

数据挖掘方法论及案例 介绍

华为技术有限公司 BI开发部



数据挖掘是BI领域的一个重要应用方向

BI指通过对行业的认知、经验，结合数学理论、管理理论、市场营销理论，利用工具软件、数学算法（如：神经网络、遗传算法、聚类、客户细分等）对企业的数、业务、市场进行分析及预测，以图表、数据分析报告的形式支撑企业决策、市场营销、业务拓展、信息运营等工作。



数据挖掘在电信行业的应用

- ◆如何发现电信客户的特征和分类？
- ◆如何预测哪些即将流失的客户？
- ◆如何评价客户的贡献价值？
- ◆如何判断客户的欺诈行为特征？
- ◆如何发掘我的潜在客户？
- ◆还有更多.....

---如何对欠费/坏账进行预测和控制

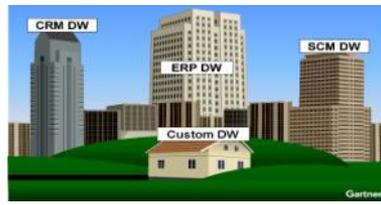
---大客户的消费行为特征是什么，人口统计学特征是什么

---如何知道公司下阶段收入情况，如何评估某一收入因素对整体收入的影响指数

数据 + 人 + 工具 + 算法 + 知识 + 预测 = 商业智能(BI)

数据挖掘

最有名的故事是：“啤酒和尿布”的故事
 最值钱的分析报告是：美国蓝德报告
 应用的最大工程是：伊拉克战争



目录

数据挖掘建模方法

数据挖掘算法介绍

数据挖掘案例分享

首先，了解数据挖掘的能力及应用

数据挖掘的能力：描述过去、预测未来。数据挖掘从算法角度分：预测类模型、非预测类模型、数据降维；从应用角度分：描述、预测、评估；常用算法包括：分类规则、聚类分析、神经网络、决策树；时间序列、回归分析、关联分析、贝叶斯网络、偏差检测；因子分析、主成分分析、数学公式

数据挖掘算法

应用领域

◆预测类模型

--连续变量

- a. 线性回归
- b. 非线性回归
- c. 时间序列

--离散变量

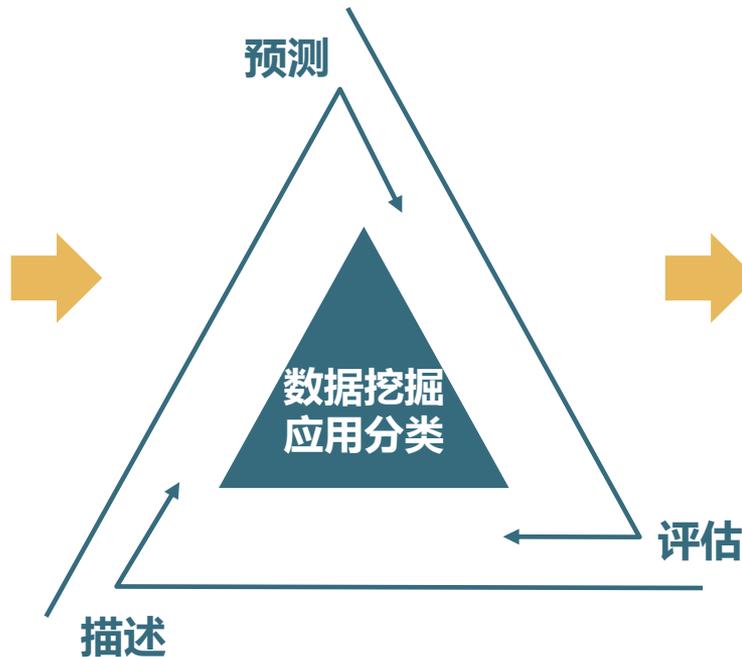
- a. 神经网络
- b. 决策树
- c. Logistic回归
- d. 贝叶斯网络

