大数据在风险管理中的应用

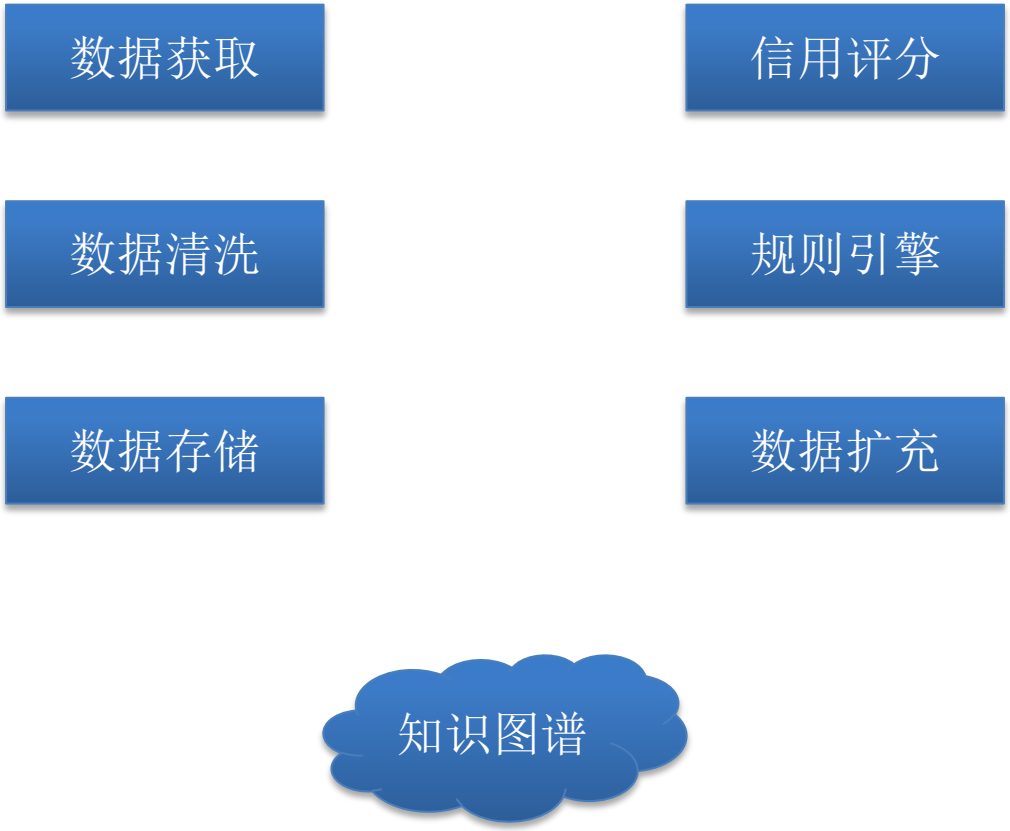
- 反欺诈
  - 借款人是否是本人？
  - 如果借款人是本人，他提交的资料是否是真的？
- 信用评估
  - 还款能力
    - 月收入是否足够覆盖还款？
  - 还款意愿
    - 是否是老赖？是否有诚信？

- 借款人是否是本人？
  - 姓名，身份证，手机号，银行卡一致
  - 人脸识别，活体检测
- 借款人是本人，信息是否真实？
  - 单位是否真实
  - 家庭地址是否真实
  - 收入，工资流水是否真实
  - ……

- 还款能力
  - 消费多是否代表有还款能力？
  - 高档消费多是否代表有还款能力？
  - 地位高是否代表有还款能力？
  - 工作好代表有还款能力？
  - 住高档小区代表有还款能力？
  - 在高档写字楼上班代表有还款能力？
  - 程序员有还款能力？
  - 微博被关注的人多有还款能力？
- 还款意愿
  - 信用卡及时还代表有还款意愿？
  - 银行贷款及时还代表有还款意愿？
  - 考试不作弊代表有还款意愿？
  - 从来不闯红灯代表有还款意愿？

- 人工信审的自动化和产能提升
  - 很多时候人是在机械的执行风险政策
  - 这些政策往往是可以量化和自动化的
  - 当数据很多时，人往往会漏掉其中有用的信息，而机器不会
  - 不同人的风险偏好不同
- 让人做人最擅长的事情
  - 收集机器无法收集的信息，比如打电话后的感受
  - 斗智斗勇……

大数据处理流程



- 用户提交申请数据
  - 当前用户提交的数据
  - 历史用户提交的数据
- 爬虫获取
  - 通过用户申请信息在网上发现的线索
  - 公司的信息
  - 各种ground truth
- 三方合作

- 数据来源非常庞杂
  - 爬虫，业务数据库（几千张表）
- 数据格式非常多样
  - 个人，公司，商铺，帖子，招聘，楼盘……
- 需要将数据统一到一个格式
  - 如何定义这个格式
  - 如何定义所有其他格式到这个格式的映射

- 定义了表达一切事务的统一数据结果
  - 接近十亿的实体
  - 上百亿的关系
- 所有风险控制的数据来源基于知识图谱
- 用Jena+HBase + ElasticSearch实现
  - Jena 用来定义Schema
  - Hbase用来存储三元组
  - ES用来索引需要查询的属性

- 知识图谱的查询 Query DSL
  - 给定一个手机号，查看关联的所有实体
  - 给定一个公司，查看关联的所有实体的某些统计信息
  - 统计不同地区的某种类型的实体的某些属性的统计信息
  - ……
- 当一个用户提交申请表后
  - 对申请表的各个字段（身份证，手机号，公司，联系人）进行扩充，找到相关的实体
  - 利用这些实体交叉验证用户的信息
  - 验证有矛盾的地方可以
    - 设计成问题，让用户回答
    - 人工电话联系客户确认

- 定义了所有风险管理的逻辑，包括
  - 要去知识图谱中查什么
  - 查出来的东西怎么用
  - 怎么交叉验证
  - 什么样叫做有矛盾
  - 做什么样的决策
- 快速验证新的逻辑
  - 当有了一个新的逻辑时，如何快速验证这个逻辑的正确性
- 如何发现新的逻辑
  - 如何基于当前的逻辑，或者一些初步的想法，去发现新的有用的逻辑

- 大数据风控的特点
  - 样本少（相对与广告，推荐系统）
    - 每个样本要通过放款来收集
    - 每个样本需要相当长时间才能收集到表现
  - 特征多，大量的missing value（相对于传统的风控）
- 这个特点造成模型很容易过拟合
  - 特征工程：较少采用原始特征，而是基于原始特征和领域知识总结出新的特征
  - 采用Random Forest训练模型
  - 多次交叉验证

- 针对电商
- 用户线上授权店铺数据
- 综合知识图谱中的大量数据
- 实时额度预估
- 实时信用评估，拒绝风险高的用户，通过风险低的
- 发现风险点，由信审人工确认（非常快）
- 实时放款

商通贷

liangxiang@creditease.cn