第四部分　采购需求说明书

第一章 背景及目标

1.1 需求背景

当前商业银行业务风险来源和影响日趋复杂，客户风险不仅源于自身，亦可通过产业链、供应链、资金链、股权链等关联方传导而来。对单一客户的风险监测已不足以满足银行风险管理的需要。通过识别企业间各类关联关系，研究各类风险的传导逻辑和路径，准确发现和预警传染风险，实现由单一风险到组合层面影响的风险预警，有助于提高我行的风险管理能力。

本项目任务聚焦于研发一套基于关联关系的由单一客户风险到组合层面的风险传染、预警模型，并实现相关系统应用建设。本项目可通过客户股权、担保、交易流水、产业链、供应链等关联关系数据，利用知识图谱、图计算、机器学习或深度学习等先进技术搭建关系识别、风险传染、风险估算模型，并结合业务逻辑准确刻画、衡量我行客户间的风险传染关系、路径及影响，解决我行对风险传染、交叉风险存在的判断难、识别难、估算难等切实业务难点，完善我行风控体系及功能，支持总分行客户风险研判、预警、监控、排查等具体风险管理工作。

本项目的具体任务为：基于客户间关联关系的风险传染模型。该子赛道侧重于通过搭建全类别关联关系图谱，全面刻画客户之间关联关系，搭建模型对客户间风险传染概率进行量化测算，圈定客户级别风险传染范围，识别风险传染路径，估算风险大小，形成精准的企业风险传染图谱，为风控预警监控、业务拓客等具体环节减负增效。

1.2 需求简述及目标

1.2.1 需求简述

本项目的任务提出者：广东省分行

服务对象：

■对公客户　■对私客户　□内部柜员　□内部管理人员　□其他_____　（可多选，其他请说明）

随着企业多元化发展、业务范围跨行业、跨区域的客户越来越多，企业间的关联关系日趋错综复杂，信用状况参差不齐，因关联企业识别不充分所造成的各种风险隐患或营销策略错误屡见不鲜。处于某链条或者是闭环上的企业节点，其自身的风险不仅受到直接相邻业务节点的影响，同时也受到外界影响通过相邻节点而传递过来的影响。有效识别企业关联关系并通过这些关系路径对风险进行量化的传导分析，对我行整个风控体系来说是非常重要的一环。

本项目基于行内目前对企业关系链的掌握，结合现有的对公授信客户综合评价结果得到的违约概率进行优化，拟构建风险传导模型，对企业风险判断不仅只关注企业本身，需要把企业关联方的风险也纳入评估并且形成量化的分析结果。

1.2.2 实施目标及计划

1.实施目标

预计实现两个功能。

第一是关联传染分图谱，就是会生成一张以风险源为核心的关联传染得分图谱，展示各种关联关系中企业、个人的关系分布图以及各主体在我行的授信余额、盘存类别及分类等情况，每个关联线路都展示出风险源与被传染对象之间最佳传导路径的传染分（由起点到终点的最终传染概率折算而来）。

第二就是依托对公授信客户综合评价结果风险源的违约概率，乘以最佳传导路径的传染概率，算出被传导对象违约概率的增加值，最终建立起综合基于客户自身风险和受传染风险的全息风险综合评价体系。

2. 实施计划

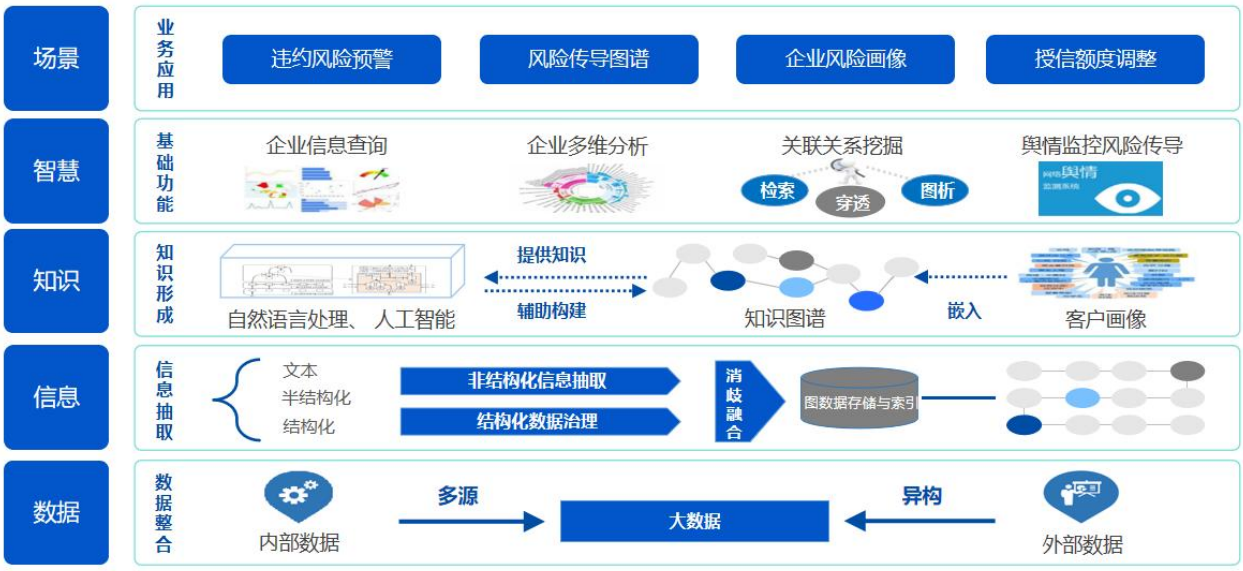
工作阶段	计划完成时间
项目立项	2024 年 6 月 11 日
提交采购、签署合同	2024 年 7 月初
需求开发、系统测试	2024 年 7 月底
系统投产	2024 年 8 月上旬

