数据挖掘

AI学院
数据分析全栈工程师

数据挖掘应用前景

2015年1月，Linkedin对全球超过3.3亿用户的工作经历和技能进行分析，公布2014年最受雇主喜欢、最炙手可热的25项技能，统计分析和数据挖掘位列榜首。

热

据艾瑞的研究报告，未来与数据分析相关的就业岗位会在1000万左右，而目前来说国内的合格的数据分析师不足5万左右

缺

企业也希望能在找到一个合格的数据分析，希望在互联网与大数据时代，把握整个企业在市场上的走向

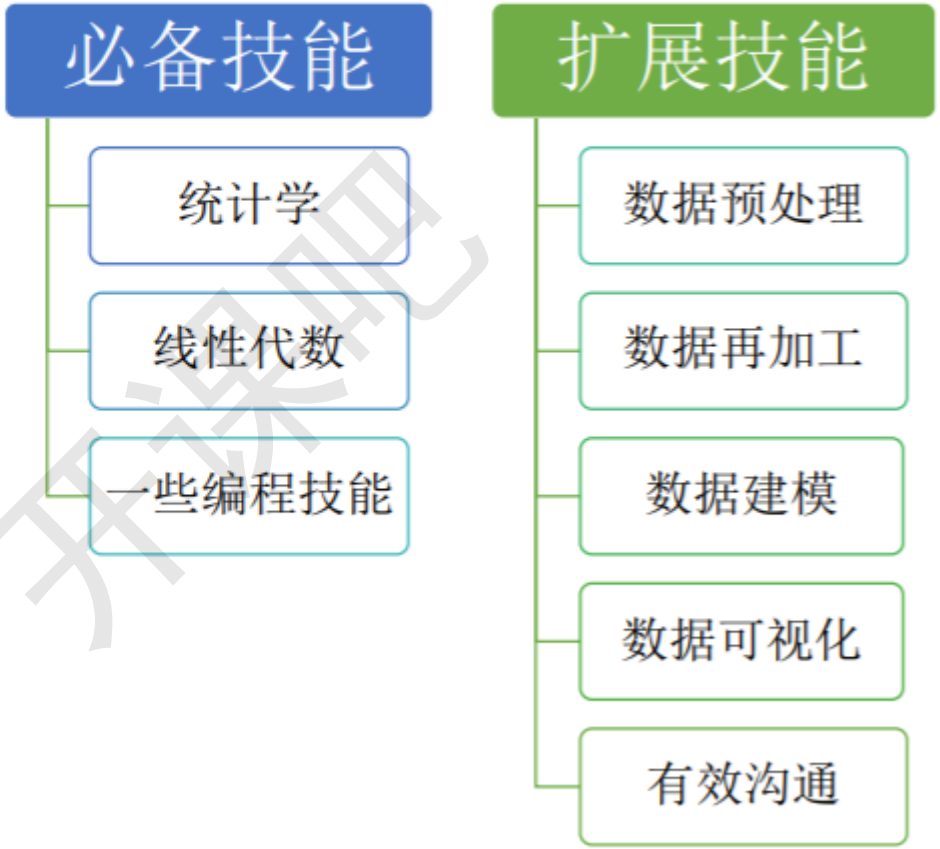
专

数据挖掘从业人员收入

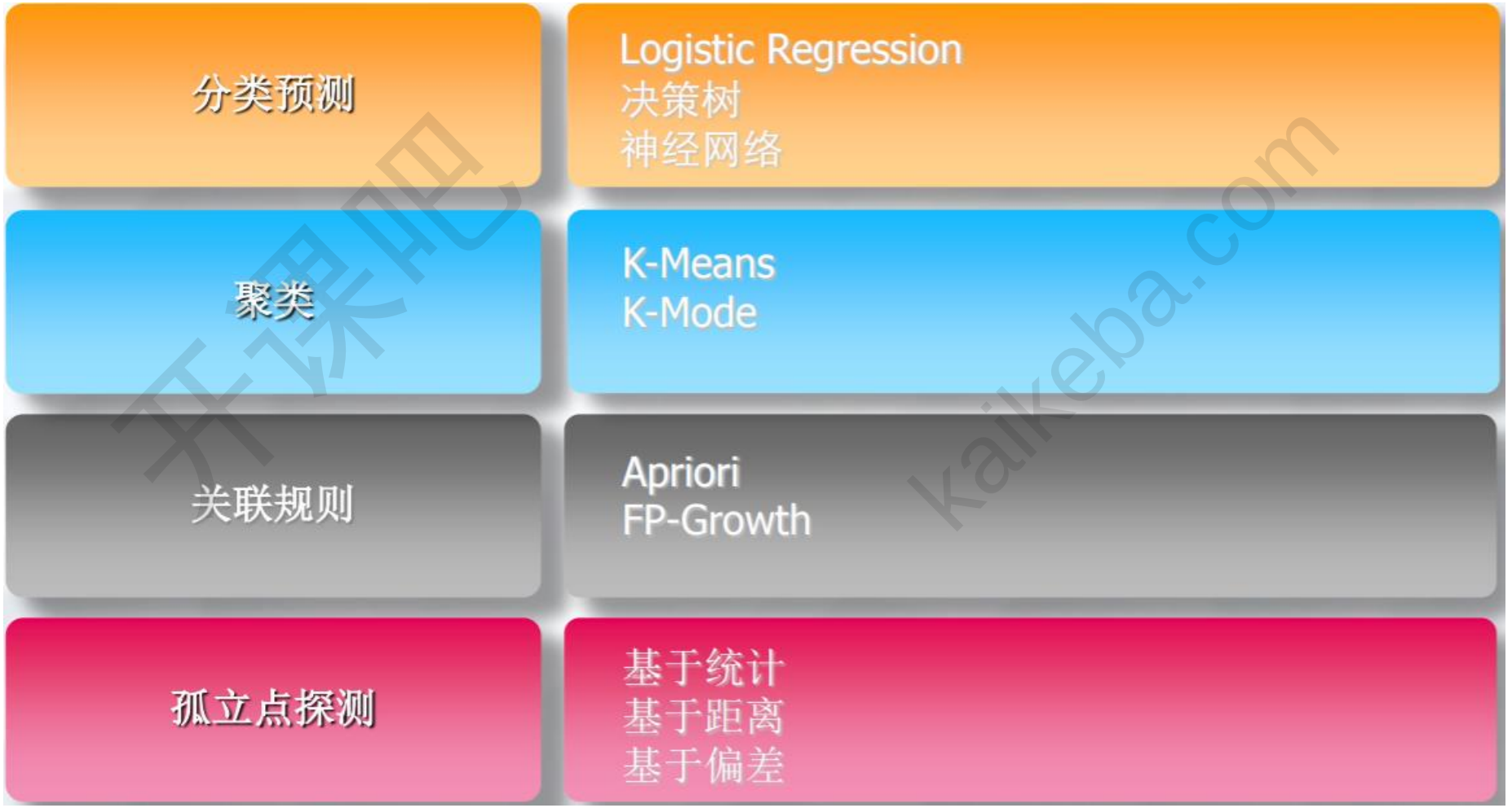
地区竞争力分析

1	北京 (3743份样本)	￥23250
2	深圳 (1361)	￥20910
3	上海 (1895)	￥20280
4	杭州 (975)	￥19190
5	苏州 (112)	￥15810
6	广州 (782)	￥15240
7	南京 (461)	￥13990
8	武汉 (223)	￥13980
9	成都 (455)	￥13010
10	西安 (142)	￥10170

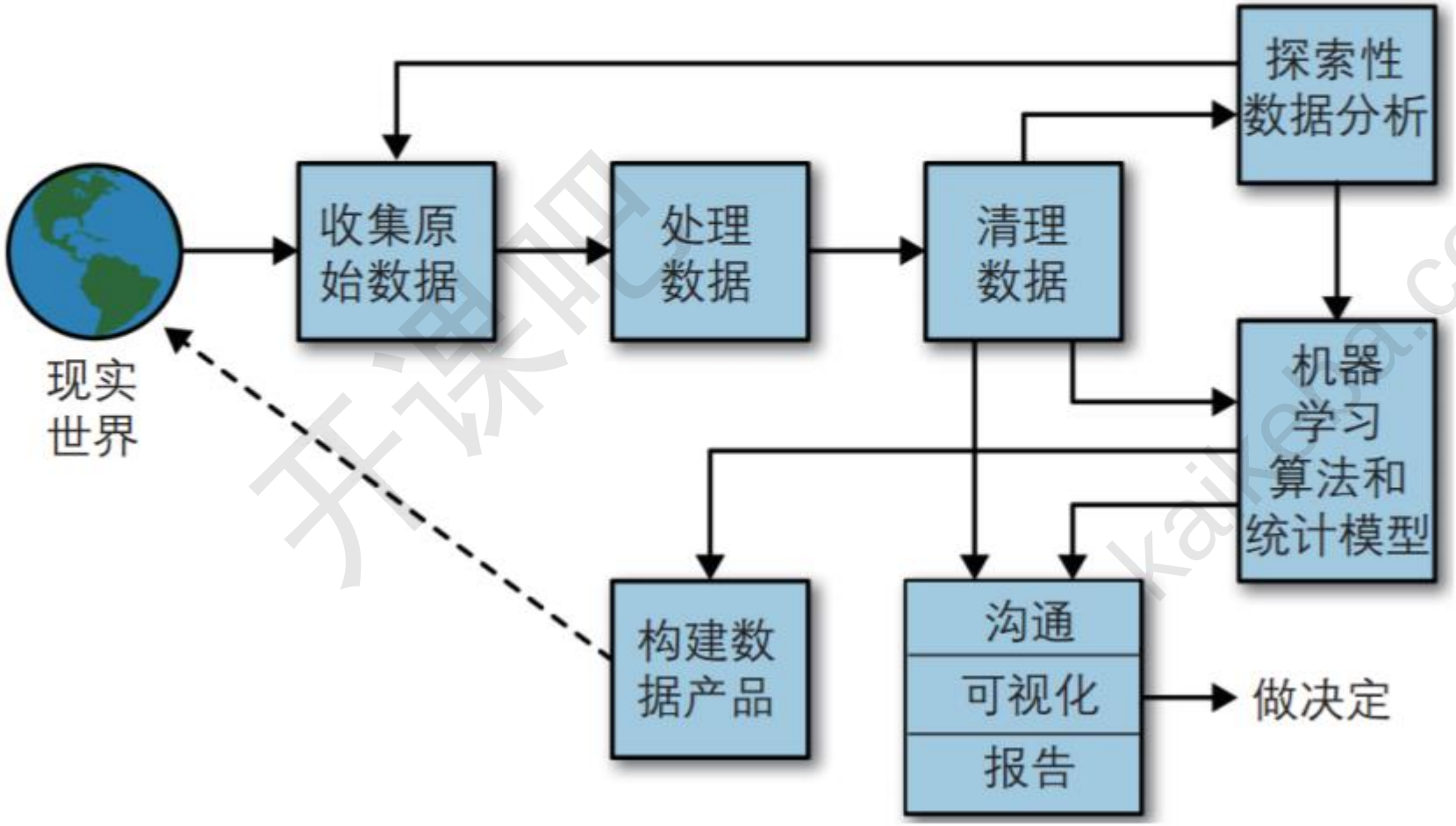
如何成为一名的数据从业者



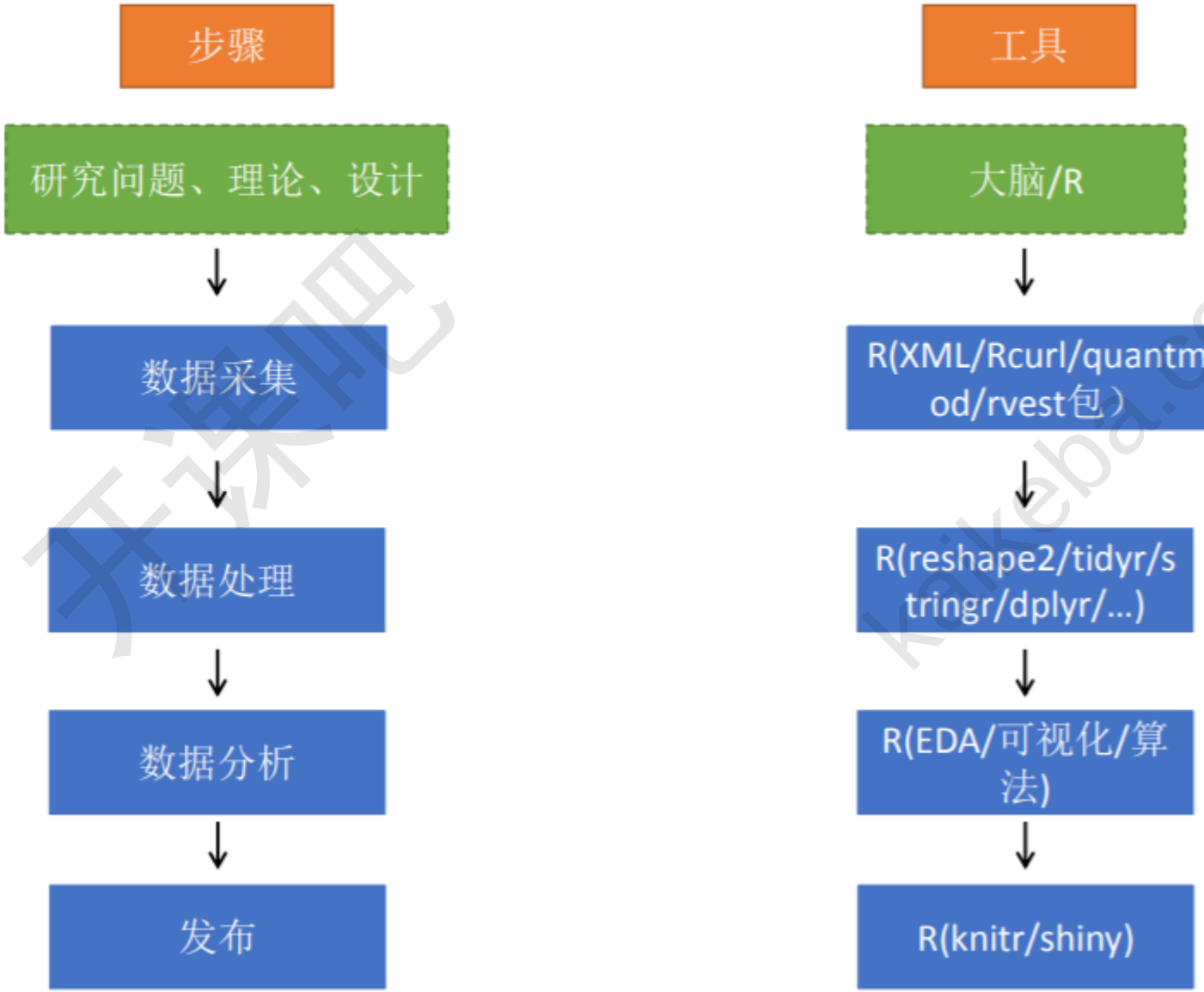
数据挖掘=模型+算法



数据科学的工作流程


探索性
数据分析
收集原
始数据
处理
数据
清理
数据
现实
世界
机器
学习
算法和
统计模型
构建数
据产品
沟通
可视化
报告
做决定

使用R进行数据挖掘



R快速入门

软件安装

- **Windows下安装R、Rstudio**

方法：从http://www.r-project.org/网站上下载R 安装文件