◆非预测类模型

- 聚类分析
- 关联分析
- 偏差检测

◆数据降维

- 因子分析
- 主成分分析
- 数学公式



客户管理

产品服务

市场运营

客户
细分

交叉
营销

市场
预测

客户
获取

资费
管理

信用
管理

客户
价值

服务
管理

欠费
管理

客户
流失

渠道
管理

异常
发现

其次，清楚数据挖掘建模方法论（CRISP-DM）

- 数据挖掘：需明确数据挖掘目标以及业务需求
- 需要在业务的基础上，给出可实现的算法
- 输出数据挖掘具体实施方案

输入：数据挖掘目标

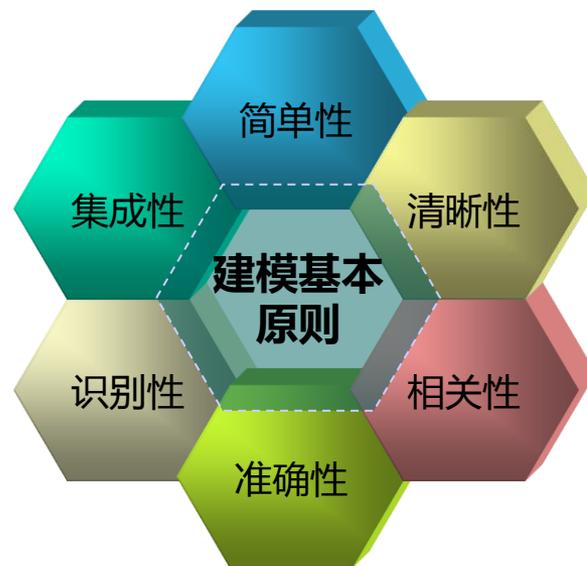
业务现状

业务需求

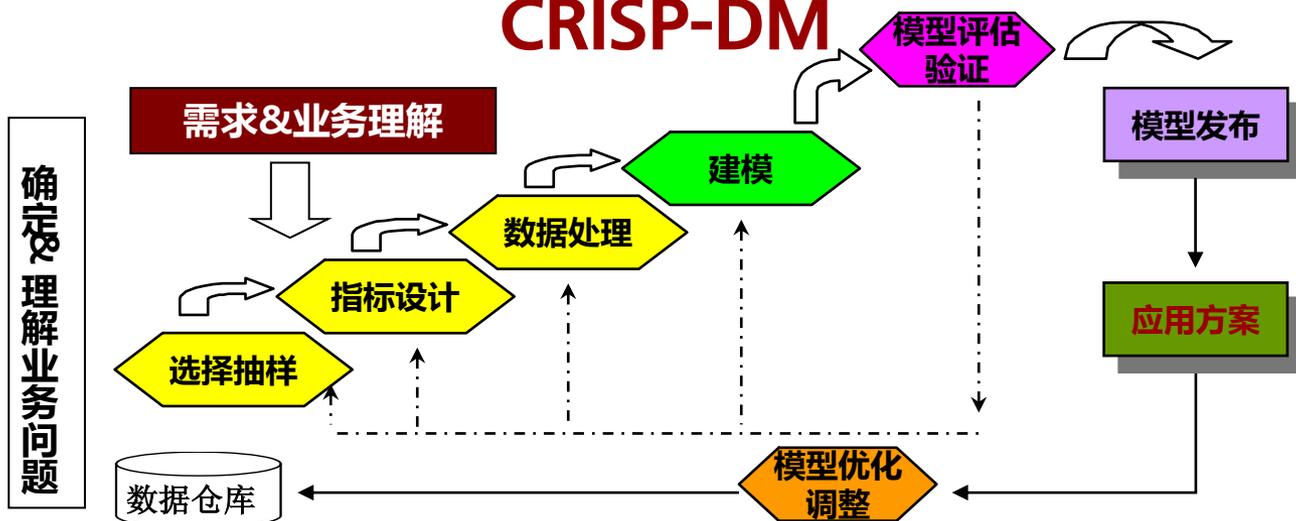
输出：实现算法

实施方案

应用方案



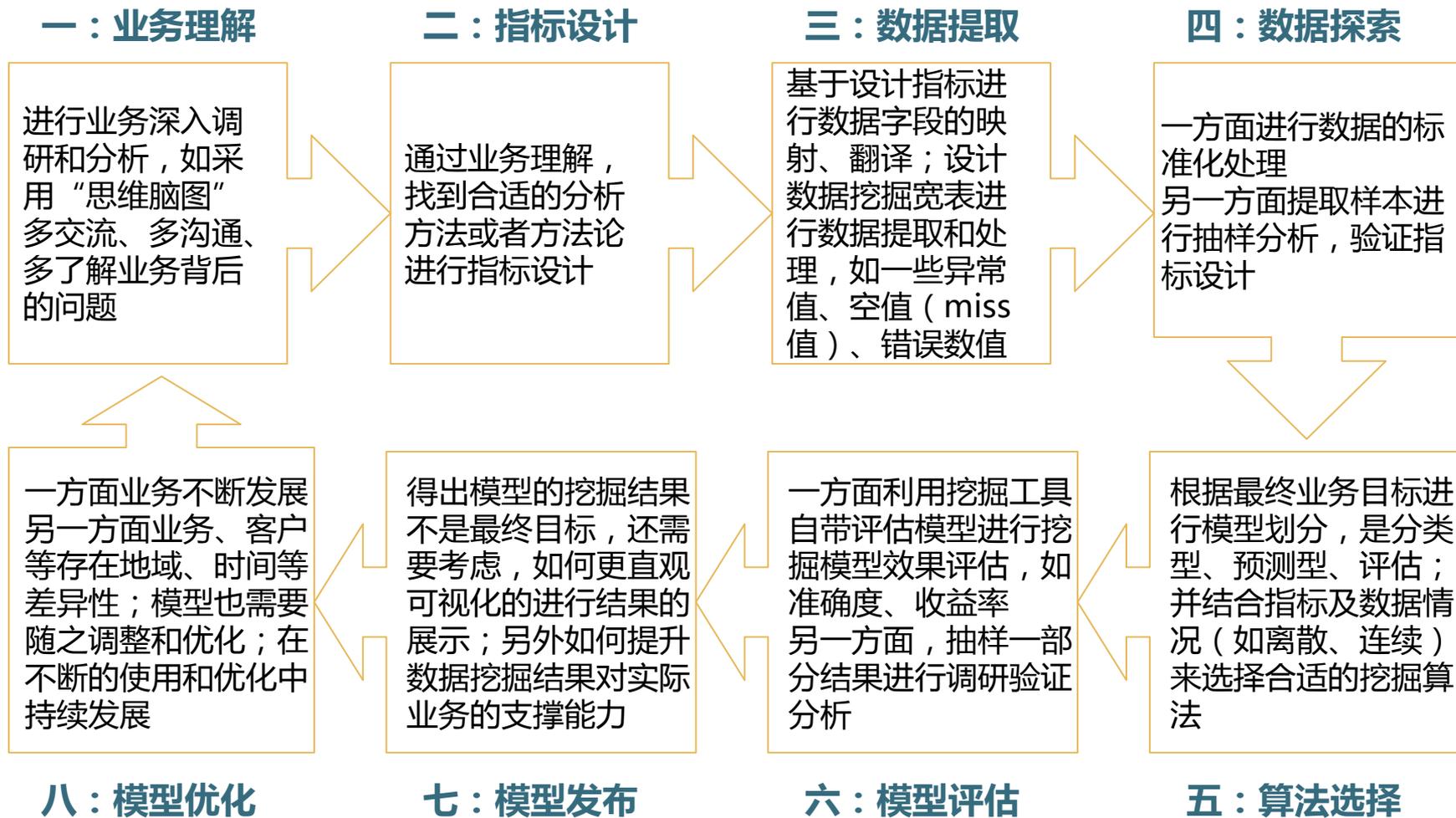
CRISP-DM



- ◆ 遵循CRISP-DM（跨行业数据挖掘标准过程）原则和建模基本原则
- ◆ 制定一套切实可行的数据挖掘实施方法论。
- ◆ 基于模型结果构建端到端的应用支撑

再次，掌握数据挖掘建模常规步骤（八步法）

数据挖掘建模八步法指：业务理解、指标设计、数据提取、数据探索、算法选择、模型评估、模型发布、模型优化



步骤一：业务理解

常见的误区：很多人以为不需要事先确定问题和目标，只要对数据使用数据挖掘技术，然后再对分析挖掘后的结果进行寻找和解释，自然会找到一些以前我们不知道的，有用的规律和知识。



我们要什么样的数据挖掘模型？

可解释的！有实际业务涵义！可使用的！



◆根据掌握的相关业务情况总结分析；思维脑图是个不错的选择

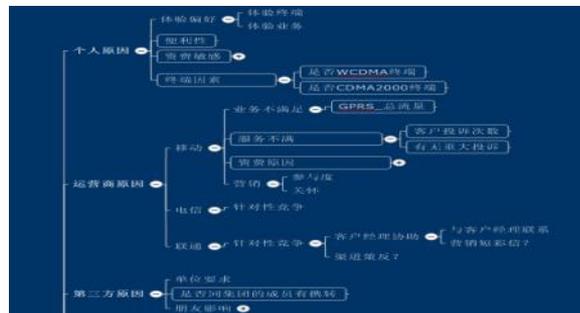
◆如：“携号转网预测模型”

1. 业务产生的背景：携号转网业务在天津、海南等地启动
2. 业务目前的发展情况：大量用户申请携转到竞争对手
3. 业务带来的影响：客户、高端客户流失
4. 需要我们做什么：找到携转倾向较高的用户，进行挽留

5. 为什么会出现这种情况

1. 个人原因
2. 运营商原因
3. 第三方原因

(找到一种合适的分析方法进行分析)



步骤二：指标设计

基于对业务问题的梳理分析，找到合适的分析方法或者方法论指导模型指标设计，确保指标体系化、全面性。
常见的一些分析方法

战略管理	SWOT分析、PEST分析、麦肯锡7s分析、五力模型、波士顿矩阵、通用矩阵、平衡计分卡、企业价值链
营销	4P-4C-4R、体验式营销、资费管理4阶段、品牌健康度、AIDA模型、精准营销、整合营销
服务、渠道	客户满意度、客户期望值管理、KANO服务质量模型
客户类心理类	马斯洛需求理论、客户画像视图、峰终定律、感觉适应定律、心理定势、决策价值链

仍以“携号转网预测模型”为例 **基于用户决策价值链“携号转网驱动力”分析进行指标设计**

认知需要

收集信息

评价选择

购买决策

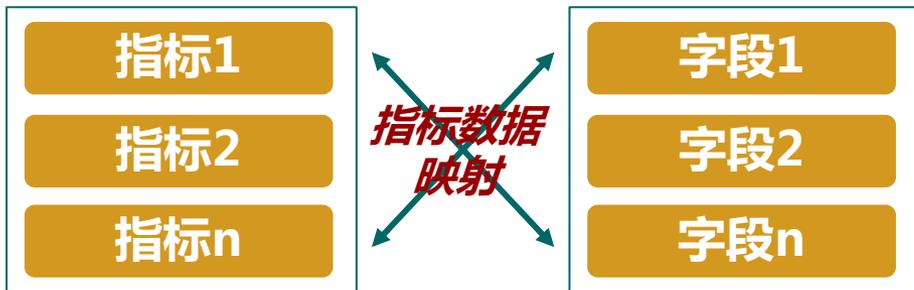
购后行为



步骤三：数据提取

数据提取确保建模数据的完整性、可用性和完整性。

如：携号转网预测指标设计与数据映射



符合转网条件	转网条件符合度	终端捆绑	是否捆绑赠机	if_ter_bind	
		债务纠纷	是否欠费用户	owe_flag	
		业务捆绑	是否捆绑营销包	if_bind	
		资料不符	检查身份证规则位数和姓名	IF_ID_ACCORD	
内在驱动力负向指数	粘性业务度		是否VPMN用户	if_vpmn	
			是否手机报用户	if_mpaper_flag	
			是否新闻早晚报用户	if_xw_mpaper_flag	
			是否号簿管家用户	if_hb_flag	
			是否139邮箱活跃用户	if_mail_flag	
			是否飞信活跃用户	if_fetion_active_flag	
			是否12580活跃用户	if_12580_flag	
			营销活动参与次数	无	
			便利性	家庭住址附近有哪家运营商	无
		内在驱动力正向指数	资费原因		话费查询次数
	漫游通话费占比			roam_fee/total_fee_all	
	长途费占比			ld_fee/total_fee_all	
	GPRS收入占比			gprs_fee/total_fee_all	
终端适配度			是否为Iphone	if_iphone	
			是否WCDMA终端	term_type	
	是否CDMA2000终端		无		
服务满意度			客户投诉次数	ts_count	
			有无重大投诉	if_uppt_ts	
理智选择	业务满意度			当月超出月租通信费金额占比	up_month_comm_fee_scale
			当月超出月租流量费金额占比	up_month_gprs_fee_scale	

- 缺失数据处理
- 极值数据处理
- 错误数据处理
- 冗余数据处理

- 数据挖掘宽表构建

数据提取

数据清洗

数据审核

数据集成

- 提取建模所需数据

- 数据统计错误审核
- 数据源错误审核
- 数据统计口径审核

步骤四：数据探索

□数据探索主要涉及两项工作：第一，进行数据检测、分析、验证是否符合指标设计初衷和业务涵义；第二，根据建模需要进行部分数据的标准化处理，使不同的指标在相同的量纲上进行数学运算。



□汇总统计：

- 频数和众数、百分位数
- 位置度量（均值、中位数）
- 散布度量（方差、极差）

□异常检测：

- 是否符合业务涵义
- 是否普遍性
- 是否存在异常值

□筛选指标：

- 利用相关性分析检查指标是否存在重复
- 是否达到数据质量要求

□指标衍生：

- 一些指标是否需要相应的处理，如幂处理、对数处理和标准化处理

数据标准化常用方法

1) 最小-最大规范化

对原始数据进行线性变化，假定minA和maxA分别为属性A的最小和最大值，最小-最大规范化通过计算：

$$v' = \frac{v - \min A}{\max A - \min A} (\text{new_max } A - \text{new_min } A) + \text{new_min } A$$

