评分卡

1. 数据概述

数据来自借点花花历史放款数据,时间窗口为 6 月 15 日至 11 月 15 日，总用户量为 728471，逾期状态取该用户在时间窗口内最后一笔借款状态．好坏客户判定标准：逾期 9 天以上的认定为坏客户，9 天以内或者未逾期的认定为好客户.

2. 特征标签

标号	特征变量	特征解释
X0	label	是否是坏客户(1,0)
X1	age	借款用户年龄
X2	education	教育程度(0表示高中及以下,1表示大专,2表示本科,3表示硕士及以上)
X3	shebao	是否有社保(1,0)
X4	vehicle_num	是否是车主(1,0)
X5	income_range	收入范围(未知:0,3000以下:1,3000-5000:2,5000-8000:3,8000-12000:4)
X6	loan_rate	贷款利率
X7	seniority	工龄(unknown:0,6个月以内:1,12个月:2,12-24个月:3,24个月以上:4)
X8	client_type	新老客(1,2)
X9	zhima_score	芝麻信用分
X10	gender_Female	是否女性(1,0)
X11	gender_Male	是否男性(1,0)
X12	marriage_marriage	是否已婚(1,0)
X13	marriage_unmarriage	是否未婚(1,0)
X14	house_nature_owned	自有住房(1,0)
X15	house_nature_rent	租房(1,0)
X16	house_nature_unknown	未知住房情况(1,0)
X17	house_nature_with_parents	同父母居住(1,0)
X18	cv_ios	苹果客户端(1,0)
X19	cv_other	其他客户端(1,0)

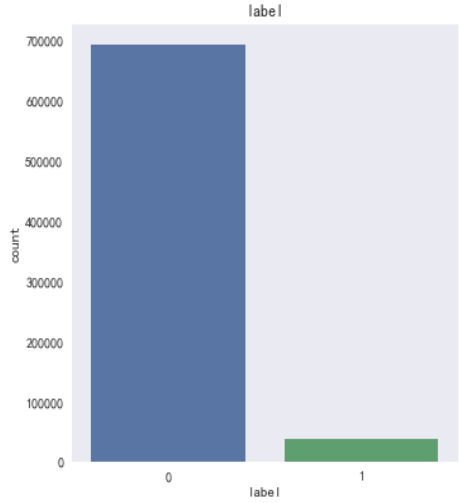
3. 变量解析

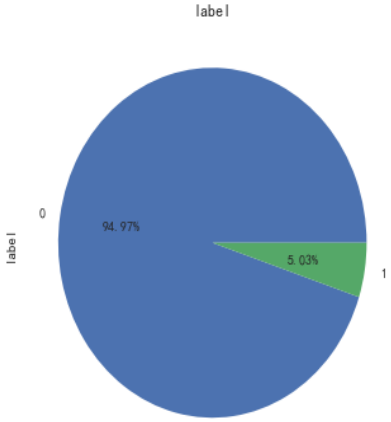
1 连续变量

	age	education	income_range	seniority	zhima_score
count	728603	728603	728603	728603	728603
mean	29	1	2	2	612
std	6	1	1	2	49
min	17	0	0	0	350
25%	24	0	0	0	582
50%	28	0	2	3	611
75%	32	1	3	4	641
max	60	3	4	4	836

借款用户年龄集中在 24 至 32 岁区间, 50%的用户教育程度均为高中及以下, 50%的客户月收入低于 5000.

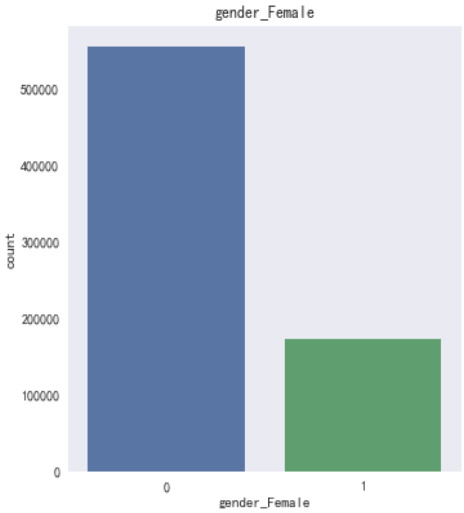
2 好坏客户比例

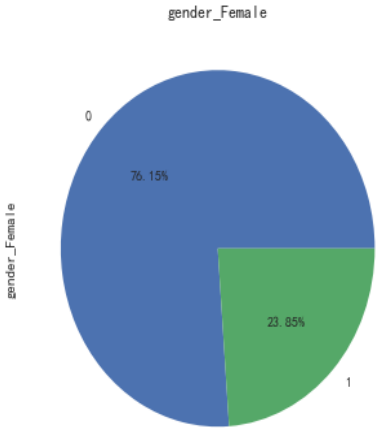




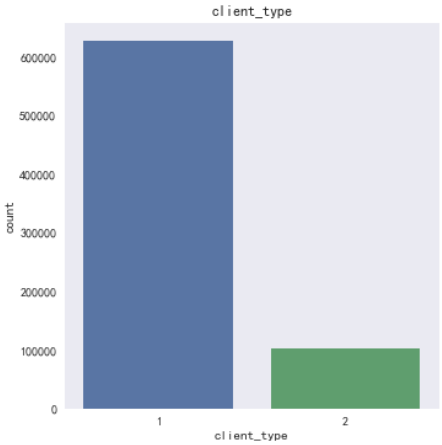
坏客户占比约为 5%,及逾期 10 天以上的客户约有 5 万户.