从http://www.rstudio.com/网站上下载RStudio安装文件

- **linux下安装R、Rsutdio**

方法：执行sudo apt-get install r-base-dev 安装R

执行wget https://download1.rstudio.org/rstudio-1.0.136-amd64.deb

sudo gdebi rstudio -1.0.136-amd64.deb安装rstudio

基本操作

- 查找帮助
- 工作空间
- 包的使用
- 数据读入

安装R包

- install.packages()
- devtools::install_github()
- RCMD INSTALL "xxx.tar.gz"
- 本地安装(通过窗口操作)

数据对象

- 向量(vector)
- 列表(list)
- 矩阵(matrix)
- 数据框(data.frame)

数据的创建

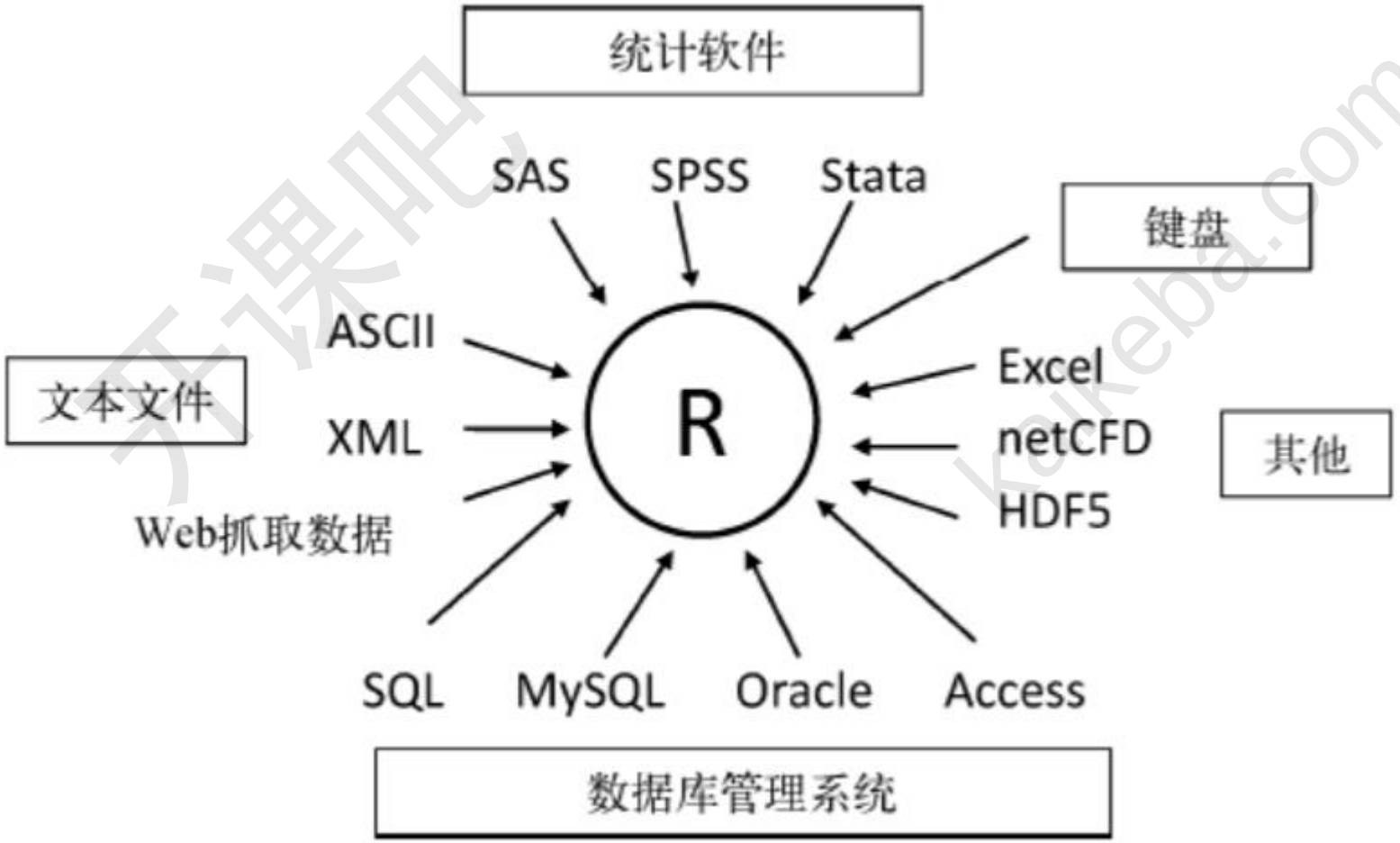
- 通俗地说，对象类型是指R语言组织和管理内部元素的不同方式。数据类型则描述了一个变量内元素取值的类型。例如，逻辑类型数据的取值是TRUE和FALSE，而数值类型的取值是实数。不同对象类型元素取值的数据类型如下表所示：

