天善智能
TIANSHAN SOFT

第5章-3: 数据整理

Python数据科学：全栈技术详解

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。

主要内容

- FRM提取行为变量
- 数据重组
  - 拆分列
  - 堆叠列
- 抽样

天善智能
TIANSHAN SOFT

FRM提取行为变量

分析客户价值的重要性

- *20%的客户占80%的销售额*
- *20%的客户提供给我们80%的利润*
- *这20%的客户就是我们的重要客户，是我们利润的来源*
- *新客户的开发成本是老客户的5倍*
- *潜力客户是目前采购产品线较窄，以后会增加的客户*
- *客户的采购金额增长率超出 公司的销售增长率*

分析客户价值的方法

根据美国数据库营销研究所Arthur Hughes的研究，客户数据库中有三个重要指标：

- 最近一次消费(Recency)

最近一次消费意指上一次购买的时间。上一次消费时间越近的顾客对提供即时的商品或是服务也最有可能会有反应。对提供即时的商品或是服务也最有可能会有反应。

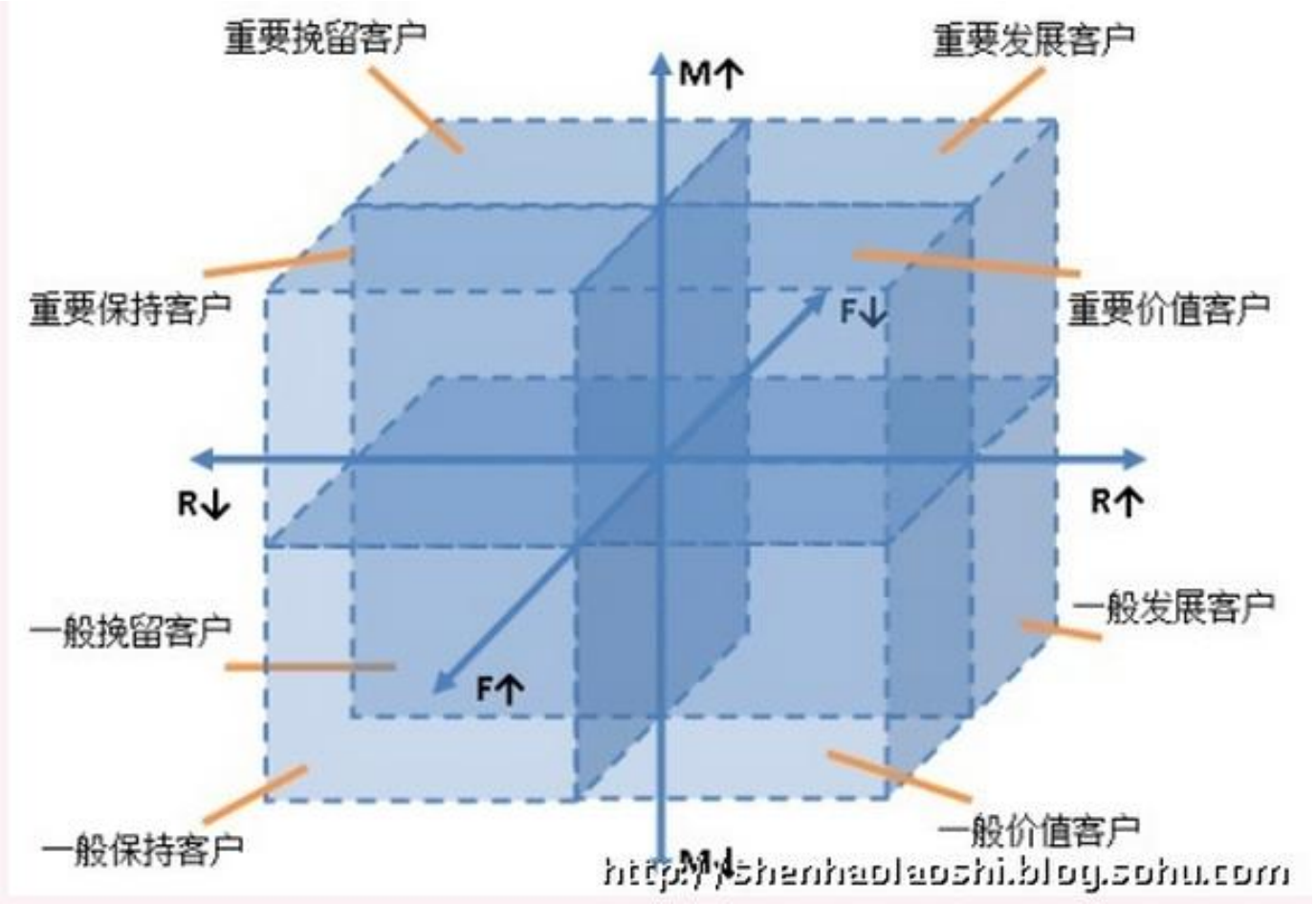
- 消费频率(Frequency)

