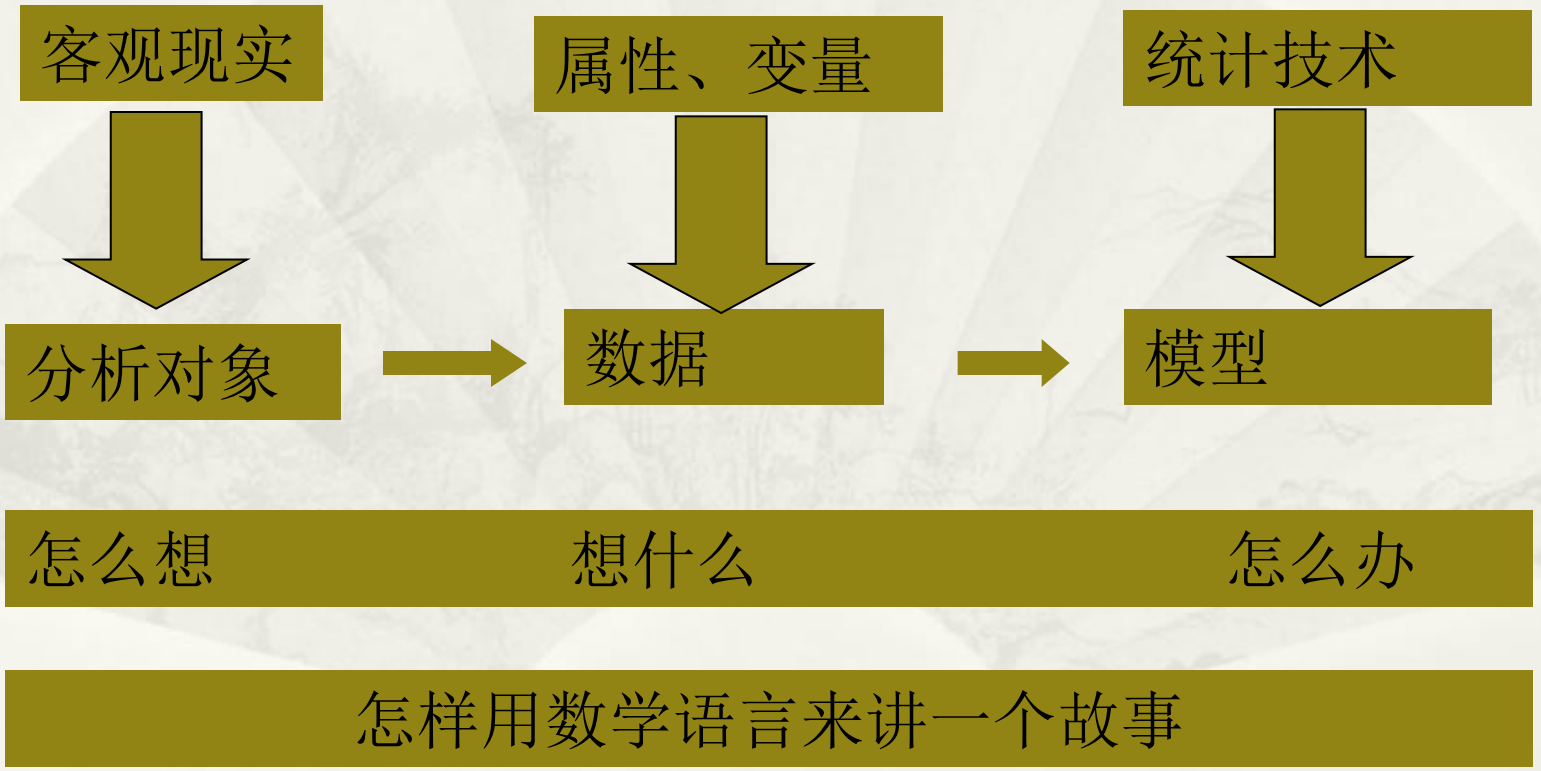
建模流程

翟祥
北京林业大学统计系
zhaixbh@126.com

数据分析（挖掘）逻辑

- 从系统的观点看：系统的记忆性，也就是某一时刻或空间进入系统的输入对系统后继行为的影响，图示如下：

数据分析（挖掘）逻辑



模型分类

* 机理模型

根据数据提供的信息体现其中的关系

* 经验模型

利用数据体现的规律进行预测（predict和forecast）

* 模拟模型

通过模型全面的展现事情发生发展的整个过程

模型分类

验证性模型	有监督训练	预测模型
探索性模型	无监督训练	模式发现

模型分类

预测模型

回归分析

决策树

神经网络

模式发现

聚类分析

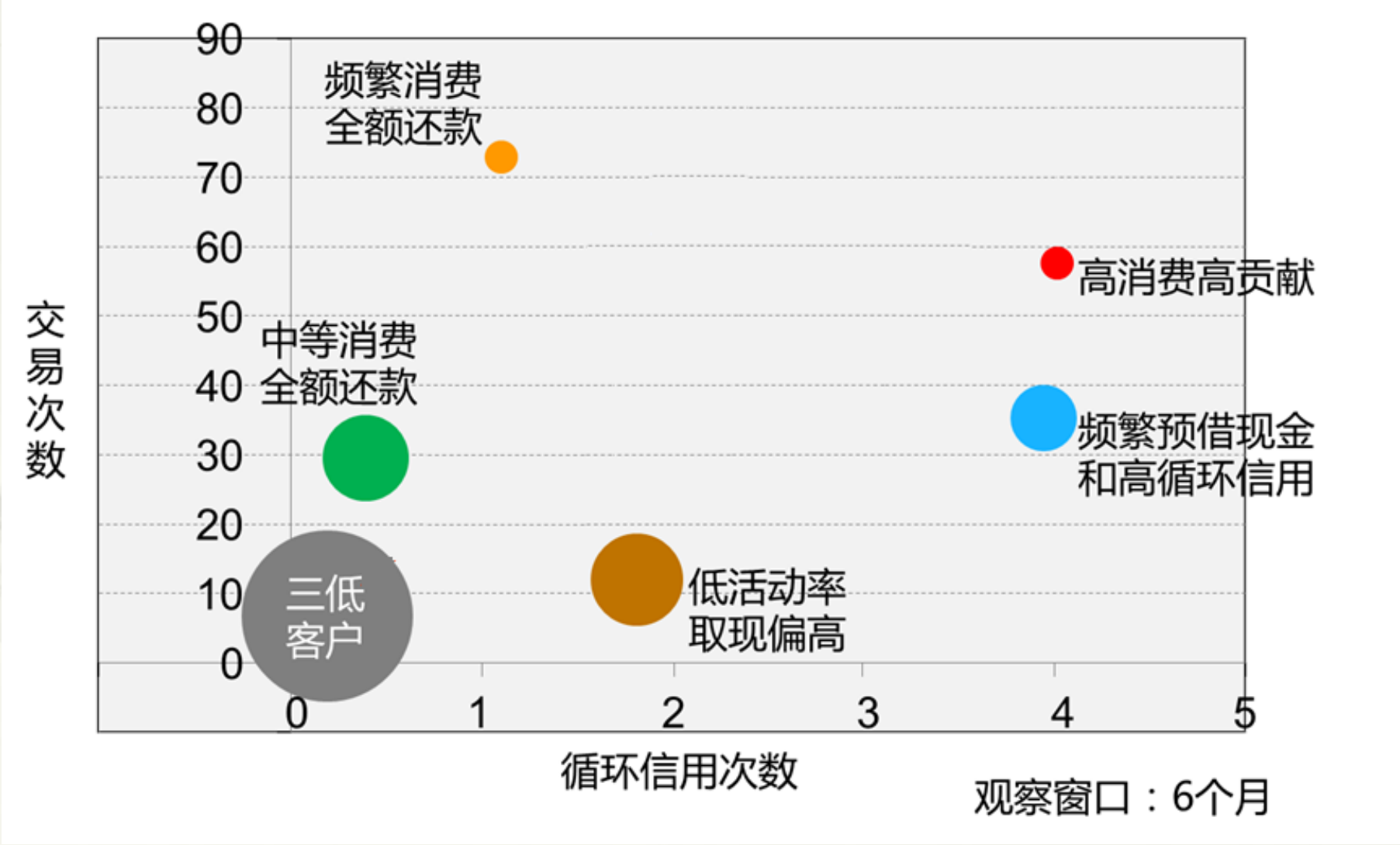
探索性因子分析

关联规则

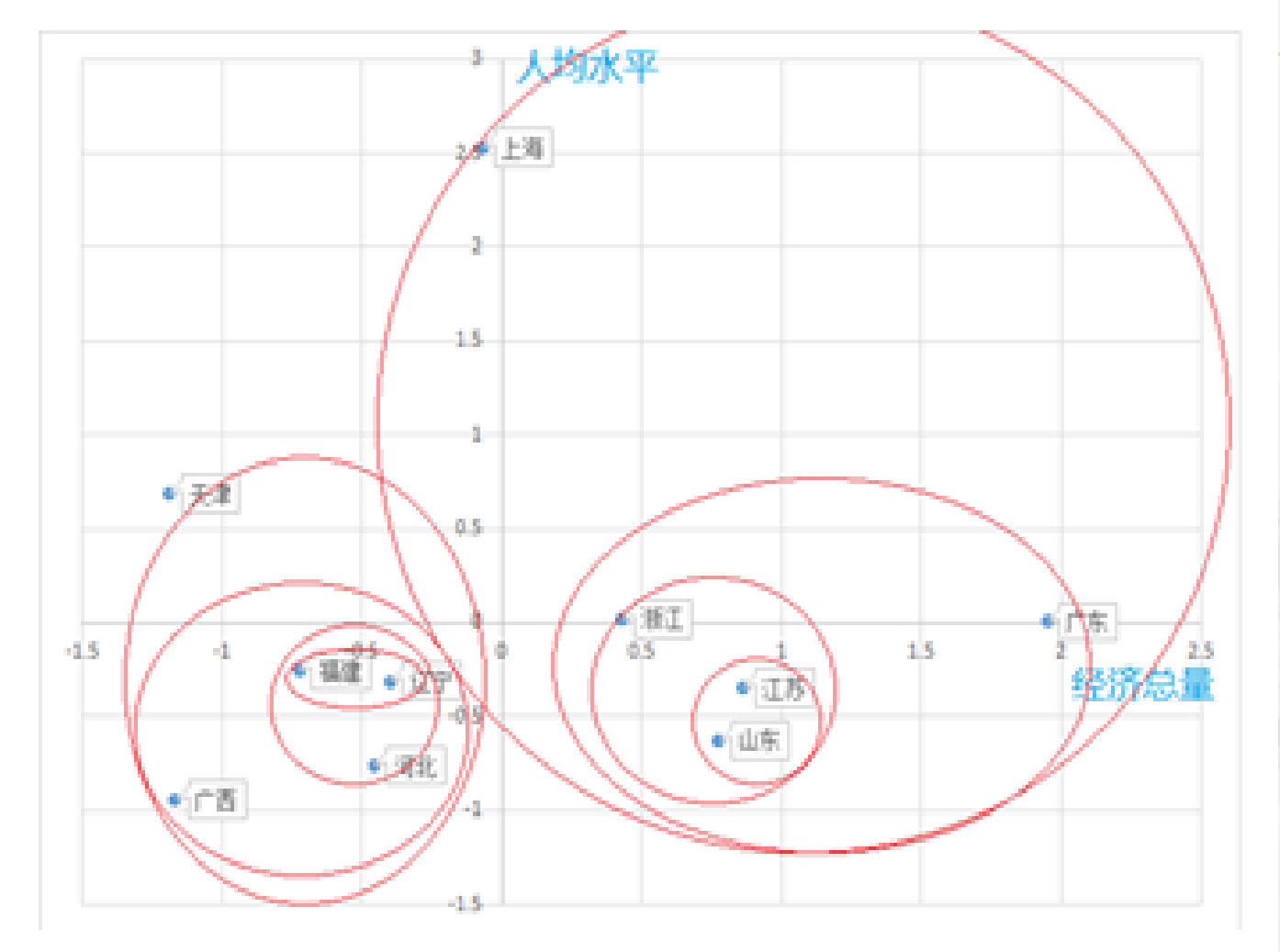
模式发现的数据挖掘方法

- **模式发现——无监督的学习（Unsupervised Learning）**
  - 根据一定规则对数据集进行分类，了解数据集的内部结构
    - 根据**个体之间的相似性**对个体进行分类，例如客户细分
    - 根据**变量之间的相关性**对变量进行分类，例如因子分析
    - 根据客户对商品的购买来发现**商品之间的相关性**，例如关联规则
    - 根据客户之间的联系来发现**社交圈**，例如社交网络分析
  - 有别于监督学习，无监督学习由于没有参照物（即没有监督），不知道分类是否正确
  - 主要算法：
    - 聚类、关联分析、因子分析、主成份分析、社交网络分析、...

数据挖掘方法——客户细分示例

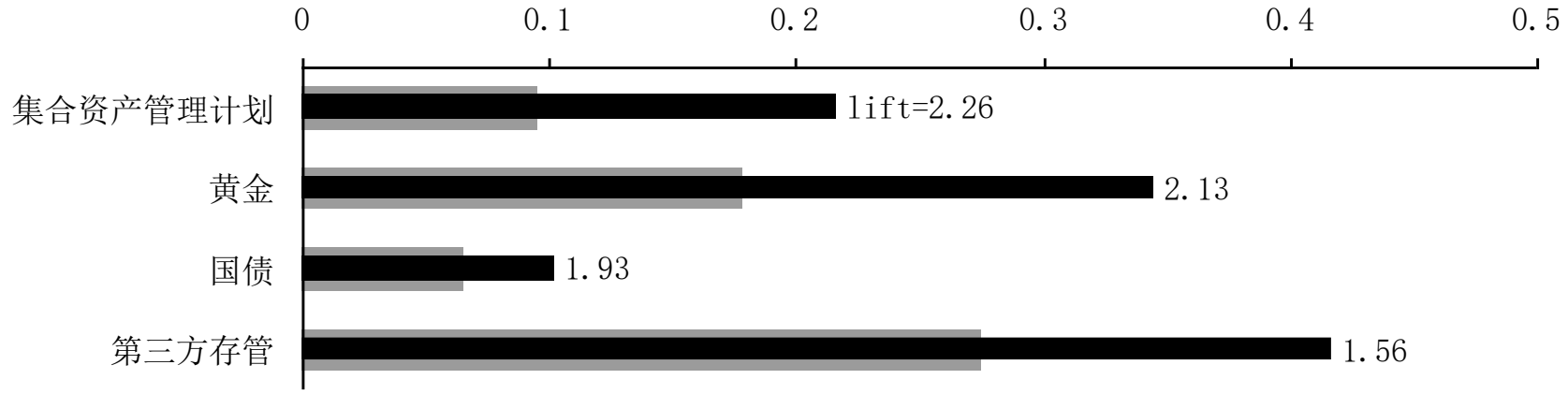


数据挖掘方法——因子分析示例

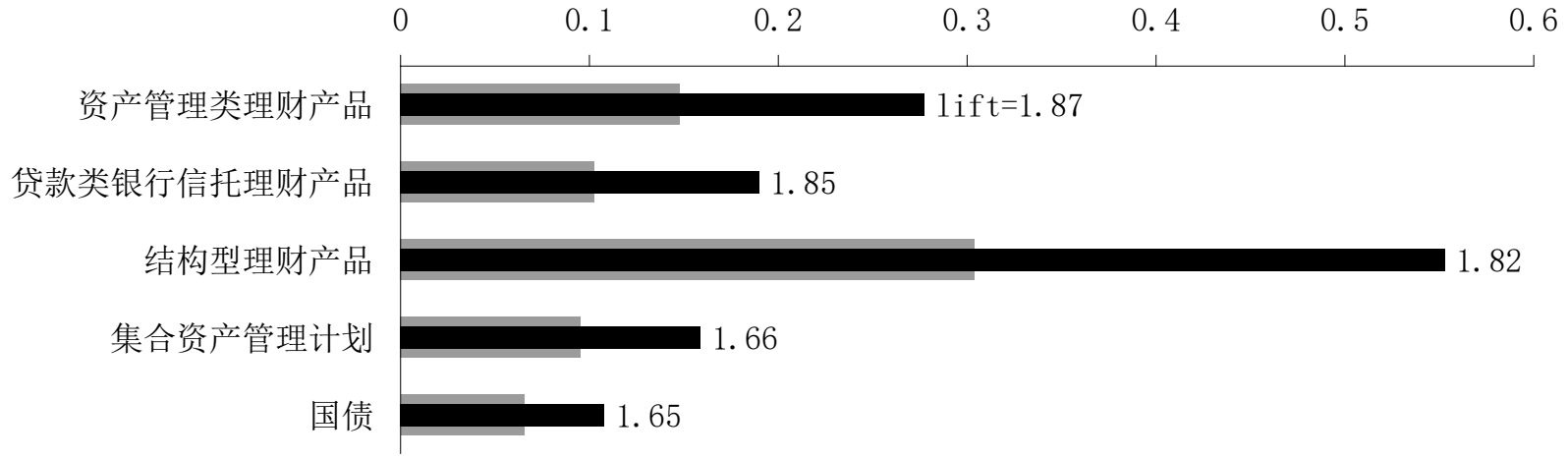


数据挖掘方法——关联规则示例

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



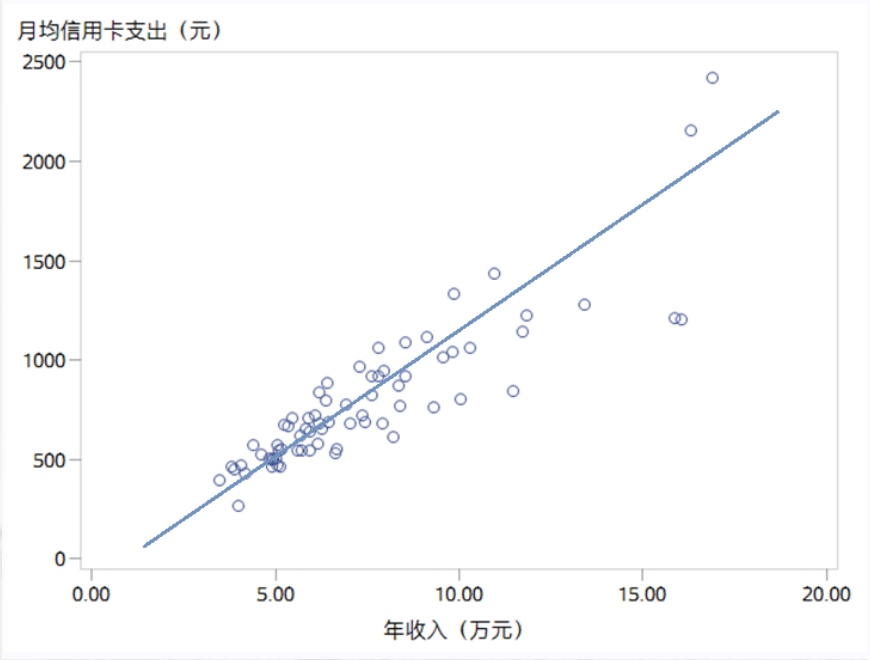
数据挖掘方法——社会网络示例

预测性的数据挖掘方法

- **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、…

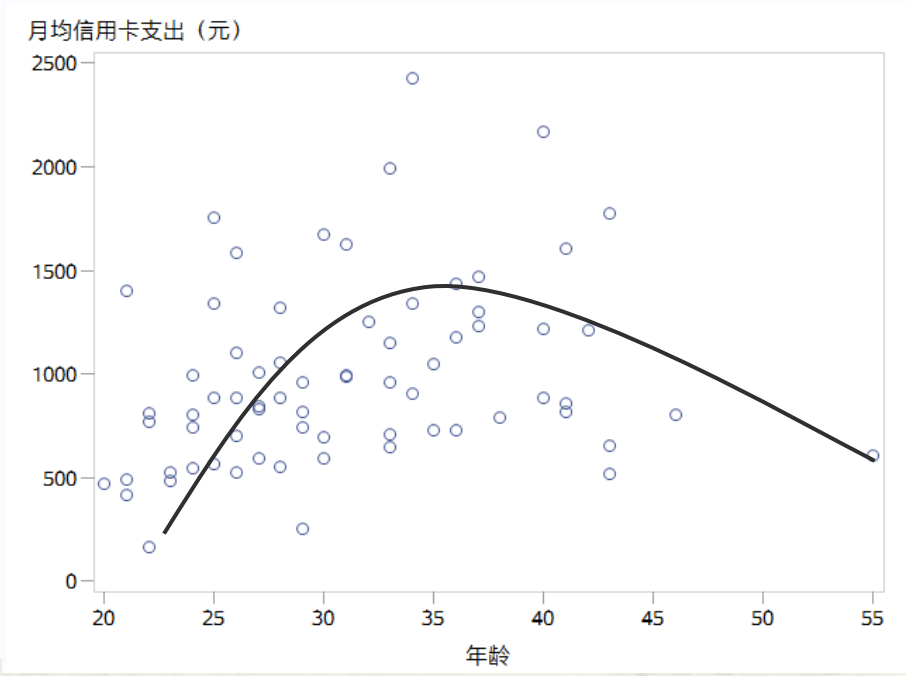
预测性的数据挖掘方法——线性回归示例

被解释和解释变量呈线性关系



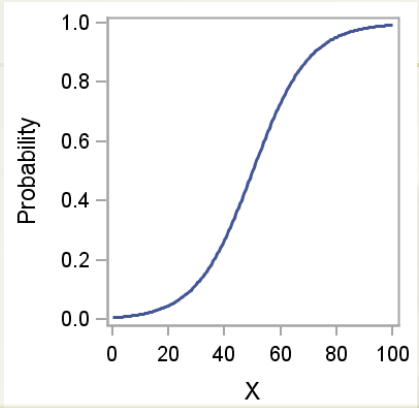
月均信用卡支出=285+98*年收入

被解释和解释变量呈非线性关系

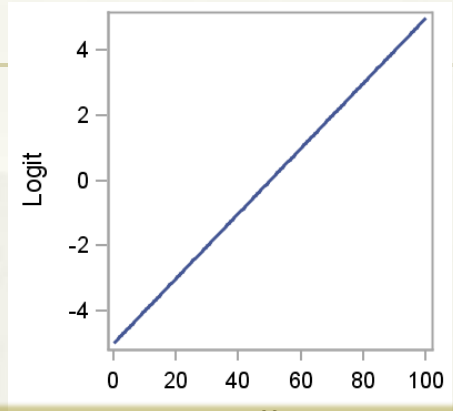


月均信用卡支出
=-1725+154*年龄-2*年龄平方

预测性的数据挖掘方法——逻辑回归示例



Logit转换



模型参数示例

等式用于 手机证券用户（订购）

```
5.433 * LN_USER_ONLINE_ID +
0.1133 * LN_AGE +
0.4784 * LN_CALL_DURATION_M +
-0.31 * LN_FREEFUNC_CALL_COUNTS +
-0.216 * LN_TOLL_CALL_DURATION_M +
-0.1277 * LN_ROAM_CALL_DURATION_M +
-0.04475 * LN_GPRS_FLUX +
0.1155 * LN_MONETMMS_COUNTS +
-0.0002821 * LN_MONETSMS_COUNTS +
-0.2101 * LN_MO_PTPMMS_COUNTS +
-0.1999 * LN_MT_PTPMMS_COUNTS +
-0.362 * LN_MO_SMS_COUNTS +
0.1097 * LN_MT_SMS_COUNTS +
2.698 * LN_WAP_SUB_COUNTS +
1.016 * LN_WAP_USE_COUNTS +
0.1452 * LN_SHOULD_FEE +
-0.1961 * LN_NEWBUSI_FEE +
-0.5202 * LN_TYPE_COUNTS +
-0.5046 * LN_AVG_CALL_DURATION_M +
0.3681 * r_busi_hours_call +
0.6235 * r_newbusi +
0.1221 * mms_use_mark +
-0.4655 * [BRAND_ID=1] +
```

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

模型结果示例

个人客户的ID	单个概率：churn=1
74961005	0.9994583966
74929842	0.9994397695
74808121	0.9991592587
74919003	0.9984661971
74855032	0.9984524769
74842778	0.9983242248
74841623	0.9981002856
74861472	0.9980441824
73417274	0.9979985216
74158095	0.9979708375
74115753	0.9974554829
74816707	0.9973049003
74927592	0.996768544
74966972	0.9963245117
74236709	0.9963224259
74932993	0.9961538611

数据分类（review）

测定层次	特征	运算功能	举例
1、定类测定	分类	计数	产业分类
2、定序测定	分类；排序	计数；排序	企业等级
3、定距测定	分类；排序；有基本测量单位	计数；排序；加减	产品质量差异
4、定比测定	分类；排序；有基本测量单位；有绝对零点	计数；排序；加减乘除	商品销售额

数据分类

* 观测数据
  * 通过调查或观测收集到的数据
* 实验数据
  * 在实验中控制实验对象而收集到的数据
* 截面数据
  * 在相同或者相近的时间点上收集的数据
* 时间序列数据
  * 在不同的时间上收集到的数据

数据和时间的关系

(1)中国1993年—1998年的GDP增长率 （%）

1993	1994	1995	1996	1997	1998
14.2	13.5	10.5	9.6	8.8	7.8

（2）横截面数据。一个或多个变量在同一时点上收集的数据。

1992年实际GDP增长

国家/地区	加拿大	智利	墨西哥	秘鲁	美国	中国	香港	日本
GDP	0.9	12.3	3.6	-1.7	2.7	14.2	6.3	1

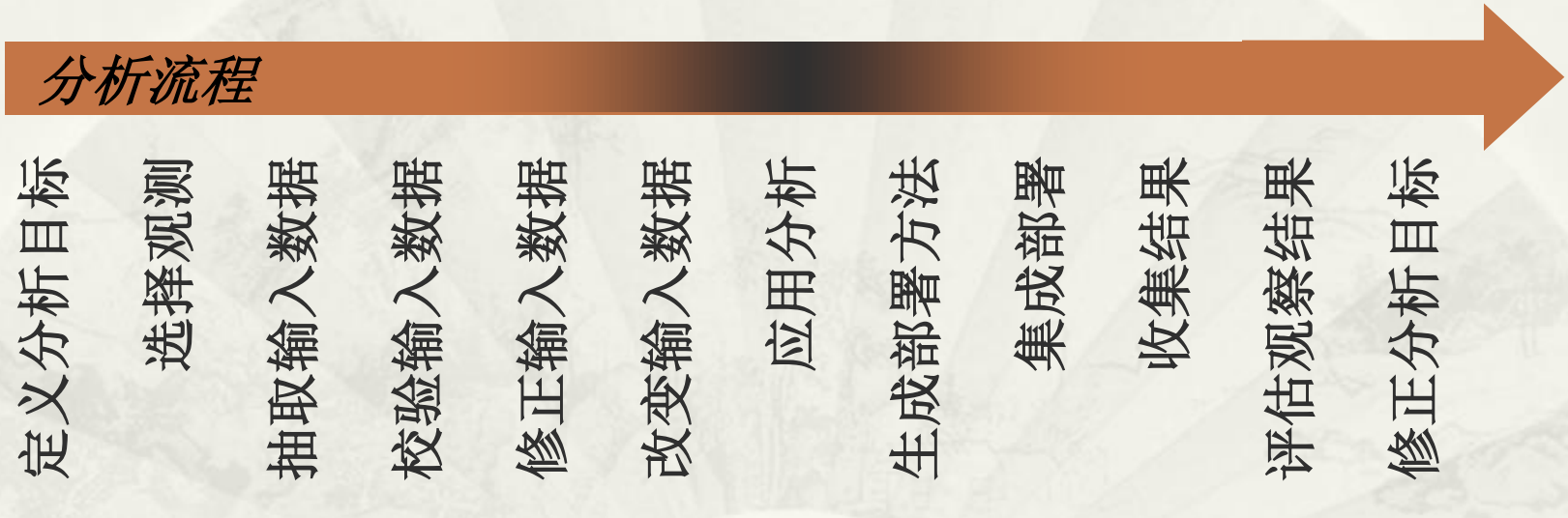
（3）混合数据

国家和地区	实际GDP增长率						
	1992年	1993年	1994年	1995年	1996年	1997年	1998年
加拿大	0.9	2.5	3.9	2.2	1.2	4.0	3.1
智利	12.3	7.0	5.7	10.6	7.4	7.1	3.4
墨西哥	3.6	2.0	4.4	-6.2	5.2	7.0	4.8
秘鲁	-1.7	6.4	13.1	7.4	2.5	6.9	0.3
美国	2.7	2.3	3.5	2.0	2.8	3.9	3.9
中国	14.2	13.5	12.6	10.5	9.6	8.8	7.8
香港	6.3	6.1	5.4	3.9	4.6	5.3	-5.1
日本	1.0	0.3	0.6	1.5	3.9	1.4	-2.8

数据分析（挖掘）支撑点

* 模型可解释性
* 模型和技术是否对数据有严格的假定
* 能否有效的抵御维度“诅咒”
* 能否稳健的应对异常值
* 数据测量层次的难度（定性数据问题）
* 缺失值是否需要事先处理
* 计算的复杂性

分析流程（SEMMA）


分析流程
定义分析目标
选择观测
抽取输入数据
校验输入数据
修正输入数据
改变输入数据
应用分析
生成部署方法
集成部署
收集结果
评估观察结果
修正分析目标

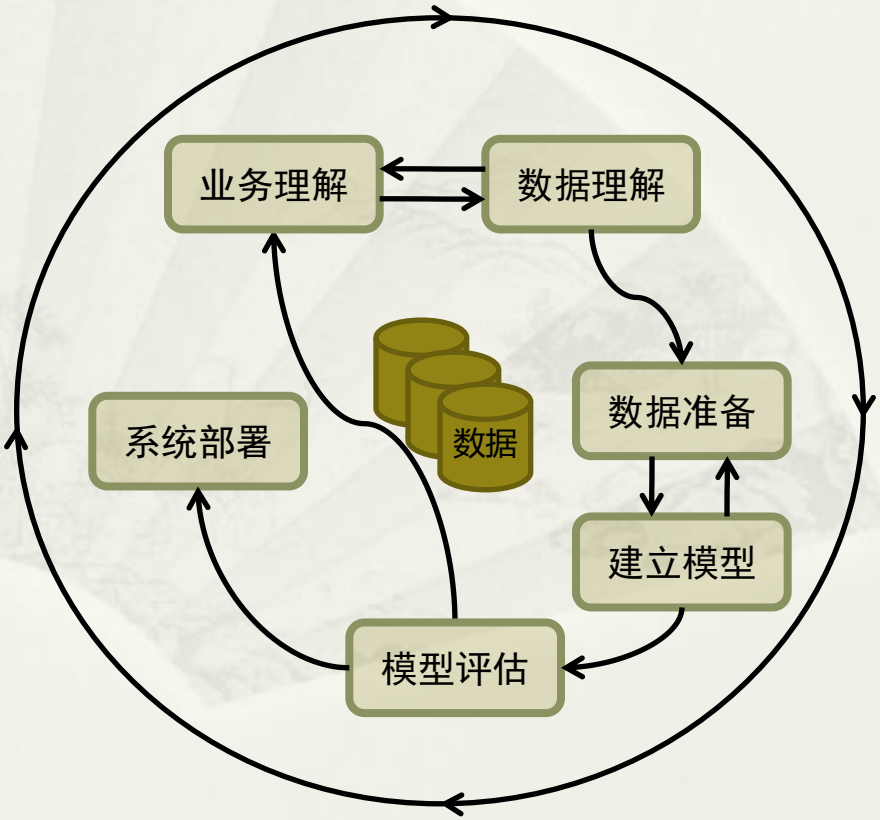
分析流程（SEMMA）

* Sample（数据选取）
* Explore（数据探索）
* Modify（数据预处理）
* Model（模型建立）
* Assessment（模型评价）
* Deploy（模型部署）
* Revise（模型修正）

模型开发步骤

- 业务理解
  - 目标定义
  - 分析窗口定义
- 数据探索
  - 基本情况
  - 单变量显著性分析
- 数据准备（预处理）
  - 数据质量检查
  - 数据预处理
  - 宽表构建
- 模型构建
  - 抽样：训练集、验证集
  - 运用相应算法训练数据

数据挖掘方法论



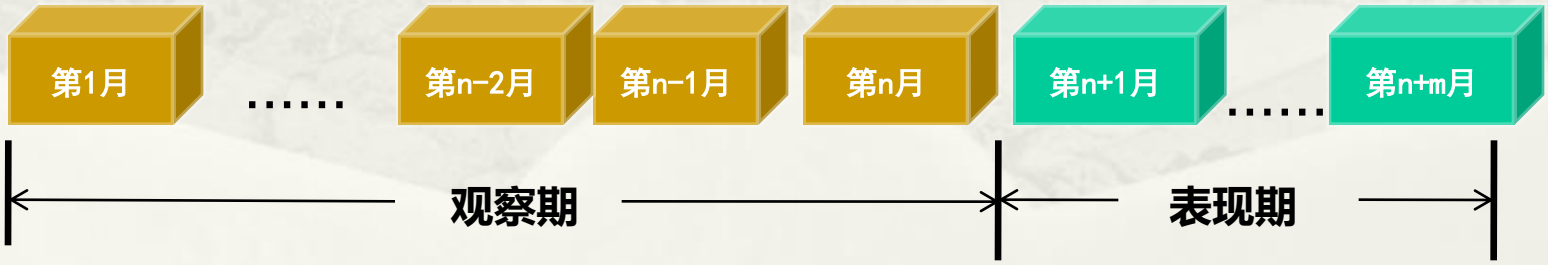
业务理解

业务问题定义

- 面对的分析对象是什么。
- 样本数据是在观察期内正常的合同，在表现期已经提前接清（流失）的合同为目标值1，否则为0。

分析时间窗口

- 基于业务特征和行业经验，我们一般把时间窗口设计为8-18个月，时间窗口设计为观察期6-12个月，表现期2-6个月。
- 模型是通过观察期的分析对象特征和持有客户的行为特征来预测分析对象在表现期的状态。



行业数据挖掘主题


分析模型
市场预测
业务量预测
收入预测
营销销售
客户分群
交叉销售
资费定价对比
渠道配置优化
套餐收入测算
客户响应
客户动因/偏好分群
服务维系
客户价值
流失预警
客户满意度
客户忠诚度
风险控制
客户信用
欺诈发现
欠费预测

分析模型介绍（基础模型—必做）

分析模型	模型应用介绍
客户细分	在客户理解和营销策划阶段，业务人员根据不同客户群自然属性和行为特征，进行差异化的客户分析和客户服务，从而指导针对性营销工作
客户响应	预测客户对新产品、套餐、营销活动的响应情况，指导新产品和套餐销售策略，提升营销活动的推出效果
流失预警	预测客户在未来一段时间内的流失可能性，采取不同的客户维系和挽留的方法，提高运营商存量保有水平，可以用于主动拆机流失、话务量流失、零次户流失、欠费停机流失等预测
客户信用	针对客户不同信用等级，在客户业务受理、业务使用、帐单催欠等环节进行不同等级的信用控制，从而提供运营商的客户信用管理水平

分析模型介绍（可选—业务需求驱动）

分析模型	模型应用介绍
业务量预测	在客户理解阶段，通过分客户群、分产品、分地域，**预测运营商**不同客户群体或者单个客户未来一段时间内的业务量的变化趋势，根据预测的不同情况，采取不同的营销措施
收入预测	在客户理解阶段，通过分客户群、分产品、分地域，**预测**运营商不同客户群体或者单个客户未来一段时间内的收入的变化趋势，根据预测的不同情况，采取不同的营销措施
客户价值	综合考虑客户当前价值和潜在价值,分析客户对运营商的贡献，根据客户价值高低，采取不同的营销、销售和服务策略
欠费预测	分析影响客户欠费的因素，进行欠费预测分析中，生成欠费预测结果，从而适时地对客户进行重点跟踪，并在必要时采取措施，以减少损失
渠道配置优化	对渠道的人、财、物的资源优化配置，分析客户在各个渠道的行为特征，结合不同渠道营销成本和渠道的特点，帮助营销执行人员选择合理的渠道，对客户进行针对性营销，降低营销成本
客户动因/偏好	通过分析客户产品偏好，客户偏好渠道等，可以及时、准确的洞察客户动因和偏好，跟踪和把握客户需求，全面深入理解客户心里需求，提高针对性营销有效性

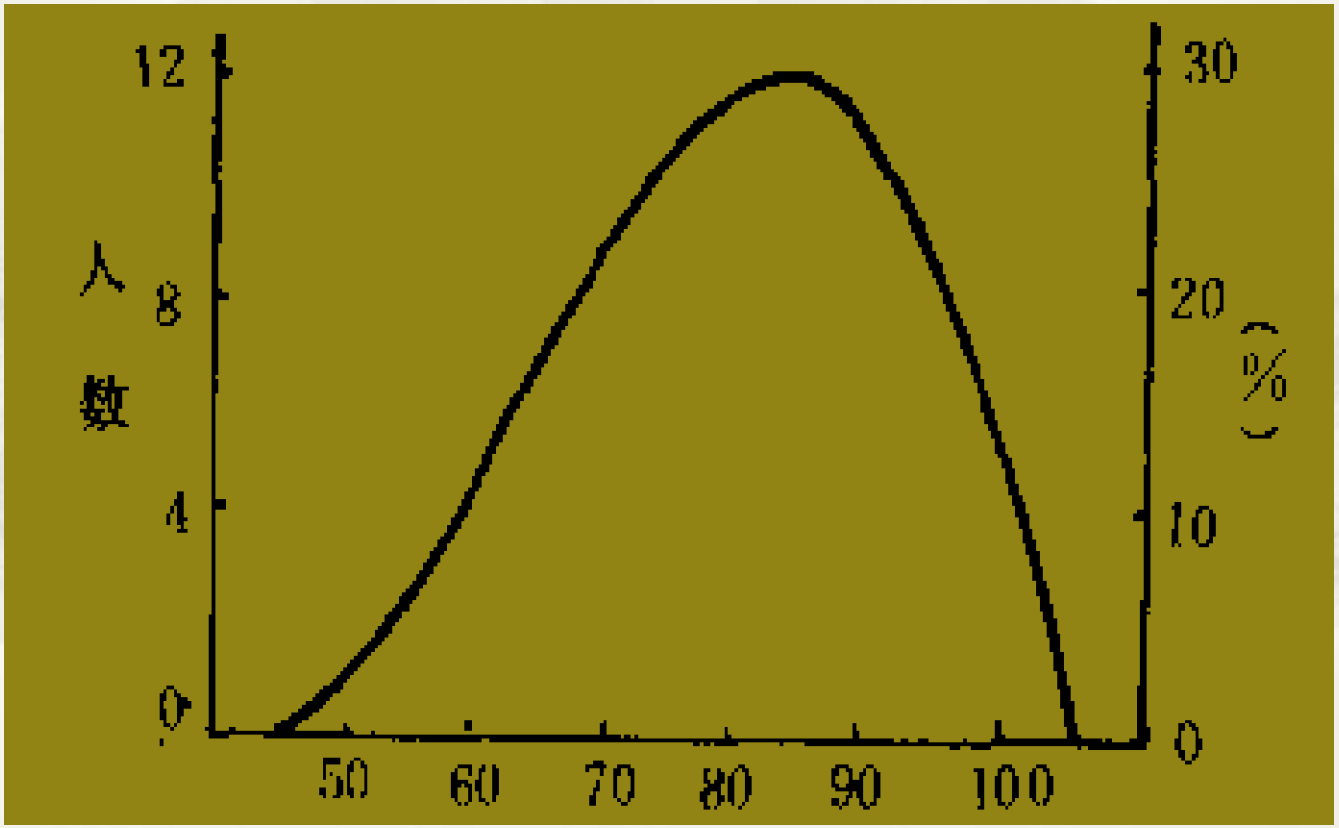
分析模型介绍（可选—业务需求驱动）

分析模型	模型应用介绍
欺诈发现	通过发现客户欺诈模式，在客户业务受理、业务使用、帐单催欠等环节进行控制，提高风险控制水平。
交叉销售	通过分析产品之间关联性，找出产品组合的规则，指导组合产品的设计，提高产品组合营销的效果，降低单产品的流失率
资费定价对比	通过套餐测算方程，研究套餐对于客户的区隔，以及套餐对不同ARPU区段客户的影响，为合理选择套餐方案和套餐分档设计提供有力的支持
套餐收入测算	完成目标客户在新资费政策下的消费行为预演，预演结果与原资费政策下用户的消费情况进行对比评估，验证资费是否与预期营销目的相符，用于电信运营商对资费政策和营销计划的调整
客户满意度	分析影响客户满意度的关键因素，找出客户满意度的评价指标，改善客户服务水平
客户忠诚度	分析客户忠诚度的影响因素，能够找出影响客户忠诚度的评价指标，提高客户对自身的依赖程度

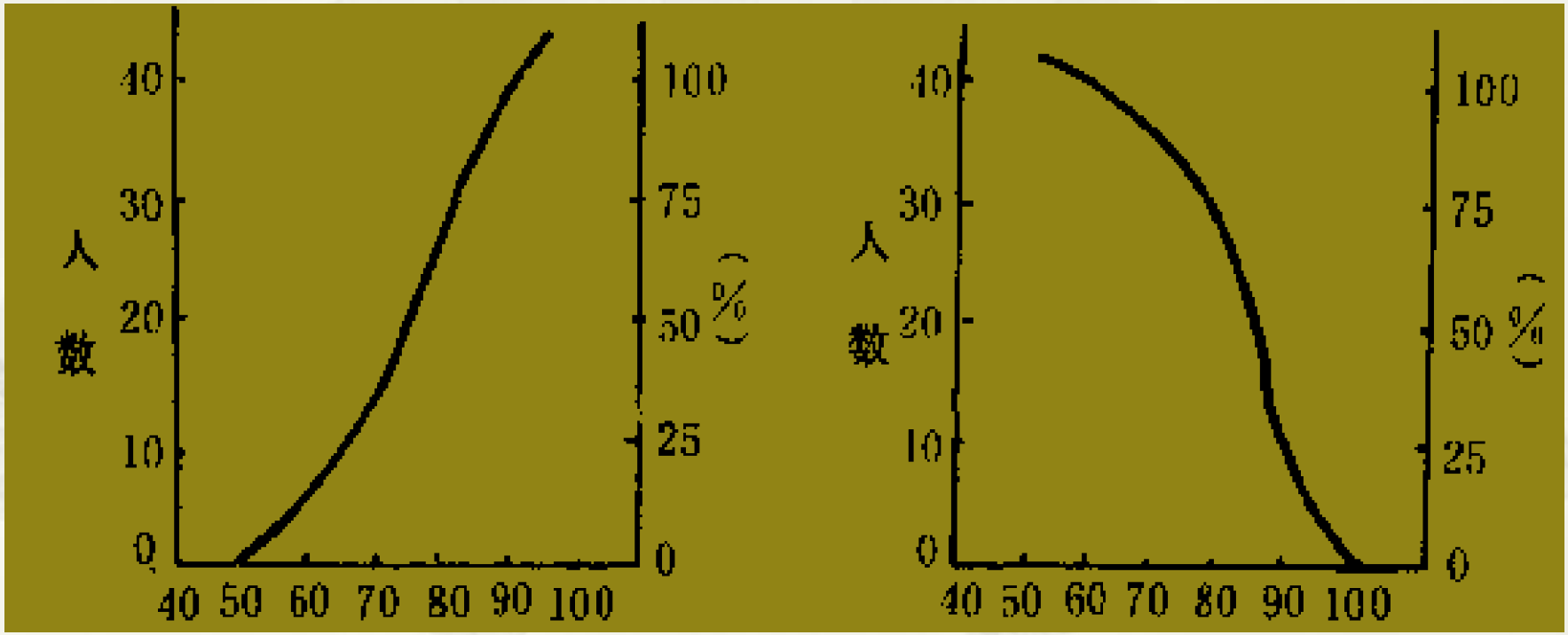
数据探索

- 重要变量的分布探索
- 密度函数(P(X=x))
- 分布函数(P(X≤x))
- 生存函数(P(X>x))
- 变量间关系探索
- 相关分析
- 维度诅咒

密度函数



分布函数和生存函数


人数
40
30
20
10
0
40 50 60 70 80 90 100
100
75
50
25
0
%
人数
40
30
20
10
0
40 50 60 70 80 90 100
100
75
50
25
0
%

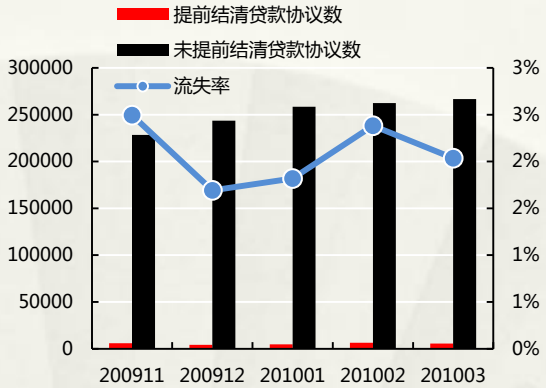
数据探索

房贷合同发展基本状况

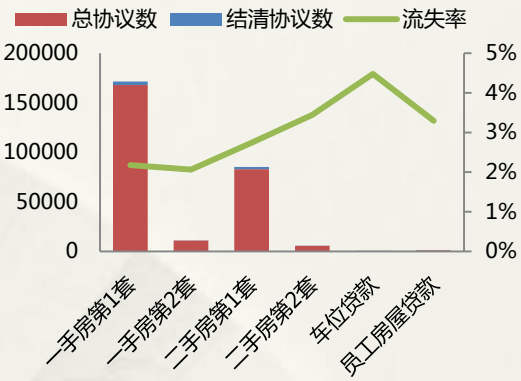
数据理解 | 业务理解

> 房贷合同结清状况趋势分析
> 房贷合同的构成分析
> 提前结清房贷合同构成分析
> ……

房贷提前结清按月发展趋势



房贷提前结清按合同类型分析



特征变量显著性分析

变量对提前结清房贷行为的影响程度

> 房贷合同基本信息

房贷合同的类型、短期期限的贷款、贷款余额较少、已还款期数较多、当前贷款执行利率较高都能显著地影响客户的提前结清还款的倾向。

> 房贷客户基本信息

客户年龄较大、高学历、已婚人士更易提前偿付贷款。

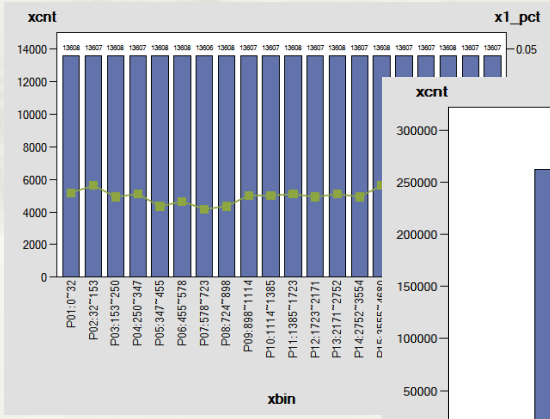
> 房贷客户交易行为

客户储蓄存款余额多寡、储蓄存款次数金额、接收转账汇款的频度和额度都能影响到客户提前还贷。

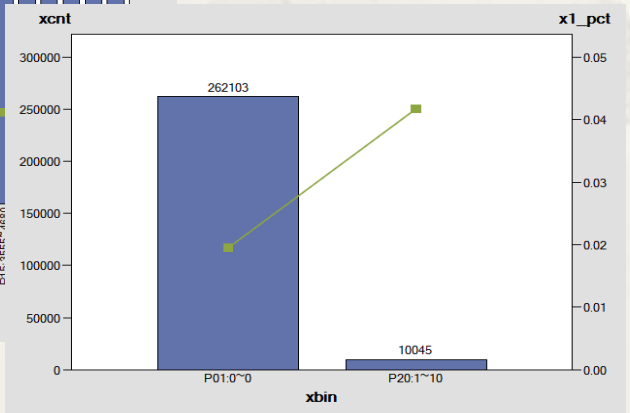
> 历史还贷情况

历史提前还贷次数、金额和单笔金额对提前结清贷款有显著影响。

储蓄账户余额



近6月个贷提前还款次数



最近半年有提前还贷行为的客户，其未来2月内提前结清贷款比例为4.17%，是平均水平2.04%的2倍多。

数据探索-相关性分析

* 散点图
* 相关系数
* 皮尔森相关系数
* 斯皮尔曼相关系数
* 肯德尔相关系数
* 霍弗丁相关系数

数据探索-相关性分析

分析目的

> 分析各特征变量间相互影响的强弱程度，剔除彼此强相关的变量，减少变量冗余程度。

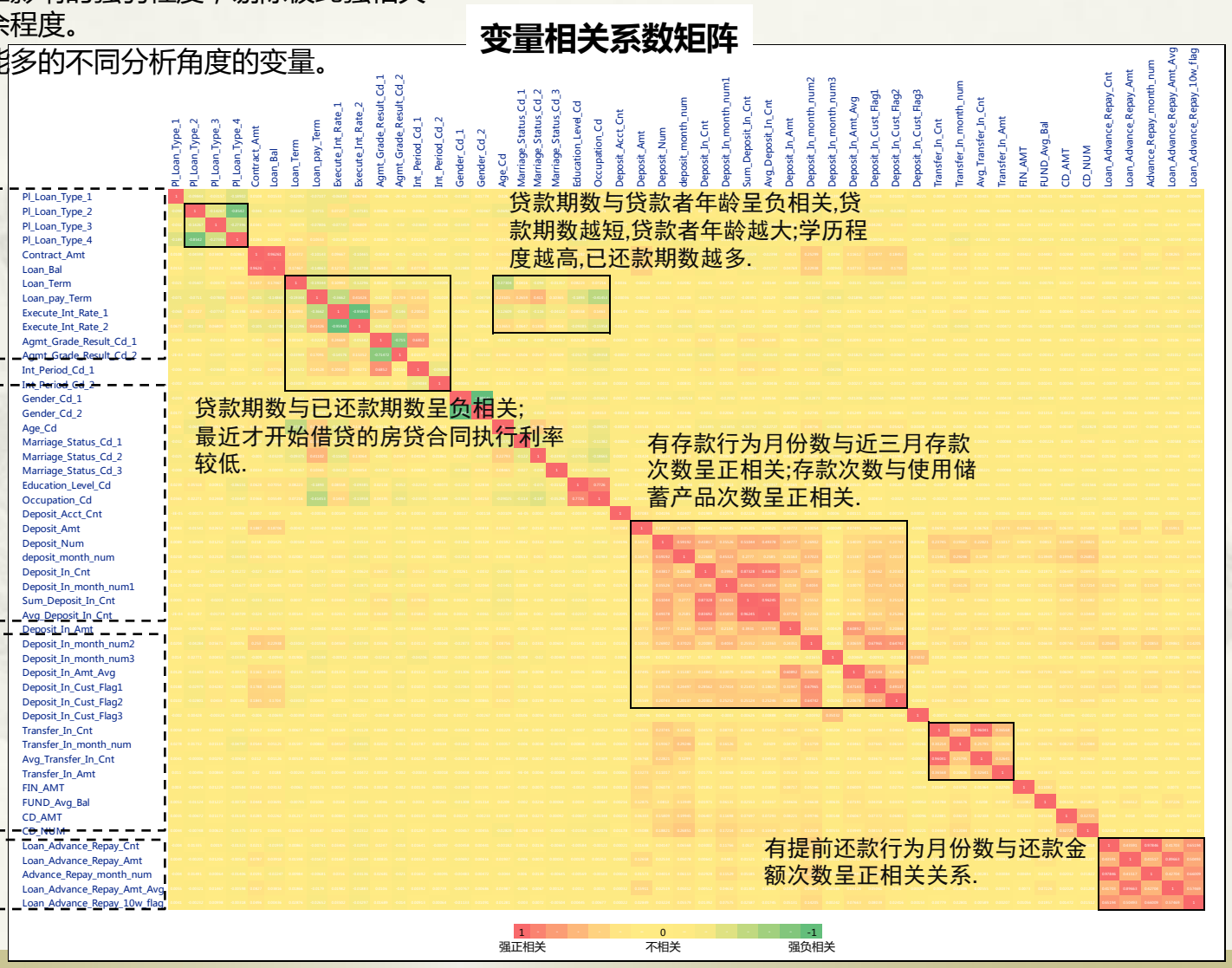
> 为模型构建选取尽可能多的不同分析角度的变量。

房贷合同基本信息

客户基本信息

客户交易信息

历史还贷记录



数据准备（预处理）


数据准备
1
客户群体选择
■ 持有某产品的客户
■ 持有某产品的客户
■ 持有某产品的客户
■ .....
2
数据抽样
■ 每一层中所有观察值有相同概率被抽样
■ 保护样本中的层的比率
■ 采用分层抽样
■ 在建模时增加分类的准确性
3
变量分析
■ 运用基本的数据统计进行变量分析
■ 变量分析包括变量的均数、方差、最大最小值和百分数等数字特征
■ 分析目的是了解数据的分布和质量
■ 分析成果作为后续数据调整的依据
4
数据处理
■ 缺失值处理
■ 异常值处理
■ 极值处理
5
数据调整
■ 变量衍生
■ 变量的数据转换
■ 变量归并
■ 变量标准化
■ 变量加权

分析客户群选择

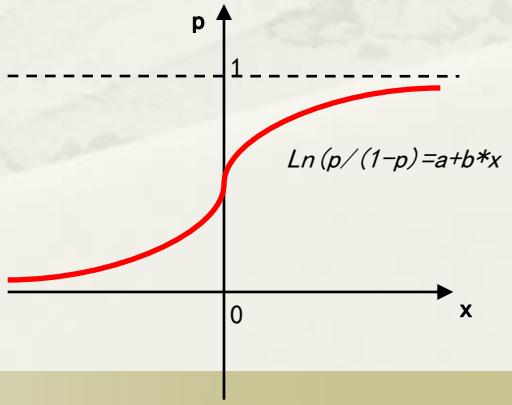
- 分析客户群
  - 分析客户群：2010年4月底状态正常的客户。
  - 分析样本观察窗口是2009年11月-2010年4月，表现窗口是2010年5-6月。

模型构建-抽样设计

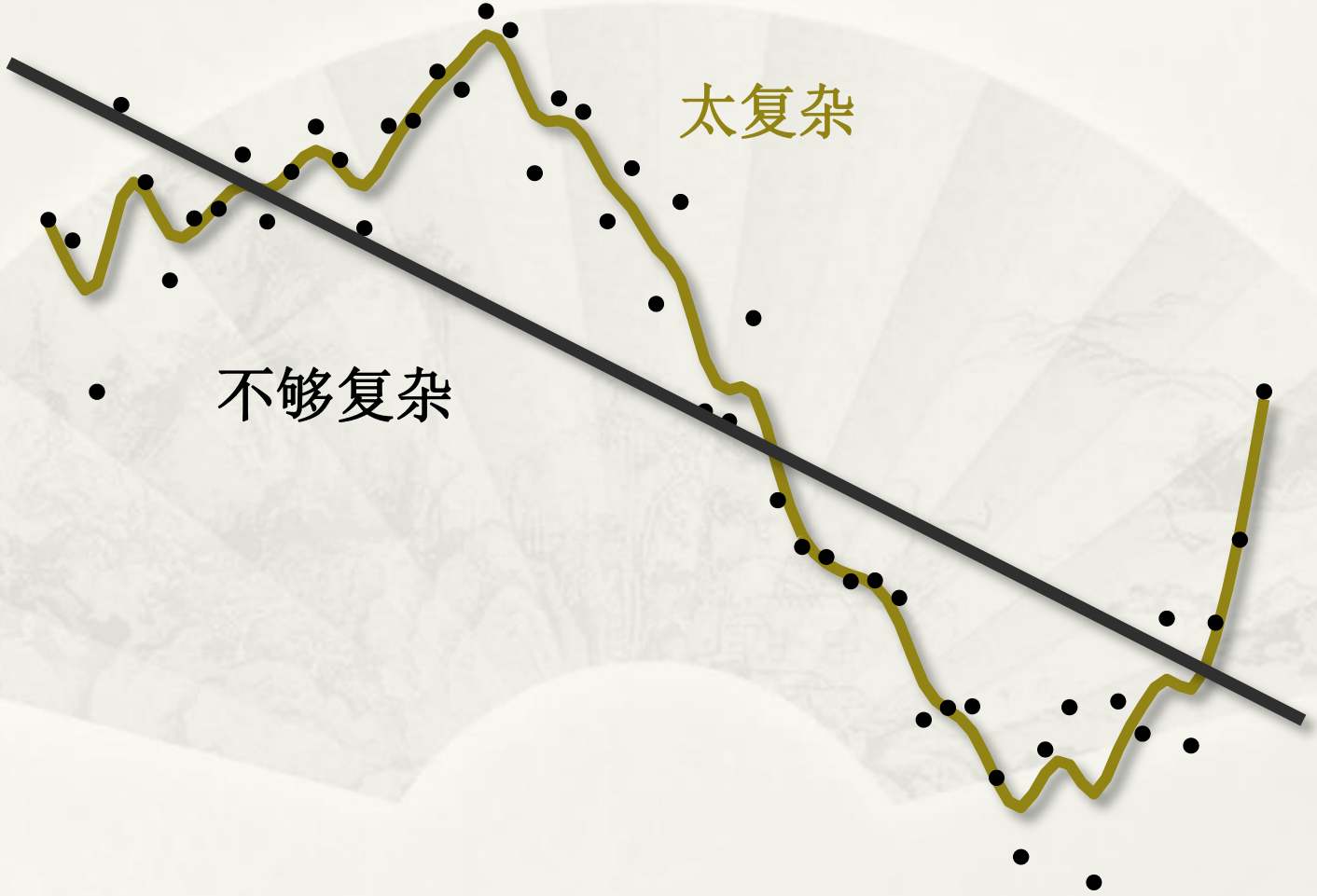
- 在建模过程中，需要使用抽样技术，从符合考察范围的用户总体中以适当方法，抽取一定量的样本，进行分析。从而保证模型的可靠性和预测模型的稳定性，同时，降低建模初期的数据准备压力。
- 针对不同的模型的实际数据的差异性，采用不同的抽样方法来进行抽取模型的样本。
  - 简单无重复随机抽样
  - 分层抽样: 分层等比例随机抽样、分层不等比例随机抽样
  - 分类抽样

模型构建-模型选择

- 逻辑回归算法Logistic Regression
  算法基本公式：
  P = exp（S）/ (1 + exp（S)）
  S = a + b1* X1+ ~ + bn* Xn
  Odds = P / (1-P)
- LOGISTIC模型主要用于处理因变量为分类变量的问题，如：客户信用的好、坏；在分析时，我们感兴趣的是因变量取各个值的概率P。
  - P表示因变量取某个值(Y=1)的概率：
    1+exp(a + b1* X1+ ~ + bn* Xn)
  - 经变换后生成对数似然比（Odds）：
    ln(p/(1-p))=a + b1* X1+ ~ + bn* Xn
    揭示了n个自变量X与响应变量Y之间的非线性关系；
  - 因变量Y是分类型变量（如二分变量0/1）；用p表示Y=1发生的概率；预测p与n个自变量X的关系。
- 逻辑斯蒂模型的优点：
  - 变量的可解释性强
  - 对变量分布正态性和方差齐性不做要求
  - 对自变量类型不做要求
  - 有较强的预测稳定性和准确性



模型复杂度



一点困扰

* 数据分区
* 业务支持

模型结果解释

评分公式

计算相应的概率：

P = exp(S)/(1 + exp(S))

其中，S= a + b1* X1+ … + bn* Xn，重要变量Xi及其参数bi如前所述。

变量分析

变量的对数似然率 odds rate 表示该变量取值增加1个单位，对目标p与非目标值1-p间的差异贡献的程度，odds rate越大，说明该变量对目标变量的区分能力强。

。

重要影响变量

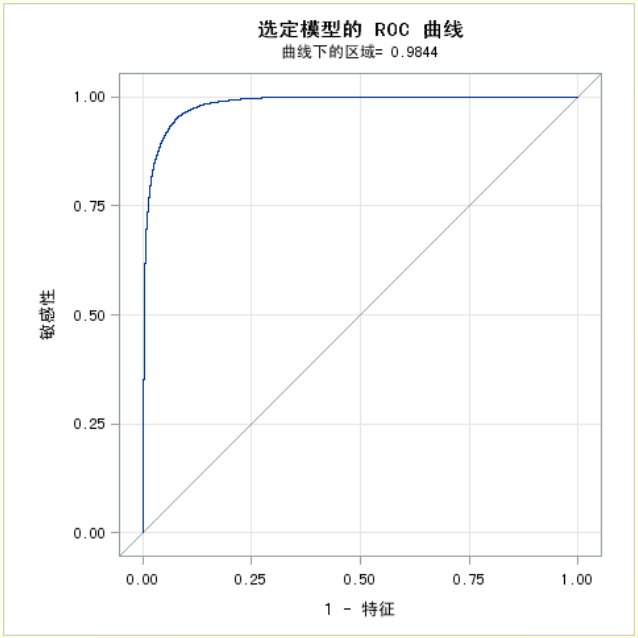
依据各变量的odds rates估计值和变量间相关关系，可知影响客户的重要变量有：

■。

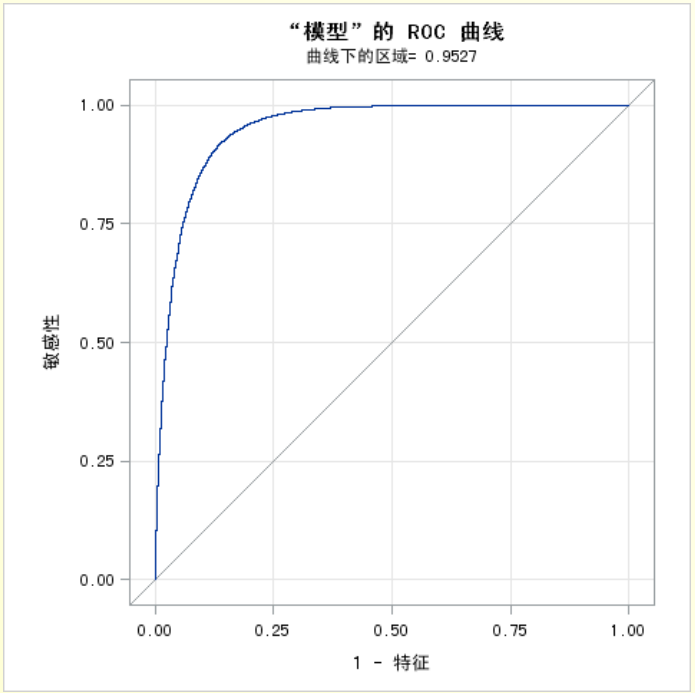
模型评估

- 验证方法
  - 样本内验证（In sample Validation）
    - 把建模样本分成两部分，如60%：40%，一部分用来建模，另一部分用来验证。
  - 样本外验证（Out of Sample Validation）
    - 把应用到建模样本以外的客户群进行验证，如：不同时间段的相似客户群。
  - 两种验证方法结合保证了模型的稳定性。

模型评估-ROC图



V.S.



逐步法:16个解释变量

因子分析法:8个解释变量

其他验证指标

- K-S曲线（图1）

国际上用K-S指标来衡量评估结果是否优于期望值，具体标准是，如果K-S大于40%，模型具有较好的预测功能，发展的模型具有成功的应用价值。

- 准确率和查全率（图2）
- Lorentz曲线（图3）

“捕获”的目标人数占目标总人数的百分比。

- 提升率、覆盖率（图4）

图1

K-S统计量代表了预测响应/非响应对象能达到的最大提升度（与随机判断比较）。

$$KS = \max_{s} |B(s) - G(s)|$$

rank	%ile for Combined Populatio	Total Account	Good	Bad	Score Range	Bad%	Cum Pct Bad	Cum Bad Ratio
0	1	2495	2263	232	474—594	9.30%	17.9%	9.3%
1	2	2451	2336	115	595—606	4.69%	26.8%	7.0%
2	3	2562	2475	87	607—614	3.40%	33.5%	5.8%
3	4	2642	2574	68	615—620	2.57%	38.7%	4.9%
4	5	2139	2084	55	621—624	2.57%	43.0%	4.5%
5	6	2443	2395	48	625—628	1.96%	46.7%	4.1%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	17	2978	2955	23	654—655	0.77%	73.5%	2.2%
17	18	1567	1557	10	656—656	0.64%	74.3%	2.2%
18	19	3226	3210	16	657—658	0.50%	75.5%	2.1%
19	20	1596	1588	8	659—659	0.50%	76.2%	2.0%

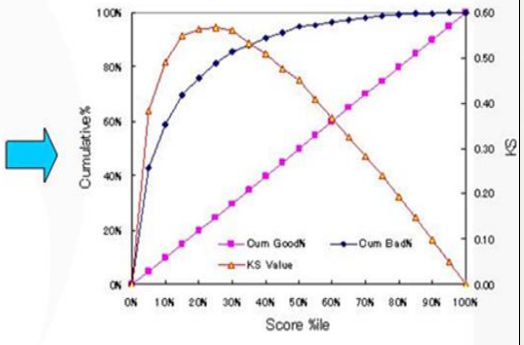
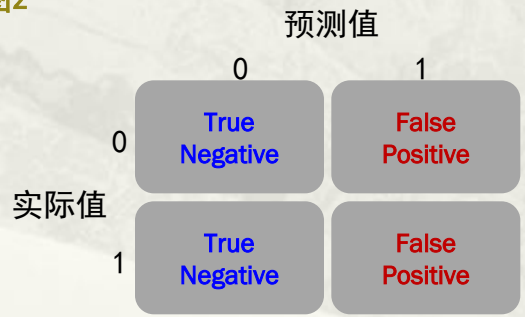


图2



准确率=TP/(FP+TP)

查全率（覆盖率）=TP/(FN+TP)

图3

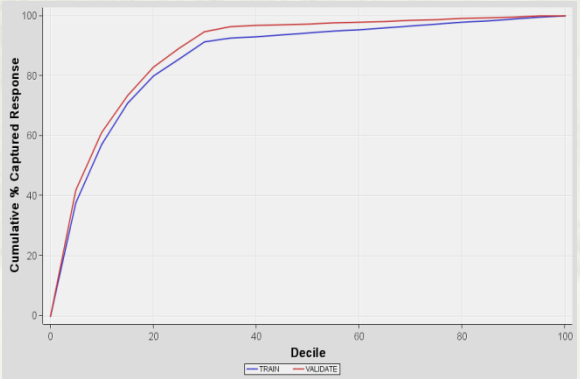
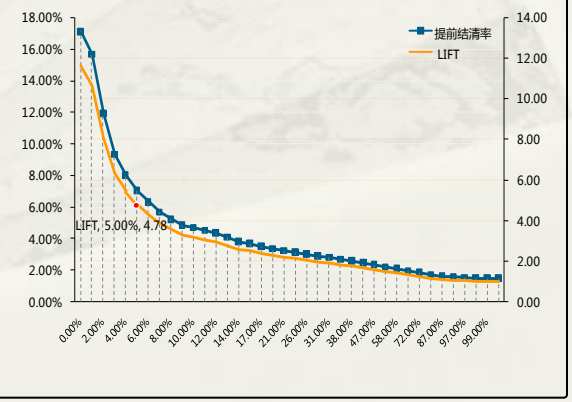
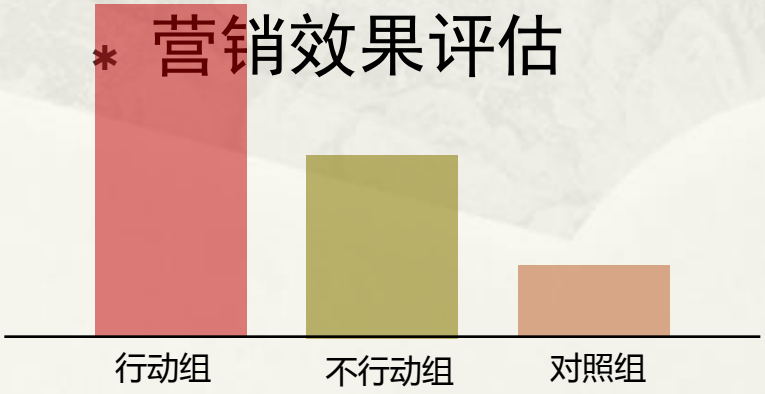


图4



模型应用

- 业务应用
  - 制定客户维系营销策略
    - 对提前接清倾向评分较高的客户，根据其年龄、性别、职业特点进行相应的理财产品推荐。
    - 捆绑销售其它个人产品
    - 对提前接清倾向评分较高、价值较低的客户提供优惠。
  - 对其它相关产品开发提供支持
- 营销活动效果评估
  - 模型预测效果评估
  - 营销效果评估



各组客户描述:

- 行动组:模型预测评分较高、参与营销活动。
- 不行动组:模型预测评分较高、不参与营销活动。
- 对照组:随机抽取、参与营销活动。

行动组和对照组营销效果对比体现了模型预测效果的优劣。
行动组营销效果与和不行动组办理业务情况对比体现了营销活动效果。

数据挖掘模型业务流程的应用