


# 第1章 数据科学家的武器库



数据科学实战：Python篇

讲师：Ben

# 自我介绍

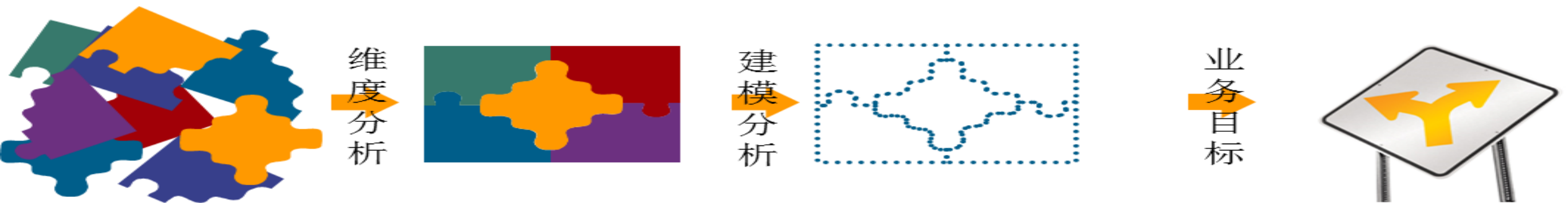
- 天善商业智能和大数据社区      讲师 –Ben
- 天善社区 ID - Ben\_Chang
- <https://www.hellobi.com> – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

- 1、数据科学的基本概念
- 2、数理统计技术
- 3、数据挖掘的技术与方法
- 4、分类模型的评估方法

# 1、数据科学的基本概念

# 数据科学的基本概念

数据科学是一个发现和解释数据中的模式，并用于解决问题的过程



数据

信息

知识

决策和行动

客户编号	交易时间	交易额	交易类型
10001	6/14/2009	58	特价
10001	4/12/2010	69	特价
10001	5/4/2010	81	正常
10001	6/4/2010	60	正常

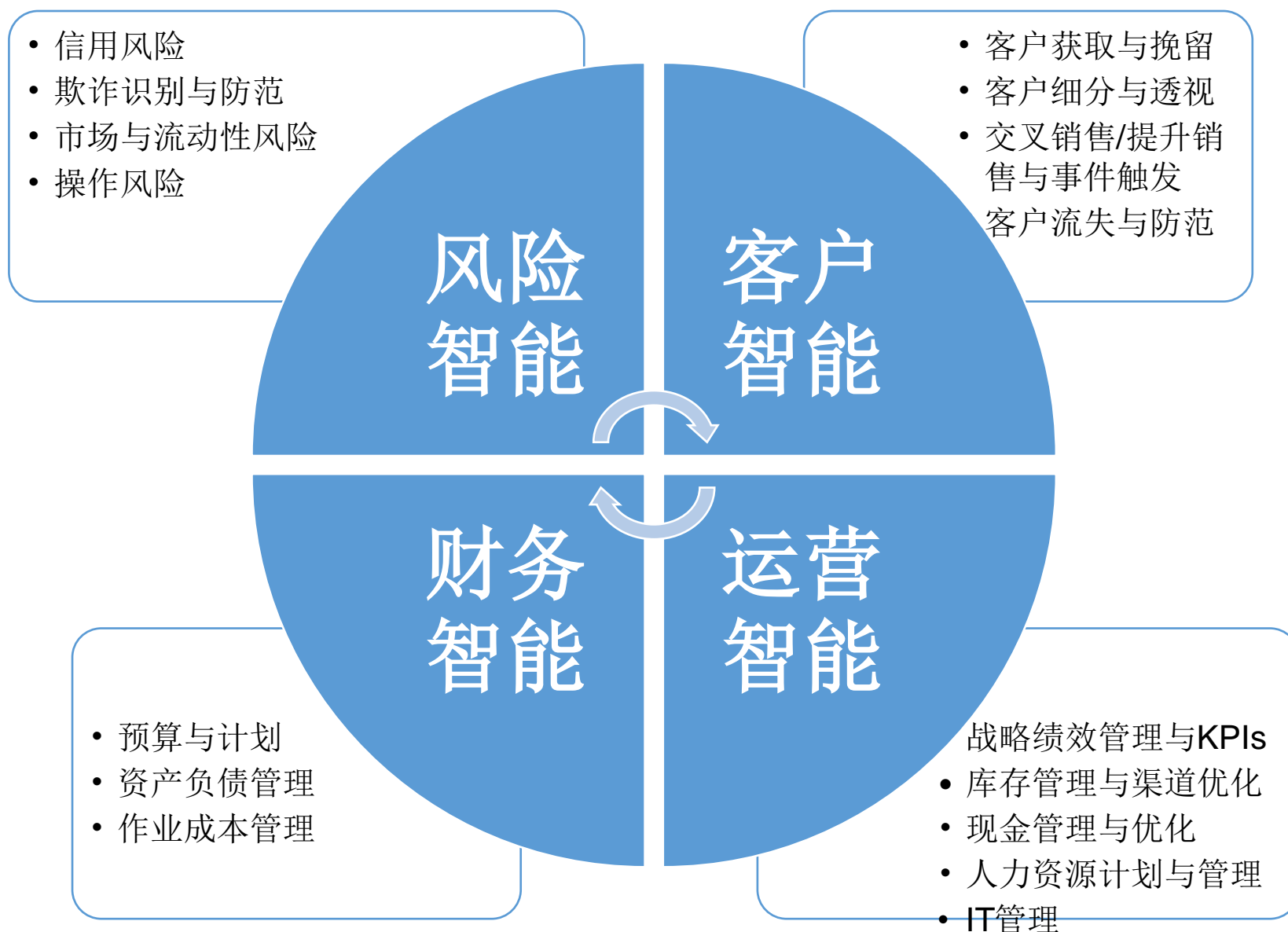
RFM模型

Index	interest	value	time_new
10001	0.118	3.33e+03	1.28e+09
10002	0	1.65e+03	1.28e+09
10003	0.0625	3.44e+03	1.28e+09
10004	0.118	3.31e+03	1.28e+09
10005	0	2.12e+03	1.28e+09
10006	0.0909	1.86e+03	1.28e+09
10007	0.0625	3.96e+03	1.29e+09

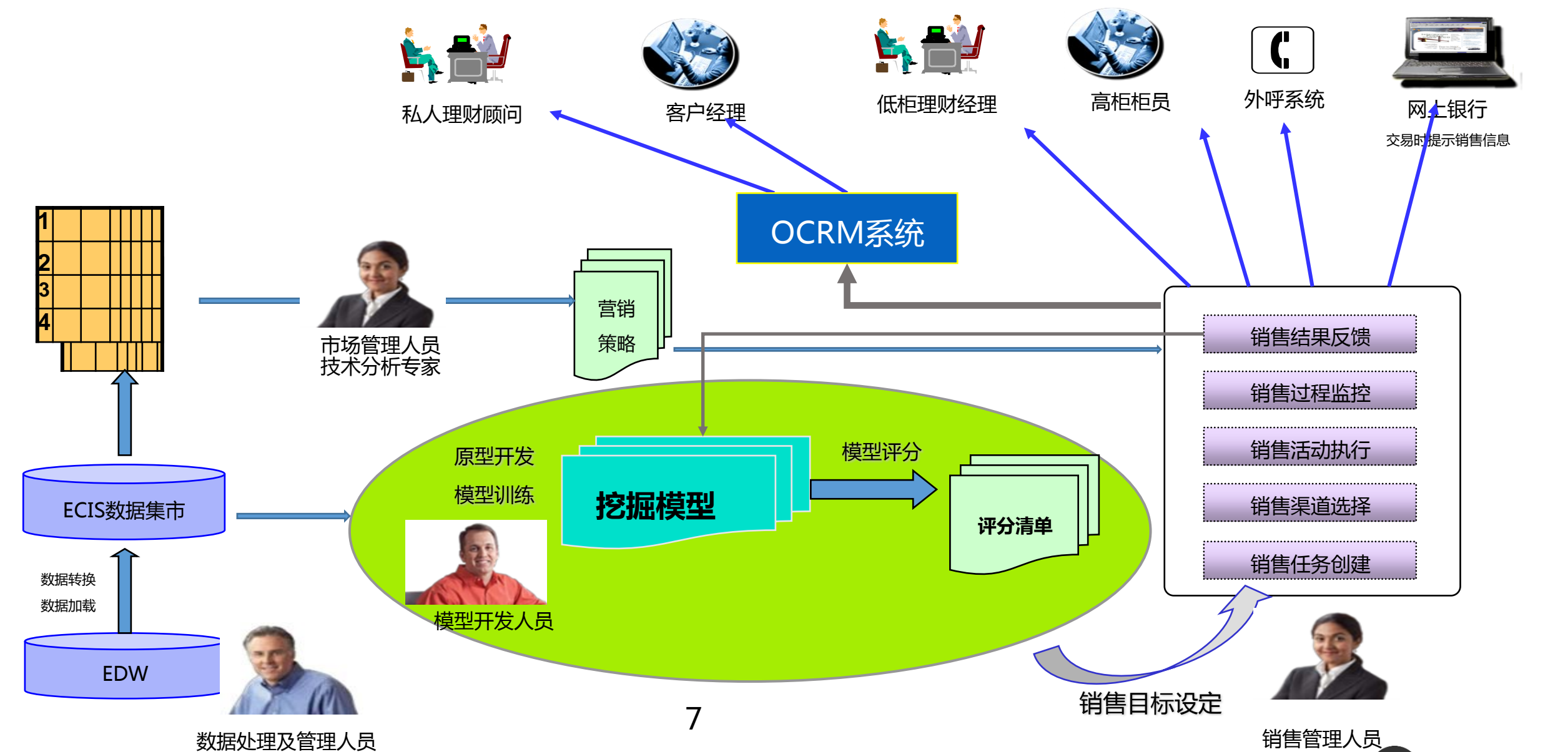
Index	interest	value	time	label
10001	1	1	1	有兴趣-高价值-活跃
10002	0	0	0	无兴趣-低价值-沉默
10003	0	1	0	无兴趣-高价值-沉默
10004	1	1	0	有兴趣-高价值-沉默
10005	0	0	0	无兴趣-低价值-沉默
10006	1	0	0	有兴趣-低价值-沉默
10007	0	1	1	无兴趣-高价值-活跃

客户类型	营销策略
无兴趣-低价值-沉默	不打扰
有兴趣-低价值-沉默	发短信
有兴趣-低价值-活跃	不打扰
无兴趣-低价值-活跃	不打扰
无兴趣-高价值-沉默	发短信
有兴趣-高价值-沉默	电话促销
有兴趣-高价值-活跃	发短信
无兴趣-高价值-活跃	不打扰

# 数据科学的运用场景从未改变



# 数据科学家的角色



# 数据科学家的能力范畴

## 数据管理能力

- 数据管理标准化能力
- 数据整合开发的能力
- 数据获取的多样性
- 业务理解的敏感性

## 统计模型能力

- 数据准备与业务一致性
- 数据探索洞见能力
- 了解开发建模体系
- 了解模型评估与验证能力
- 模型结果与业务可解释性
- 业务理解与模型选择匹配度

基于分析工作角度对员工的三大分类



## 建模分析能力

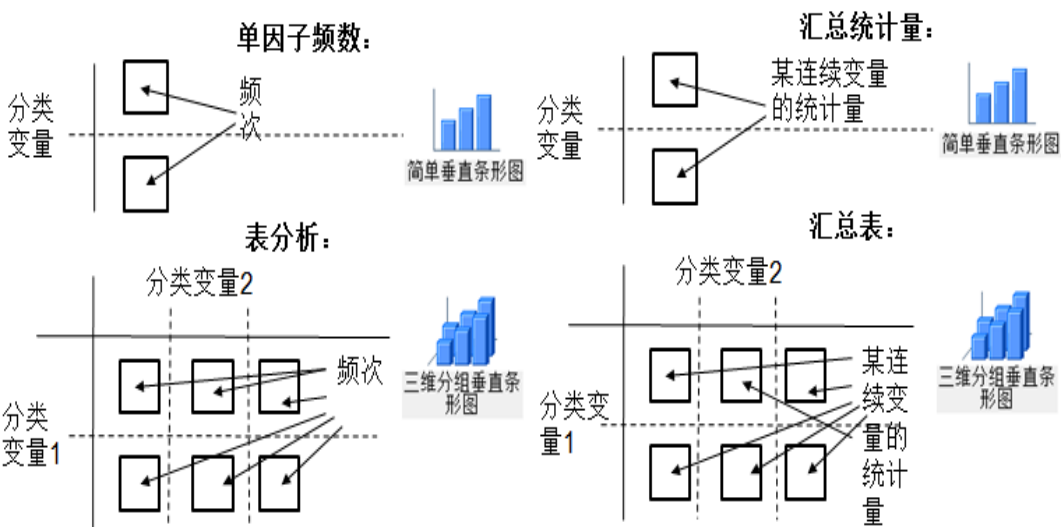
- 领先大数据算法理解能力
- 无监督模型的认知
- 模型与业务结果一致性分析
- 建模步骤条理清晰度
- 熟悉开发建模体系
- 熟悉模型评估与验证能力
- 模型构建业务应用价值



## 2、数理统计技术

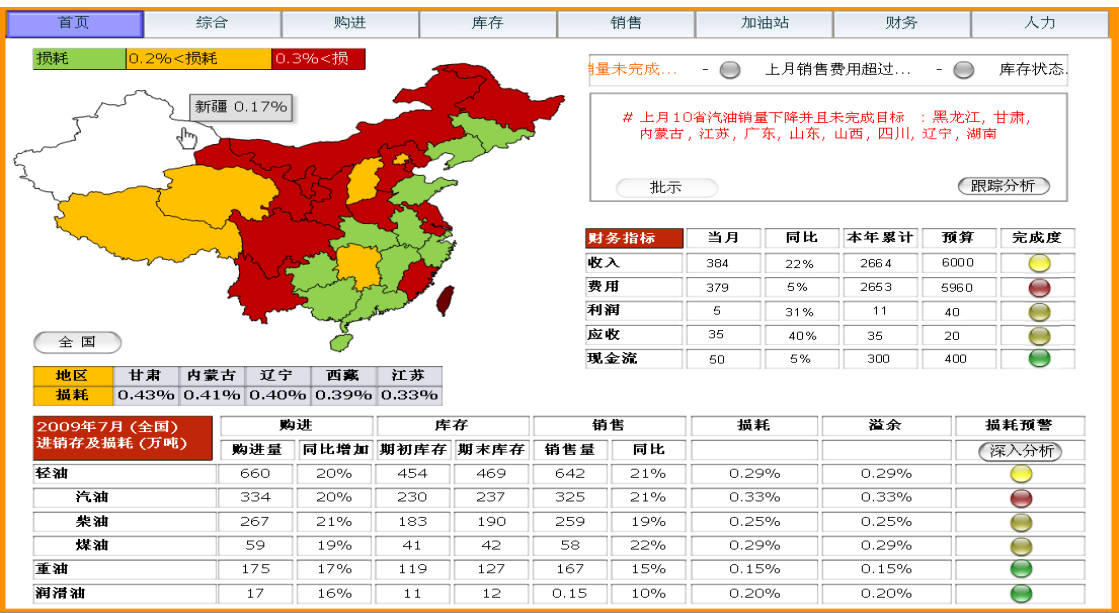
# 数理统计技术

Numpy-数值计 Pandas-数据处理 Matplotlib-统计制图



Statemodels-统计分析

Echat、Hchart-数据可视化



预测变量X \ 被预测变量Y		分类（二分）	连续
		分类（二分）	连续
单个变量	分类（二分）	列联表分析 卡方检验	双样本t检验
	分类（多个分类）	列联表分析 卡方检验	单因素方差分析
	连续	双样本t检验	相关分析
多个变量	分类	逻辑回归	多因素方差分析 线性回归
	连续	逻辑回归	线性回归

# 3、数据挖掘的技术与方法

# 数据挖掘/机器学习

Statemodels-统计分析  
Pyspark ML-分布式机器学习

Sklearn-机器学习  
Pyspark Graphframes-复杂网络

Tensorflow-深度学习

## 预测模型

### 分类模型

排序类  
(评分卡)

Y-主观  
体现权衡

逻辑回归

决策类  
(分类器)

Y-客观  
强调精确

决策树  
神经网络  
组合算法

贝叶斯网络  
KNN  
SVM

### 估计模型 (回归)

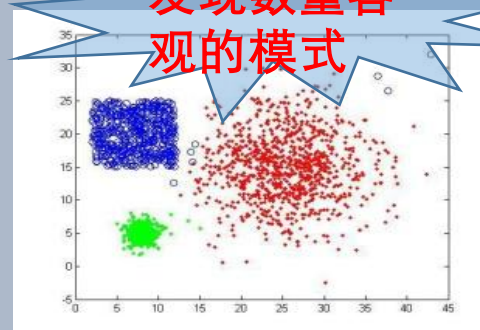
线性回归  
回归树  
神经网络  
...

推荐算法

## 聚类模型

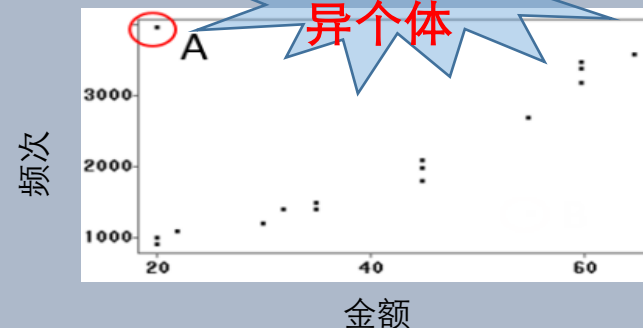
族群发现  
(客户细分)

发现数量客  
观的模式



识别异常

寻找有差  
异个体



复杂网络

时间序列

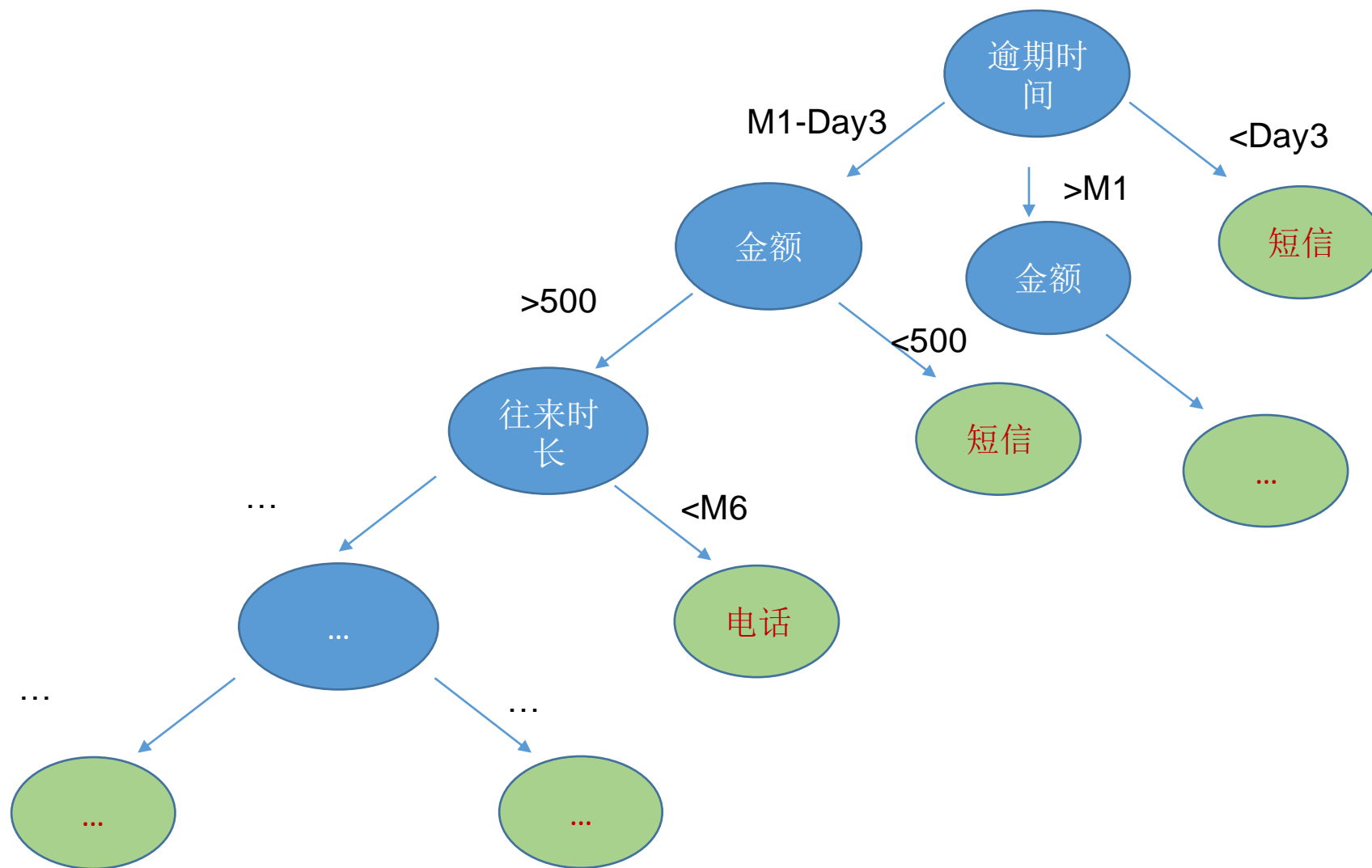
# 分类模型的数据挖掘方法

- **预测性——有监督的学习 ( Supervised Learning )**

- > 机器学习的方法
- > 以历史数据为训练资料，从中学习并建立模型，将此模型运用到当前的数据上，推测未来的结果
- > 训练数据由自变量 (  $X$  ) 和因变量 (  $Y$  ) 组成
  - $Y$ 是连续值，通常称为回归
  - $Y$ 是分类值，通常称为分类
- > 主要算法：
  - 决策树、线性回归、Logistic回归、神经网络、判别分析、...

# 分类模型常用分类算法举例

## 用决策树做催收策略模型



# 分类模型常用分类算法举例

## 规则与贝叶斯网络做进件筛选

规则1: IF 手机IP在国外 THEN 拒绝

规则2: IF 芝麻分小于600 THEN 拒绝

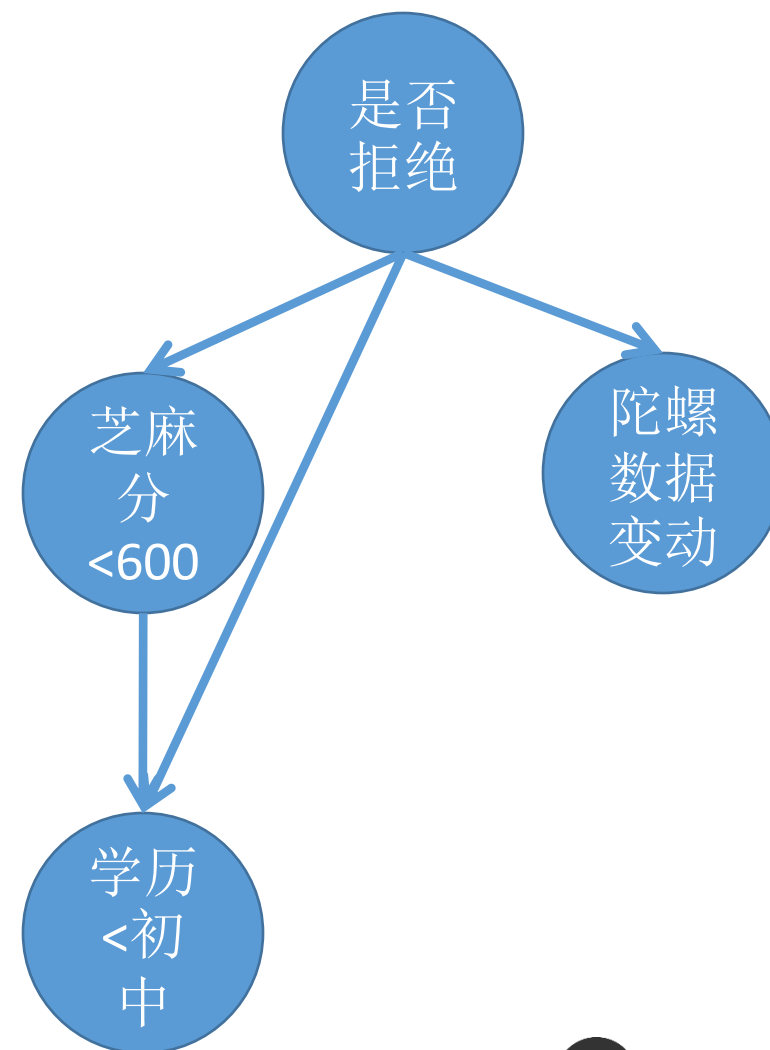
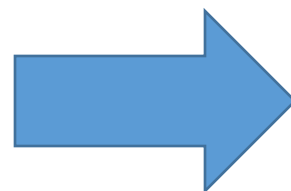
规则3: IF 学历低于初中 THEN 拒绝

规则4: IF 手机的陀螺数据在填写时无变化  
THEN 拒绝

规则5: IF 最近1天, 同一个IP地址申请数量  
超过2个 THEN 拒绝

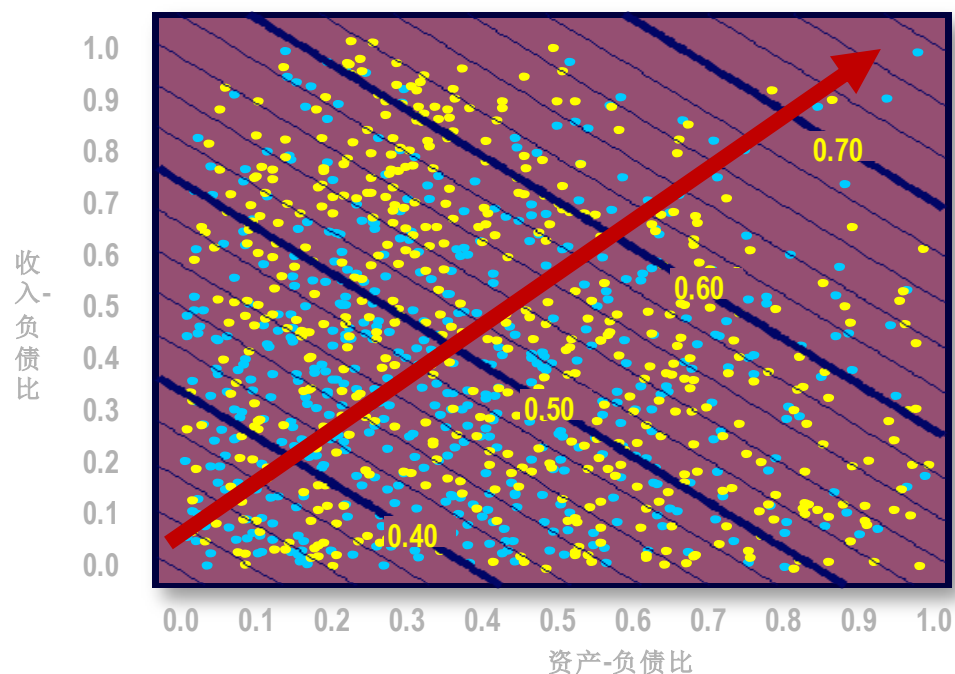
规则...

规则组合



# 分类模型常用分类算法举例

## 用逻辑回归做信用评级



(黄色代表违约的人)

可以根据历史客户的申请数据（作为X）和还款表现（作为Y），寻找到一个合适的公式，作一把尺子（打分），分值高，违约的可能性越高。

逻辑回归是使用最广泛的分类算法，该方法拟合了一条 $P(y=1)$ 的等高线。该值越高，说明Y等于1的可能性越大。

$$\text{logit}(\hat{p}) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

比如公式为 $0.04 \times \text{收入-负债比} + 1.2 \times \text{资产-负债比}$

则收入-负债比为2，资产-负债比为0.3的新申请客户，其违约概率为0.608



# 分类模型常用分类算法举例

## 用神经网络做信用评级的“金模型”或欺诈模型

logit 等式

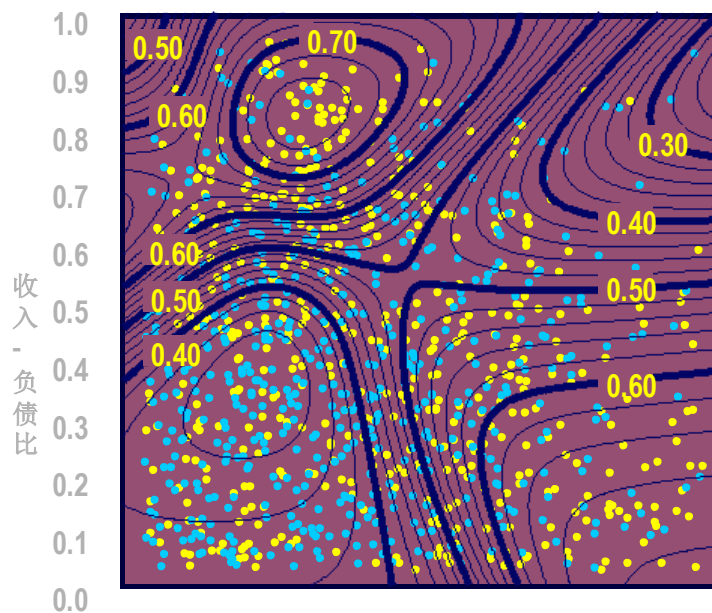
$$\text{logit}(\hat{p}) = -0.5 + -2.6 H_1 + -1.9 H_2 + 0.63 H_3$$

$$H_1 = \tanh(-1.8 + 0.25 x_1 + -1.8 x_2)$$

$$H_2 = \tanh(2.7 + 2.7 x_1 + -5.3 x_2)$$

$$H_3 = \tanh(-5.0 + 8.1 x_1 + 4.3 x_2)$$

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$



资产-负债比

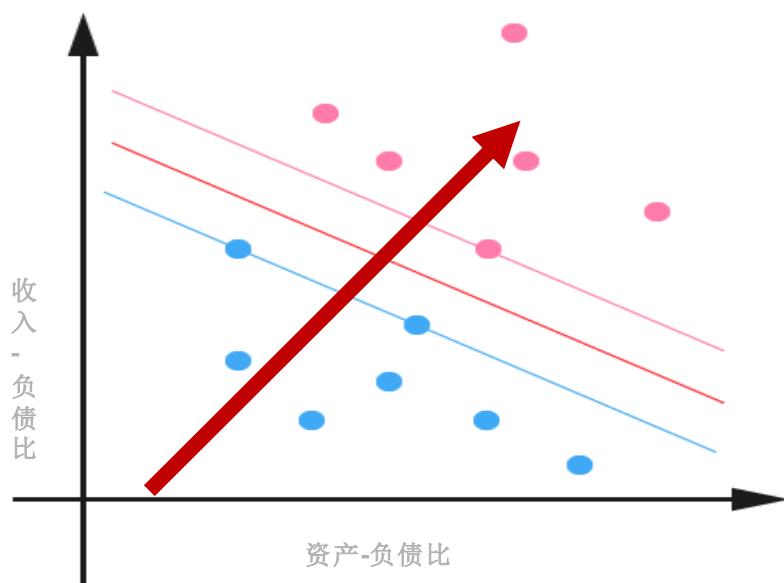
(黄色代表违约的人)

逻辑回归作出的等高线有可能是不精确的。大家都知道收入-负债比和资产-负债比不会是等比换算的。

为了得到这种精确的预测结果，神经网络被发明和运用。该方法省略了部分数据探索的工作，只要将解释因素放入模型，自然的得到解释因素和结果之间复杂的关系。

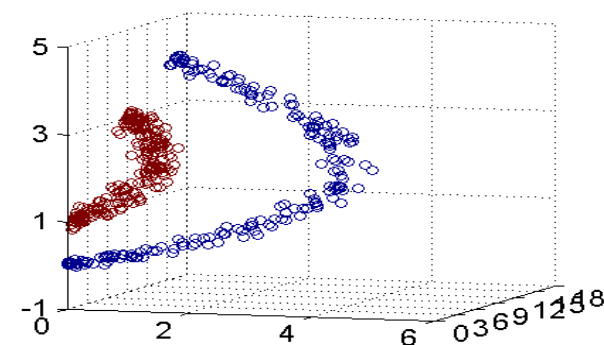
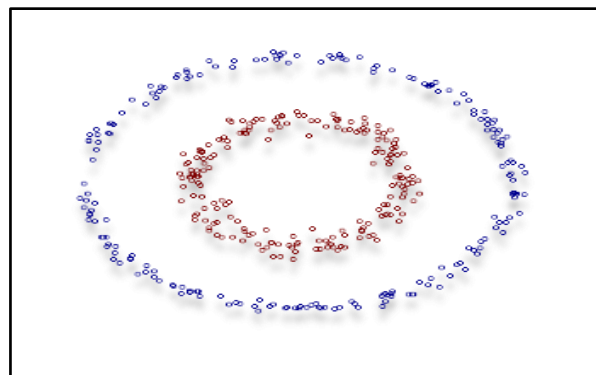
# 分类模型常用分类算法举例

## 使用支持向量机(SVM)做欺诈模型



(粉色代表会违约的人)

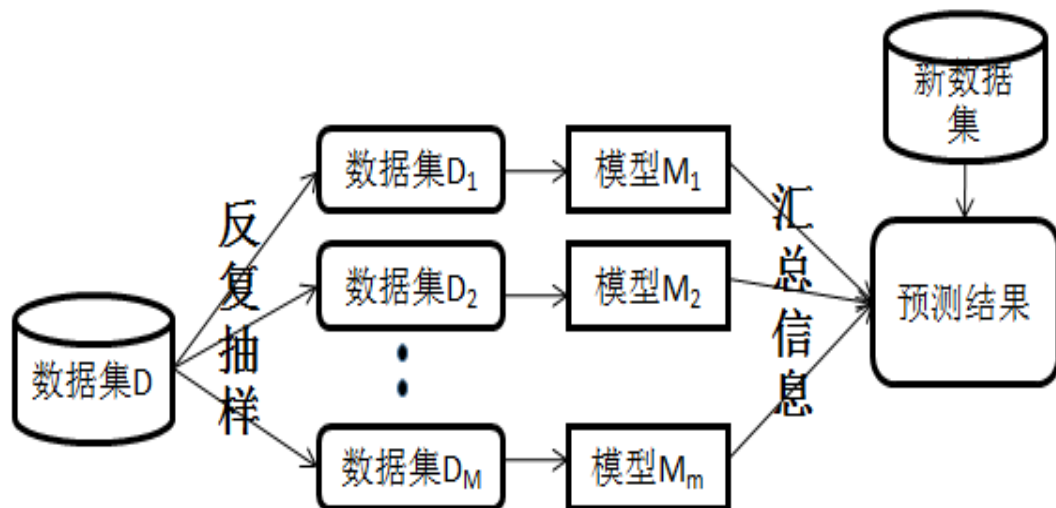
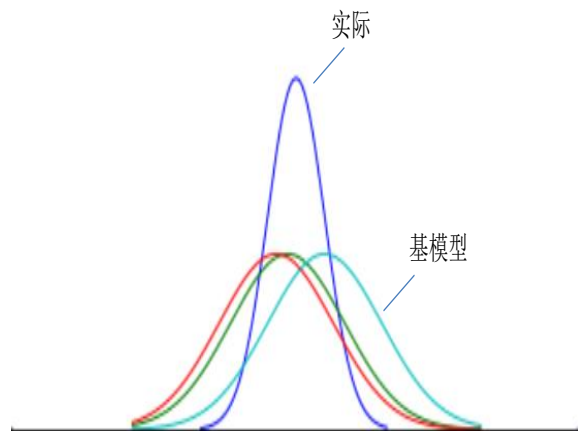
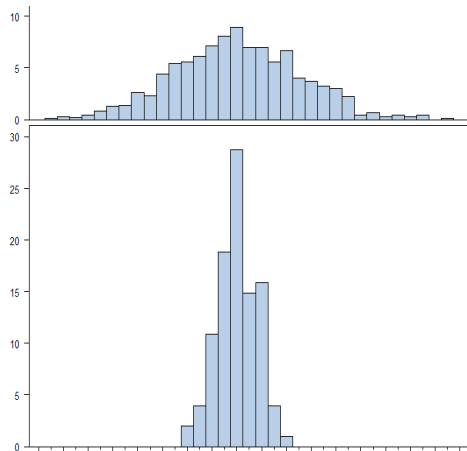
神经网络得到的分隔方式是随机的，不是最优的。SVM方法寻找最优的一个分离平面。有人会觉得左图所示的可以分隔的情况太特殊了。其实这不难，只要将低维度的数据（线性，比如 $x$ ）映射到高维度（非线性，比如 $x^2$ ）就可以了。技术问题早有人解决了，我们直接使用即可。



图片摘自: <http://www.chinakdd.com/article-W82k0g2822JE712.html>

# 分类模型常用分类算法举例

## 使用组合方法做欺诈模型



集成学习概述

装袋法 (Bagging)

-随机森林

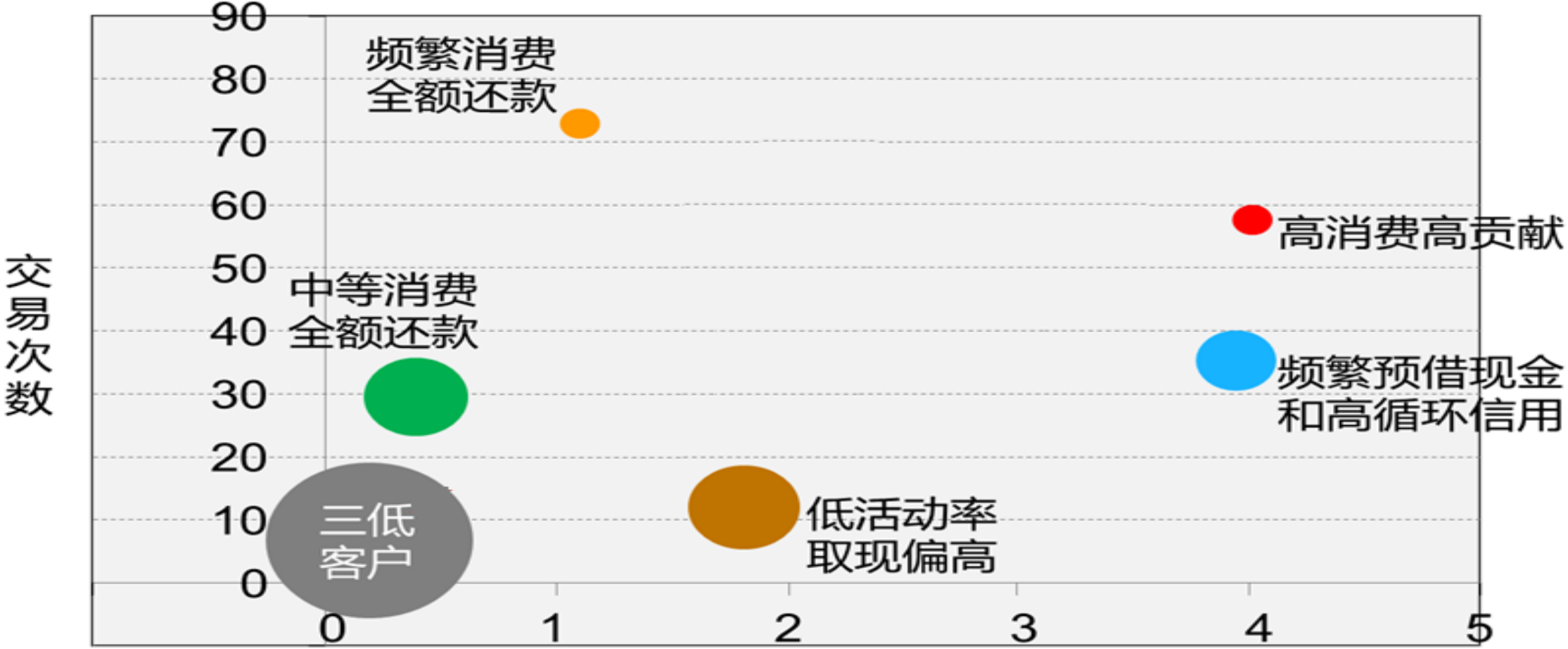
提升 (boosting)

-Adaboost算法

-GBDT和XGBoost

这就是组合方法的思想，该方法被称为预测能力最强，并且最稳健的模型，其原理体现了“兼听则明”的传统观点。该方法不求作出一个大而准的模型，而是通过反复的自抽样，构造不同的分类模型，每个小模型可以都是决策树或神经网络，也可以每个小模型使用的方法都不一样。每个预测样本打分为所有模型预测的均值或众数。

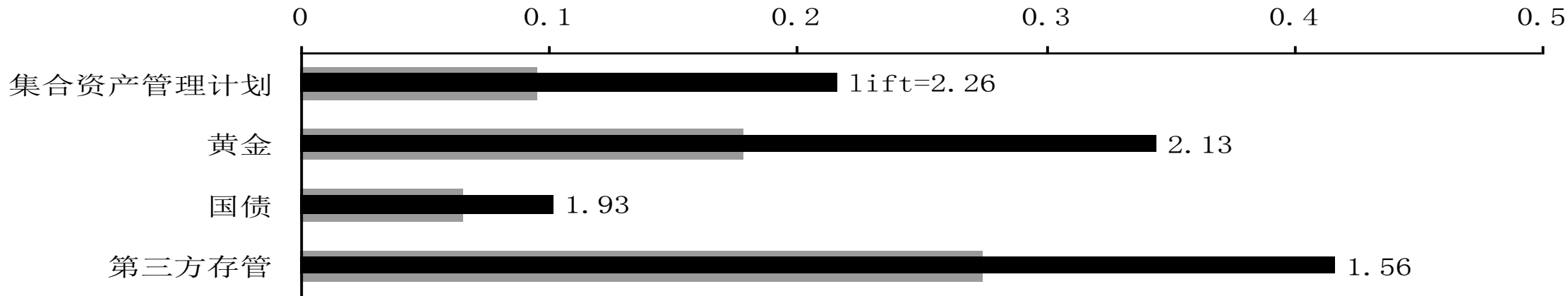
# 聚类模型——客户细分示例



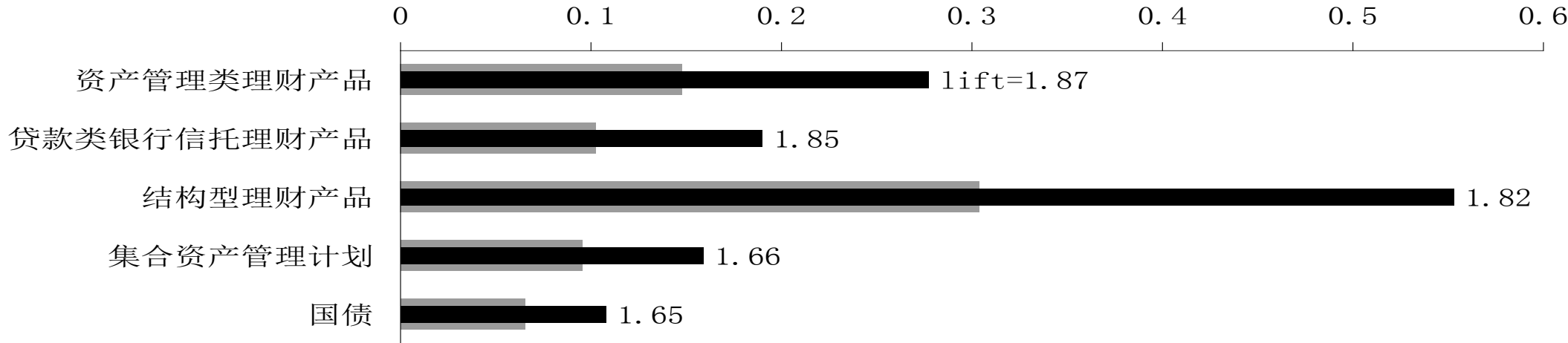
观察窗口：6个月

# 关联规则示例

购买了基金(28%)的客户，还购买下列产品的可能性

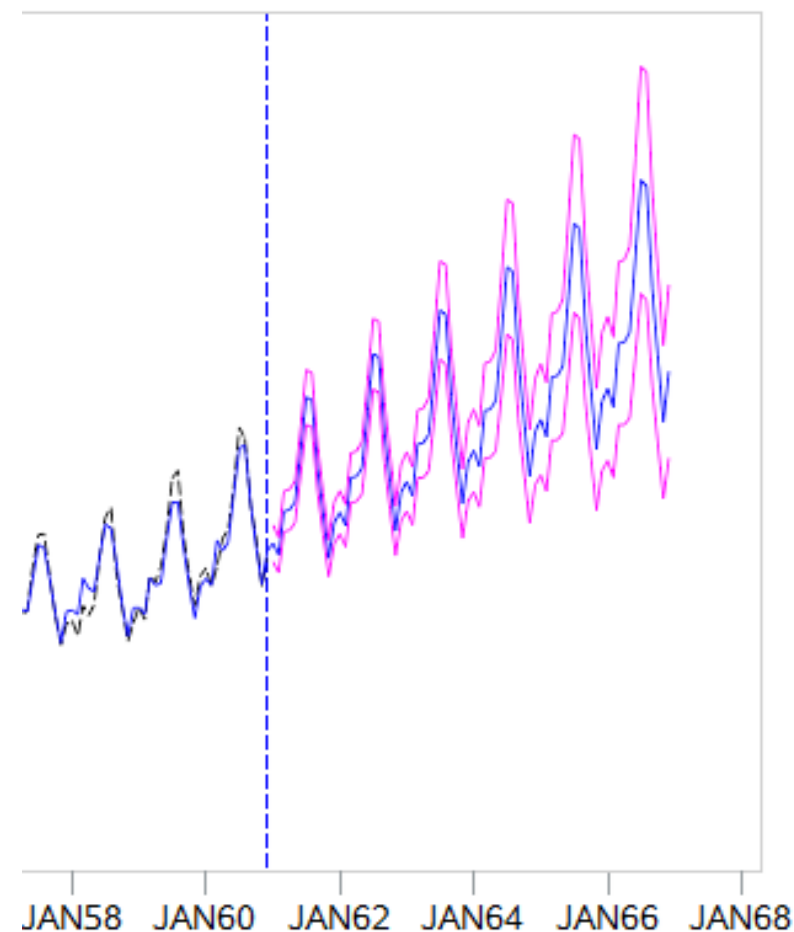
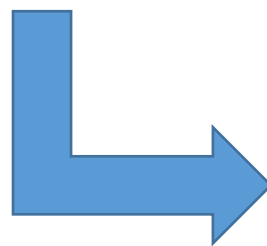
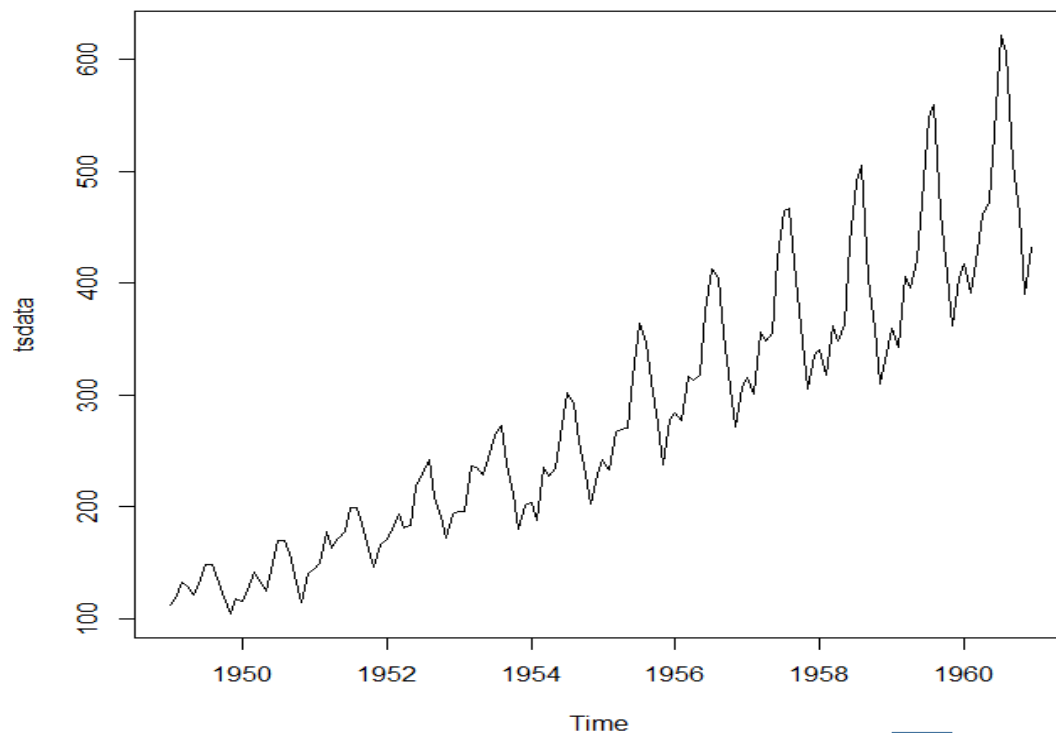


购买了固定收益类理财产品(39%)的客户，还购买下列产品的可能性

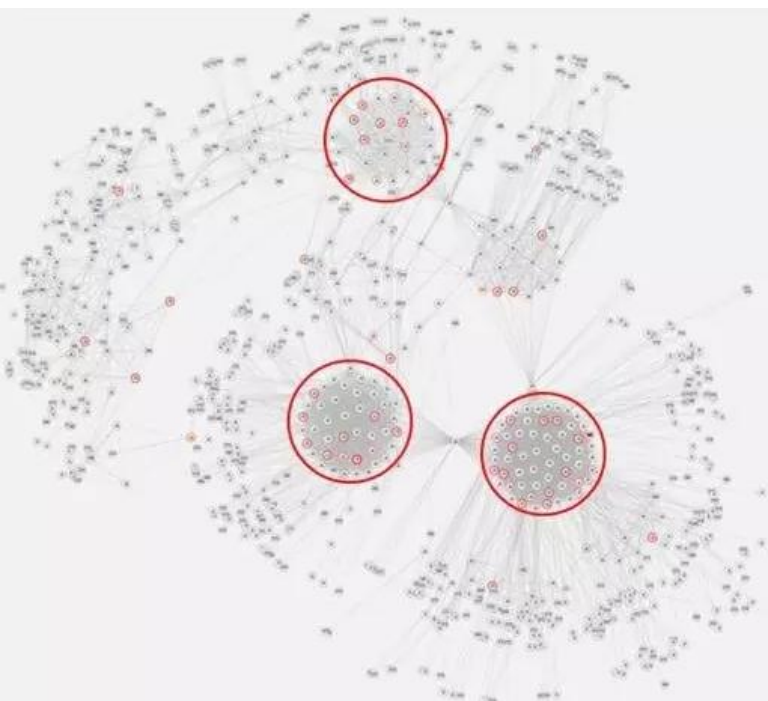


# 时间序列

DATE	AIR
1/1/1949	112
2/1/1949	118
3/1/1949	132
4/1/1949	129
5/1/1949	121
6/1/1949	135
7/1/1949	148
8/1/1949	148
9/1/1949	136
10/1/1949	119
11/1/1949	104
12/1/1949	118
1/1/1950	115
2/1/1950	126
3/1/1950	141

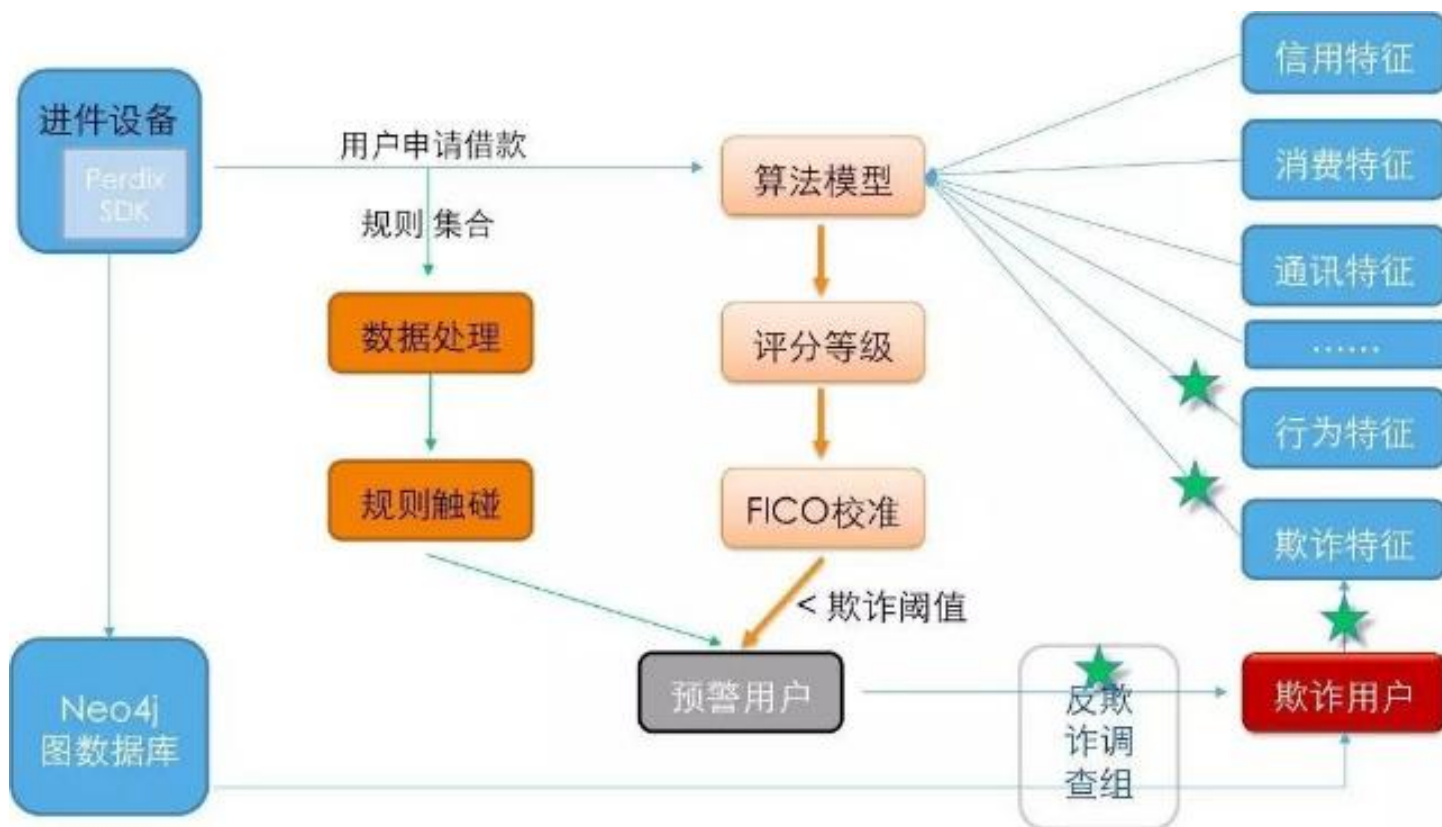


# 复杂网络



众安依照“物以类聚、人以群分”的思路，发现欺诈高发的群体；  
执行“射人射马、擒贼擒王”的方法，排除黑中介。  
腾询构建聊天人群的复杂网络，用于进行传销等事件的侦测。

宜人贷将传统反欺诈技术与知识图谱、复杂网络相融合。



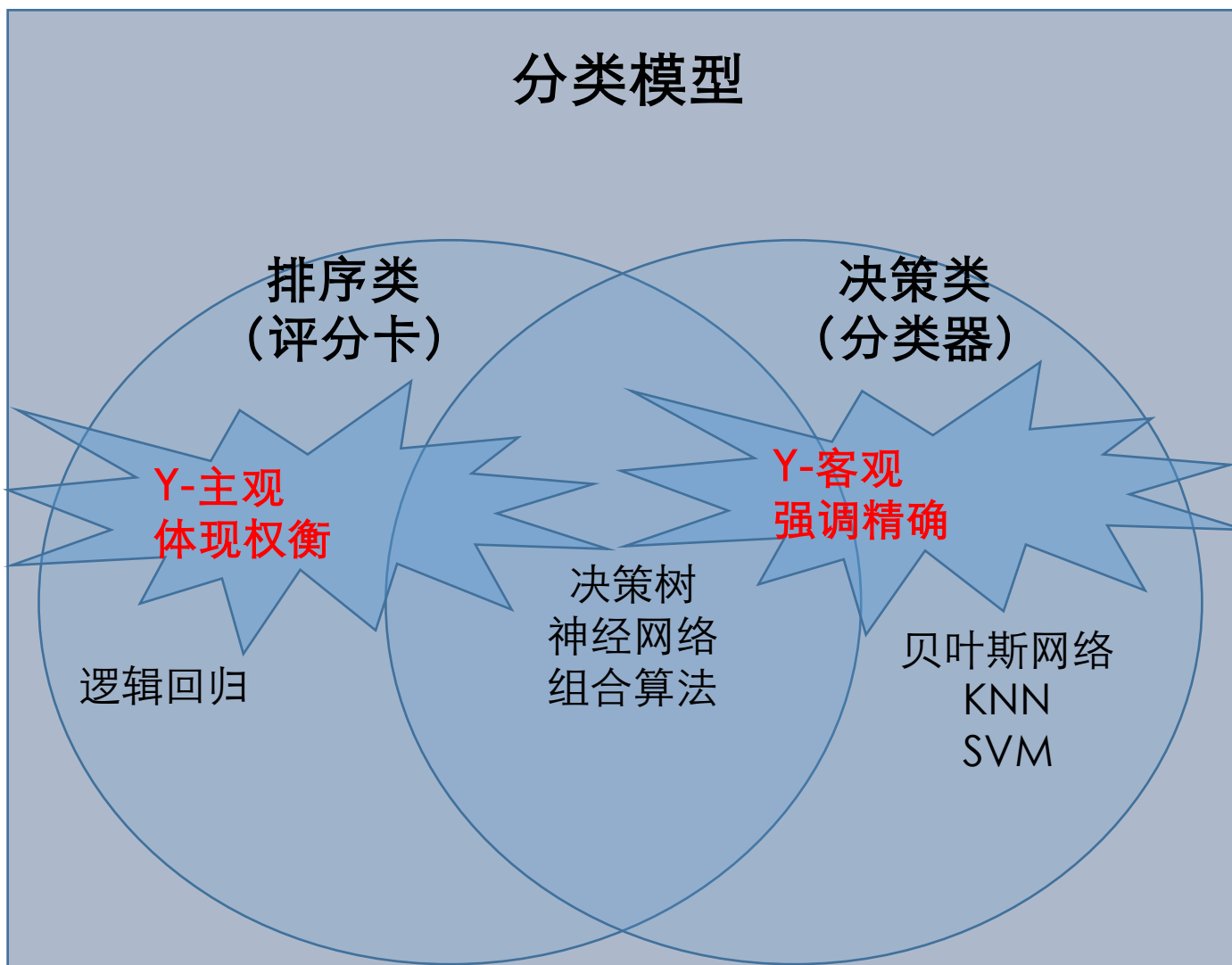
参考: <http://www.docin.com/p-1957163916.html>



## 4、分类模型的评估方法



# 分类模型：排序类和决策类模型辨析

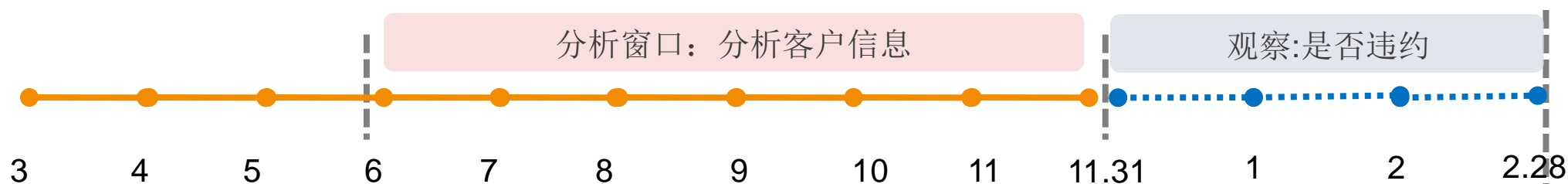


# 分类算法的模型评估

样本内评估：使用训练集同期的数据



样本外评估：使用下一期的滚动数据



## 评估指标汇总

预测类型	统计量
决策 (Decisions)	准确率/误分类 利润/成本
排序 (Rankings)	ROC 指标 (一致性) Gini 指数 K-S统计量 提升度

# 分类算法的模型评估

## 决策类模型评估

该类模型的需求是回答“是不是？”。比如判别持身份证办业务的人是否为证件所有者。

混淆矩阵： 每给定一个阈值，就可以做出一个混淆矩阵		打分值		
		反应（预测=1）	未反应（预测=0）	合计
真实结果	呈现信号（真实=1）	A（击中） True Positive	B（漏报） False Negative	<b>A + B</b>
	未呈现信号（真实=0）	C（虚报） False Positive	D（正确否定） True Negative	<b>C + D</b>
合计		<b>A + C</b>	<b>B + D</b>	<b>A + B + C + D</b>

1. 正确率 =  $(A+D)/(A+B+C+D)$
2. 召回率（覆盖率recall；灵敏度Sensitivity） =  $A/(A+B)$
3. 命中率（精确度 Precision、PV+） =  $A/(A+C)$
4. 特异度 (Specificity；负例的覆盖率)=  $D/(C+D)$
5. 负命中率( PV-) =  $D/(D+B)$

# 分类算法的模型评估

## 评估指标汇总

预测类型	统计量
决策（Decisions）	精确性/误分类 利润/成本
排序（Rankings）	ROC 指标 Gini 指数 K-S统计量 提升度

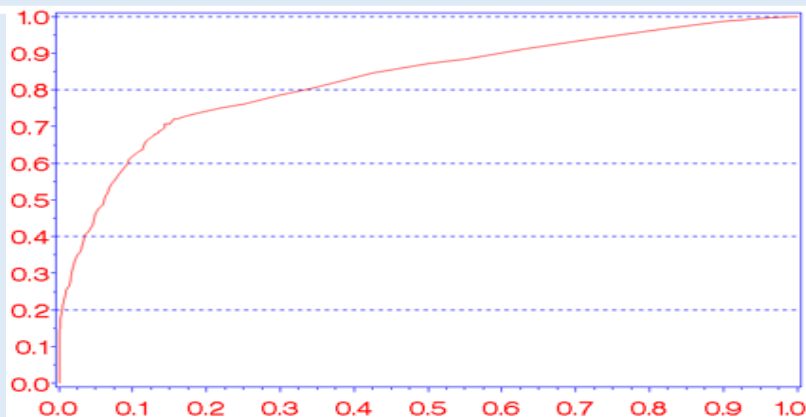
# 分类算法的模型评估

## 排序类模型的评估指标

该类模型的需求是回答“会不会？”。比如预测一下客户违约的概率、营销响应的概率。

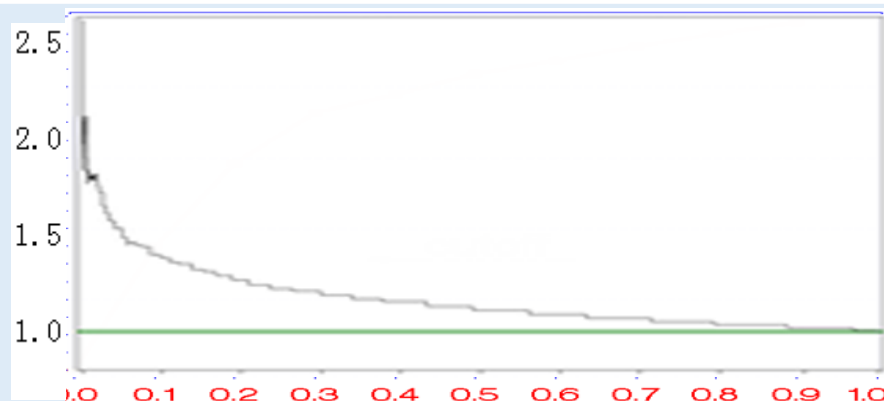
**ROC曲线**：用来描述模型分辨能力,对角线以上的图形越高模型越好

X:特异度  
Y:灵敏度



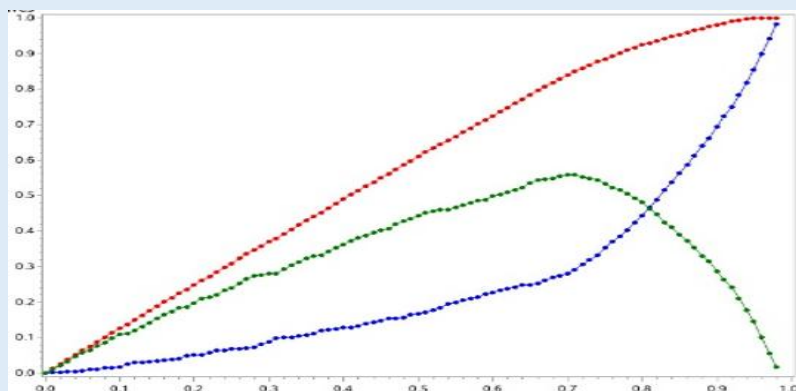
**累积提升曲线**：由于展示使用模型预测结果与随机情况下获取显性样本的能力比较

X:深度  
Y: 正例的  
累积密度  
除以基准  
概率



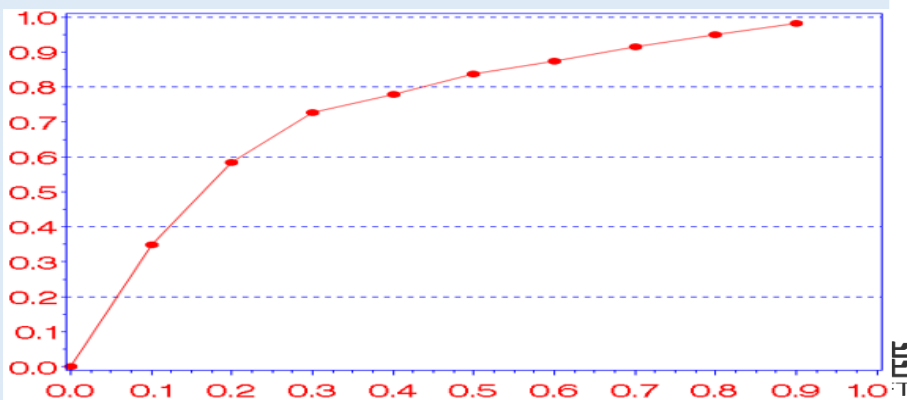
**K-S曲线**：用来描述模型对违约客户的分辨能力

X:深度  
Y红: 正例的  
累积密度  
Y蓝: 负例的  
累积密度  
Y率: K-S值



**洛伦兹曲线**：用来描述预期违约客户的分布

X:深度  
Y: 正例的  
累积密度



—— 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  
数据仓库**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仓库模型的搭建  
• 高质量的数据库模型和技巧，以及丰富的例子  
• 数据仓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**  
—— 循序渐进，案例为王，诠释全面，思路制胜 ——  
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