法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

申请评分卡中的数据预处理和特征衍生（II）

目录

分箱的注意点(续)

特征信息度的计算和意义

单变量分析

多变量分析

分箱的注意点(续)

分箱的注意点(续)

对于连续型变量，

- 使用ChiMerge进行分箱(默认分成5个箱)
- 检查分箱后的bad rate单调性；倘若不满足，需要进行相邻两箱的合并，直到bad rate为止
- 上述过程是收敛的，因为当箱数为2时，bad rate自然单调
- 分箱必须覆盖所有训练样本外可能存在的值！

分箱的注意点(续)

对于类别型变量，

- 当类别数较少时，原则上不需要分箱
- 当某个或者几个类别的bad rate为0时，需要和最小的非0bad rate的箱进行合并
- 当该变量可以完全区分目标变量时，需要认真检查该变量的合理性
- 例如：“该申请者在本机构历史信用行为”把客群的好坏样本完全区分时，需要检查该变量的合理性(有可能是事后变量)

目录

特征信息度的计算和意义

特征信息度

IV(Information Value), 衡量特征包含预测变量浓度的一种指标

	Good	Bad	Good% (1)	Bad% (2)	WOE Log(1/2)	IV (1-2)*WOE
Group 1	G_1	B_1	G_1/G	B_1/B	$\log(\frac{G_1/G}{B_1/B})$	$(G_1/G-B_1/B)*\log(\frac{G_1/G}{B_1/B})$
Group 2	G_2	B_2	G_2/G	B_2/B	$\log(\frac{G_2/G}{B_2/B})$	$(G_2/G-B_2/B)*\log(\frac{G_2/G}{B_2/B})$
…						
Group N	G_N	B_N	G_N/G	B_N/B	$\log(\frac{G_N/G}{B_N/B})$	$(G_N/G-B_N/B)*\log(\frac{G_N/G}{B_N/B})$
Total	$G=\sum G_i$	$B=\sum B_i$				$\sum(\frac{G_i}{G}-\frac{B_i}{B})\times\log(\frac{G_i/G}{B_i/B})$

特征信息度的计算和意义

特征信息度的解构

$$\mathrm{IV}_i = (G_i - B_i) \times \log\left(\frac{G_i}{B_i}\right) = (G_i - B_i) \times WOE_i$$

其中，G_i，B_i代表箱i中好坏样本占全体好坏样本的比例

WOE: 衡量两类样本分布的差异性

$(G_i - B_i)$：衡量差异的重要性

例如：$G_1 = 0.2, B_1 = 0.1$ 与 $G_2 = 0.02, B_2 = 0.01$

$WOE_1 = WOE_2 = \log(2)$

$IV_1 = (0.2 - 0.1) \times \log(2) = 0.1 \times \log(2)$

$IV_2 = (0.02 - 0.01) \times \log(2) = 0.01 \times \log(2)$

特征信息度的计算和意义

□ 特征信息度的作用

挑选变量

- 非负指标
- 高IV表示该特征和目标变量的关联度高
- 目标变量只能是二分类
- 过高的IV，可能有潜在的风险
- 特征分箱越细，IV越高
- 常用的阈值：

<=0.02: 没有预测性，不可用

0.02 to 0.1: 弱预测性

0.1 to 0.2: 有一定的预测性

0.2 +: 高预测性

目录

信用风险中的单变量分析

多类别离散变量和连续变量的分箱的注意点

多类别离散变量

- 以bad rate代替原有值，转化成连续型变量再分箱
- 例：UserInfo_2，原始值有327个城市，分箱后有5个组别

连续型变量

- 把特殊值单独化为一组
- ThirdParty_Info_Period4_1：$-1 \cup [0, 1506]$
- -1单独分为一组

信用风险中的单变量分析

- 单变量分析

以分箱后的WOE为值

一、用IV检验有效性

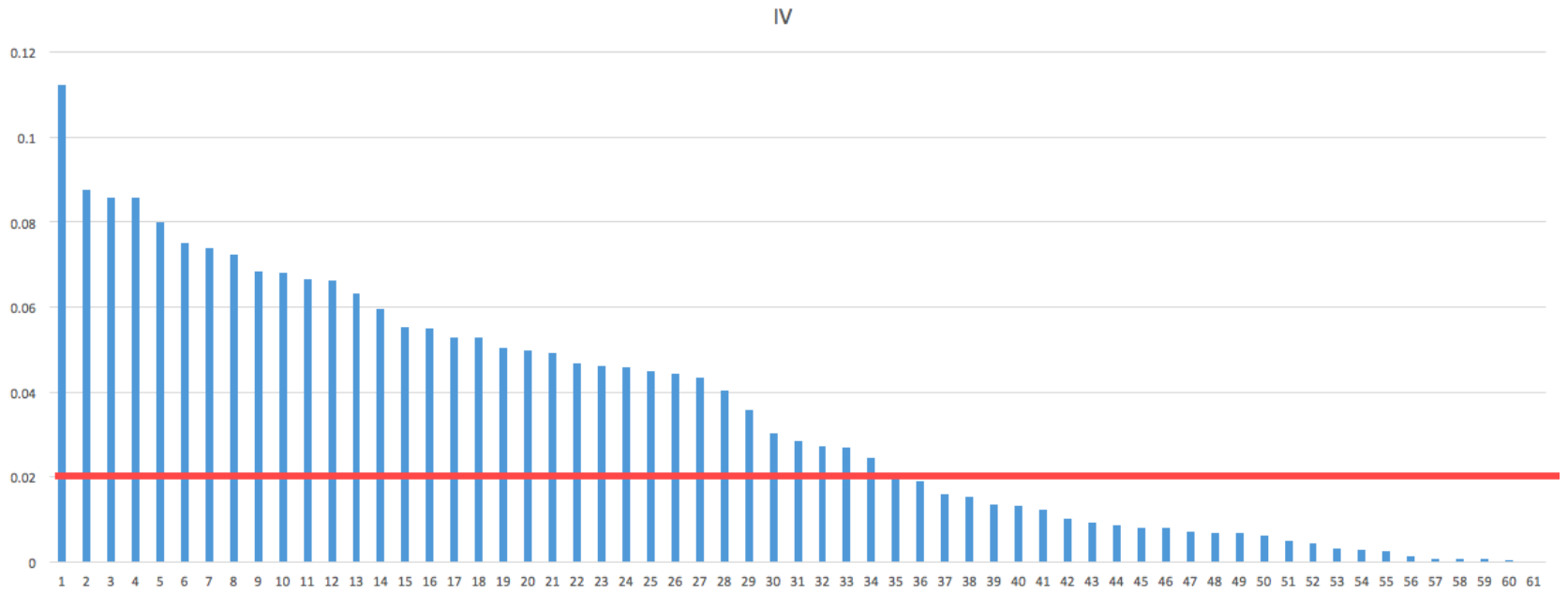
二、连续变量bad rate的单调性(可以放宽到U型)

三，单一区间的占比不宜过高

信用风险中的单变量分析

IV分布

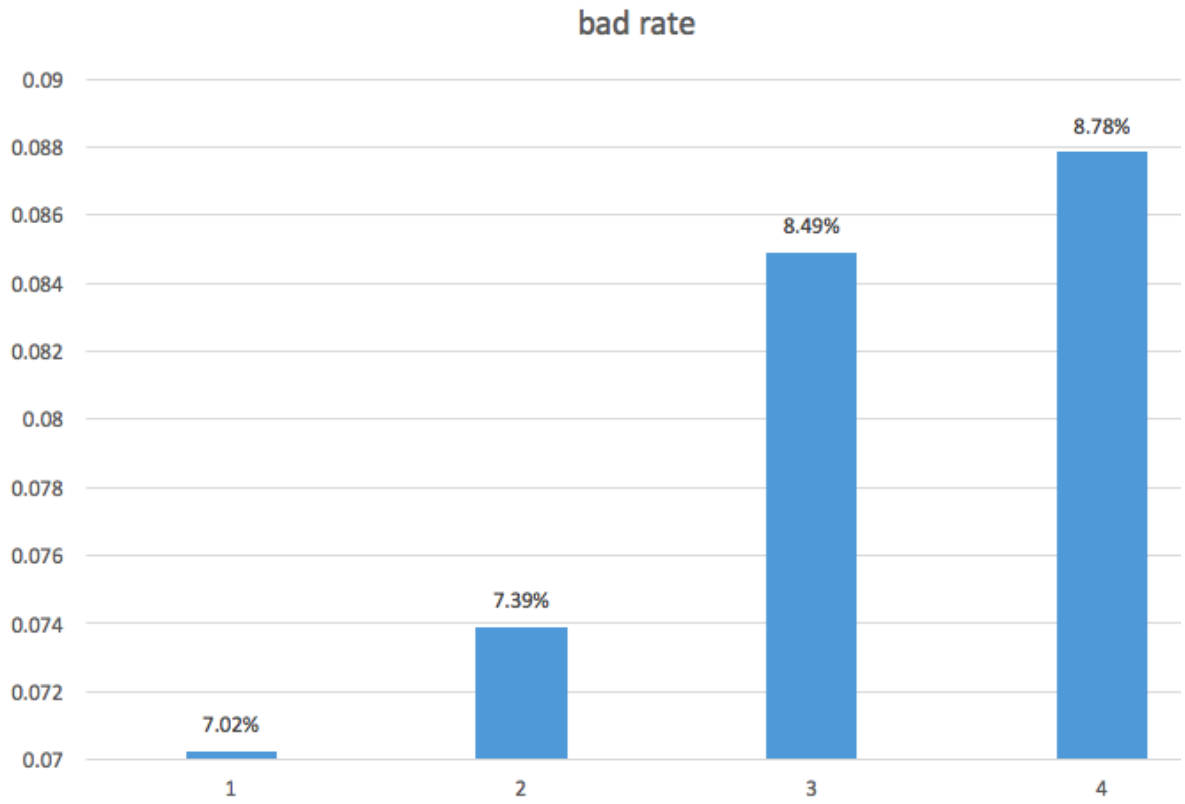
随机挑选的60个变量的IV值(分箱后)



信用风险中的单变量分析

连续变量bad rate单调性

以分箱后的‘ThirdParty_Info_Period2_7’为例

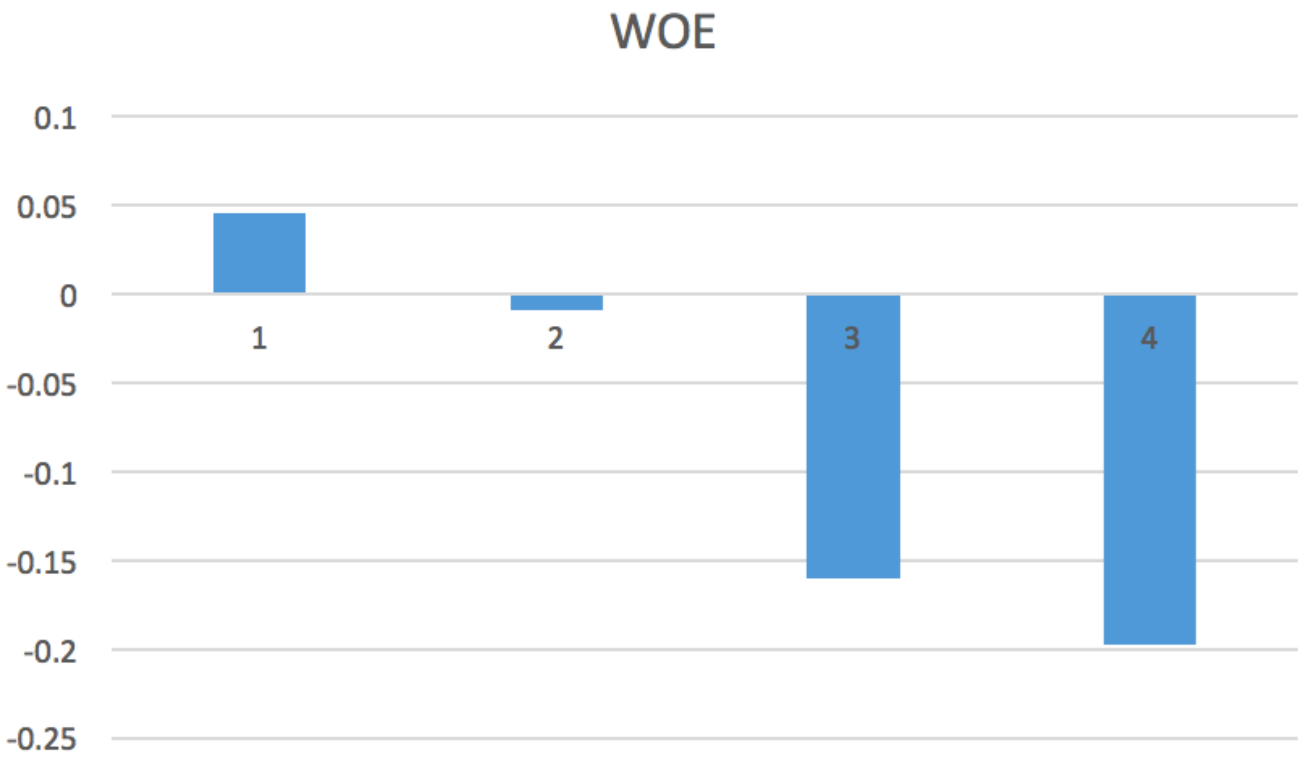


信用风险中的单变量分析

连续变量的WOE

以分箱后的‘ThirdParty_Info_Period2_7’为例

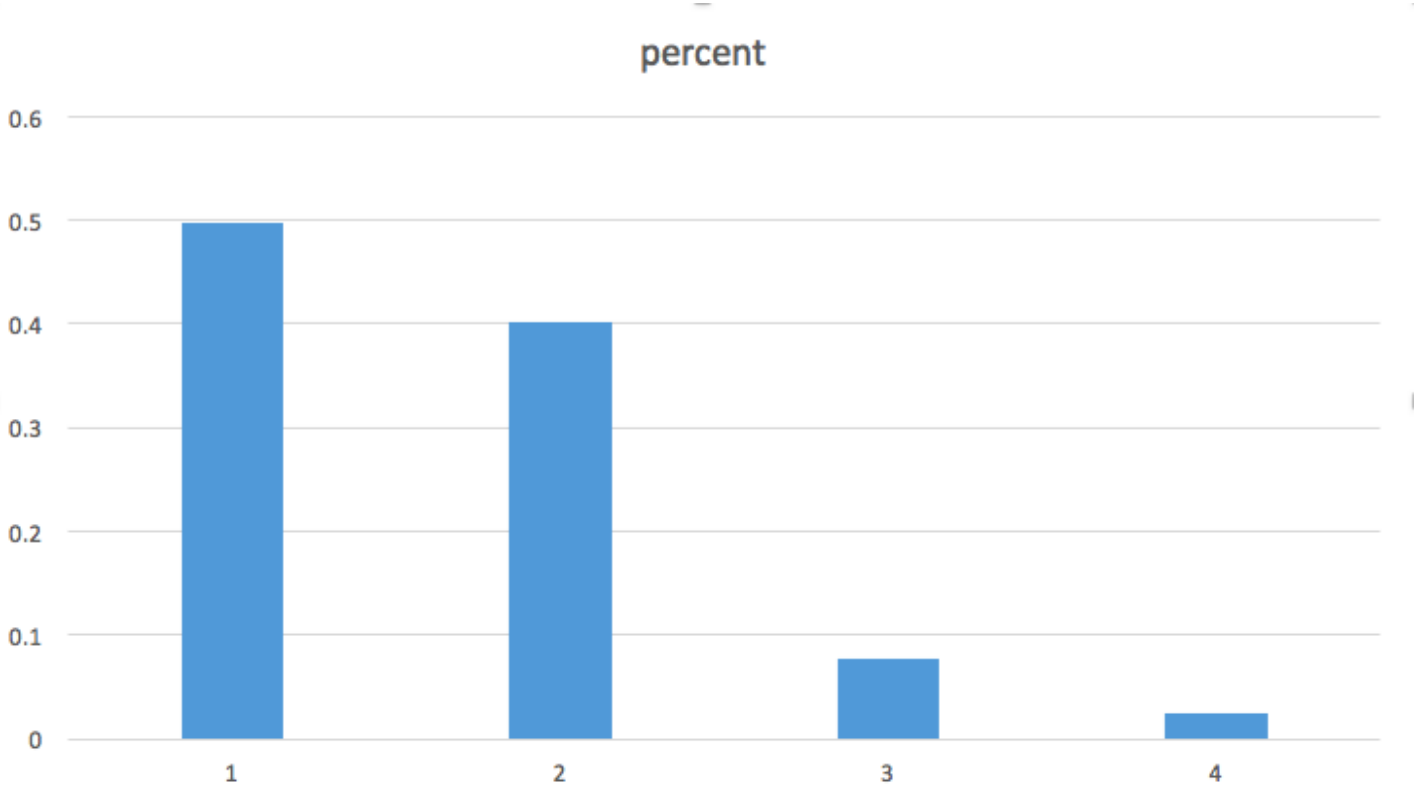
注：在某些场合下，要求占比最多的箱的WOE为0.此时需要整体平移。



信用风险中的单变量分析

单一区间的占比

以分箱后的‘ThirdParty_Info_Period2_7’为例



目录

多变量分析

- 多变量分析：变量的两两相关性

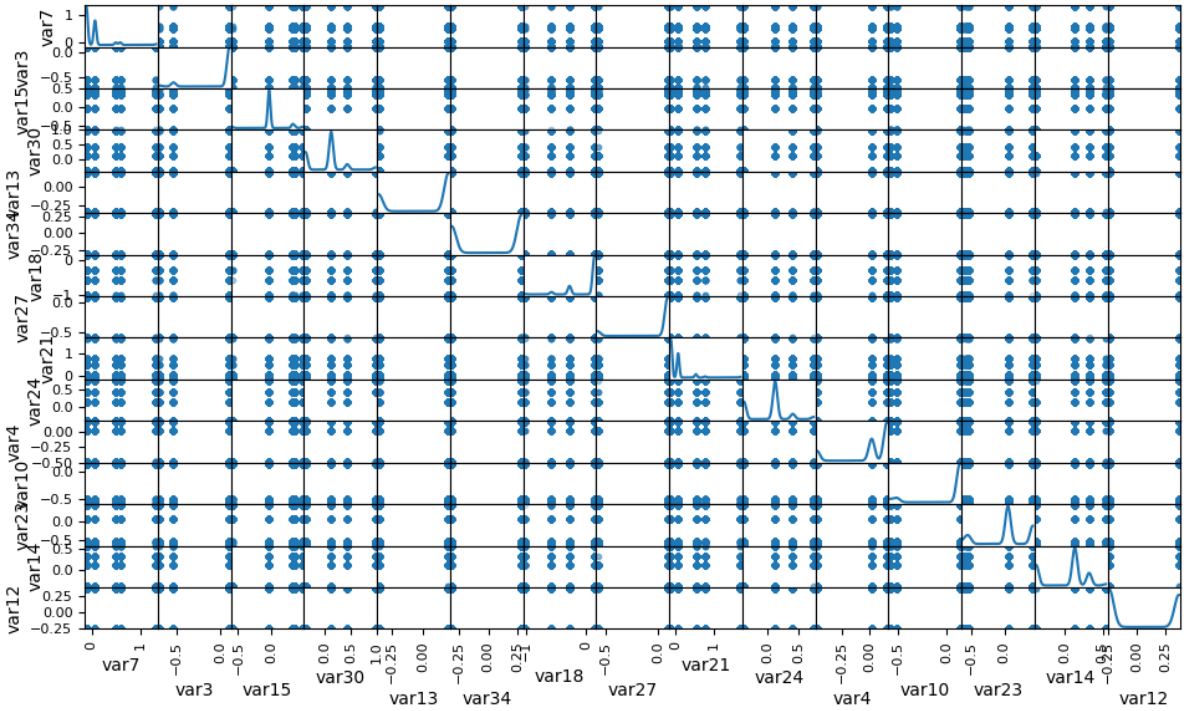
当相关性高时，只能保留一个：

- 可以选择IV高的
- 可以选择分箱均衡的

多变量分析

- WOE相关性矩阵

(从IV>0.02的变量中随机挑选的15个)



多变量分析

- 多变量分析：变量的多重共线性

通常用VIF来衡量，要求VIF<10

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

其中R_i^2是$\{x_1, x_2, .., x_{i-1}, x_{i+1}, x_{i+2}, \ldots, x_N\}$对$x_i$的线性回归的$R^2$

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop