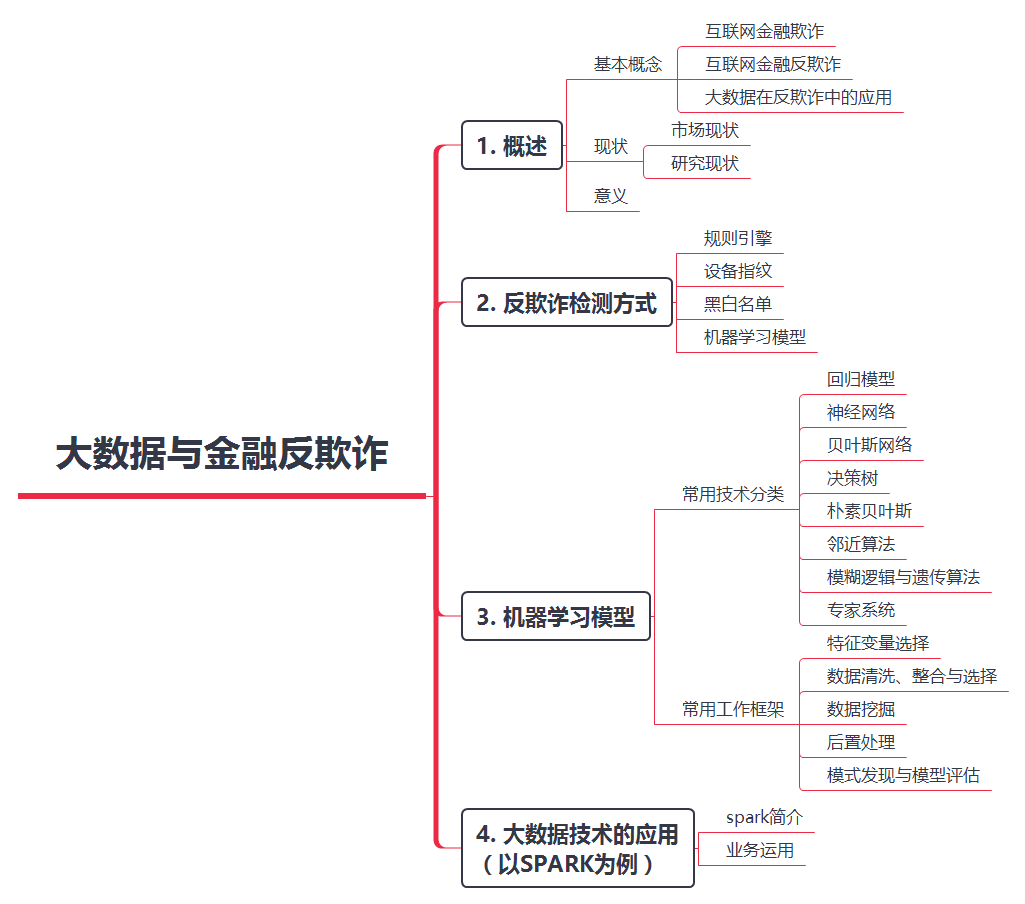
大数据与金融反欺诈

傅琦佳/张炽 2017.08.22



# 1.概述

## 1.1基本概念

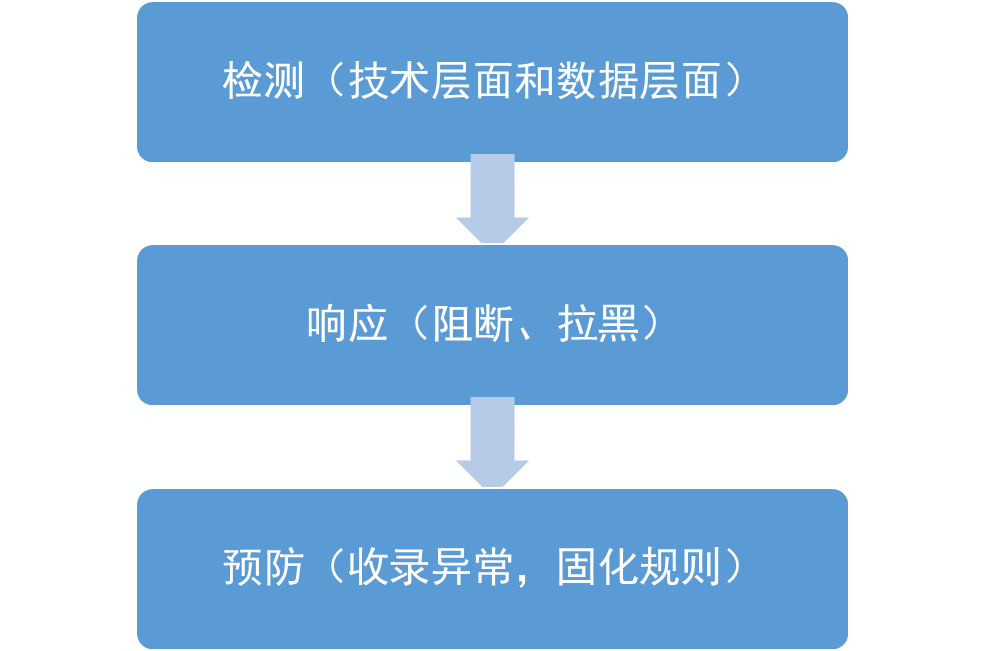
### 1.1.1互联网金融欺诈

互联网金融是传统金融机构与互联网企业利用互联网技术和信息通信技术实现资金融通、支付、投资和信息中介服务的新型金融业务模式。由于互联网的匿名性、产品客户的逆向选择与道德风险、快速抢占市场的业务压力、基础IT运维能力不足、从业者经验不足考虑不周等，互联网金融欺诈时有发生。金融欺诈属于操作风险，主要包括盗窃、伪造、黑客攻击等外部欺诈风险。例如典型的黑产交易，通过制作病毒木马、各种钓鱼手段、黑客攻击方式获取用户信息，进行电信诈骗、盗取装备、盗卡交易等非法牟利活动。

### 1.1.2 互联网金融反欺诈

互联网金融反欺诈的本质是通过分析历史欺诈行为梳理完善风险特征库，如异地登录、非常用设备登录等风险信号，建立一系列的风险规则判定集，预测用户行为背后的欺诈概率。一般来说，互金公司会在互联网服务过程中收集用户信息，结合线上线下数据，借助Apache等开源框架下一系列大数据处理工具进行风控。

反欺诈的业务流程如下图所示：



图x-金融反欺诈业务流程

互联网金融企业应对互联网金融欺诈的常用方法：

|  |  |  |
| --- | --- | --- |
| 方法 | 描述 | 例子 |
| 大数据 | 记录用户的网络行为，据需求利用一些隐私相关的业务信息 | 根据客户端的行为轨迹和访问的业务API，区分正常用户和恶意用户；分析访问内容，形成用户画像 |
| 身份识别 | 利用短信验证码、语音验证码、图形验证码、拖动滑块式验证码、人脸识别、活体检测等手段识别身份 | 身份证或存档照片比对等场景虹膜识别 |
| 求助外援 | 第三方解决方案提供商可以是开发者、伪装工具提供者、培训者、实施者、变现者等 | 整体的风控外包商提供风控基础构架定制化开发；SAAS风控解决方案提供商整合各家数据提供黑名单服务 |

### 1.1.3大数据在反欺诈业务中的应用

大数据是金融风控的基础。互金公司通常会在服务过程中收集用户信息（如身份信息、信用信息、社交信息、消费信息、行为信息和合作机构数据），结合申请者提供的信息借助大数据处理工具建立风控体系，对用户的欺诈风险、还款意愿、还款能力等进行有效判别。进行风控。以现金贷为例，通过勾稽比对或交叉检验，利用多维度数据验证用户真实性和可靠性；对权重较大的变量进行强特征筛选；通过验证与申请人相关联个体的信息，考察风险关系；利用用户的行为数据鉴别金融欺诈。应用于互联网金融反欺诈的大数据通常来自电商、社交、同业合作的征信系统，例如：

|  |  |  |
| --- | --- | --- |
| 产品 | 数据源 | 类型 |
| 京东金融 | 京东商城和京东物流 | 电商数据 |
| 芝麻信用 | 支付宝、天猫和淘宝数据 | 电商数据 |
| 融360 | 天机系统 | 同业合作 |
| 腾讯征信 | 腾讯QQ、微信、QQ空间等 | 社交数据 |
| 宜信、聚信立、闪银 | Talking Data | 移动数据 |

## 1.2现状

### 1.2.1市场现状

传统的金融企业主要是利用来自央行、商业银行、企业、通信运营商乃至黑市的数据，通过“黑白名单”、基于规则的防范机制、自有业务数据建模等方式进行风控。在移动互联网时代，现代金融公司更注重来自移动端的行为数据，采用互联网思维，采集用户的社会行为和社会属性等多维数据识别欺诈客户。

### 1.2.2研究现状

有关金融欺诈监测的研究主要集中于信用卡欺诈和反洗钱研究。在信用卡欺诈研究方面，主要采用的数据挖掘技术是分类（如神经网络、决策树）和关联规则（如Apriori算法）。在反洗钱方面，主要应用的数据挖掘技术有孤立点分析、序列模式挖掘、分类研究、聚类分析等。借助数据挖掘技术，美国金融犯罪执法网络的FAIS系统实现了对可疑交易报告的分析和评估，澳大利亚交易分析与报告中心开发了 Screen IT 系统自动筛选可疑交易报告。

## 1.3意义

随着互联网金融的爆发性增长，各类信贷产品的消费群体范围日益扩大，互联网金融行业风险越来越高，每年因欺诈导致的损失达500 亿元，金融风险防控受到前所未有的重视。全国人大会议连续三年就互联网金融进行了探讨，2016年重点工作明确指出，“加快改革完善现代金融监管体制，实现金融风险监管全覆盖；规范发展互联网金融；严厉打击金融诈骗等违法犯罪活动。”中国人民银行副行长亦表示，银行与公安机关、工信部等部门合力构建电信网络欺诈风险管理体系。在“互联网+金融”的新型模式下，大数据技术必将与金融业的各类应用场景深度结合，进而催生出巨大的基于大数据的金融反欺诈服务市场。

2. 反欺诈检测方式

反欺诈检测方式包括：**规则引擎、设备指纹、黑白名单、机器学习模型**。其中，规则引擎和学习模型是传统反欺诈系统中重要的两个基本组成部分。大部分金融反欺诈机构和公司会结合公司实际需求和反欺诈方案，综合运用上述方法，搭建出一套最适合自己公司业务、产品以及用户类型的反欺诈管理系统。

## 2.1 规则引擎

规则引擎由推理引擎发展而来，是一种嵌入在应用程序中的组件，实现了将业务决策从应用程序代码中分离出来，并使用预定义的语义模块编写业务决策。接受数据输入，解释业务规则，并根据业务规则做出业务决策。

规则引擎的优势在于，可以直接从数据库中导入数据，挑选黑名单和其他坏的列表。当每一个新的诈骗情况发生后，分析师会增加一个新规则以保证公司在可预见范围内免于欺诈风险，可以避免一些周期性出现的欺诈。但是这些甄别欺诈行为的规则依赖于从大量历史案例中总结出来的“专家知识”，不适用于大规模的欺诈案例。

## 2.2 设备指纹

设备标识（指纹）是基于设备的多种存储或路由介质生成的不重复、唯一的设备标识，和现实世界里政府用身份证来标识公民情况类似。相对于传统的cookie，设备指纹具备不受浏览器兼容性限制、用户无法修改、不会被浏览器清除、可跨应用追踪等诸多优点。现阶段的设备指纹主要采集设备上网信息、用户实名信息及场景信息，用于大数据信贷风险控制。

## 2.3 黑白名单

指在商业银行及金融机构间建立银行黑名单机制，将有欺诈记录的不法分子、可疑的网站IP地址将被列入黑名单在机构间共享，建立案件资料库，对已经发生的案件进行整理并在机构间共享。

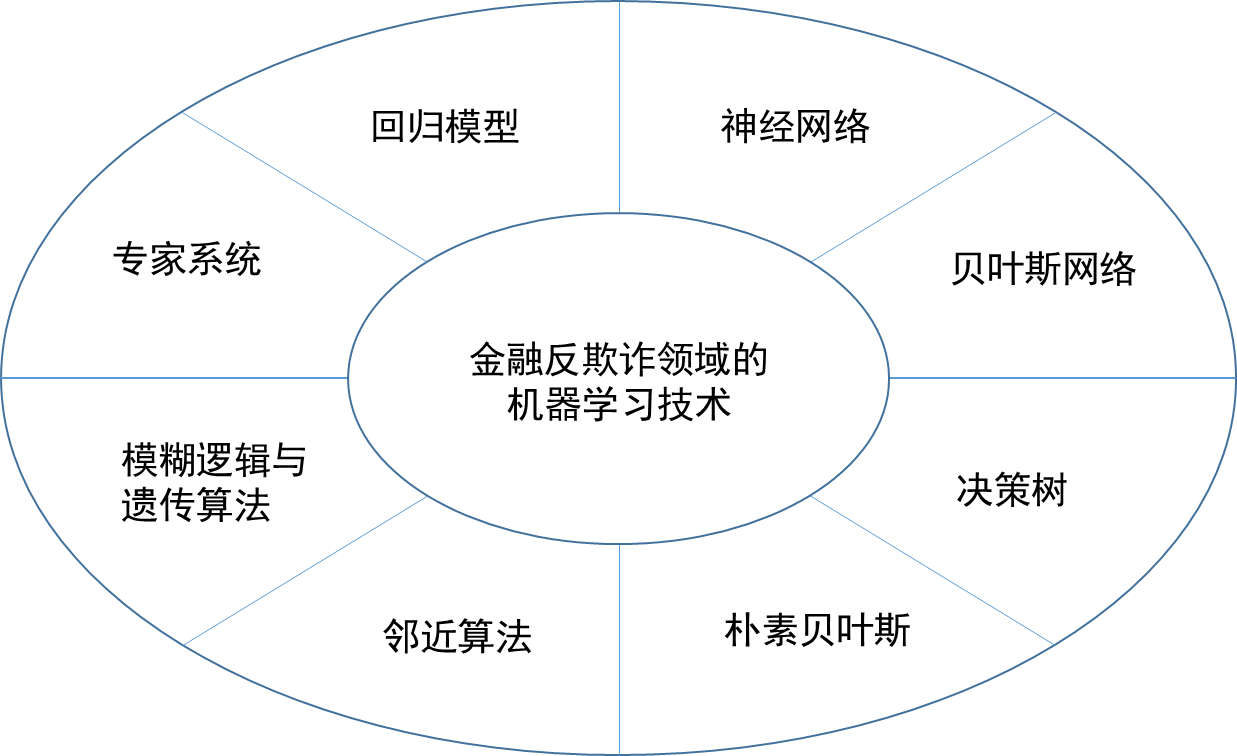
黑名单来源于民间借贷、线上P2P、信用卡公司、小额借贷等公司的历史违约用户及催收公司。灰名单是逾期但是还没有达到违约的客户，或多头借贷用户。黑名单和灰名单是很好的风控方式，但是各个征信公司所拥有的名单仅仅是市场总量的一部分，很多互联网金融公司不得不接入多个风控公司，来获得更多的黑名单来提高查得率。

## 2.4 机器学习模型

采用数据挖掘方法，基于历史数据而建立的分类模型。这类模型的训练往往需要大量数据。机器学习模型包括有监督和无监督两类，下一部分重点对有广泛应用的机器学习模型进行介绍。

# 3. 机器学习模型在金融反欺诈中的应用

## 3.1 常用技术分类

****

**图x-金融反欺诈领域的机器学习技术**

上图列出了常用的数据挖掘技术与应用方法。需要指出的是，许多的研究与工程人员尝试综合地运用上述各项数据挖掘技术来进行金融反欺诈监测，诸如决策树、神经网络、贝叶斯网络、K聚类算法等。算法之间的结合，可以达到更好的检测效果。目前中国在金融反欺诈的应用与研究还是以单一的逻辑回归为主。

### 3.1.1 回归模型 Regression Models

回归模型是金融反欺诈领域最常用的方法，主要基于逻辑回归（logistic regression）、逐步逻辑回归（stepwise-logistic regression）、多准则决策方法（multi criteria decision making method）和广义第二类贝塔分布（exponential generalized beta two (EGB2)）。

逻辑回归和聚类分析可以一同建立金融反欺诈模型，通过识别财务指数、公司管理、财务风险和相关趋势四个方面的数据，对可能出现的金融欺诈行为进行提前判断。逻辑回归模型算法简单，构造方便，且分类效果好。据研究指出，基于逻辑回归的金融反欺诈系统可以达到95.1%的正确率。

*Chen, G., Zhanjia L., & Feng, H. (2007). Empirical Study on detecting financial statements Fraud- based on empirical data of public companies. [J] Auditing Study, 2007.*

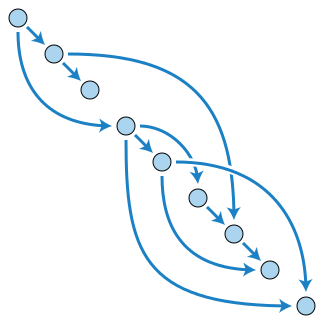
### 3.1.2 神经网络 Neural Networks

神经网络算法是仿照人脑结构构建的非线性统计建模工具，具有很好的适应性（adaptive）、鲁棒性（robust），且较易引入新的训练权重。通过学习内部训练集的特征，神经网络模型在管理水平上预测企业金融欺诈行为有很好的表现。

对比研究显示，和决策树、贝叶斯网络、逻辑回归等算法相比，神经网络系统表现出了更高的学习效率。

*Fanning, K., & Cogger, K. (1998). Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management, vol. 7, no. 1, pp. 21- 24, 1998.*

### 3.1.3 贝叶斯网络 Bayesian Belief Network

贝叶斯网络通过有向非循环图表示随机变量之间的关系，每一个节点代表一个随机变量，节点之间的有向线段代表随机变量之间的条件独立性（conditional independencies）。贝叶斯网络常用在信用卡业务、移动终端保险以及公司金融反欺诈领域。在一研究中，其分类正确率达到90.3%，超过神经网络和决策树方法。

*Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements, Expert Systems with Applications 32 (4) (2007) 995–1003.*

### 3.1.4 决策树 Decision Trees

在决策树的树状结构中，每一个节点代表一次对属性的测试，每一个枝杈代表可能产生的结果，通过这种方式将观测值划分成互相独立的多组值。决策树模型建立在其他的算法之上，诸如ID3，CART和C4.5 等。决策树同样在金融反欺诈领域有很广泛的应用。

### 3.1.5 朴素贝叶斯 Naïve Bayes

朴素贝叶斯是一种简单的概率分类器。它基于贝叶斯条件概率公式，引入特征条件独立的强假设，建模简单，且有较为稳定的分类效率。在属性相关性较小时，朴素贝叶斯性能良好。朴素贝叶斯方法在银行和金融机构的反欺诈系统中广泛应用。

### 3.1.6 邻近算法 K-Nearest Neighbor algorithm

kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。邻近算法多用于车险欺诈和信用卡违约领域。

### 3.1.7模糊逻辑与遗传算法 Fuzzy logic and Genetic Algorithm

遗传算法应用在分类器系统中，通过对内审人员在欺诈条件下的决策行为进行建模，与支持向量机结合，解决那些没有被完全发现和定义的信用诈骗问题。

模糊逻辑是一种描述主观逻辑推理和分类的数学方法。它的基本原理是通过数据从属某个群组的概率对数据进行分组或聚类。模糊逻辑能通过量化欺诈风险因子，实现对变量有效的建模。研究显示，模糊逻辑模型的预测正确率可达86.7%

*Lenard, M. J., & Alam, P. (2004). The use of fuzzy logic and expert reasoning for knowledge management and discovery of financial reporting fraud. In H. Nemati and C. Barko (Eds.), Hershey, PA: Idea Group, Inc.*

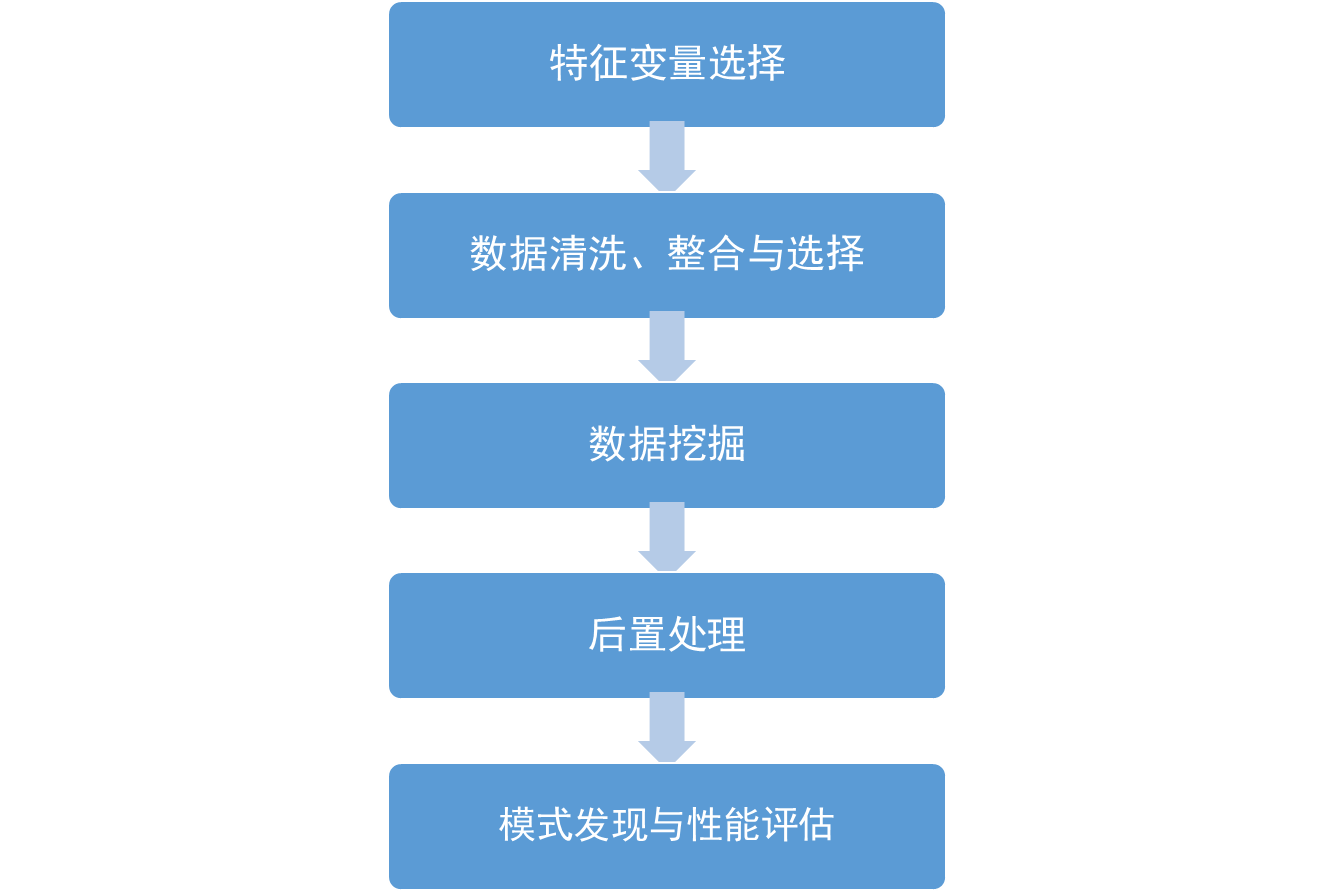
### 3.1.8 专家系统 Expert Systems

专家系统（Expert System)是一种在特定领域内具有专家水平解决问题能力的程序系统。它能够有效地运用专家多年积累的有效经验和专门知识，通过模拟专家的思维过程，解决需要专家才能解决的问题。

通过专家系统，审计人员可以获得在不同内容和水平下更好的欺诈监测能力，并通过合理的审计程序，给出更可靠的审计建议。

## 3.2 金融反欺诈的工作框架

一个常用的基于机器学习的反欺诈工作框架如下表所示：



图x-基于机器学习的反欺诈工作框架

### 3.2.1 特征变量选择

特征变量的筛选主要是从个人行为数据中，选择能够反映个人信用的变量。互联网提供了丰富且繁杂的个人行为，有些变量能很好体现个人信用度，有些则不明显。对变量进行适当地选择，能够提高判断信用度的准确性。

常用的变量主要包括以下几个维度：

* 用户数据：包括用户的年龄，性别，婚姻，职业，教育程度，收入情况等。
* 信用数据：包括用户在银行的征信记录，用户在银行或其他征信公司的征信记录。
* 交易数据：包括用户的交易金额，交易频率，交易地点，交易账户等。
* 消费数据：包括用户的消费时间，消费地点，消费习惯，消费金额等。
* 社交数据：包括用户的好友数量，好友的信用评级，好友的身份特征等。

同时，变量之间的组合可以共同反映用户的某种特性，在选择变量时应该尽可能从多个维度来刻画用户的特征。

### 3.2.2 数据清洗、整合与选择

数据的来源包括财务报表、审计报告、内控系统等公司财务数据渠道以及社交软件、征信记录、消费记录等个人信息渠道。这些数据的来源不同，需要对其进行统一的处理和整合。同时对数据进行相应的预处理，包括标准化、缺失值、合并、分组、异常值、变量衍生（交叉组合）、降维等方法。数据集的选择主要指对于训练集与测试集的选择。

### 3.2.3 数据挖掘

金融反欺诈问题的实质是一个有监督的二分类模型问题，利用上述所列出的机器学习的8种方法技术，对数据进行分类、聚类、预测、异常值检验、回归和可视化处理等。

### 3.2.4 后置处理

对于模型判断为欺诈的用户，需要人工对其进一步审核，减少误判的概率。方法包括信息回访、实地调研、资料查询。

### 3.2.5模式发现与性能评估（Pattern and Performance Evaluation）

模式评估指的是根据某种兴趣度度量，识别代表知识的有价值的模式，并用于后续的决策。模式兴趣度有两种主要的度量方式，客观度量基于所发现模式的结构和关于它们的统计量，主观度量则基于用户对于数据的期望程度。

通常评估模型有两种方法，一种是试受者工作特征曲线（ROC），另一种是预测值F-Score指标。ROC曲线是反映特异性和敏感性联系变量的综合指标，它通过将变量设定多个连续的阔值，计算出一系列特异性和敏感性的值，图示法揭示特异性和敏感性之间的制约关系。F-Score指标代表模型精确度和召回率的调和平均数，在一定程度上对模型的预测结果给出了综合的评价。

# 4. 大数据技术的应用（以spark为例）

## 4.1 spark简介

Apache Spark是一个开源和兼容Hadoop的集群计算平台，是Apache旗下的顶级项目，可以让程序提高100倍的内存计算速度或者10倍的磁盘计算速度(Hadoop)。相较于Java语言编写的MapReduce过程，Spark的代码更为简短，且提供脱机的交互式shell(REPL)逐行检测，实现即时数据分析。此外，Spark还提供Scala、Java、Python和R的API接口，整合Hadoop生态系统和数据储存系统(HDFS, Amazon S3, HIVE, HBase, Cassandra等)。Spark核心组件能够与SparkSQL、Spark Streming、 MLlib、GraphX等高效的软件库无缝连接使用。

## 4.2 spark在互联网金融反欺诈中的运用

在互联网金融领域，Spark堆栈技术可以获取大量的历史数据和其他一些外泄数据以及一些连接/请求信息(IP地理信息或时间信息)，从而在信用诈骗和风险管控系统的应用中取得非常好的模型结果。Spark还可以简化处理大规模数据的步骤流程，并将许多复杂的功能(如机器学习算法、图算法)结合起来。运算速度快如闪电的Spark是大数据时代的产物，它的数据处理技术可以利用海量、多源、异构数据准确识别欺诈行为，线上渠道实时反欺诈技术可以快速应对欺诈行为，关联关系挖掘技术深入研判愈发复杂的欺诈场景，人工智能技术灵活地挑战不断翻新的欺诈形式。

明略数据提出的在线实时决策和离线机器学习是较为成熟的案例，一方面利用Spark分布式集群计算引擎，对数据进行协处理、流式处理、交互式分析等，基于账户信息、信贷信息、行业信息等数据建立知识图谱，对显性或者隐性的关联性风险进行预警，对资金流、风险的传导进行判断，从而识别潜在的风险行业和欺诈客户。另一方面，基于 Spark 架构的数据挖掘和机器学习平台，离线构建规则自学习模型，为客户画像系统和数据挖掘平台提供计算能力，发现新型欺诈行为并产生对应的反欺诈规则，自动挖机检验生成标签规则，实现高效实时的反欺诈。如明略数据为某大型国有银行信用卡中心建立的实时反欺诈系统，借助spark处理海量高并发的线上行为数据（历史数据60TB和增量数据1500万条/天），应对虚假账号、伪装账号、异常登录、频繁登录等新型风险和欺诈行为。