互联网金融的大数据应用

面包君 2017/6/20

DT时代

目录

DT时代

定义

互联网金融在美国也叫做“FINTECH” 。互联网金融的定义有广义和狭义之分。

从广义来说，互联网金融是所有使用技术来提供金融服务的代名词。无论是提供全球数字汇款服务的 SWIFT 系统还是股票交易所或清算所的门户入口，都是互联网金融的一部分。

从狭义角度来看，互联网金融多指近年来由新兴公司和金融服务公司等提供的创新的具有破坏性的新型金融服务和产品，主要指的是运用新型技术提供服务的银行业务、公司金融业务、资本市场业务、金融数据分析、支付、个人财富管理。

DT时代

美国互联网金融发展

上世纪90年代前后

- 传统金融机构和业务信息化

上世纪90年代中后期

- 传统业务和互联网融合：SFNB/PayPal货币基金

本世纪以来

- 新型互联网金融业务发展：P2P/移动支付/虚拟货币

DT时代

美国互联网金融生态

我国互联网金融的变革

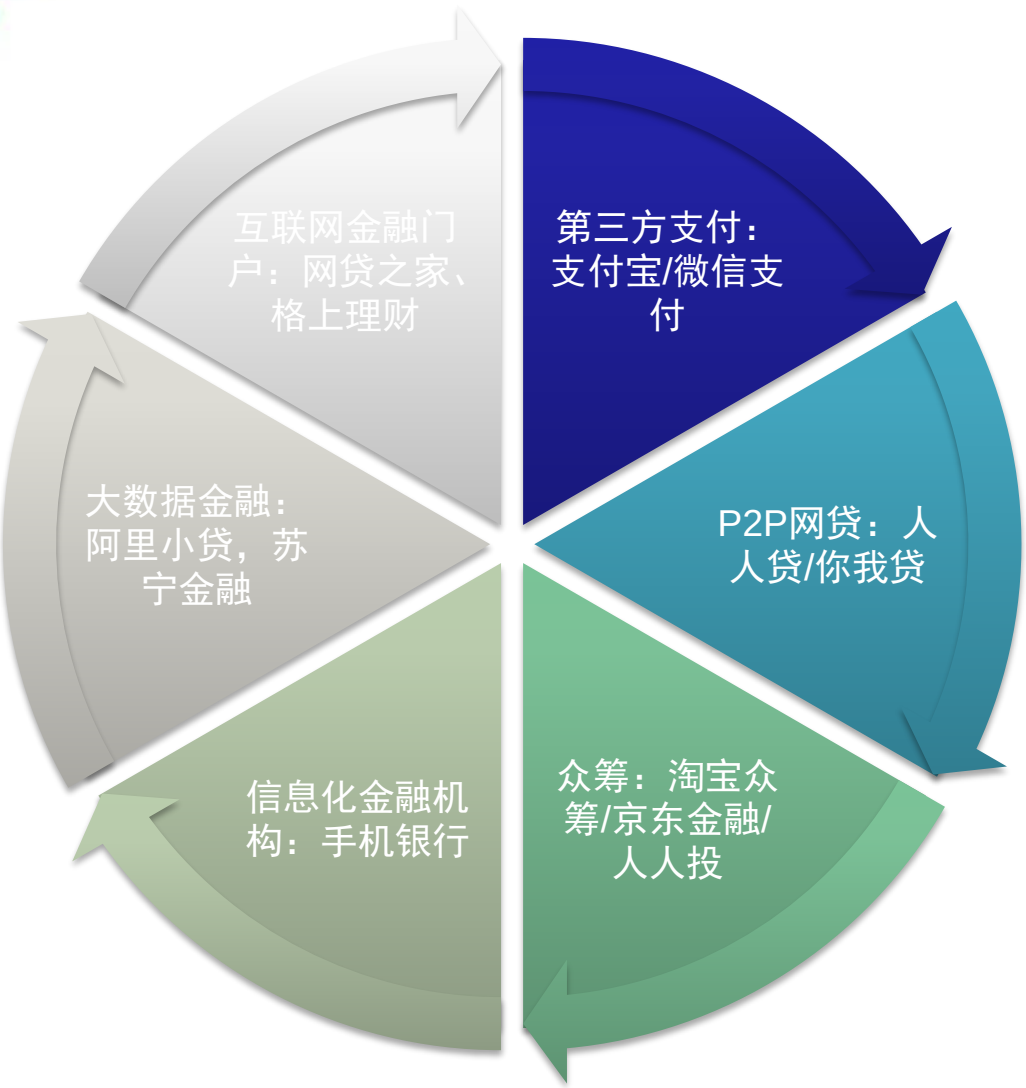
DT时代

传统金融的互联网化

互联网的金融服务

DT时代

我国互联网金融生态



DT时代

互联网金融的数据应用

IT时代

安全日志

系统日志

对账系统

管理系统

安防系统

DT时代

客户画像

个性化推荐

风控系统

反作弊欺诈

……

DT时代

大数据在个人信用体系中的应用

DT时代

大数据在个人信用体系中的应用

FICO Score

