

平安银行算法实践

潘鹏举 (ppj) - 算法团队负责人

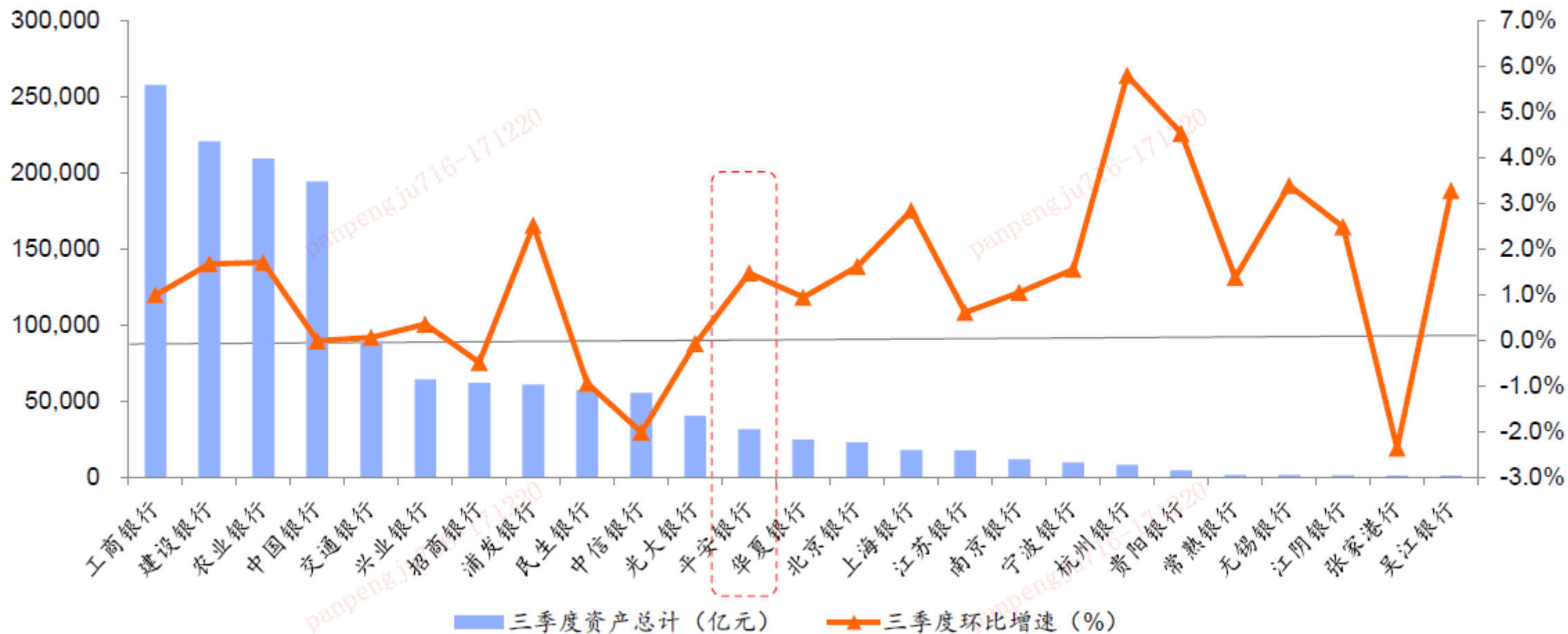
自我介绍-PPJ

- 平安银行大数据AI算法团队Leader
- 携程酒店技术算法组Leader
- 全栈数据
- 码农+管理

业务背景

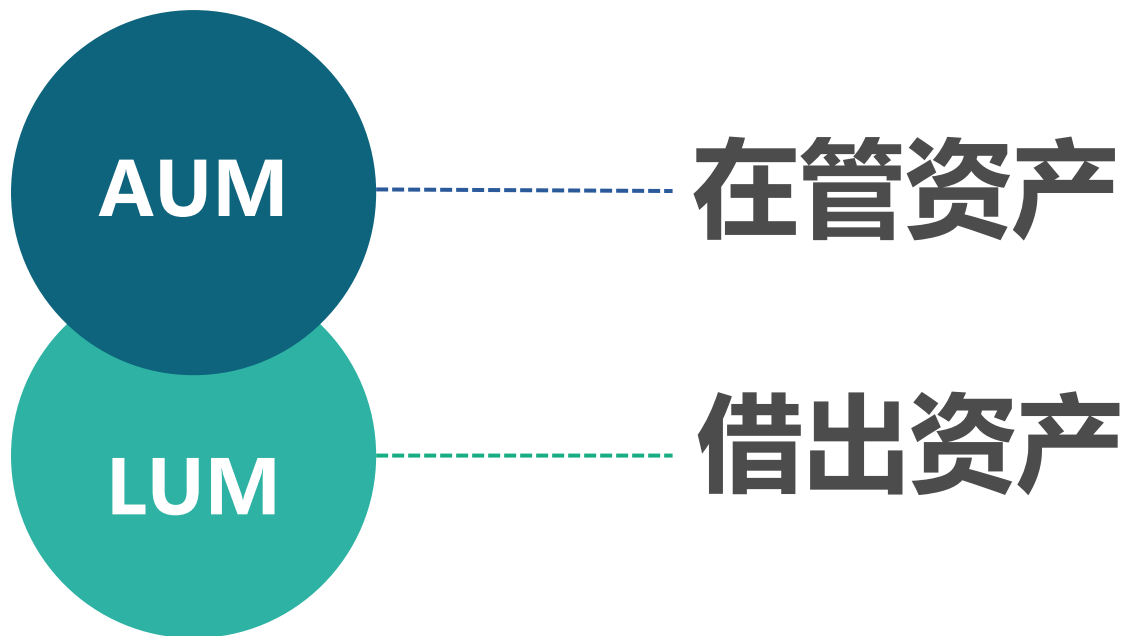


平安银行资产排名





银行核心KPI



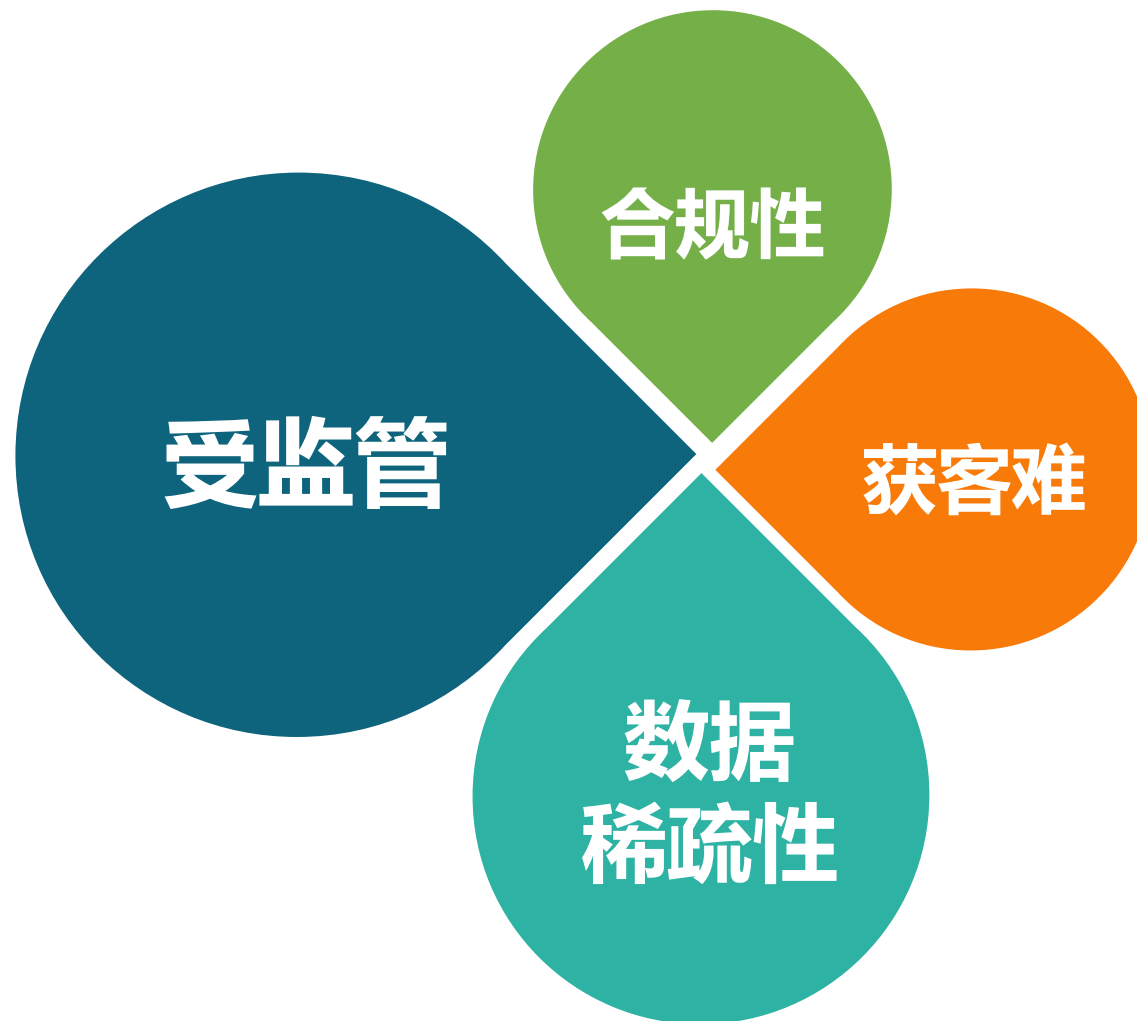
AUM: Asset Under Management

LUM: Loan Under Management





一些挑战



算法实践



总体架构图



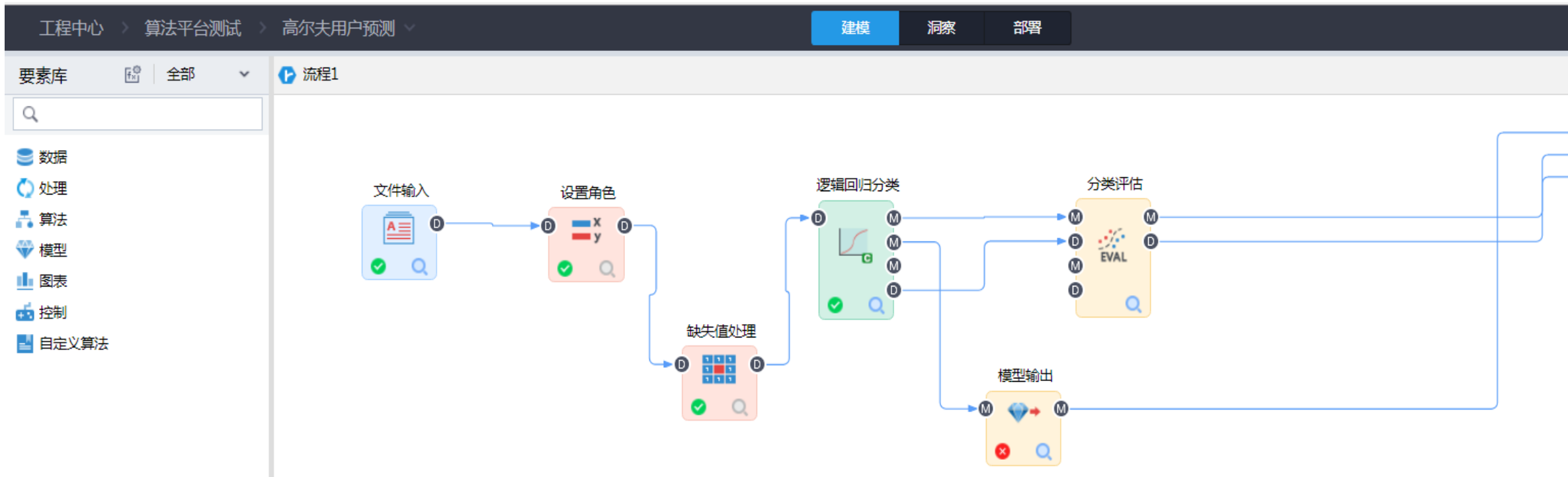


应用概览





算法平台



新的框架？



算法组件

• 数据源->处理->算法->评估->部署

数据

▼ 输入

○ 文件输入

○ 关系数据库输入

○ HIVE输入

○ HBase输入

○ HDFS输入

○ 同步输入

○ 样例数据

▼ 输出

○ 关系数据库输出

○ HIVE输出

○ HBase输出

○ HDFS输出

处理

算法

模型

处理

▼ 行

○ 数据过滤

○ 排序

○ 随机抽样

○ 数据平衡

▼ 列

○ 设置角色

○ 重命名

○ 属性过滤

○ 属性生成

○ 随机数/ID生成

○ 缺失值处理

○ 数值型属性变换

○ 字符型属性变换

○ 日期型属性变换

▼ 表

○ 表转置

○ 数据连接

○ 数据追加

○ 数据拆分

○ 数据分解

○ 分类汇总

▼ 高级

○ 数据标准化

算法

▶ 回归

▼ 分类

○ 逻辑回归分类

○ 朴素贝叶斯

○ 贝叶斯网络分类

○ 神经网络分类

○ 随机森林分类

○ 支持向量机分类

○ CART

○ ID3分类

○ C45+决策树分类

○ 梯度提升决策树分类

○ L1/2稀疏迭代分类

○ RBF神经网络分类

○ KNN

○ 线性判别分类

○ Adaboost分类

○ bagging分类算法

数据

处理

算法

模型

图表

▼ 基本

○ 条线图

○ 圆饼图

○ 散点图

▼ 分组

○ 分组散点图

○ 分组折线图

▼ 统计

○ 直方图

○ 方差分析

○ 相关系数

○ 典型相关分析

○ 偏相关分析

○ 变量选择

○ 相似度

○ 描述数据特征

○ 概率单位回归

○ P-P图

○ Q-Q图

控制

自定义算法

数据

处理

算法

模型

▼ 评估

○ 回归评估

○ 分类评估

○ 聚类评估

○ 时间序列评估

○ 自动择参

▼ 其他

○ 模型输出

○ 模型读取

○ 模型利用

图表

控制

自定义算法



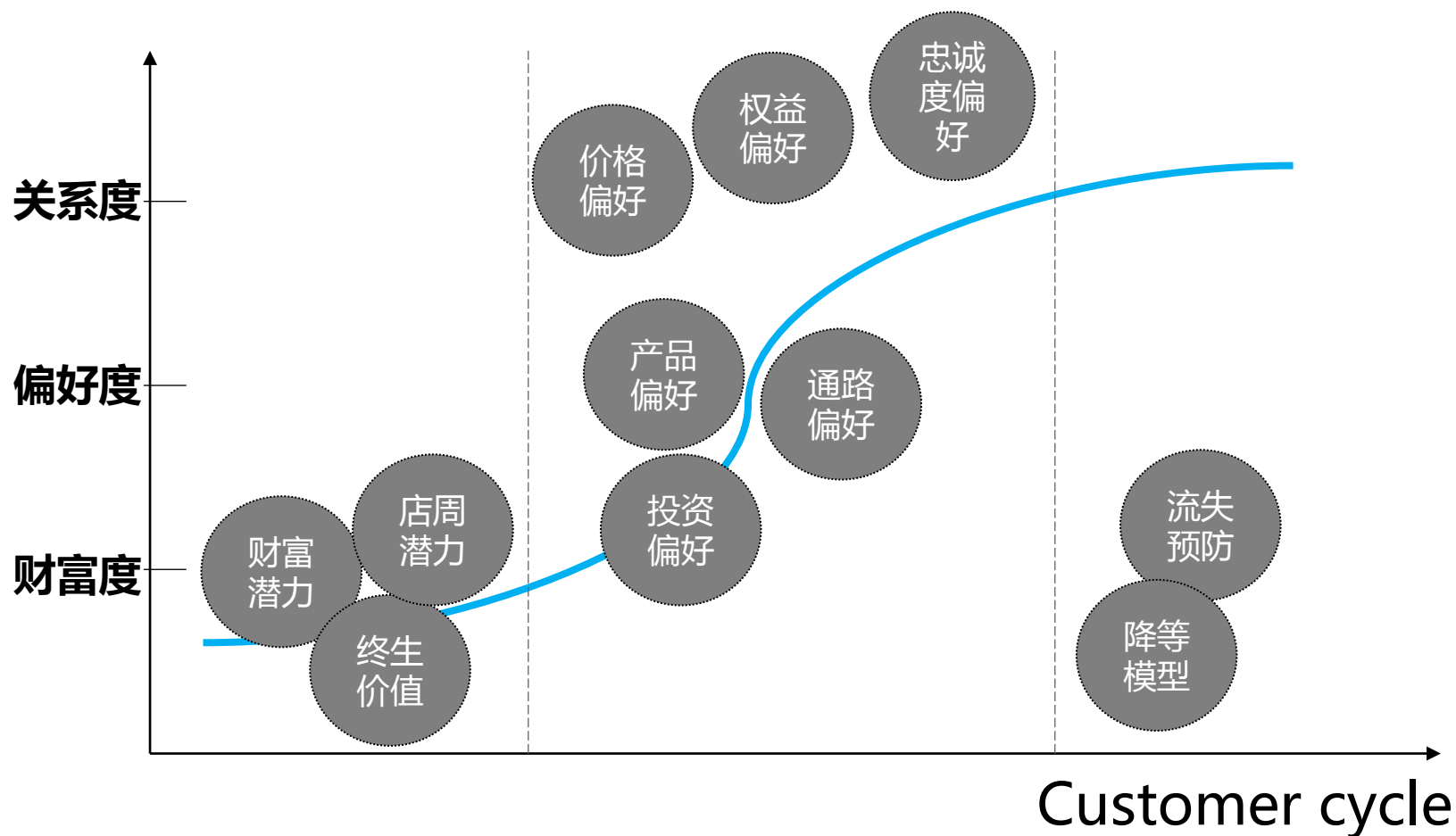
Case:客群管理

- 基于生命周期进行管理

获客模型

迁移模型

流失模型



模型	运用方式
财富潜力	预测客户现行财富
终生价值	预测客户未来财富
店周潜力	寻找分行周边的潜在客户
产品偏好	预测客户的购买特定产品的机率
流失预防	预测可能流失的客户



Case:客户智能分析

中国平安

保险·银行·投资

平安银行

PINGAN BANK

AI智能分析 反馈评价 评价

AI智能分析试用版运行啦，需要您的评价才能支持更好哦。

Hi, 我是AI【深潜】
以下是我的客户智能分析报告

女 张然然 18周岁

对接大数据标签库，集成客户产寿信用卡信息

集团	客层：潜力私行
	资产层级：M1
寿险	客层：A类
	持卡等级：钻石
产险	客层：A类
	客户车辆价值：20-50万
信用卡	负债累积月日均：¥8,748.00

投资偏好

线上风测 稳健型

临柜风测等级 中低风险

和盈倾向度 高
预测客户未来一个月将购买和盈的概率

SOW等级 [0,10K]
客户可支配资金金额分层

基于模型计算客户和盈购买倾向和可支配资金区间

推荐产品

每月基于推荐算法推荐可能性最高的产品

金领通 推荐1
集团非银无抵押贷款客户

定期存款-3年 推荐2
集团非银无抵押贷款客户

和盈资产管理类-187天以上 推荐3
集团非银无抵押贷款客户

对接推荐产品，支持具体推荐规则展示，直接跳转详情页，产品购买闭环。

流失预测

流失概率走势图
每周一更新该客户（当前财富等级）两个月后降级的概率

2017-9-22 流失概率:76.3%

7/01 7/08 7/10 7/13 7/15 7/22 8/01 8/08 8/15 9/22

以下是造成流失的关键异常因子

面访次数小于3/每季度 高

通话时长小于60s/次 偏高

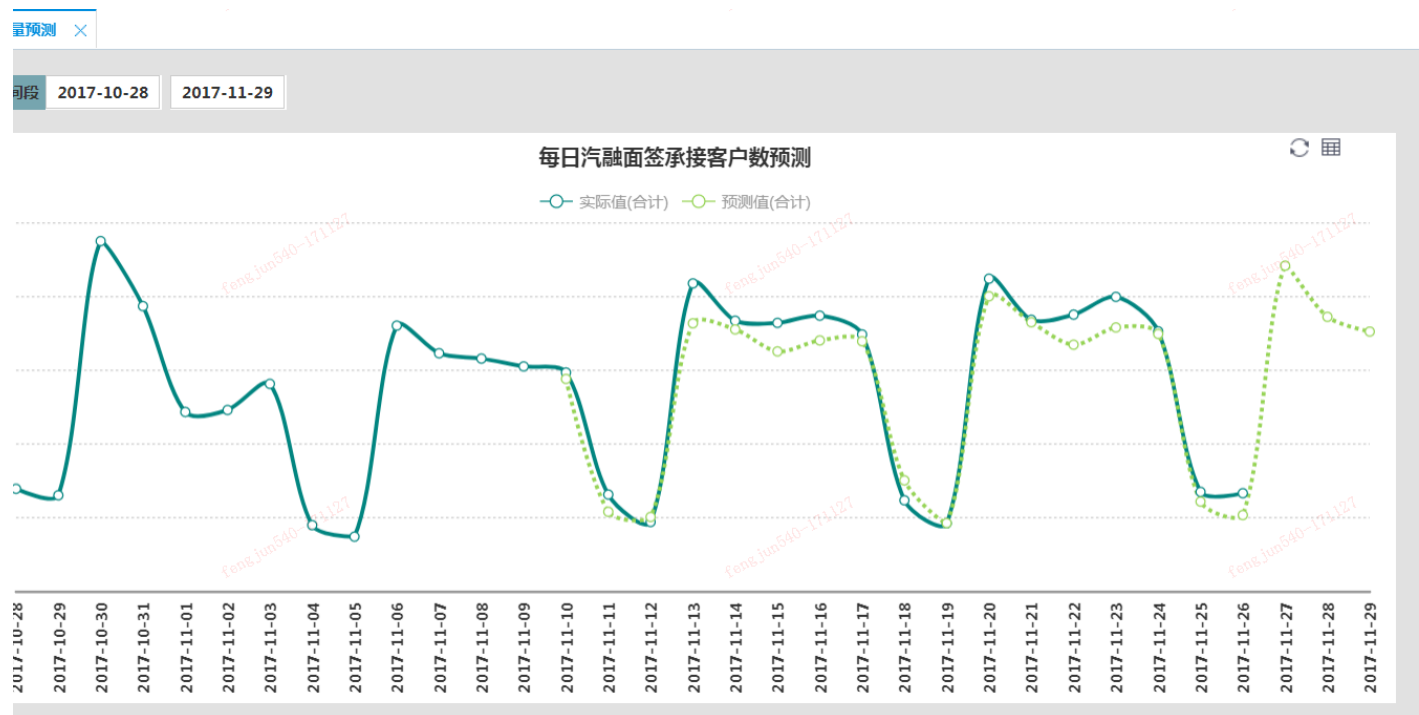
计算客户未来两个月流失概率，挖掘异常流失因子，指引接触动作。

反馈评价



Case:业务预测

某业务进件预测，提前预测D+1~D+3，绝对误差<9%



训练数据：不到半年

预测方法：ensemble

- Arima
- Tar
- 指数平滑
- 历史同天平均值
- 上月同一天

回归

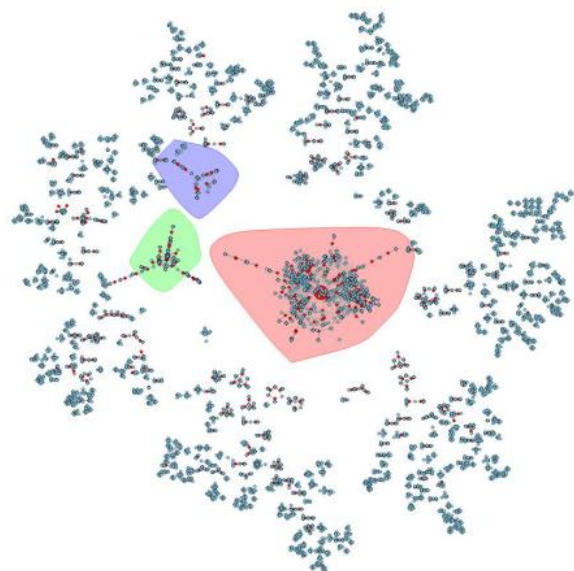
规则

最终结果

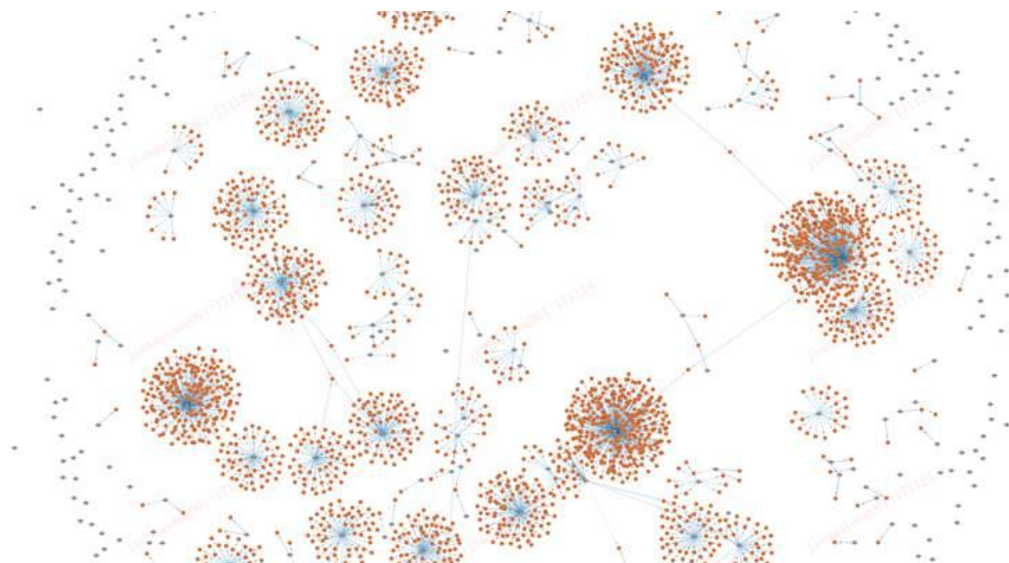


知识图谱在复杂风控中的应用：风险聚类，案件探索，复杂规则挖掘。

疑似羊毛党



账户、卡号、电话、进件的关系

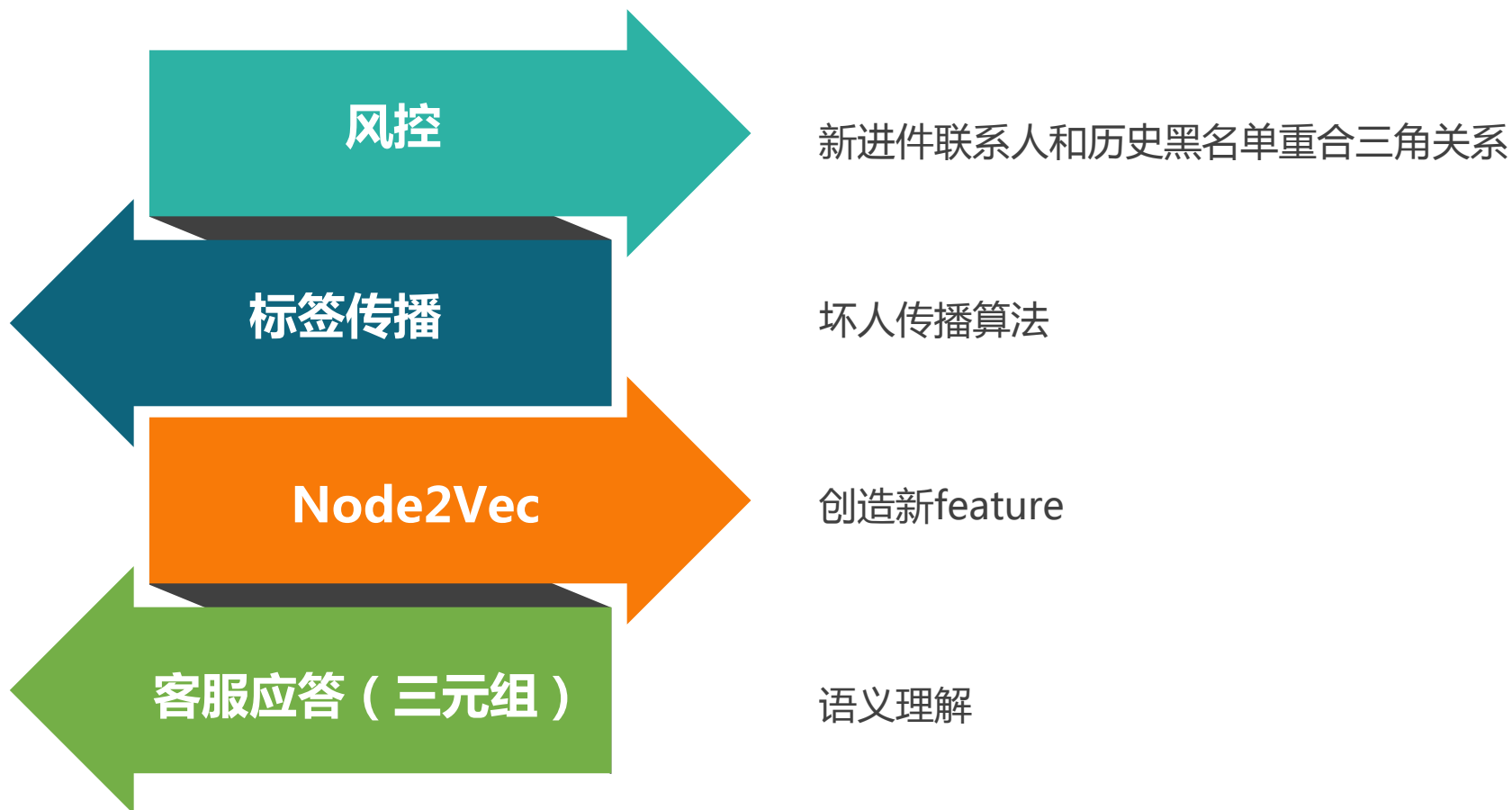


上亿的节点、10+种的实体关系





图谱应用



Node2vec: <https://github.com/aditya-grover/node2vec>





图数据库

OrientDB

企业级应用，可定制开发

Neo4J

扩展性低，独立索引体系

JanusGraph

存储：HBase，索引：ES

GraphX的定位？





地址标准化？

ADDR: 上海浦东福山路455号全华信息大厦9楼

问题	例子
全角地址	重庆市九龙坡区红狮大道1 0号3幢2 6 - 6
重复地址	深圳市宝安区观澜镇大水坑村鸿观科技园深圳市宝安区观澜镇大水坑村鸿观科技园
近义词	重庆九龙坡红狮大道10号3栋26-6 重庆市九龙坡区红狮大道10号3幢26-6
填写的区域不是官方的行政区域	['成都市', '高新', '西区', '合作', '路', '888', '号'] (高新西区不是行政区)
道路的识别	['沈阳市', '苏家屯区', '金桔', '二路', '8', '-', '1', '号', '3', '2', '室'] ['湖南', '长沙', '雨花', '桂花', '路', '187', '号']
地址填写顺序不一致，缺少部分行政层级	['', '吕', '城镇', '圣旨', '西路', '41', '号'] (缺少省和市) ['沈阳市', '苏家屯区', '金桔', '二路', '8', '-', '1', '号', '3', '2', '室'] (缺少省)
直辖市有多一级或少一级行政区域的情况	['广东', '深圳', 'null', '罗湖区', '华裕', '花园', 'A19B'] ['广东', '深圳', '罗湖区', '华裕', '花园', 'A19B']
同一地址多种表述	['广东', '深圳', 'null', '南山区', '西丽镇', '腾飞', '苑', 'B座', '303'] ['广东', '深圳', '南', '山西', '丽新围', '村', 'B', '栋', '303']
缺少省、市	东直门外大街天恒大厦8层
地址纠错	





地址标准化！

Restful API:

http://10.14.221.88:8889/api/v1000/single_address?address=上海浦东福山路455号全华信息大厦9楼

Address NER

Result:

原始地址	上海浦东福山路455号全华信息大厦9楼
返回地址	上海市浦东新区福山路455号全华信息大厦9楼
国家	中国大陆
省	上海市
市	上海市
区/地级市	浦东新区
县/区县/镇	
村	
路	福山路
门牌号	455号
小区	全华信息大厦
楼号	9楼

Enter your address:

算法

规则



CRF

Bi-LSTM+CRF?



地址相似度



Restful API:

`http://10.14.221.88:9090/rest/nlp/v1/address_similarity?addr1='上海福泉路458号'&addr2='上海福山路456号'`



算法

规则



图谱



LR

AI产品经理

欢迎加入 平安银行

NLP算法工程师
图像算法工程师
语音算法工程师
机器学习工程师
AI产品经理
数据产品经理
大数据架构师
各种数据人才~~~



潘鹏举

immortalness 



Scan the QR Code to add me on WeChat



– THANKS –