对象类型	数据类型	是否允许出现不同数据类型
向量	数值型、复数型、字符型、逻辑型	不允许
因子	数值型、复数型、字符型、逻辑型	不允许
数组	数值型、复数型、字符型、逻辑型	不允许
矩阵	数值型、复数型、字符型、逻辑型	不允许
数据框	数值型、复数型、字符型、逻辑型	相同列内元素，其数据类型必须相同； 不同列之间的数据类型可以不同
列表	数值型、复数型、字符型、逻辑型	任何元素的数据类型均可不同
时间序列	数值型、复数型、字符型、逻辑型	不允许

- 对于未知类型的对象，在R中有3个函数可以查看对象的类型：class()、mode()、typeof()。

可供R导入的数据源

- R可以从键盘、文本文件、Microsoft Excel和Access、流行的统计软件、特殊格式的文件，以及多种关系型数据库中导入的数据。

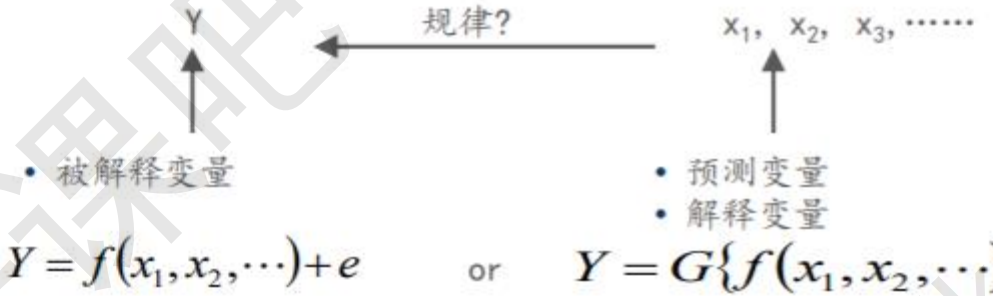


逻辑回归

AI学院
数据分析全栈工程师

预测性模型是什么

- 预测性建模基于用户的**历史信息**去预测其**将来**的行为
- 预测性模型是帮助提高营销活动的一个工具，能针对用户实现精准营销



类型	被解释变量	函数形式	
线性回归	连续变量 例如 消费，收入	$Y_i = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + e_i = X\beta + e$	e是误差项
逻辑回归	事件是否发生（0或1） 例如 是否响应促销	$\ln(Y_i / 1 - Y_i) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + e_i = X\beta + e$	e是误差项

逻辑回归的数理原理

- **应用场景**
  - 逻辑回归被广泛应用在目标变量是二值变量的场合（0，1）
- **公式**
  - P(y=1|x) 表示 y = 1的概率
  - 从而得到 y = 1 对 y = 0 概率的比值

