

案例2：精准营销的两阶段预测模型

《Python数据科学：技术详解与商业实践》

讲师：Ben

自我介绍

- 天善商业智能和大数据社区 讲师 – Ben
- 天善社区 ID - Ben_Chang
- <https://www.hellobi.com> – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

- 总体思路
- 分类变量的压缩(压缩单变量水平数)
 - 重编码（概化）
 - WOE转换
- 连续变量的压缩(压缩变量个数)
 - 主成分分析
 - 变量聚类

总体思路

客户营销的业务理解

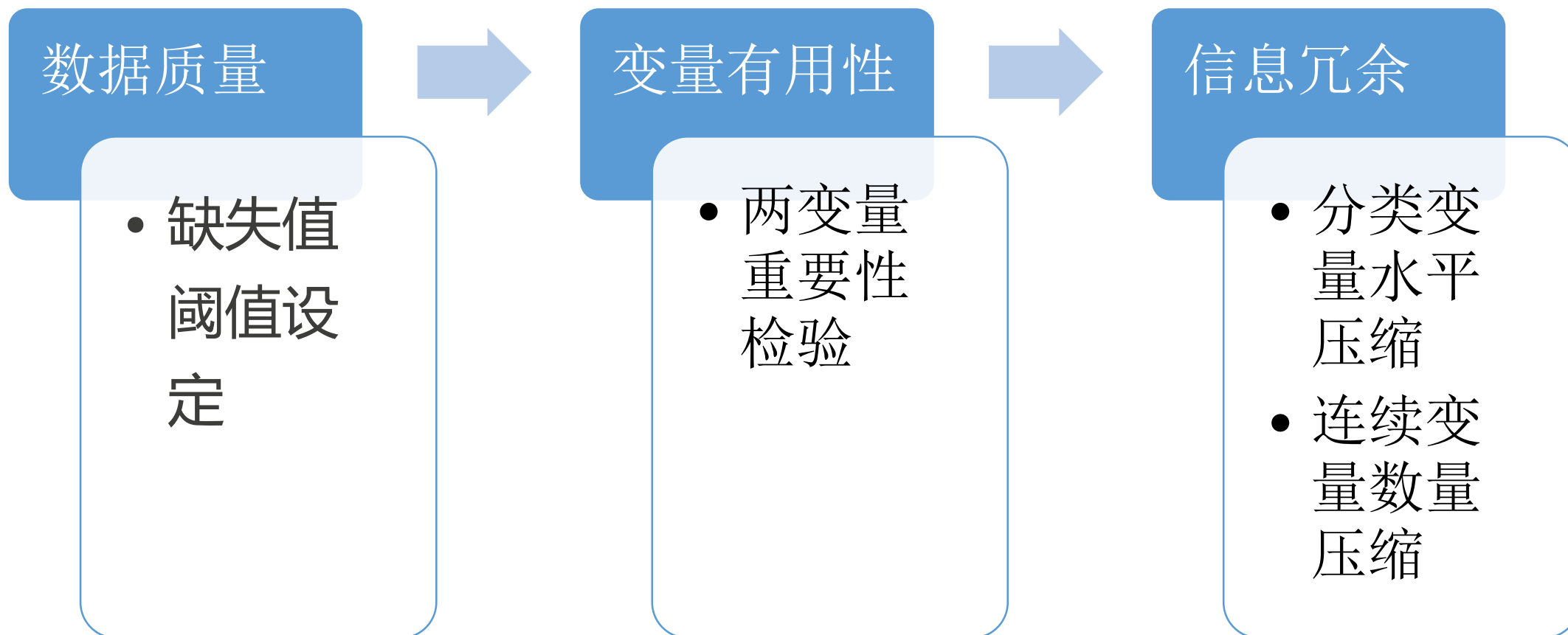


背景:有一个老兵社会组织主要通过发信件和邮寄小礼物的形式募集善款。为了减少成本,该组织决定仅向最有可能提供捐款的人发放信件和礼物。目前该组织有350万条历史的营销记录,并详细的记录了营销信息与响应结果。其中该组织最感兴趣的是最近12-24月有过捐款行为的人,并希望通过数据分析完成以下两个任务:

