大数据风控实践

杨青
技术委员会执行主席

度小满金融
Du Xiaoman Financial

QCon+ 案例研习社

扫码学习大厂案例

学习前沿案例，向行业领先迈进

40 个
热门专题
—
行业专家把关内容筹备，
助你快速掌握最新技术发展趋势

200 个
实战案例
—
了解大厂前沿实战案例，
为 200 个真问题找到最优解

40 场
直播答疑
—
40 位技术大咖，每周分享最新
技术认知，互动答疑

365 天
持续学习
—
视频结合配套 PPT
畅学 365 天

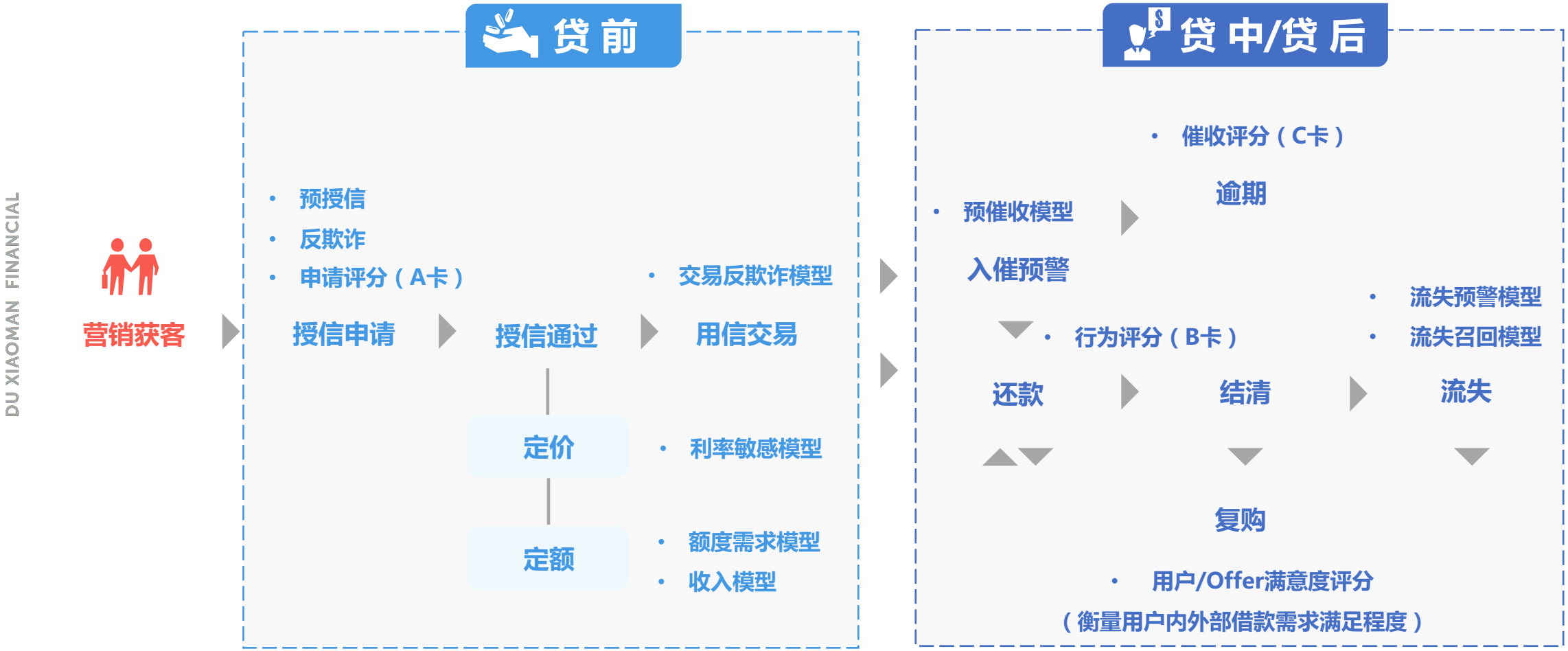
DU XIAOMAN FINANCIAL

目录

Catalogue

Du
Xiaoman

信贷风控流程



大数据风控挑战

复杂场景下的可信赖智能金融风控服务挑战

面临挑战

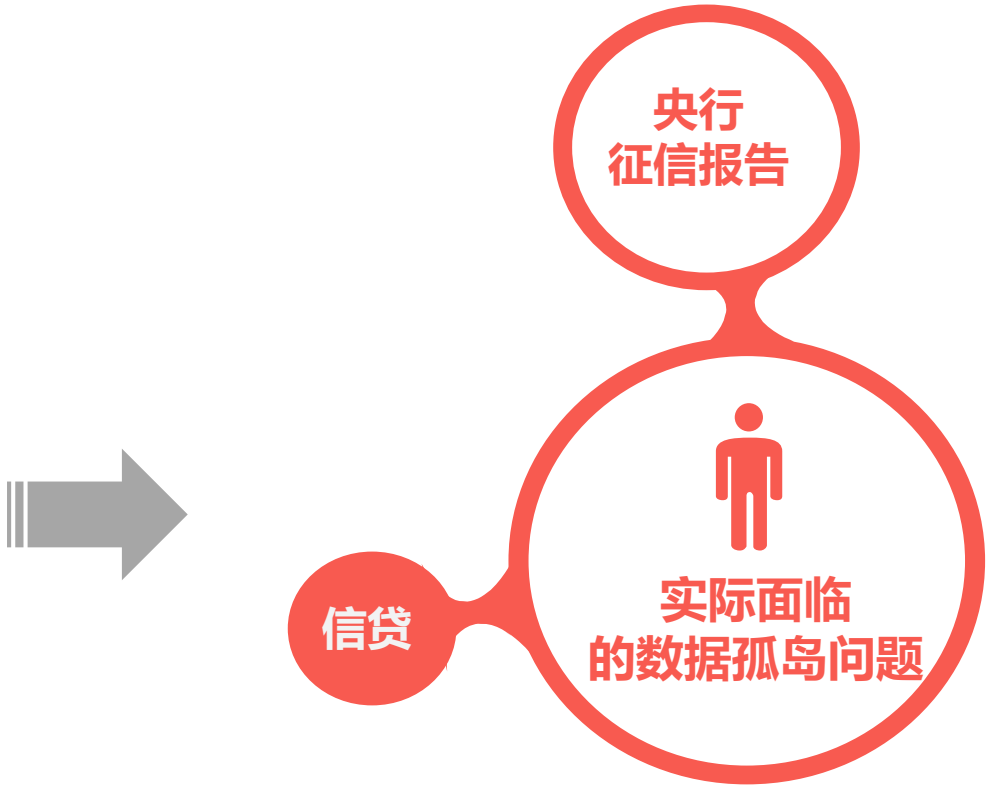
- 数据孤岛
- 非结构化信息
- 复杂模型
- 内外部多场景个性化建模需求

解决方案

- 联邦学习
- 多源融合
- 可解释性
- Auto-ML

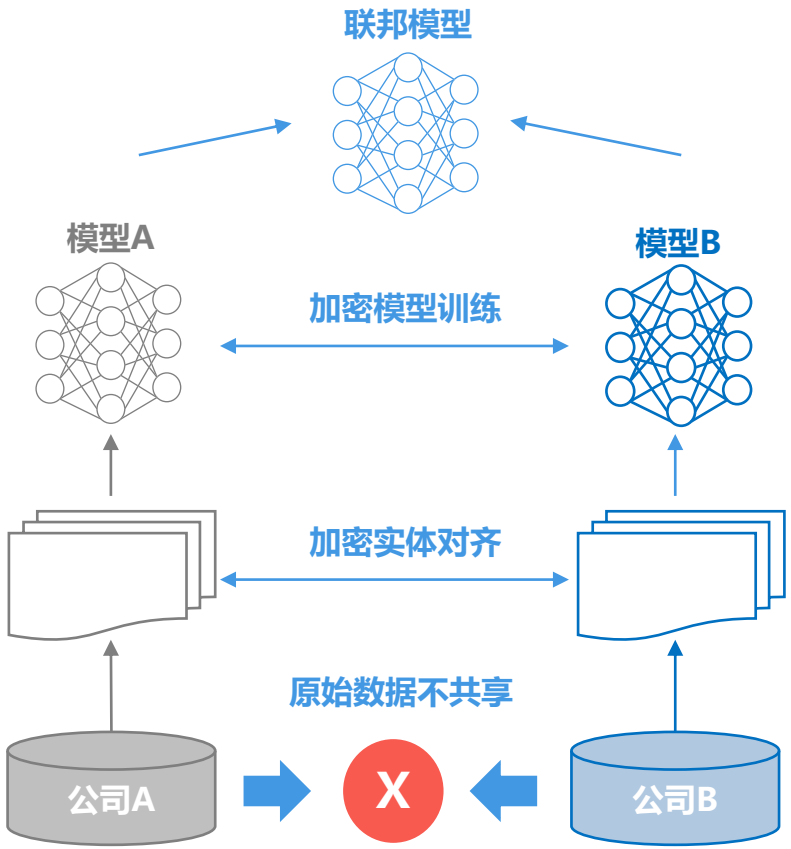
数据孤岛

全面的个人还款能力及风险行为数据



- 数据隐私保护趋严，更高的安全要求
- 同行业数据难以共享
- 跨行业数据价值无法发挥

联邦学习技术



纵向联邦学习系统架构

互联网+获客
联合客户价值建模

- 用户响应提升
- 营销效率提升

互联网+信贷
联合信贷风控建模

- 数据合作壁垒下降
- 模型效果提升2个ks

	模型效果	隐私保护	开放生态
原始数据/标签	建模没有数据方面障碍, 理论效果最好	泄漏原始数据, 数据未脱敏	一般见于私有项目
通用评分	由于客群等的差异，可能效果比较一般	基于合规要求加工产出的脱敏数据字段或模型分	不开放
联合实验室	效果受限于参与联合实验室建模的样本规模和特征丰富程度	数据出库, 有数据安全风险	双方提供特征和样本数据, 在共享的实验室环境联合建模产出高维子模型分输出
联邦学习	相同样本和特征条件下, 联邦学习建模和放在一处建模效果相同, 或相差不大	各方数据都保留在本地, 不泄漏隐私也不违反法规	共有模型, 共同获益

非结构化信息

DU XIAOMAN FINANCIAL

大数据的特点

数据体量巨大　　数据类型繁多

价值密度低　　数据结构复杂

如何从海量信息中挖掘出能区分用户风险、发掘用户需求的信息是度小满大数据风控一直不断努力的方向

结构化信息

性别、年龄、学历、婚姻状态
资产负债特征

文本数据信息

文本相关的信息(家庭地址、公司名称),
挖掘其中有价值的文本信息

图网络信息

同一位置、同一城市、同一公司等构成不同网络

时序信息

用户风险敞口、负债压力、资金需求等的都随时间发生变化,捕捉其中信息和规律,预测未来趋势.

图像信息

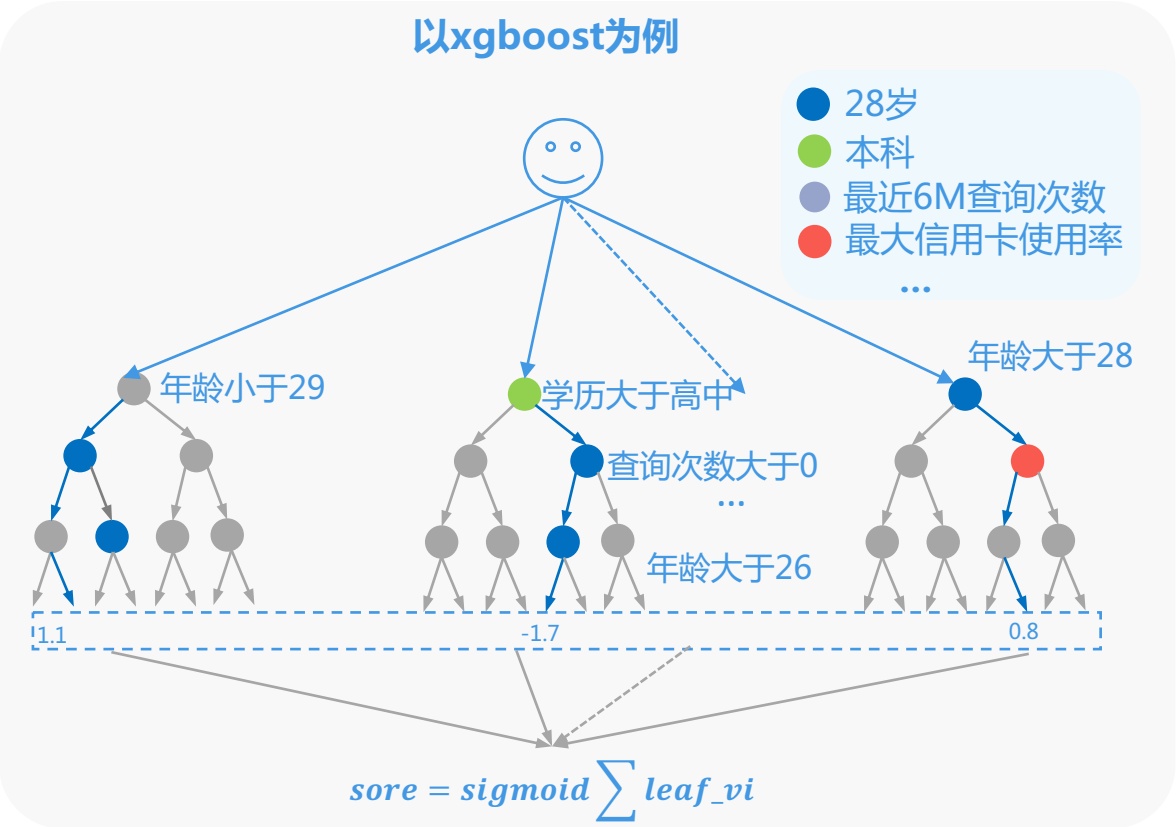
身份证照片、学历扫描
保单照片、营业执照照片

语音信息

信审电话记录
催收电话记录

复杂AI模型下的可解释性

	模型形式	特征与目标	变量特点	风险区分效果
传统金融模型	逻辑回归	独立/强相关/线性组合	少、数值	一般
复杂AI模型	Xgboost/FM/DNN	不独立/弱相关/高阶组合	多、数值、文本、序列	强



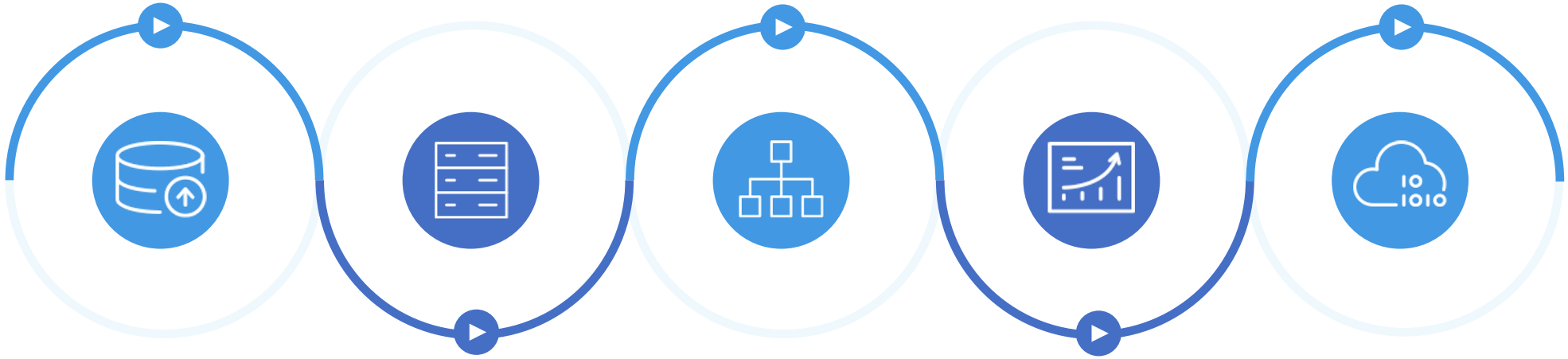
特征	分裂次数	基尼增益	覆盖
年龄	6	612.2	41184
学历	12	2812.7	127451.1
最大信用卡使用率	41	4493.1	185665.5

容易获知：主要特征列表以及基于统计值的重要性
难以判断：该用户为什么打分高或低，年龄对该用户判断风险起到多大作用

- 复杂但可解释模型（EBM(2019)、TABNET(2019)）
- 基于代理模型或工具的模型诊断（LIME(2016)、SHAP(2017)）
- 通过挖掘特殊例子来解释模型的行为（Alibi(2020)、DiCE (2020)）
- 神经网络的模型解释，例如基于梯度或热图的方法（CAM(2016)、exBert(2019)）

AutoML助力建模

利用自动建模技术提高建模效率及建模效果



数据管理

- **基准测评**
  维护标准数据集
  支持人工模型评估

数据处理

- **自动特征工程**
  特征工程全自动化
  无需人工参与

算法选择

- **丰富模型库**
  XGBoost/LightGBM
  /DNN等基分类器
- **支持多种框架**
  TF/Torch/MXNet等

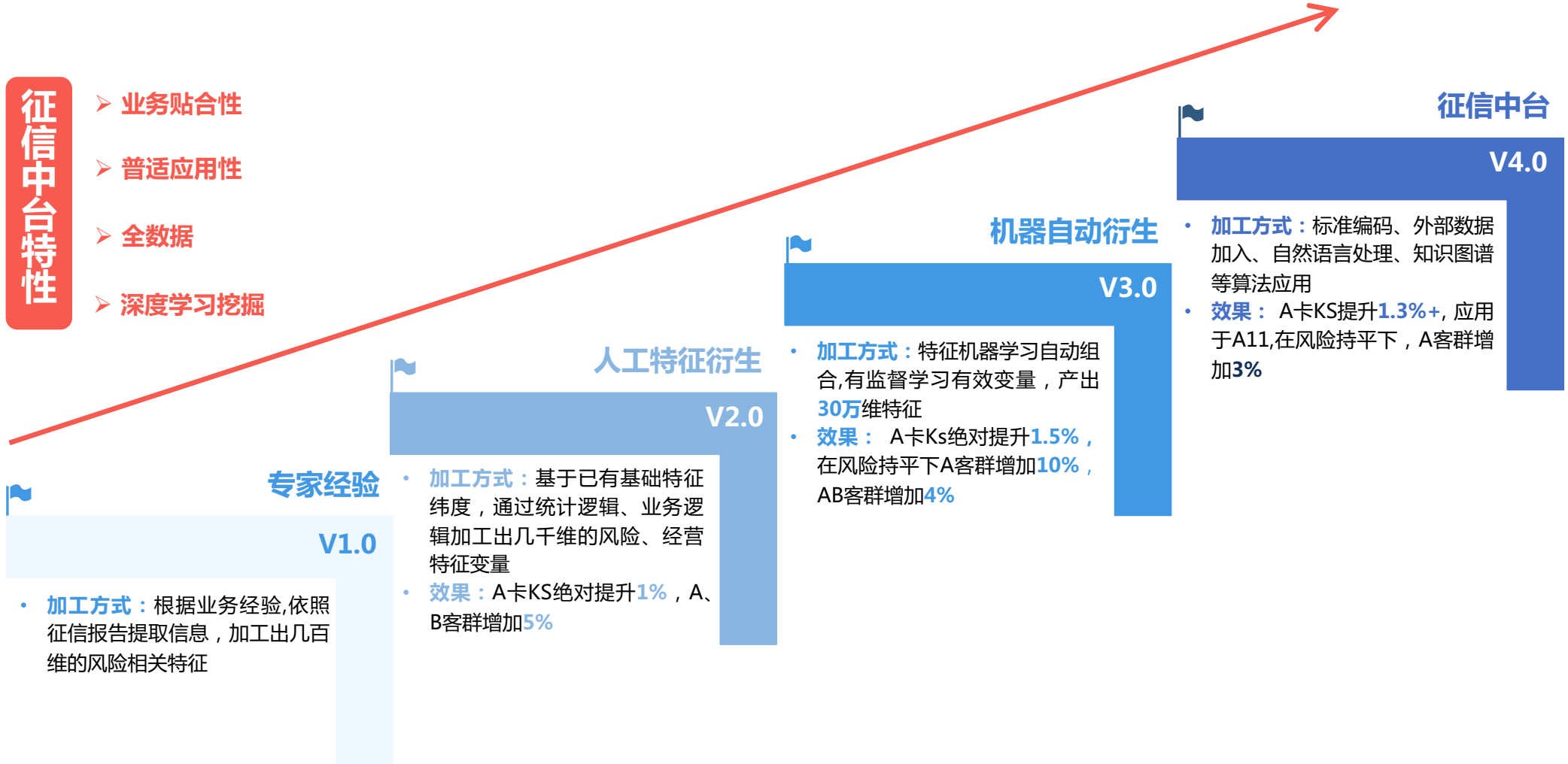
自动调优

- **超参寻优**
  TPE/ASHA等先进方
  法搜索全局最优解
- **高效调参**
  GPU硬件和分布式计
  算对调参加速

发布上线

- **跨语言跨平台**
  输出标准描述模型
  支持Python/C/JAVA
  Windows/Linux
  多语言多平台

征信报告解读：技术发展趋势

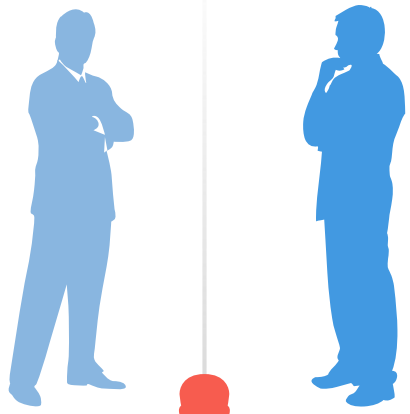


基于业务经验的特征加工

DU XIAOMAN FINANCIAL

用户还款意愿

- 贷款账户近3个月最高逾期期数
- 贷款逾期至今天数
- 连续逾期期数
- 贷款最长逾期月份数

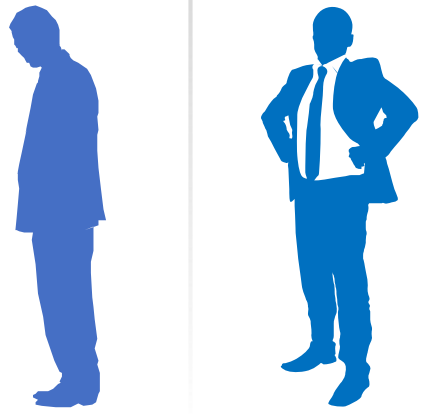


用户满意度

- 等额本息利率
- 信用卡额度最大值
- 贷款非结清房贷车贷最近12个月总期数的均值
- 所有贷款授信额度总和

用户还款能力

- 公积金月缴存额
- 学历
- 房贷车贷总期数的均值
- 用户负债总额



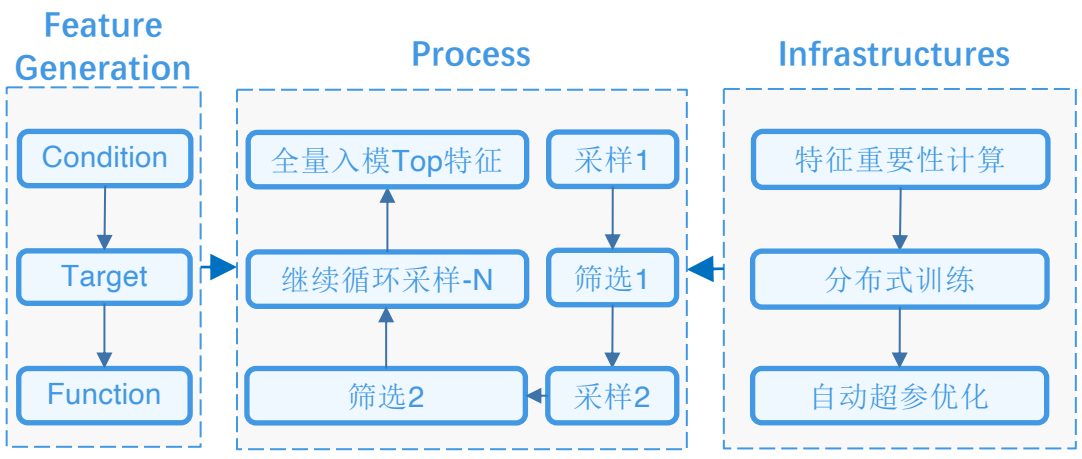
用户借贷需求

- 信用卡审批查询次数
- 查询机构数
- 最近两年贷后管理查询次数
- 最近借款频率

机器自动衍生背景及范式介绍

DU XIAOMAN FINANCIAL

机器自动衍生



征信范式举例

模式	Condition(贷款状态+还款类型)* target(还款金额)* fun(均值)
特征名	zx_l_isNoCloS_isRepayModeMth_getMthActRpyAmt_mean
特征含义	征信_贷款_未结清_按月归还_本月实际还款额_均值

征信范式

Condition

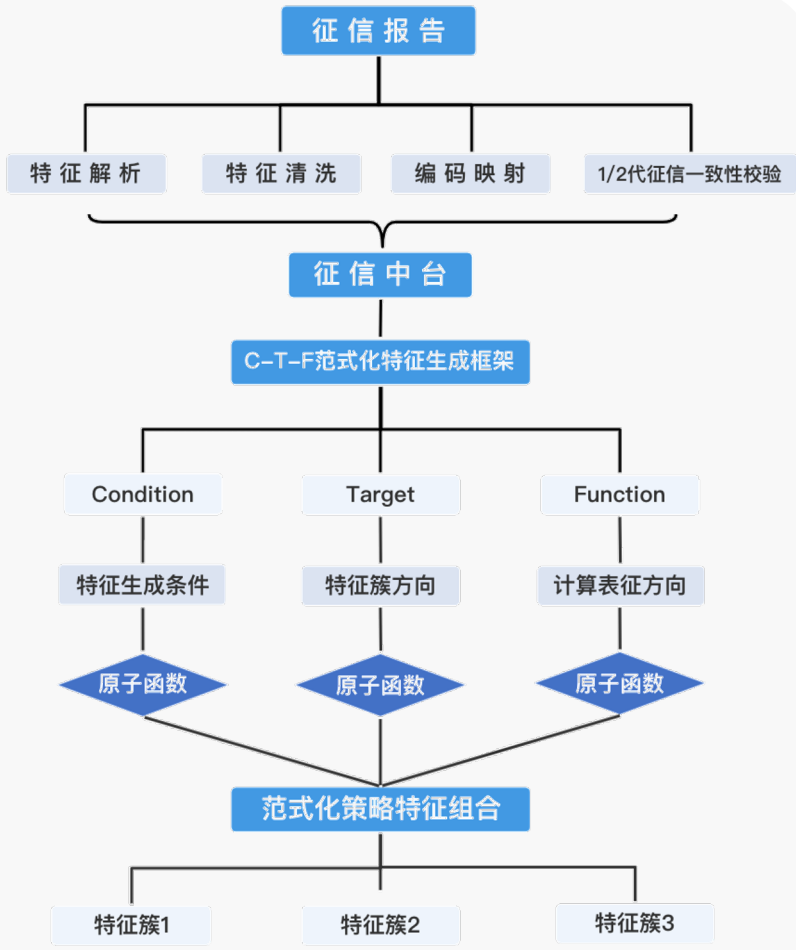
贷款状态
贷款机构类型
贷款类型
时间周期

Target

本月应还金额
本月实还金额
贷款金额
贷款期数

Function

Sum
Max
Min
Average



征信中台

聚焦央行征信报告所有信息

致力于描绘全面、精准、高效的用户画像

帮助信贷、支付、保险、金科、理财定位用户风险、挖掘用户需求、支持业务全方位发展

DU XIAOMAN FINANCIAL

“C-T-F”特征变量库

- Condition
- Target
- Function

机器学习算法

- 知识图谱
- 时序特征挖掘
- 地理位置稳定性探索

200+	30万+	1万+	20%	5%	1+
类别型数据标准编码	衍生变量	风险模型入模变量	NLP效果提升	相同风险条件下授信通过率提升	标准化工程开发平台

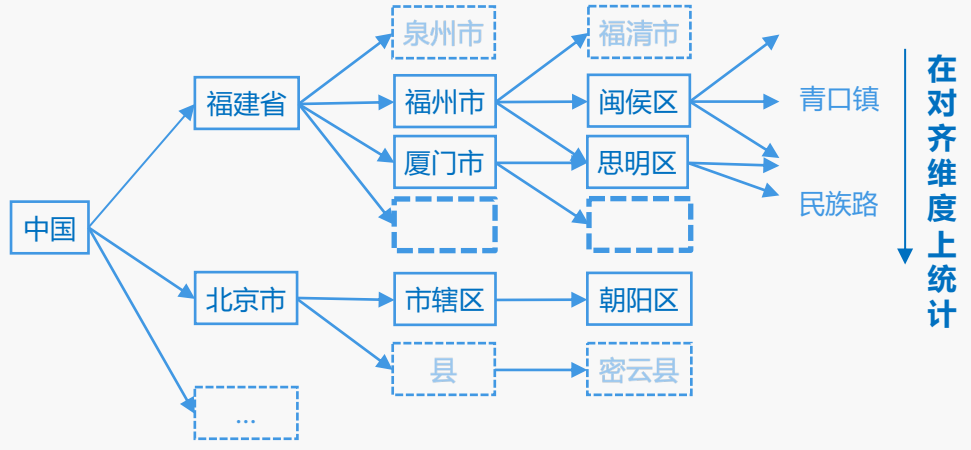
地址信息挖掘应用

DU XIAOMAN FINANCIAL

原始数据

编号	居住地址/单位地址	更新日期
1	福州市闽侯区尚干镇五虎路XX号	2006.09.26
2	福建省福州市闽侯县青口镇沪屿街XX号	2010.11.30
3	厦门市思明区民族路XX号	2018.06.29

事实数据（维度归一化）



居住地稳定性

- 居住省份数
- 城市数
- 不同区数量
- 最长居住时间

公司稳定性

- 不同公司个数
- 最长公司工作时间
- 最短公司工作时间
- 平均工作时间

行业稳定性

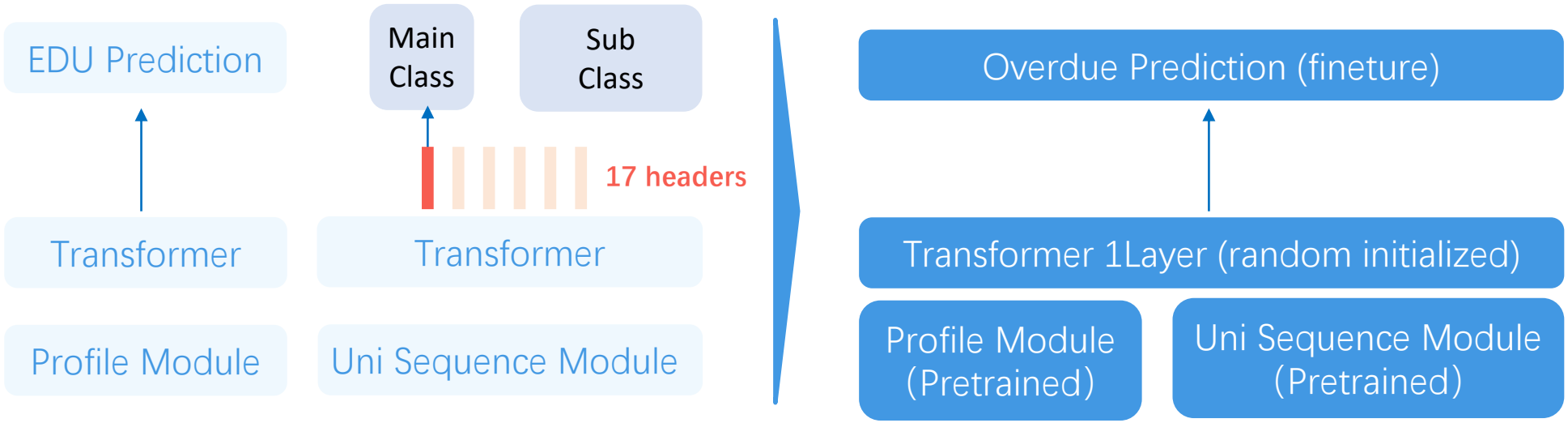
- 不同行业个数
- 最长行业时间
- 最短行业时间
- 平均行业时间

工作地稳定性

- 在不同省份单位数量
- 不同城市单位数量
- 不同区单位数量

自监督学习应用

通过自监督学习内在语义信息，建立容错性高及可迁移学习的基础模型，大幅提升最终模型效果



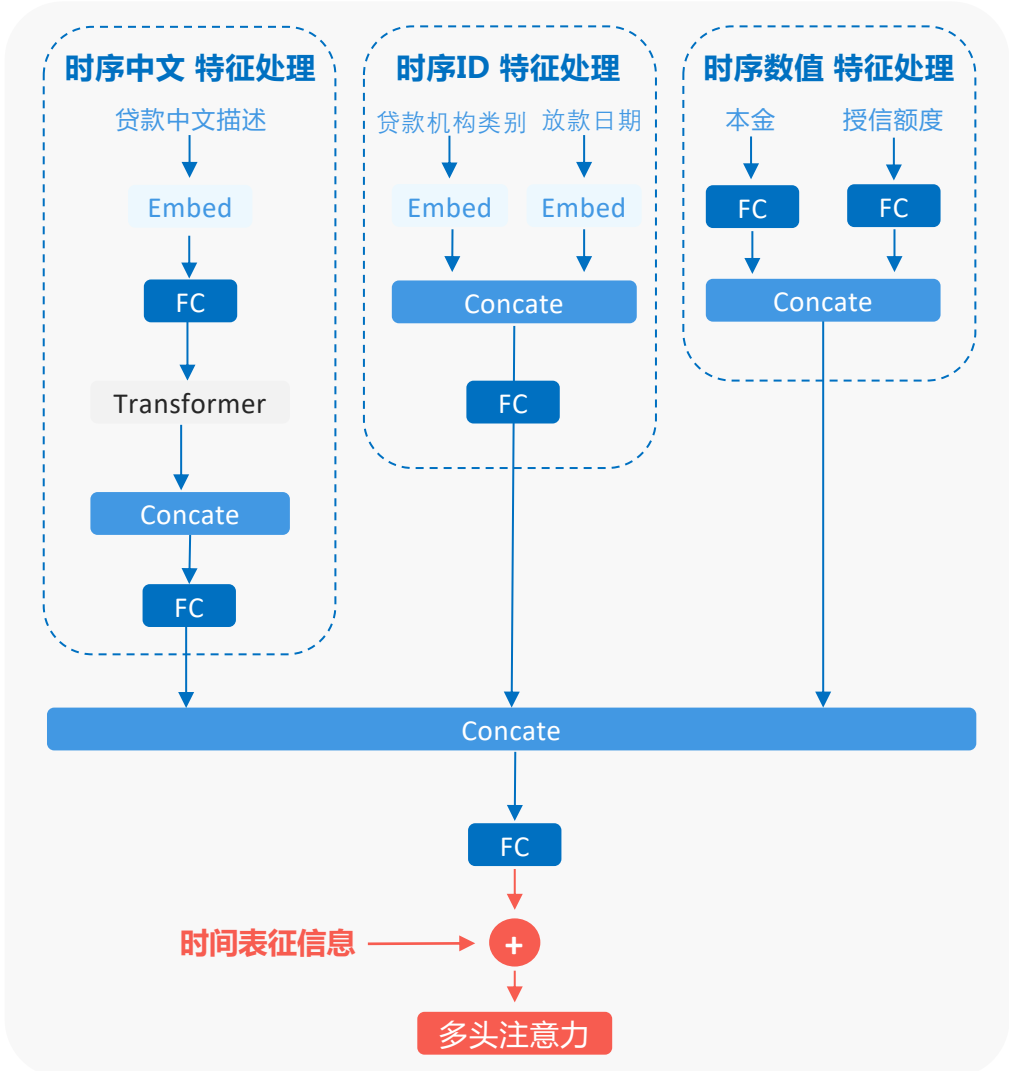
对画像信息进行自监督

- 基于其他属性:公司列表年龄居住信息等
- 预测本科及以上的概率

对信贷序列进行自监督预训练

- 对信贷序列统一编码合并为统一时间序列
- 对信贷的一条记录离散化编码转化为有限状态
- 对状态利用共现进行聚类构造层次预测目标

交叉时序信息的模型优化



A用户 (观察日:2019年5月3日)		
发放日	发生	状态
2007/12/02	发放2000元农户贷款	结清
2008/11/11	发放2000元农户贷款	结清
2017/08/25	发放贷记卡授信额度29000	无逾期记录 当前已用4198
2017/11/10	发放贷记卡授信额度16w	无逾期记录 当前已用14w
2018/11/30	发放贷记卡授信额度8000	无逾期记录 当前已用719
2019/1/10	发放1w个人消费贷	20期按月归还
2019/1/26	发放5000个人消费贷	13期按月归还
Mob12m2+:2020年5月31日真实逾期30天以上		

信息

- 无违约记录
- 农户身份
- 近十年缺失信贷信息
- 授信额度高 & 使用率高
- 贷记卡转消费贷（时序）
- 金额小 & 分期数大

独立特征加工+xgb	
有无违约	农户贷款次数
最高授信额度	最改消费贷金额
平均使用率	…

XGB模型

无特殊风险

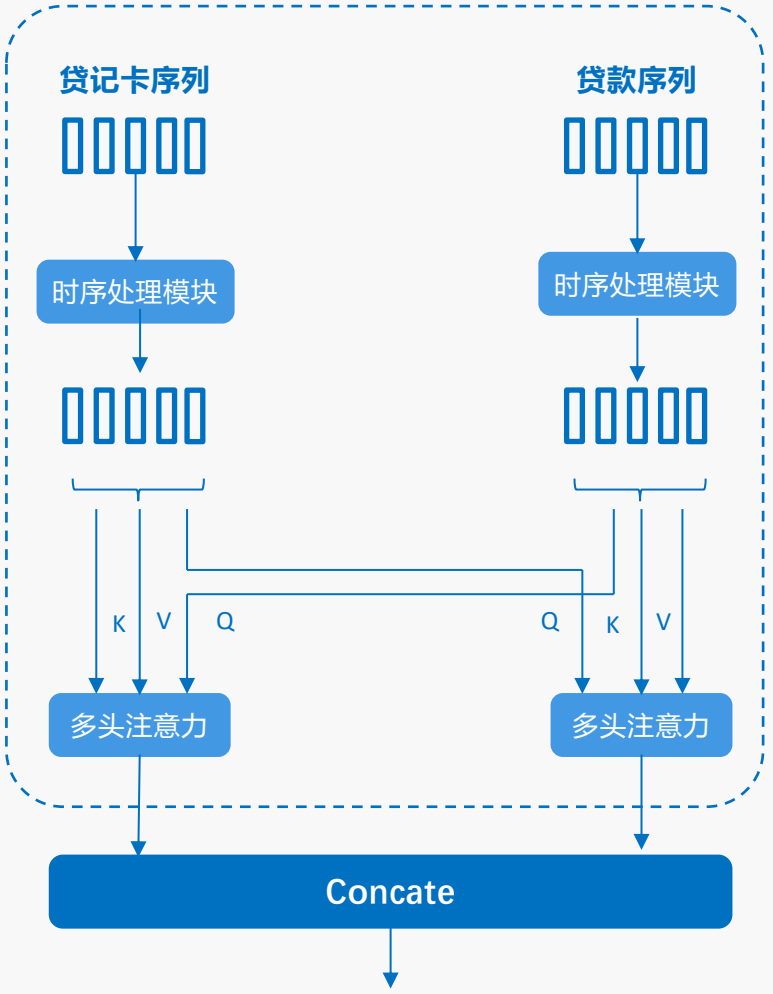
时间序列模型

- 记忆历史信息细节，捕捉全生命周期信息变化趋势
- 自动挖掘识别类别及数值隐藏信息
- 挖掘细粒度趋势信息

时间序列模型

整体中和负债失控概率较大，风险较高

交叉时序信息的模型优化



B用户以贷养贷挖掘推演

贷记卡

开卡日期	授信额度	已用额度	本月应还	账单日
2012/9/9	50,000	41476	4437	2019/10/8
2018/2/11	51,000	47790	11000	2019/9/10
2019/2/25	93,125	88076	5300	2019/9/4

时序模型
坏人

贷款

放款日期	本金	机构类别
2019/2/25	4000	小贷
2019/3/9	9350	商业银行
2019/3/24	1900	商业银行
2019/5/2	2500	商业银行
2019/9/9	11000	小贷
2019/9/22	8000	消金

XGB模型
好人

时间序列交叉网络对比XGB模型

- 交叉网络不仅能捕捉单个模块的时序信息,还能捕捉到多个模块交叉信息
- 针对一条记录关联到其他模块中的相关记录(时间、金额、风险表现)

图模型发展及征信中应用介绍

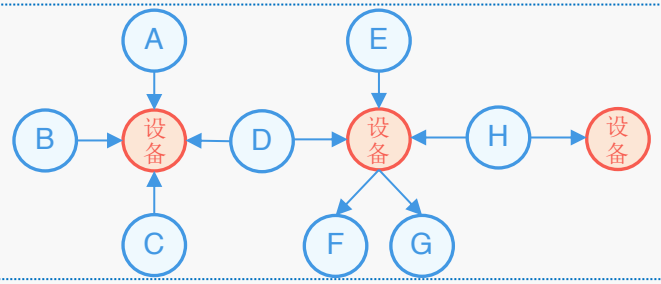
DU XIAOMAN FINANCIAL

关联网络2.0riskSage算法优势

- 适用异构网络，线上线下信息整合。
- 基于关系和属性的端到端模型，根据历史样本自动学习特征高阶组合和参数，而非有限专家经验挖掘的特征模式。

图关联网络在征信中应用实例

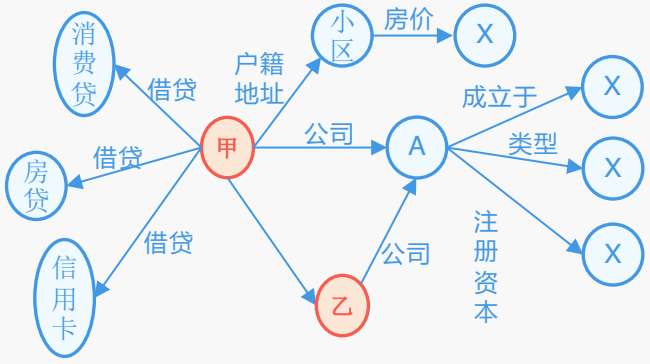
关联网络



通过关联分析和聚类等方法

发现团伙欺诈行为

图模型

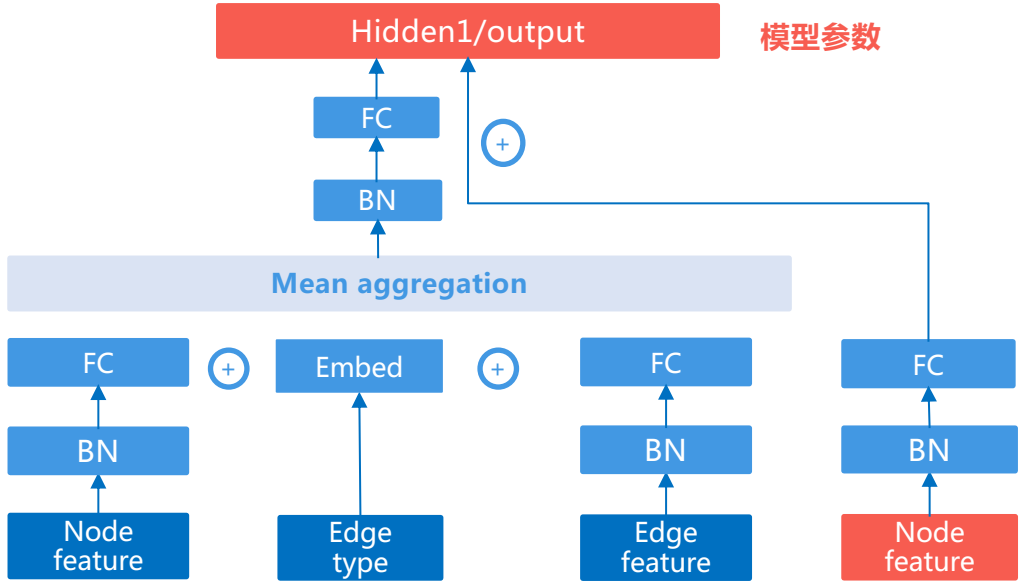


基于异构图和底层特征

建模挖掘并识别风险特征

不局限于欺诈信息

类目	属性/类别	规模
用户节点	性别、年龄、学历、query风险分、app风险分、征信风险分等	10亿+
公司节点	注册资本金、公司类型(个体、有限公司等)、行业、成立时间、注册地、企业状态、处罚信息	2亿+
位置节点	建筑年代、建筑类型、物业费用、开发商、楼栋数、户数、均价	1亿+
边	公积金缴纳关系、所在单位、居住地关系、工作地关系、申请地关系、公司法人、股东、管理员关系	10亿+

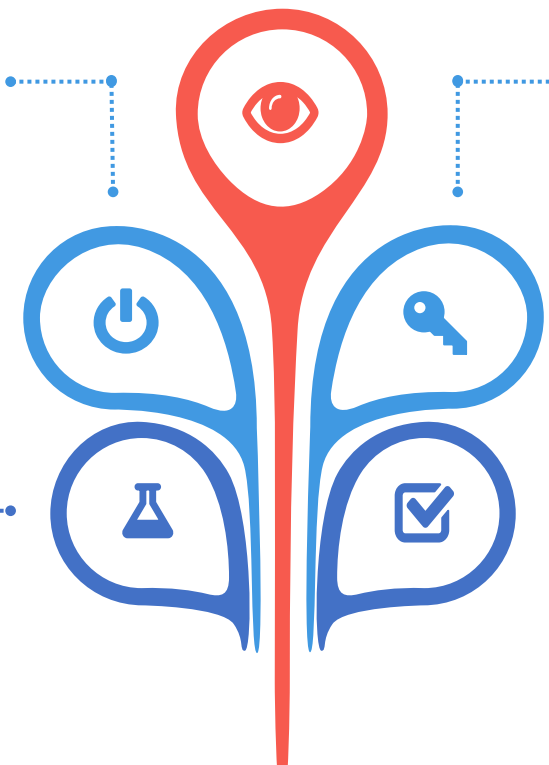


征信挖掘未来探索方向

异构图挖掘
优化基于图表达的模型结构
充分挖掘内部及内外部结合
的关系信息

排序模型
将逾期的先后顺序
纳入模型

多份征信报告挖掘
针对多次查得的征信
报告学习差异信息



预训练
通过大类无标签数
据的自学习来提升
模型的信息抽取能力

模型可解释性
优化对复杂模型的归因方法
提升模型适用性

Thanks

度小满金融
Du Xiaoman Financial

2021 InfoQ 技术大会近期会议推荐

—— 盘点一线大厂创新技术实践

北京站

GMTC 全球大前端技术大会

时间：2021年07月04-05日

地点：北京 · 国际会议中心

扫码查看完整日程

深圳站

ArchSummit 全球架构师峰会 th

时间：2021年07月23-24日

地点：深圳 · 大中华喜来登酒店

扫码查看大会专题