$$odds = \frac{P}{1-P}$$

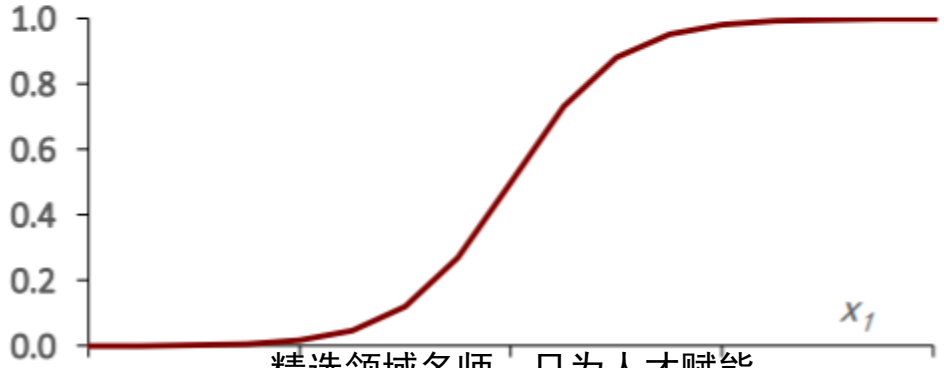
  - 定义逻辑变换：

$$\text{logit}(P) = \ln(odds) = \ln(\frac{P}{1-P})$$

$$\text{logit}(P) = \beta_0 + \beta^T X$$

- **模型估计**
  - 极大似然估计
- **模型阐释/评估**
  - 一个解释变量的阐释图（如右）
  - C值，Lift图

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1)}}$$



逻辑回归模型能够解决哪些商业问题（一）

一、扩大市场占有率

挑战：

- 人力成本不断攀升
- 客户流失不断加剧

方案：克隆模型
获客模型

收益：1. 找出潜在客户

2. 扩大市场份额

逻辑回归模型能够解决哪些商业问题（二）

二、活动响应预测

挑战：

- 市场活动预算有限
- 提高活动针对性

方案：活动响应模型
产品响应模型

收益：1. 在预算范围内最大化收益

2. 对响应概率低的客户制定响应的市场活动

逻辑回归模型能够解决哪些商业问题（三）

三、流失预警/客户赢回

挑战：

- 客户自然流失率很高
- 来自竞争对手的威胁

方案：客户流失模型
客户赢回模型

收益：1. 确认流失概率高的客户，制定流失防范措施

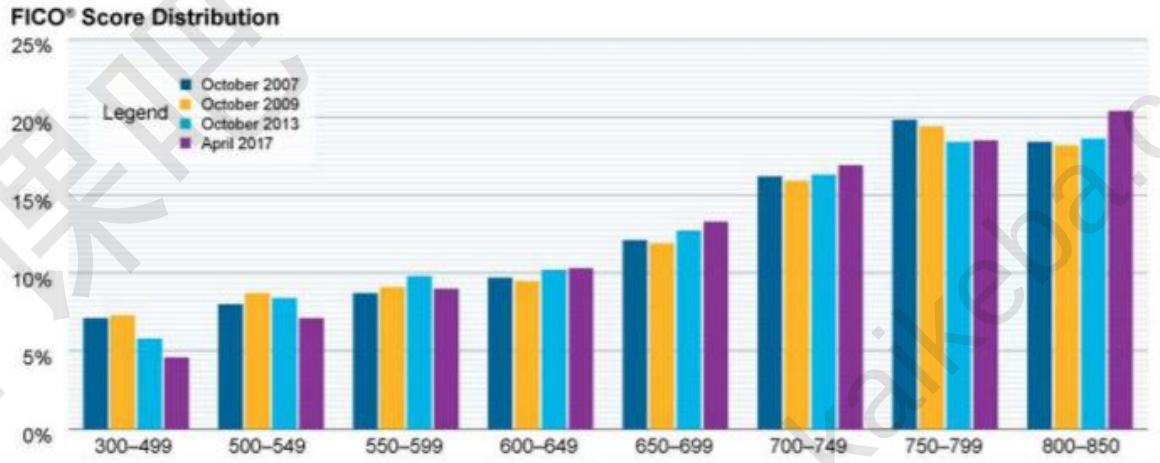
2. 对赢回概率高的客户指定优惠的市场活动

信用评分

AI学院
数据分析全栈工程师

美国信用体系

美国信用体系中,个人信用评分系统主要有FICO公司完成.评分值越高,违约率越低,依此可以通过信用评分进行信用卡的发放或者信用额度等决策,美国信用总体状况如下:



	PERCENT OF POPULATION												
FICO® Score 8	October 2005	October 2006	October 2007	October 2008	April 2009	April 2010	April 2011	April 2012	April 2013	April 2014	April 2015	April 2016	April 2017
300–499	6.6	6.5	7.1	7.2	7.3	6.9	6.3	5.7	5.6	5.4	4.9	4.6	4.7
500–549	8.0	8.0	8.0	8.2	8.7	9.0	8.7	8.5	8.4	8.1	7.6	7.1	6.8
550–599	9.0	8.8	8.7	8.7	9.1	9.6	9.9	10.0	9.9	9.6	9.4	9.0	8.5
600–649	10.2	10.2	9.7	9.6	9.5	9.5	9.8	10.1	10.1	10.2	10.3	10.3	10
650–699	12.8	12.5	12.1	12.0	12.0	11.9	12.1	12.2	12.2	12.8	13.0	13.3	13.2
700–749	16.4	16.3	16.2	16.0	15.9	15.7	15.5	16.0	16.3	16.4	16.6	16.9	17.1
750–799	20.1	19.8	19.8	19.6	19.3	19.5	19.6	19.0	18.9	18.2	18.2	18.5	19
800–850	16.9	17.9	18.4	18.7	18.2	17.9	18.1	18.5	18.5	19.3	19.9	20.4	20.7
TOTAL*	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