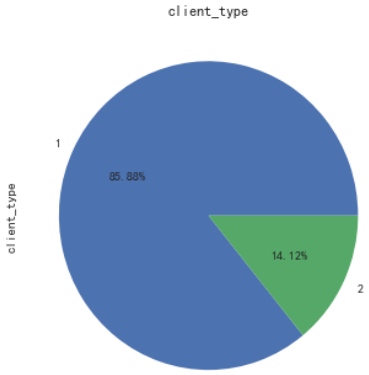
3 性别分布





4 新老客占比

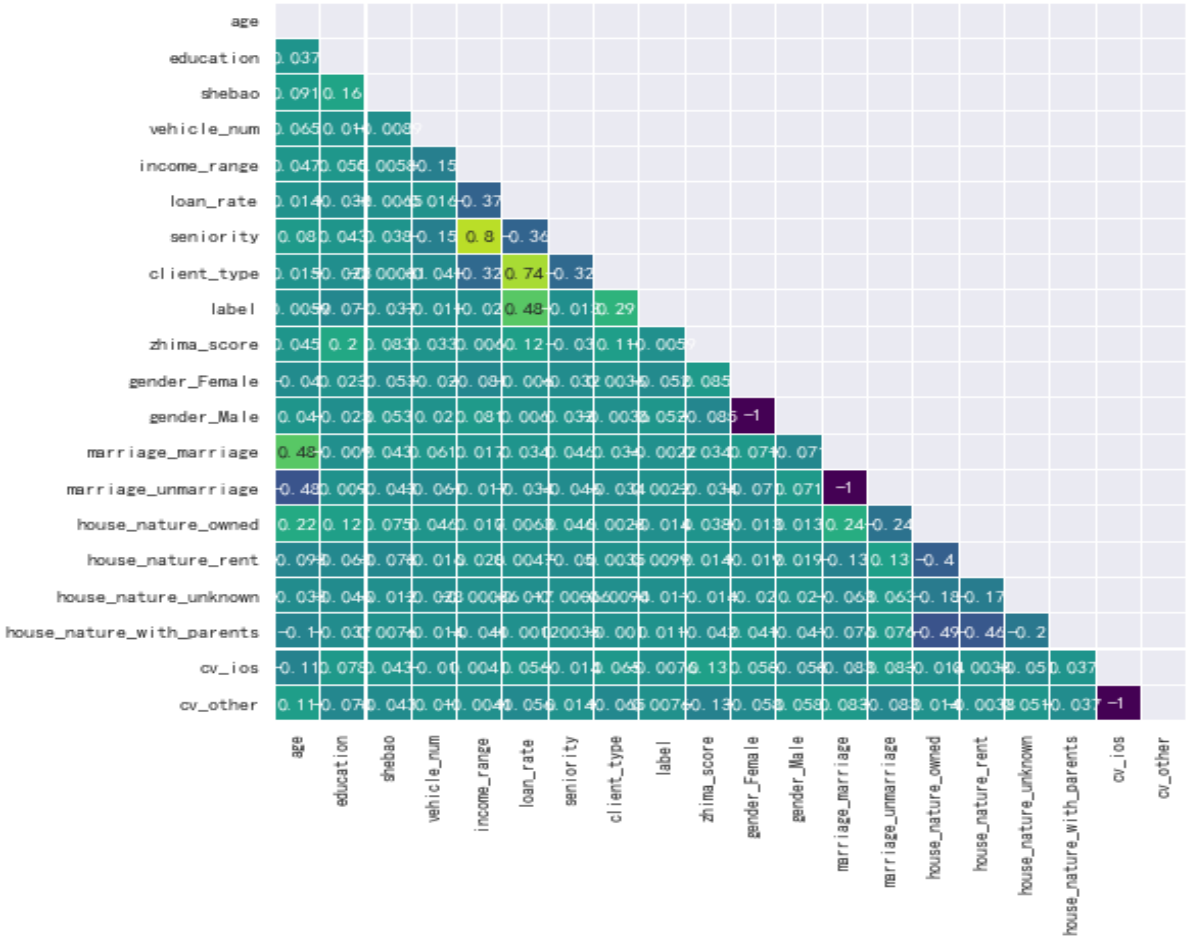




4. 变量筛选

1 相关系数分析

变量之间相关系数越大,两个变量之间互相影响程度较大，为了保证变量间独立性，理应删除其中一个变量，设定阈值为正负 0.5.



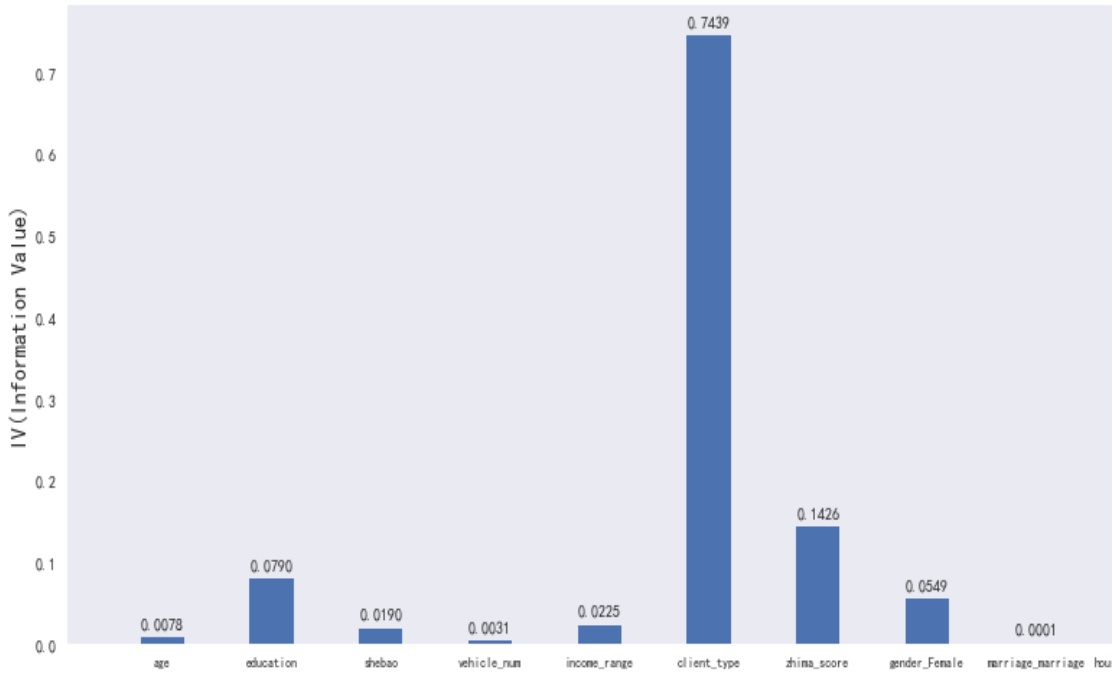
删除变量: seniority,loan_rate, gender_male, marriage_unmarriage, cv_ohter

2 信息价值判断

计算每个变量的 Infomation Value（IV)。IV 指标是一般用来确定自变量的预测能力,并删除信息价值低于 0.02 的变量.

通过 IV 值判断变量预测能力的标准是：

IV值	预测能力
<0.02	无
0.02-0.1	较弱
0.1-0.3	中等
0.3-0.5	较强
>0.5	强



删除变量: age, shebao, vehicle_num, marriage_marriage, house_nature_owned, house_nature_rent, house_nature_unknown, house_nature_with_parents, cv_ios.

剩余变量特征:

标号	特征变量	特征解释
X0	label	是否是坏客户(1, 0)
X2	education	教育程度(0表示高中及以下, 1表示大专, 2表示本科, 3表示硕士及以上)
X5	income_range	收入范围(未知:0, 3000以下:1, 3000-5000:2, 5000-8000:3, 8000-12000:4)
X8	client_type	新老客(1, 2)
X9	zhima_score	芝麻信用分
X10	gender_Female	是否女性(1, 0)

3. 特征权重(WOE)转换

	count	mean	std	min	25%	50%	75%	max
X2	728603.00000	-0.58860	0.35902	-1.32700	-0.79000	-0.27900	-0.27900	-0.27900
X5	728603.00000	-0.26514	0.20502	-0.47700	-0.47700	-0.39300	-0.03100	0.05600
X8	728603.00000	1.59900	0.00000	1.59900	1.59900	1.59900	1.59900	1.59900
X9	728603.00000	-0.05996	0.43332	-1.89600	0.16700	0.16700	0.16700	0.16700
X10	728603.00000	-0.54000	0.00000	-0.54000	-0.54000	-0.54000	-0.54000	-0.54000

5. 构建逻辑回归模型

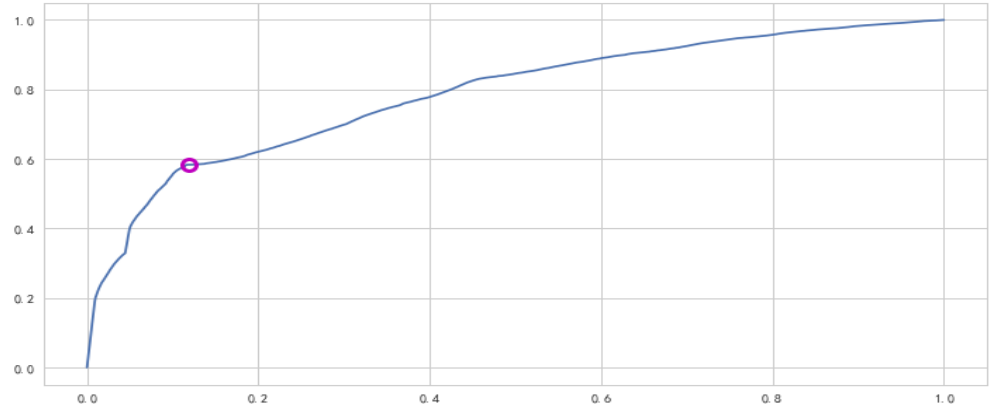
处理步骤: 数据标准化(极差法)--处理不平衡样本,---及通过过采样, 使得好坏客户在数量上相同---初步构建逻辑回归模型(默认参数)—模型参数优化(网络搜索)—重构逻辑回归模型—模型评估(AUC,KS)—输出特征系数

模型评估

	precision	recall	f1-score	support
0	0.68	0.88	0.77	207118
1	0.83	0.58	0.69	208073
avg / total	0.76	0.73	0.73	415191

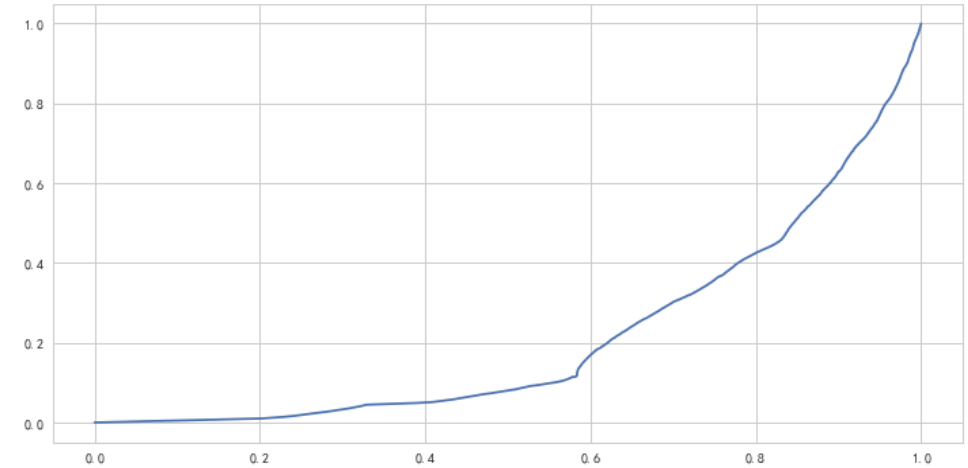
模型精度评估

模型精度：0.786078383429



即在全部机审的情况下,新增一个借款人时,判断准确率为 78.6%.

好坏客户区分度KS值：0.467006438426



在评分卡模型中, KS>0.2 是模型可用与否的界限, 该模型 KS 值为 0.467, 模型区分好坏客户能力一般.

6 信用评分

1 模型特征系数:
coe = [-0.0451483,-1.55711183 ,0.29976425,2.45474657,-0.88762834,-0.75027511]
分别对应的参数为:
coe_name = [intercept, education, income_range, client_type, zhima_score, gender_Female]

2 信用评分
取 600 分为基础分值, PDO(比率翻倍的分值)为 20（每高 20 分好坏比翻一倍），好坏比取 20。

计算公式:
p = 20/math.log(2)
q = 600 - 20 * math.log(20) / math.log(2)
baseScore = round(q + p *coe[0], 0)
个人总评分=基础分+各部分得分
baseScore = 512

2. 评分卡

baseScore(基础得分)=512	

education(学历)	score
高中及以下	-37
未知	-37
大专	-4
本科	15
硕士及以上	49

client_type(新老客)	score
新客	0
老客	113

gender_Female(女性)	score
是女性	12
不是女性	0

income_range(收入范围)	score
未知	-4
3000以下	-4
3000-8000	0
8000以上	3

zhima_score(芝麻信用分)	score
<550	-37
[550, 650)	-4
[650, 750)	15
>=750	49

3 评分检验(取 10%的数据集)

个人总得分 = 学历分+收入分+新老客分+芝麻分+性别分+基础分

逻辑回归模型KS=0.46,AUC=0.78						
score(得分)	0(good)	1(bad)	total	bad_rate	ratio	备注
(:,450)	1538	0	1538	0.00%	2.11%	模型失效
[450,470)	5407	268	5675	4.72%	7.79%	
[470,490)	22354	705	23059	3.06%	31.65%	
[490,510)	15668	372	16040	2.32%	22.01%	
[510,530)	10793	124	10917	1.14%	14.98%	
[530,550)	4284	27	4311	0.63%	5.92%	
[550,570)	892	5	897	0.56%	1.23%	
[570,590)	2751	1048	3799	27.59%	5.21%	
[590,610)	1303	295	1598	18.46%	2.19%	
[610,630)	2256	491	2747	17.87%	3.77%	
[630,650)	1503	205	1708	12.00%	2.34%	
[650,670)	490	45	535	8.41%	0.73%	
[670,690)	25	1	26	3.85%	0.04%	
[690,710)	8	0	8	0.00%	0.01%	
[710,:)	2	0	2	0.00%	0.00%	
合计	69274	3586	72860	4.92%	100.00%	

education	income_range	client_type	zhima_sco	gender_Fe	备注
高中及以下或未知	分别占比:17%,10%,50%,19%,2%	新客	<549	92%,男性	模型失效

7 总结

教育程度,收入范围均是用户手动输入,真实性不确定，另外运营商数据均无法提取，若要进一步提高模型准确度，仍需更多数据源，比如在网时长，月均话费,月均通话次数,月均家庭通话时长，可用余额，多头借贷笔数，公积金认缴时长等特征变量. 后续待催记标签上线并积累一定数据量，方可进行催收评分卡建模.