


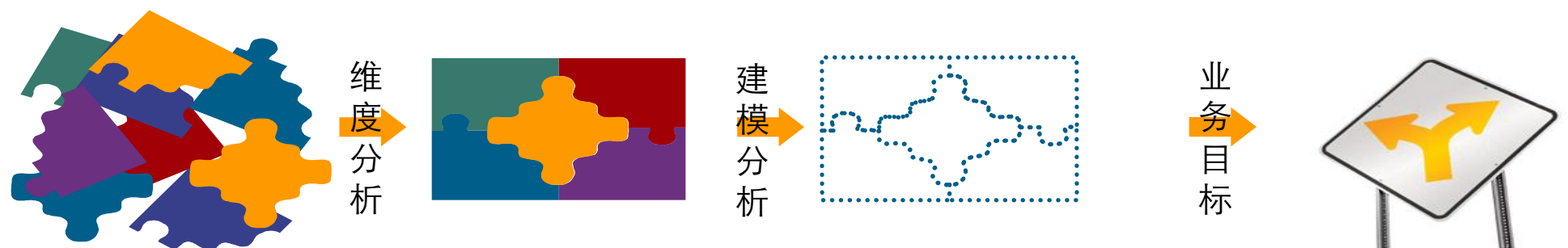
案例：个人贷款违约预测模型



数据科学实战：Python篇

数据科学方法论

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 $\xrightarrow{+主题}$ 信息 $\xrightarrow{+规则}$ 知识 $\xrightarrow{+业务经验}$ 决策和行动



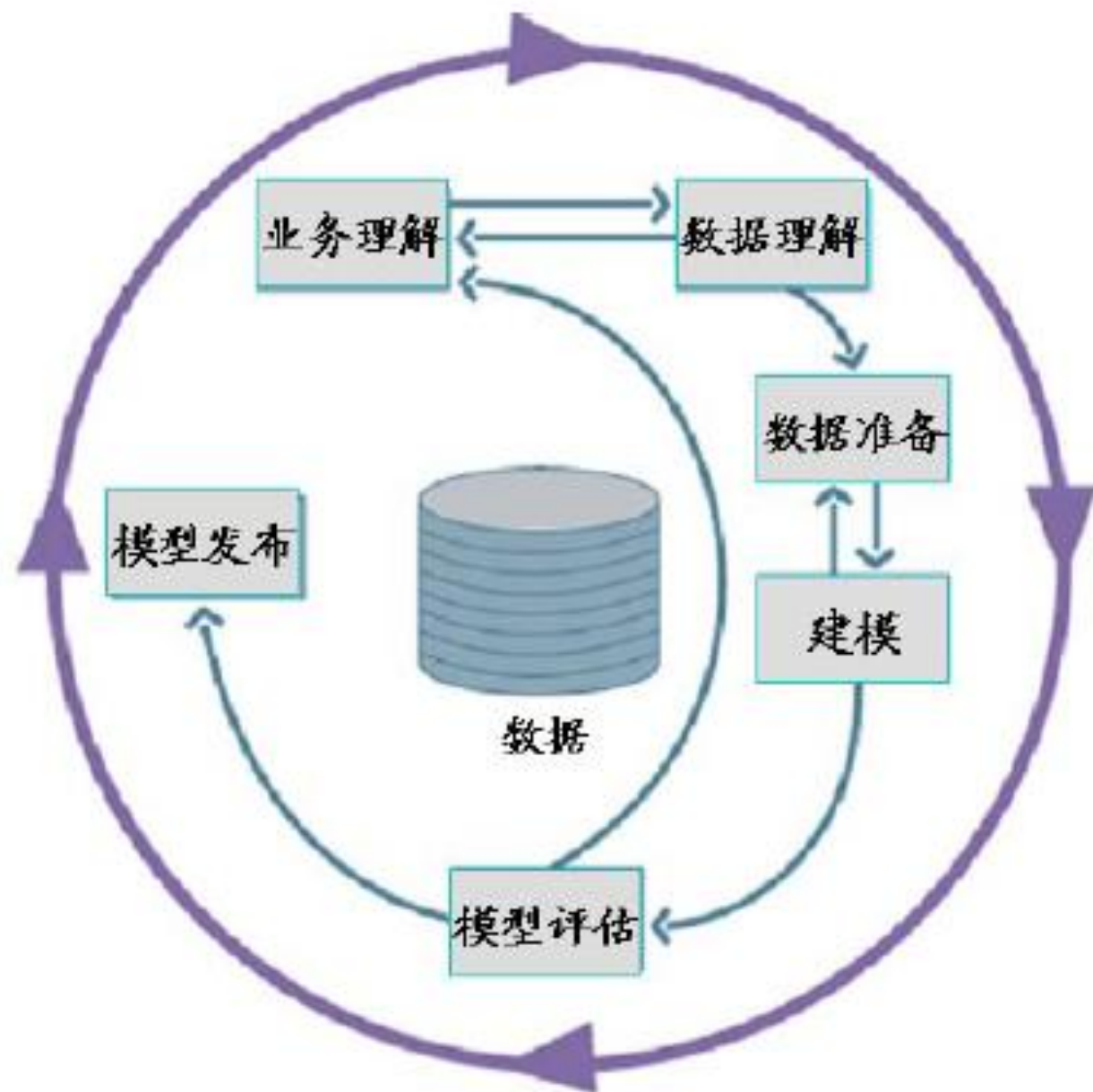
发现:

找出隐藏在数据背后的模式，这些模式能把数据转化为知识

部署:

应用已发现的知识达成实用的目的 - 例如：预测

数据挖掘实施路线图：CRISP-DM



1 数据理解



案例背景

- 本数据为一家银行的个人金融业务数据集，可以作为银行场景下进行个人客户业务分析和数据挖掘的示例。这份数据中涉及到5300个银行客户的100万笔的交易，而且涉及700份贷款信息与近900张信用卡的数据。通过分析这份数据可以获取与银行服务相关的业务知识。例如，提供增值服务的银行客户经理，希望明确哪些客户有更多的业务需求，而风险管理的业务人员可以及早发现贷款的潜在损失。
- 可否根据客户贷款前的属性、状态信息和交易行为预测其贷款违约行为？

案例背景

- 截取自一家银行的真实客户与交易数据；
- 涉及客户主记录、帐号、交易、业务和信用卡数据；
- 一个账户只能一笔贷款，“loan”表中记录了客户贷款信息。

loan_id	account_id	date	amount	duration	payments	status
5314	1787	1993-07-05	96396	12	8033	B
5316	1801	1993-07-11	165960	36	4610	A
6863	9188	1993-07-28	127080	60	2118	A
5325	1843	1993-08-03	105804	36	2939	A
7240	11013	1993-09-06	274740	60	4579	A
6687	8261	1993-09-13	87840	24	3660	A
7284	11265	1993-09-15	52788	12	4399	A
6111	5428	1993-09-24	174744	24	7281	B

案例背景

贷款表(Loans)		
名称	标签	说明
disp_id	权限号	(主键)
loan_id	贷款号	
account_id	账户号	
date	发放贷款日期	
amount	贷款金额	
duration	贷款期限	
payments	每月归还额	
status	还款状态	A代表合同终止，没问题；B代表合同终止，贷款没有支付；C代表合同处于执行期，至今正常；D代表合同处于执行期，欠债状态。

•根据以往的贷款数据，状态为B和D的为违约客户，A为正常客户，C的最终状态还不明确。

A	B	C	D
203	31	403	45

数据说明

账户表(Accounts)

- 每条记录描述了一个账户的静态信息
- 条数: 4500

客户信息表 (Clients)

- 每条记录描述了一个客户的特征信息
- 条数: 5369

账户表(Accounts)

名称	标签
account_id	账户号(主键)
district_id	开户分行地区号
date	开户日期
frequency	结算频度 (月, 周, 交易之后马上)

客户信息表 (Clients)

列名	标签
client_id	客户号(主键)
Sex	性别
birth_date	出生日期
district_id	地区号 (客户所属地区)

数据说明

权限分配表(Disp)

- 每条记录描述了客户和账户之间的关系，以及客户操作账户的权限
- 条数: 5369

支付命令表 (Orders)

- 每条记录代表描述了一个支付命令
- 条数: 6471

权限分配表(Disp)		
名称	标签	说明
disp_id	权限设置号	(主键)
client_id	顾客号	
account_id	账户号	
type	权限类型	只用"所有者"身份可以进行增值业务操作和贷款

支付订单表 (Orders)		
名称	标签	说明
order_id	订单号	(主键)
account_id	发起订单的账户号	
bank_to	收款银行	每家银行用两个字母来代表,用于脱敏信息
account_to	收款客户号	
amount	金额	
K_symbol	支付方式	

数据说明

交易表 (Trans)

- 每条记录代表每个账户上的一条交易
- 条数:1056320

贷款表 (Loans)

- 每条记录代表某个账户的上的一条贷款信息
- 条数: 682

贷款表(Loans)

名称	标签	说明
disp_id	权限号	(主键)
loan_id	贷款号	
account_id	账户号	
date	发放贷款日期	
amount	贷款金额	
duration	贷款期限	
payments	每月归还额	
status	还款状态	A代表合同终止，没问题； B代表合同终止，贷款没有支付； C代表合同处于执行期，至今正常； D代表合同处于执行期，欠债状态。

交易表 (Trans)	
名称	标签
trans_id	交易序号(主键)
account_id	发起交易的账户号
date	交易日期
type	借贷类型
operation	交易类型
amount	金额
balance	账户余额
K_Symbol	交易特征
bank	对方银行
account	对方账户号

数据说明

信用卡(Cards)

- 每条记录描述了一个账户上的信用卡信息
- 条数: 892

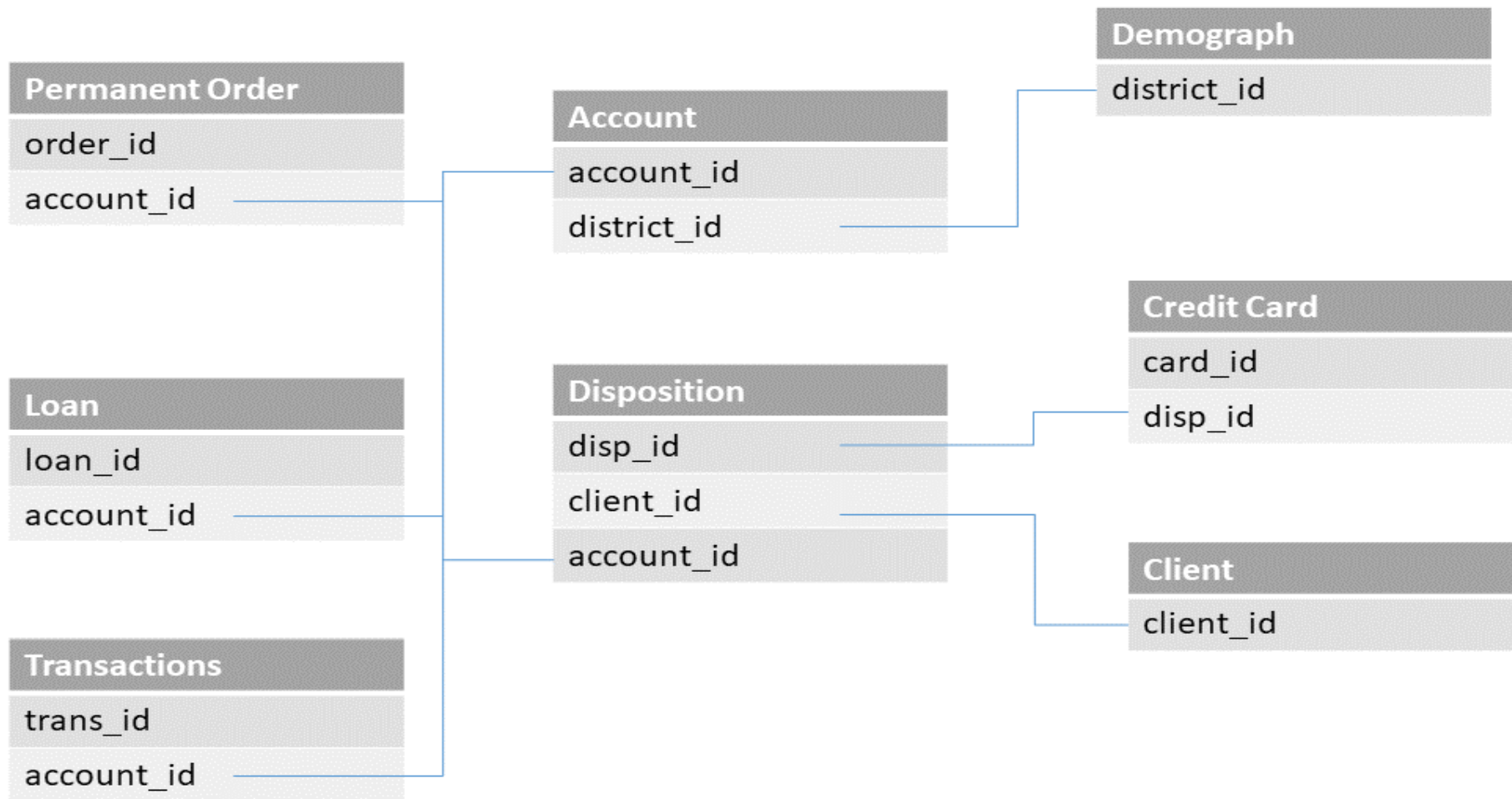
人口地区统计表 (District)

- 每条记录描述了一个地区的人口统计学信息
- 条数: 77

信用卡(Cards)表	
名称	标签
card_id	信用卡id(主键)
disp_id	账户权限号
type	卡类型
issued	发卡日期

人口地区统计表 (District)	
名称	标签
A1 = district_id	地区号(主键)
GDP	GDP总量
A4	居住人口
A10	城镇人口比例
A11	平均工资
A12	1995年失业率
A13	1996年失业率
A14	1000人中有多少企业家
A15	1995犯罪率(千人)
A16	1996犯罪率(千人)

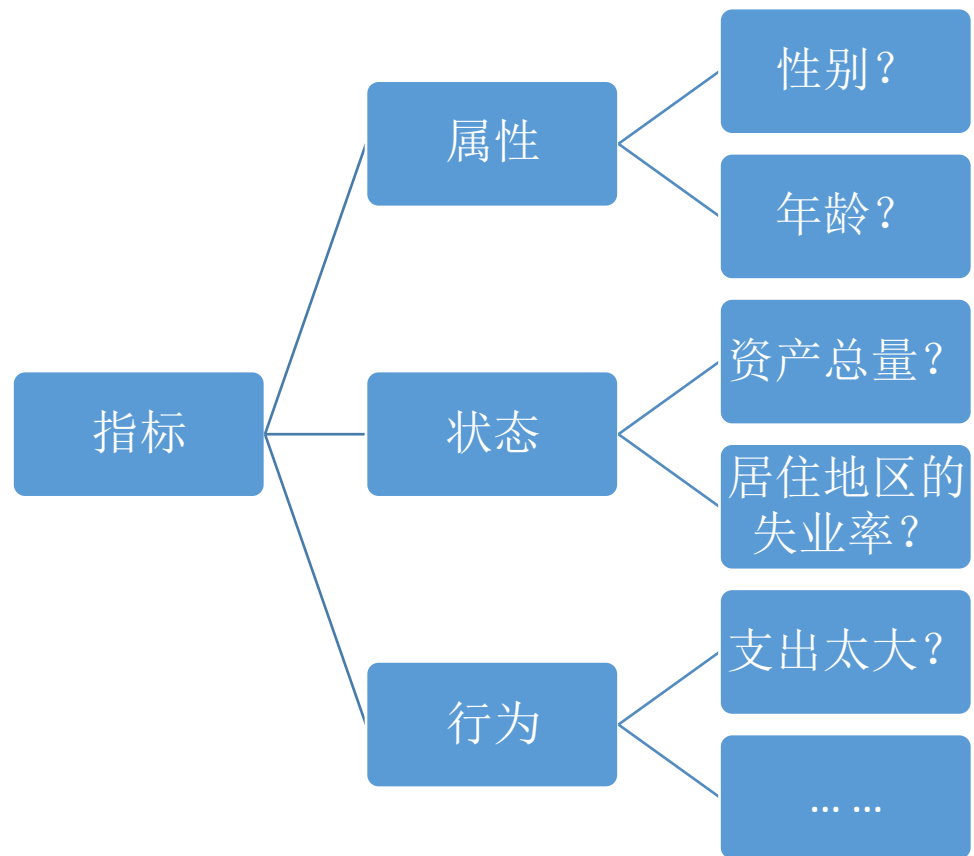
表的关系：数据的实体-关系图(ER图)



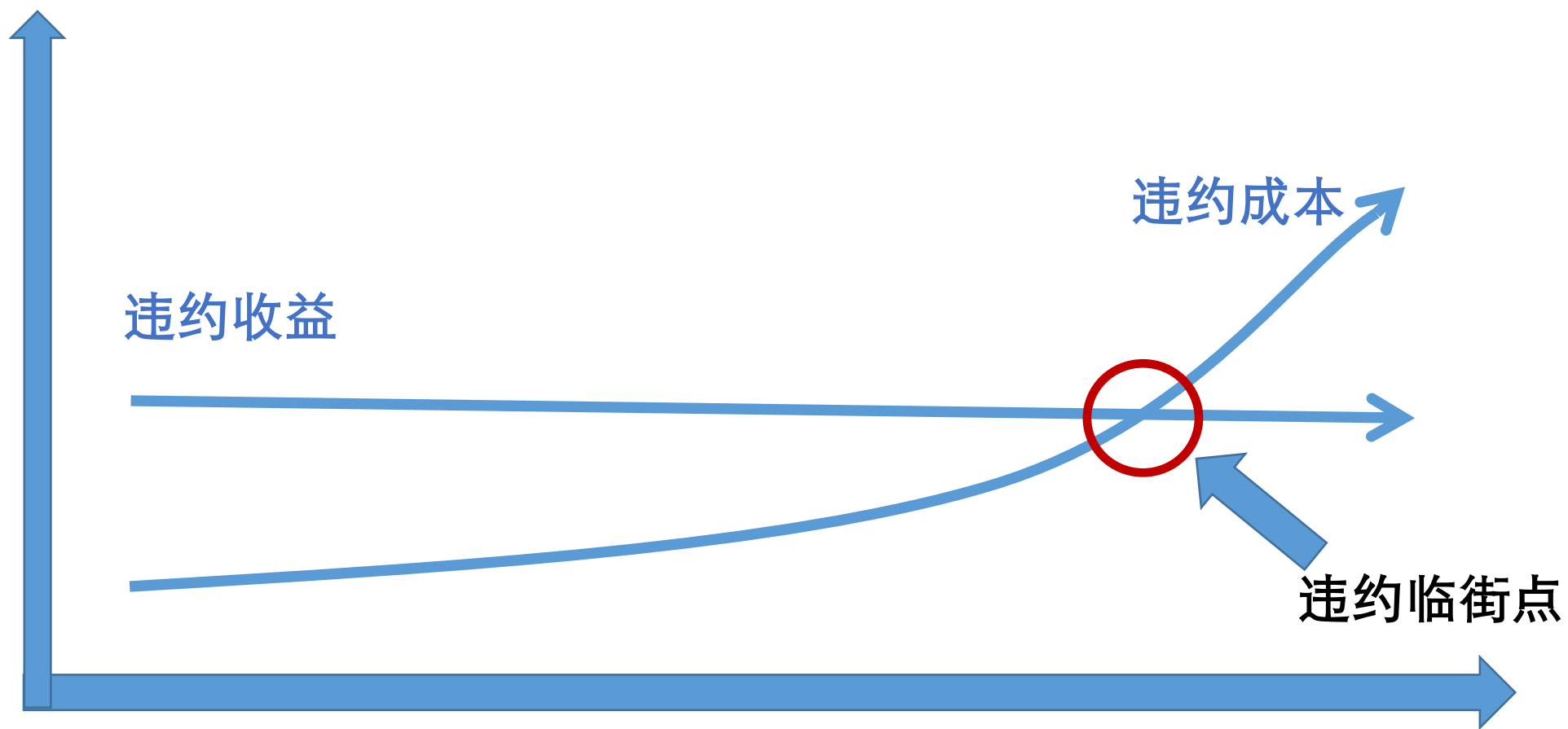
2 业务理解



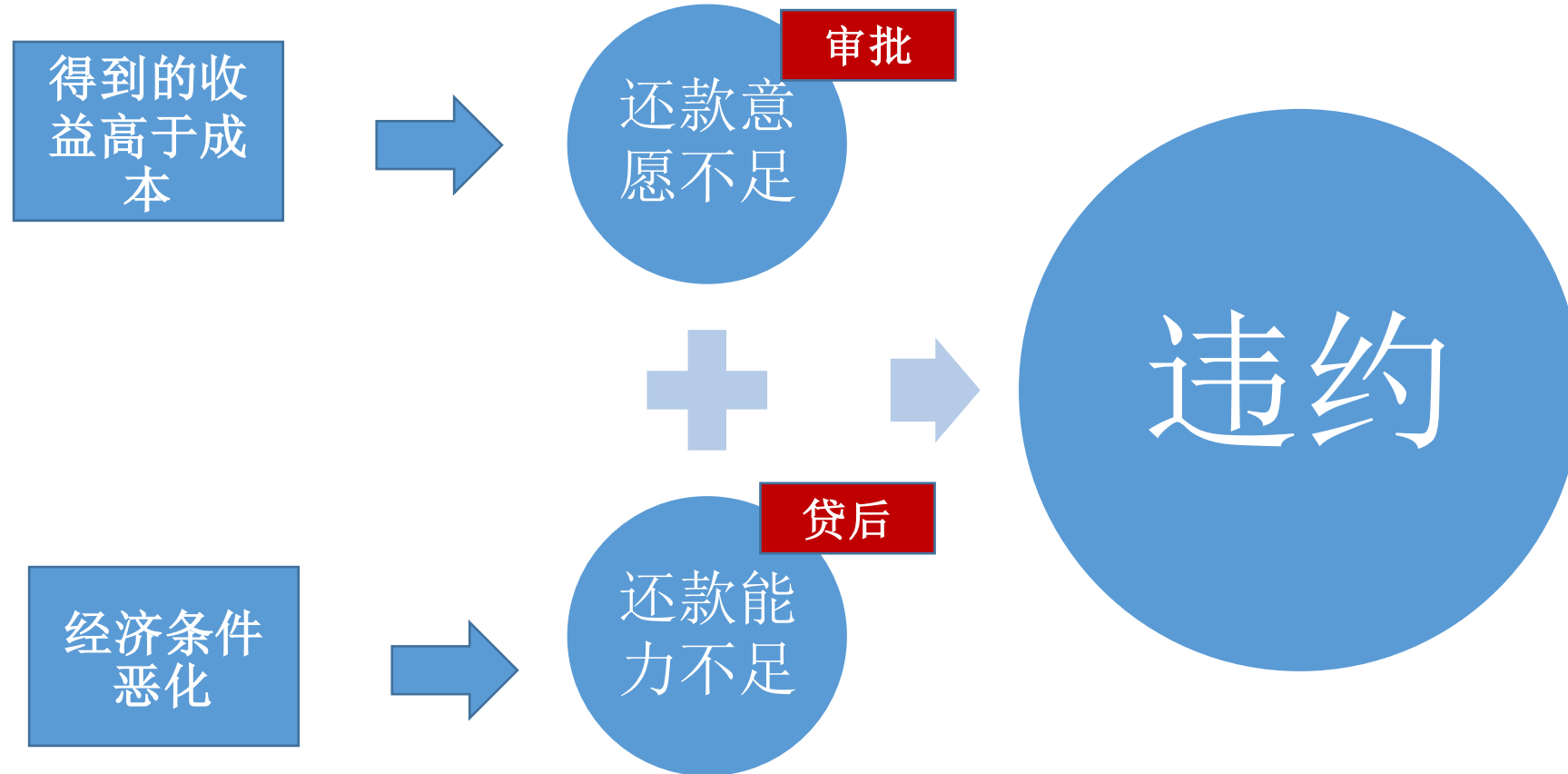
什么指标有预测能力？



客户为什么不还钱？

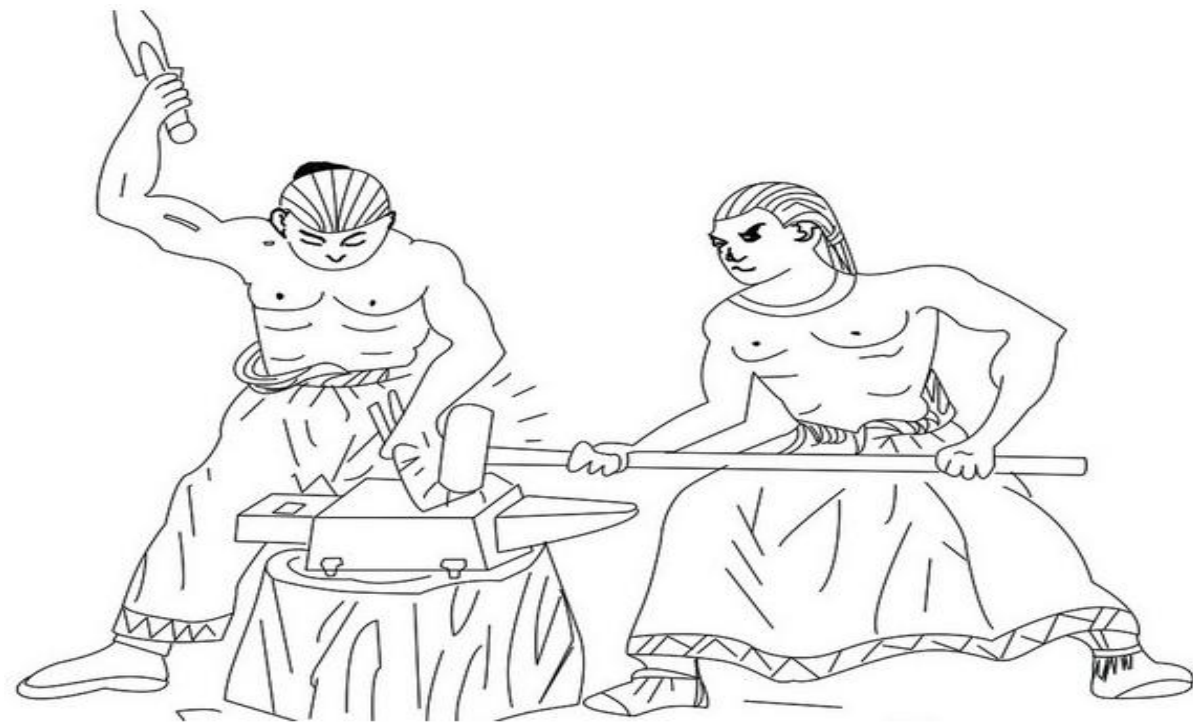


思路整理





- 有预测价值的变量基本都是衍生变量：
 - 一级衍生，比如资产余额；
 - 二级衍生，比如资产余额的波动率、平均资产余额；
 - 三级衍生，比如资产余额的变异系数。



