
面试指南之
大数据风控建模师
京东数科-个人服务群组-智能模型实验室
郑邦祺

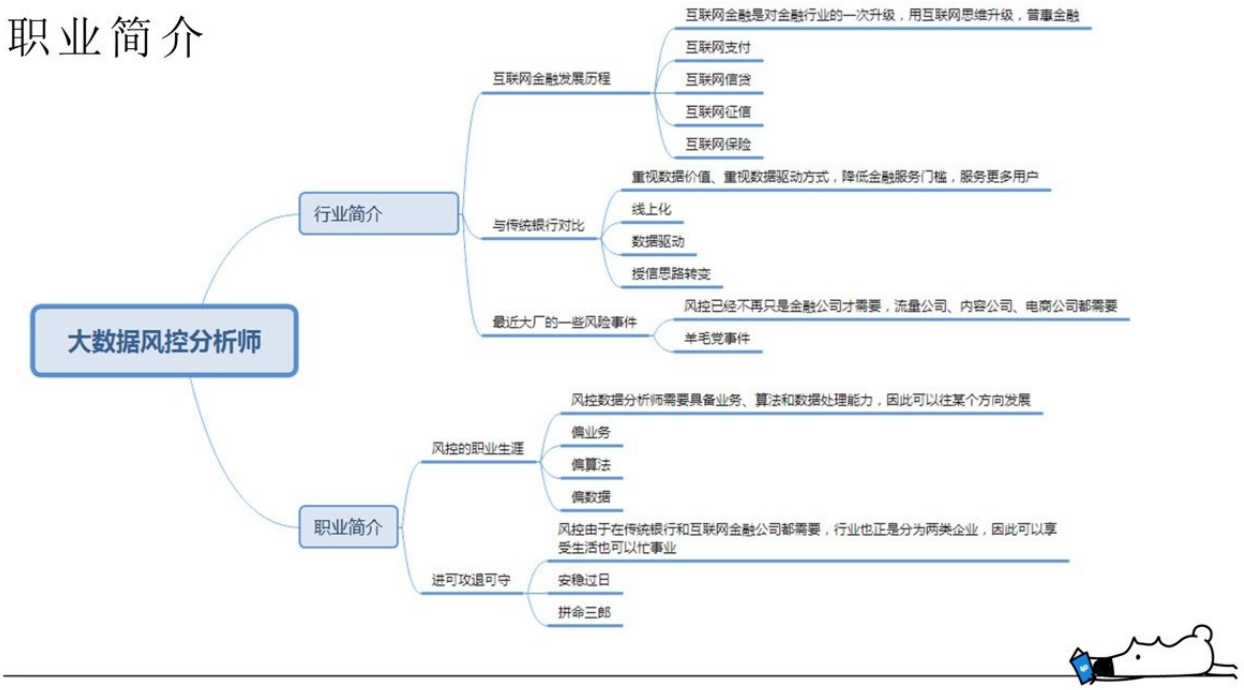

目录
• 职业介绍（主要安利大数据风控建模师）
• 面试准备（面试官在面试的时候到底在面什么）
• 核心技能准备（除了算法原理，还需要什么）

第一部分——职业简介

职业简介

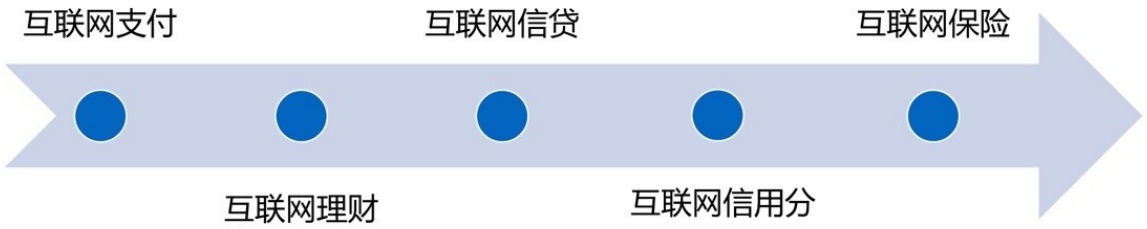
- 互金发展历程
- 互金与银行在风控方面对比
- 大数据风控建模师的职业规划

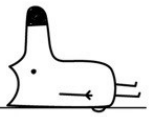
职业简介



互金发展史

互联网金融=互联网用户思维+金融服务





与传统银行相比

- **线上化 vs 线下化**
  - 获客渠道线上化
  - 数据来源线上化
  - 审批自动化
- **数据驱动 vs 经验驱动**
  - 数据驱动更能反映当下环境的一个最优解
  - 经验驱动更有指向性和灵活性
- **证明你有钱 vs 证明你能还钱**
  - 高净值人群，关于银行有句笑话，当你能证明你不缺这笔钱的时候，柜台经理很愿意把钱借给你
  - 普通大众，从日常数据证明你是正常能还钱的好人

风控职业规划

- **职业道路**
  - 偏业务

    数据产品经理，经常会和一些运营人员交流，获悉产品、行业情况
  - 偏技术

    算法工程师，需要对深度学习、无监督算法等常见算法原理了解
  - 偏数据

    大数据工程师，集群管理、数据清洗、变量加工
- **就业类型　风控-进可攻，退可守**
  - 拼事业or忙生活

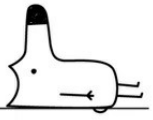
第二部分——简历准备

写在面试前

- **实习的意义**
  - 遮瑕双非、想要进入核心团队、想要避开笔试、面试准备期
  - 海龟熟悉国内技术要求，简历可以写清楚实习长短
- **实习的长短**
  - 长期：方便接触核心项目（能长期实习，可以在简历重点备注）
  - 短期：可以在简历刷一笔经验，增加简历丰富度
- **投递渠道**
  - 实习僧、学校bbs、技术社群
- **如何选择公司、如何选择实习内容**
  - 核心项目判断，很多团队都会招聘算法工程师，是偏向数据、业务还是算法
  - 转正机会判断，通过看团队的年龄来判断团队是否有校招传统、转正概率

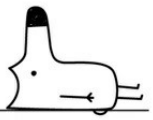
简历准备

- **个人简介**
  - 姓名、邮箱、电话、学历、求职意向
- **教育背景**
  - 985双一流可以备注，相关专业也可以备注专业，专业排名、绩点和核心课程
- **项目经验/课题经验**
  - 应届生，可以多提技术术语，和在做课题过程中尝试过的方向，以及未来的方向
  - 社招生，需要和应届生形成差异，突出项目背景，实际效果，项目尽量有深度，有延展的
- **加分项**
  - 专业技术：python、机器学习算法、hadoop框架原理
  - 数据比赛/文章：会参考，但不同团队的认可度可能不一样
  - 实际落地的项目：实习项目或者在校和老师一同完成的实际项目
  - 组织技术论坛：可以周末组织学校同学或者感兴趣的人员参与，并记录技术感悟



面试流程

- **提前批**
  - 尽量争取，小组直面，免笔试，有机会和未来同事多请教团队问题
- **笔试**
  - 诚信为本，提早准备、提前刷题（LeetCode、 Lintcode）、网上搜索历届笔试题，了解常见题目类型（排序 & 查找、链表反转、插入、二叉树遍历、动态规划、分治 & 递归、字符串、图）
- **一二面**
  - 技术面试，一般考察简历上所列的项目经验和额外技术考察点，不仅仅是算法原理（工程实现能力和算法原理）
  - 项目：提前回忆项目细节，熟悉数据量，数据字段、项目过程，项目结果
  - 技术：列出逻辑，慢慢回答，循序渐进，把知道的都说出来，展现技术栈的宽度与深度
- **终面**
  - 更看重个人素质，反映能力、思考深度、灵活性、韧性、情商
- **Hr面**
  - 对职业生涯有明确规划、对公司文化认同，突出稳定性，是否能脚踏实地

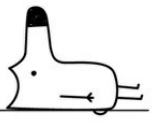


面试常见问题

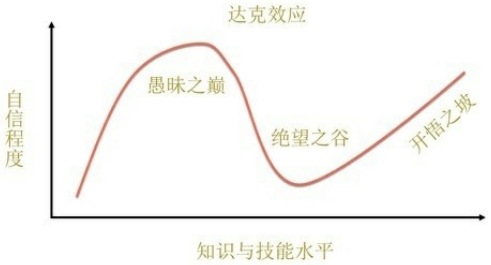
- **简单介绍下自己**
  - 在最简短的时间内，突出亮点，教育背景，学科背景、教研室背景，业余项目，注意控制语速
- **为什么要跳槽**
  - 首先回答原有岗位为什么不能满足你，技术、领导、薪资等因素，可以透露对应聘公司业务的了解
- **职业规划，为此目标当下有哪些行动？**
  - 提前思考，最好顺序回答最近一年，三年和五年规划，让面试官知道你是有规划的人
- **与上级意见不一致时，你将怎么办**
  - 坚持需要坚持的，放弃不必要坚持的
- **有没有收到其他offer**
  - 收到了其他offer大可列出来，增加信服度，没有则也可以回答这是第一家
- **你还有什么问题**
  - 别只问面试表现或者技术问题，多问组织架构、问业务情况、问团队情况、问项目情况

offer选择

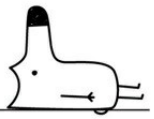
- **工作内容|技术方向**
- **怎么确认团队氛围|直接领导很重要**
  - 询问面试官联系方式，如果是提前批可能会有联系方式，正式校招比较困难
  - 脉脉/linkedin/社群
- **大公司还是小公司**
  - 不出意外，大部分应届生最适合去大公司
- **国企还是私企**
  - 两种生活态度的选择，国企可能有户口、闲暇时间也较多，生活与工作能较好平衡，私企机会多，项目多，锻炼人
- **地域选择，一线还是老家**
  - 选择安逸舒服的生活状态还是努力与拼搏的，前几年可以多去闯闯，再回到老家



职业生涯规划



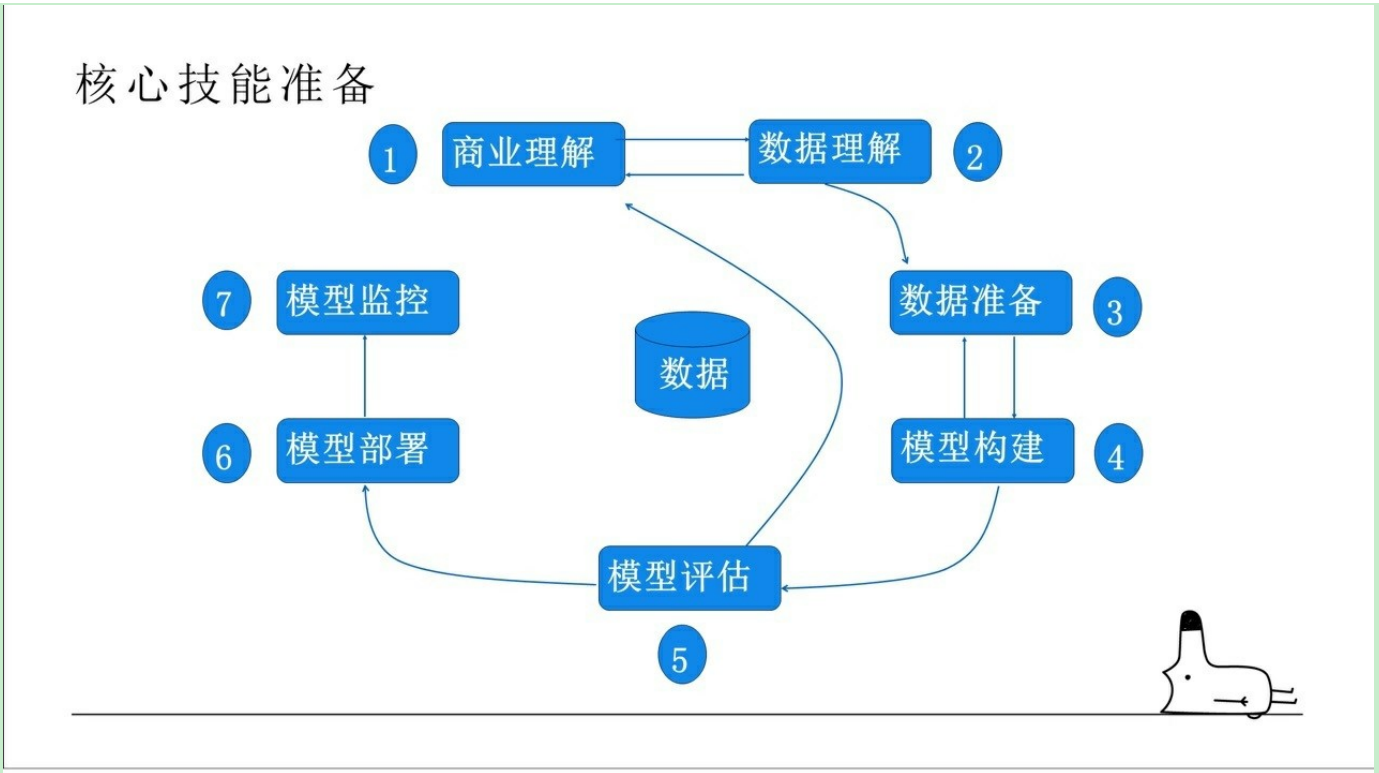
- **如何增值**
  - 核心团队、核心项目是最好的保值方法
  - 自我总结：总结项目经过、规划职业生涯、学会自我思考、要有独立思考的能力
  - 自我学习：不可否认，学习知识的转好途径还是通过项目
  - 站在行业高度：积极参加行业论坛、组织行业聚会，关注行业热点、产品、技术、动向

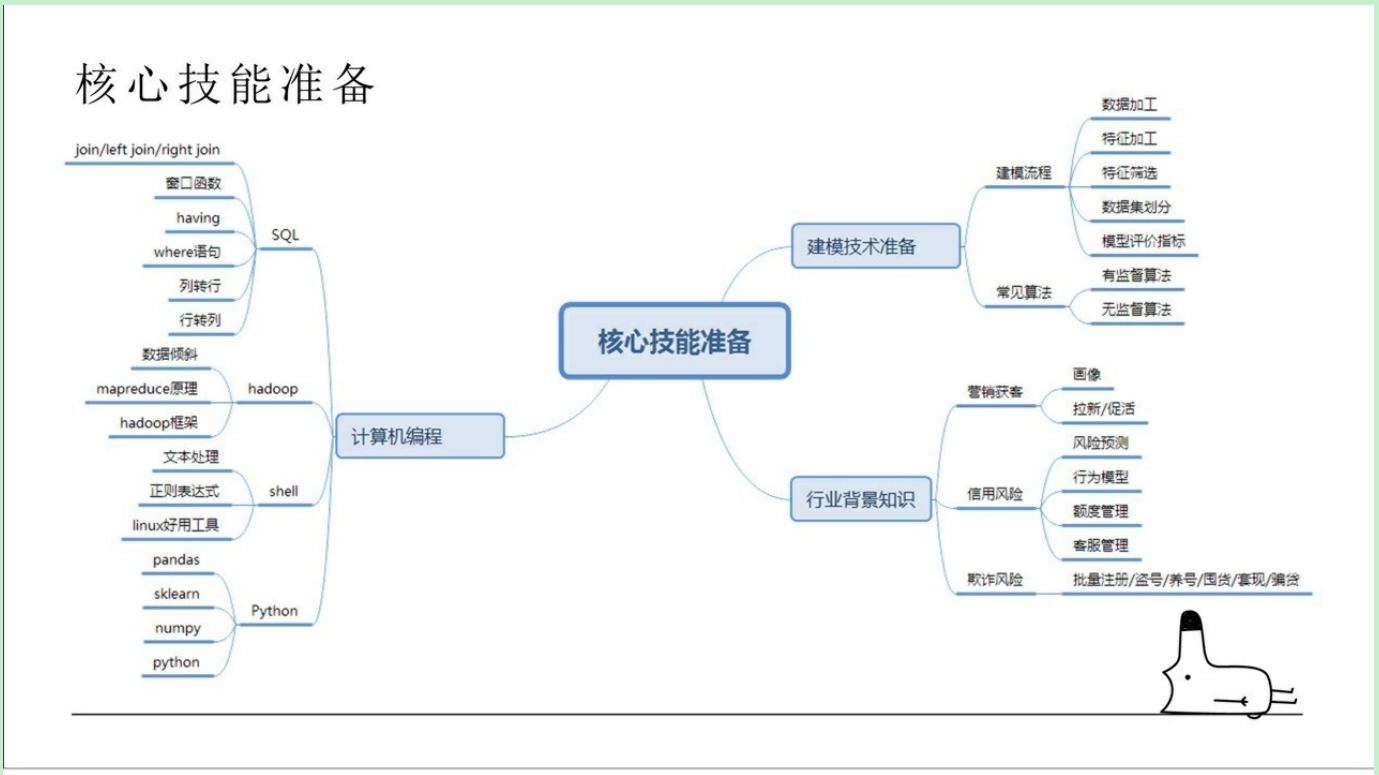


第三部分——核心技能准备

核心技能准备

- 建模技术准备
- 计算机编程
- 行业知识





建模流程与原理

1. 建模流程。

1）离群点有什么方式解决？比如有些用户是企业用户，日常消费金额高，次数也多。是否可以考虑对金额做log处理，或者woe变换。

2）异常值对xgb有影响吗？对LR有影响吗？为什么？

3）分类变量如何加工？one-hot?embeding?还是特征聚类？它们之间有什么区别和常用的方法吗？

4）样本不均衡问题如何处理？什么是上采样？什么是过采样？还有其他哪些采样方式？如果使用聚类方法和其他无监督方法加工特征时，还需要做什么预处理加工呢？

5）有哪些常见的去量纲的方式？

2. 特征工程

1）从实体角度看来特征。

一般以预测问题参与实体为基准。

例如预测下单模型，这种场景，就使用用户，商品，用户与商品的关系，每种实体都可以思考有什么属性可以挖掘？

用户维度。就有 最近xx天，月，年的白条使用次数，金额占比，账龄，活跃天数等。

商品维度。商品价格，评论，品牌。最近1个月，xxx 白条支付的次数和金额等等。

用户与商品交互维度。下单的时间，下单的优惠信息，下单地址。

任何场景建模，都要考虑清楚有哪些实体。再看看实体有哪些可用的属性。

2）数据维度。

流水数据。例如：消费订单，一般按时间维度汇总，比如xxx天，月，年。金额、次数、最大值、最小值、均值、方差、

序列数据。例如 点击流数据，需要对关键动作进行分析，注册、登录、浏览商品页、修改资料、加工的时候，可以关注间隔的时间、次数、路径。

例如 注册 xxx天后 修改资料，修改资料 xxx天后购买商品，登录xx天后修改资料。

属性数据。年龄，性别，籍贯等。直接使用交叉特征，或者one-hot。比如高富帅

3. 特征筛选

1）特征降维。xgb,lr对模型变量重要性进行输出。使用PCA，深度学习，单变量筛选的方式。

有哪些常见的单变量筛选指标？单变量筛选和模型重要变量有哪些区别？各自适应哪些场景？

2）数据集化。一般划分 train,test,valid，一般特征带有时间属性的时候，valid按照时间线进行训练。防止时间穿越，查看过拟合。

3）模型评价。验证集 AUC，KS，PSI对多个模型进行比较。在欺诈场景是否应该用AUC评价模型好坏？

4）转化成业务指标。点击率，转化率，下单率，DAU,MAU等指标。这些指标往往比模型评价指标更有用。

模型提升一个点，要转化成业务指标，如通过率等。人数每周增加1000K，DAU提升1个点。让项目可以量化。

问题：如果线下指标提升很高，但线上甚至没有提升，这可能的原因是什么呢？

建模样本和实际样本是有偏差的，或者前序模型的影响，导致样本有偏差，或者有一些营销活动，多种情况，导致下线指标不错，而线上商业指标一般般。

4. 算法原理

SVM，KNN，朴素贝叶斯，xgb,lgb,fm,deepfm，要了解原理和核心超参数。常见考点

1）有监督算法。XGB和LR的区别。XGB和GBDT区别，XGB和RF区别，什么是wide and

deep，过拟合和欠拟合分别调整哪些参数？xgb的核心参数有哪些？

2）无监督算法。K-means，孤立森林，社区发现，知识图谱。无监督学习要多关注问题，比如k-means如何构建特征空间？能否使用fund框架对特征进行加工？

30分钟学习正则表达式。

建模流程与原理

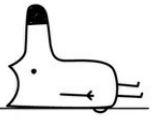
- **建模流程**
  - 数据处理、一些缺失值、离群点、分类变量加工方式、样本抽样、特征去量纲的小技巧
  - 特征筛选、降低模型复杂度，提升速度与稳定性
  - 数据集划分、建模新手最易遇到的问题，时间穿越问题可以了解下
  - 评价指标、你会从哪些角度对一个模型进行评价
- **特征工程**
  - 流水数据
  - 序列数据
  - 属性数据
- **常见算法**
  - 有监督算法、老生常谈，xgb、lr、wide&deep、cnn
  - 无监督算法，有志向做反欺诈或者其他高对抗性的项目的同学可以了解下

建模流程与原理

- **建模流程**
  - 对于离散特征，例如城市、商品品类，该怎么用特征表示
  - 线上和线下效果不一致可能的原因，怎么解决
  - 如果模型训练正负样本不平衡该如何处理，什么是上抽样、下采样，还有哪些方式？
  - 分类模型输出值越大代表什么，两个分类模型的数值可以对比吗？
  - 怎么进行变量筛选，用单变量筛选合适吗
  - 如果是为聚类或者其他无监督算法加工特征，需要做什么预处理加工？有哪些常见去量纲方式？
  - 什么是过拟合现象，有什么常用解决方法，什么是欠拟合现象，怎么判断是过拟合还是欠拟合
  - 什么是auc、psi、ks，怎么计算他们，还有哪些常见模型评价指标？
- **常见算法**
  - xgboost和随机森林、gbdt原理对比
  - Xgboost有哪些超参数，一般你会调节哪些参数，过拟合怎么办
  - 深度学习和xgboost的优劣势
  - 聚类个数选择方法
  - 聚类是不是特征个数越多越好

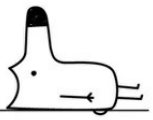
基本编程能力

- **SQL**
  - Join/left join/having/where/group by
  - 窗口函数，row_number、rank、lag、FIRST_VALUE
  - 列转行，
  - 行专列，
- **Hadoop**
  - 数据倾斜、为啥hive的reduce阶段一直停留在99%
    Join函数的map和reduce过程
  - Mapreduce原理
    map、reduce、shuffle过程
  - Hadoop 框架
    数据采集、数据存储、数据计算（实时、离线）、数据查询和其他框架



基本编程能力

- **SQL**
  - 用户流水交易表（包含用户、交易金额、交易时间），如何选择当月消费金额最高的用户
  - 用户流水交易表，如何按月和用户汇总交易额、交易次数
  - 用户流水交易表，如何算出每笔交易，占总交易额的比例，并选出超过1%占比的交易流水
  - 用户流水交易表，如何新增一个字段，指示该笔交易是否凌晨交易
  - 用户流水交易表，如何保留用户每月最后三笔交易
  - 用户流水交易表，如何筛选一个用户相邻两笔订单的交易间隔超过1年的订单
  - 用户流水交易表，如何分层抽样每月消费流水，每月各保留1w条记录
- **Hadoop**
  - 数据倾斜、为啥hive的reduce阶段一直停留在99%，怎么解决
  - 当join两张表的时候，有哪些场景的优化方式，什么是map端join
  - Join过程的Mapreduce原理
  - 有哪些hive sql的优化技巧
  - 如何提升hive sql的map个数，reduce个数
  - 怎么查看hive数据表每个分区数据大小
  - 怎么查看job的占用资源情况
  - 什么是快照表，什么是拉链表，各有什么优势



基本编程能力

- **Shell**
  - 文本处理
    awk、sed、grep
  - 正则表达式
    特殊字符、非打印字符、限定符和定位符
  - 好用工具与命令
    重定向、管道、screen、nohup、vim、shell
- **Python**
  - Pandas
  - Numpy
  - Sklearn
  - Python

基本编程能力

- **Shell**
  - 如何在文件的第四行后添加一行
  - 将 /etc/passwd 的内容列出并且列印行号，同时，请将第 2~5 行删除！
  - 搜索 /etc/passwd有root关键字的行
  - 将文件中第一个tab键替换为逗号，将全部tab替换逗号
  - 在当前目录中，查找后缀有 file 字样的文件中包含 test 字符串的文件，并打印出该字符串的行
  - Awk、sed、grep之间的区别，和适合的场景
  - 如何脱机执行程序，定时执行脚本
- **Python**
  - 在python中如何创建包含不同类型数据的dataframe
  - dataframe将日期列和时间列进行拼接
  - 如何对dataframe进行缺失值填充
  - Sklearn如何划分训练集和测试集
  - 如何对离散特征进行woe变化
  - 如何利用numpy确定分位数
  - 如何对array进行等频、等距分箱
  - 大数据的文件读取，迭代器遍历：for line in file
  - 如何在python中复制对象，copy与deepcopy区别

行业知识

- **营销获客**
  - 画像体系
    - 客群画像：蓝领、白领、学生、宝妈
    - 地址画像：工作地址、家庭地址、居住商圈
    - 资产画像：收入、房车、消费力、消费等级
    - 学历画像：大学、大专、985院校、211院校、硕士、博士
    - 设备画像：设备价值、设备指纹、风险设备、生物探针
    - 共债画像：多头注册、多头借贷、借旧换新
  - 促活、拉新
    - 响应率模型
    - 用户价值模型
    - 生命周期管理
    - 偏好预测模型

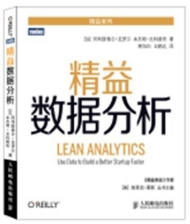
行业知识

- **信用风险**
  - 风险预测
    - 申请评分卡
    - 行为评分卡
    - 催收评分卡
  - 行为模型
    - 多头预测模型
    - 借新还旧模型
  - 优化模型
    - 额度管理
    - 人效管理
- **欺诈风险**
  - 场景：批量注册/盗号/养号/囤货/刷单/套现/骗贷
  - 技术：手机墙、模拟器、改机软件

参考书籍


MANAGING A CONSUMER LENDING BUSINESS
消费金融真经


机器学习
MACHINE LEARNING
机器学习


精益
数据分析
LEAN ANALYTICS
O'REILLY


O'REILLY
流畅的
Python


Hive
编程指南
O'REILLY