1.2.3 可行性简述

项目可行性分析简述，包括但不限于：市场状况与发展趋势、同业同类产品实现情况、国家政策影响、业务现状/问题/改进建议及业务价值、用户使用情况、风险情况、战略符合度、市场需求趋势、客户体验提升情况、监管要求等。

1.3 方案架构说明

本项目拟包含以下工作项：



1. 行内外数据融合

基于行内已有的客户基础信息、资金流水、担保缓释、财务信息等行内数据，引入总行通过数创平台下传的工商、司法、舆情、税务、海关、环保等外部数据，通过 ETL 把内外部数据通过客户号、外部商事主体唯一码、统一社会信用代码或企业名称关联存表，生成风险标签数据、供应链上下游数据、

产业链数据、企业关系数据等信息。

2.构建全面的企业动态风险事件监测机制

通过引入外部风险大数据，围绕企业本身及对企业的经营造成重大影响的人、事、物，对其发生的风险事件进行智能化的捕捉、加工，构建风险事件监测模型，将负面事件归集为信用风险、内部治理风险、应收账款风险、合规风险、经营风险、对外投资风险、行业风险等不同维度，形成外部数据风险事件库。

3.构建客户风险传染关系网络模型

依据客户基本信息数据、关联关系数据和资金流水数据，完善企业客户间的关系类型，构建企业知识图谱，包括股权投资关系、人员任职关系、集团关系、担保关系、资金交易关系、上下游关系、隐蔽性关系等多种关系类型，借助图数据计算引擎及时识别相关企业和关联实体的风险等级及传导路径，输出受影响的客户名单，辅助客户风险排查。

4.构建基于企业关系网络的风险传染模型

基于知识图谱技术的企业风险监测及传导分析方法，构建基于数据驱动的风险传导模型，设计 N 个特征变量作为模型的输入，同时，模型需要针对各种风险类型分别建模计算。在模型训练方面，拟采用 LightGBM 算法，它是一种具备训练效率高、内存使用低、支持直接使用类别特征的机器学习算法。同时，专家经验方法和机器学习算法两者结合，可以避免人工规则及方法的片面性与主观性，也可以弥补深度学习算法在解释性方面的不足，更方便于模型的更新迭代。

5.构建风险可视化展示平台

对于企业和行业的各类风险事件，结合图谱关系传导计算，实现 7*24 小时企业风险监测和传导预警，支持关注标的风险信息从授信业务部门、风险管理部门到行内管理层的多级推送。构建风险可视化平台，相关业务人员可以对风险预警信号进行风险源追溯，溯源具体的风险事件及传导链路。

第二章 需求概述

2.1 对公授信客户综合评价功能

2.1.1 行内数据处理

从“客户状况分析”（流动性、效益性、安全性）、“客户性质分析”（所有制分类、行业分类）、“授后管理难易分析”（同业结构、授信规模+签约品种、销售归行率）三大板块八大维度量化分析，实现对客户风险状况的综合评价。

客户状况分析的流动性、效益性两个维度通过财务信息对客户经营的成果实施量化评价，用 10 个财务指标反映客户的基本面，以国务院国资委每年发布的企业绩效评价标准值作为衡量标准；安全性维度通过缓释条件信息对客户在我行的担保品实施量化评价，展示我行授信的安全程度。具体如下表：

得分	1-1 客户状况分析		
	流动性	效益性	安全性
1	严重超负荷	严重亏损	信用或押品覆盖低于 10%

得分	1-1 客户状况分析		
	流动性	效益性	安全性
2	超负荷	亏损	押品覆盖不足 50%或有一定担保
3	较高负荷	微利	押品覆盖不足但超 50%
4	偏高负荷	低于行业平均	押品覆盖不足但超 70%
5	合理负荷	接近行业平均	押品基本覆盖或有较高实力担保人
6	偏低负荷	超过行业平均	有实力的担保人
7	较低负荷	行业中上水平	有效押品足值覆盖
8	低负荷	行业先进	有效押品足值覆盖且变现能力较好
9	超低负荷	行业盈利标杆	有效押品足值覆盖且变现能力极强

客户性质分析通过对所有制分类、行业归属两个维度量化对股东背景和客户所处行业的周期稳定性实施客观评价。具体如下表：

得分	1-2 客户性质分析	
	所有制分类	行业分类
1	不良企业	过剩行业
2	评级较低企业	房地产
3	评级中等企业	政府融资平台
4	一般上市公司	批发零售
5	评级较高企业	制造业
6	优质上市企业	其他行业
7	金融机构	公共事业
8	世界 500 强及其控股企业	交通行业
9	央企及其控股企业	电力行业

授后管理难易分析通过同业结构、授信规模+签约品种、销售归行率三个维度，反映客户的多头授信情况、我行的管理资源投入、客户的信息透明度等情况。具体如下表：

得分	1-3 授后管理难易分析		
	同业结构	授信规模 +签约品种	销售归行率
1	9 家及以上	（1）授信规模： （3 亿，+∞]：1 分 （2 亿，3 亿]：2 分 （1 亿，2 亿]：3 分	（-∞,0%]
2	8 家		(0%,25%)
3	7 家（以及为空）		[25%,50%)
4	6 家		[50%,75%)