- 1) 什么人哪类人更有可能成为潜在的捐献人;
- 2) 这类人中各人的捐献数额可能是多少哪类人捐献的额度更多。



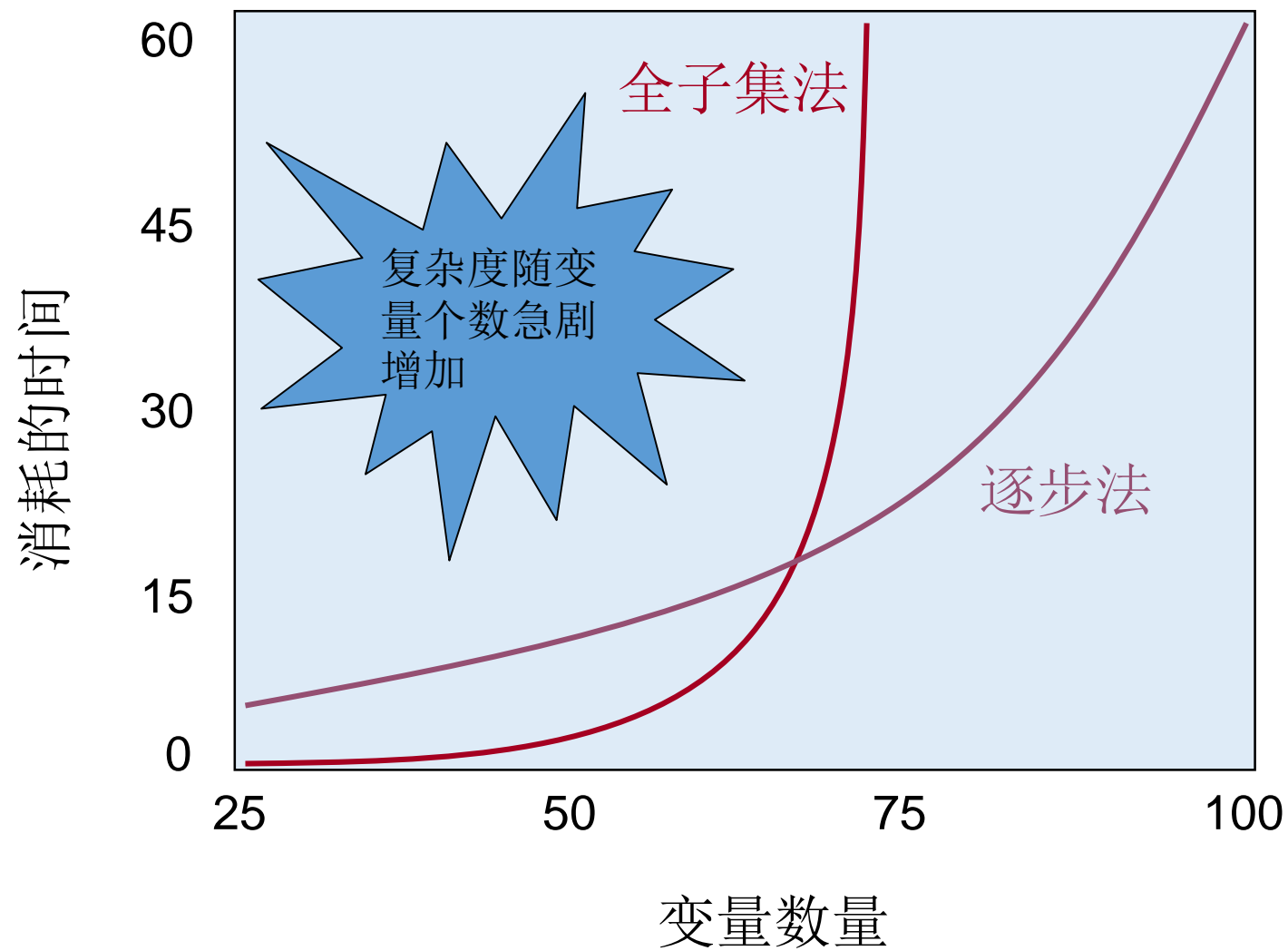
数据准备步骤



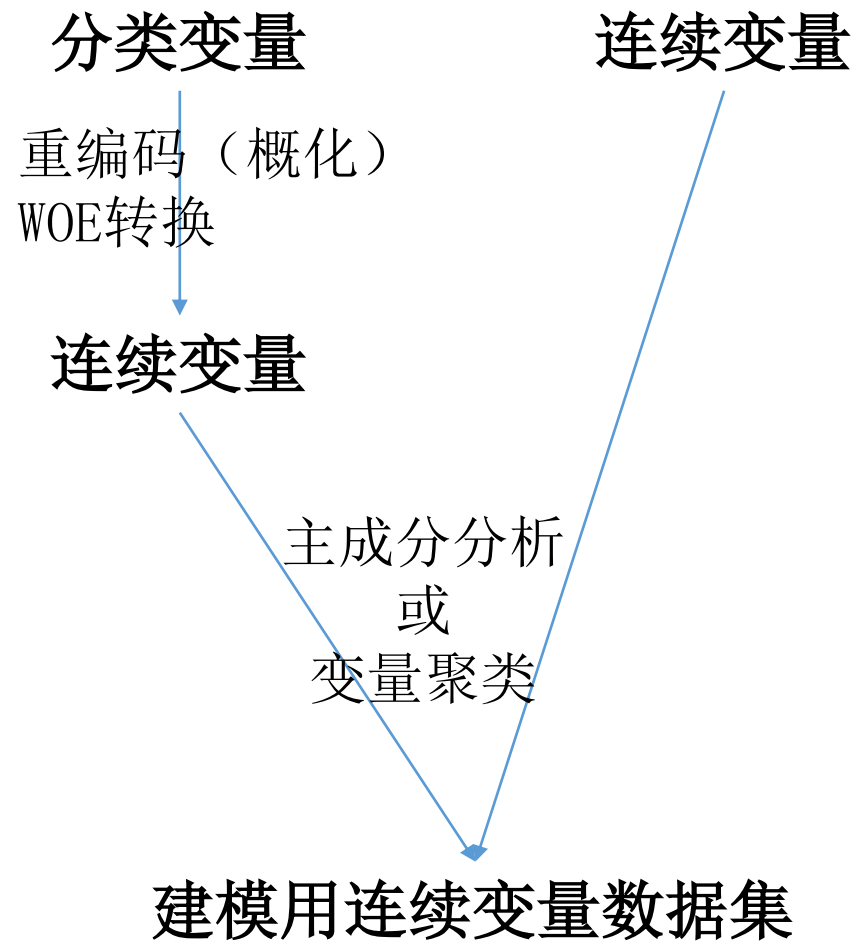
发现数据问题类型

- 脏数据或数据不正确
 - 比如 '0' 代表真实的0，还是代表缺失；Age = -2003
- 数据不一致
 - 比如收入单位是万元，利润单位是元，或者一个单位是美元，一个是人民币
- 数据重复
 - 这个问题在前面已经解决
- 缺失值
- 离群值