消费频率是顾客在限定的期间内所购买的次数。最常购买的顾客，也是满意度最高的顾客。这个指标是“忠诚度”很好的代理变量。

- 消费金额(Monetary)

消费金额是最近消费的平均金额。是体现客户短期价值重要变量。如果你的预算不多，而且只能提供服务信息给2000或 3000个顾客，你会将信息邮寄给贡献40％收入的顾客，还是那些不到1％的顾客？数据库营销有时候就是这么简单。这样的营销所节省下来的成本会很可观 。

分析客户价值的一个示例



分析客户价值的一个示例

分析的起点—原始交易记录

无论是超市、电商、制造商、电信公司和银行，总会有一张类似于以下的这张表。其中主要的变量有：订单号、客户编码、订单时间、产生金额和交易类型。这个表为“RFM_TRAD_FLOW”

记录ID	客户编号	收银时间	销售金额	销售类型
00023351	010006	27SEP09:20:10…	58.00	特价
00023372	030031	27SEP09:21:33…	69.00	特价
00023447	040102	28SEP09:21:12…	69.00	特价
00023448	020173	28SEP09:21:12…	69.00	特价
00023449	040017	28SEP09:21:13…	69.00	特价
00023464	030132	28SEP09:21:37…	81.00	正常
00023465	020145	28SEP09:21:40…	60.00	正常
00023466	030087	28SEP09:21:45…	72.00	正常
00023467	020102	28SEP09:21:46…	51.00	正常
00023469	040051	28SEP09:21:51…	60.00	正常
00023472	010299	28SEP09:22:00…	60.00	正常
00023496	010128	29SEP09:19:11…	72.00	特价
00023499	040239	29SEP09:19:16…	51.00	特价
00023500	030108	29SEP09:19:16…	60.00	特价