得分	1-3 授后管理难易分析		
	同业结构	授信规模 +签约品种	销售归行率
5	5 家	（0.5 亿，1 亿]：4 分 （0，0.5 亿]：5 分 （2）签约品种： 品种为 1：5 分 品种为 2：4 分 品种为 3：3 分 品种为 4；2 分 品种为 5 以上：1 分 （3）组合折成 9 分。	[75%,100%)
6	4 家		[100%,150%)
7	2 家		[150%,200%)
8	1 家		[200%,500%)
9	3 家		[500%,+∞)

2.1.2 行内数据纠偏

根据企业在我行的财务数据、履约合作记录，利用 Z 值、Fscore、Mscore 等财务模型和调整还款计划、展期、冻结、还款延迟、涉及网贷、资金流向异常等履约合作记录共 19 个指标，对行内数据分析初步结果进行纠偏，更加客观反映企业的风险程度。具体如下表：

序号	主题	预警指标名称	风险等级
1	财务类	过度融资	高
2	财务类	流动性一票否决	高
3	财务类	效益性一票否决	高
4	财务类	资不抵债	高
5	内部履约合作	调整还款计划	高
6	内部履约合作	展期	高
7	内部履约合作	企业在我行账户或资金被冻结	高
8	内部履约合作	还款延迟	高
9	内部履约合作	涉及网贷	高
10	财务类	Z 值预警（小于阈值 1）	高
11	财务类	Z 值预警（小于阈值 2）	中
12	财务类	财务粉饰嫌疑——Fscore 模型	中
13	财务类	财务粉饰嫌疑——Mscore 模型	中
14	财务类	财务粉饰嫌疑——收入和费用不匹配	中
15	财务类	财务粉饰嫌疑——现金流与收支不匹配	中
16	财务类	短贷长用	中

序号	主题	预警指标名称	风险等级
17	财务类	高额现金异常	中
18	内部履约合作	代发薪预警	中
19	内部履约合作	资金流向异常	中

2.1.3 外部数据处理

根据总行提供的企业外部数据，从工商、司法、舆情、税务、海关、环保等 6 大主题 20 个外部数据指标进行预警分析。每个对公客户每触发一条外部信息，则相应增加一条预警记录，触发的外部信息预警越多，代表该客户风险越大。具体指标如下表：

序号	主题	预警指标名称	风险等级
1	工商类	企业清算	高
2	工商类	企业简易注销	高
3	工商类	企业经营异常	高
4	海关类	海关评级为失信企业	高
5	税务类	企业税务非正常户	高
6	税务类	税务评级过低	高
7	司法类	失信被执行人	高
8	司法类	企业严重违法	高
9	司法类	涉案标的金额超过净资产的 30%	高
10	司法类	被金融机构起诉	高
11	工商类	企业股权冻结	中
12	环保类	企业环保处罚红牌	中
13	工商类	企业股权冻结	中
14	工商类	股权质押比例过高	中
15	工商类	企业抽查检查异常	中
16	海关类	海关行政处罚	中
17	环保类	企业环保处罚黄牌	中
18	环保类	企业环保行政处罚	中
19	税务类	税务评级过低	中
20	舆情类	负面新闻舆情	中

2.1.4 利用机器学习得出违约概率

对行内对公授信客户的各维度特征进行数据清洗，利用机器学习技术，通过 LightGBM 算法综合预测风险（包括发生逾期、纳入盘存、分类下迁 3 大类别）发生的概率，即最终输出每个客户的违约概率。通过夏普里值解释器对不同客户的预测结果进行可视化分析，输出客户风险特征在预测结果的权重情

况，当特征处于一个正权重时，表示该特征倾向于表达客户处于高风险状态。反之亦然。方便业务人员精准识别客户风险，针对性地采取相关措施，保障我行授信资产的安全。

2.2 功能说明

在信用风险管理活动中，风险监控和预警是重要一环，在上述对单一授信主体风险识别的基础上，还应通过构建内外部数据融合的企业知识图谱，综合考虑企业自身属性和企业之间关联关系，有助于快速定位风险传导强关联的企业群体，从而在标的公司发生风险事件时，及时识别相关企业和关联实体的风险等级及传导路径，并通过这些关系路径对风险进行量化的传导分析，实现提前介入和有效防范。

2.2.1 构建企业风险传导路径

在企业的关系网络中，企业是其中的一个节点，而能够对该节点施加影响的可能是单一节点，也有可能是多个节点共同影响，所以需要集合具体场景，把这个节点放在一个特定的网络中进行整体评估；另外从影响的层级上看，企业面临的可能是最近一层节点的影响和跨层级节点的影响。所以这都要求实现风险传导模型的第一步需要挖掘企业关联关系数据，如控股股东、实际控制人、关联公司、担保对象、交易对象、上下游供应商、客户等和企业均构成关联关系，本方案定义以下风险传导路径种类：