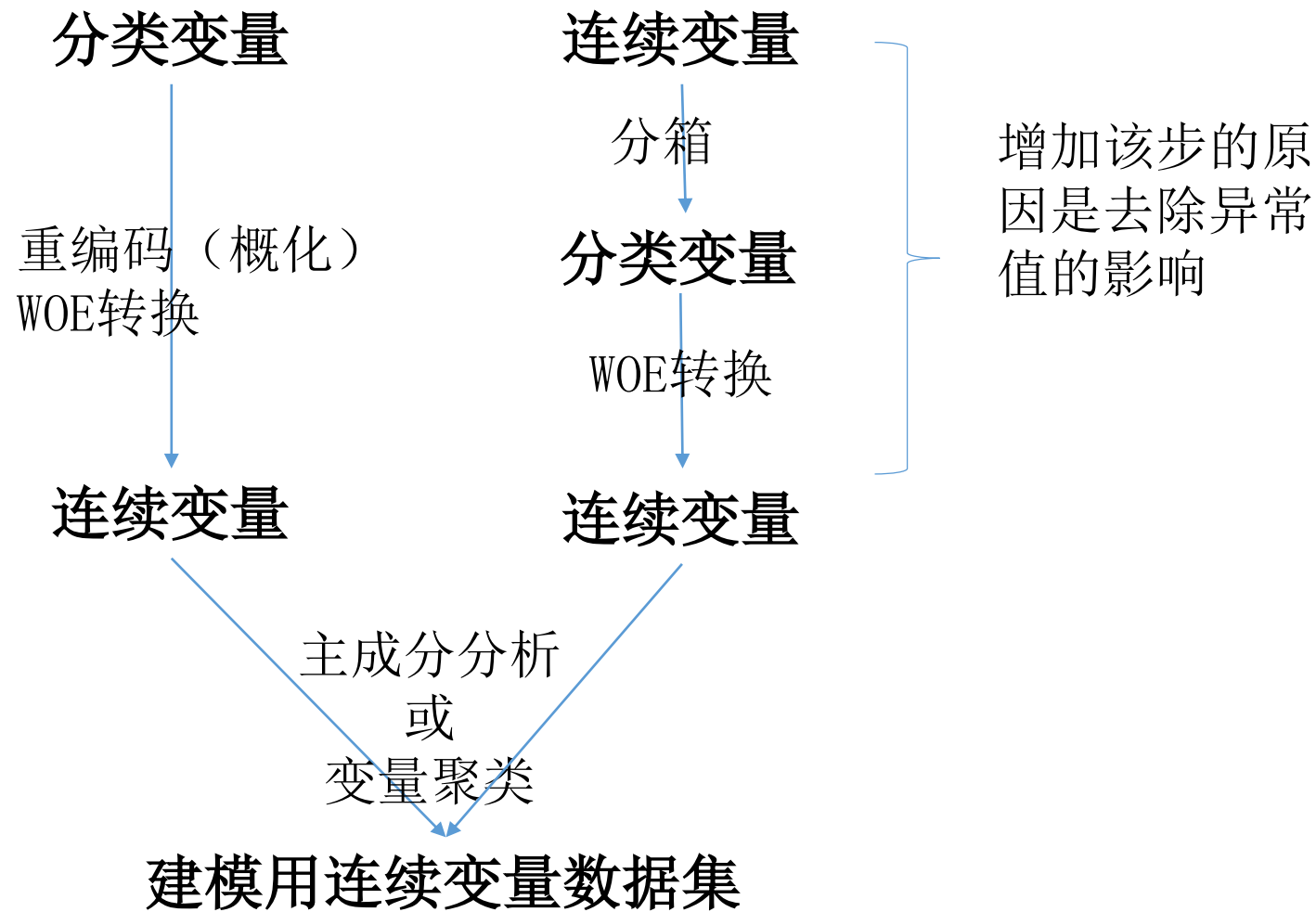
不要将变量筛选全放到建模的时候



解决方案（简单流程）



解决方案（建模标准流程）



- 分类变量的压缩(压缩单变量水平数)
 - 重编码（概化）
 - WOE转换
- 连续变量的压缩(压缩变量个数)
 - 主成分分析
 - 变量聚类

分类变量的压缩

001000

分类变量重编码（概化）

$x = w(\sum y_i)$

基于目标变量的转换-WOE

分类变量重编码（概化）



分类变量的哑变量编码法

-

虚拟变量

等级	值	标签	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
教育等级	1	1st	1	0	0	0
	2	2nd	0	1	0	0
	3	3rd	0	0	1	0
	4	4rd	0	0	0	1