把属性A的值映射到[new_minA,new_maxA]区间内；

2) z-score规范化（或零-均值规范化）

属性A的值基于A的平均值和标准差标准化，A的值v标准化v' 由下试计算得到：

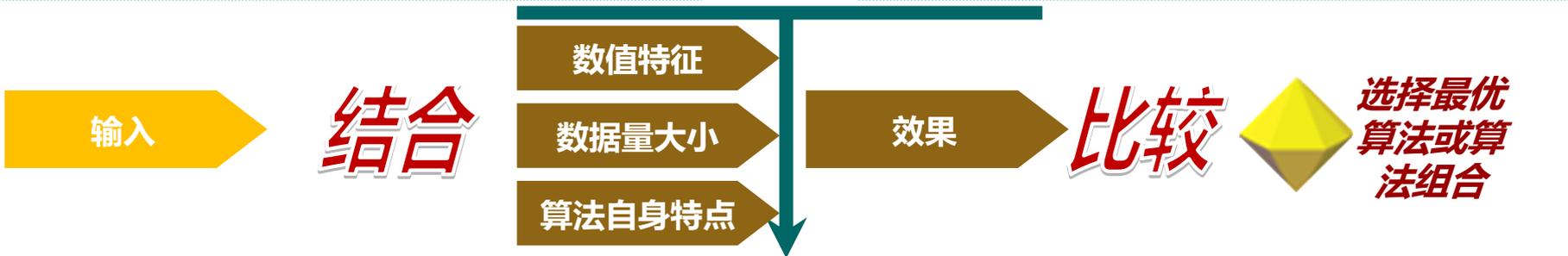
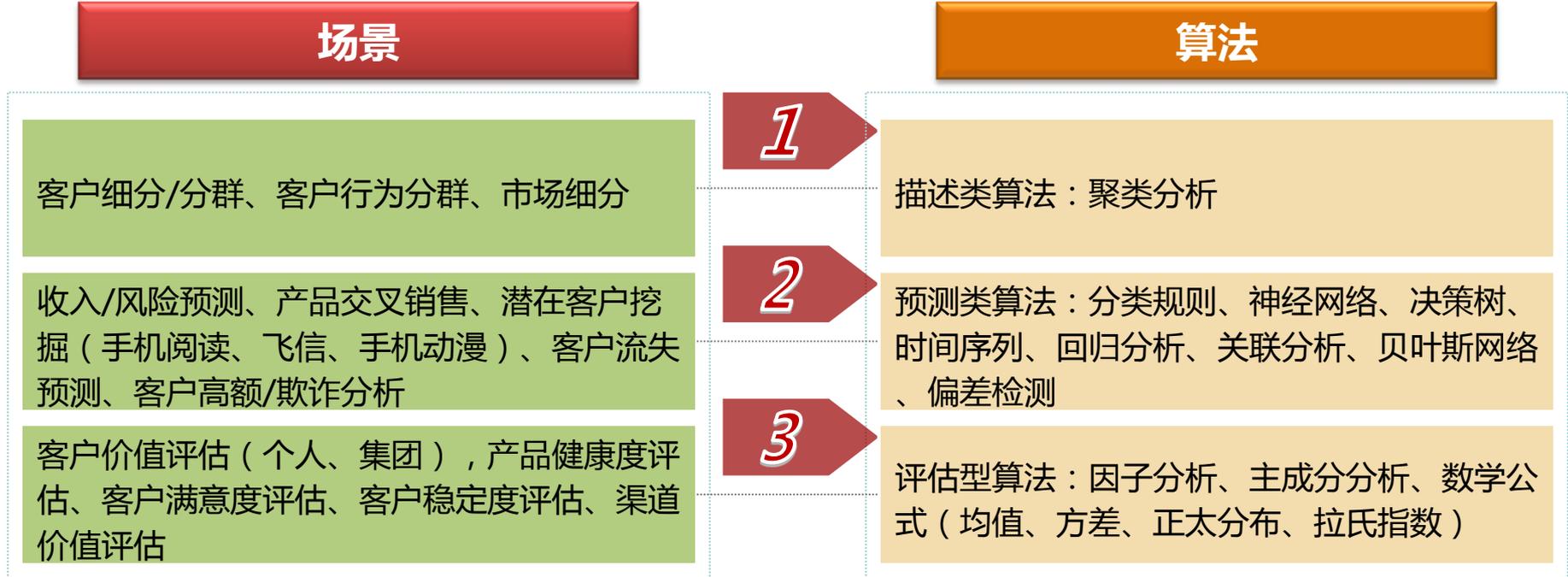
$$v' = \frac{v - \bar{A}}{\sigma_A} \quad \bar{A} \text{ 和 } \sigma_A \text{ 分别为A的均值和标准差；}$$

3) 小数定标规范化

$$v' = \frac{v}{10^j} \quad j \text{ 是使得 } \max(|v'|) < 1 \text{ 最小的整数；}$$

步骤五：算法选择

□根据建模场景进行算法选择：如：描述类有分类规则、聚类分析，预测类有、神经网络、决策树、时间序列、回归分析、关联分析、贝叶斯网络、偏差检测，评估类有因子分析、主成分分析、数学公式；并结合数据情况（如离散值、连续值，数据量大小）等选择合适的算法



步骤六：模型评估

- 模型评估目的在于：什么样的模型是有效的？模型的实际应用效果如何？
- 根据样本数据，模型结果实际效果反馈数据进行模型评估

评价标准

- 采用工具的“分析”输出节点和“评估”图形节点来进行评分
- 一般来讲“分析”节点的准确率高于75%、覆盖率高于85%为有效模型，
- “评估”节点中的提升度高于2为有效模型。

评估方法

- 确定评估对象为非C、R中的用户，设评估组和参照组。
- 参照组参照依据为当月T中转网申请率即{X/T的统计量},即参照组的准确率为转网申请率；
- 评估组的选择对象考虑用模型预测置信度90%以上的用户（且满足R的选择条件），其预测准确率为评估指标。原则上该指标>经验值即为可接受的

评估工具

- **评估分析**：使用分析节点，可以对模型生成准确预测的能力进行评估。
- **增益图**：(分位数中的匹配数量/全部匹配数量) × 100%
- **提升图**：(在分位数中的匹配/在分位数中的记录) / (全部匹配/全部记录)。

输出字段 APP_XIECHU_FLAG 的结果

比较 \$C-APP_XIECHU_FLAG 与 APP_XIECHU_FLAG

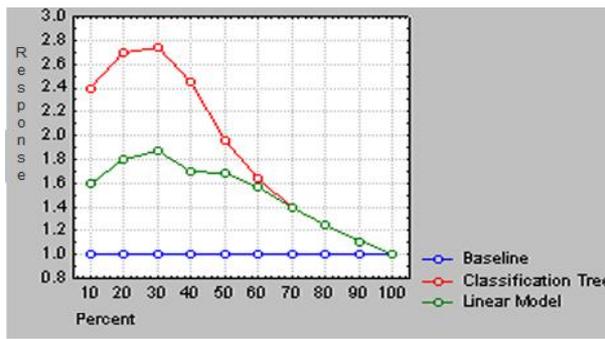
分区	1_训练	2_测试		
正确	171,704	73.24%	19,093	72.77%
错误	62,729	26.76%	7,143	27.23%
总计	234,433		26,236	

\$C-APP_XIECHU_FLAG 的重合矩阵 (行表示实际值)

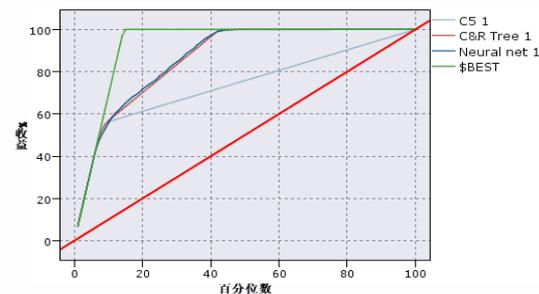
分区 = 1_训练	0	1
0	157,853	57,997
1	4,732	13,851

分区 = 2_测试	0	1
0	17,646	6,504
1	639	1,447

评估分析



lift图



增益图

步骤七：模型发布

□ 聚焦业务问题提供端到端的专题解决方案；提高数据挖掘应用的效果和价值



模型发布是：一套端到端、完整的数据挖掘专题解决方案、而非单纯的数据挖掘结果

步骤八：模型优化

模型初步构建

模型优化

模型带动业务

业务带动模型

模型
初期

- 模型初步构建进行模型验证

模型上
升期

- 根据模型验证和业务情况进行模型优化

模型成
熟期

- 模型准确率达到相应精度、稳定成熟引领业务发展

模型衰
退期

- 伴随业务的发展模型不再适用新的业务环境，逐步停下脚步。



一个生命力强、可持续应用的模型离不开“模型优化”的浇灌

目 录

数据挖掘建模方法

数据挖掘算法介绍

数据挖掘案例分享

算法介绍：聚类算法

聚类就是对大量未知标注的数据集，按数据的内在相似性将数据集划分为多个类别，使类别内的数据相似度较大而类别间的数据相似度较小。

模式相似性测度：1) 聚类测度 2) 相似测度 3) 匹配测度

1. 欧氏(Euclidean)距离
$$d(\bar{x}, \bar{y}) = \|\bar{x} - \bar{y}\| = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$$

2. 绝对值距离
$$d(\bar{x}, \bar{y}) = \sum_{i=1}^n |x_i - y_i|$$

3. 切氏(Chebyshev)距离
$$d(\bar{x}, \bar{y}) = \max_i |x_i - y_i|$$

4. 明氏距离
$$d(\bar{x}, \bar{y}) = \left[\sum_{i=1}^n |x_i - y_i|^m \right]^{1/m}$$

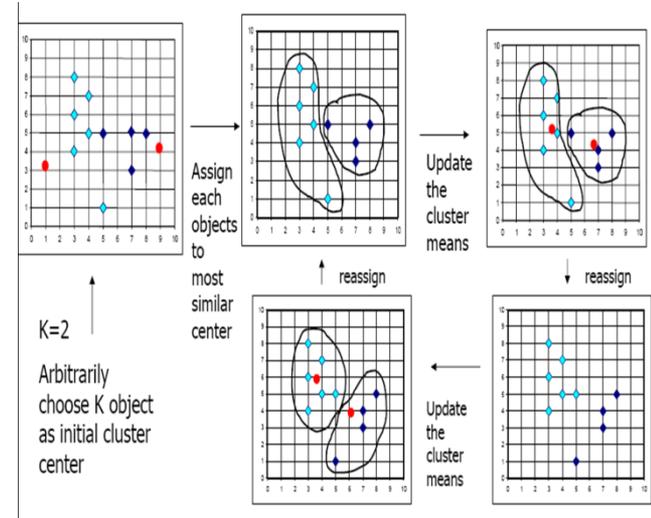
5. 马氏距离
$$d^2(\bar{x}_i, \bar{x}_j) = (\bar{x}_i - \bar{x}_j)' V^{-1} (\bar{x}_i - \bar{x}_j)$$

6. 向量余弦夹角
$$\cos(\bar{x}, \bar{y}) = \frac{\bar{x}'\bar{y}}{\|\bar{x}\|\|\bar{y}\|} = \frac{\bar{x}'\bar{y}}{[(\bar{x}'\bar{x})(\bar{y}'\bar{y})]^{1/2}}$$

7. 相似系数
$$sim(x, y) = \frac{\sum_{s \in S_{xy}} R_{xs} R_{ys}}{\sqrt{\sum_{s \in S_{xy}} R_{xs}^2} \sqrt{\sum_{s \in S_{xy}} R_{ys}^2}}$$

聚类的方法 (1) 最近距离法 (2) 最远距离法 (3) 中间距离法 (4) 重心距离法 (5) 平均距离法 (6) 离差平方和法

判别分类结果好坏的一般标准：**类内距离小，类间距离大**



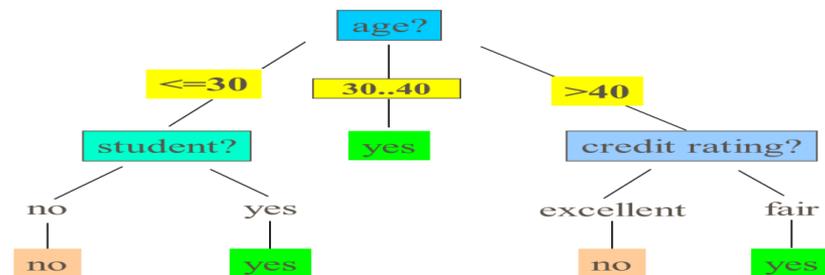
如：K-means算法

- 1 设置初始类别中心和类别数;
- 2 根据类别中心对数据进行类别划分;
- 3 重新计算当前类别划分下每类的中心;
- 4 在得到类别中心下继续进行类别划分;
- 5 如果连续两次的类别划分结果不变则停止算法;否则循环2 ~ 5;

算法介绍：决策树

□决策树一般都是自上而下的来生成的。每个决策或事件（即自然状态）都可能引出两个或多个事件，导致不同的结果，把这种决策分支画成图形很像一棵树的枝干，故称决策树。

□决策树主要是提取分类规则，进行分类预测。



```

IF age = "<=30" AND student = "no" THEN buys_computer = "no"
IF age = "<=30" AND student = "yes" THEN buys_computer = "yes"
IF age = "31..40" THEN buys_computer = "yes"
IF age = ">40" AND credit_rating = "excellent" THEN buys_computer = "no"
IF age = ">40" AND credit_rating = "fair" THEN buys_computer = "yes"
    
```

优点：

- 使用者不需要了解很多背景知识，只要训练事例能用属性→结论的方式表达出来，就能用该算法学习；
- 决策树模型效率高，对训练集数据量较大的情况较为适合；
- 分类模型是树状结构，简单直观，可将到达每个叶结点的路径转换为IF→THEN形式的规则，易于理解；
- 决策树方法具有较高的分类精确度。

算法介绍：Logistic回归

logistic回归是一个概率型模型，因此可以利用它预测某事件发生的概率。例如在临床上可以根据患者的一些检查指标，判断患某种疾病的概率有多大。

线性回归模型

$$\hat{E}(Y) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k \dots \dots \dots (1)$$

因为 $Y=0$ 或 1 两个分布，而 $E(Y)=P(Y=1)=P$ 是连续的；对于概率来讲其区间是 $[0, 1]$ 。显然线性模型不能达到这一点。我们可以通过对 P 的一种变换（LOGIT变换）

$$\text{Logit}(p) = \ln(p/(1-p))$$

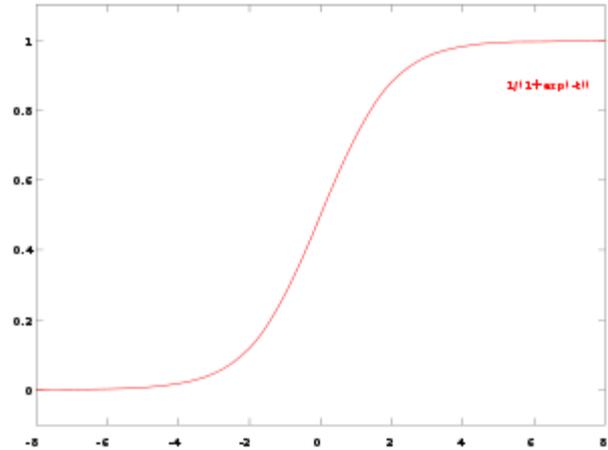
使得logit(p)与自变量之间存在线性相关的关系

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m =$$

$$P = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m)]}$$

常数项 表示因素为0时个体发生与不发生概率之比的自然对数。

回归系数 $\beta_j (j = 1, 2, \dots, m)$ 表示自变量 X_j 改变一个单位时logit P 的变量。



Logistic回归范围【0,1】

参数估计原理：最大似然 (likelihood) 估计

$$L = \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1 - Y_i}$$

↓

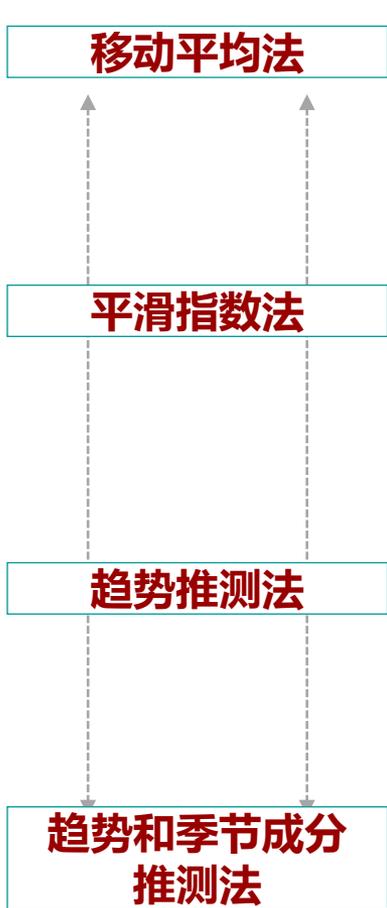
$$\ln L = \sum_{i=1}^n [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)]$$

↓

$$b_0, b_1, b_2, \dots, b_m$$

算法介绍：时间序列

- 时间序列分析法是根据过去的变化趋势预测未来的发展,它的前提是假定事物的过去延续到未来。
- 时间序列分趋势、循环、季节和不规则四种成分；主要方法有移动平均法、平滑指数法、趋势推测法、趋势和季节成分推测法。



$$\text{移动平均数} = \frac{\text{最近}n\text{期数据之和}}{n}$$

把若干历史时期的统计数值作为观察值，求出算术平均数作为下期预测值

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t$$

$$F_4 = \alpha Y_3 + (1 - \alpha) F_3 = \alpha Y_3 + (1 - \alpha) [\alpha Y_2 + (1 - \alpha) Y_1]$$

$$= \alpha Y_3 + \alpha(1 - \alpha) Y_2 + (1 - \alpha)^2 Y_1$$

F_{t+1} —— $t+1$ 期时间序列的预测
 Y_t —— t 期时间序列的实际值；
 F_t —— t 期时间序列的预测值；
 α ——平滑常数 ($0 \leq \alpha \leq 1$)

$$\text{线形方程: } T_t = b_0 + b_1 t$$

π_t —— t 期时间序列的趋势值；
 b_0 ——线性趋势的截距；
 b_1 ——线性趋势的斜率；
 t ——时间。

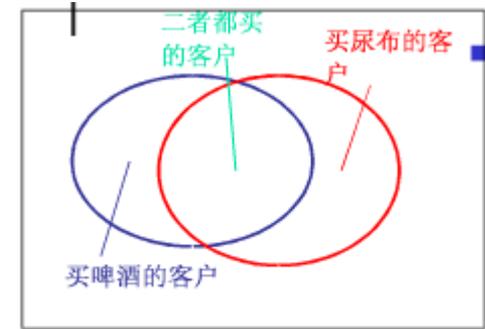
$$Y_t = T_t \times S_t \times I_t$$

Y_t -时间序列的数值
 T --趋势成分
 S -季节成分
 I --不规则成分

算法介绍：关联分析

关联规则挖掘是寻找数据项中的**有趣联系**，决定哪些事情将**一起发生**。如当一个事务中顾客购买了一样东西{钢笔}(这里X=“钢笔”)则很可能他同时还购买了{墨水}(这里Y=“墨水”)，这就是关联规则。

记录号	所购物品清单
1	啤酒、尿布，婴儿爽身粉，面包，雨伞
2	尿布，婴儿爽身粉
3	啤酒、尿布，牛奶
4	尿布，啤酒，洗衣粉
5	啤酒，牛奶，可乐饮料



尿布和啤酒赫然摆在一起出售。但是这个奇怪的举措却使尿布和啤酒的销量双双增加了。这不是一个笑话，而是发生在美国沃尔玛连锁店超市的真实案例，并一直为商家所津津乐道。

■ 期望可信度 (是否有意义)

$$Expected\ confidence\ (B) = P(B/总) = B发生的次数占事务的总和$$

■ 支持度 (关联规则重要性)

$$support\ (A \Rightarrow B) = P(A \cup B) =$$

$$\frac{A和B同时发生的次数}{事务的总和}$$

■ 置信度 (关联规则准确率)

$$confidence\ (A \Rightarrow B) = P(B/A) =$$

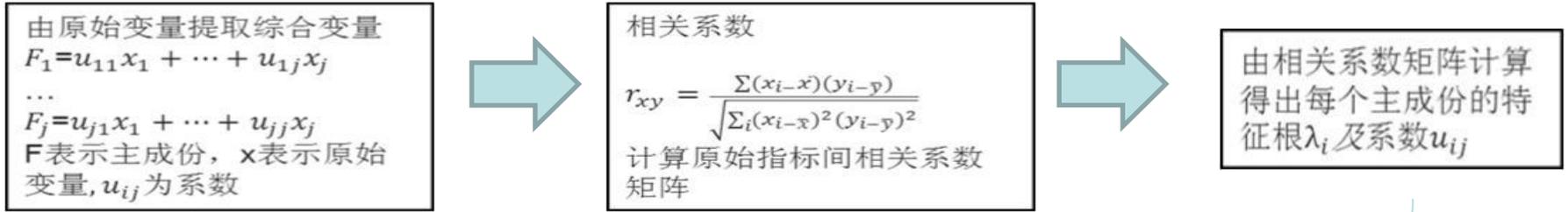
$$\frac{A和B同时发生的次数}{A发生的次数}$$

■ 提升度 (效果) = 置信度 / 期望可信度

算法介绍：因子分析/主成分分析

□当反映事物的方面太多时，过多的指标会对所描述的对象造成混乱，往往得不到正确的结论。因此，应当把相关的维度进行总结概括，尽量降低数据的维度（指标），简要对事物特征进行描述。

□通过主成份分析进行指标降维得到综合指标；再有综合指标与原始指标关系确定各指标与综合指标的系数



计算步骤：

- 计算各指标间相关系数矩阵
- 根据相关系数矩阵计算特征根及对应的特征向量（主成份与原始指标的系数）
- 计算方差贡献率，并根据方差贡献率选取主成份个数
- 计算各主成份的得分
- 根据各主成份贡献率及其对应的主成份得分计算出综合得分来反映各个集团的综合情况

通过累计方差贡献率 = $\frac{\text{选择的特征根之和}}{\text{特征根总和}}$
 确定主成份数量，默认80%以上即可反映原来指标的所有信息。

综合得分 = $\sum_{i=1}^n (\text{第}i\text{主成份得分} * \text{方差贡献率}_i)$

综合得分 = $\sum_{ij=1}^n (u_{i1}x_1 + \dots + u_{ij}x_j) * \text{方差贡献率}_i$
 $= (u_{11} + \dots + u_{1j}) * \text{方差贡献率}_1 * x_1 + \dots + (u_{ij} + \dots + u_{ij}) * \text{方差贡献率}_i * x_j$

得出原始指标与综合得分的关系系数
 系数 = $\sum_{ij=1}^n (u_{i1} + \dots + u_{ij}) * \text{方差贡献率}_i$

优势：
 采用指标间相关性计算优势是：当样本量达到一定程度后，相关的结果受样本量的变化的影响很小。

算法介绍：数学公式

■方差：一个较大的方差，代表大部分的数值和其平均值之间差异较大；一个较小的方差，代表这些数值较接近平均值

■均值 $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ ■方差 $\sigma = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$