信用分数是利用数学模型依据个人的信用报告评估银行风险大小的一个数值，一般来说数值越高风险越小。信用分数的数学模型有许多种，在银行届运用最广泛的就是FICO分数（300～850分之间），这是由FICO公司拥有的一种计算模型，其具体细节并没有批露，但是其信用分数的组成已经被总结出来了，就是下面那张图。从下图可以看出， 以下几点综合，决定了您的信用分数。

DT时代

大数据在个人信用体系中的应用

计算逻辑：

1.Payment history（按时还款）：
- 各种信用账户的还款记录
- 公开记录及支票存款记录
- 逾期偿还的具体情况，包括天数、金额、次数和时长

2. Amount Owed（信用账户）：
仍需要偿还的信用账户总数、分类账户数、余额、使用率、分期付款偿还率

3. Length of Credit History（信用年限）

4. New Credit Account（新账户数）

5. Types of credit used（信用种类）

DT时代

大数据在个人信用体系中的应用

1.身份特征（15%）
公安实名认证
身份信息
信息稳定性
……
2.信用历史（35%）
信用卡还款历史
微贷还款记录
水电煤缴费
罚单
……
3.履约能力（20%）
支付账户余额
余额宝余额
车产信息
房产信息
……
4.人脉关系（5%）
关系圈
朋友圈信用水平
社交影响力
……
5.行为偏好（25%）
账户活跃度
消费层次
缴费层次
消费偏好
……

数据来源：
1.阿里的生态系统喽：阿里小贷、淘宝、天猫、B2B业务、新浪微博数据；
2.政府公共部门：公安、工商、税务、移动……等；
3.合作机构：金融机构、同业征信等

DT时代

大数据在个人信用体系中的应用

芝麻信用，代表的是基于个人消费者及小企业互联网大数据的一系列信用评估产品，解决陌生人之间及商业交易场景中最基本的**身份可信性**问题，同时帮助**识别风险与商机**。

信用数据 → **信用技术** → **场景/结果**

3亿实名用户，覆盖近一半的中国网民

涵盖购物、支付、投资、生活、公益等**上百种**场景数据

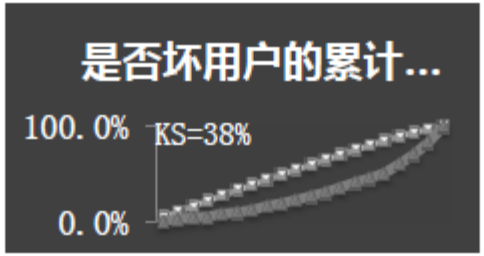
每天**PB级**数据，相当于5000个国家图书馆信息量

- 自动用户评估系统
- 用户画像信息档案
- 关系识别与评价
- 个人评分系统

账户匹配度高

互联网用户识别覆盖度60%

评分区分能力好



评分对于信用区分能力强
图为某小额贷款的场景下的评分效果校验

DT时代

大数据在风控反作弊的应用

风控概念：

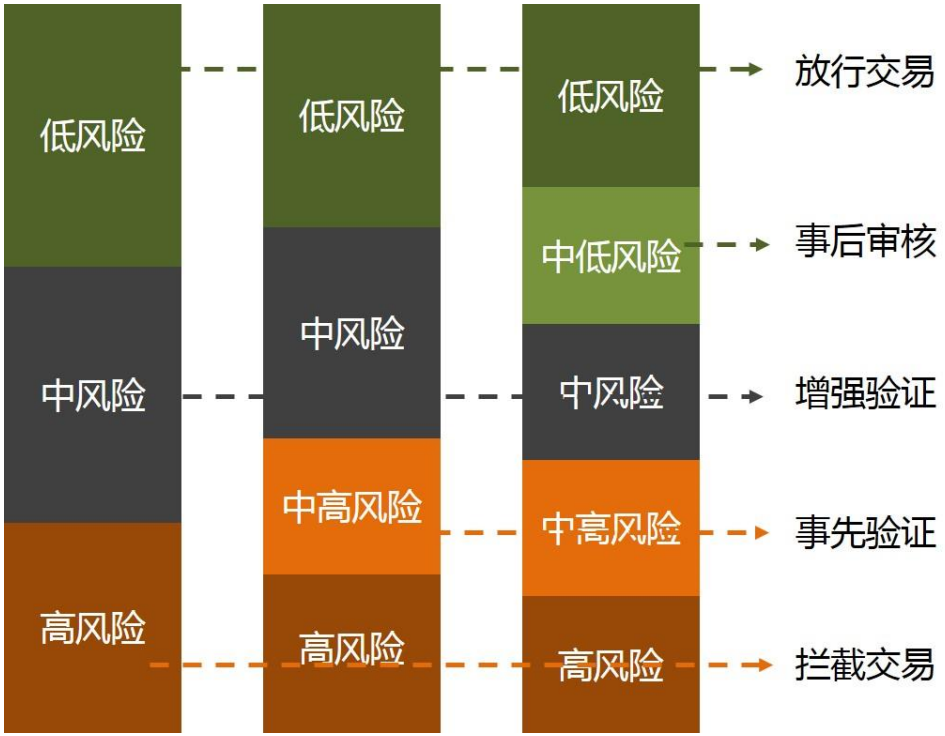
• 顾名思义，风控就是风险控制，最大程度地控制作弊和欺诈的发生，保障网站的正常运营和用户体验。支付风控涉及到多方面的内容，包括反洗钱、反欺诈、客户风险等级分类管理等。 其中最核心的功能在于对实时交易进行风险评估，或者说是欺诈检测。如果这个交易的风险太高，则会执行拦截。由于反欺诈检测是在交易时实时进行的，在要求不能误拦截的同时，还有用户体验上的要求，即不能占用太多时间，一般要求风控操作必须控制在100ms以内，对于交易量大的业务，10ms甚至更低的性能要求都是必须的。 这就需要对风控模型进行合理的设计。一般来说，要提升风控的拦截效率，就需要考虑更多的维度，但这也会带来计算性能的下降。在效率和性能之间需要进行平衡。

• 风险和作弊行为的发现、识别和处置

• 风控和反作弊是持续的博弈过程，cat-and-mouse game，时效性强，对抗性强

DT时代

大数据在风控反作弊的应用



目前主流的风险等级划分有三种方式， 三等级、四等级、五等级。

- 三等级的风险分为 低风险、中风险和高风险。大部分交易是低风险的，不需要拦截直接放行。 中风险的交易是需要进行增强验证，确认是本人操作后放行。 高风险的交易则直接拦截。
- 四风险等级，会增加一个中高风险等级。此类交易在用户完成增强验证后，还需要管理人员人工核实，核实没问题后，交易才能放行。
- 五风险等级，会增加一个中低风险等级。此类交易是先放行，但是管理人员需要进行事后核实。 如果核实有问题，通过人工方式执行退款，或者提升该用户的风险等级。

DT时代

大数据在风控反作弊的应用

三板斧：

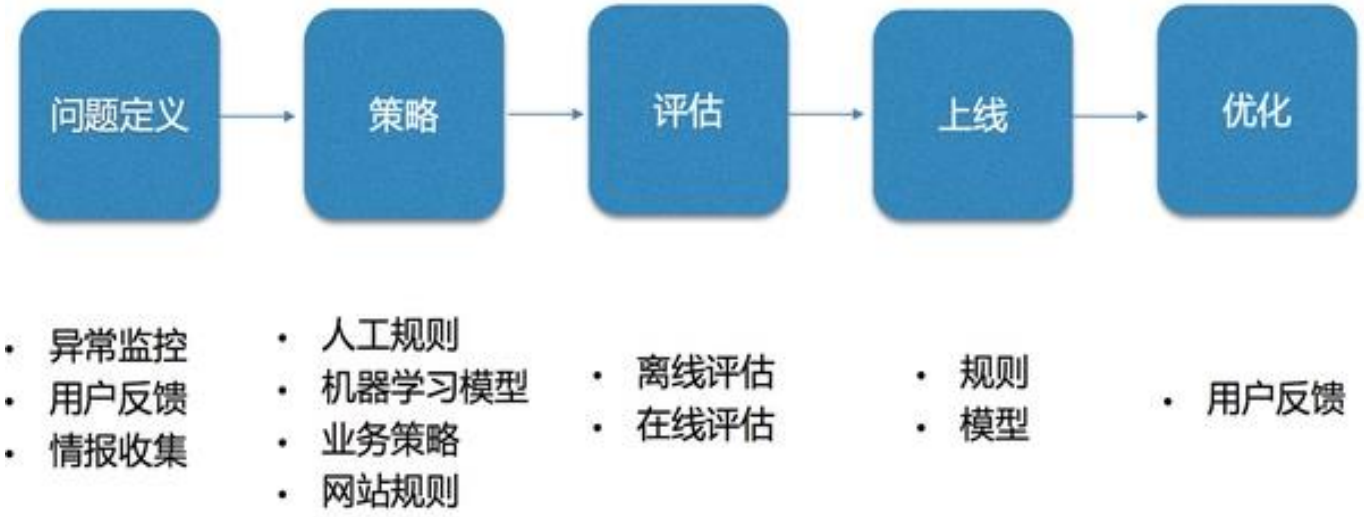
- rules；

（名单规则、操作规则、业务规则、行为规则、历史规则）

- models；

（逻辑回归、决策树、评分模型）

- strategy



DT时代

大数据在风控反作弊的应用

- 数据收集
  - 网络行为数据
  - 企业服务范围内行为数据
  - 用户内容偏好数据
  - 用户交易数据
  - 授权数据源
  - 第三方数据源
  - 合作方数据源
  - 公开数据源
- 行为建模
  - 文本挖掘
  - 自然语言处理
  - 机器学习
  - 预测算法
  - 聚类算法
- 构建画像
  - 基本属性
  - 购买能力
  - 行为特征
  - 兴趣爱好
  - 心理特征
  - 社交网络
- 风险定价
  - 申请模型
  - 行为监控模型
  - 违约模型
  - 催收策略模型

大数据在风控反作弊的应用

数据：

一般业务数据：用户、商品、交易、点击、浏览、搜索、评价、服务、处罚等
安全业务数据：设备数据（UA、cookie、MAC、Umid、IMEI、IMSI）、位置数据（IP/LBS/GPS）、行为信息、生物信息、其他

算法：

机器学习：分类、聚类、graph算法异常检测
图像算法：人脸识别、OCR、图像搜索
绝大多数场景使用RF/GBDT+LR/C5.0

注意点：
- 部分高风险业务，可以投入人力审核，追求更高的准确率/召回率
- 风险(异常)占比少，属于非平衡数据集
- 对抗意识强，模型衰减快，需要结合处置手段
- 风控的成本与回报意识，平衡人力和风险
- 能够采用更复杂的算法，但需要平衡用户体验和可解释性

DT时代

大数据在风控反作弊的应用

决策树模型

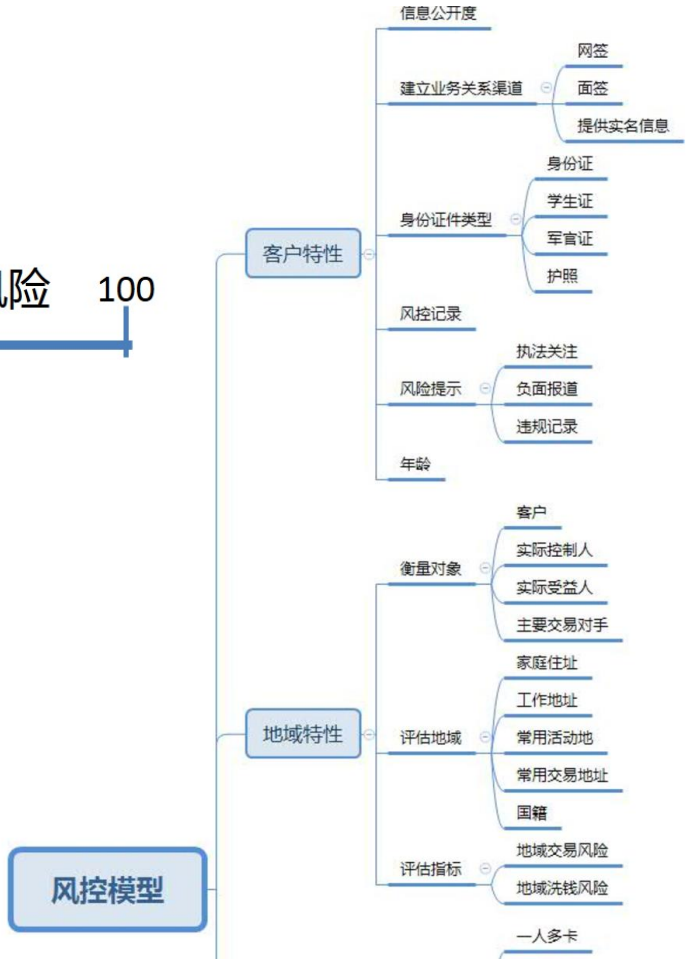
DT时代

大数据在风控反作弊的应用

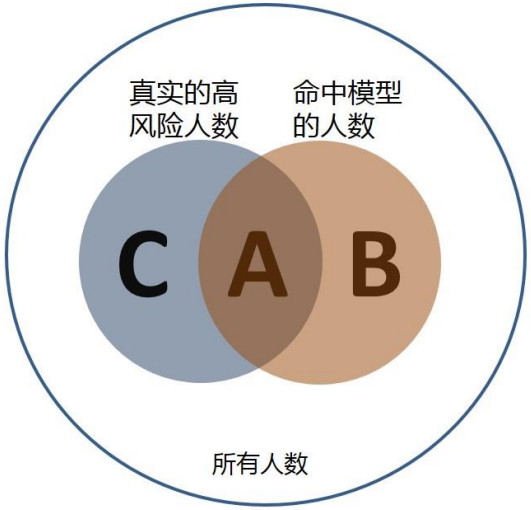
评分模型

评分模型的优势在于：
1.性能比较高，针对交易进行指标计算，按照区间来确定风险。
2.相对于规则，如果指标设置合理，其覆盖度高，不容易被嗅探到漏洞。
3.理解和分析也比较容易。 如果交易被拦截了，可以根据其各项打分评估其被拦截的原因。

存在的问题：
1.模型真的很难建立。指标的选择是一个挑战。
2.各个参数的调优是一个长期的过程。



大数据在风控反作弊的应用



非高风险

$$P_{准确率} = \frac{A}{A+B}$$

$$R_{召回率} = \frac{A}{A+C}$$

$$F_1 = \frac{2PR}{P+R}$$

以评估高风险人群的效果为例，

- Precision, 准确率，也叫查准率，指模型发现的真实的高风险人数占模型发现的所有高风险人数的比例。
- Recall，召回率，也叫查全率，指模型发现的真实的高风险人数占全部真实的高风险人数的比例。

理想情况下，我们希望这两个指标都要高。实际上，往往是互斥的，准确率高、召回率就低，召回率低、准确率高。如果两者都低，那就是模型不靠谱了。 对于风控来说，需要在保证准确率的情况下，尽量提高召回率。 那怎么发现实际的高风险人数呢？ 这就需要借助规则模型，先过滤一遍，再从中人工遴选。

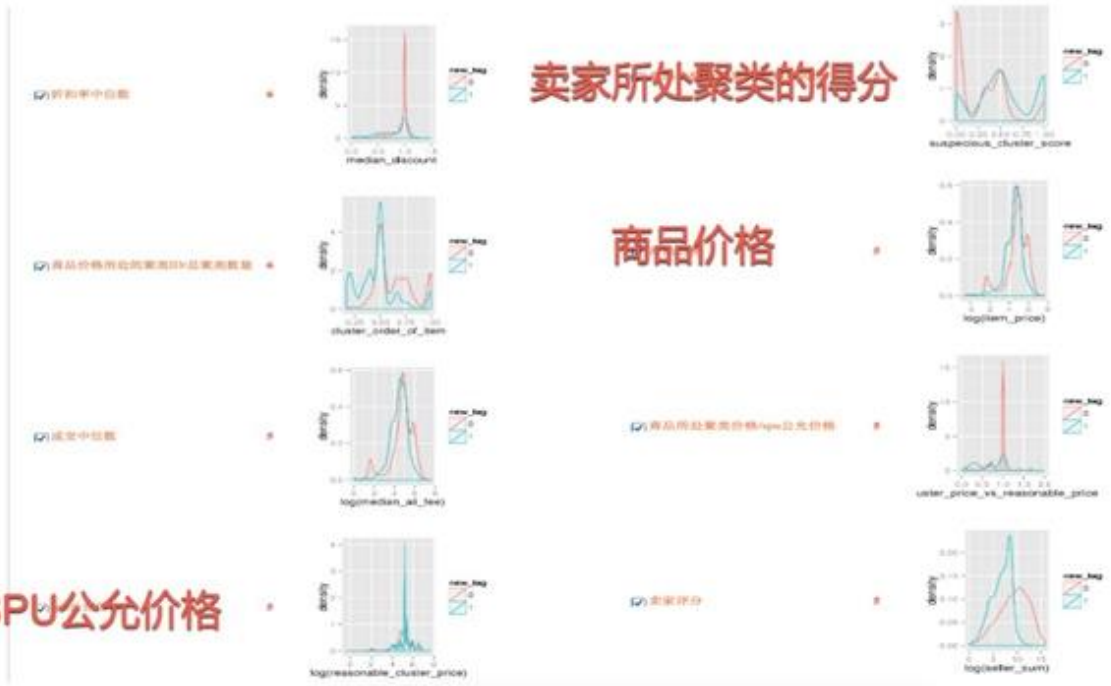
DT时代

大数据在风控反作弊的应用

DT时代

大数据在风控反作弊的应用

Case 1：如何判断某笔交易是否虚假？



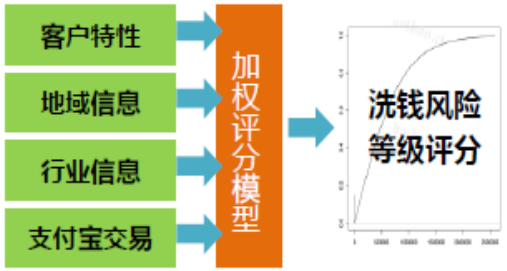
DT时代

大数据在风控反作弊的应用

Case 2：

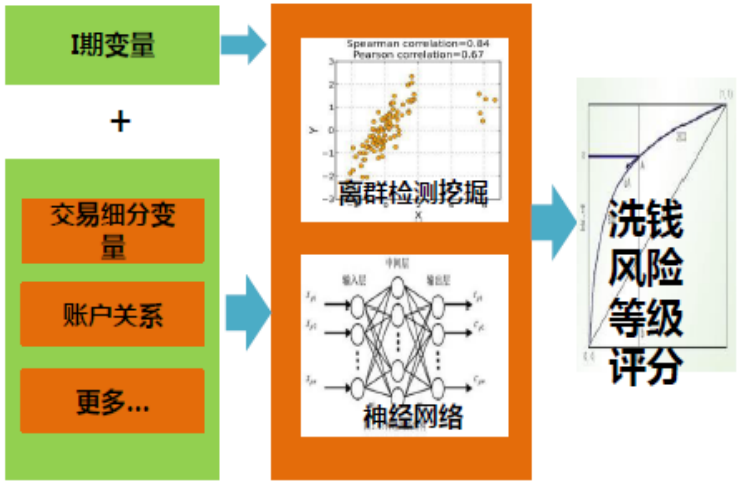
- 建立客户洗钱风险等级模型，识别高风险客户和交易行为。

模型I期：



级别	分数段	占比
低风险	[1,30)	46.4%
中低风险	[30,35)	43.7%
中风险	[35,45)	9.7%
中高风险	[40,45)	0.16%
高风险	[45,100]	0.005%
总计	**[0,100]**	**100%**

模型II期：



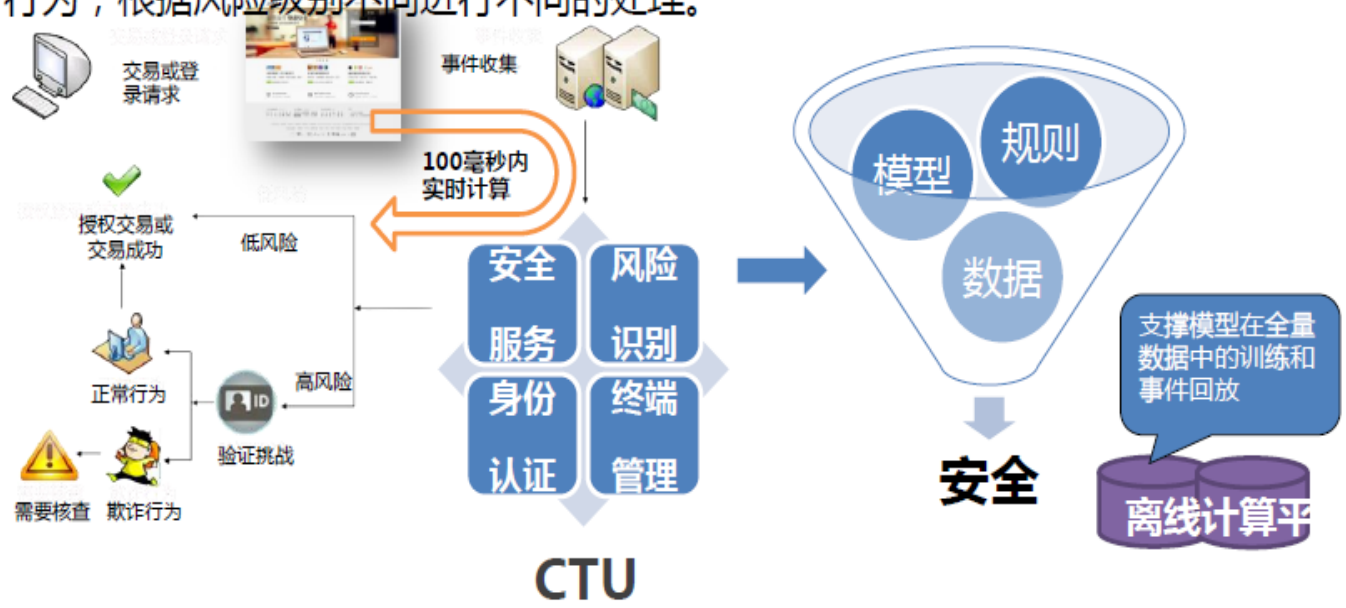
- 更多维度变量
- 更智能的方法
- 更精准的结果

DT时代

大数据在风控反作弊的应用

Case 3：

智能实时风险监控系统是目前国内最先进的网上支付风险实时监控系统之一，通过数据分析、数据挖掘等技术进行风险嫌疑数据的抓捕，发现异常的或有风险的操作行为，根据风险级别不同进行不同的处理。



DT时代

大数据在小微贷中的应用

基于大数据的信贷模式是阿里小贷的最大优势，在这个模式下，客户立即申请立即获贷，不良率低于传统银行。

客户筛选
- 交易预测模型
- 经营风险模型
- 卖家分层模型

资信调查
- 定向营销
- 批量授信

贷款审批
- 信用评级模型
- 信贷决策引擎
- 人工辅助

贷后监控
- 贷后预警系统
- 内/外催收体系
- 司法诉讼

数据 + 模型 = 高效率 + 低风险

DT时代

大数据在保险产品中的应用

- 2010年6月上线，解决退货纠纷
  - 5毛钱的保费，10块钱的保额，堪称“小而美”金融产品的典范
- 4年来，保单量以每年100%的速度增长
- 2013年全年保单量突破**10亿**单，双11当天保单出单量**1.2亿笔**
  - 覆盖淘宝全网所有实物商品类目，占实物订单量16%

DT时代

大数据在保险产品中的应用

运费险定价演进

一口价时代

- 保费按5%费率统一收
- 保额10元，保费0.5元

简单粗糙，劣币驱逐良币

精算定价时代

- 以历史出险率为唯一定价因子

解释方便，定价偏差大

数据定价时代

- 以30+因子统计建模，预测退货率

预测较准，解释性差

大数据定价时代

- 百万ID特征，实时特征
- 广告机器学习方法

预测很准，难于解释

DT时代

大数据在量化投资中的应用

DT时代

大数据在量化投资中的应用

DT时代

大数据在量化投资中的应用

从事量化交易后要考虑的几个问题：

- 如何找出合适的策略？
- 回测前辨别策略的优劣？
- 策略的回测？
- 怎么执行策略？
- 交易过程中扩大规模增加收入？
- 管理无法避免的亏损？

大数据在量化投资中的应用

- 开发工具：
MATLAB/Python/C++

- Python环境：
pandas数据统计，numpy数值处理，matplotlib绘图，sklearn机器学习库

- 编辑器：
Eclipse+Pydev

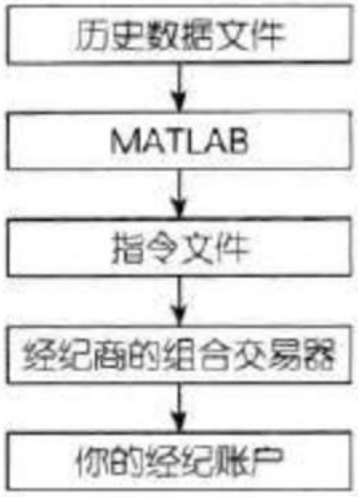
- 数据来源：
东方财富/wind/新浪财经/集思录/优矿等

- 推荐资料
https://zhuanlan.zhihu.com/p/24220361

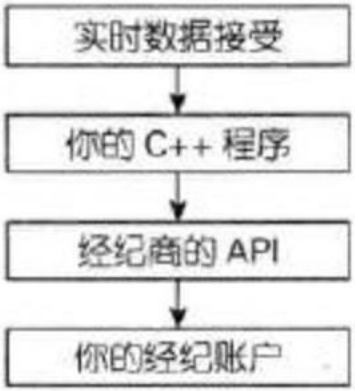
DT时代

大数据在量化投资中的应用

半自动交易系统

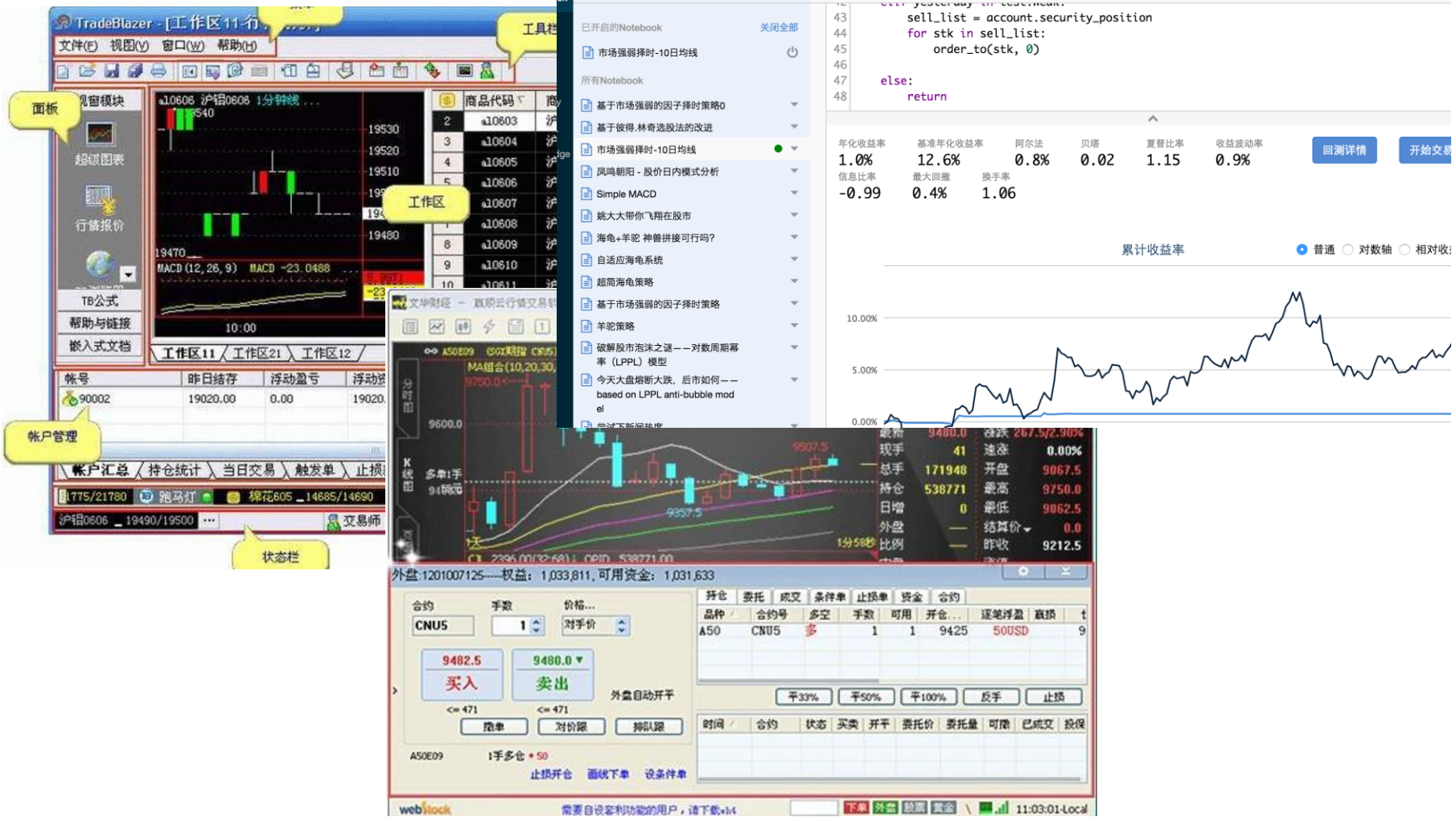


全自动交易系统



DT时代

大数据在量化投资中的应用



感谢您的关注

知乎专栏：数据分析侠