最后一个水平的哑变量不放入模型中，默认作为对照组。

当水平数较多时

vehicle_make	freq
FORD	1112
CHEVY	654
DODGE	533
TOYOTA	417
	299
CHEVROLET	265
PONTIAC	226
HONDA	209
JEEP	199

<i>Level</i>	D_{B1}	D_{B2}	D_{B3}	D_{B4}	D_{B5}	D_{B6}	D_{B7}	D_{B8}	...
FORD	1	0	0	0	0	0	0	0	0
CHEVY	0	1	0	0	0	0	0	0	0
DODGE	0	0	1	0	0	0	0	0	0
TOYOTA	0	0	0	1	0	0	0	0	0
NA	0	0	0	0	1	0	0	0	0
CHEVR.	0	0	0	0	0	1	0	0	0
PONTIAC	0	0	0	0	0	0	1	0	0
HONDA	0	0	0	0	0	0	0	1	0
...	0	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0

汽车制造商（vehicle_make）这个变量有155个水平，要生成154个哑变量。

Quasi-Complete问题（似不完整数据问题）

汽车制造商

vehicle_make	freq
3HYUNDAI	1
B50	1
BUIUCK	1
CADALLIC	1
CADDY	1
CADI	1
CALERA	1
CEHV	1

汽车制造商

是否违约

	0	1
	237	62
3HYUNDAI	0	1
ACCURA	2	0
ACURA	10	1
AUDI	16	1
B50	1	0
BMW	13	2
BUICK	85	17
BUIUCK	1	0
CAD	1	1
CADILLAC	1	0

由于某个水平中，Y缺乏变异而导致无法计算。

由于这份数据中的许多汽车制造商只有一条记录，因此无法直接将这个变量直接纳入分类模型中。

合并不同水平

<i>Level</i>	<i>N_i</i>
FORD	1112
CHEVY	654
DODGE	533
TOYOTA	417
...	...
WV	1
ZX2	1

示例

<i>Level</i>	<i>p_i</i>
CHRY	0.4
MITSUBISHI	0.36
MERC	0.318
DAEWOO	0.30
SUZUKI	0.29
ISUZU	0.29
MAZDA	0.28
PONTIAC	0.25
SUBARU	0.24
...	...

方法1：将频次少的水平简单合为一类，看上去简单，但是精度降低不大，问题是水平数依然不少。

方法2：根据每个水平 $Y=1$ 的占比，将值接近的划分为一类。本步骤手工处理比较麻烦，将来学了决策树之后，可以使用该技术进行处理。

合并不同水平的问题

<i>Level</i>	N_i
FORD	1112
CHEVY	654
DODGE	533
TOYOTA	417
...	...
WV	1
ZX2	1



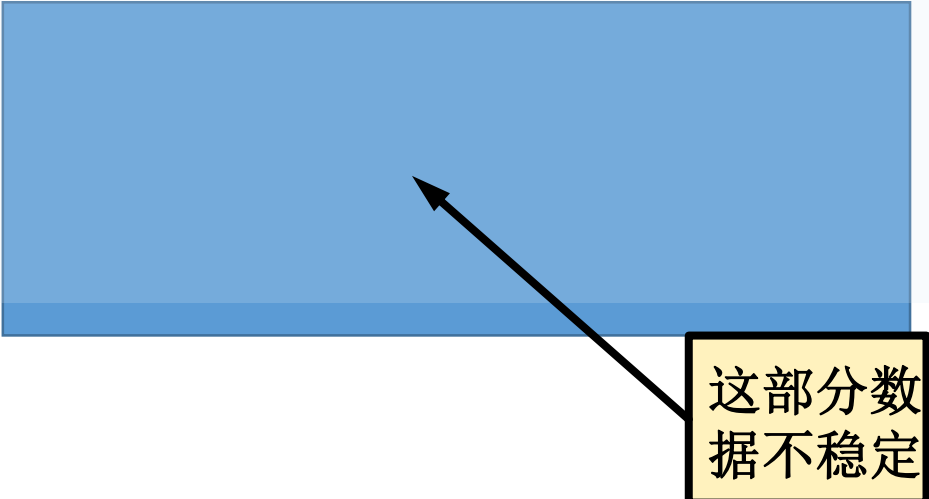
<i>Level</i>	p_i
CHRY	0.4
MITSUBISHI	0.36
MERC	0.318
<u>DAEWOO</u>	<u>0.30</u>
SUZUKI	0.29
ISUZU	0.29
MAZDA	0.28
PONTIAC	0.25
<u>SUBARU</u>	<u>0.24</u>
...	...