■概率密度函数：对于随机变量X，如果存在非负可积函数 $f(x)$ ($-\infty, +\infty$)，使得对任意实数x，有

$$F(x) = \int_{-\infty}^x f(t) dt \quad P(X \leq x)$$

则称X为连续型随机变量，称 $f(x)$ 为 x 的概率密度函数，简称为概率密度。

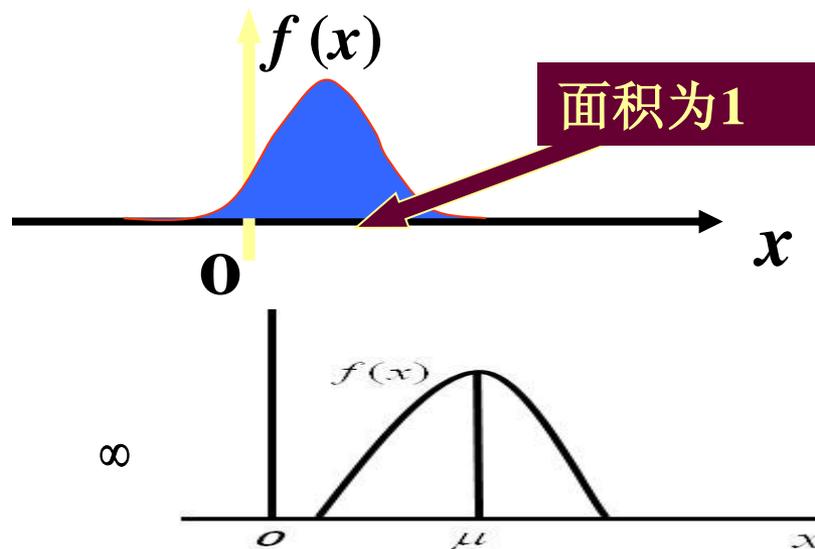
■概率密度函数性质

1. $f(x) \geq 0$

2. $\int_{-\infty}^{+\infty} f(x) dx = 1$

■正态分布函数

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



目 录

数据挖掘建模方法

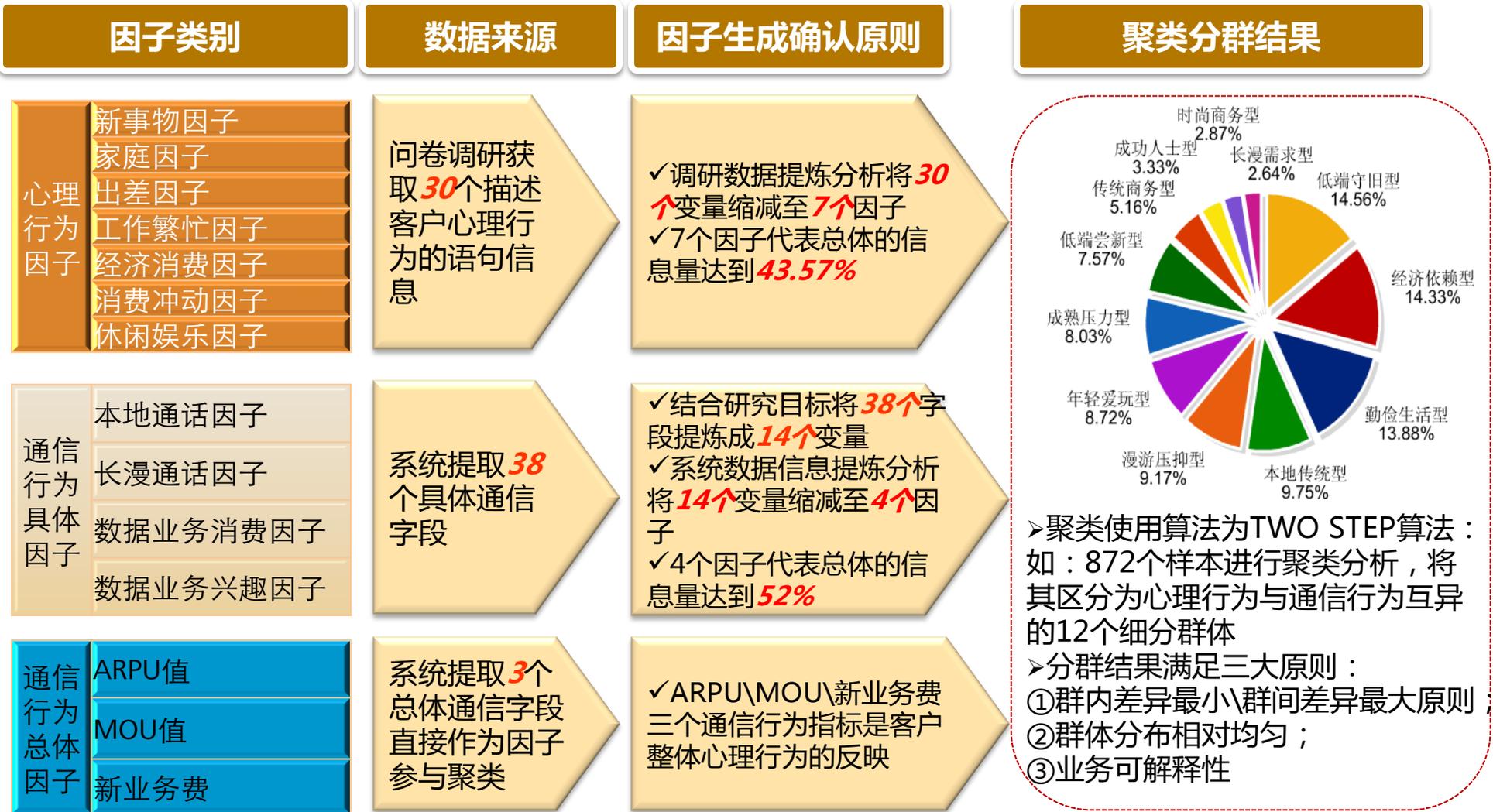
数据挖掘算法介绍

数据挖掘案例分享

华为数据挖掘模型在电信行业的应用

维度		模型
客户	个人	客户行为分群（消费、使用、活动、接触）、客户价值分群、客户流失预测、客户信用度评估、客户高额/欺诈分析、客户来源及离网去向分析、核心客户保有
	集团	集团客户流失预测、集团成员流失预测、集团客户价值评估、集团客户健康度评估、集团客户识别模型
	家庭	家庭客户识别模型、家庭客户小区定位模型
产品	增值业务	产品关联性分析（交叉销售）、产品价值分析、业务/产品健康度分析
	集团产品	集团业务粘性模型、集团业务健康度模型
	资费产品	资费产品生命周期识别模型、资费产品健康度模型
资源		智能资源管理模型、定制终端潜在客户挖掘模型、定制终端效益评估
合作伙伴		SP欺诈识别模型
渠道	自营渠道	自营渠道效益评估模型、自营渠道价值评估模型
	社会渠道	社会渠道价值评估模型、社会渠道违规监控模型、社会渠道流失预警模型
	电子渠道	电子渠道分流模型、电子渠道传播能力评估模型
内部运营		收入风险监控模型、收入预测模型、收入诊断模型、垃圾短信识别模型

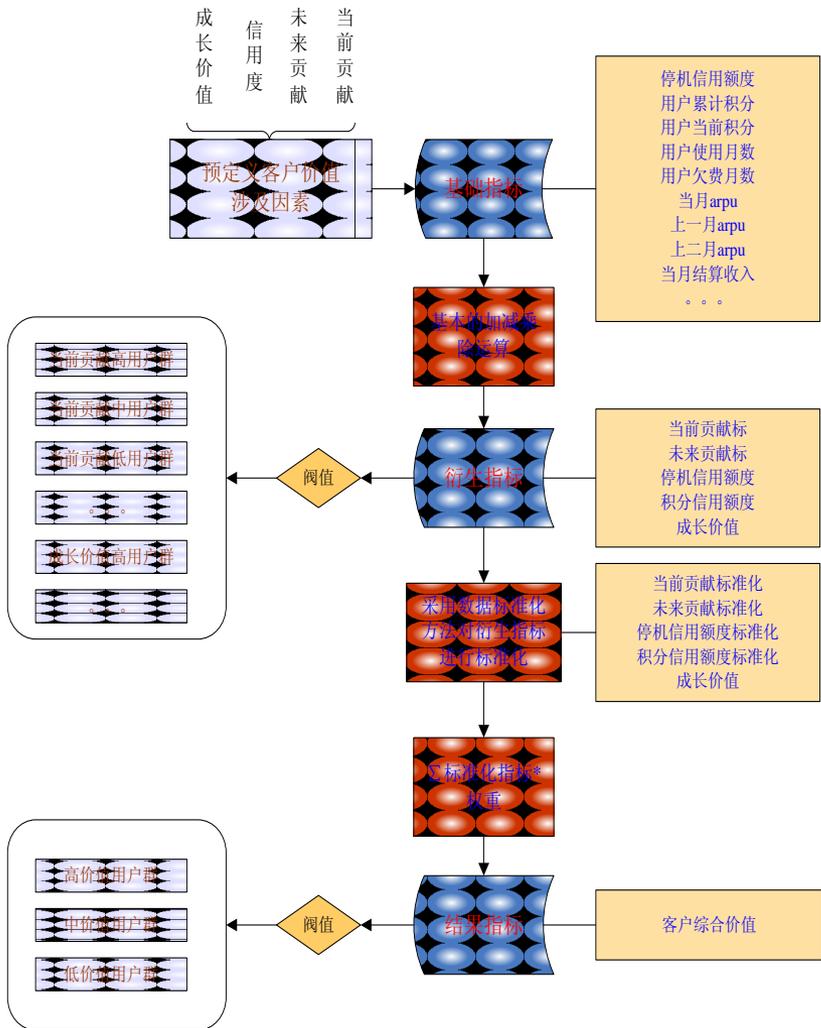
案例1：客户细分模型



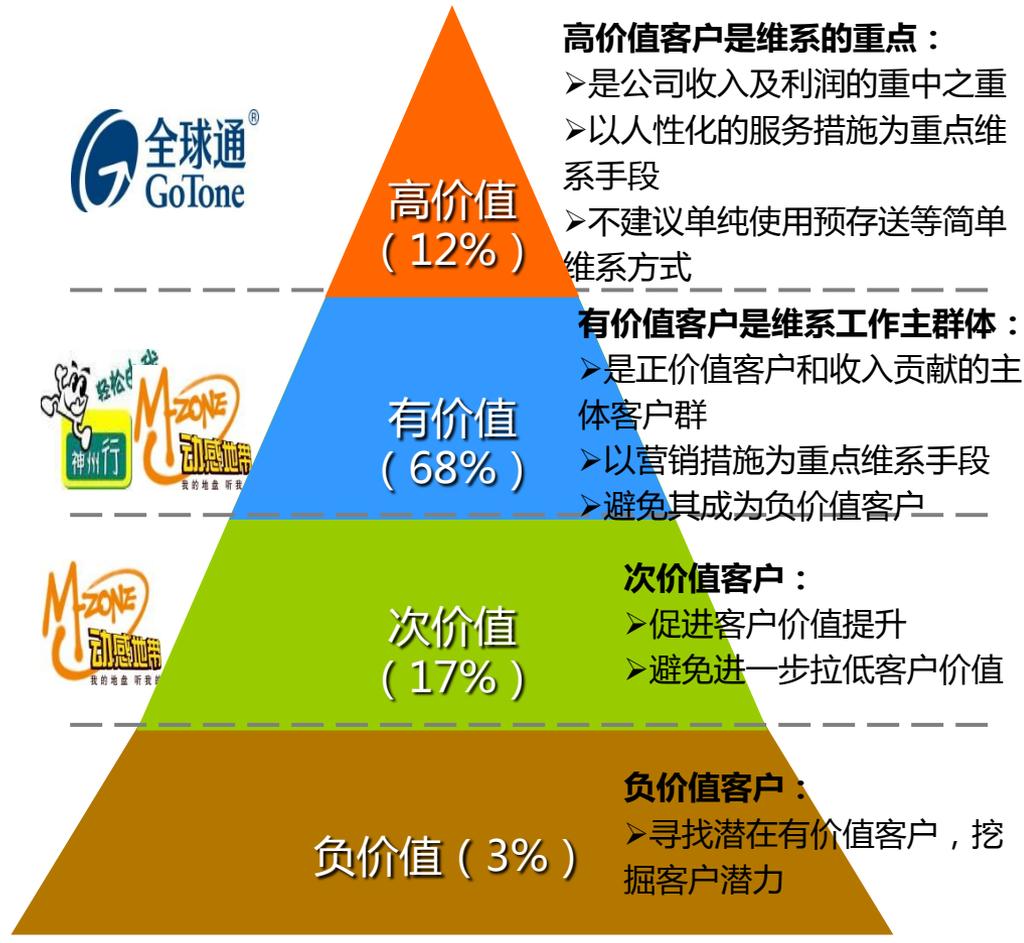
如时尚商务型：此类用户对语音、数据业务及新兴产品的需求匹配度都很高，采取营销产品应全面渗透策略

案例2：客户价值评估

从四个方面来评价客户的综合价值，按照客户贡献价值高低，确定四类价值客户：负价值、次价值、有价值、高价值群。

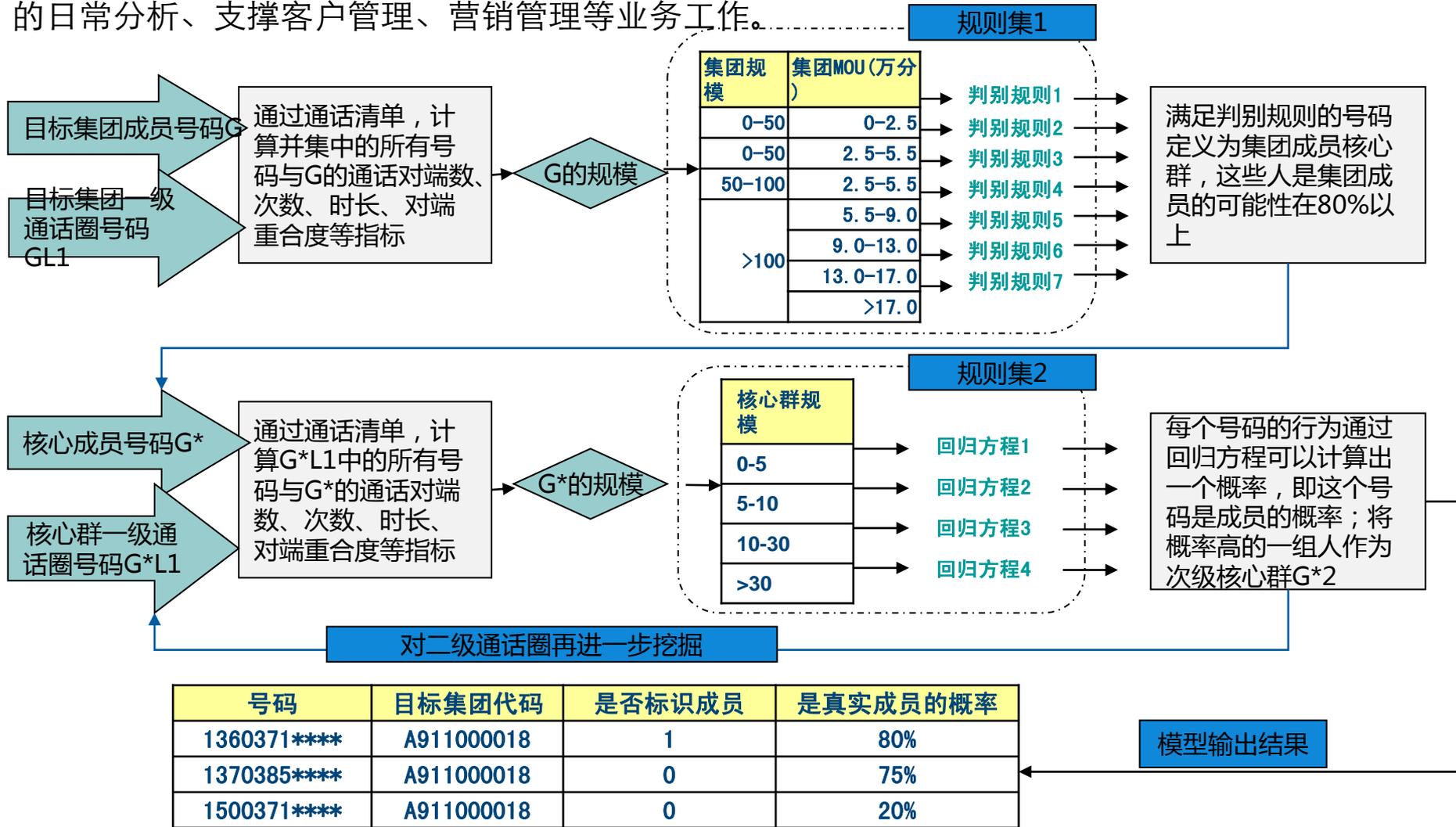


策略：重点维系高价值、有价值客户；提升次价值、负价值客户贡献



案例3：集团客户识别模型

基于集团成员交往圈，采用决策树、回归分析算法构建集团客户识别模型；辅助集团客户信息的日常分析、支撑客户管理、营销管理等业务工作。



案例4：利用协同过滤算法进行手机图书智能营销

基于相似性算法进行用户行为分群模型和内容偏好模型构建；并基于协同过滤算法进行图书推荐；提升客户获取效率和质量。

行为分群模型

根据用户阅读的行为方式，将用户分为深度活跃、付费欣赏、免费欣赏、登陆无欣赏、包月无欣赏5个类别，并进行特征刻画和数据业务关联分析。

图书内容偏好模型

从频度、粘度、费用3个层面来综合分析不同内容偏好客户的阅读次数、PV数、图书订购等主要行为特征，对用户不同内容的偏好程度进行打分评价。

$$sim(x, y) = \frac{\sum_{s \in S_{xy}} R_{xs} R_{ys}}{\sqrt{\sum_{s \in S_{xy}} R_{xs}^2} \sqrt{\sum_{s \in S_{xy}} R_{ys}^2}}$$

相似性算法

$$P_{xp} = \bar{R}_x + \frac{\sum_{i \in I_x} (R_{ip} - \bar{R}_i) sim(x, i)}{\sum_{i \in I_x} sim(x, i)}$$

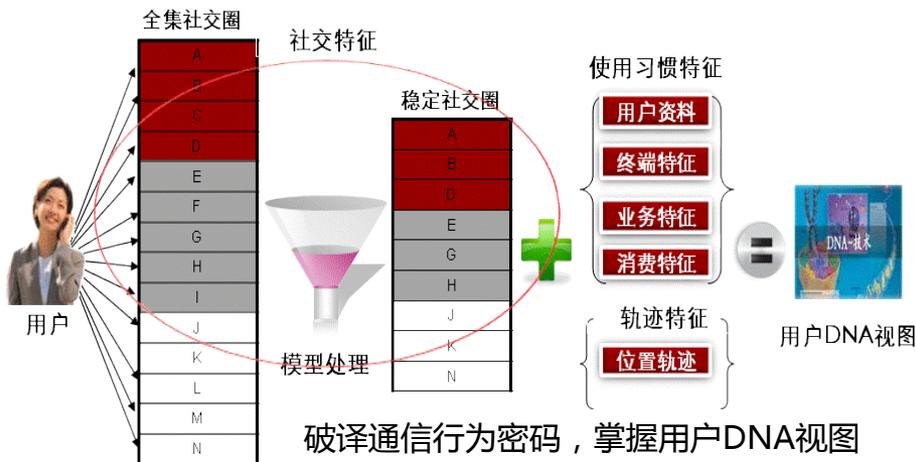
协同过滤算法

图书推荐模型

利用协同过滤推荐技术，根据手机阅读业务之间的关联相似性、用户之间的偏好相似性，预测评估用户对未阅读图书的潜在偏好程度，最终根据偏好程度评分的排序对每个用户进行TOP-N的图书业务推荐。

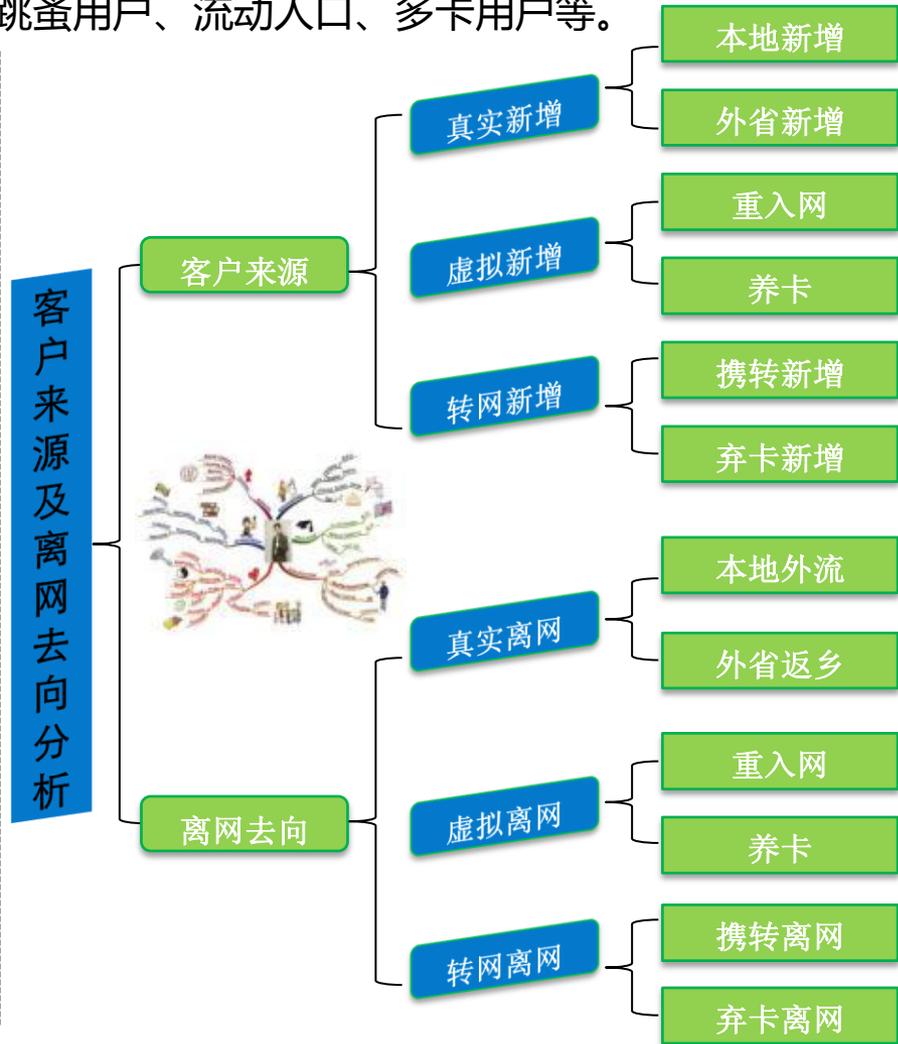
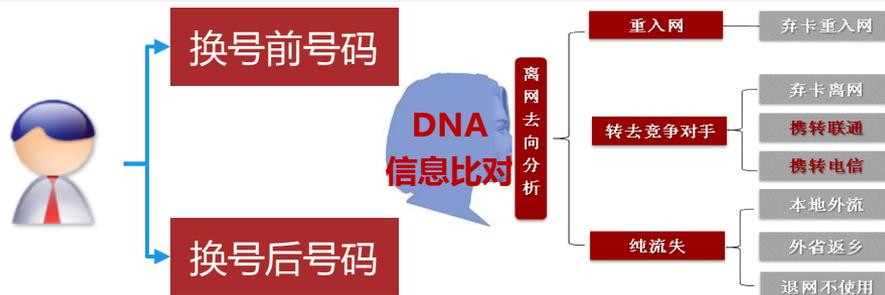
案例5：客户来源及离网去向分析模型

破译通信行为密码，掌握用户DNA视图；达到对用户入网来源、离网去向细分的目的，实现一些特殊目标用户的判断，如重入网用户、养卡用户、跳蚤用户、流动人口、多卡用户等。



破译通信行为密码，掌握用户DNA视图

- 用户DNA定义：从用户全集社交圈中找到核心、稳定的交往圈；结合用户使用习惯特征和位置轨迹特征；形成用户独有的特征链，进而把各类属性的特征链组合在一起，最终形成用户的“DNA”。
- 用户DNA特征：相对稳定性、个体差异性



案例6：增值业务健康度分析模型

□搭建业务健康度评估模型，开展业务健康度管理，从消费健康度、活跃健康度、营销健康度等方面综合评价业务发展面临的风险，提升业务发展的质量。



业务/指标
到达客户普及率
付费客户普及率
客户规模均衡度
保有率
活跃度
活跃客户数同比增幅
新入网客户占使用客户的比重
手机报客户占彩信客户的比重
高级会员数占比
户均使用量
户均流量
业务量同比增幅
户均收入
收入占总收入的比重
收入占数据增值业务收入的比重
收入均衡度
收入同比增幅

- 1、将指标值 进行归一变换，变换后的值 服从标准正态分布
计算公式：
- 2、对指标变换后值 计算其概率密度：
计算公式：
- 3、采用标准分的计算方法，将各指标值标准正态分布概率密度进行线形变换，转换为标准评分

$$score = \min_score \times (1 - f(y)) + \max_score \times f(y)$$

案例7：产品交叉销售



业务问题

每项业务能有多大贡献；在规模指标和业务质量指标上是否健康？哪些内容或者增值业务是可以打包的；哪些不能打包？如何提升用户的活跃度？

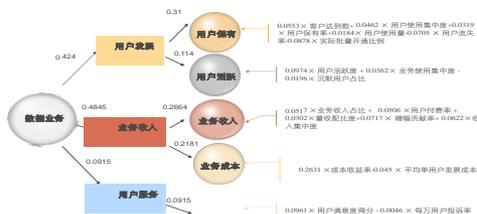
应用价值

主要是为提升业务健康度服务和为交叉销售服务，不健康的业务可及时下线或者调整在客户端上的位置。梳理关联关系，同一类用户喜欢或者用户有可能先购买A再购买B，基于这些关联规则可以向用户打包推荐

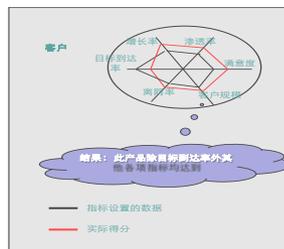
业务评估和关联分析

应用意义

1 结合业务特色梳理关键指标树



2 关键指标综合评分

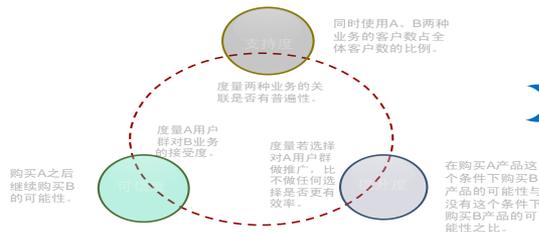


360度信息评估，并对评估结果进行回溯，辅助开展业务优化

1 用户业务订购、使用行为梳理



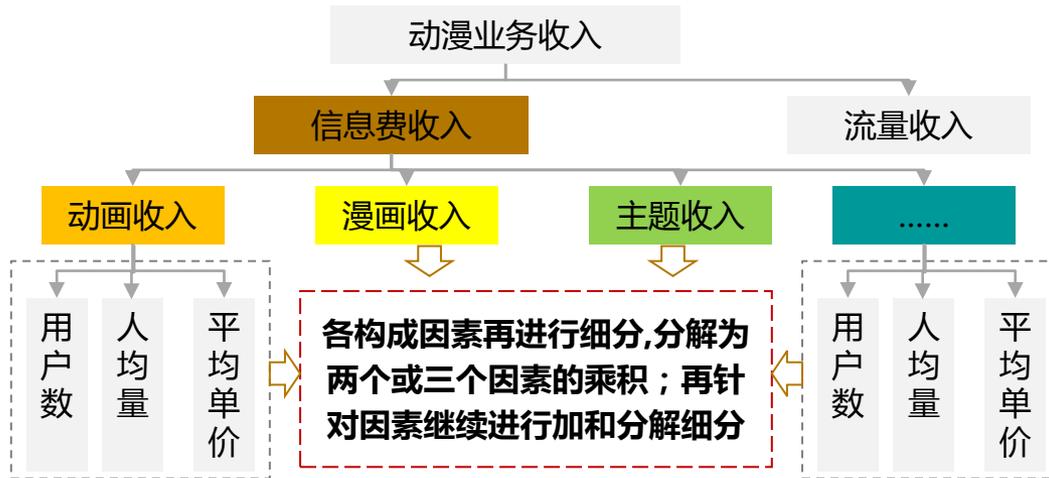
2 强相关业务挖掘



强相关产品进行交叉推荐、捆绑销售，提升用户活跃度和价值

案例8：收入路径及异常预警模型

- 通过对动漫业务收入进行结构化分解，建立其收入分析路径图；使动漫收入结构、路径一目了然
- 基于动漫业务收入路径利用差额法和连环替代分析法构建收入异常预警模型



假设：总收入增幅 θ ，各项收入本期分别为 X_1, Y_1, \dots, Z_1 ，上期分别为 X_0, Y_0, \dots, Z_0 ，上期总收入 M ； A 为实际发生的用户数， B 为人均量， C 为平均单价， A_1, B_1, C_1 为本期， A_0, B_0, C_0 分别代表影响本期、上期动漫业务收入构成的三个因素。

加和结构化影响度分析

$$\theta = \{(X_1 - X_0) + (Y_1 - Y_0) + \dots + (Z_1 - Z_0)\} / M = (\text{本期各费用值} - \text{上期各费用值}) / \text{上期总运营收入} \times 100\%$$

乘积结构化影响度分析：连环替代分析法，各因素顺序按数量在前质量在后（单价属质量）

$$\text{收入变动} \alpha = (A_1 \times B_1 \times C_1 - A_0 \times B_0 \times C_0) = (A_1 \times B_0 \times C_0 - A_0 \times B_0 \times C_0) \text{--- 用户数变化对收入变化的影响} \\ + (A_1 \times B_1 \times C_0 - A_1 \times B_0 \times C_0) \text{--- 人均量变化对收入变化的影响} \\ + (A_1 \times B_1 \times C_1 - A_1 \times B_1 \times C_0) \text{--- 单价变化对收入变化的影响}$$

变动影响度 β ：

- A 的影响度 = $(A_1 \times B_0 \times C_0 - A_0 \times B_0 \times C_0) / \text{上期总运营收入}$
- B 的影响度 = $(A_1 \times B_1 \times C_0 - A_1 \times B_0 \times C_0) / \text{上期总运营收入}$
- C 的影响度 = $(A_1 \times B_1 \times C_1 - A_1 \times B_1 \times C_0) / \text{上期总运营收入}$



THANK YOU!