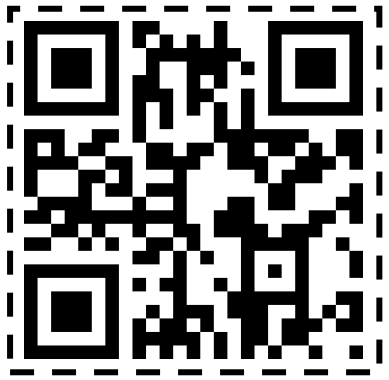
规则篇

《100天成为风控专家》

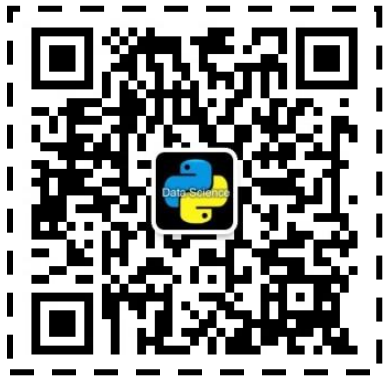
规则生成(2)：交叉表(含实操)

出品：东哥起飞

解锁风控课程



关注我的公众号



目录

三个步骤

交叉表生成规则和单变量相同，一般也是三个步骤：

对变量进行描述性统计分析，进行初步筛选；

对筛选后保留的变量进行分箱处理，计算分箱下的统计量和指标，基于IV值再次对变量进行筛选；

对再次筛选后的变量进行交叉分析，可以是二维交叉，也可以是多维交叉；

扫码加我微信

一、交叉表介绍

1.1. 交叉表的概念

什么是交叉表?

交叉表，顾名思义，就是两个或者两个以上的变量进行交叉判断。

比如下面这个示例，一个变量是“最近6个月新开信贷账户数”，另一个变量是“当前公积金状态”，这就是两个变量的交叉表形式，也叫“**二维交叉表**”，如果是两个以上的变量就是“**多维交叉表**”。

交叉表的形成，本质上就是变量的“笛卡尔积”。

Bad_rate(%)	当前公积金状态		
最近6个月新开信贷账户数	-9999	1	2
(-0.001,0.5]	4.24%	1.60%	2.97%
(0.5,1.5]	12.60%	6.12%	13.04%
(1.5,2.5]	21.74%	8.82%	17.95%
(2.5,28.0]	19.18%	12.66%	34.00%

1.2. 交叉表的特点

交叉表有什么特点?

按照规则的**复杂度**和**数据维度**两个角度来看，交叉表规则处于单变量规则和评分卡模型之间的中间形态。

与单变量规则相比，交叉表拥有更多的维度，对于客户风险评估更加准确。

与评分卡模型相比，交叉表虽变量维度少，但复杂度更低，迭代开发速度更快。

在所有的工具中，交叉表属于一种中间的形态，同时兼顾了维度和复杂度两点。

简单、维度少

一维变量

bins	Bad_rate(%)
(-0.001,0.5]	2.69%
(0.5,1.5]	9.80%
(1.5,2.5]	15.03%
(2.5,28.0]	20.30%

二维变量

Bad_rate(%)	当前公积金状态		
最近6个月新开信贷账户数	-9999	1	2
(-0.001,0.5]	4.24%	1.60%	2.97%
(0.5,1.5]	12.60%	6.12%	13.04%
(1.5,2.5]	21.74%	8.82%	17.95%
(2.5,28.0]	19.18%	12.66%	34.00%

多维变量

...

逻辑回归评分卡

复杂、维度多

1.3. 交叉表的前置条件

要生成二维交叉表，有3个前提条件：

1）基于IV筛选出预测效果好的变量池，从中选择交叉所需的变量组。一般的原则是：交叉变量最好是不同维度的，且相互间的相关性不高，这样综合效果才会达到最优。

2）对变量进行分箱操作，连续型变量需要有排序性；

仍以下面的二维交叉表为例，我们看到“最近6个月新开信贷账户数”是连续型变量，“当前公积金状态” 是离散性变量。这里公积金状态有三个离散值，因此不需要分箱；而最近6个月新开账户数由于是连续型变量，是需要做分箱处理的。

Bad_rate(%)	当前公积金状态		
最近6个月新开信贷账户数	-9999	1	2
(-0.001,0.5]	4.24%	1.60%	2.97%
(0.5,1.5]	12.60%	6.12%	13.04%
(1.5,2.5]	21.74%	8.82%	17.95%
(2.5,28.0]	19.18%	12.66%	34.00%

1.3. 交叉表的前置条件

3）总样本和坏样本数量足够多。交叉表通过两两组合，有更多的格子。比如下面一维变量只有4个格子，而二维交叉表有12个格子，而总数量和总坏客户数是相同的，那么经过稀释后交叉表的每个格子数据量会变少。如果总样本数和坏客户数不够的话，那么分散到每个格子的数量就可能出现过少，或者没有数据的情况，导致无统计意义无法分析。因此如要保证每个格子都有足够的数据，总样本和坏样本数就必须足够多。

一维变量

bins	Bad_rate(%)
(-0.001,0.5]	2.69%
(0.5,1.5]	9.80%
(1.5,2.5]	15.03%
(2.5,28.0]	20.30%

二维变量

Bad_rate(%)	当前公积金状态		
最近6个月新开信贷账户数	-9999	1	2
(-0.001,0.5]	4.24%	1.60%	2.97%
(0.5,1.5]	12.60%	6.12%	13.04%
(1.5,2.5]	21.74%	8.82%	17.95%
(2.5,28.0]	19.18%	12.66%	34.00%

扫码加我微信

二、交叉表规则生成与评估

2.1. 交叉表规则生成：三个步骤

交叉表规则制定一般有以下三个步骤：

基于透视表统计坏客户数和总客户数；

基于坏客户数和总客户数统计量，计算出区间坏账率(badrate)和客户数占比；

基于格子的区间坏账率和客户占比制定规则。

2.2. 交叉表规则生成(1)：透视表

步骤一：透视表统计坏客户数和总客户数。

客户id	最近6个月新开	当前公积金状	当前逾期
0	1	1	0
1	3	1	0
2	0	1	0
3	0	1	0
4	2	1	0
5	0	2	0
6	1	2	0
7	0	2	0
8	4	-9999	1
9	1	1	0
10	0	1	0
11	2	-9999	0
12	7	2	1
13	1	2	0
14	0	1	0

透视表

透视表

对当前逾期“求和”

坏客户数	当前公积金状态			
最近6个月新开信贷	-9999	1	2	ALL
(-0.001,0.5]	19	11	8	38
(0.5,1.5]	16	9	9	34
(1.5,2.5]	10	6	7	23
(2.5,28.0]	14	10	17	41
ALL	59	36	41	136

对当前逾期“计数”

总客户数	当前公积金状态			
最近6个月新开信贷	-9999	1	2	ALL
(-0.001,0.5]	429	677	261	1367
(0.5,1.5]	111	138	60	309
(1.5,2.5]	36	62	32	130
(2.5,28.0]	59	69	33	161
ALL	635	946	386	1967

2.3. 交叉表规则生成(2)：计算指标

步骤二：基于坏客户数和总客户数统计量，计算出区间坏账率(badrate)和客户数占比。

坏客户数	当前公积金状态			
最近6个月新开信贷	-9999	1	2	ALL
(-0.001,0.5]	19	11	8	38
(0.5,1.5]	16	9	9	34
(1.5,2.5]	10	6	7	23
(2.5,28.0]	14	10	17	41
ALL	59	36	41	136

badrate	当前公积金状态			
最近6个月新开信贷账户数	-9999	1	2	ALL
(-0.001,0.5]	4.43%	1.62%	3.07%	2.78%
(0.5,1.5]	14.41%	6.52%	15.00%	11.00%
(1.5,2.5]	27.78%	9.68%	21.88%	17.69%
(2.5,28.0]	23.73%	14.49%	51.52%	25.47%
ALL	9.29%	3.81%	10.62%	6.91%

总客户数	当前公积金状态			
最近6个月新开信贷	-9999	1	2	ALL
(-0.001,0.5]	429	677	261	1367
(0.5,1.5]	111	138	60	309
(1.5,2.5]	36	62	32	130
(2.5,28.0]	59	69	33	161
ALL	635	946	386	1967

客户占比	当前公积金状态			
最近6个月新开信贷账户数	-9999	1	2	ALL
(-0.001,0.5]	21.81%	34.42%	13.27%	69.50%
(0.5,1.5]	5.64%	7.02%	3.05%	15.71%
(1.5,2.5]	1.83%	3.15%	1.63%	6.61%
(2.5,28.0]	3.00%	3.51%	1.68%	8.19%
ALL	32.28%	48.09%	19.62%	100.00%

区间坏账率=每个格子的坏客户数/对应格子的总客户数，是上下两个交叉表每个格子对应位置的计算，比如蓝色框示例，4.43%=19/429；

客户占比=每个格子的客户数/总客户数，只需总客户数一个交叉表即可，比如红色框示例，3%=59/1967；

2.4. 交叉表规则生成(3)：制定和评估

步骤三：基于格子的区间坏账率和客户占比制定规则。

badrate	当前公积金状态			
最近6个月新开信贷账户数	-9999	1	2	ALL
(-0.001,0.5]	4.43%	1.62%	3.07%	2.78%
(0.5,1.5]	14.41%	6.52%	15.00%	11.00%
(1.5,2.5]	27.78%	9.68%	21.88%	17.69%
(2.5,28.0]	23.73%	14.49%	51.52%	25.47%
ALL	9.29%	3.81%	10.62%	6.91%

客户占比	当前公积金状态			
最近6个月新开信贷账户数	-9999	1	2	ALL
(-0.001,0.5]	21.81%	34.42%	13.27%	69.50%
(0.5,1.5]	5.64%	7.02%	3.05%	15.71%
(1.5,2.5]	1.83%	3.15%	1.63%	6.61%
(2.5,28.0]	3.00%	3.51%	1.68%	8.19%
ALL	32.28%	48.09%	19.62%	100.00%

生成规则：“最近6个月新开信贷账户数在(2.5,28]之间”且“当前公积金状态为2”，触发则拒绝，反之通过。

扫码加我微信

三、交叉表应用场景

3.1. 策略D类调优

扫码加我微信

➢ 背景介绍

某机构发现，近期市场环境不好，客户的贷后逾期率不断升高，业务部门提出需求：需要风控策略人员对贷前审批策略进行收紧，降低逾期风险，但同时不降低太多通过率，因为业务规模是本年的考核指标。

➢ 策略方案

该需求属于策略D类调优。可新增二维交叉表规则，比如右侧这条规则，命中率仅为1.68%，但拒绝客户中一半以上都是坏客户。

如果使用单变量规则，比如最近6个月新开信贷账户数>=3时拒绝，区间坏账率为25.47%，命中率则为8.19%，会降低很大通过率 。

badrate	当前公积金状态			
最近6个月新开信贷账户数	-9999	1	2	ALL
(-0.001,0.5]	4.43%	1.62%	3.07%	2.78%
(0.5,1.5]	14.41%	6.52%	15.00%	11.00%
(1.5,2.5]	27.78%	9.68%	21.88%	17.69%
(2.5,28.0]	23.73%	14.49%	51.52%	25.47%
ALL	9.29%	3.81%	10.62%	6.91%

客户占比	当前公积金状态			
最近6个月新开信贷账户数	-9999	1	2	ALL
(-0.001,0.5]	21.81%	34.42%	13.27%	69.50%
(0.5,1.5]	5.64%	7.02%	3.05%	15.71%
(1.5,2.5]	1.83%	3.15%	1.63%	6.61%
(2.5,28.0]	3.00%	3.51%	1.68%	8.19%
ALL	32.28%	48.09%	19.62%	100.00%

二维交叉表规则： “最近6个月新开信贷账户数在(2.5,28]之间”且“当前公积金状态为2”，触发则拒绝，反之通过。

扫码加我微信

四、Python代码案例实操

4.1. 代码示例