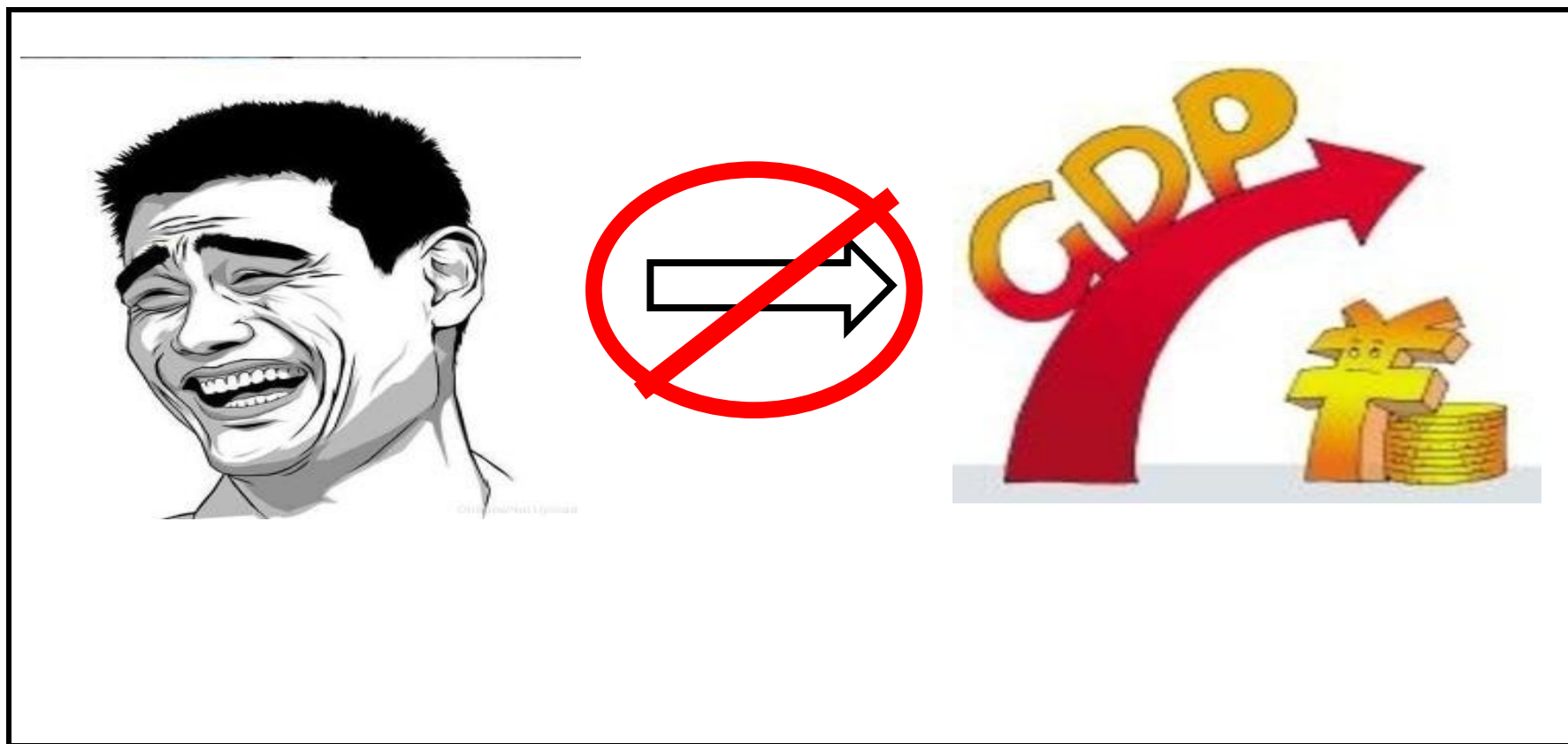
3 数据提取



目标

- 相关与因果之间的关系
- 注意构建模型时数据选取的标准。

相关关系 vs 因果关系



上实际80年代，姚明鞋的尺寸和GDP总量明显正相关。

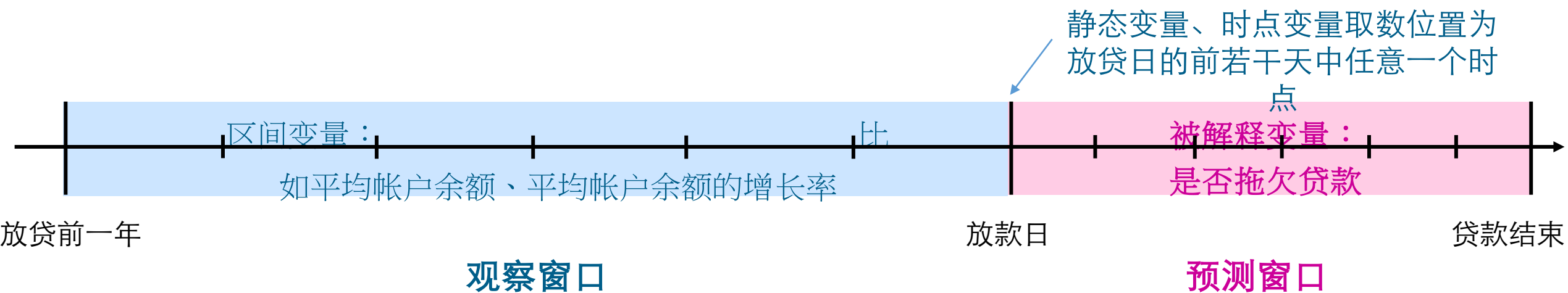
建立因果关系模型

- 我们分析的变量按照时间变化情况可以分为动态变量和静态变量
- 属性变量（比如性别、是否90后）一般是静态变量，行为、状态和利益变量均属于动态变量。
- 动态变量还分为时点变量和区间变量，状态变量（比如当前帐户余额、是否破产）和利益变量（对某产品的诉求）均属于时点变量。行为变量（存款频次、平均帐户余额的增长率）为区间变量。

贷款违约预测的取数规则

- 模型框架

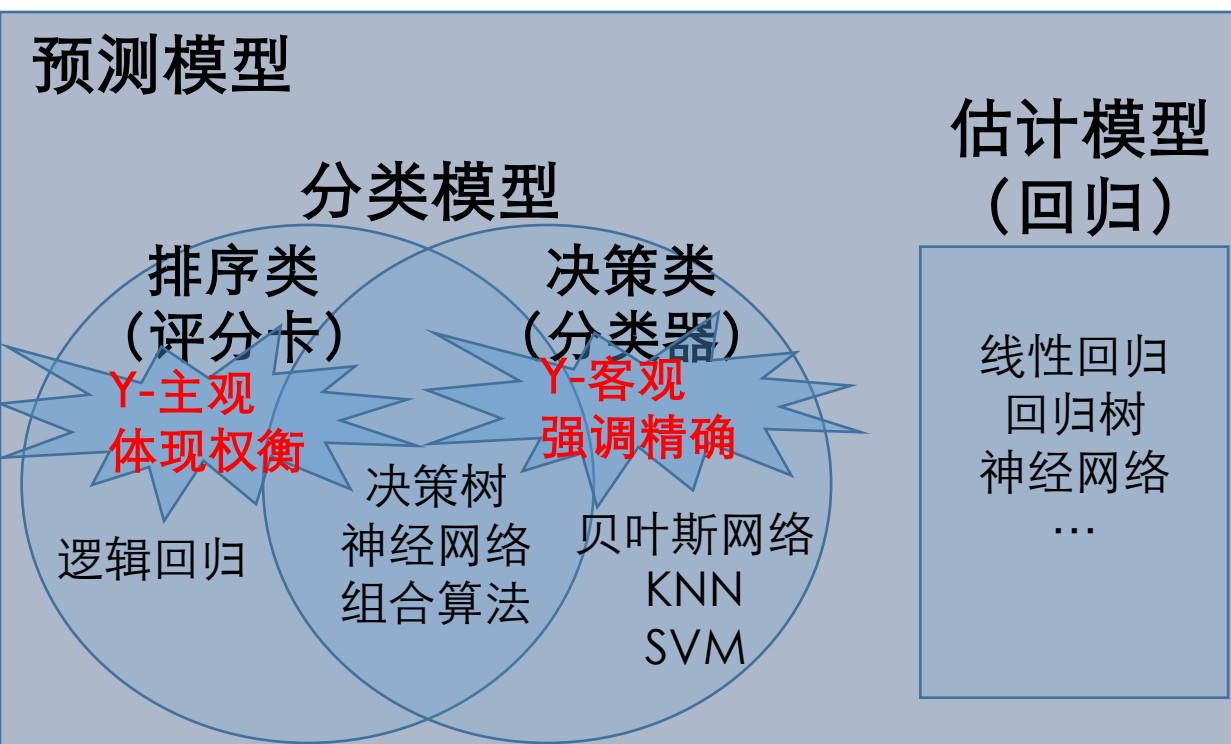
- > 根据客户基本信息、业务信息、状态信息
- > 预估履约期内贷款客户未来一段时间内发生违约的可能



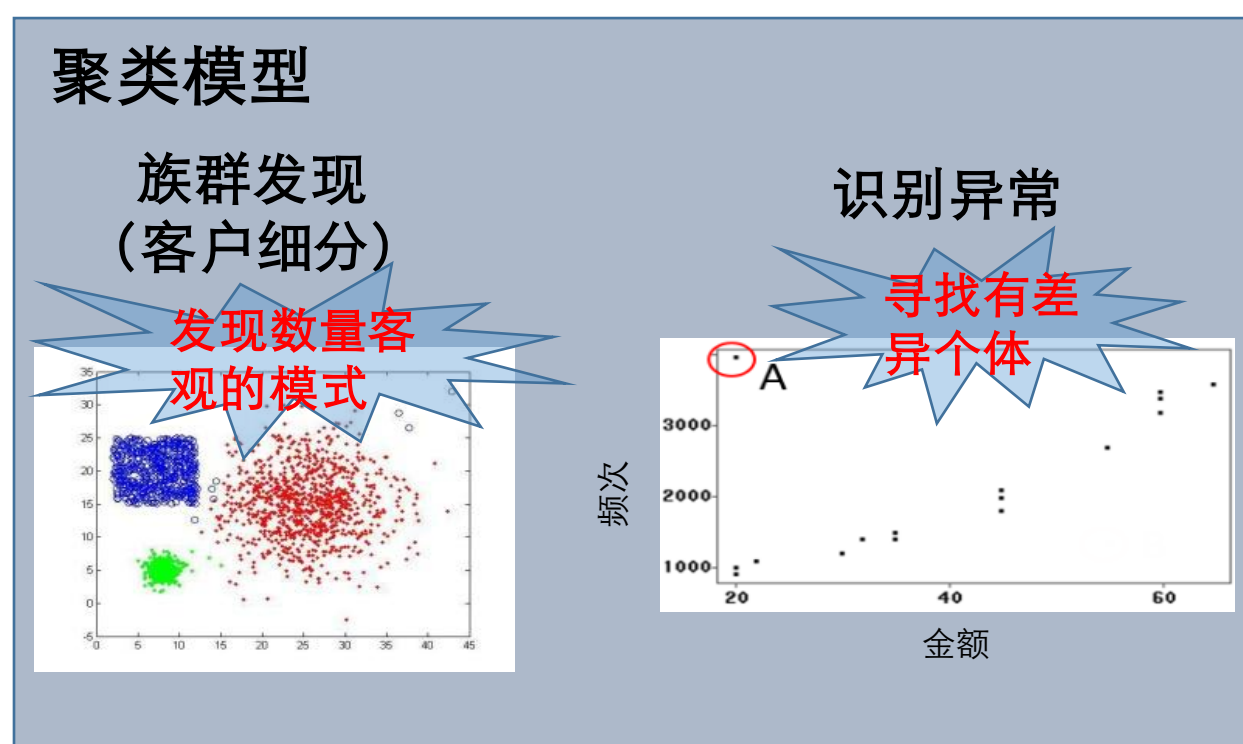
4 模型构建与评估



数据挖掘模型分类

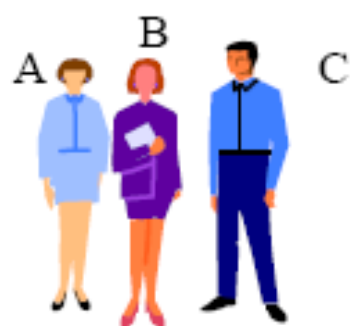


关联规则



时间序列

使用逻辑回归建立行为评分卡



行为评分



客户号	违约概率P
A	0.87
...	
B	0.36
...	
C	0.12

$$p(\text{default}) = \frac{e^{b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n}}{1 + e^{b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n}}$$

仅示例

变量名	回归系数	相关系数
BAL_NUM_P3	0.5528	0.2886
BAL_PCT_P6	-1.3811	-0.2636
CDT_LMT_AMT	-0.1421	-0.1413
CSM_CNT	-0.0616	-0.2674
DQT_LVL_CDE_3_M1_Dummy	1.0501	0.0483
LMT_AMT_PCT_P6	0.2308	0.1477
LST_FNL_DYS	0.0067	0.3369
LST_PMT_DYS	0.0038	0.2356
MTL_STS_CDE_MARR	-0.3511	-0.1331
PMT_OF_BAI_PCT_3	-0.3489	-0.3513

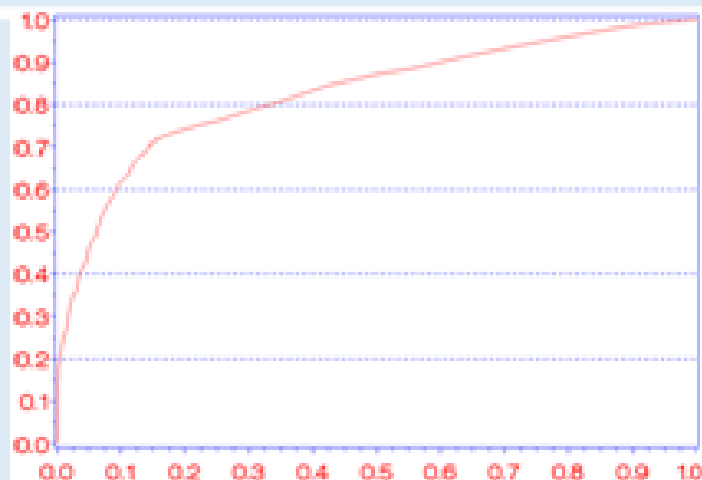
评估指标汇总

分类模型类型	统计指标
决策（Decisions、二分类器）	精确性/误分类/召回率/准确度 利润/成本
排序（Rankings、二分类器）	ROC曲线（一致性） Gini指数 K-S曲线（分离度） PR曲线 提升度曲线
估计（Estimates、回归）	误差平方均值 R方

评分卡模型的评估指标

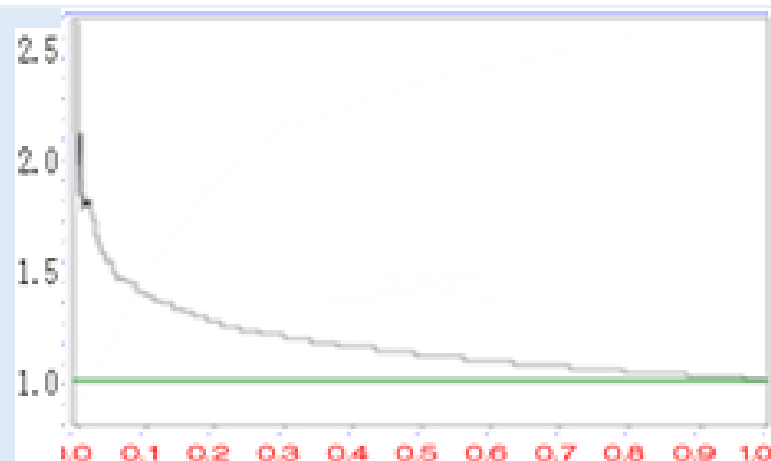
ROC曲线：用来描述模型分辨能力,对角线以上的图形越高模型越好

X:1-特异度
Y:灵敏度



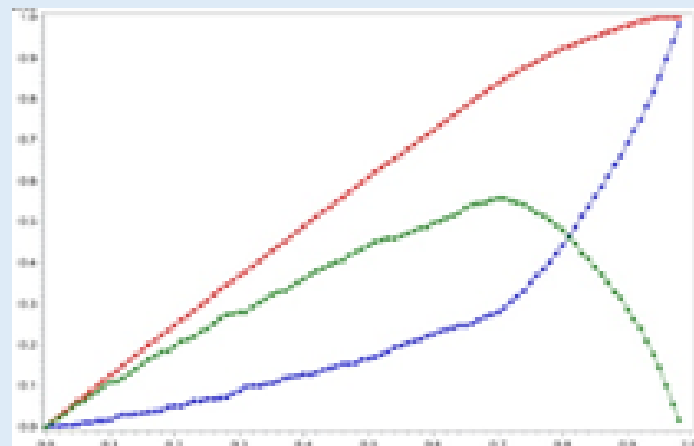
累积提升曲线：由于展示使用模型预测结果与随机情况下获取显性样本的能力比较

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