压缩之后的分类变量还是会生成若干个哑变量。从道理上讲，在后续建模中，由一个分类变量生成的哑变量要么同时在模型中，要么都不在模型中。但是在模型选择变量时不能满足这个要求，因此较好的方法是将分类变量转换为连续变量。

基于目标变量的转换-WOE

基于目标变量的转换-WOE

不要使用原始变量进行WOE转换

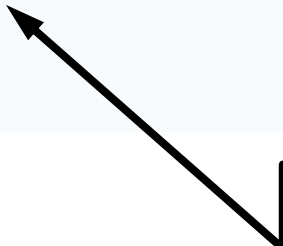
<i>Level</i>	<i>N_i</i>	<i>N(Y_i=1)</i>	<i>N(Y_i=0)</i>	<i>P(Y_i=1)</i>	<i>P(Y_i=0)</i>	<i>log(p_i/1-p_i)</i>
FORD	1112	253	859	0.211	0.184	0.134
CHEVY	654	128	526	0.107	0.113	-0.056
DODGE	533	112	411	0.102	0.088	0.142
TOYOTA	417	78	339	0.065	0.073	-0.113
...
VE	1	1	0			
WV	1	0	1			
TT	1	1	0			
ZX2	1	1	0			
SUM	5845	1197	4648			

基于目标变量的转换-WOE

要使用重分组后变量进行WOE转换

<i>Level</i>	<i>N_i</i>	<i>$N(Y_i=1)$</i>	<i>$N(Y_i=0)$</i>	<i>$P(Y_i=1)$</i>	<i>$P(Y_i=0)$</i>	<i>$\log(P(Y_i=1)/P(Y_i=0))$</i>
FORD	1112	253	859	0.211	0.184	0.134
CHEVY	654	128	526	0.107	0.113	-0.056
DODGE	533	112	411	0.102	0.088	0.142
TOYOTA	417	78	339	0.065	0.073	-0.113
...
VE	4	3	1	0.0025	0.0002	2.5
WV						
TT						
ZX2						
SUM	5845	1197	4648			

样本少的
归为一类

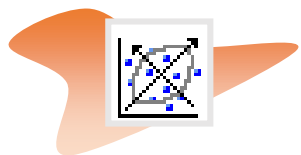


说明：

基于目标变量的转换是一种思路，实际工作中的实现方法很多，目前本节讲的是思路最简单，但是操作很麻烦的做法，实际工作中并不经常使用。

连续变量的压缩

连续压缩变量的思路方法



在建模之前使用主成分、变量聚类



在建模时使用逐步法或全子集法

主成分分析的思路

- 主成分分析，是考察多个变量间相关性一种多元统计方法，研究如何通过少数几个主成分来揭示多个变量间的内部结构，即从原始变量中导出少数几个主成分，使它们尽可能多地保留原始变量的信息，且彼此间互不相关。

1-标准化变换

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, p$$
$$\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}, s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}$$

2-相关系数矩阵

$$R = [r_{ij}]_{p \times p} = \frac{Z^T Z}{n-1}$$
$$r_{ij} = \frac{\sum z_{kj} \cdot z_{ki}}{n-1}, i, j = 1, 2, \dots, p$$

