xx 消费贷准入策略项目报告

版本 1.2

版本迭代

版本	提交日期	提交者	说明
1.0	2017 年 3 月 1 日	xxx	-
1.1	2017 年 4 月 5 日	xxx	迭代更新
1.2	2017 年 5 月 2 日	xxx	新增同盾分、删除 xx 特征

目录

一、目标

本项目为消费贷准入策略项目，通过本项目希望实现以下目标：

（1） 引进先进理念和工具，一方面提升 xxx 公司的决策效率和技术水平，另一方面控制公司的风险敞口。

（2） 通过充分运用准入策略引擎，推进公司精细化、科学和标准化运作，并逐步构建量化风控体系，促进和支持 xxx 公司快速稳健的发展。

项目中根据公司业务模式，参考同行先进经验，结合公司内部和外部信息，构建风险准入模型，从而实现标准化用户准入。针对消费贷的特定环境、申请群体、申请数据和征信信息等方面，选取可以高效划分风险程度的因素，再结合现有的样本以及与业务部门的交流共同构建模型。本文档将详细描述准入策略引擎的开发过程和结果。

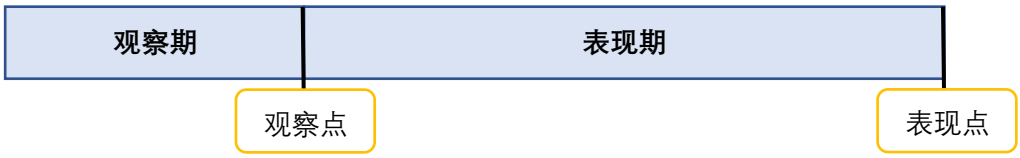
二、模型设计

1. 数据定义

模型需要确定用户的分类，包括观察期、表现期，好、坏的定义。

2. 观察期和表现期

下面为示例图：



观察期

在观察点这一时间点上，用户之前产生的数据作为构建模型使用，观察期是一个滚动窗口，一般选取 1-6 个月的数据，这样才能保证一定量的数据来开发模型。

表现期

表现期是对观察点上的用户进行监控的时间周期。用户在表现点后完成一次贷款周期才可以判定是否违约，有时候为了加强用户的区分度，会适当删除表现点后 1-16 天内的违约用户（灰色用户）。

3. 好坏和不确定定义

好坏用户是对立关系，坏用户一般根据逾期状况来定义，一般选取违约超过 16 天之后；不确定用户一般是违约在 1-16 天，这是通过逾期迁移率总结而来。

4. 开发条件

为了使模型能对未来有决策作用，模型开发需要有足够的样本，并且样本当前的状态和未来差异不大。对于一些特殊用户，如其行为和资料异常；信息缺失严重；国家规定无法方行等，出现这些情况则会特别说明。

综上所述，满足模型开发的条件如下：

1) 表现窗口一般为 6-24 个月，具体根据信贷周期决定；
2) 临近程度：数据越接近当前时期，则模型对未来的预测能力越强；
3) 代表性：模型需要遵循统计假设，即历史数据能代表未来；
4) 稳定性：历史和未来的数据分布差异不大；
5) 坏样本数量：一个稳健的模型一般需要 1500 个或以上的坏样本数量；
6) 可获性和低成本：主要考虑数据获得难易程度和成本。

5. 算法选择

本次构建模型所用到的算法是决策树（Decision Tree），决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。

由于决策树解析性强、复杂度可控和训练迅速，能够生成决策层，对于模型训练和部署来说效率很高，很适合作为准入策略的开发。

三、数据准备

1. 变量说明

明确项目开发方向，和各部门商讨后，结合 xx 公司在市场的具体情况，从个人基本信息、社会公开信息、征信数据以及第三方数据这几个方面，选取可以区分用户好坏的特征，具体看**附录 A：变量说明**

2. 窗口划分

观察期	观察月份数	表现期	表现期月份数
2016/3-2017/3	12	2016/3-2017/5	14

以下是消费贷账户的统计结果：

说明	定义	样本数量	占比
正常还款但未到最后还款日	好	123208	38.3%
到期付清	好	159880	49.7%
还款期间违约，处于宽限期	不确定	11	0.0%
逾期 1-16 天	不确定	2770	0.9%
逾期 16 天以上	坏	35823	11.1%

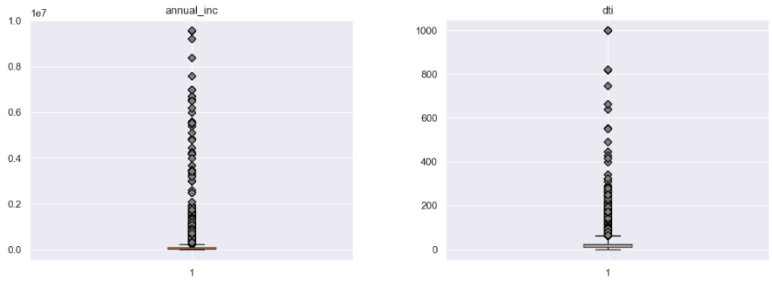
3. 缺失值和异常值处理

1) 缺失值分不同情况做不同处理：
   a) 缺失比例超 80%以上，这些变量已经严重影响模型构建，需要全部删除；
   b) 缺失值为一种状况，这些变量因为用户没有这种状况发生而缺失，经过探讨，这些变量用最大值替换更合适
   c) 缺失值没有明确状态，属于系统缺失，根据不同特点的变量给予不同填充。

2) 异常值处理：

有些变量偏态严重，且只有少数个案离群严重，对于这些连续型变量统一限制在 3 倍标准差以内。

下面为示例图：



具体填充方式看[附录 B：缺失值和异常值处理]

4. 卡方分箱

有些分类型变量由于离散值过多而不利于构建模型，需要进行分箱，这里用到分箱计数为卡方分箱（Chi Square）：

$$x^2 = \sum \frac{(observed - expected)^2}{expected}$$

卡方分箱对于比较离散的变量可以很好的合并，合并同时保持最大区分度，分箱的结果看[附录 C：卡方分箱结果]

四、变量筛选

1. 共线性筛选

对于相关性高于 0.7 以上的变量，仅保 IV 值较高的变量，共线性比较严重需要删除的变量如下：

funded_amnt_inv	tot_hi_cred_lim	revol_util
total_rev_hi_lim	num_tl_op_past_12m	open_rv_24m
pub_rec_bankruptcies	tot_cur_bal	num_rev_tl_bal_gt_0
rec_cr_line_month	bc_util	open_il_24m
total_bc_limit	funded_amnt	loan_amnt

具体查看[附录 D：变量相关性]

2. IV 筛选

IV 是一个可以辨别变量区分能力强弱的指标，一般高于 0.02 才可以认为变量对模型有预测效果。

$$IV = \sum_{i=1}^{n} \left(\frac{bad_i}{bad_T} \div \frac{good_i}{good_T} \right) \times WOE_i$$

这里将排除 IV 低于 0.05 的变量，最后进入模型训练的变量如下：

installment	all_util	dti	mo_sin_rcnt_tl
sub_grade	inq_fi	inq_last_6mths	mort_acc
home_ownership	inq_last_12m	pub_rec	mths_since_recent_bc
annual_inc	acc_open_past_24mths	revol_bal	mths_since_recent_inq
verification_status	avg_cur_bal	open_acc_6m	num_actv_rev_tl
purpose	mo_sin_old_rev_tl_op	open_il_12m	percent_bc_gt_75
addr_state	mo_sin_rcnt_rev_tl_op	mths_since_rcnt_il	verified_inc
open_rv_12m	max_bal_bc		

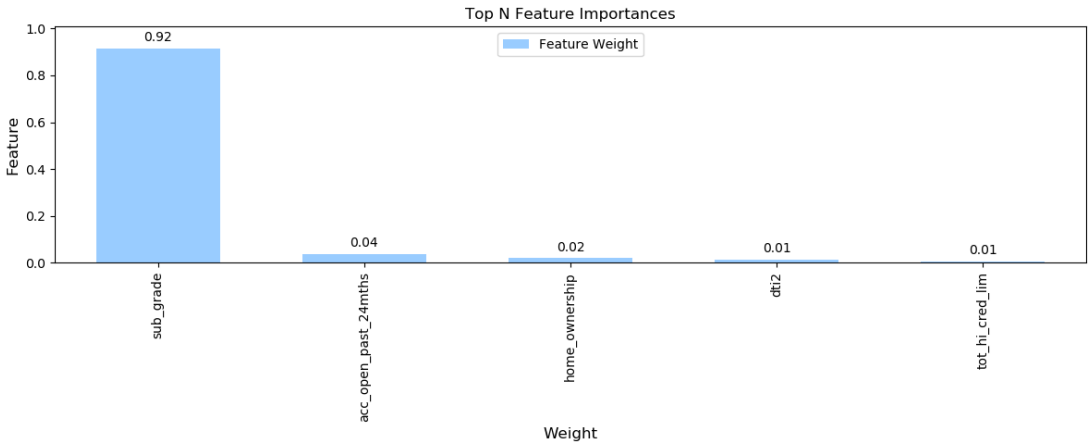
五、模型评估

1. 模型参数和变量

1) 经过模型训练后，最终确定模型最优参数为：

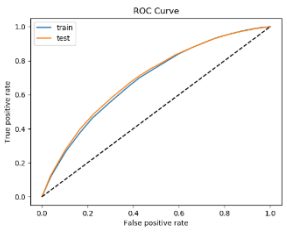
DecisionTreeClassifier(ccp_alpha=0.0, class_weight='balanced', criterion='gini',
max_depth=5, max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=0.041578947368421056,
min_samples_split=2, min_weight_fraction_leaf=0.0,
presort='deprecated', random_state=None,
splitter='best')

2) 参与决策的前五个变量为：



其中征信评分是主要策略条件，其重要程度占到 90%以上

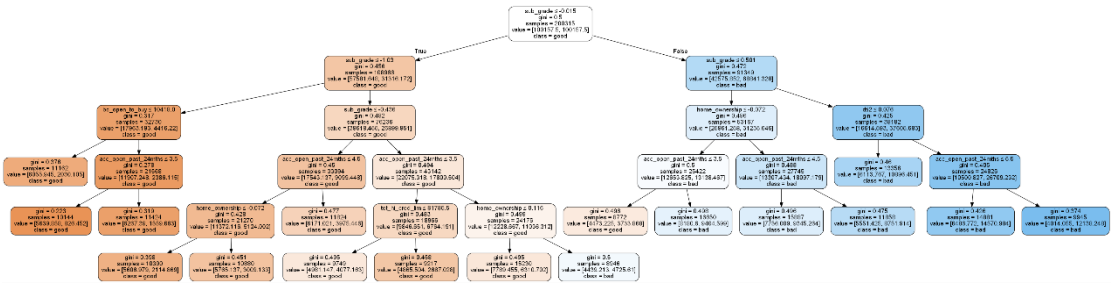
3) 模型在训练集和测试集 AUC、KS 和 Recall 得分分别如下：



AUC_train:0.687 | AUC_test:0.695
KS_train:0.272 | KS_test:0.285
Recall_train:0.697 | Recall_test:0.715

训练集和测试集评估相近，模型稳定且没有过拟合现象。

4) 准入策略详细看[附录 E：策略树]



2. KS 报告

根据测试集的 KS 报告显示，编号为 15 的一行，这一行占总人数 5%，违约率 27%（总体违约率为 11.1%），odds 为 36%。若将这一行申请用户剔除，将会减少 5%的准入账户，同时减少未来 13%的违约人数

编号	min	max	bads	goods	total	bad_rate	good_rate	odds	bad_prop	good_prop	total_prop	cum_bads	cum_goods	cum_total	cum_bads_prop	cum_goods_prop	cum_total_prop	ks
0	13%	13%	131	6,674	6,805	2%	98%	2%	1%	6%	6%	131	6,674	6,805	1%	6%	6%	-0.05
1	20%	20%	196	6,787	6,983	3%	97%	3%	2%	6%	6%	327	13,461	13,788	3%	12%	11%	-0.10
2	25%	25%	234	6,691	6,925	3%	97%	3%	2%	6%	6%	561	20,152	20,713	5%	18%	17%	-0.14
3	27%	27%	291	6,440	6,731	4%	96%	5%	2%	6%	6%	852	26,592	27,444	7%	24%	23%	-0.17
4	34%	34%	401	6,794	7,195	6%	94%	6%	3%	6%	6%	1,253	33,386	34,639	10%	31%	29%	-0.20
5	36%	36%	273	4,361	4,634	6%	94%	6%	2%	4%	4%	1,526	37,747	39,273	13%	35%	32%	-0.22
6	39%	39%	470	7,270	7,740	6%	94%	6%	4%	7%	6%	1,996	45,017	47,013	16%	41%	39%	-0.25
7	45%	45%	1,017	12,116	13,133	8%	92%	8%	8%	11%	11%	3,013	57,133	60,146	25%	52%	50%	-0.28
8	45%	45%	466	5,011	5,477	9%	91%	9%	4%	5%	5%	3,479	62,144	65,623	29%	57%	54%	-0.28
9	52%	52%	470	4,386	4,856	10%	90%	11%	4%	4%	4%	3,949	66,530	70,479	32%	61%	58%	-0.29
10	53%	53%	1,104	9,029	10,133	11%	89%	12%	9%	8%	8%	5,053	75,559	80,612	41%	69%	66%	-0.28
11	55%	55%	1,203	8,901	10,104	12%	88%	14%	10%	8%	8%	6,256	84,460	90,716	51%	77%	75%	-0.26
12	61%	61%	1,074	6,471	7,545	14%	86%	17%	9%	6%	6%	7,330	90,931	98,261	60%	83%	81%	-0.23
13	64%	64%	1,526	7,183	8,709	18%	82%	21%	13%	7%	7%	8,856	98,114	106,970	73%	90%	88%	-0.17
14	69%	69%	1,809	6,874	8,683	21%	79%	26%	15%	6%	7%	10,665	104,988	115,653	87%	96%	95%	-0.09
15	**75%**	**75%**	**1,526**	**4,197**	**5,723**	**27%**	**73%**	**36%**	**13%**	**4%**	**5%**	**12,191**	**109,185**	**121,376**	**100%**	**100%**	**100%**	-

详细看[附录 F：KS 报告]

六、稳定性评估

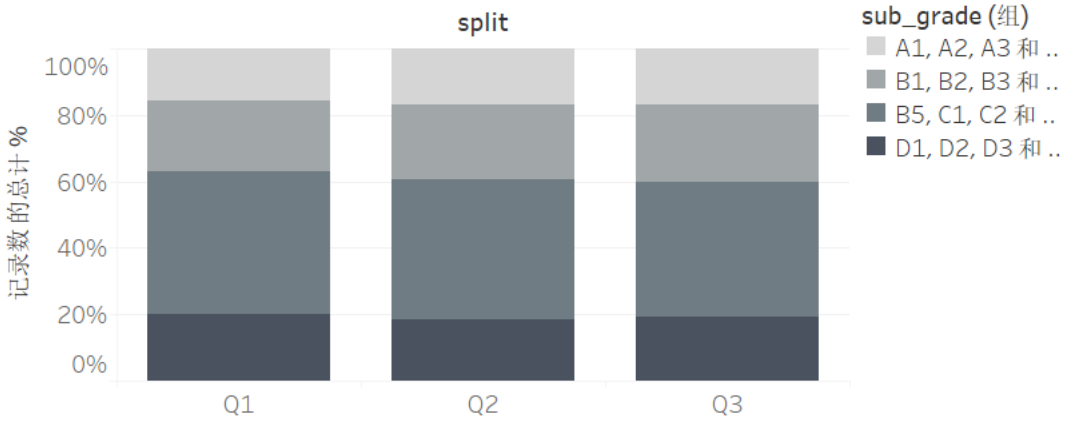
1. PSI 指标

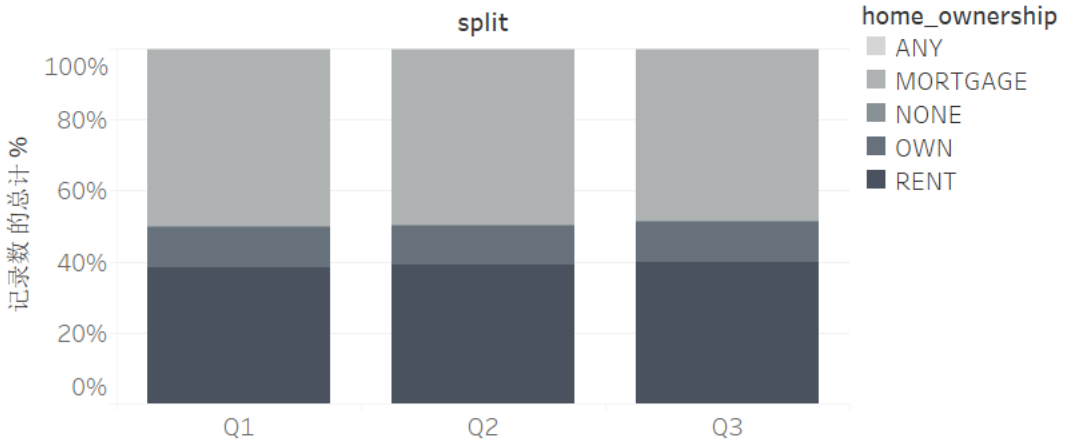
各变量 PSI 均小于 0.1，在正常范围

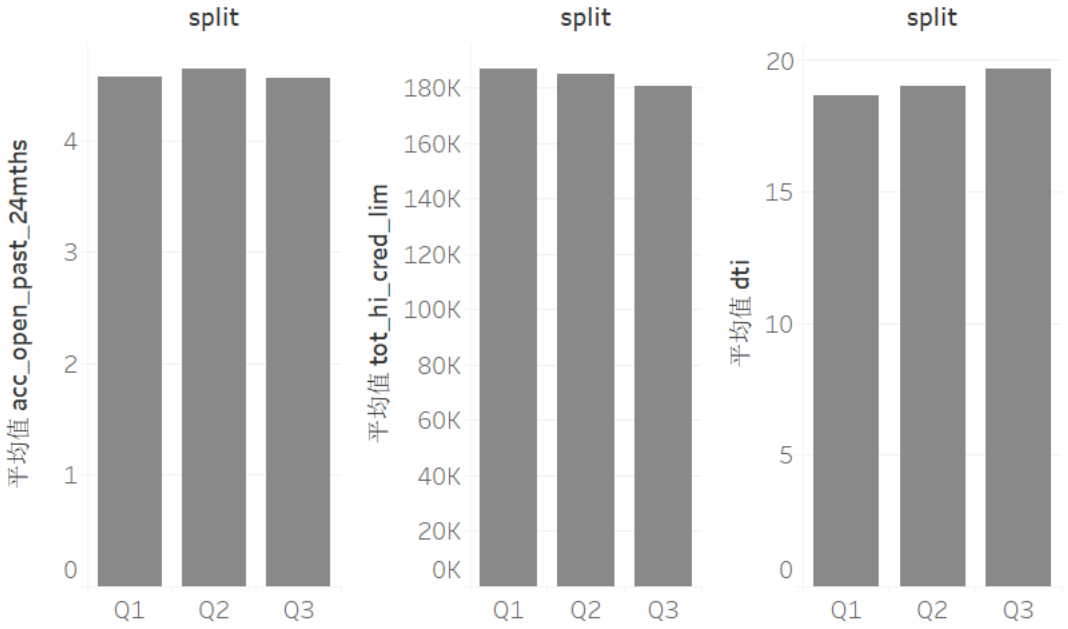
变量	PSI
sub_grade	0.00
acc_open_past_24mths	0.00
home_ownership	0.00
dti2	0.01
tot_hi_cred_lim	0.00

2. 占比和均值

各变量占比和均值稳定，波动范围在 5%以内







七、策略输出

详细看[附录 G：准入策略]

八、附件