1

申请评分卡中的数据预处理和特征衍生

目录

特征信息度的计算和意义

信用风险中的单变量分析和多变量分析

特征信息度的计算和意义

☐ 变量挑选

在评分卡模型中，变量挑选是非常重要的工作

- ✓ 变量间的共线性、线性相关性
  - 信息冗余
  - 降低了显著性，甚至造成符号失真
- ✓ 加剧了后期验证、部署、监控的负担
- ✓ 业务上含义不充分

特征信息度的计算和意义

- 变量挑选的依据
  - 带约束：LASSO
  - 特征重要性：随机森林
  - 模型拟合优度和复杂度：基于AIC的逐步回归
  - 变量信息度：IV

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

特征信息度的计算和意义

信用风险中的单变量分析和多变量分析

单变量分析和多变量分析

☐ 单变量分析

- 目的

根据变量某些属性，从初选名单(long list)中筛选出合适的变量进入缩减名单(short list)。

- 需要分析的变量属性

✓ 变量的显著性(高IV)

✓ 变量的分布

✓ 变量的业务含义

单变量分析和多变量分析

单变量分析

以分箱后的WOE为值

一、用IV检验有效性

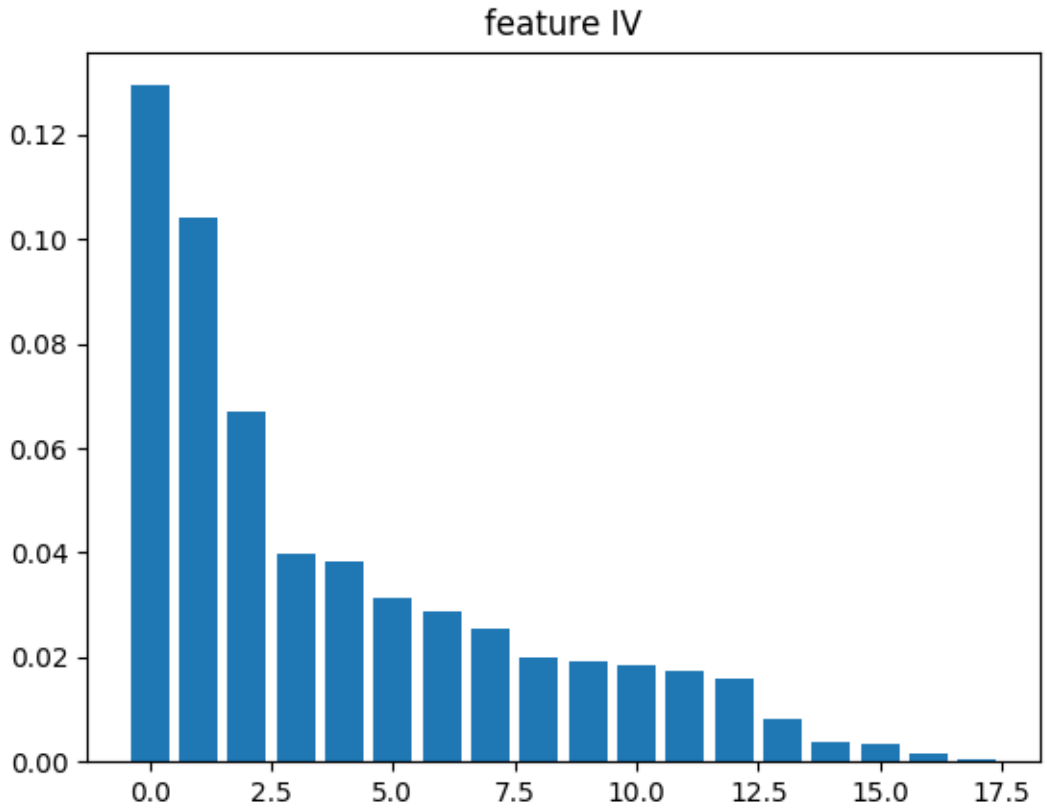
二、连续变量bad rate的单调性(可以放宽到U型)

三，单一区间的占比不宜过高

单变量分析和多变量分析

IV分布

22个变量的IV值(分箱后)



单变量分析和多变量分析

- 多变量分析：变量的两两相关性

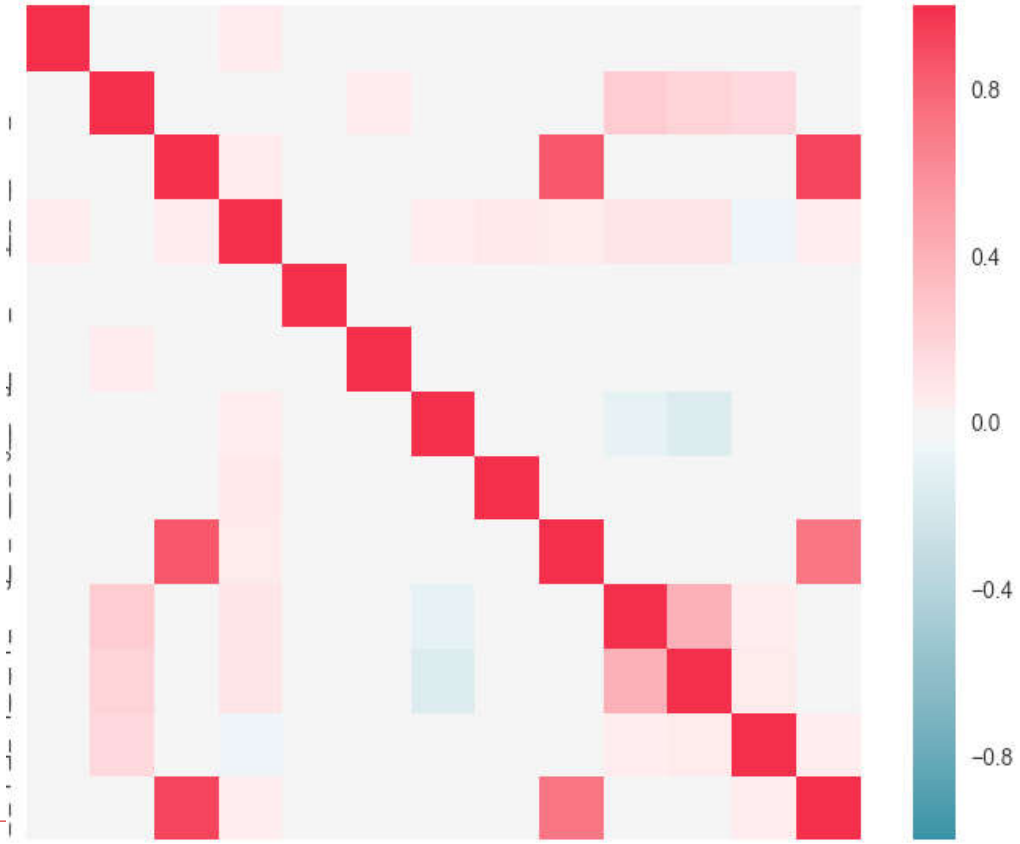
当相关性高时，只能保留一个：

- 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

(基于IV>0.01的变量)



单变量分析和多变量分析

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$R^2$

当某个变量的VIF超过10，需要逐一剔除解释变量。当剔除掉x_k时发现VIF低于10，从$\{x_k, x_i\}$中剔除IV较低的一个。

疑问

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

课程视频资源扫码：

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：小象学院
- 新浪微博：小象AI学院