天善智能
TIANSHAN SOFT

数据重组

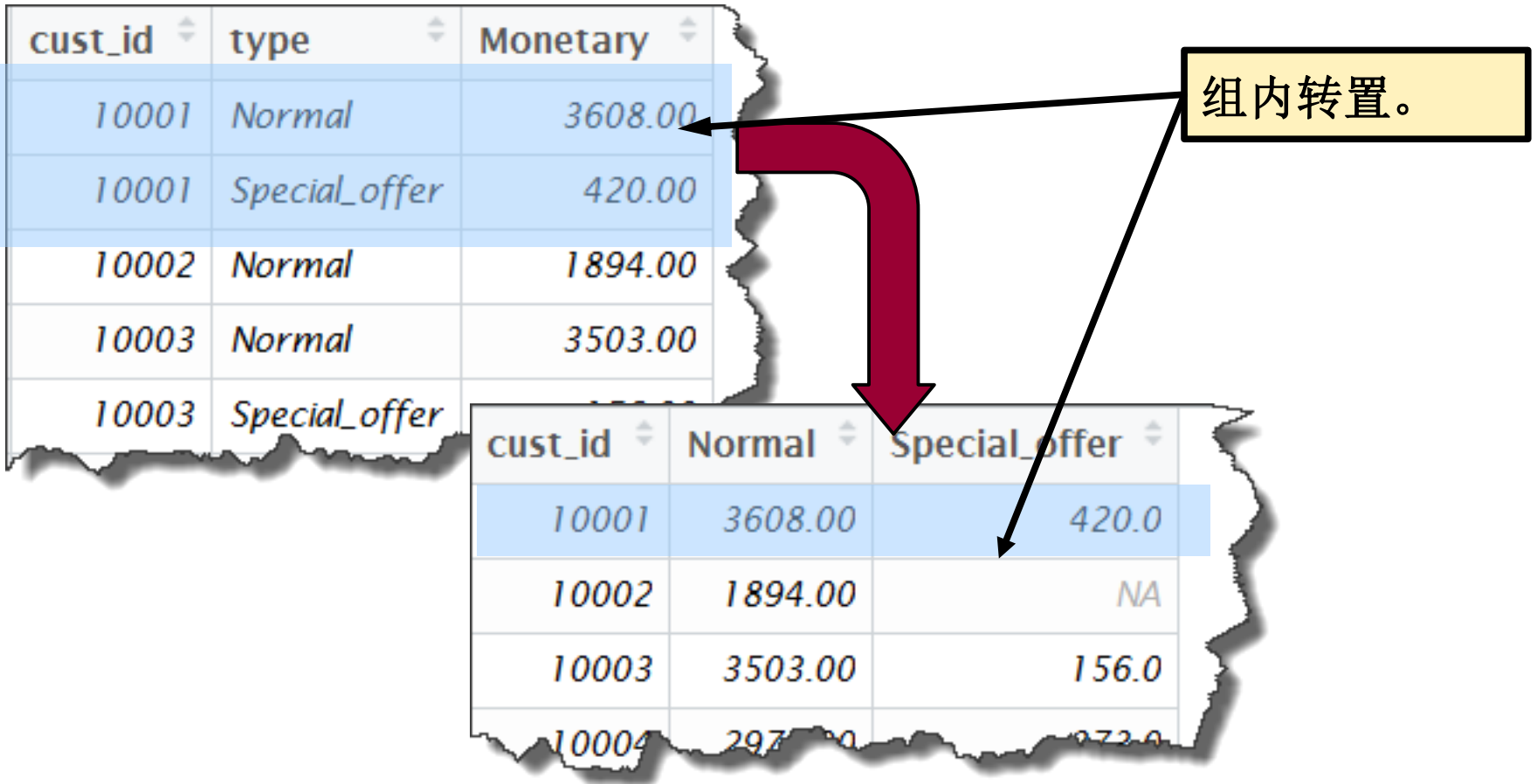
天善智能
TIANSHAN SOFT

拆分列

继续上一节案例，找出偏爱打折的客户：

cust_id	Normal	Special_offer	Special_offer_ratio
10001	3608.00	420.0	0.104270109
10002	1894.00	0.0	0.000000000
10003	3503.00	156.0	0.042634600
10004	2979.00	373.0	0.111276850
10005	2368.00	0.0	0.000000000
10006	2534.00	58.0	0.022376543
10007	4021.00	179.0	0.042619048

商业情景



•拆分列操作即按组做某个变量的转置，需要告知三个主要变量：1、分组依据；2、用于做变量名的列； 3、需要拆分的列； 。

天善智能
TIANSHAN SOFT

堆叠列

堆叠列

cust_id	Normal	Special_offer
10001	3608.00	420.0
10002	1894.00	NA
10003	3503.00	156.0
10004	297	

组内转置。

cust_id	type	Monetary
10001	Normal	3608.00
10001	Special_offer	420.00
10002	Normal	1894.00
10003	Normal	3503.00
10003	Special_offer	156.00