** All columns may not add up to 100.0% due to rounding.*

Source: FICO Blog

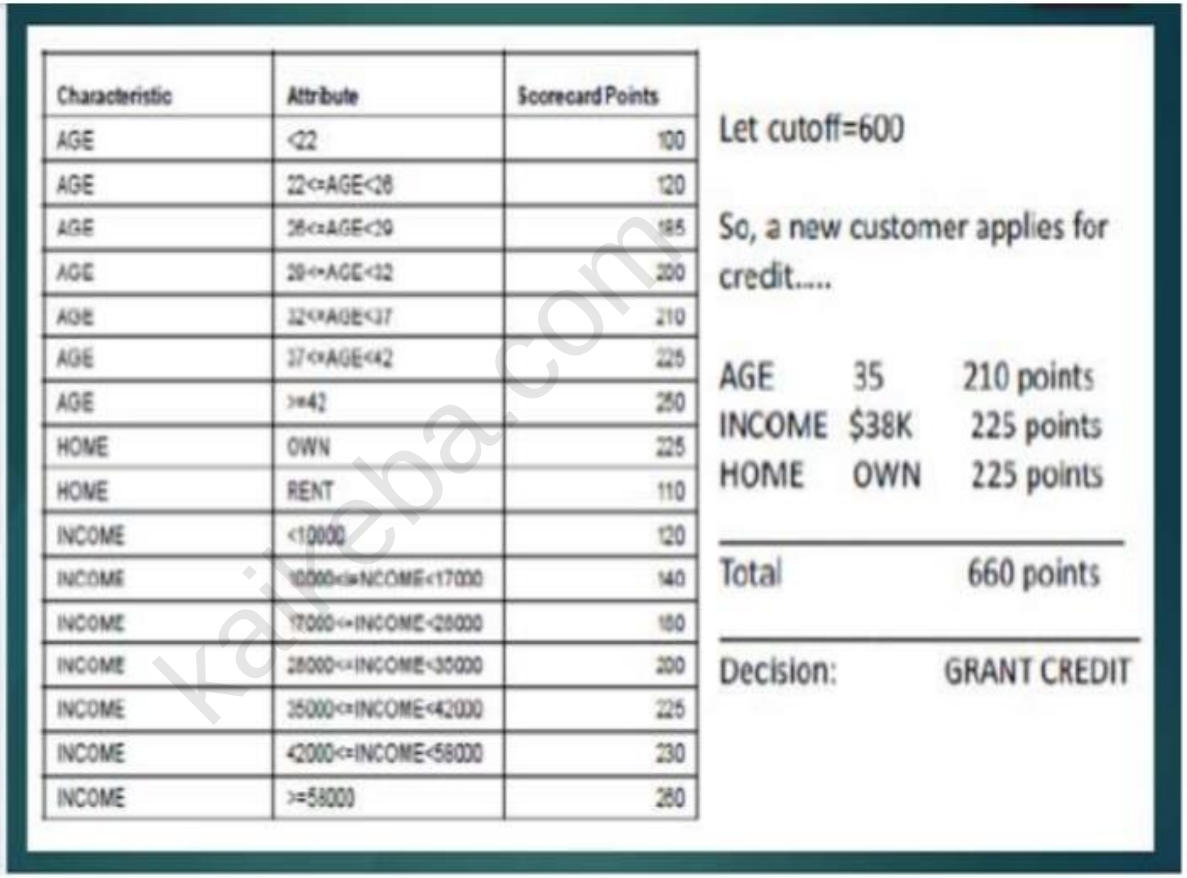
FICO信用评分考虑因素

分类:偿还历史35%,信用账户数30%,信用年限15%,新开账户10%,信用类型10%

评分的分类依据是基于一般个体中各个分类的重要性,对于特定的群体(例如刚开始使用信用卡的人群),每个分类的重要性可能会不同

比较类别	芝麻信用分	FICO 分	评论
考量维度	5 个维度：信用历史、行为偏好、履约能力、身份特质、人脉关系	5 个维度：支付记录、欠款金额、信用历史长度、信用类型、新卡申请数量	芝麻分没有在其官网公布 5 个维度的详细权重；FICO 分注重消费者的信贷经历。5 个维度中，支付记录和欠款金额的计算权重较高，分别为 35%和 30%
分值范围	350 分到 950 分	300 分到 850 分	两者的分数取值范围接近。芝麻分高于 650 分为优秀或极好；FICO 分高于 660 分才能方便办理信用卡业务，FICO 分高于 740 分可以享受贷款利率优惠

信用评分卡



Characteristic	Attribute	Scorecard Points
AGE	<22	100
AGE	22<=AGE<26	120
AGE	26<=AGE<29	185
AGE	29<=AGE<32	200
AGE	32<=AGE<37	210
AGE	37<=AGE<42	225
AGE	>=42	250
HOME	OWN	225
HOME	RENT	110
INCOME	<10000	120
INCOME	10000<=INCOME<17000	140
INCOME	17000<=INCOME<28000	160
INCOME	28000<=INCOME<35000	200
INCOME	35000<=INCOME<42000	225
INCOME	42000<=INCOME<58000	230
INCOME	>=58000	260

Let cutoff=600

So, a new customer applies for credit.....

AGE	35	210 points
INCOME	$38K	225 points
HOME	OWN	225 points
Total		660 points
Decision:		GRANT CREDIT

- 作用:1.决策类:是否放贷,是否同意信用卡申请;2.数额类:放贷额度,信用卡额度
- 优点:1.便于业务人员傻瓜式操作;2.便于监管部门监管(防止性别种族的歧视);3.易监控和调整

建模流程和统计量

- 建立评分模型基本流程
  - 输入变量的分箱
  - 建模,一般使用logistic回归建立模型
  - 指定业务参数将logistic回归系数转化为评分
  - 模型检验
- WOE(Weight of Evidence):证据权重,与违约比例同方向变动,可以看到不同分箱的重要性
- IV(information Value):信息值,表示变量的重要性.
- $IV < 0.02$,对预测几乎无帮助;$0.02 \leq IV < 0.1$,具有一定帮助
- $0.1 \leq IV < 0.3$,对预测有较大帮助;$IV \geq 0.3$,具有很大帮助

WOE和IV

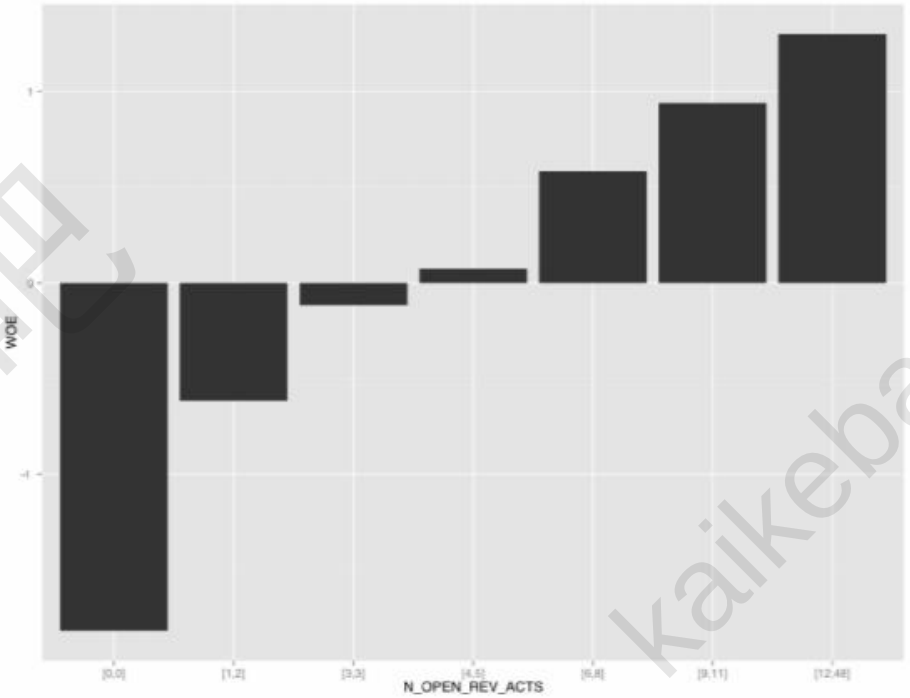
Categorized variable:Age

Age Group	Number of Goods	Number of Bads	Total N	Distribution of Goods	Distribution of Bads	Cumulative Information Value (IV)	WoE weight of
(-inf,21>	63	42	105	0.09000	0.14000	0.022	-44.18
(21,24>	82	52	134	0.11714	0.17333	0.022	-39.18
(24,31>	188	87	275	0.26857	0.29000	0.002	-7.68
(31,34>	90	23	113	0.12857	0.07667	0.027	51.70
(34,42>	128	46	174	0.18286	0.15333	0.005	17.61
(42,inf)	149	50	199	0.21286	0.16667	0.011	24.46
All Groups	700	300	1000	1.00000	1.00000	0.089	

$$WOE = ln\frac{yPctGood}{yPctBad} \times 100$$

$$IV = \sum_{i=1}^{n}(yPctGood - yPctBad) \times ln\frac{yPctGood}{yPctBad}$$

WOE分箱原则



- 分箱数适中,不宜过多过少
- 各个分箱内记录数合理
- 分箱应该体现出明显的趋势特性
- 相邻分箱的差异不宜过大

生成信用评分模型

- 评分需要控制在一定范围内(例如0-1000)
- 对于特定分数,好客户和坏客户有一定的比例关系,即优比(odds),$odds = \frac{xPctGood}{xPctBad}$,例如800分时比值是50:1
- 增加一定评分值时优比增加一倍,例如增加45分,odds增加一倍(从50:1到100:1)

$$Score = Offset + Factor \times ln(odds)$$

$$Score + pdf = Offset + Factor \times ln(2 \times odds)$$

pdo:points to double the odds

初始化: 例如Score=800对应odds=50,pdo=45.

即可算出对应的Offset和Factor

Q&A

AI学院
数据分析全栈工程师

备用页:

我们为什么不直接用这个WOE绝对值的加和来衡量一个变量整体预测能力的好坏，而是要用WOE处理后的IV呢。
我们这里给出两个原因。IV和WOE的差别在于IV在WOE基础上乘以的那个$(py_i - pn_i)$，
我们暂且用pyn来代表这个值。
第一个原因，当我们衡量一个变量的预测能力时，我们所使用的指标值不应该是负数，否则，
说一个变量的预测能力的指标是-2.3，听起来很别扭。从这个角度讲，乘以pyn这个系数，保证了变量每个分组的结果都是非负数，
你可以验证一下，当一个分组的WOE是正数时，pyn也是正数，当一个分组的WOE是负数时，pyn也是负数，而当一个分组的WOE=0时，pyn也是。

上面的原因不是最主要的，因为其实我们上面提到的 $WOE = \sum_i^n |WOE_i|$ 这个指标也可以完全避免负数的出现。
更主要的原因，也就是第二个原因是，乘以pyn后，体现出了变量当前分组中个体的数量占整体个体数量的比例，对变量预测能力的影响。

A	响应	未响应	合计	响应比例	WOE	IV
1	90	10	100	90%	4.3944492	0.0390618
0	9910	89990	99900	10%	-0.00893	7.937E-05
合计	10000	90000	100000	10%	4.4033788	0.0391411

从这个表我们可以看到，变量取1时，响应比达到90%，对应的WOE很高，但对应的IV却很低，原因就在于IV在WOE的前面乘以了一个系数$(py_i - pn_i)$，而这个系数很好的考虑了这个分组中样本占整体样本的比例，比例越低，这个分组对变量整体预测能力的贡献越低。相反，如果直接用WOE的绝对值加和，会得到一个很高的指标，这是不合理的。