3-求解特征值

$$|R - \lambda I_p| = 0$$

$$\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j} \geq 0.85$$

4-主成分表达

$$U_{ij} = z_i^T b_j^o, j = 1, 2, \dots, m$$

- U1称为第一主成分
- U2 称为第二主成分
- , ...,
- Up 称为第p 主成分

5-主成分评价

- 对m 个主成分进行加权求和，即得最终评价价值，权数为每个主成分的方差贡献率

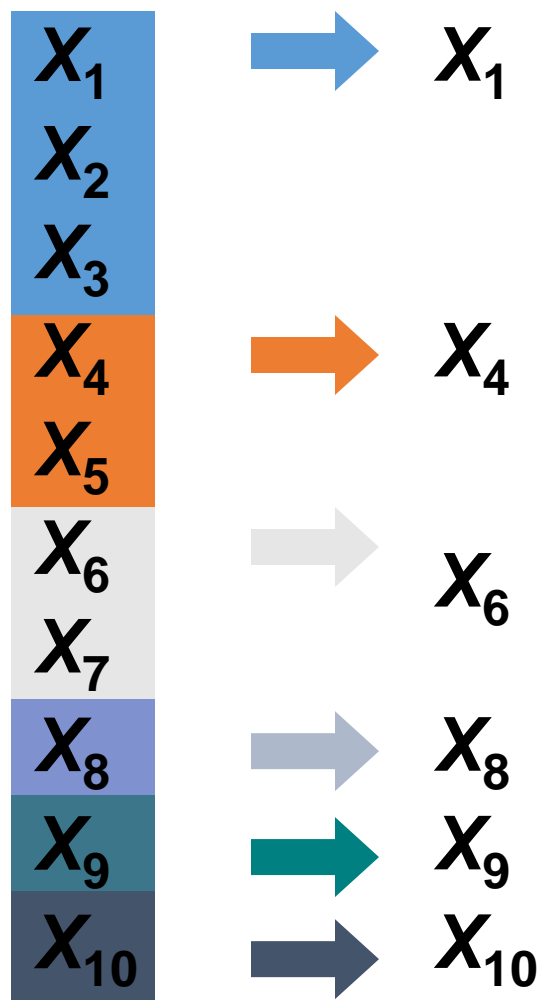
优点:

- ①可消除评估指标之间的相关影响
- ②可减少指标选择的工作量

缺点:

- ①对提取的主成分必须给出符合实际背景和意义的解释
- ②主成分的解释其含义一般多少带有点模糊性

变量聚类思路



变量聚类之后，选择输入变量：

- 对聚类的代表性
- 专家指定
- 与被解释变量的相关性

秦路主讲

七周成为数据分析师

七周为期，Get一条数据分析师职业黄金通道！



Python

数据分析与挖掘

集Python爬虫、数据采集、数据处理、数据分析与数据挖掘于一体，打造Python全栈工程师

主讲老师：韦玮

VIP会员群+在线答疑+录播复习+1年反复观看

案例为师，实战为王

开启Python机器学习之路

科学规划全套课程体系，从入门到进阶，从理论到技巧，嵌入丰富课程案例讲解，逐步推进

讲师：唐宇迪 深度学习领域多年一线实践研究专家

独一无二的数据库建模指南系列教程升级版

- 从企业视角进行数据规划以及数据库模型的搭建
- 高质量的数据库模型和技巧，以及丰富的例子
- 数据库架构理论和实践要领

资深讲师：BAO胖子 15年+BI从业经验
涉足电力、快消品、医药、信息服务行业的BI老兵

业务知识一站通

技术+业务，挣钱有门路！

讲师：陈文



自己动手 丰衣足食

Python3网络爬虫实战案例

一循序渐进，案例为王，诠释全面，思路制胜一

讲师：崔庆才 北航硕士，百万级热度爬文博主



讲师 丘祐玮

人人都爱数据科学家

Python数据科学精华实战课程

数据分析报告制作

秘籍升级版

讲师：陈丹奕 知乎大神，前百度资深数据分析师

先机致胜 破冰AI

深度学习模型/框架与实战

讲师：唐宇迪 同济大学硕士
深度学习领域多年一线实践研究专家



BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI