# 法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

# 互联网金融业贷款申请评分卡的介绍

# 目录

# 信用风险和评分卡模型的基本概念

## 什么是信用风险

交易对手未能履行约定契约中的义务而造成经济损失的风险，即受信人不能履行还本付息的责任而使授信人的预期收益与实际收益发生偏离的可能性，它是金融风险的主要类型。

## 组成部分

| | |
|---|---|
| PD | 违约概率 |
| LGD | 违约条件下的损失率 |
| EAD | 违约风险下的敞口暴露 |
| RWA | 风险权重资产 |
| EL | 期望损失 |

# 信用风险和评分卡模型的基本概念

## 坏样本的定义

- M3 & M3+ 逾期
- 债务重组
- 个人破产
- 银行主动关户或注销
- 其他相关违法行为

## M0，M1，M2的定义

- M0：最后缴款日的第二天到下一个账单日
- M1：M0时段的延续，即在未还款的第二个账单日到第二次账单的最后缴款日之间
- M2：M1的延续，即在未还款的第三个账单日到第三次账单的最后缴款日之间

。。。

# 信用风险和评分卡模型的基本概念

## 什么是评分卡

信贷场景中的评分卡

- 以分数的形式来衡量风险几率的一种手段
- 是对未来一段时间内违约/逾期/失联概率的预测
- 有一个明确的(正)区间
- 通常分数越高越安全
- 数据驱动
- 反欺诈评分卡、申请评分卡、行为评分卡、催收评分卡

非信贷场景中的评分卡

- 推荐评分卡
- 流失评分卡

# 信用风险和评分卡模型的基本概念

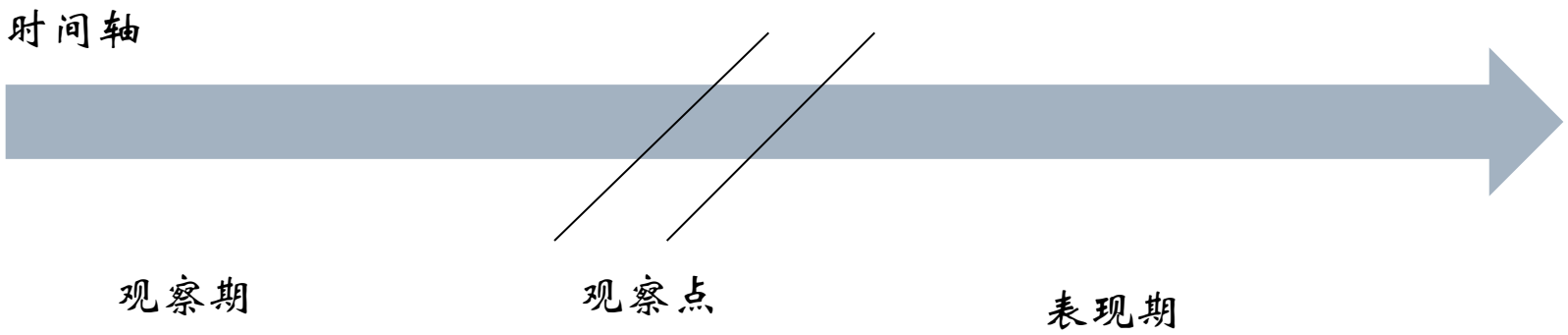
## 观察期与表现期

### 观察期

- 搜集变量、特征的时间窗口，通常3年以内
- 带时间切片的变量

### 表现期

- 搜集是否出发坏样本定义的时间窗口，通常6个月～1年



# 信用风险和评分卡模型的基本概念

## □ 评分卡模型开发步骤

I. 立项

II. 数据准备与预处理

III. 模型构建

IV. 模型评估

V. 验证/审计

VI. 模型部署

VII. 模型监控

# 信用风险和评分卡模型的基本概念

## 评分卡开发的常用模型

- 逻辑回归

  优点：简单，稳定，可解释，技术成熟，易于监测和部署

  缺点：准确度不高

- 决策树

  优点：对数据质量要求低，易解释

  缺点：准确度不高

- 其他元模型

- 组合模型

  优点：准确度高，不易过拟合

  缺点：不易解释；部署困难；计算量大

# 信用风险和评分卡模型的基本概念

- 模型监控的指标

AR

KS

PSI

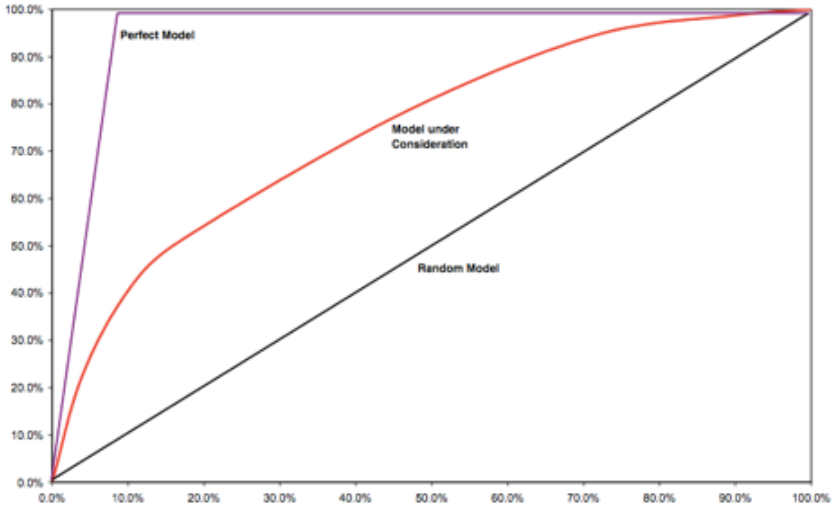
Kendall’ Tau

Migration Matrix

# 信用风险和评分卡模型的基本概念

## AR(Accuracy Ratio)

衡量分数预测能力的指标，需要一个完整的表现期。取值位于－1～1之间。




先把样本按分数由低到高排序，X轴是总样本的累积比例，Y轴是坏样本占总的坏样本的累积比例。AR就等于模型在随机模型之上的面积除以理想模型在随机模型之上的面积。计算中可以用梯形近似逼近曲线下面积来计算，AR越高说明模型区分效果越好。

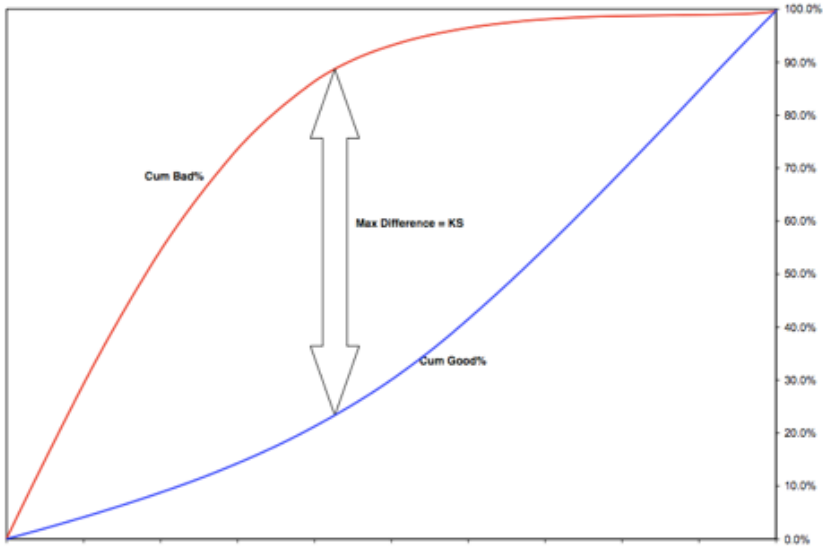
$x_k, y_k$代表分数的第k个分位点对应的累积总样本及相应的坏样本的比例。设总的坏样本的比例为$B_0$，令 $(x_k, y_k) = (0,0)$

$$AR = \frac{\sum_{k=1}^{10} \frac{1}{2}(x_k - x_{k-1})(y_k + y_{k-1}) - \frac{1}{2}}{\frac{1}{2}(1 - B_0)}$$

# 信用风险和评分卡模型的基本概念

## KS(Kolmogorov-Smirnov)

衡量分数区分能力的指标




把样本按分数由低到高排序，X轴是总样本累积比例，Y轴是累积好、坏样本分别占总的好、坏样本的比例，两条曲线在Y轴方向上的相差最大值即KS。KS越大说明模型的区分能力越好。

$Bad_k$和$good_k$分别为分数累积到第k个分位点的坏样本个数和好样本个数，KS计算公式是：

$$KS = \max\{\frac{Bad_k}{Bad_{total}} - \frac{Good_k}{Good_{total}}\}$$

# 信用风险和评分卡模型的基本概念

## PSI (Population Stability Index)

衡量分数稳定性的指标

令$R_i$是现在样本中第i个组占总样本的百分比，$B_i$是模型开发时第i个分组占总样本的百分比。PSI取值越小说明分数的分布随时间变化越小。

$$PSI = \sum_i (R_i - B_i) * \ln(R_i / B_i)$$

# 信用风险和评分卡模型的基本概念

## Kendall's Tau

正确有效的评分卡模型中，低分段的实际逾期率应该严格大于高分段的实际逾期率。我们将分数从低到高划分为10组，每组的实际逾期率记作$r_1,r_2,..,r_{10}$。对所有的$(r_i,r_j)$的组合，如果$r_i<r_j$且i<j或者$r_i>r_j$且i>j，则记作一个discordant pair, 否则记作concordant pair. Kendal's Tau的计算公式为：

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

Kendall's Tau越接近1或者等于1，说明逾期率在分数上的单调下降性越明显，反之说明分数的变化与逾期率的变化的一致性得不到保证。

# 信用风险和评分卡模型的基本概念

## Migration Matrix

迁移矩阵是衡量分数迁移的指标，对相同的人群，观察在相邻两次监控日期（一周）分数的迁移变化。迁移矩阵中元素$M_{jk}$代表上次监控日期分数在j组中的人群在当前迁移到第k组的概率。实际计算中可把分数平均分成10组，计算这十组之间的迁移矩阵。

| | | Jan'2017 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 350～450 | 451～550 | 551～650 | 651～750 | 751～850 | 851～950 |
| Jul'2016 | 350～450 | 85.00% | 5.00% | 5.00% | 1.00% | 0.80% | 3.20% |
| | 451～550 | 5.00% | 75.00% | 5.00% | 10.00% | 2.00% | 3.00% |
| | 551～650 | 4.00% | 15.00% | 70.00% | 5.00% | 3.00% | 3.00% |
| | 651～750 | 6.00% | 6.00% | 5.00% | 75.00% | 5.00% | 3.00% |
| | 751～850 | 4.00% | 3.00% | 3.00% | 15.00% | 65.00% | 10.00% |
| | 851～950 | 3.00% | 7.00% | 2.00% | 3.00% | 5.00% | 80.00% |

# 目录

# 申请评分卡的重要性和特性

## 互联网金融特性与产品

- 传统金融机构 + 非金融机构
- 传统金融机构：传统金融业务的互联网创新以及电商化创新、APP软件等
- 非金融机构：　利用互联网技术进行金融运作的电商企业

  (P2P)模式的网络借贷平台

  众筹模式的网络投资平台

  挖财类(模式)的手机理财APP(理财宝类)

  第三方支付平台等。

## 为什么要开发申请评分卡

- 风险控制
- 营销
- 资本管理

# 申请评分卡的重要性和特性

- 评分卡的特性

稳定性

区分性

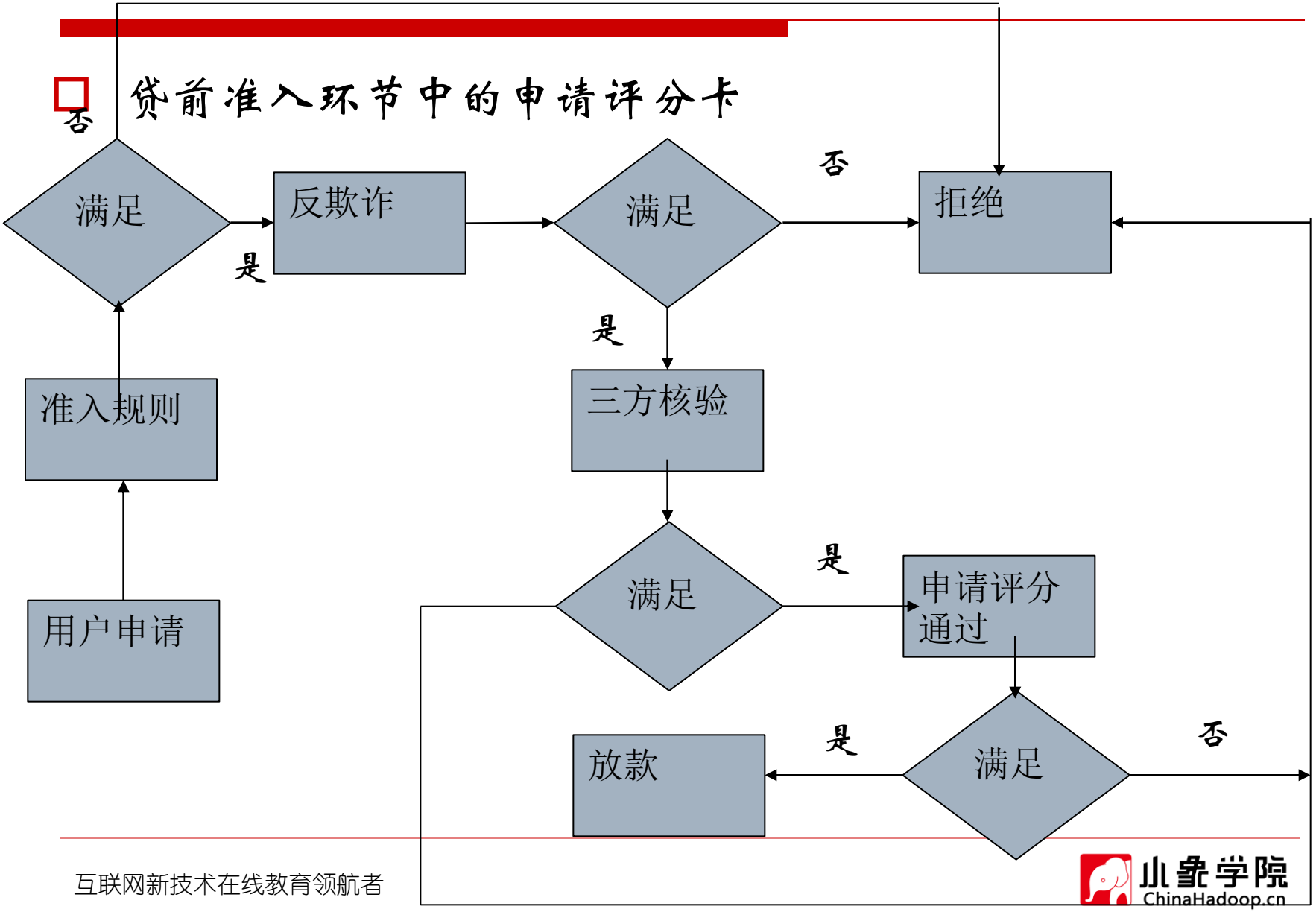
预测能力

和逾期概率等价

# 目录

# 贷款申请环节的数据介绍和描述

- 贷前准入环节中的申请评分卡



# 贷款申请环节的数据介绍和描述

## 申请评分卡常用的特征

| | |
|---|---|
| 个人信息 | 学历 性别 收入 |
| 负债信息 | 在本金融机构或者其他金融机构负债情况 |
| 消费能力 | 商品购买纪录，出境游，奢侈品消费 |
| 历史信用记录 | 历史逾期行为 |
| 新兴数据 | 人际社交 网络足迹 出行 个人财务 |

# 贷款申请环节的数据介绍和描述

## 数据字典

每一行代表一个样本（一笔成功成交借款），每个样本包含200多个各类字段

**Master** idx：每一笔贷款的unique key，可以与另外2个文件里的idx相匹配。

UserInfo_*：借款人特征字段

WeblogInfo_*：Info网络行为字段

Education_Info*：学历学籍字段

ThirdParty_Info_PeriodN_*：第三方数据时间段N字段

SocialNetwork_*：社交网络字段

LinstingInfo：借款成交时间

Target：违约标签（1 = 贷款违约，0 = 正常还款）

# 贷款申请环节的数据介绍和描述

- **Log_Info**

借款人的登陆信息

ListingInfo：借款成交时间

LogInfo1：操作代码

LogInfo2：操作类别

LogInfo3：登陆时间

idx：每一笔贷款的unique key

**Userupdate_Info**

借款人修改信息

ListingInfo1：借款成交时间

UserupdateInfo1：修改内容

UserupdateInfo2：修改时间

idx：每一笔贷款的unique key

# 目录

# 非平衡样本问题的定义和解决方法

## 非平衡样本的定义

在分类问题中，每种类别的出现概率未必均衡

信用风险:正常用户远多于逾期/违约用户

流失风险: 留存客户多于流失客户

## 非平衡样本的隐患

降低对少类样本的灵敏性

# 非平衡样本问题的定义和解决方法

## 非平衡样本的解决方案

**过采样**

- 优点：简单，对数据质量要求不高
- 缺点：过拟合

**欠采样**

- 优点：简单，对数据质量要求不高
- 缺点：丢失重要信息

**SMOTE(合成少数过采样技术)**

- 优点：不易过拟合，保留信息
- 缺点：不能对有缺失值和类别变量做处理

# 非平衡样本问题的定义和解决方法

## SMOTE算法

I. 采样最邻近算法，计算出每个少数类样本的K个近邻

II. 从K个近邻中随机挑选N个样本进行随机线性插值

III. 构造新的少数类样本

$$New = x_i + rand(0,1) \times (y_j - x_i), j = 1,2,..,N$$

其中$x_i$为少类中的一个观测点，$y_j$为k个邻近中随机抽取的样本

IV. 将新样本与原数据合成，产生新的训练集

# 疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

# 联系我们

## 小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop