同盾大学
xue.tongdun.cn

让|诚|信|生|活|更|美|好

标签定义和特征工程

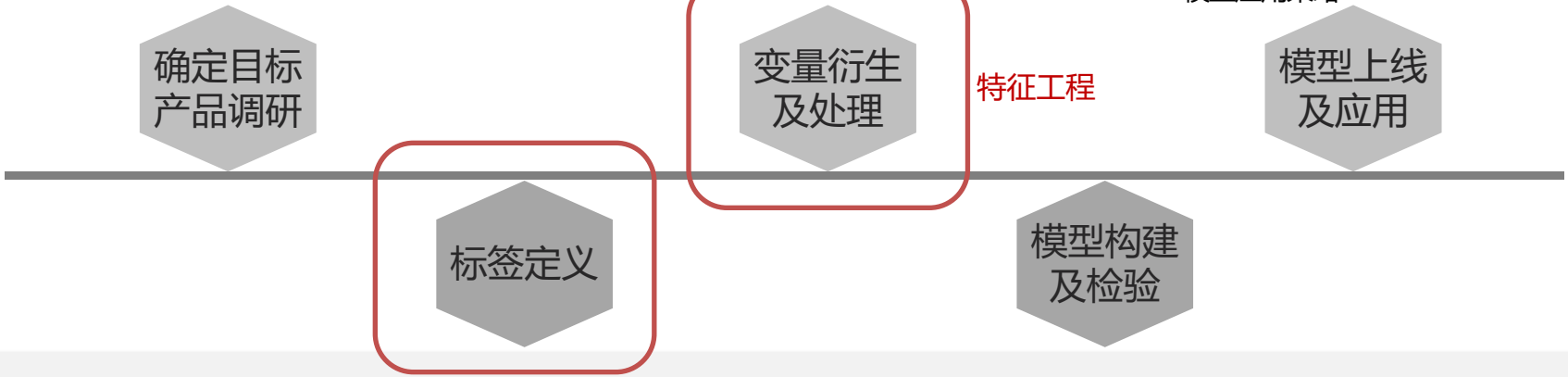
C 课程目录
ONTENTS

回顾——评分卡示例

了解产品基本信息，现有流程，放款量，产品期限等产品要素，历史政策变化情况，确定入模的样本

根据原始数据生成衍生变量，分箱并计算每个变量的区分能力

在系统里配置模型上线，结合业务实际流程和风险偏好制定模型应用策略



对客户贷款表现进行vintage分析，roll rate分析，根据结果以及产品期限在结合行方产品实际情况确定，观察期，好坏客户的定义

模型训练，并测试模型的稳定性，区分度。

回顾——评分卡示例

变量分类	变量名	变量解释	变量分组	分数
企业基本情况	setup_y	企业营业时间	缺失	
			小于9年	
			「9，13）年	
			大于等于13年	
	total_asset	企业总资产	缺失	
			4252万以下（小于）	
			[4252，5758）万	
			[5758，14185）万	
			大于等于14185万	
存款情况	last3m_cnt	最近3个月本行内累计贷方发生笔数（收入）	缺失	
			小于10次	
			[10，28）次	
			[28，99）次	
			大于等于99次	
	td_self_bankacc_12mth_osavg	企业最近1年在本行的平均存款额	缺失	
			小于等于12522元	
			「12522，88414）元	
			[88414，583204）元	
			大于等于583204元	
	td_self_bankacc_curr_osavg	企业最近1个月在本行的平均存款额	缺失	
			小于等于4221	
			[4221，71013）元	
			[71013，1118644）元	
			1118644元	
贷款信用卡情况	loan_balance	本行当前贷款余额	小于等于88万	
			[88，202）万	
			[202，500）万	
			大于等于500万	
	td_Loan_emi	每个月应还本行贷款金额	小于等于45054元	
			[45054，197667）元	
			[197667，689791）元	
			大于等于689791元	
	td_owner_cc_6mth_util_avg	企业主最近6个月信用卡平均使用额度	缺失	
			小于等于60%	
			[60%，96%）	
			大于等于96%	