天善智能
TIANSHAN SOFT

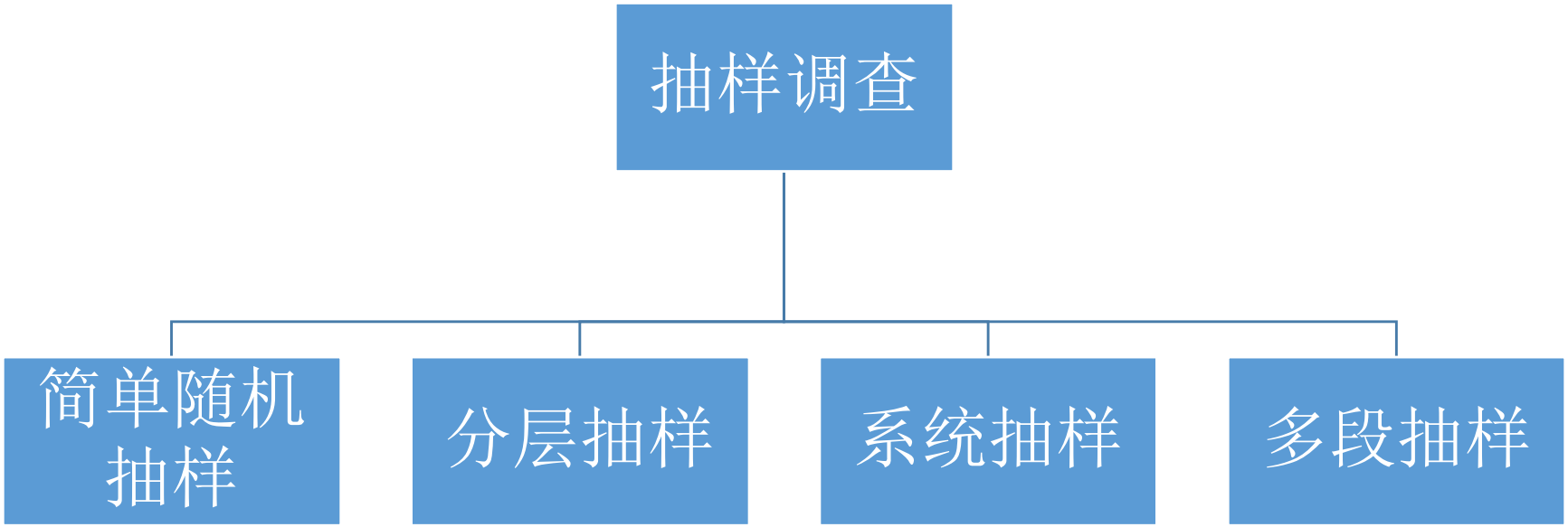
抽样

商业情景——客户调查

- 在每个地区随机选择5名客户进行满意度调查
- 预期数据格式（部分）

client_id	sex	birth_date	district_id
1	女	1970-12-13	18
2	男	1945-02-04	1
3	女	1940-10-09	1
4	男	1956-12-01	5
5	女	1960-07-03	5
6	男	1919-09-22	12
7	男	1929-01-25	15
8	女	1938-02-21	51

抽样调查分类



简单随机抽样（SPS）

从总体中不加任何分组、划类、排队等，完全随机地抽取调查单位。

特点：

- 每个样本单位被抽中的概率相等，样本的每个单位完全独立，彼此之间无一定的关联性和排斥性。
- 简单随机抽样是其他各种抽样形式的基础。通常只是在总体单位之间差异程度较小和数目较少时，才采用这种方法。

局限性：

- 当总体单位数很大时，就难以实现简单随机抽样，且抽样误差较大

分层抽样（STR）

也称类型抽样，总体分成不同的“层”，然后在每一层内进行抽样。

二种方法：

（1）等数分配法

（2）等比分配法

例：

- **企业按照大中小微型分类**
- **对家庭收入分为高收入、中等收入、低收入等**

系统抽样

也称等距抽样，其步骤如下：

（1）按某一标志值的大小将总体单位进行排队并按顺序编号

（2）根据确定的抽样比例确定抽样间距

（3）随机确定第一个样本单位

（4）按顺序总体等间距地抽取其余样本单位

- 系统抽样的随机性主要体现在第一个样本单位的选取上，因此一定要保证抽取第一个样本单位的随机性。
- 该方法适用于总体情况复杂，各单位之间差异较大，单位较多的情况。

多段抽样

将调查分成两个或两个以上的阶段进行抽样，第一阶段先将总体按照一定的规范分成若干抽样单位，称之为一级抽样单位，再把抽中的一级抽样单位分成若干个二级抽样单位，从抽中的二级抽样单位中再分三级抽样单位等，这样就形成一个多阶段抽样过程，分成若干个阶段逐步进行。

例：从某省抽取100人组成样本单位

省→市→县—100人

放回抽样

放回抽样又称重复抽样，每次从总体中随机地抽取一个样本单位，观察登记其标准值后又放回总体中，如此进行N次的抽样方法，

特点：

- **在重复抽样的过程中，被抽取的总体单位总数始终保持不变，每一次抽样中各总体单位被抽到的机会都相同，每次抽样结果相互独立；**
- **每一总体单位都有被重复抽取的可能。**

不放回抽样

不放回抽样也称不重复抽样，指被抽到的单位不再放回总体，每次仅在余下的总体单位中抽取下一个样本的抽样方法。

特点：

- **任意总体单位都不会被重复抽到；**
- **可以一次抽取所需要的样本单位数。**
- **在实际应用中通常采用的都是不重复抽样方法**

抽样在挖掘中的作用

- **快速获得数据的基本特征**
- **数据量较大，建模速度较慢**
- **数据不足时**
- **数据平衡**
- **数据分为训练集，测试集，验证集**

抽样与bagging集成算法

Bagging**的基本思路：**

第一步：从总体N**中重复抽样选取数据集**N1,N2,…**形成**{NI}

第二步：基于{NI}**，建立统计模型，分别给出每个模型的预测结果**

第三步：对结果进行投票

适用范围：

- **训练集之间相互独立，可以并行运行模型**

更多商业智能BI和大数据精品视频尽在 www.hellobi.com

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

天善智能
TIANSHAN SOFT