Data Analysis Tools and Practice(Using R)

2021.05.18

信用评分II

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

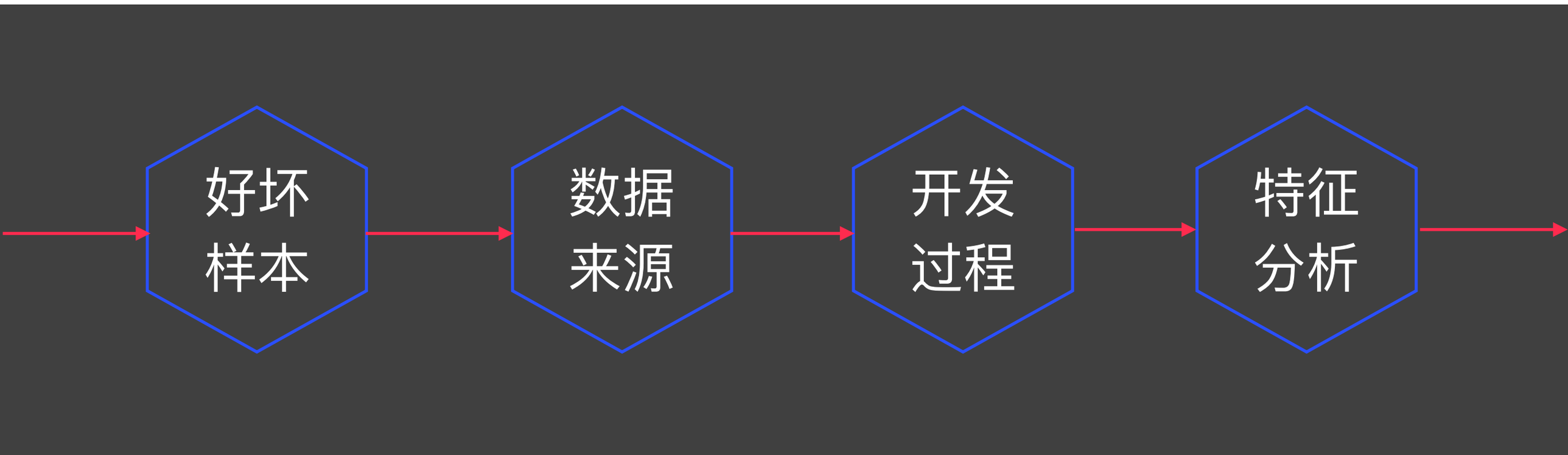
课堂测试10

先用电脑完成
40分钟 然后誊抄纸上

- 1、针对Give Me Some Credit 数据
  - ✳ （1）通过可视化开分析缺失值和异常值
  - ✳ （2）处理缺失值和异常值
  - ✳ （3）分析变量的相关性
  - ✳ （4）通过分箱、WOE和IV来检查各变量预测能力
  - ✳ （5）变量和特征选择
  - ✳ （6）用逻辑回归建立一个模型
  - ✳ （7）检验模型有效性（F1、ROC）

可以使用
ScoreCard包

评分卡建模



好坏样本

Good

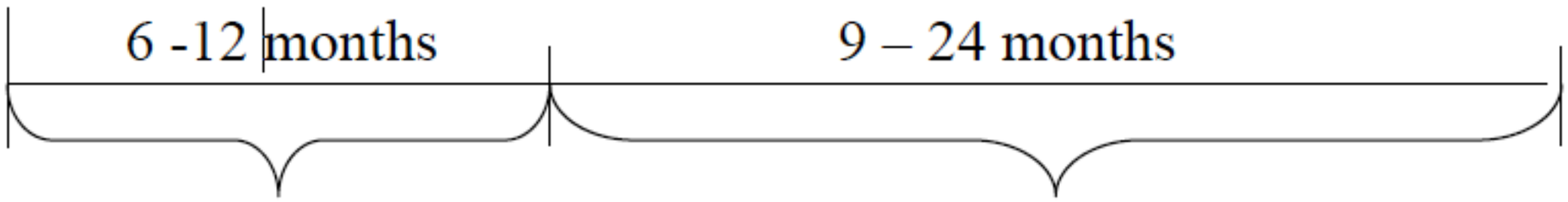
各自1500

Bad

GD?

最好20000-50000，所有的坏用户

随机性 | 正确性 | 实时性 | 完整性 | 合法性



季节

长度

数据来源

申请表

年龄

婚姻

地址

工作

征信报告

账户

抵押

司法

借贷

银行纪录

账户

流水

贷款

新数据

社交媒体

手机使用

短信使用

地理位置

Credit Scoring

评分卡例子

Residential status	
Attribute	Score
Owner	30
Tenant	17
Living with Parents	20
Other	0

Age	
Attribute	Score
18–25	5
26–35	10
36–43	15
44+	20

Loan purpose	
Attribute	Score
New Car	31
Used Car	9
Home Improvement	14
Other	0

Time at present address (years)	
Attribute	Score
< 2	4
2–5	9
6–11	16
12+	18

小于50 | 大于等于50

小于48 | 48到53 | 大于等于54

20岁+和父母住+买车+2年居住: **43** (5+20+9+9)

55岁+自有住房+女儿婚礼+17年居住: **68** (30+20+0+18)

SuperPass

SuperFail

风险定价

准入条件

提高风险管理

减少业务花费

丰富客户服务

获取一致性

数据采集

Table 2.2. *Some reasons for data collection.*

Purpose	Examples
To identify customer	Name, address, date of birth
To be able to contract with customer	Name, address, date of birth, loan amount, repayment schedule, interest rate
To process/score the application	Scorecard characteristics
To get a credit bureau report	Name, address, date of birth, previous address
To assess marketing effectiveness	Campaign code, date of receipt of application, application channel, loan amount, gender, date of birth, address
To effect interbank transfers of money	Bank account number, bank branch details
To develop scorecards	Any information legally usable in a scorecard (laws vary from country to country)

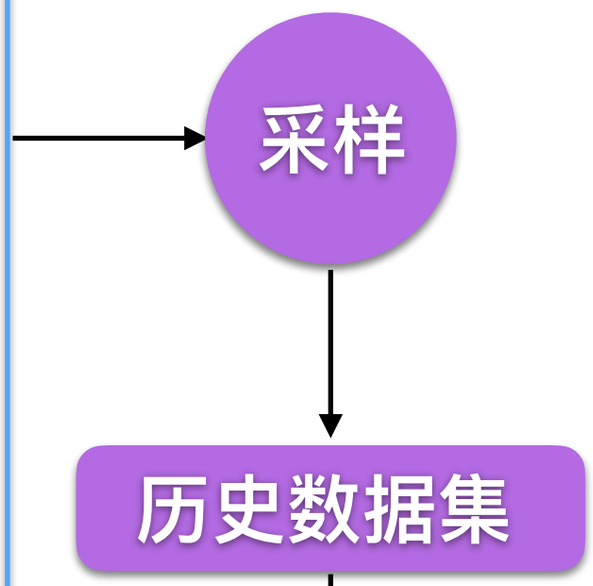
申请数据　征信数据　自有数据　第三方数据　新数据源

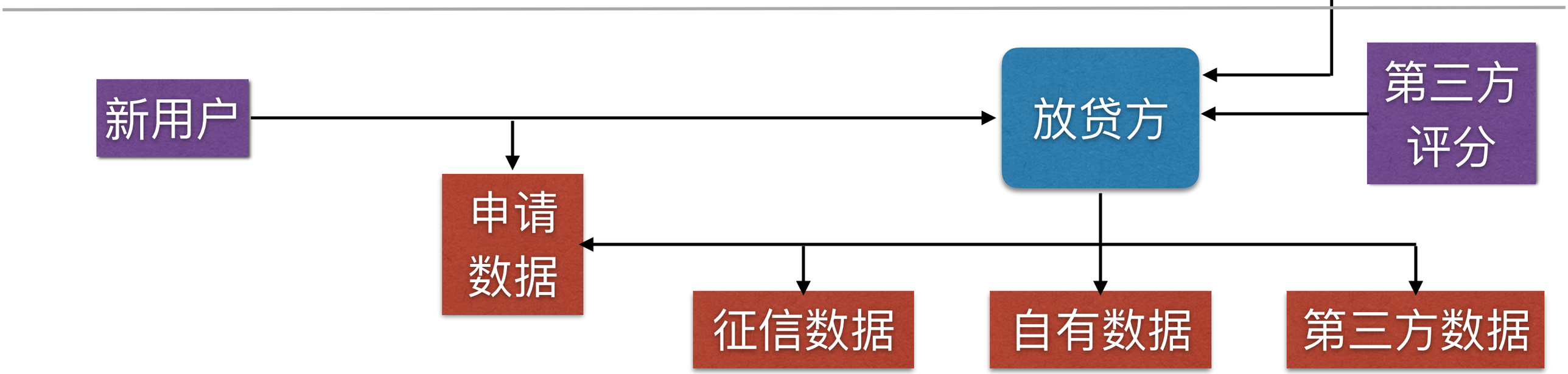
准确性　可用性　法律要求　文化影响　数据保护

数据处理

历史信息


用户1
属性1
属性2
…
属性n
评分
用户2
属性1
属性2
…
属性n
评分
…
…
…
…
…
…
用户m
属性1
属性2
…
属性n
评分


采样
历史数据集


新用户
放贷方
第三方
评分
申请
数据
征信数据
自有数据
第三方数据

好坏客户

Good

一次未还款

不能确定好坏的客户

Good

Bad

损失

Bad

没有足够经历的客户

小于6个月还款记录

三次以上未还款　　抵押

两次连续未还款　　贷款

Bad

Good

损失

贝叶斯

$$p(\mathbf{x}|G) = \frac{\text{Prob(applicant is Good and has attributes } \mathbf{x})}{\text{Prob(applicant is Good)}}.$$

$$p(G|\mathbf{x}) = \frac{p(\mathbf{x}|G)p_G}{p(\mathbf{x})}.$$

$$p(G|\mathbf{x}) = \frac{\text{Prob(applicant has attributes } \mathbf{x} \text{ and is Good)}}{\text{Prob(applicant has attributes } \mathbf{x})},$$

$$s(\mathbf{x}) = \ln\left(\frac{p(G|\mathbf{x})}{p(B|\mathbf{x})}\right) = \ln\left(\frac{p_G p(\mathbf{x}|G)}{p_B p(\mathbf{x}|B)}\right) = \ln\left(\frac{p_G}{p_B}\right) + \ln\left(\frac{p(\mathbf{x}|G)}{p(\mathbf{x}|B)}\right)$$

$$\text{or } s(\mathbf{x}) = s_{pop} + \text{woe}(\mathbf{x}).$$

$$p(\mathbf{x}|G) = p(x_1|G)p(x_2|G)\ldots p(x_p|G) \text{ and } p(\mathbf{x}|B) = p(x_1|B)p(x_2|B)\ldots p(x_p|B).$$

$$s(\mathbf{x}) = \ln\left(\frac{p(G|\mathbf{x})}{p(B|\mathbf{x})}\right) = \ln\left(\frac{p_G p(\mathbf{x}|G)}{p_B p(\mathbf{x}|B)}\right) = \ln\left(\frac{p_G}{p_B}\right) + \ln\left(\frac{p(x_1|G)}{p(x_1|B}\right)$$
$$+ \ln\left(\frac{p(x_2|G)}{p(x_2|B)}\right) + \cdots + \ln\left(\frac{p(x_p|G)}{p(x_p|B)}\right),$$

假设变量独立

例子

	Owner		Not owner		Total	
Age	G	B	G	B	G	B
30−	100	10	200	40	300	50
30+	500	10	100	40	600	50
Total	600	20	300	80	900	100

$$s_{pop} = \ln(900/100) = 2.20,$$

$$\text{woe}(30-) = \ln\left(\frac{300/900}{50/100}\right) = \ln(2/3) = -0.41,$$

$$\text{woe}(30+) = \ln\left(\frac{600/900}{50/100}\right) = \ln(4/3) = 0.29,$$

$$\text{woe}(\text{owner}) = \ln\left(\frac{600/900}{20/100}\right) = \ln(10/3) = 1.20,$$

$$\text{woe}(\text{not owner}) = \ln\left(\frac{300/900}{80/100}\right) = \ln(5/12) = -0.88,$$

$$s(\mathbf{x}) = s_{pop} + \text{woe}(x_1) + \text{woe}(x_2).$$

线性回归

$$w_0 + w_1 X_1 + w_2 X_2 + \cdots + w_p X_p = \mathbf{w}^* \cdot \mathbf{X}^{*T},$$
$$\text{where } \mathbf{w}^* = (w_0, w_1, w_2, \ldots, w_p),\ \mathbf{X}^* = (1, X_1, X_2, \ldots, X_p),$$

$$p_i = w_0 + x_{i1} w_1 + x_{i2} w_2 + \cdots + x_{ip} w_p \quad \text{for all } i.$$

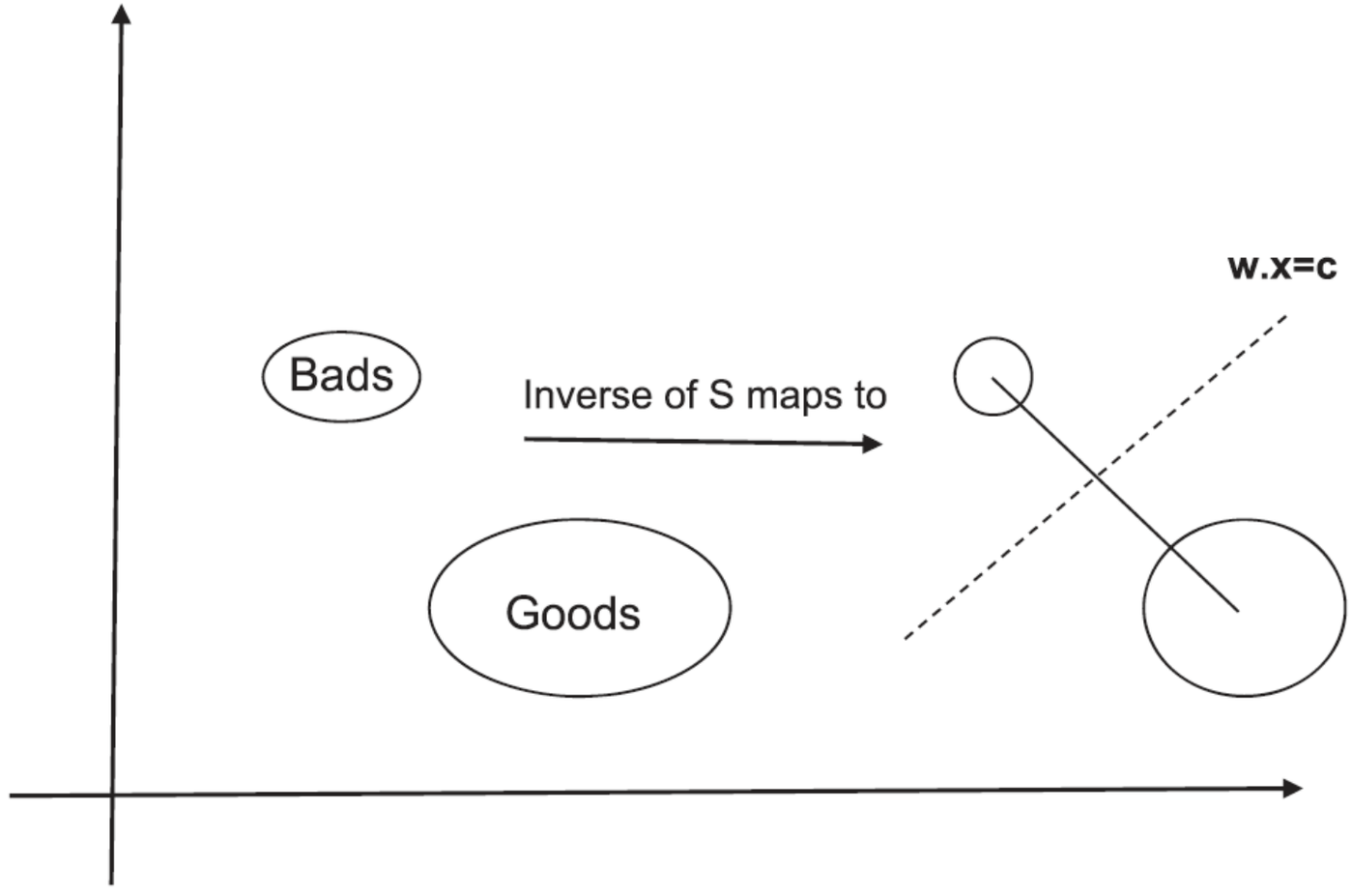


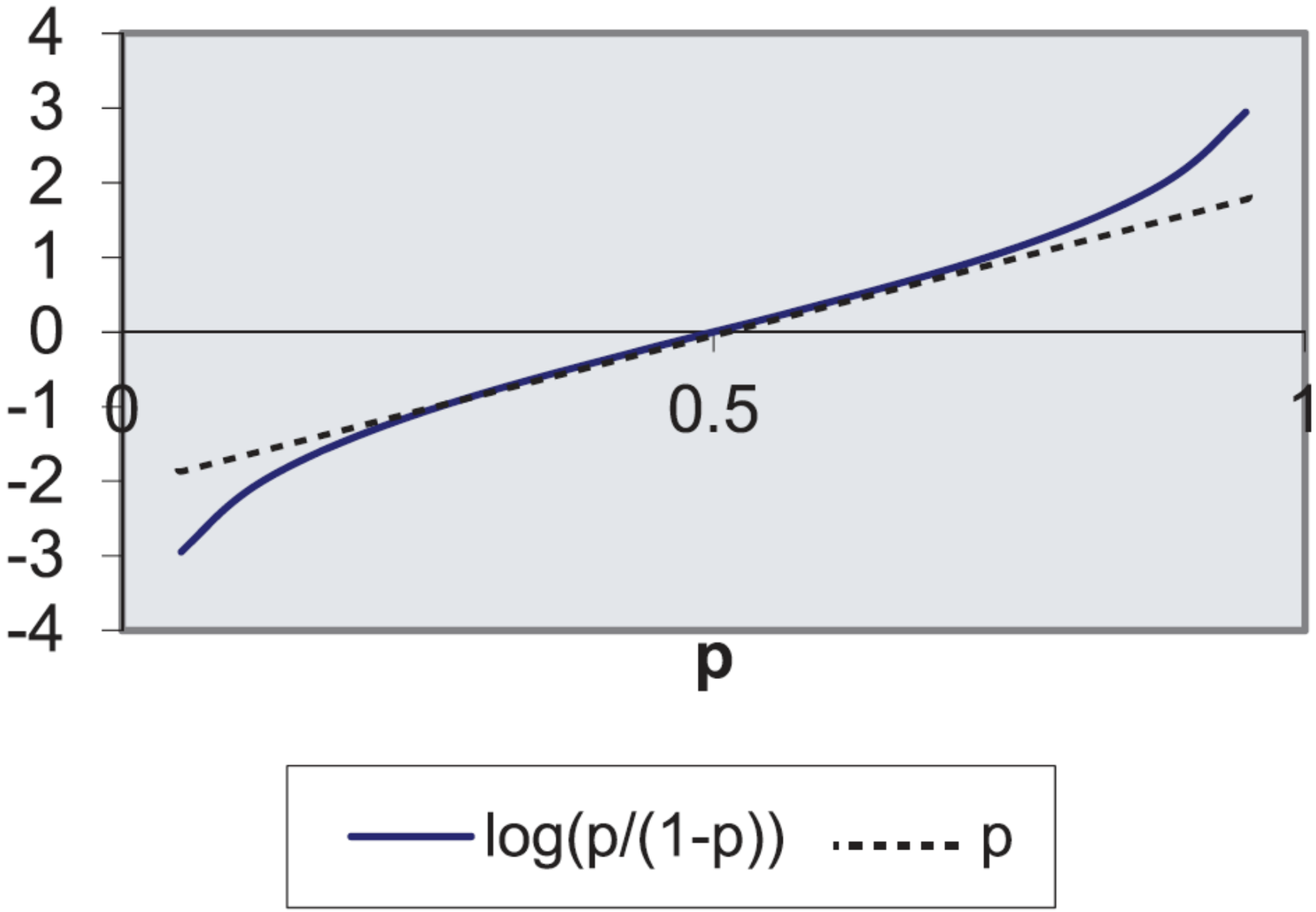
Figure 3.2. *Line corresponding to scorecard.*

连续变量

Credit Scoring

逻辑回归

$$s(\mathbf{x}) = \log\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p = \mathbf{w} \cdot \mathbf{x}^T.$$



离散变量

Figure 3.3. *Graph of* $\log(p/(1-p))$ *and* $ap + b$.

分类树

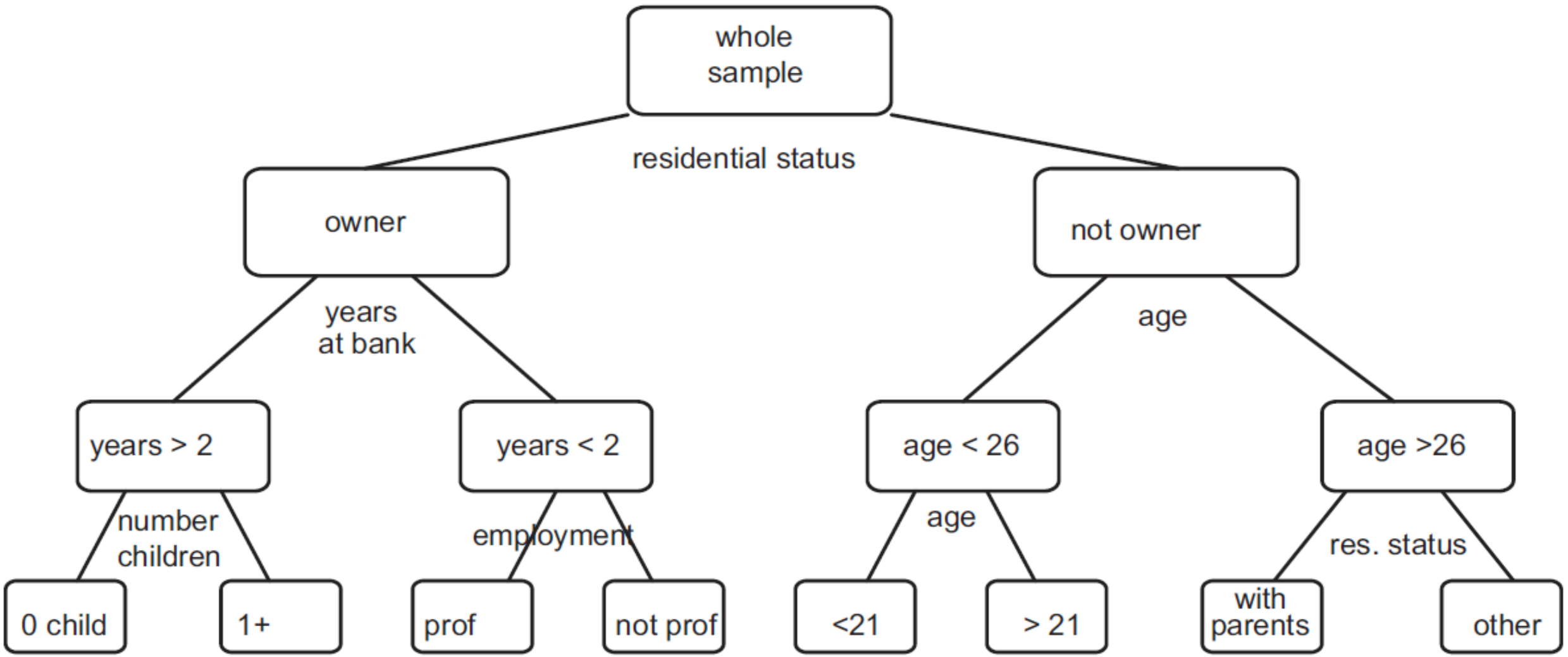


Figure 3.4. *Classification tree.*

划分规则　停止规则　分配规则

其余方法

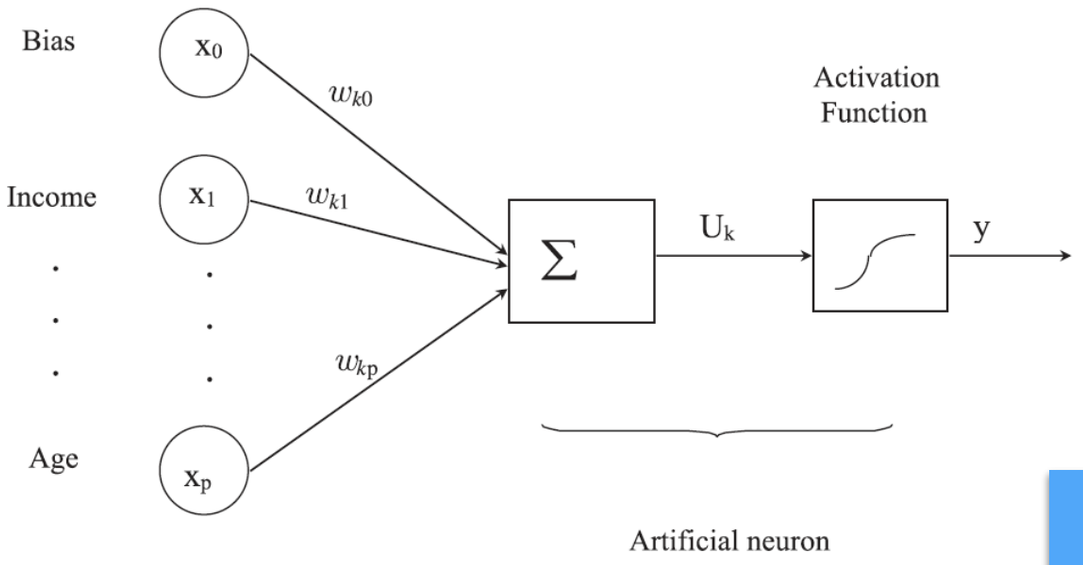


Figure 4.3. *A single-layer neural network.*

神经网络

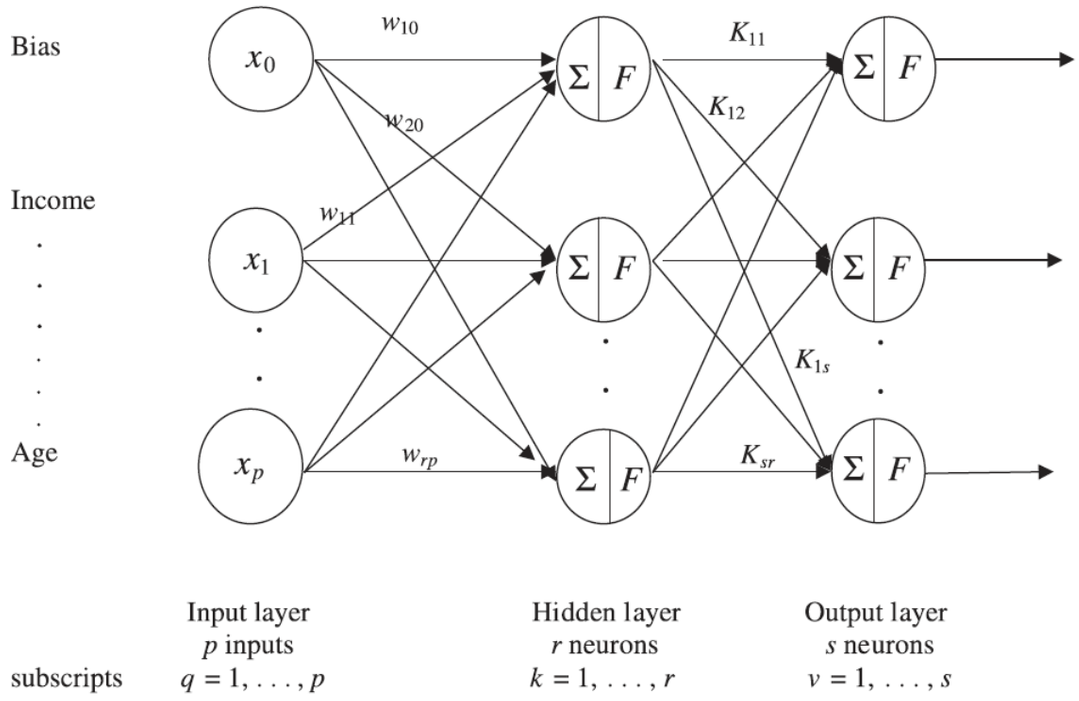


Figure 4.5. *A multilayer perceptron.*

SVM

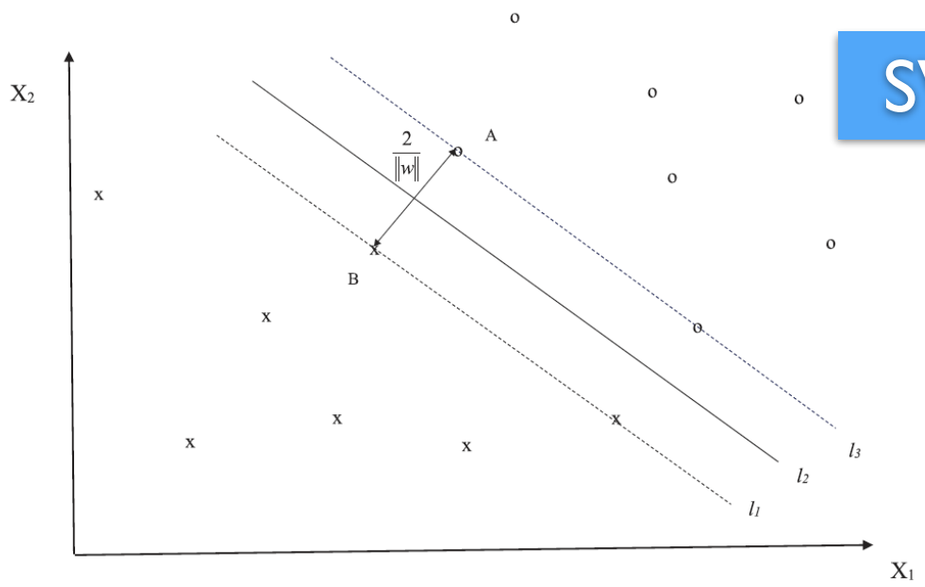


Figure 4.10. *Support vectors for separable classes.*

KNN

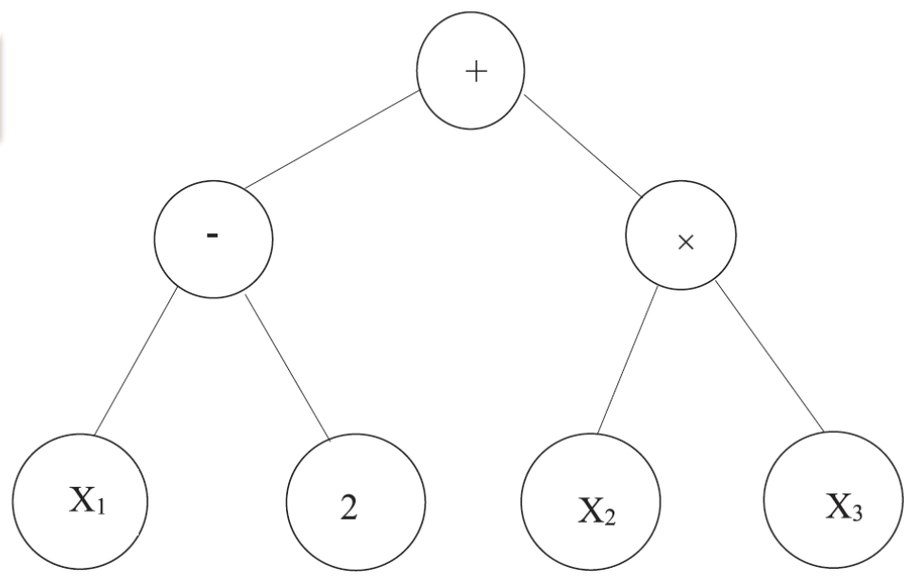


Figure 4.13. *An example of a genetic program tree.*

Credit Scoring

特征变量

Table 6.1. *Characteristics in three application forms.*

Characteristic	Finance house	U.S. credit card	U.K. credit card
Zip code/postal code	X	X	X
Time at address	X	X	X
Residential status	X	X	X
Occupation	X	X	X
Time at employment	X	X	X
Appl. monthly salary	X	X	X
Other income	X	X	
No. dependents	X	X	
No children	X	X	X
Checking account/current account	X	X	X
Savings account	X	X	
Credit card	X	X	X
Store card	X	X	X
Date of birth			X
Telephone		X	X
Monthly payments	X		
Total assets	X		
Age of car	X		

缺失 核实 填写 职业

歧视： 种族、肤色、宗教、血统、性别、婚姻状态、年龄

Credit Scoring

征信报告

中国人民银行
征信中心
CREDIT REFERENCE CENTER,
THE PEOPLE'S BANK OF CHINA

NO.B2016081200010323514

企业信用报告

（自主查询版）

评分

名称：报告样本公司

机构信用代码：G1111010811677 9***

中征码：1101080000000***

报告日期：2016-08-12

信贷记录

这部分包含您的信用卡、贷款和其他信贷记录。金额类数据均以人民币计算，精确到元。

信息概要 逾期记录可能影响对您的信用评价。

	资产处置信息	保证人代偿信息
笔数	1	2

	信用卡	住房贷款	其他贷款
账户数	7	3	4
未结清/未销户账户数	4	2	3
发生过逾期的账户数	4	1	1
发生过90天以上逾期的账户数	4	0	0
为他人担保笔数	0	0	1

资产处置信息

1. 2010年11月8日东方资产管理公司接收债权，金额400,000。最近一次还款日期为2011年1月8日，余额20,000。

保证人代偿信息

1. 2008年10月5日富登融资租赁担保公司进行最近一次代偿，累计代偿金额400,000。最近一次还款日期为2011年1月8日，余额20,000。
2. 2009年6月21日平安保险公司进行最近一次代偿，累计代偿金额200,000。最近一次还款日期为2011年4月5日，余额135,000。

信用卡

发生过逾期的贷记卡账户明细如下：

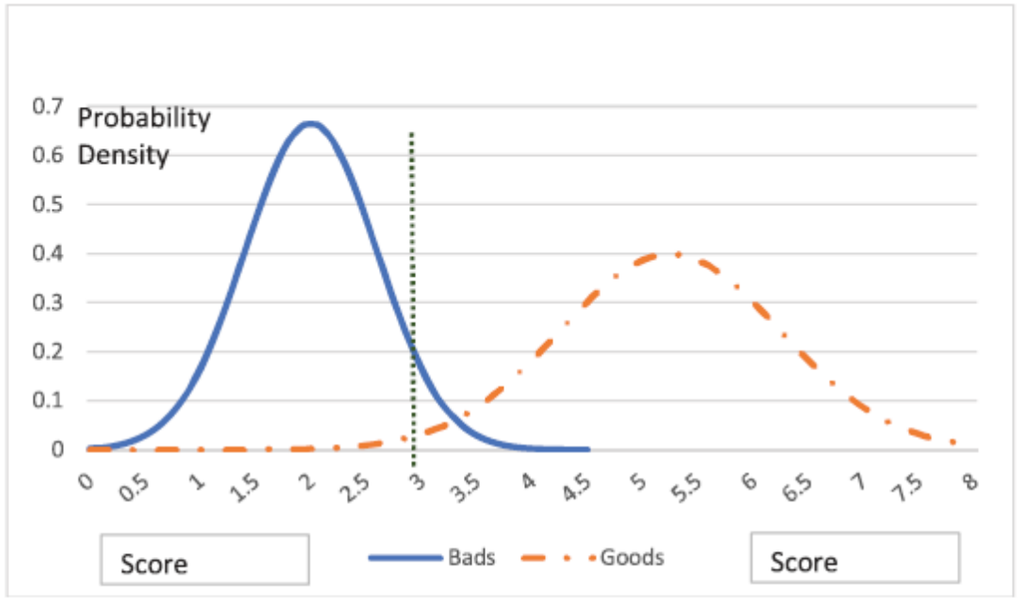
1. 2004年8月30日中国工商银行北京分行发放的贷记卡（人民币账户）。截至2010年10月，信用额度10,000，已使用额度500，逾期金额500。最近5年内有11个月处于逾期状态，其中5个月逾期超过90天。
2. 2003年4月1日中国民生银行信用卡中心发放的贷记卡（人民币账户），2009年12月销户。最近5年内有7个月处于逾期状态，其中3个月逾期超过90天。
   2010年3月，该机构声明：该客户委托XX房地产开发公司偿还贷款，因开发公司不按时还款导致出现多次逾期。

透支超过60天的准贷记卡账户明细如下：

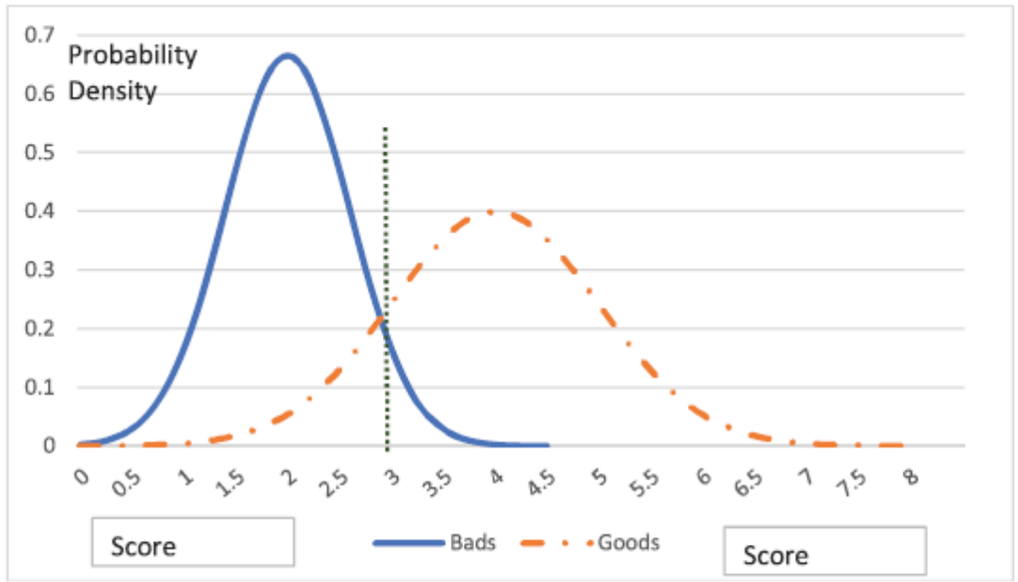
3. 2007年6月30日中国银行北京分行发放的准贷记卡（人民币账户）。截至2010年10月，信用额度10,000，透支余额5,000。最近5年内有6个月透支超过60天，其中3个月透支超过90天。
4. 2006年3月10日上海浦东发展银行北京分行发放的准贷记卡（人民币账户），2009年12月销户。最近5年内有20个月透支超过60天，其中16个月透支超过90天。

公开信息 | 查询信息 | 贷款信息 | 违约信息 | 账户信息 | 汇总信息 | 司法信息 | 其余信息

好坏分类



(a)



(b)

Figure 8.1. (a) *Means of Goods and Bads are apart.* (b) *Means of Goods and Bads are close.*

拒绝推断

Good

Bad

被拒绝

核准

三组

增补

外推

坏客户

增加样本

改变策略

新产品

好客户

其余应用

预先审核

预先批准

防范欺诈

住房贷款

小企业

风险定价

交易授权

债务偿还

坏账

出口担保

直销

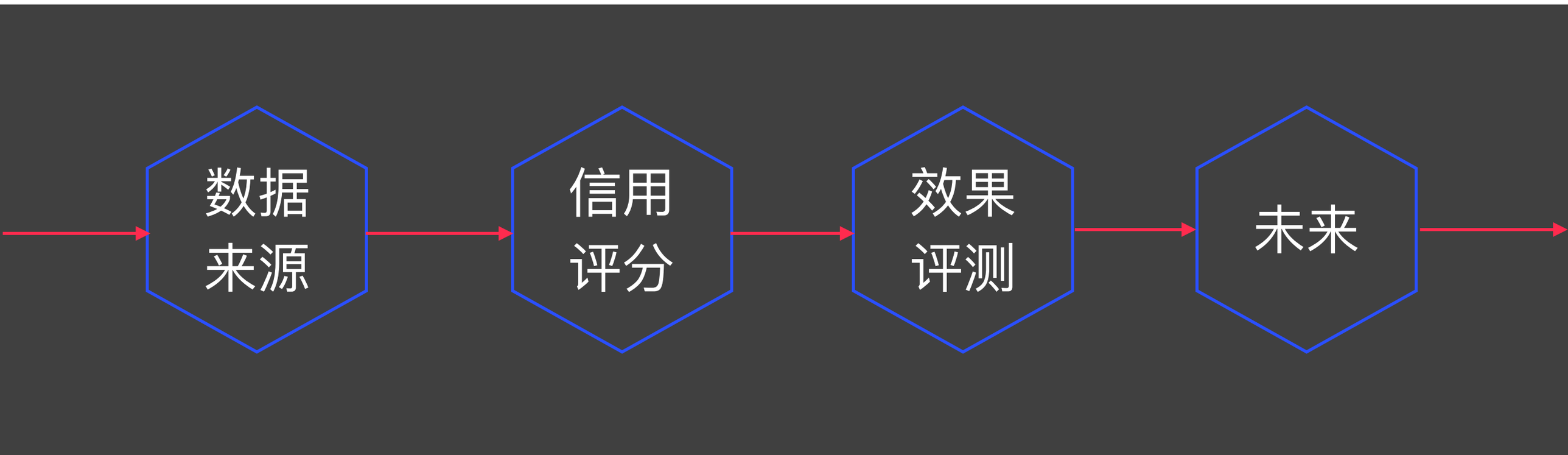
利润评分

税务检查

罚款

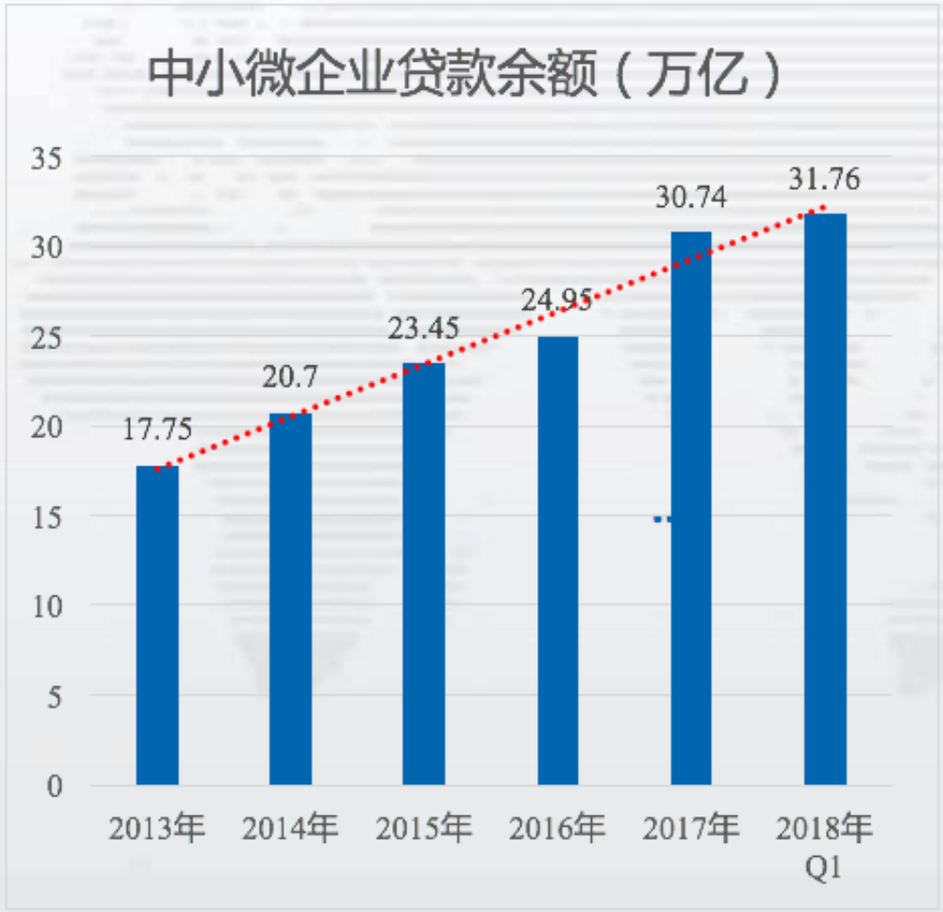
假释

中小微企业
信用评分



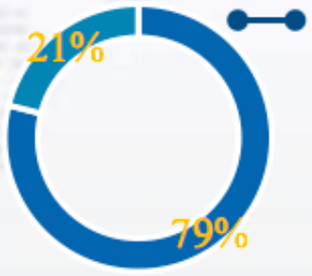
中小微企业

中国中小微企业贷款规模快速增长



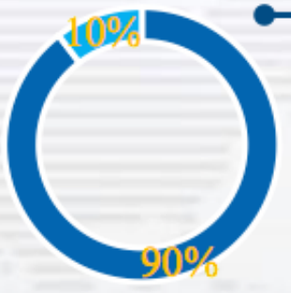
中小微企业风控服务市场潜力巨大

中小微企业贷款需求高



2017 年 7 月小微企业数量 7328.1 万户，小微企业贷款户数约 1545 万户，约占21%

江苏中小微企业贷款余额占比高



截至 2018 年一季度末江苏省小微企业贷款余额为3.29万亿元，占全国小微企业贷款总量的9.95%，占全省贷款余额的28.49%

国家频繁发布政策，加大对中小微企业的扶持和支持力度

2018年11月9日主持召开国务院常务会议

- 国务院总理李克强要求加大金融支持缓解民营企业特别是小微企业融资难融资贵。
- 从大型企业授信规模中拿出一部分，用于增加小微企业贷款。

2018年6月25日,五部委联合发布银发〔2018〕162号)

- 加大金融科技等产品服务创新。银行业金融机构要加强对互联网、大数据、云计算等信息技术的运用，改造信贷流程和信用评价模型，降低运营管理成本，提高贷款发放效率和服务便利度。

中小微企业数据

- **企业信息类8个特征变量**

企业经营年限、企业注册资本、企业所在区县、一般纳税人资质、企业类型、所属行业、法人持股比例、商变更情况

- **实际控制人6个特征变量**

法人年龄、婚姻状况、子女情况、户籍种类、住房情况、申请人本行业从业年限

- **经营发票数据14个特征变量**

销项发票计算的销售额、主要销售地区省内、主要下游客户经营年限、主要下游客户企业类型、主要下游客户行业、红冲发票比例、无效发票比例、专票占比、近24个月月波动率、近24个月季度波动率、近24个月交易方一致性、近24个月集中度、销售额全国企业中排名、销售额行业排名

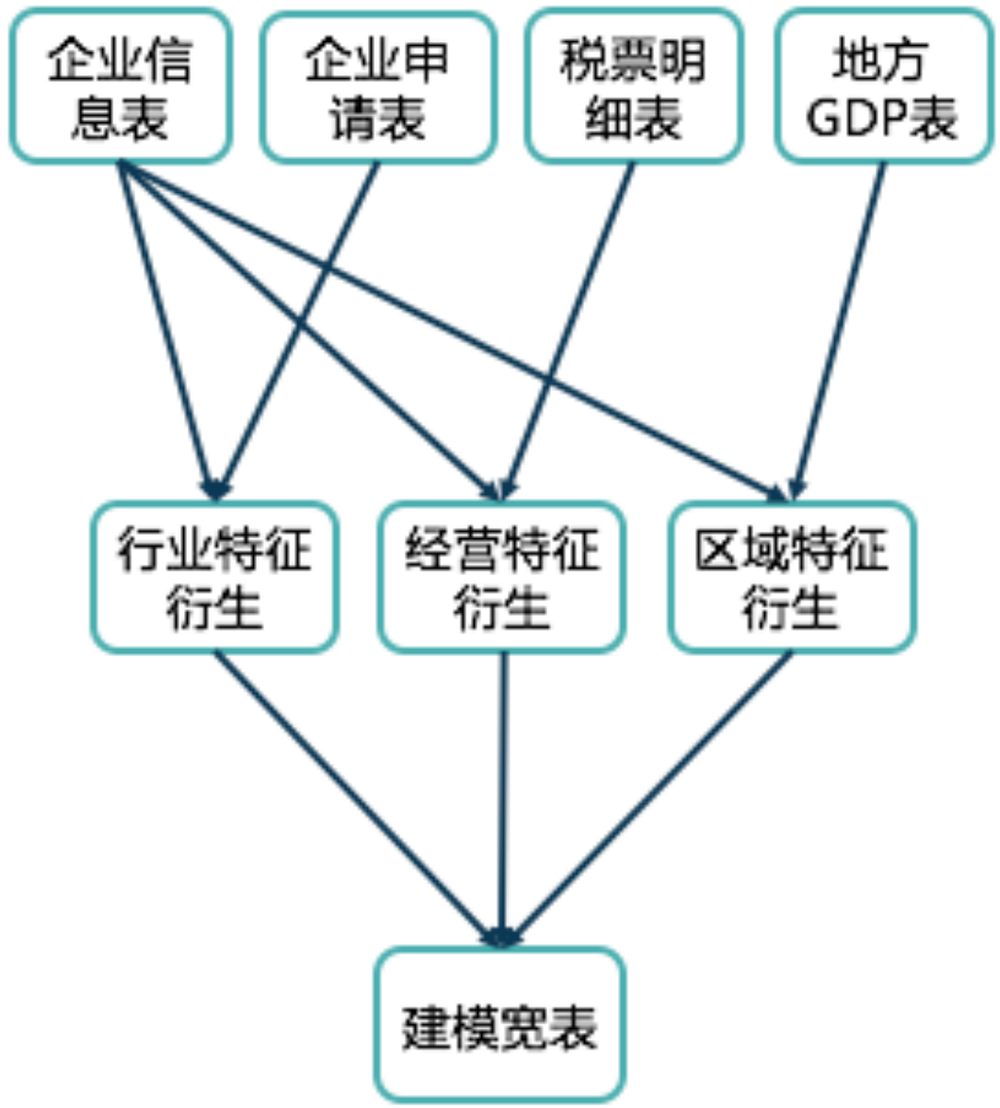
- **企业风险信息类2个特征变量**

企业有不良声誉记录的（如上过有关机构或部门的黑名单），

企业实际控制人有赌博、吸毒等不良嗜好的

- **综合评价类2个特征变量**

企业一致性指数(企业销售客户稳定性（下游）)、企业授信倾向分。



Credit Scoring 信用评分

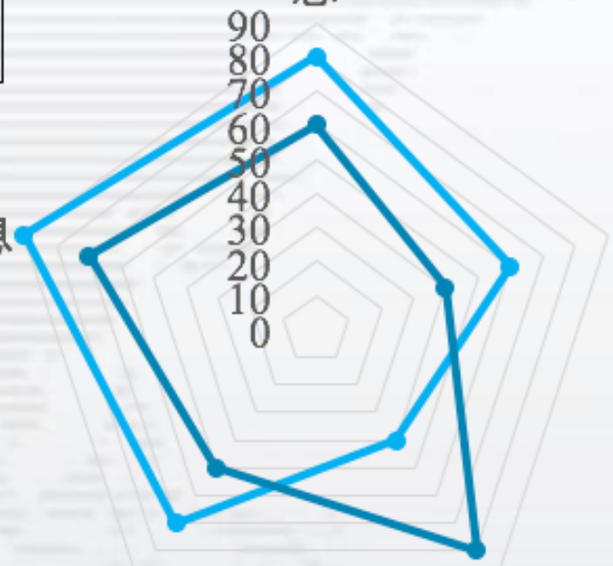
基于计量经济分析，通过大数据以及机器学习算法，计算中小微企业信用评分

- ❐行业
- ❐规模
- ❐注册年份
- ❐股东及出资信息
- ❐地区
- ❐企业性质
- ❐法定代表人
- ❐……

企业基本信息

- ❐诉讼信息
- ❐失信被执行人
- ❐行政处罚
- ❐环保处罚
- ❐税收违法
- ❐统计失信
- ❐环保失信
- ❐食药品抽检不合格
- ❐环保违法
- ❐税票机开机情况
- ❐……

风险信息



企业经营情况

- ❐年税票总额
- ❐主营商品
- ❐平均月税票额
- ❐年度同比税票额增长
- ❐季度数票额方差
- ❐最大季度税票额
- ❐最小季度税票额
- ❐最大月度税票额税
- ❐最小月度税票额
- ❐有税票月份数
- ❐无税票月份数
- ❐专用增值税票金额
- ❐普票金额
- ❐ 红冲税票金额
- ❐无效税票金额
- ❐下游企业家数
- ❐……

竞争力

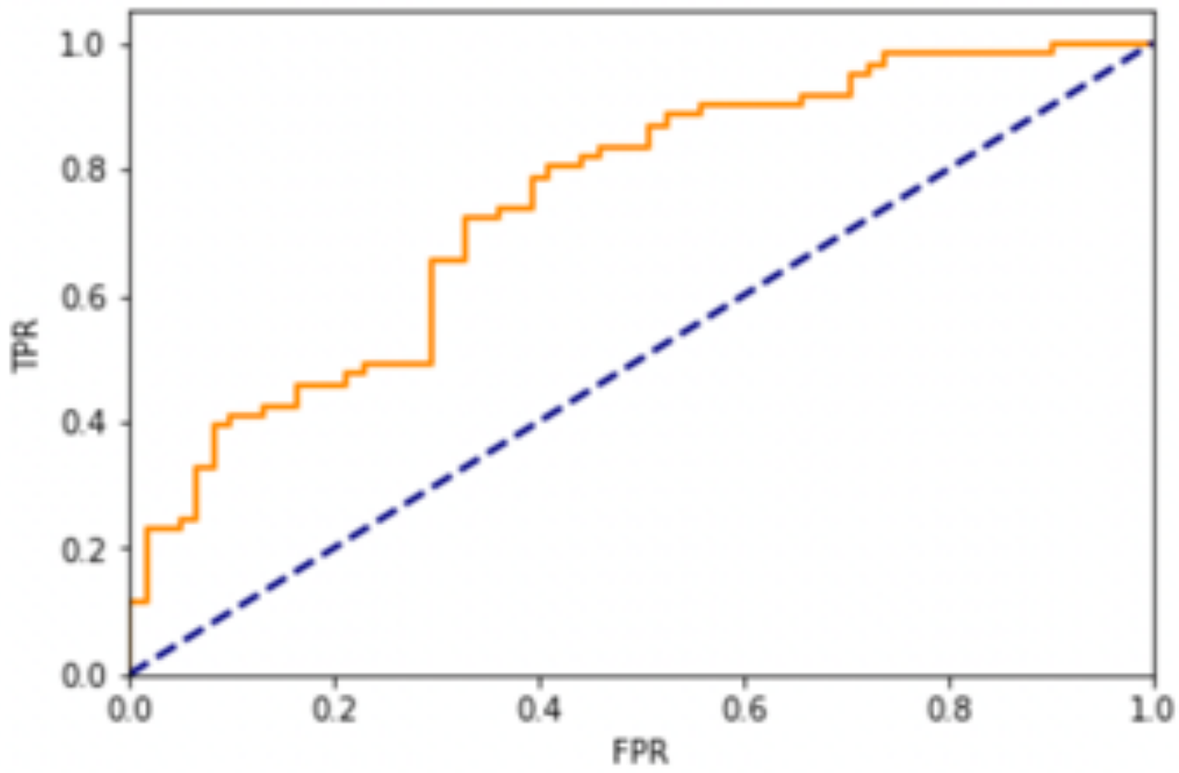
- ❐行业排名
- ❐规模排名
- ❐经营年限排名
- ❐地区排名
- ❐增长率排名
- ❐……

经营环境

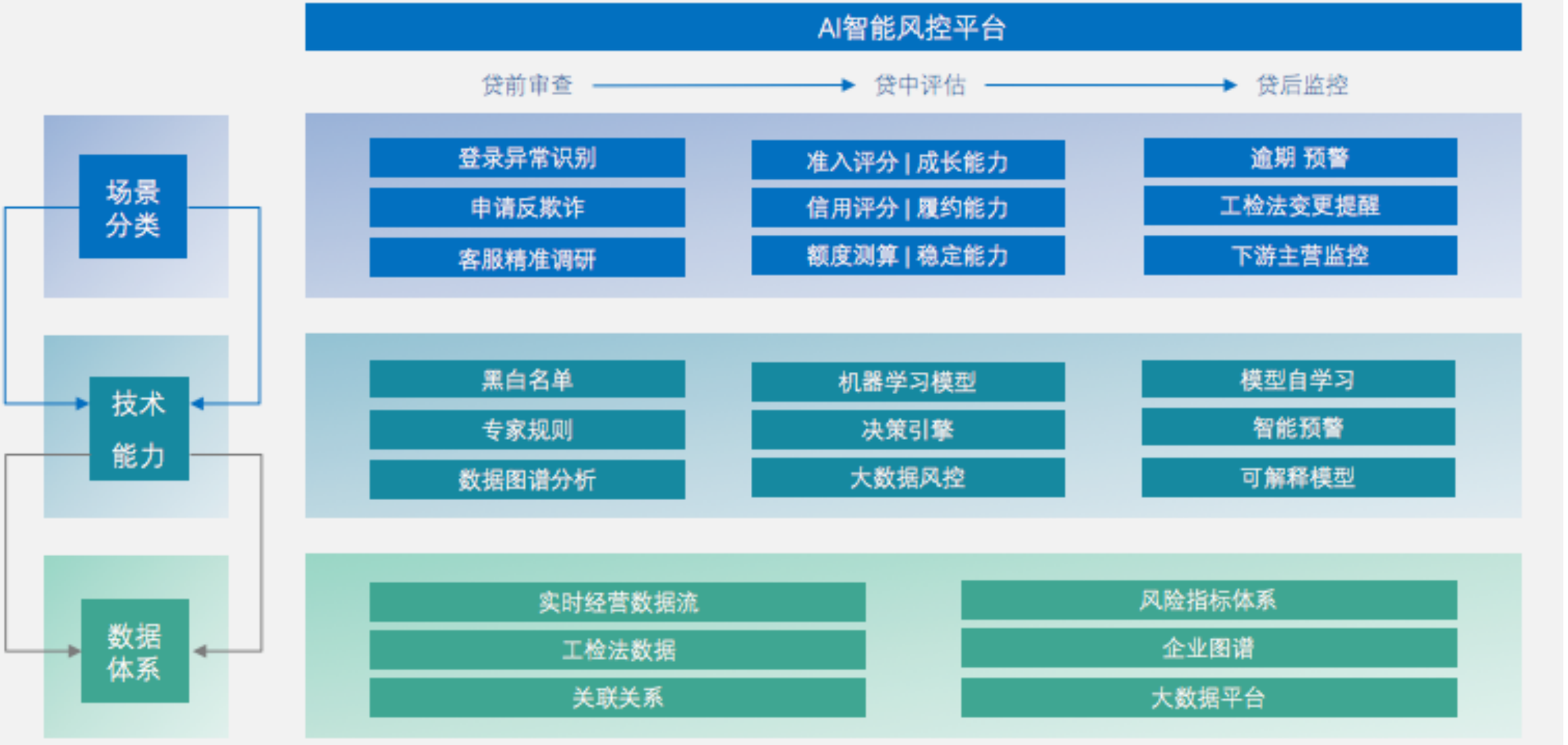
- ❐行业增长率
- ❐地区GDP
- ❐地区经济增长率
- ❐地区PPI
- ❐行业利润率
- ❐地区人口
- ❐地区CPI
- ❐……

效果评测

- **模型超参数：**
  学习速率：0.05
  决策树最大深度：3
  gbdt中决策树的棵数：76
- **模型效果：**
  准确率：0.664
  AUC：0.747
  正样本精确率：0.61,
  正样本召回率：0.90,
  负样本精确率：0.81,
  负样本召回率：0.43



未来

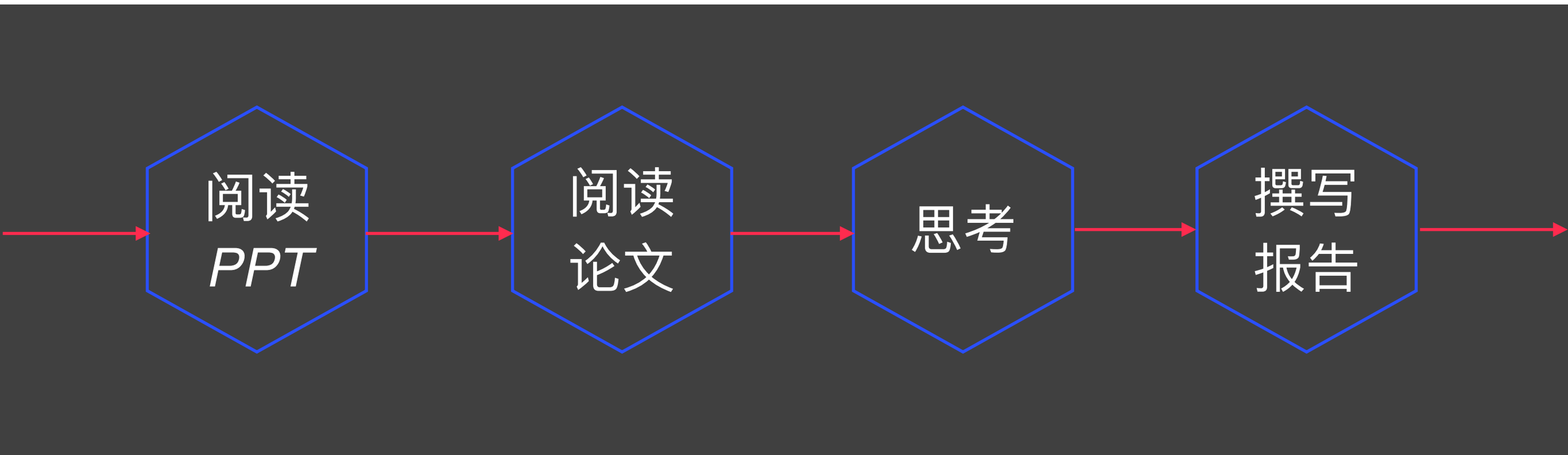

AI智能风控平台
贷前审查
贷中评估
贷后监控
场景分类
登录异常识别
申请反欺诈
客服精准调研
准入评分 | 成长能力
信用评分 | 履约能力
额度测算 | 稳定能力
逾期 预警
工检法变更提醒
下游主营监控
技术能力
黑白名单
专家规则
数据图谱分析
机器学习模型
决策引擎
大数据风控
模型自学习
智能预警
可解释模型
数据体系
实时经营数据流
工检法数据
关联关系
风险指标体系
企业图谱
大数据平台

提问时间！

孙惠平

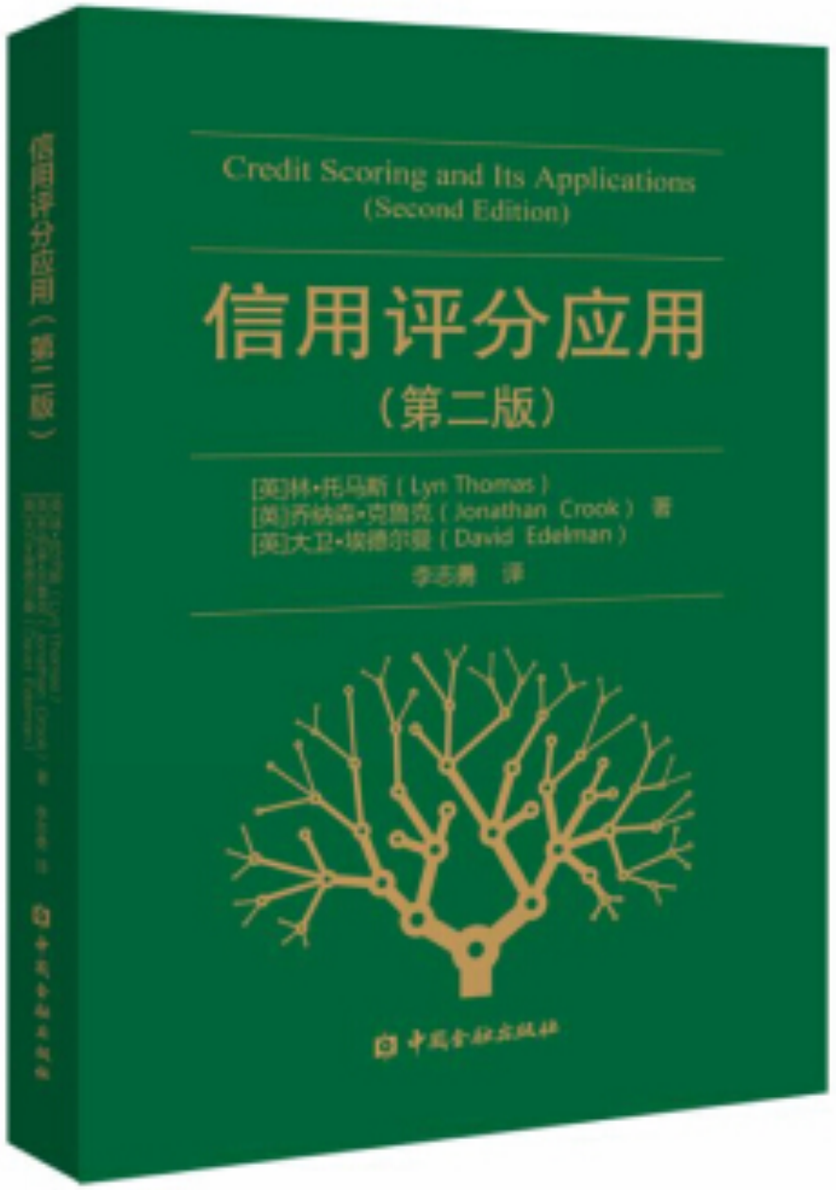
sunhp@ss.pku.edu.cn

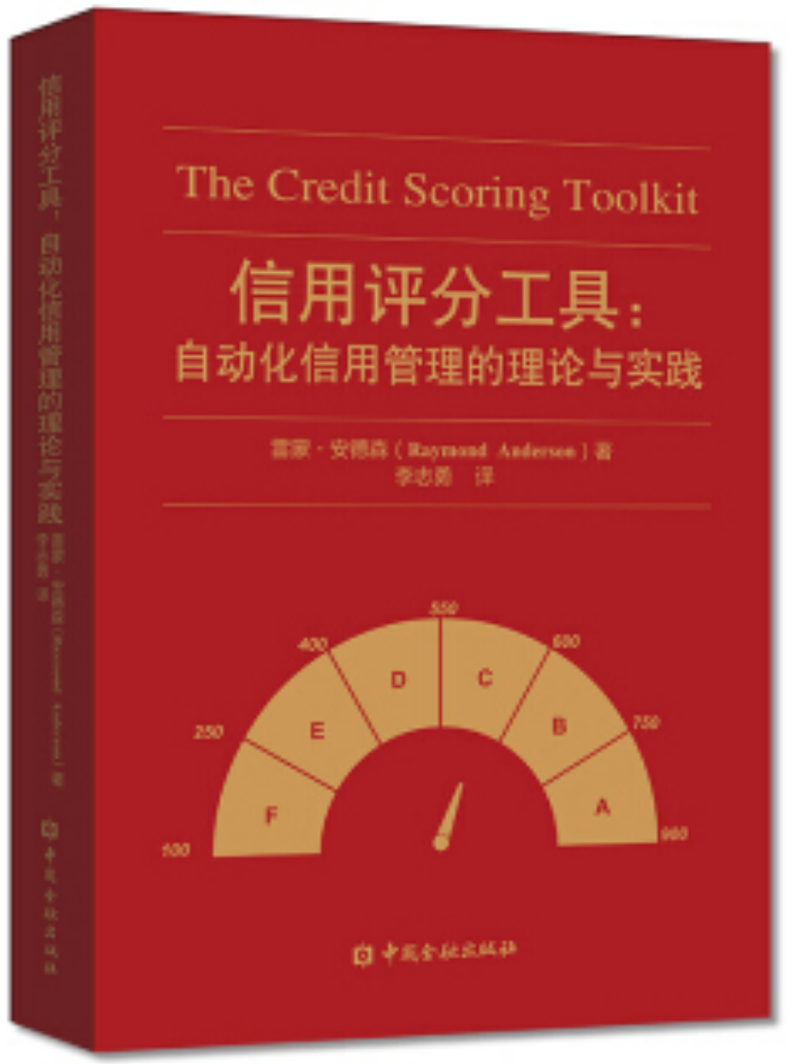
课后作业



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