课程目录
CONTENTS

不同评分卡标签定义

■ 申请评分卡（A卡）

样本为所有的申请客户,包括通过和拒绝客户。通过客户利用实际贷款表现来打标，拒绝客户利用拒绝推断来打标。

	描述
观察期	申请前一段时间
表现期	放款后一段时间
目的	预测客户在放款后一段时间内变坏的概率

■ 行为评分卡（B卡）

样本为某时刻点（如年末）所有满观察期和表现期且在该时刻点上还款状态为正常（无逾期），且在观察期内无严重逾期记录的客户。按在一段贷款表现达标。

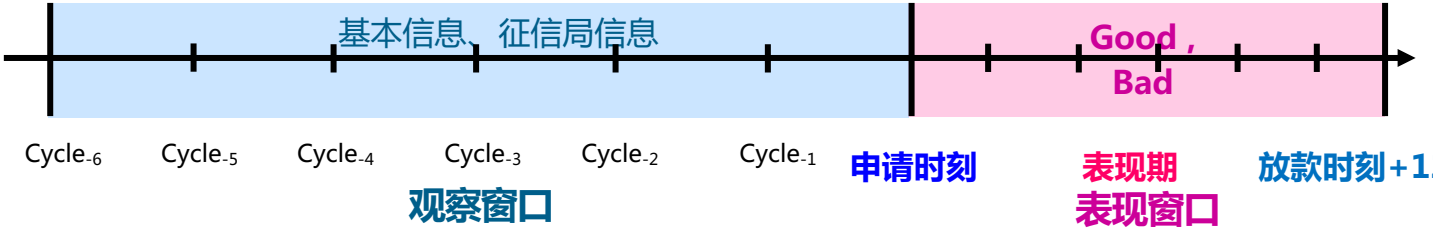
	描述
观察期	某时刻点前一段还款期
表现期	某时刻点后一段还款期
目的	用正常客户之前还款行为预测之后一段时间的还款行为

■ 催收评分卡（C卡）

样本为某一时刻点（如年末）所有满观察期和表现期且在该时刻点上还款状态为关注（轻度逾期），且在观察期内无严重逾期记录的客户。按在短时间内信用恶化情况打标。

	描述
观察期	某时刻点前一段还款期
表现期	某时刻点后一段还款期
目的	用轻度逾期客户之前还款行为预测之后一段时间内迅速变坏的概率

- **模型框架**
  - 根据客户基本信息、标的信息和过去的征信
  - 预估未来12个月内，发生严重拖欠的概率



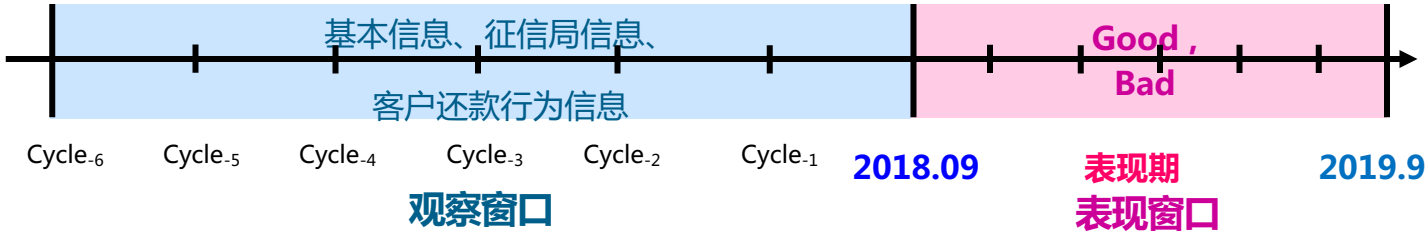
全部客户群

- 客群：所有申请客户（通过拒绝
- 排除客户：信用政策差异，欺诈客户……

坏客户定义

- 在放款后12个月内发生过M2+的客户
- 在拒绝推断中找到的坏客户

- **模型框架**
  - > 根据客户实时基本信息、三方数据信息，截止计算时刻的征信、还款历史
  - > 预估未来表现期n个月内(较长期) ，发生严重拖欠的概率
  - > 更适用于还款期较长，信用卡循环贷



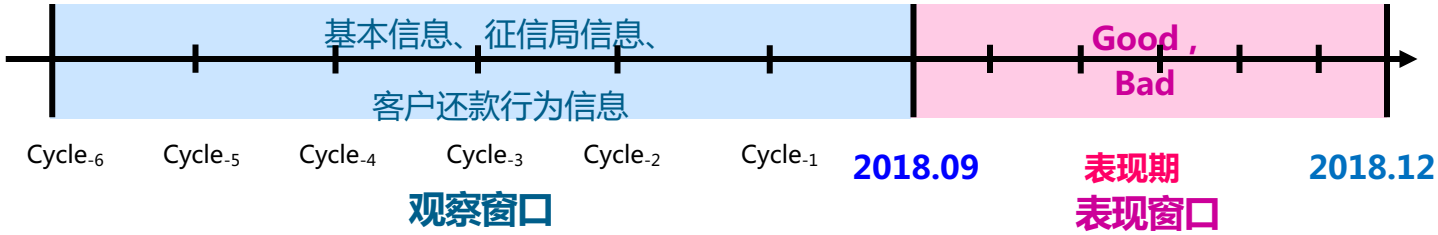
全部客户群

- 客群：在2018年9月没有过逾期的正常客户
- 排除客户：**MOB小于n个月客户，历史有过严重逾期的客户**，欺诈客户，政策还款方式不同.....

坏客户定义

- 在表现期内12月内发生过M2+的客户

- **模型框架**
  - 根据客户基本信息、标的信息和过去的征信、还款数据
  - 预估未来n个月内（短期），发生严重逾期的概率



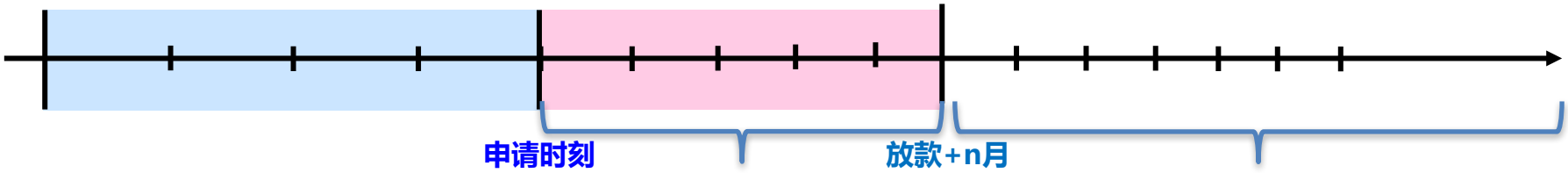
全部客户群

- 客群：在2018年9月有过逾期客户
- 排除客户：**MOB小于n个月客户，历史有过严重逾期的客户**，欺诈客户，政策还款方式不同……

坏客户定义

- 在表现期内从M1变到M3
- 从M2到M3并且未来两个月未还款

不同评分卡标签定义——ABC卡的应用



正常：A卡

正常：B卡

逾期：C卡

C 课程目录
ONTENTS

数据收集

收集历史上已收集的数据，或请三方数据公司回溯。
行方内部数据：申请表，PBOC数据等；
外部数据：运营商、多头等

数据清理

质量和准确性（业务上有没有验证是否客观）；二义性（有二义性的建立数据标准）；覆盖率（过低的去除）；极值处理（Capping and Flooring）；缺失值处理（mean，mode等）

特征编码

机器学习模型需要的数据是数字型的，因为只有数字类型才能进行计算。因此，对于各种特殊的特征值，我们都需要对其进行相应的编码，也是量化的过程

预筛选

对单一变量分别进行筛选；
在组合的层面上对变量整体进行筛选。

Step1. 数据收集：

对于X:收集行方已保存的数据，或请三方数据公司回溯。行方内部数据：申请表，PBOC数据等；

外部数据：运营商、多头等；

对于Y：收集行方已保存的还款表现数据，申请时间放款时间，审批结论，拒绝原因，催收纪要。

Step2. 数据清洗：

质量和准确性：业务上有没有验证？是否客观？

二义性：有二义性的建立数据标准？

覆盖率：是否过低？

极值处理：去除不合理的极值（Capping and Flooring）

缺失值处理：用mean，mode等代替，单独分为一组，并入坏账率相似的一组。

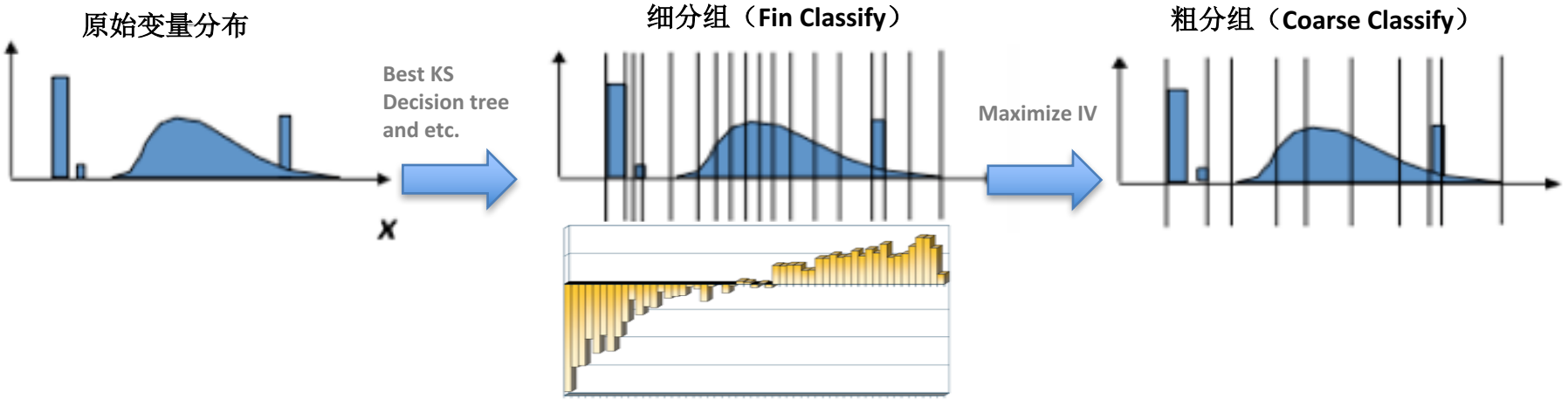
特征编码

1. **标准化**：去除量纲，0-1化处理
2. **二值化**：设定阈值，非0即1.
3. **类别编码**：将有序的离散型变量映射成数字等级，比如学历：小学为1，初中为2，高中为3，大学为4，研究生为5，博士为6....
4. **哑编码**：哑编码是一种状态编码，属于一对多的特征映射。简单点讲，它将性别映射为两个变量：是否是男性、是否是女性。它解决了 LabelEncoder 中序的问题，比如在 LabelEncoder 中，女性用2表示，但明显不可能是2倍的男性。
5. **WOE编码**：WOE 即证据权重, WOE的值越大代表对应的变量对“是好人”的贡献就越大，反之，越小就代表对应的变量对“是坏人”的贡献越大。所以WOE值可以作为特征的一种编码方式

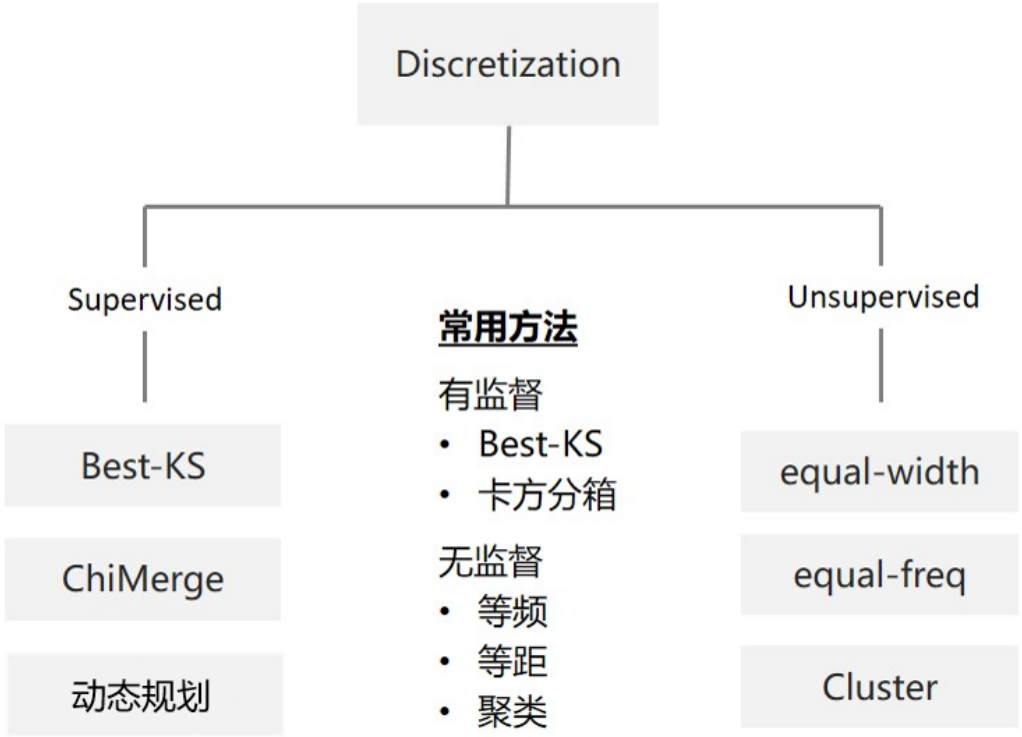
分箱

定义：将连续变量离散化，找到几个节点分成几个组。

目的：通过分箱可以增强模型的稳定性、健壮性，增强模型的解释性。



常见变量分箱算法



1. Best-KS 算法-自上而下

1.1 设置初始参数: 分组数量、分组最小占比、分组最小样本数

1.2 缺失值单独分为一组

1.3 统计每一个指标取值下的 **坏样本数、**好样本数**、总样本数、bad_rate，并升序排列**.

1.4 找到ks最大的点，根据该点将样本切分为2组

1.5 判断切分后的2组是否满足初始参数条件，如果满足，则重复1.4 步骤，否则停止分箱.

1.6 通过1.4,1.5 步骤, 我们可以获得一系列最佳ks点对应的切分点. 然后我们根据最终需要的分组数从已有的切分点中遍历每一种组合，选择iv最大的作为最优切分点.

1.7最后根据最优的切分点计算出相应的分组信息.

2. 卡方算法-自下而上

1.1 设置初始参数: 分组数量、分组最小占比、分组最小样本数、卡方统计量阈值

1.2 缺失值单独分为一组

1.3 统计每一个指标取值下的 **坏样本数、**好样本数**、总样本数、bad_rate，并升序排列**.

1.4 遍历所有的组, 计算每相邻两组的卡方统计量, 选择最小的卡方统计量, 如果最小的卡方统计量大于卡方统计量阈值，则停止分箱，否则将该最小卡方统计量对应的两组合并，并判断是否满足初始条件.

1.5 重复1.4步骤可以逐步减少总分组数量，直至达到初始设定的分组数量

1.6最后根据最优的切分点计算出相应的分组信息

3. 动态规划算法-全局最优

1.1 设置初始参数: 分组数量、分组最小占比、分组最小样本数

1.2 缺失值单独分为一组

1.3 统计每一个指标取值(n个取值)下的 **坏样本数、**好样本数**、总样本数、bad_rate，并升序排列**.

1.4 然后我们目标要找到递归方程式，令 **M[n][k]** 代表指标在含有n个不同取值切切分为k组的最大因此我们可以发现递归方程式如下:

$$M[i][j] = \max_{0 \le index < i} \left(M[index][j-1] + \left(\frac{pg[index:]}{total_{good}} - \frac{pb[index:]}{total_{bad}}\right) / \ln\left(\frac{pg[index:]}{total_{good}} * \frac{total_{bad}}{pb[index:]}\right) \right)$$

$total_{good}$ 是总好样本数, $total_{bad}$ 是坏样本数. pg 每个指标取值下的好样本数列表, pb每个指标取值下的坏样本数列表.

切满足初始条件:

$$M[i][1] = 0 \; for \; each \; i; M[1][j] = 0 \; for \; each \; \mathbf{j}$$

1.5 根据初始条件和迭代方程式我们可以求解方程. **M[n][piece]** 就是满足条件的最优分组的iv, 对应的切分点几位最优切分点.

1.6最后根据最优的切分点计算出相应的分组信息

为什么要对变量进行离散化？

1. 离散特征的增加和减少都很容易，易于模型的快速迭代；
2. 稀疏向量内积乘法运算速度快，计算结果方便存储，容易扩展；
3. 离散化后的特征对异常数据有很强的鲁棒性：比如一个特征是年龄>30是1，否则0。如果特征没有离散化，一个异常数据“年龄300岁”会给模型造成很大的干扰；
4. 逻辑回归属于广义线性模型，表达能力受限；单变量离散化为N个后，每个变量有单独的权重，相当于为模型引入了非线性，能够提升模型表达能力，加大拟合；
5. 离散化后可以进行特征交叉，由M+N个变量变为M*N个变量，进一步引入非线性，提升表达能力；
6. 特征离散化后，模型会更稳定，比如如果对用户年龄离散化，20-30作为一个区间，不会因为一个用户年龄长了一岁就变成一个完全不同的人。当然处于区间相邻处的样本会刚好相反，所以怎么划分区间是门学问；
7. 特征离散化以后，起到了简化了逻辑回归模型的作用，降低了模型过拟合的风险。

Step4. 变量预筛选

变量筛选规则如下（可以增加筛选规则）：

优先级顺序	规则名称	规则详细逻辑	规则参数设置	规则状态	备注
CS01	单一阈值筛选	1、计算每个变量的每种取值占比 2、设定单一阈值阀值 3、遍历每个变量： 如果max(变量的不同取值占比)> 单一阈值阀值，则剔除变量；否则保留变量	单一阈值=0.9	启用	每一次筛选都是承接上一次筛选结果
CS02	最小变量分组数量	1. 变量分组默认分为5组，如果存在缺失值，则分6组，缺失值单独分为一组；分组要求，每组数量占比不得低于5% 2. 如果除去缺失值，剩下的无法分出大于等于“最小变量分组数量”的组数，则剔除变量;否则保留	最小变量分组数量=2	启用	
CS03	IV筛选	1、计算每个变量的最优分组 2、设定IV阀值 3、遍历每个变量： 如果变量IV< 单一阈值阀值，则剔除变量；否则保留变量	IV阀值=0.02	启用	
CS04	PSI筛选	1. 如果建模样本的时间点跨度几个月，甚至更长，可以按月切分或者更长步长切分 2. 选择时间最早的一组作为基本样本，剩下的作为验证样本，计算组间的PSI，剔除PSI大于"PSI阀值"的变量	PSI阀值=0.05	启用	该部分视实际样本而定
CS05	相关系数筛选	1、计算所有变量之间的相关性 2、设定相关系数阀值 3、遍历每个变量： 选择出与该变量相关系数大于相关系数阀值的所有变量，只保留这些变量中IV最大额那个变量，其余变量删除	相关系数阀值=0.6	启用	

模型搭建—特征工程

CS06	AR筛选	1、设定随机切分比例 2、随机切分数据集为训练集、测试集，计算训练集、测试集每个变量的AR值 3、遍历每个变量： 如果该变量在训练集、测试集上的 AR值负号相反，则剔除变量，否则保留	随机切分比例=6:4	未启用	
CS07	MIC筛选	1、计算所有变量之间的最大互信息数MIC 2、设定MIC阈值 3、遍历每个变量： 如果该变量的MIC<MIC阈值，则剔除；否则保留	MIC阈值=0.1	未启用	
CS08	皮尔逊卡方检验筛选	1、计算所有变量的皮尔逊卡方检验统计量的pvalue 2、设定置信水平alpha 3、遍历每个变量： 如果该变量的pvalue>alpha，则剔除；否则保留	alpha=0.05	未启用	
CS09	似然比检验筛选	1、计算所有变量的似然比检验统计量的pvalue 2、设定置信水平alpha 3、遍历每个变量： 如果该变量的pvalue>alpha，则剔除；否则保留	alpha=0.05	未启用	
CS10	F检验筛选	1、计算所有变量的F检验统计量的pvalue 2、设定置信水平alpha 3、遍历每个变量： 如果该变量的pvalue>alpha，则剔除；否则保留	alpha=0.05	未启用	

Note: 该部分只展示部分常见的变量筛选规则，需要根据实际场景制定相应的变量筛选规则体系。

同盾科技
www.tongdun.cn

THANKS