图表 5 教育程度变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
婚姻状态	已婚	438510	299798	68.37%
婚姻状态	单身/未婚	438510	55258	12.60%
婚姻状态	同性恋	438510	36529	8.33%
婚姻状态	离异	438510	27251	6.21%
婚姻状态	遗孀	438510	19674	4.49%

图表 6 婚姻状态变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
居住方式	房子/公寓	438510	393791	89.80%
居住方式	和父母住	438510	19075	4.35%
居住方式	市政公寓	438510	14212	3.24%
居住方式	租公寓	438510	5973	1.36%
居住方式	办公公寓	438510	3920	0.89%
居住方式	合作公寓	438510	1539	0.35%

图表 7 居住方式变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
是否有手机	是	438510	438510	100.00%
是否有手机	否	438510	0	0.00%

图表 8 是否有手机变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
是否有工作电话	否	438510	348121	79.39%
是否有工作电话	是	438510	90389	20.61%

图表 9 是否有工作电话变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
是否有电话	否	438510	312320	71.22%
是否有电话	是	438510	126190	28.78%

图表 10 是否有电话变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
是否有邮件	否	438510	391063	89.18%
是否有邮件	是	438510	47447	10.82%

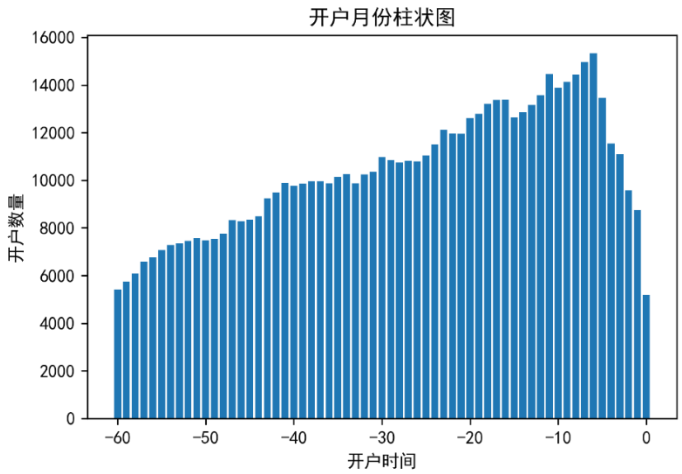
图表 11 是否有邮件变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
职业	未知	438510	134193	30.60%
职业	劳工	438510	78231	17.84%
职业	核心员工	438510	43000	9.81%
职业	销售员	438510	41094	9.37%
职业	经理	438510	35481	8.09%
职业	司机	438510	26090	5.95%
职业	高技能职业	438510	17285	3.94%
职业	会计	438510	15983	3.64%
职业	医务人员	438510	13518	3.08%
职业	厨师	438510	8076	1.84%
职业	保安	438510	7993	1.82%
职业	清洁工	438510	5843	1.33%
职业	私人服务员	438510	3455	0.79%
职业	低技能职业	438510	2140	0.49%
职业	秘书	438510	2044	0.47%
职业	写手/酒吧服务员	438510	1665	0.38%
职业	房地产经纪人	438510	1041	0.24%
职业	人事招聘员	438510	774	0.18%
职业	信息技术员	438510	604	0.14%

图表 12 职业变量诊断统计表

变量	分类取值	总样本数	分类样本数	分类占比
借款状态	C	15277108	6398779	41.88%
借款状态	0	15277108	5430795	35.55%
借款状态	X	15277108	3247736	21.26%
借款状态	1	15277108	160932	1.05%
借款状态	5	15277108	22066	0.14%
借款状态	2	15277108	10664	0.07%
借款状态	3	15277108	3696	0.02%
借款状态	4	15277108	2440	0.02%

图表 13 借款状态变量诊断统计表



图表 14 开户月份柱状图

从上以上的图表可以发现职业这一变量出现缺失值共计 134193 个样本。这里将缺失值另外做一个类别“未知”。

1.2 连续变量诊断

连续变量诊断统计表如图表 15 所示。

变量	中位数	总样本数	样本缺失值	最大值	均值	最小值	标准差
年收入	160940	438510	0	6750000	187525	26100	110089
孩子个数	0	438510	0	19	0	0	1
家庭人数	2	438510	0	20	2	1	1
年龄	43	438510	0	69	44	21	11
工作年限	4	438510	0	48	-166	-1001	380

图表 15 连续变量诊断统计表

由图表 16 可以看出工作年限的最小值不符合常识，查看原数据，开始工作时间的信息里有 75324 条样本数为 365243 这一异常值，这里可以是失业的一种表达形式，将其分在“失业”类别。对开始工作时间为非缺失值的数据进行重新诊断统计，如下方图表 16 所示。

变量	中位数	总样本数	样本缺失值	最大值	均值	最小值	标准差
工作年限	5	363186	75324	48	7.196533	0	6.585014

图表 16 工作年限变量诊断统计表

1.3 数据合法性诊断

数据合法性诊断统计表如图表 17 所示。

条件	ID个数	备注
孩子人数>=家庭人数	99	

图表 17 数据合法性诊断表

由图表 17 可知孩子人数大于等于家庭人数的不合法样本有 99 条，这里将其做删除处理。

1.4 表关联诊断

跨表 ID 匹配诊断表如图表 18 所示。

表	ID个数	备注
只在申请记录表出现	13229	
只在信用记录表出现	204305	
共同出现	425182	

图表 18 跨表 ID 匹配诊断表

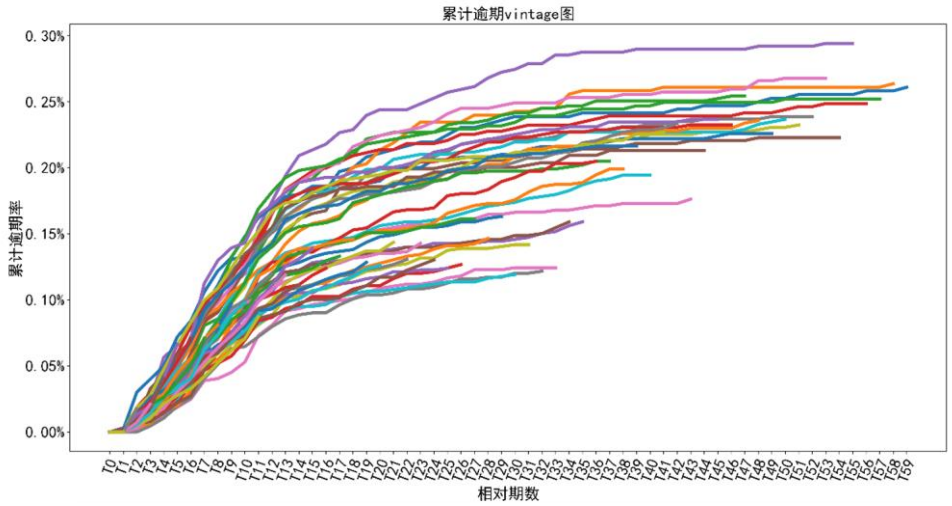
由图表 18 可以看到共同出现的 ID 数有 425182 个，这里将不共同出现的 ID 分别在申请表和信用记录表中删除。

二、坏客户的预测

2.1 定义坏客户

在规定表现期内达到规定的逾期状态的客户被定义为坏客户。

其中表现期一般由 vintage 的累计逾期图得到，累计逾期 60 天的 vintage 图见图表 19。图中的相对期数指的是距离开户的月数（横轴的 T1 表示距离开户一个月，以此类推）。图中每条线表示的是不同月份开户的客户在相对期数的累计逾期超过 60 天的客户数。通过 vintage 图可以看出在 20 周期以后的累计逾期率的曲线斜率几乎平稳，即在 20 周期的观察窗口中绝大多数用户暴露了自己的信用情况。在开户 20 周期内，用户累计逾期率的曲线斜率依旧较大，说明观察期小于 20 周期不足以使用户完全暴露其信用情况。开户 20 周期后，累计逾期率有缓慢的提升，虽然观察窗口越长，负样本比例越多，越容易训练得到精确的模型，但是总样本的数量也在不断减少，使得模型的适用范围大大受限。综合考虑，选择 20 周期作为最终观察窗口。



图表 19 累计逾期 60 天 vintage 图

根据巴塞尔协议的评级标准，在表现期内一旦达到 90 天逾期就视为违约。但是此次 90 天逾期的坏用户样本数过少（见图表 20），不利于建模，并且一般坏用户样本的占比控制在 3%，故选择表现期内达到 60 天逾期的客户作为风险用户的定义。

综上所述，定义开户 20 个月内有达到 60 天逾期状态的客户为坏客户。

逾期情况		样本数	总样本数	样本占比
>1	目标Y=1	216048	247498	87.29%
	非目标Y=0	31450	247498	12.71%
>30	目标Y=1	30566	247498	12.35%
	非目标Y=0	216932	247498	87.65%
>60	目标Y=1	4059	247498	1.64%
	非目标Y=0	243439	247498	98.36%
>90	目标Y=1	1773	247498	0.72%
	非目标Y=0	245725	247498	99.28%
>120	目标Y=1	1278	247498	0.52%
	非目标Y=0	246220	247498	99.48%

图表 20 坏用户定义与坏用户个数关系表

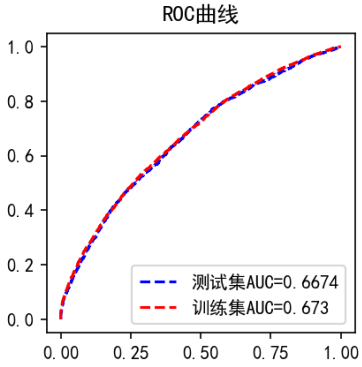
2.2 建立逻辑回归模型预测坏客户

建立逻辑回归模型 ，Y 为 1 表示坏客户，0 表示好客户。将通过分箱得到变量的 IV 值，筛选 IV 值较大的 X 变量进行模型的建立。去除显著为 0 的变量，最终模型展示见图表 21。

变量		回归系数	备注
截距		0.000	
信用局评分		-0.005	
女性		-0.558	基准值：男性
拥有房产		-0.311	基准值：没有房产
收入类别	养老金	7.254	基准值：其他
	工作	-0.385	
	商业合伙人	-0.467	
婚姻状况	未婚/遗孀	1.130	基准值：已婚
	离婚	0.850	
	同性恋	0.219	
职业	司机	-0.298	基准值：其他
	劳工	-0.261	
	经理	-0.186	
	销售	-0.309	
	未知	-0.276	
年收入	108225元以下	0.393	基准值：225382元以上
	108225-225382元	0.126	
年龄	35岁以下	-0.303	基准值：43岁以上
	36-43岁	-0.544	
家庭人数	1人	-1.345	基准值：3人以上
	2人	-0.476	
	3人	-0.503	
工作年限	失业	-7.584	其他

图表 21 逻辑回归模型展示

模型预测效果见图表 22 的 ROC 曲线。ROC 曲线越平滑，说明其拟合性越好。AUC 值越高说明预测效果越好。本次测试集的 AUC 为 0.67，曲线较为平滑，且和训练集拟合效果差不多，故模型预测能力较好。



图表 22 ROC 曲线

2.3 将逻辑回归值转为分数

将逻辑回归模型得到的概率转化为 400-900 区间的分数。转化公式见公式 1，p 表示逻辑回归得到的违约率，通过设定基础分为 525 分，违约概率与正常概率的比值为 5%，违约概率与正常概率的比值翻番时分数的变动值为 10 分，得到公式 1 的两个系数，将分数映射到 400-900 区间的分数。

$$\text{风险得分} = 616.95 - 36.07 * \ln\left(\frac{p}{1-p}\right) \qquad \text{公式（1）}$$

2.4 模型验证结果

模型稳定性监控报表见图表 23。稳定性系数越小表示模型越稳定，由表可见该模型稳定性系数为 0.00007，稳定性良好。

分数	建模样本数	百分比1	评分样本数	百分比2	ln(百分比比例)	百分比变化	IV值
400-600	76	0.04%	22	0.03%	-0.392	-0.01%	0.00006
600-700	788	0.45%	333	0.45%	-0.014	-0.01%	0.00000
700-800	135227	78.05%	58087	78.23%	0.002	0.18%	0.00000
800-900	37157	21.45%	15808	21.29%	-0.007	-0.16%	0.00001
合计	173248	100.00%	74250	100.00%	稳定性系数（PSI）		0.00007

图表 23 模型分数稳定性监控报表

模型预测能力的监控报表见 [(1)] 图表 24，可以看出随着分数的升高，坏客户率大大降低。模型预测区分能力高。

细分单元	最高分	最低分	建模样本坏客户率	评分样本坏客户率
1	600	400	96.05%	95.45%
2	700	600	9.90%	9.61%
3	800	700	1.81%	1.83%
4	900	800	0.63%	0.74%

图表 24 模型预测能力监控报表

三、营销响应客户的预测

3.1 X 变量的生成

本次营销响应预测模型是逻辑回归模型，对于 X 变量的范围包括客户交易信息，余额信息，产品使用情况，渠道使用情况，客户风险信息，客户价值信息（RFM）。其中对 RFM 指标划分定义见图表 25-图表 27。

对 R 指标进行等级划分，见图表 25。由于客户在 R 指标的客户数几乎分布在 0 天，而其他天数的客户很少，故分成两档。

客户分组	最近消费距离天数分档	客户数
高	0天	4440
低	0天以上	50

图表 25 R 指标分档

对 F 指标进行等级划分，见图表 26，分档的界限为 F 指标的中位数 40 次。

客户分组	半年交易次数分档	客户数
低	40次以下	2879
高	40次以上	1611

图表 26 F 指标分档

对 M 指标进行等级划分，见图表 27。划分的依据是二八定律，即公司 80%的收入来自于 20%的用户。

客户分组	半年交易总金额	客户数
低	310000元以下	3543
高	310000元以上	947

图表 27 M 指标分档

RFM 的客户价值分层模型展示见图表 28。由图表可见客户主要为重要价值/发展和一般价值/发展客户。

划分群体类型	R（最近一次消费时间）	F（消费频率）	M（消费金额）	成交客户等级	客户数
重要价值客户	高	高	高	A	616
重要发展客户	高	低	高	A	331
重要保持客户	低	高	高	B	0
重要挽留客户	低	低	高	B	0
一般价值客户	高	高	低	B	995
一般发展客户	高	低	低	B	2498
一般保持客户	低	高	低	C	0
一般挽留客户	低	低	低	C	50

图表 28 RFM 客户价值分层模型展示

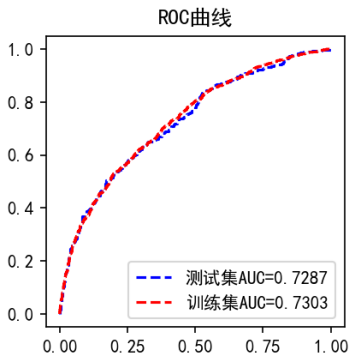
3.2 建立逻辑回归预测营销响应

建立逻辑回归模型，因变量 Y 取 1 表示该用户的营销响应，Y 取 0 表示不响应。将经过筛选的变量带入模型，并去除显著为 0 的变量，得到的模型展示见图表 29。

变量	取值	系数	备注
截距项		0.000	
表现期的交易次数		0.027	
利息交易次数占比		1.613	
贷款交易次数占比		1.887	
现金支取次数占比		0.795	
家庭人数		-1.821	
收款次数占比		2.253	
孩子个数		2.110	
工作年限		0.011	
年龄		-0.006	
有保险交易		1.382	基准值：没有保险交易
RFM	一般发展客户	-0.349	基准值：其他
	重要价值客户	0.244	
	重要发展客户	-0.367	
是否有房贷	是	0.955	基准值：没有房贷
婚姻	已婚	0.151	基准值：其他
	离婚	0.219	
	单身	-1.994	
	遗孀	-2.221	
性别	男	0.591	基准值：女
职业	厨师	-0.396	基准值：其他
	司机	-0.519	
	人事招聘	1.013	
	高技能技术人员	-0.264	
	低技能劳动者	-1.287	
	经理	0.168	
	秘书	0.777	

图表 29 逻辑回归模型展示

模型预测效果见图表 30 的 ROC 曲线。本次测试集的 AUC 为 0.73，曲线较为平滑，并且和训练集拟合效果差不多，故可认为模型预测能力良好。



图表 30 ROC 曲线

3.3 逻辑回归值映射为分数

将逻辑回归模型得到的概率转化为 400-900 区间的分数。转化公式见公式 2，p 表示逻辑回归得到的营销响应率。

$$收益得分 = 724.49 + 64.92 * \ln\left(\frac{p}{1-p}\right) \quad 公式（2）$$

3.4 模型验证结果

将逻辑回归得到的响应概率值转化为 400-900 区间的分数，模型稳定性监控报表见图表 31，稳定性系数为 0.0063，即模型稳定性良好。

分数	建模样本数	百分比1	评分样本数	百分比2	ln(百分比比例)	百分比变化	IV值
400-500	4	0.13%	3	0.22%	1.75	0.10%	0.0017
500-600	496	15.78%	227	16.85%	1.0679	1.07%	0.0114
600-700	1950	62.04%	787	58.43%	0.9417	-3.62%	-0.0341
700-800	606	19.28%	288	21.38%	1.1089	2.10%	0.0233
800-900	87	2.77%	42	3.12%	1.1264	0.35%	0.0039
合计	3143	100.00%	1347	100.00%	稳定性系数（PSI）		0.0063

图表 31 模型稳定性监控报表

模型预测能力的监控报表见图表 32，可以看出随着分数的升高，建模样本和评分样本的坏客户率都大大降低，模型预测区分能力高。

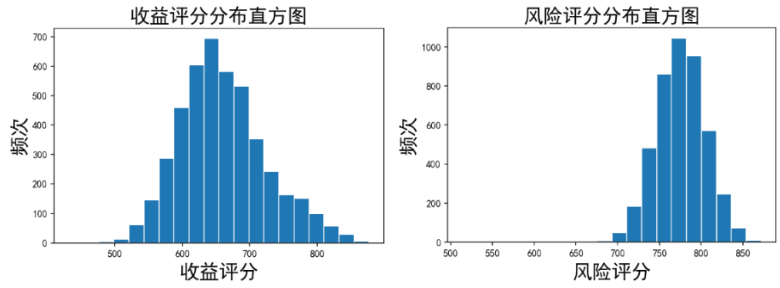
细分单元	最高分	最低分	建模样本响应率	评分样本响应率
1	500	400	0.00%	0.00%
2	600	500	8.27%	6.61%
3	700	600	24.92%	24.90%
4	800	700	55.94%	57.64%
5	900	800	79.31%	80.95%

图表 32 模型预测能力监控报表

四、营销策略的制定

4.1 营销单元的细分

表现窗口定在 1998 年的下半年，用户的风险分数分布和收益分数分布都呈正态分布状（见图表 33），故将风险评分和收益评分分别根据响应率差异最大化分成三组，使每组的区分度尽可能地高。分组结果见图表 34 和图表 35。



图表 33 风险/收益分数分布直方图

变量	变量分组	变量取值	总客户数	响应客户数	未响应客户数	响应率	woe	iv
收益评分	低	610分以下	953	82	871	8.60%	-1.5146	0.3273
收益评分	中	610-690分	2302	570	1732	24.76%	-0.2630	0.0335
收益评分	高	690分以上	1235	694	541	56.19%	1.0974	0.3770
收益评分	—	合计	4490	1346	3144	29.98%	—	0.7378

图表 34 收益评分分组表

变量	变量分组	变量取值	总客户数	坏客户数	好客户数	坏客户率	woe	iv
风险评分	高	735分以下	360	11	349	3.06%	1.3063	0.2757
风险评分	中	735-755分	711	10	701	1.41%	0.5136	0.0543
风险评分	低	755分以上	3419	17	3402	0.50%	-0.5354	0.1696
风险评分	—	合计	4490	38	4452	0.85%	—	0.4996

图表 35 风险评分分组表

4.2 营销策略的制定

本次营销策略的指定见图表 36。对于低风险用户，对高收益用户和中收益用户提升较大的额度；对于中风险用户，只对高收益用户提升少许额度；对于高风险用户，对银行造成的损失往往大于收益，故不提升额度。

风险评分	收益评分	提升额度	客户数	客户占比
高	高	3500	168	3.74%
	中	1000	149	3.32%
	低	—	43	0.96%
中	高	500	236	5.26%
	中	—	357	7.96%
	低	—	118	2.63%
低	高	—	868	19.35%
	中	—	1734	38.65%
	低	—	813	18.12%

图表 36 不同客户细分单元的策略制定

附录：改进建议

〔1〕 低于 600 分的客户坏客户率达到了 96.05%，建议该分段的客户必须拒绝开通信用卡。高于 800 分的用户的坏客户率只有 0.63%，建议该分段的客户可以自动通过信用卡开通的申请。