Data Analysis Tools and Practice(Using R)

2021.05.11

信用评分

北京大学 软件与微电子学院
School of Software and Microelectronics, Peking University

Huiping Sun(孙惠平)

sunhp@ss.pku.edu.cn

课堂测试09

先用电脑完成
40分钟 然后誊抄纸上

- 1、下表是一个一个村庄儿童年龄和平均身高的统计数据
  - ✳ （1）画出平均身高height和年龄age关系的散点图
  - ✳ （2）建立回归模型并提取结果输出，在(1)中的图中表示生成的模型

平均身高与年龄			
年龄（月）	平均身高（厘米）	年龄（月）	平均身高（厘米）
18	76.1	24	79.9
19	77	25	81.1
20	78.1	26	81.2
21	78.2	27	81.8
22	78.8	28	82.8
23	79.7	29	83.5

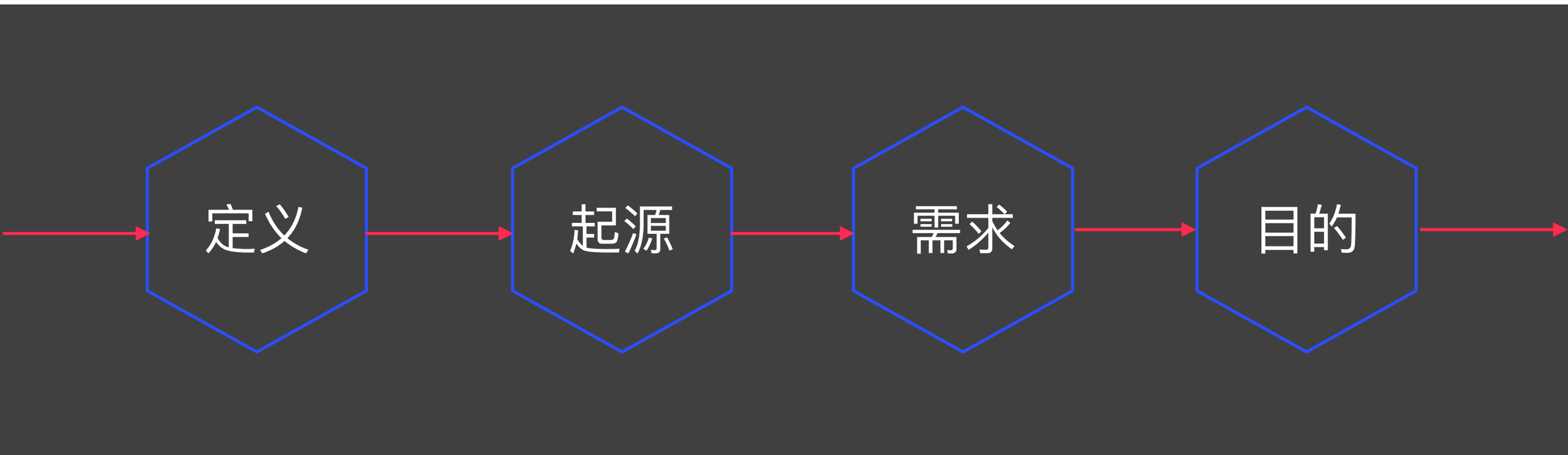
- 2、 revenue.txt中记录了财政收入(y)和第一产业GDP X_1、第二产业GDP X_2、第三产业GDPx_3、人口数X_4、社会消费品零售总额X_5、受灾面积X_6、等情况的统计数据。要求:写出多元线性回归模型。

课堂测试09

先用电脑完成
40分钟 然后誊抄纸上

- 3、某公司想要了解消费者购买牙膏时更追求什么样的目标,于是通过商场拦访对30个人进行访谈，用7级里克特量表询问他们对以下陈述的认同程度(即1表示非常不同意，7表示非常同意,V1:购买预防蛀牙的牙膏是重要的;V2:我喜欢使牙齿亮泽的牙膏；v3:牙膏应当保护牙龈；V4:我喜欢使口气清新的牙膏；V5:预防坏牙不是牙膏提供的一项重要功效；V6:购买牙膏时最重要的考虑是富有魅力的牙齿:
  - ✴ 将调查样本存储于文本文档 yagao.txt。请使用R函数factanal对数据进行因子分析，根据载荷系数矩阵，写出因子和原变量之间的线性关系式。
- 4、某地区农业生态经济系统的各区域单元相关指标数据在文本文件agriculture.txt中，使用R中的主成分分析的函数princomp选取更少的指标来描述该地区的农业生态经济系统。写出主成分和原变量之间的线性关系式。

信用评分简介



信用评分定义

- Credit Scoring is decision support systems used in consumer credit aims at assessment of potential borrowers and existing borrowers.

- Default probability is predicted from observed borrowers characteristics on the basis of the analysis of known performance pf previous customers.

- Risk / creditworthiness is usually measured by default probability.

Credit Scoring 信用评分起源

- Character
- Capacity
- Collateral
- Capital
- Conditions

- 自动化
- 快速
- 一致
- 客观
- 不安全

信用评分发展

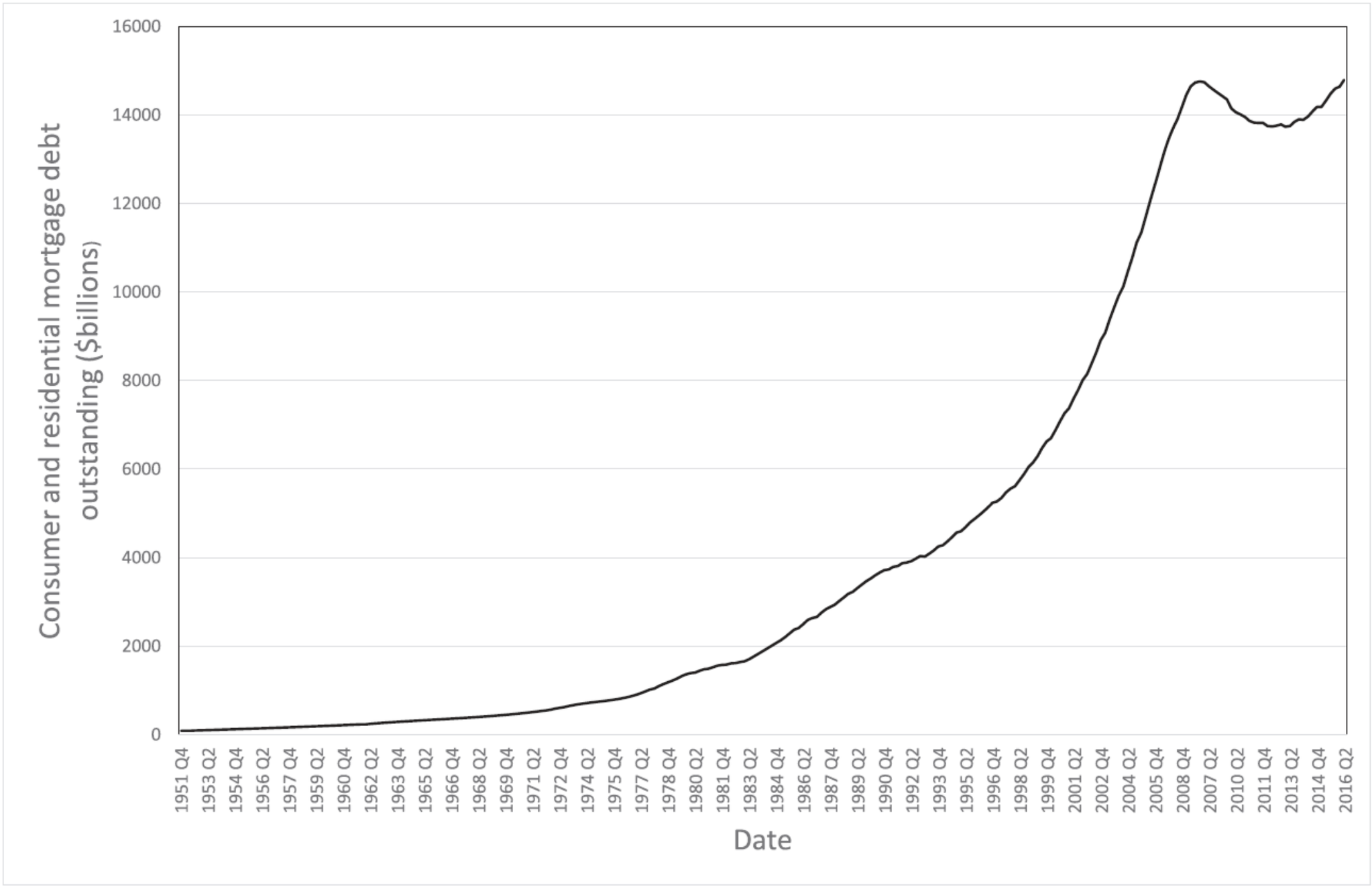


Figure 1.1. *U.S. household debt. Source: Board of Governors of Federal Reserve System.*

Credit Scoring

信用评分需求

个人贷款	企业贷款	中小微企业贷款
贷款额度小	贷款额度大	贷款额度?
贷款客户多	贷款客户少	贷款客户?
主要是预测	主要是因果	关注?
研究较少	大量研究	模型?
教科书少	大量教科书	研究?
管理＋数据	金融＋会计	学科?

Credit Scoring

信用评分目的

申请客户

- 债务违约
- 产品使用
- 用户流失

已有客户

- 信用更新
- 交叉销售
- 再次申请

问题客户

- 预警
- 催收
- 坏账

风险定价

抵押担保

利润评分

客户评价

资本充足

风险度量

IFRS9

Credit Scoring II

为什么需要信用评分

风险评估

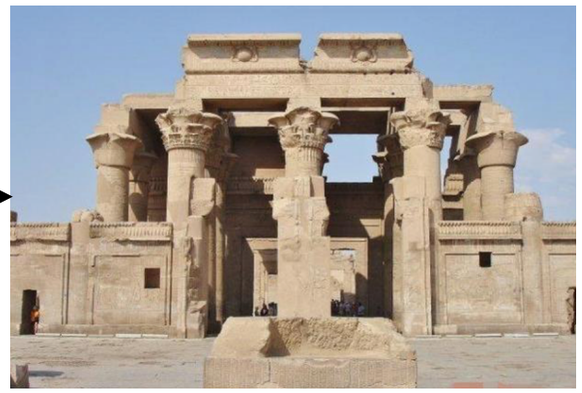
凭直觉	凭关系	凭信誉	封闭环境
担保抵押	偿还能力	借方特征	使用目的
长期训练	经验丰富	本地Office	风险保守

信用评分

销售产品	业务拓展	利益最大化	业务数量
电话公司	电话购物	电力公司	供水公司
信用卡	分期	抵押	透支

Credit Scoring

信用历史



...

Credit Scoring 信用评分历史

…… ……	第一个专家系统
1950	FICO,
1975	Equal Credit Opportunity Acts
1980	Bank, Logistic regression
1992	Credit Scoring Conference, CSCC
2000	巴塞尔协议, 1988, 2005, 2010
2008	次级房贷危机

Credit Scoring

信用评分例子

评分卡

神经网络

随机森林

支持向量机

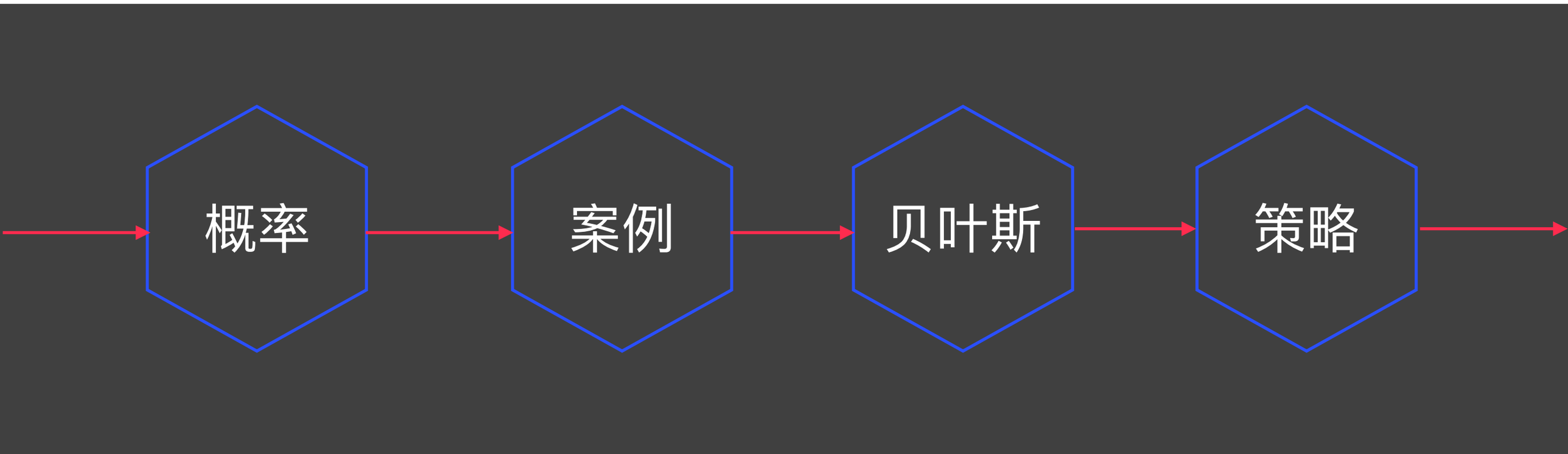
种族 | 地域 | 性别 | 年龄

健康 | 保险 | 工作 | 房屋

籍贯 | 孩子 | 婚姻 | 负担

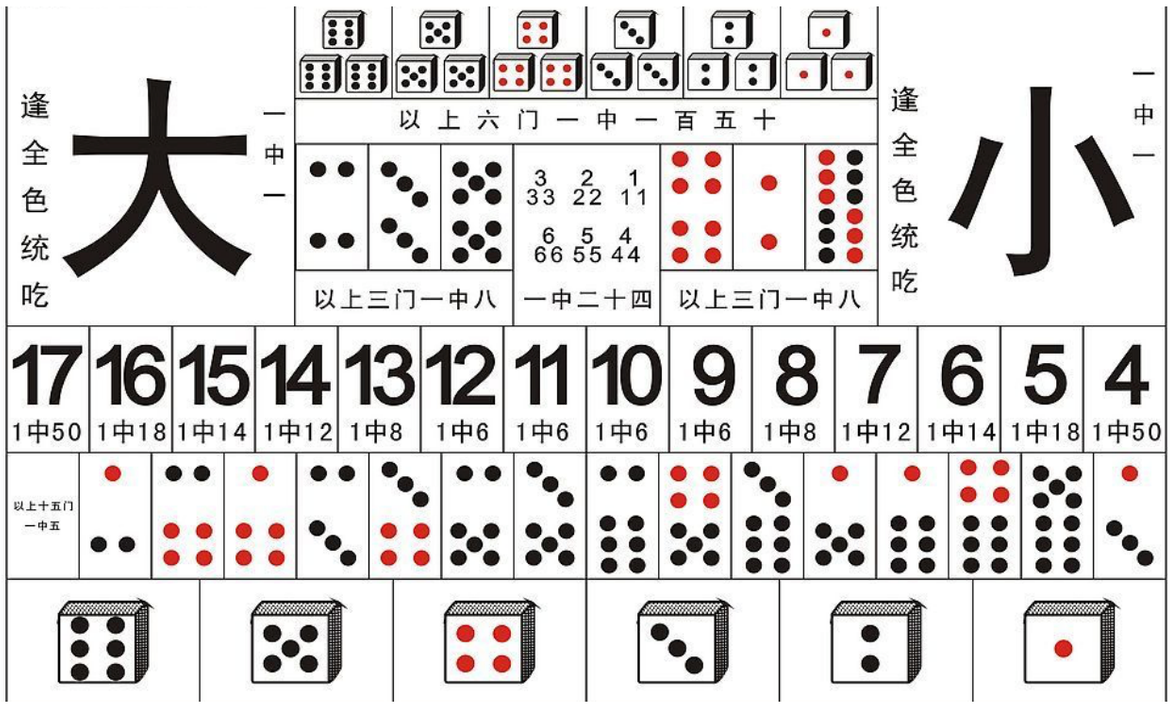
现金 | 账户 | 支票 | 负债

评分卡概述



Credit Scoring

概率和信用评分



Good

Bad

重复实验N次

R次结果是Good

P(G)、P(B)

P＝G/(G+B)

G: B＝P/(I－P)

O(G)=P(G):P(B)

Credit Scoring 信用评分案例

银行有8000客户申请贷款，一年后7000是好的，1000是坏的，好客户平均收益一千，坏客户平均损失一万

好客户弥补坏客户、损益平衡点：10:1

群体好坏概率：7:1

Marital Status:

	Good	P(x|G)	Bad	P(x|B)	Marginal Odds	
Married	4900	0.7	400	0.4	49 : 4	12.25:1
Not married	2100	0.3	600	0.6	21: 6	3.5:1
Total	**7000**	**1**	**1000**	**1**		

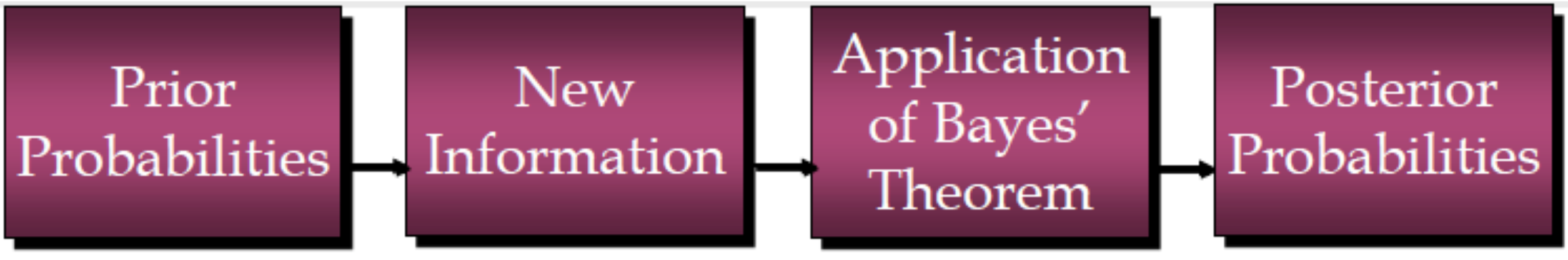
Marginal Odds of Married = 0.7 : 0.4 × 7: 1 = 12.25

Marginal Odds of NM = 0.3 : 0.6 × 7: 1 = 3.5

Information Odds

NB: Marginal Odds = Information Odds × Population Odds

信用评分中的贝叶斯理论



Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ be characteristics (variables) of the borrowe such as age, marital status, etc.

$\mathbf{x} = (x_1, x_2, \ldots, x_m)$ be outcomes/ attributes of characteristics.

P(G) and *P(B)* are *prior probabilities.*

Posterior probabilities:

$P(G|\mathbf{x})$ is the probability of being Good given certain attributes

$P(B|\mathbf{x})$ is the probability of being a Bad customer given certain attributes

工作时间

Marital Status:

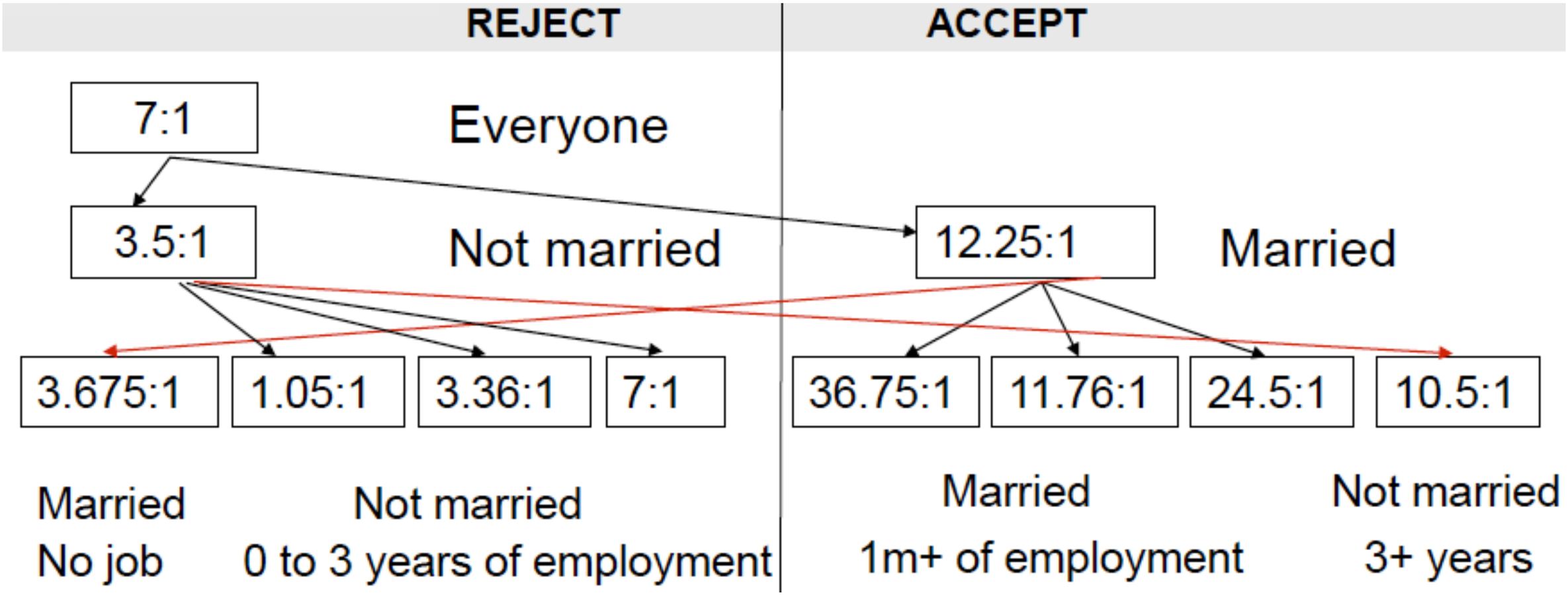
	Good	$P(x|G)$	Bad	$P(x|B)$	**Marginal Odds, $O(G|x)$**	
Married	4900	0.7	400	0.4	49 : 4	12.25:1
Not married	2100	0.3	600	0.6	21: 6	3.5:1

Time in Employment :

0	1050	0.15	500	0.5	105 : 50	2.1:1
up to 6 m	1680	0.24	250	0.25	168 : 25	6.72:1
6m - 3y	1960	0.28	140	0.14	196 : 14	14:1
3y+	2310	0.33	110	0.11	231 : 11	21:1
Total	**7000**		**1000**			

贝叶斯评分卡

Pop Odds × Info Odds (Char 1) × … × Info Odds (Char n)



Odds of Married and No Job = $7/1 \times 0.7/0.4 \times 0.15/0.5$ =

= $7 \times 1.75 \times 0.3 = 3.675$

独立性假设

Odds of Not Married and 3+ years of employment = ?

= $7/1 \times 0.3/0.6 \times 0.33/0.11 = 7 \times 0.5 \times 3 = 10.5$

贝叶斯评分卡

Time at current address	Less than 6 months	6m – 2 years	2 – 6 years	6 - 10 years	10 + years	Unknown	
	0	3	6	13	25	0	
Resi-dential Status	Owner	Tenant	With parents	Unknown			
	15	5	2	0			
Banking	Current account	Saving account	Current and saving	No account	Unknown		
	5	10	14	0	0		
Occupa-tion	Retired	Full-time	Part-time	Self-employed	Student	Other	Un-known
	21	16	7	6	5	10	0
Age	18-25	26-31	32-40	41-54	55+	Unknown	
	5	10	15	20	25	0	

属性 性能 风险

历史

评分

Credit Scoring 策略

36.75:1	Married	3+ years of employment
24.5:1	Married	6m – 3y of employment
11.76:1	Married	up to 6m of employment
10.5:1	Not married	3+ years of employment
7:1	Not married	6m – 3y of employment
3.675:1	Married	No job
3.36:1	Not married	up to 6m of employment
1.05:1	Not married	No job

数据

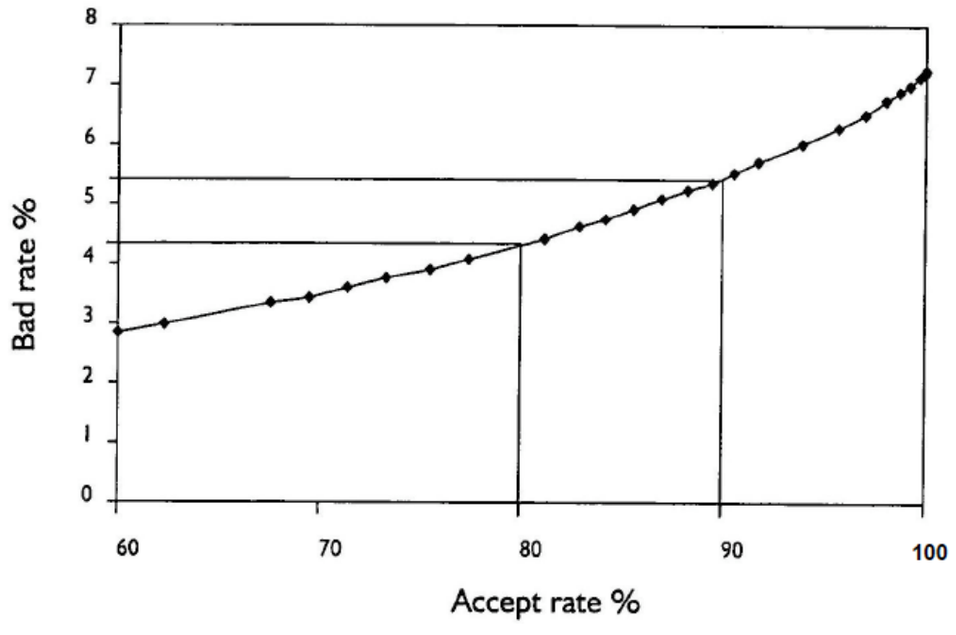
易获取

自动化

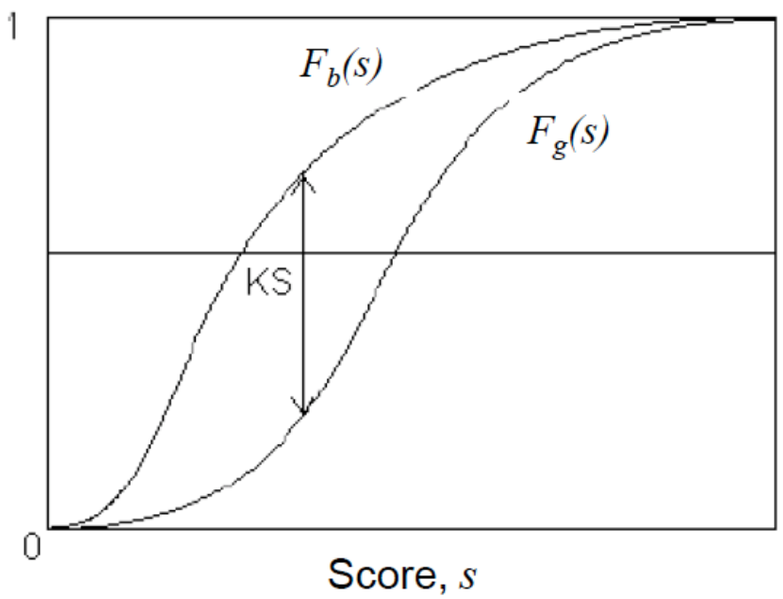
廉价

假设：未来和过去相似

信用评分是预测，不是可解释的



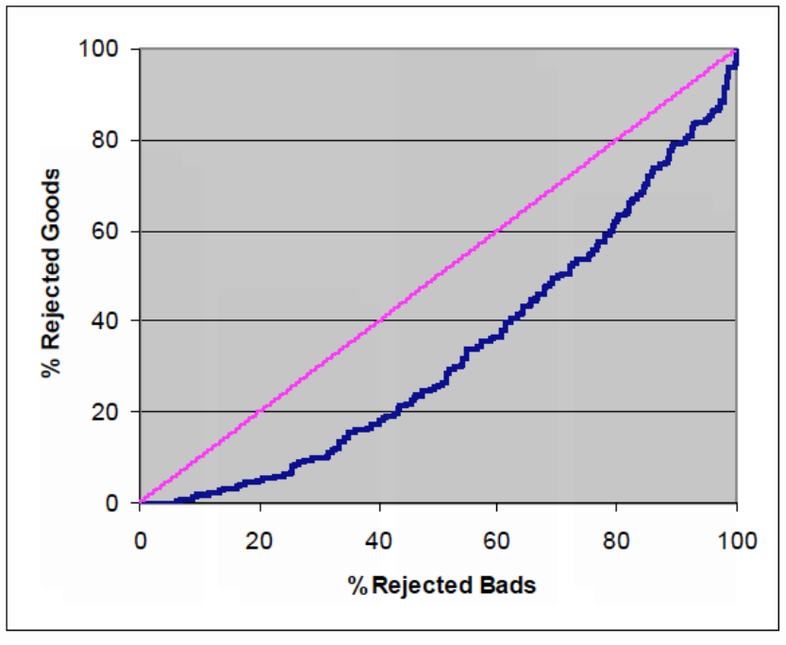
评价



KS

基尼系数

$F_g(s)$

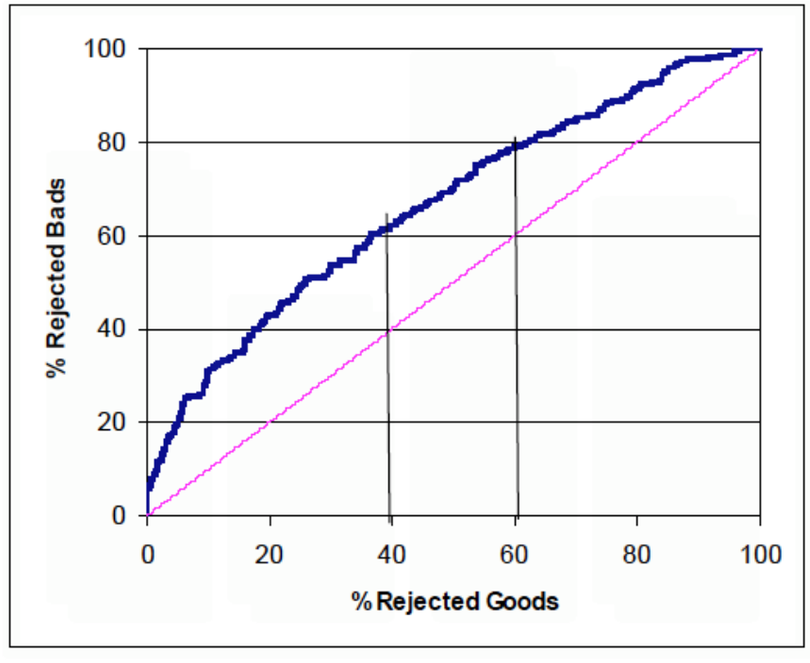


$F_b(s)$

or Lorenz diagram

ROC

$F_b(s)$



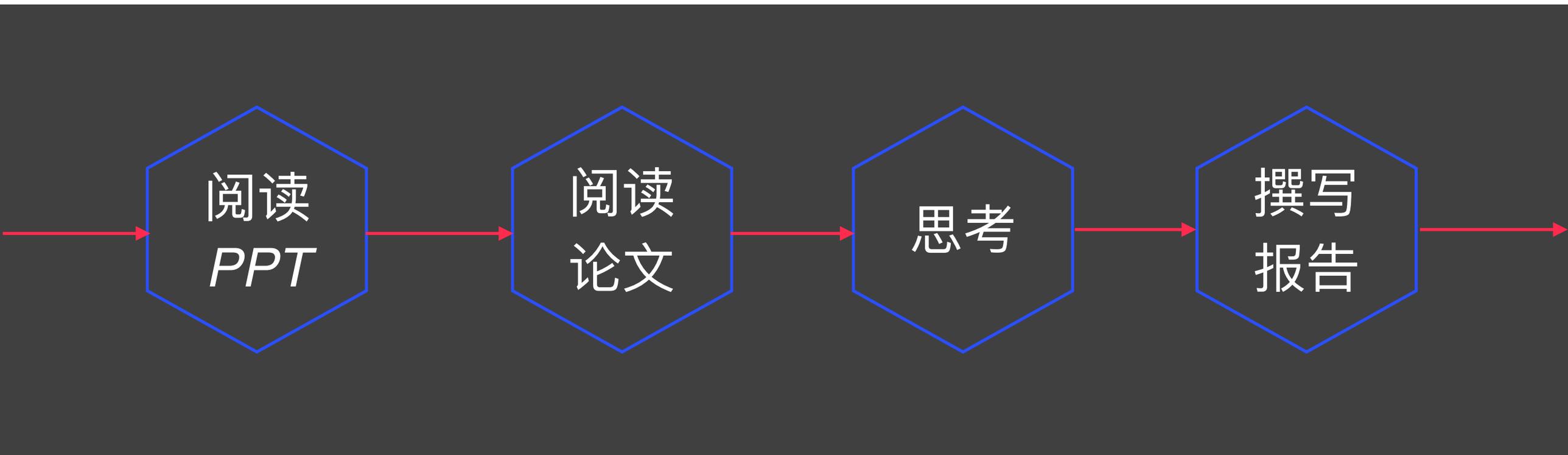
$F_g(s)$

提问时间！

孙惠平

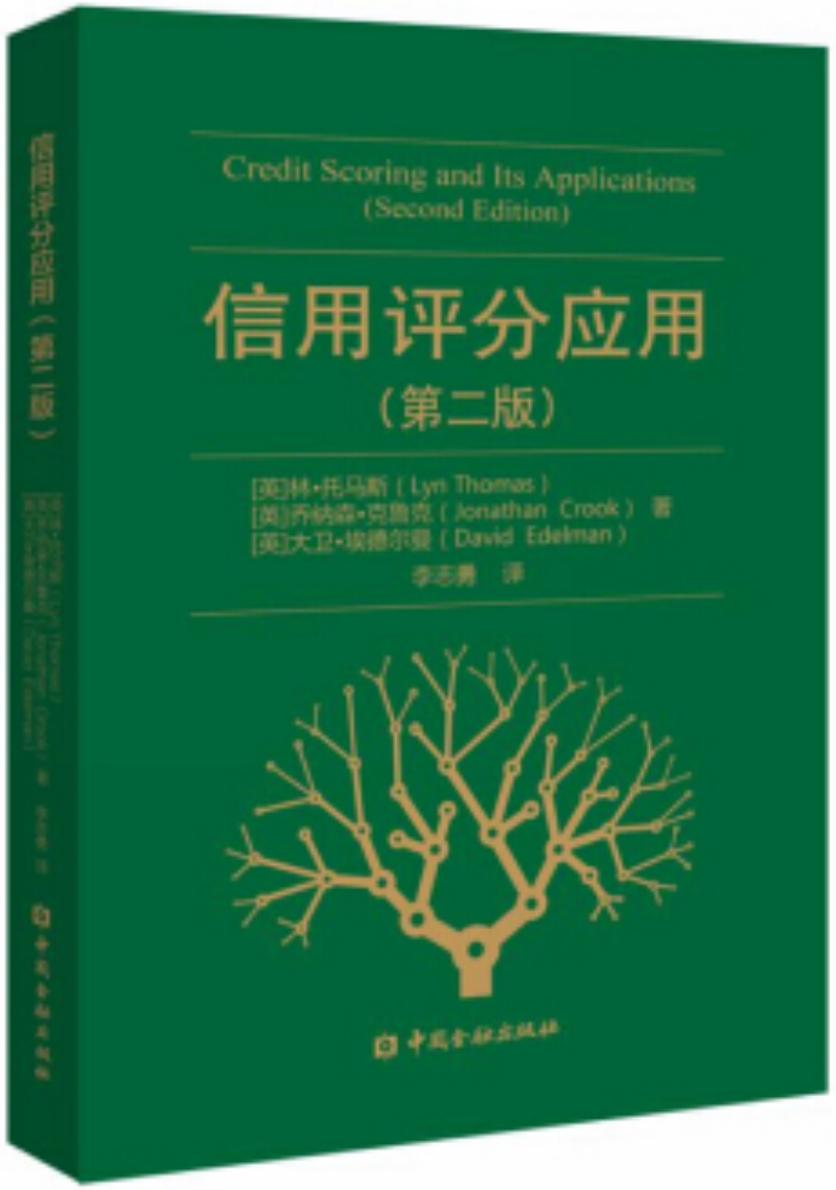
sunhp@ss.pku.edu.cn

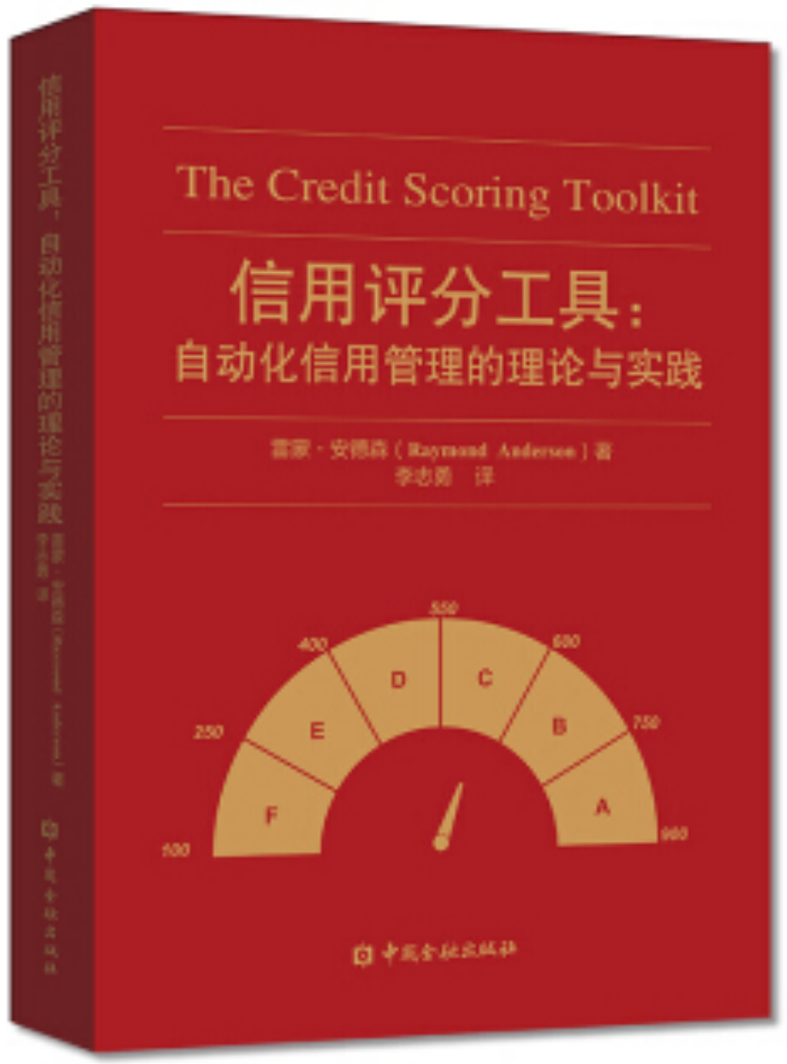
课后作业



Homework

推荐书





Homework

阅读教材

练习 - 0032

Give Me Some Credit 数据

https://www.kaggle.com/c/GiveMeSomeCredit

数据描述

缺失值处理

异常值处理

好坏样本选择

特征选择

特征工程

模型构建

逻辑回归模型

模型评测

Lending Club数据

提交代码和报告

谢谢！

孙惠平

sunhp@ss.pku.edu.cn