2.2.1.1 股权关系路径

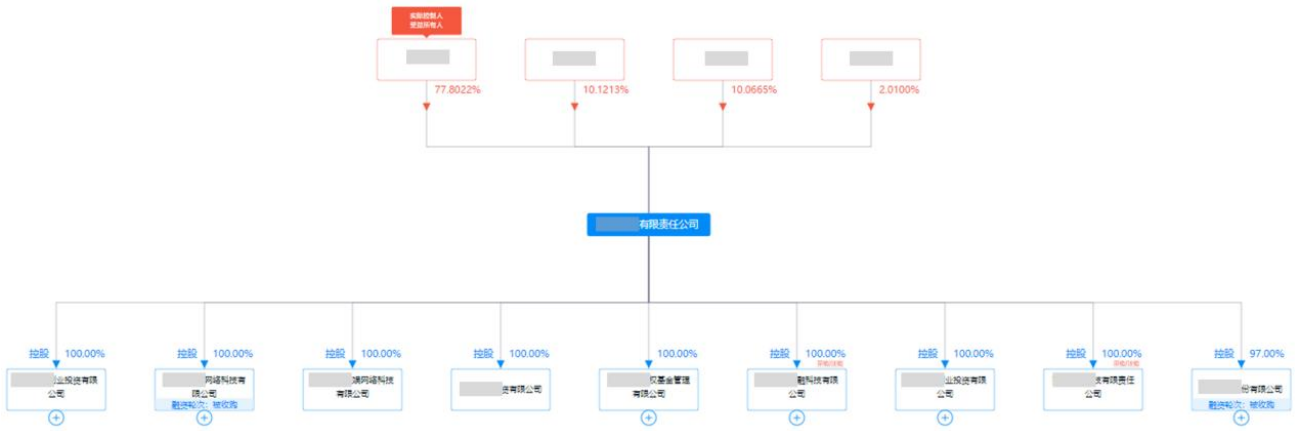
关系定义：股权投资关系，是根据工商数据中的股权投资以及股比数据进行构建，最终构建出全体企业的股权投资关系。

加工规则：股权投资关系根据企业的股东，以及每个股东的股比进行挖掘。使用企业的股权投资信息找到目标企业的股东后，再通过查找目标企业股东的股权投资信息，来构建该企业的股东关系结构树。按照这种方式，不断深挖企业的股权投资信息，获得企业最终的股权投资关系。

数据来源：工商数据，外部数据作为补充。

关系示例：

股东关系模型支持挖掘出目标企业完整的股权投资路径。



2.2.1.2 关键人员关系路径

关键人员定义：企业法定代表人、董监高、财务代表、企业网银经办、企业网银复核等员工。

企业关键人员作为风险传染关系网络中的重要节点，关键人员是否有正在进行诉讼、仲裁或其他法律程序，是否曾有过破产、犯罪、欺诈、不当商业行为等风险，都会对企业造成影响。

基于工商信息计算，构建企业关键人员关系，挖掘自然人担任关键人员以及控股的企业，识别人-企关系链路。

数据来源：以工商登记数据为主，以模型或算法推理出的数据为辅（需要对人员进行唯一性识别）。

关系示例：

挖掘目标企业核心管理人员任职或控制的其他公司

发现与目标企业具有相同法定代表人、实际控制人……的关联企业

2.2.1.3 集团家谱关系路径

以特定客户(企业)出发，以我行现有的集团客户成员为基础，并增加依据以下业务规则识别集团关系。

关系定义：利用全量工商股权数据，结合《大额风险暴露管理办法》关联客户识别方法进行建模，深入挖掘企业间的股权投资关系，分析企业集团派系。

加工规则：通过股权关联关系分析判断目前企业与其他企业直接控制\被直接控制\间接控制\被间接控制\亲属联合控制等关系，实现基于集团系的集团关系识别。

数据来源：以行内统一授信维护的集团家谱为主，行外数据作为补充。

2.2.1.4 实际控制人关系路径

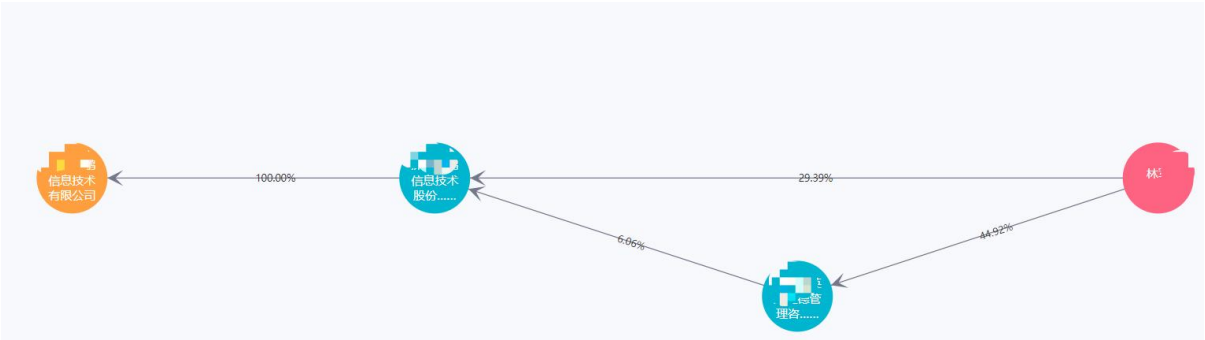
企业法定代表人不等于实际控制人，控股股东也不一定是实际控制人。实际控制人作为网络中的关键节点，拥有公司的控制权，可以对公司的生产经营、投资借款、利润分配等重大事项产生重大影响。识别实际控制人不仅是信用风险监控的基础，也是审查实际控制人是否操纵关联交易的前提条件。

关系定义：通过直接持有公司的股份或者通过投资关系、协议等情况，能够对股东大会的决议产生重大影响或能够实际支配公司决策的人。

加工规则：在控制人关系的构建中，通过无限层级的穿透股东以及其股比信息，计算并汇总企业间深层的持股内容。在企业持股、控股的内容中不断上查找目标企业的控股股东以及控股股东企业的股东，来挖掘目标企业的控制人关系。控制人关系区分直接控制人与间接控制人：直接控制人为对企业持股比例大于 50%股比的股东；间接控制人为针对目标企业大于 30%总持股，且在持股的所有路径中均超过 30%的股东。

数据来源：工商数据。

关系示例：



2.2.1.5 资金交易关系路径

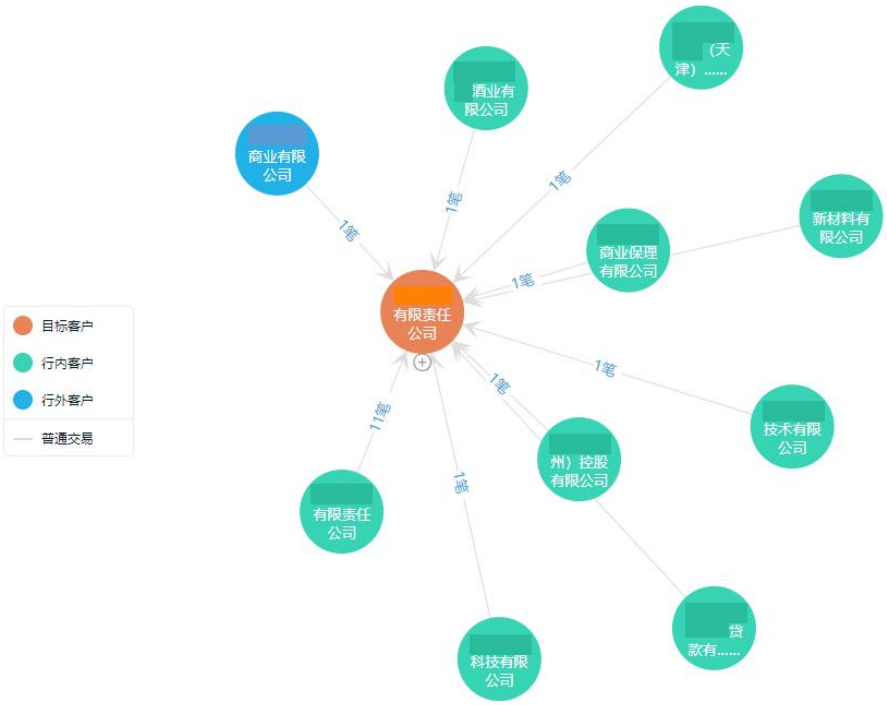
主体之间存在长期稳定的交易关系，在一定程度上说明彼此存在紧密的联系，两者之间可能发生风险传导的概率就越大。风险可能在资金链交易对手之间传导，例如下游企业财务出现困难，无法按时支付款项，如果上游企业资金长期无法回笼，可能面临资金链断裂的风险；上游供应商的财务困境、生产中断、质量问题或交付延迟可能直接影响到下游企业的生产和交付能力。

关系定义：资金交易关系包括交易主体、交易金额、资金转入，资金转出、交易时间、交易账户、交易类型等信息，使用行内的交易数据进行分析，通过挖掘分析行内所有的企业交易数据信息，最终得到行内企业交易关系。目前行内已经构建的交易关系路径如下：

1. 贷款（含贸易融资、贴现等）资金被挪用
2. 借款人参与民间借贷
3. 借款人虚增账户结算量
4. 授信客户交易对手互为上下游
5. 授信客户的交易对手之间存在三方的交易循环
6. 大额资金往来
7. 贷款资金流向集中
8. 涉嫌利用搭桥资金还旧借新
9. 同一客户多次转入资金代供款
10. 贷款供款资金来源集中
11. 贷款资金受托支付后产生回流

数据来源：内部交易流水数据。

关系示例：



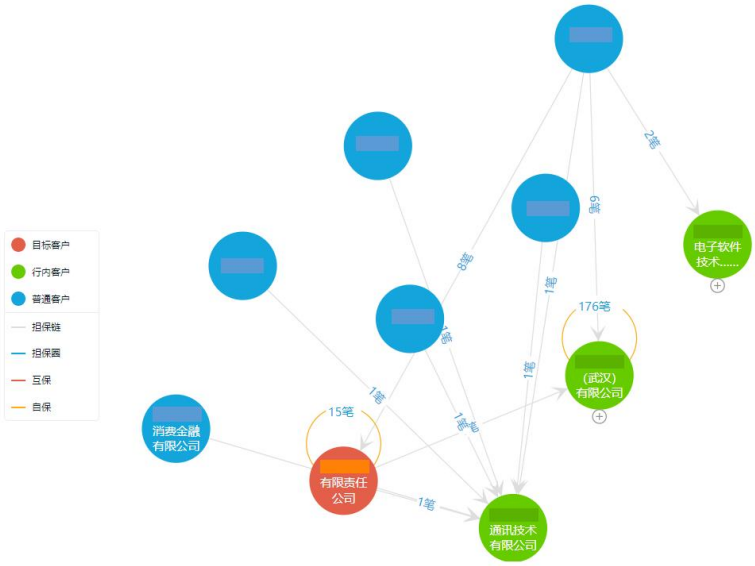
2.2.1.6 担保关系路径

担保链、担保圈而引发的信用风险，在全国频繁发生。互保、循环担保、交叉担保等关系复杂，往往导致担保虚化。信贷风险会通过担保链条在担保圈内循环、传递、放大，一旦某个链条断裂，可能引发多米诺骨牌效应，传染性强、化解难度大，甚至引发区域性金融风险。

关系定义：担保关系的构建，使用行内 GRMS 的担保数据，通过挖掘行内所有的企业担保数据的担保数据信息，最终计算出行内企业担保关系，识别担保链、担保圈、互保等特殊担保关系。目前行内已构建企业对应的保证人链路基础数据。

数据来源：GRMS 数据。

关系示例：



2.2.1.7 供应链关系路径

供应链关系挖掘是通过招投标公告、上市企业公告、法律文书、新闻舆情等数据挖掘，展现企业与上游供应商、下游客户、竞争对手的关系。基于核心企业可以展开一级供应商、二级供应商、三级供应商。

数据来源：外部供应链数据+内部交易数据。目前行内已构建资金流入前十大交易对手、资金流出前十大交易对手及其交易量数据。

2.2.1.8 产业链关系路径

产业图谱定义：产业链的划分标准很复杂，不同类型的产业链可以参考国家、各部委、地方的相关的产业报告及产业/产品分类方法，并结合券商、咨询机构等的研究报告加以梳理而成。

企业上链：通过企业主营业务挖掘模型，确定企业和产业的最终关联度。对于上市和部分发债企业主要使用其公告的年报数据中的主营业务匹配；针对非上市公司，则根据上市公司及相关产品的龙头公司的企业画像总结了产品词的映射字典，并搭建了产品企业关联度模型，使用模型对包括但不限于工商主营、招投标、投融资、资质、知识产权等在内的数据源进行了挖掘匹配，并搭配了人工抽检、复核等方式来保证数据的准确性。

数据来源：外部产业链数据+内部客户数据。目前行内已构建资金流入前十大交易对手、资金流出前十大交易对手及其交易量数据。

2.2.1.9 事件关系路径

事件关系的构建主要数据来源是舆情新闻事件内容，通过对舆情事件的分析与总结，将事件中出现的企业进行整合和挖掘，将所有出现在同一个事件中的企业进行关联与呈现。即将一篇舆情中所有实体提取，并打上舆情主题标签，同一个舆情里面的主体产生关联关系。

事件关系中，总行采购的外部数据的每篇新闻带有一个情感属性，分别是负面、非负，以及有相应的预警分类重要性评分，便于评估事件对客户的影响，从繁杂的新闻数据中做“减法”，让用户的关注点更为聚焦。

默认呈现 3 个月内的事件关系列表，最多支持 6 个月，新闻事件一般是第三方媒体的客观报道，通过新闻事件可间接的评估企业的经营状态，以及面临的风险等。

数据来源：舆情数据。

2.2.1.10 知识产权关系路径

通过知识产权数据，包括专利数据、著作权数据、商标数据等，挖掘两家企业共同申请了某项专利、软件著作权，定义为具有共同知识产权关系，具备共同知识产权关系的企业可能存在比较密切的商业合作，也可能同属一个集团派系，其中企业若发生风险，也有一定几率发生传导。

2.2.1.11 隐性关系挖掘（高阶）

有些关联方虽然在股权、任职等方面与目标企业毫无关系，但是在经营地址、联系方式、网站域名、邮箱等对外公布的联系渠道中某项信息是相同的，或者是共同申请了某项专利、软件著作权，或者在某场司法诉讼中，共同作为被告或者原告，这样的“蛛丝马迹”可能代表着背后有隐性关联关系，需要逐一排查这些存在相同点的企业是否与目标企业发生过交易，以及是否存在实质性的关联关系。

关系定义：隐性关系，通过大数据分析提取相同经营地址、电话号码、邮箱、域名、邮箱域名共用关键字、专利、法律文书等各类疑似数据维度共同特征，并将特征内容计算关联，围绕目标企业呈现疑似存在关系的企业，深度挖掘、追查隐性关联关系。

数据来源：工商数据、司法涉诉数据、知识产权数据、招投标数据等。

2.2.1.12 识别原发性风险，建立风险事件、社区风险传导模型（高阶）

通过总行的外部数据，必要时引入外部风险大数据，构建风险事件监测模型，将负面事件归集为内部治理风险、对外投资风险、信用风险、应收账款风险、经营风险、合规风险、不可抗力（宏观事件）风险等 7 个风险大类、27 个风险小类，以事件对企业继续经营的威胁程度作为评估标准，对负面事件的严重程度进行评估，实现监控数据的分类分级管理，助力我行提前预警客户高危风险事件，并通过

风险传染网络，监测风险波及的客户范围，提前干涉，减少无效时间投入与人力成本，提升风控质效。

公司治理风险：包括不限于高层人事变动、股东减持或退出、实际控制人变更、人员规模缩减等事件。

对外投资风险：包括不限于投资企业疑似退出市场、投资企业出现债务危机、投资企业受到监管处罚、投资企业经营行为异常、投资企业纳税行为异常等事件。

信用风险：包括不限于疑似或实际发生违约、财产和行为被限制、拒绝或无能力偿还债务等风险事件。

应收账款风险：包括不限于疑似或实际被违约、债务可能面临损失等风险事件。

经营风险：包括不限于虚假经营嫌疑、流动资金不足、生产质量或进度问题、当前或预期业绩不佳、供应商疑似断供、疑似退出市场等风险事件。

合规风险：包括不限于监管关注和处罚、市场抽查结果异常、经营行为异常等风险事件。

不可抗力（宏观事件）风险：包括不限于社会异常事件、公共卫生事件、自然灾害、行业风险等。

以下为风险事件类型的具体定义：

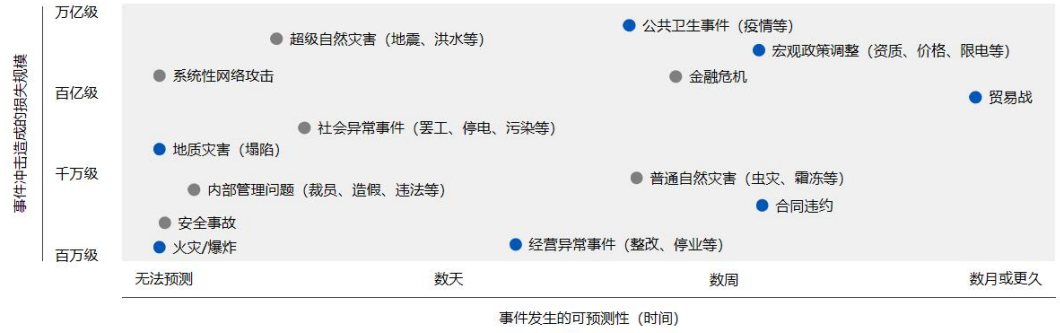
大类	小类	负面事件类型
公司治理风险	高层人事变动	法人变更
		高层变更
		高管无法履职
		职位变动
	股东减持或退出	大股东变更
		股份减持
	实际控制人变更	实际控制人变更
	人员规模缩减	裁员消息
		参保人数骤降
对外投资风险	投资企业疑似退出市场	破产清算
		简易注销
		非经营状态
	投资企业出现债务危机	被纳入失信名单
		无能力偿还债务
		发生债务违约
	投资企业受到监管处罚	行政处罚
		环保处罚
		公开谴责
		立案调查
		市场禁入
		黑名单

大类	小类	负面事件类型
	投资企业经营行为异常	经营异常
		严重违法失信
	投资企业纳税行为异常	欠税信息
		非正常户
		重大税收违法
信用风险	疑似或实际发生违约	被他人起诉
		被起诉且败诉
		票据承兑逾期
		近五年违约规模较大
	财产和行为被限制	股权被冻结
		资产被冻结
		限制高消费
		限制招投标
	拒绝或无能力偿还债务	被强制执行
		资产被司法拍卖
		拒绝偿还债务
		无能力偿还债务
		发生债务违约
应收账款风险	疑似或实际被违约	起诉他人
		起诉且胜诉
	债务可能面临损失	申请财产保全
		债务人疑似失踪或死亡
		申请强制执行
		申请他人破产
经营风险	虚假经营嫌疑	法人任职异常
		高层任职异常
		注册地址异常
		托管、代办地址注册
		联系电话异常
		治理结构异常
		集中申请变更
	流动资金不足	动产抵押

大类	小类	负面事件类型
		股权质押
		土地抵押
		知识产权出质
		股权出质
	生产质量或进度问题	客户投诉
		产品问题
		产品召回
		项目延期
		停工停产
		安全事故
		无证施工
		货物未准入境
	当前或预期业绩不佳	注册资本下降
		业绩亏损/下降
		列入观察名单
		评级展望负面
		评级下调
		证券戴帽
		退市风险
		暂停上市
	供应商疑似断供	整改或停业
		疑似退出市场
		电力供应问题
		空气重污染
		毁灭性伤害
		新冠疫情
		破坏性灾害
		一般灾害
	疑似退出市场	破产清算
		简易注销
		非经营状态
合规风险	监管关注和处罚	行政处罚

大类	小类	负面事件类型
		环保处罚
		安全监管
		虚假宣传
		垄断
		贪污/职务侵占
		行贿受贿
		监管函
		监管约见
		警示
		通报批评
		责令改正
		公开谴责
		立案调查
		市场禁入
		公益诉讼
		违法犯罪
		黑名单
	市场抽查结果异常	抽查检查异常
		双随机抽查异常
	经营行为异常	经营异常
		严重违法失信
	纳税行为异常	欠税信息
		非正常户
		重大税收违法
宏观事件风险	社会异常事件	产业政策变化
		行业景气下降
		贸易制裁
	公共卫生事件	新冠疫情
	自然灾害	破坏性灾害
		一般灾害

其中宏观事件风险并不是起源于单点企业传染，而是形成一个原发性的风险事件，对行内某个客户群体产生影响。



1. 通过行业事件识别原发性风险

识别原发性风险，比如某个宏观政策、行业风险，会影响产业链等，从而传染到相应的企业。具体而言，由于供应链金融不是单一企业融资，且围绕着同一家核心企业的上下游融资企业天然存在风险传导关系，尤其关注自然灾害、公共卫生安全、产业政策变化等因素是否会产生供应链风险，明确风险的传导路径，评估事件发展走向与影响。比如当产业结构进行调整时，国家出台政策支持或限制某个产业的发展，被限制的产业生产规模缩小；市场供需变化导致抵质押资产价格波动或是大幅度贬值；行业性重大风险事件发生，如自然灾害、瘟疫、国际间贸易摩擦升级，都会对被影响产业的供应链金融造成打击。

通过舆情信息的采集、NLP 清洗、分析，可智能监控公共卫生安全、自然灾害、环境保护等供应链风险事件（风险类型持续增加），及其是否对监控对象产生影响。

2. 构建行业景气度模型

产业景气指数从市场关注度、政策的倾向度、战略重要度、资本流向热度和产业竞争度等维度对产业景气状况进行评价。

产业景气时，产业市场关注度、专业关注度、资本流热度等提高；对应产业企业的经济状态趋于上升或改善，处于景气状态。产业不景气时，产业市场关注度、专业关注度、资本流热度等下降，对于产业企业的经济状况处于下降或恶化，处于不景气状态。比如某个行业景气度下降到某个值，会对处在这个领域的核心企业有影响，进而影响到核心企业的上下游，形成传染风险。

例如：我们监测到某个细分产业（行业）景气度骤降，可识别出这个产业链上行内客户有哪些，受影响（传染）的企业（存量客户）有哪些？

受影响的企业群体（社区）在行内的业务结构和授信情风险况如何，可结合行内数据做进一步分析。

综上所述，我们通过搭建基本的显性关联关系图谱，进一步挖掘隐含关系，还对原发性风险识别，建立风险事件、社区风险传导模型进行了探索，为下一步的风险奠定了基础。

2.2.2 客户关系图谱构建

图谱构建就是将各种数据源中获取的数据进行分析和融合，转化成具有实体、关系、属性的基础的图数据结构。总体来说图构建包含三个主要的步骤：

2.2.2.1 构建实体

在定义的实体范围内从多源异质的数据源中抽取出相关的实体信息，然后通过算法引擎对这些实体进行对齐、消歧和融合，最后通过实体构建技术构建出具有唯一性的企业和自然人实体，并最终给每个实体赋予唯一的实体 ID。

2.2.2.2 构建实体间关系

根据构建出来的实体和已有的数据，建立实体和实体之间的关系，并确定关系对应的属性值。实体之间的关系主要从三个方面获取。一方面可以通过已有的结构化数据中直接获取实体之间的关系以及关系对应的属性。第二个方面就是需要通过 NLP 算法、机器学习算法、深度学习算法引擎从非结构化信息中挖掘出实体之间的关系。第三个方面就是在已有的实体和关系的基础上，利用图挖掘算法，获取隐藏在数据中的深层次的关系。

客户关系知识图谱构建的数据来源分为行内自有数据和外部数据，需提取出相关联实体之间关系，关系包括但不限于股权关系、交易关系、担保关系、企业上下游关系、其它疑似关系等，其中股权类至少包括实际控制人识别、股权控制结构、集团客户识别、企业集团架构等；交易类至少包含关联交易分析；担保类关系至少覆盖担保链、担保圈、互保关系等分析；上下游关系类至少包括应收应付异常识别、应收应付风险预警分析。

2.2.2.3 构建实体属性

根据已有的实体及其相关的数据，构建实体的标签属性。实体的标签属性主要从三个方面来构建。第一个方面是通过已有的结构化的数据中直接抽取。第二个方面就是利用 NLP 算法、机器学习算法和深

度学习等算法构建实体属性挖掘模型，获取更深层次的实体属性。第三个方面就是根据特定的业务，例如预警信号等，构建实体的动态智能标签。

2.2.3 客户关联图检索和展示

最终构建完成的客户关系知识图谱的实体与关系数量规模巨大，如何在可接受时间内完成海量知识检索是一个巨大的挑战。本项目的关键之一是实现面向大规模知识图谱的信息检索方法。

图展现能展示出图构建中构建的实体、关系和属性。以单一客户、集团客户、系客户、行业客户、地域客户等的形式展现知识图谱。

展现的方式支持网络装展示和树状展示。网格状展示：除展示与中心实体关联关系外，还可展示其他关联实体之间的关联关系。树状展示：展示中心实体关联的对应实体。

支持点击知识图谱中某个实体，查看实体详情（包括法定代表人、注册资本、注册时间、社会统一信用码、经营范围等）。

2.2.4 关于图数据库选型

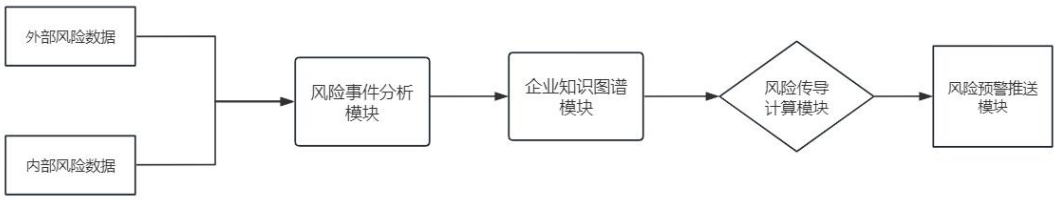
传统的关系型数据库并不能很好地表现数据的联系，而一些 NoSQL（非关系型数据库）数据库又不能表现数据之间的联系。同样属于 NoSQL 范畴的图数据库是以图的结构形式来存储数据的，它所存储的就是联系的数据，是关联数据本身。Neo4j 作为主流的图数据库之一，专门用于网络图的存储。在图数据库领域，除 Neo4j 之外，还有其他如 OrientDB、Giraph、AllegroGraph 等各种图数据库。跟这些图数据库相比，Neo4j 的优势表现在以下两方面：

1.Neo4j 是一个原生图计算引擎，它存储和使用的数据自始至终都是使用原生的图结构数据进行处理的；

2.Neo4j 是一个开源的数据库，其开源的社区版吸引了众多第三方的使用和推广，同时也得到了更多开发者的拥趸和支持，聚集了丰富的可供交流和学习的资源与案例。

所以本项目拟选择 Neo4j 用于企业知识图谱关系数据存储。

2.2.5 基于企业关系网络的风险传导模型设计



风险传导的实现包括四要素：监控标的、风险事件、传导链路以及风险传导结果量化。

监控标的（关系网络上的节点）：包括企业本身、企业的董事、监事、高管，企业的重要资产等能对企业的经营造成重大影响的人、事、物。

风险事件识别：围绕监控标的，对其发生的风险信息进行智能化的捕捉、加工，形成风险标签信息。

传导链路：传导链路即为企业间的关联关系，如本文上部所述各式各样的关联关系。常见的传导链路有产业链关联关系、集团关联关系、以股权债权为基础的投资关联关系等。

风险结果量化：针对传导链路上的不同节点，根据不同企业的关系强弱情况与其在传导链路中的不同位置，利用机器学习+专家经验，量化风险事件对其的影响。

2.2.5.1 模型算法设计

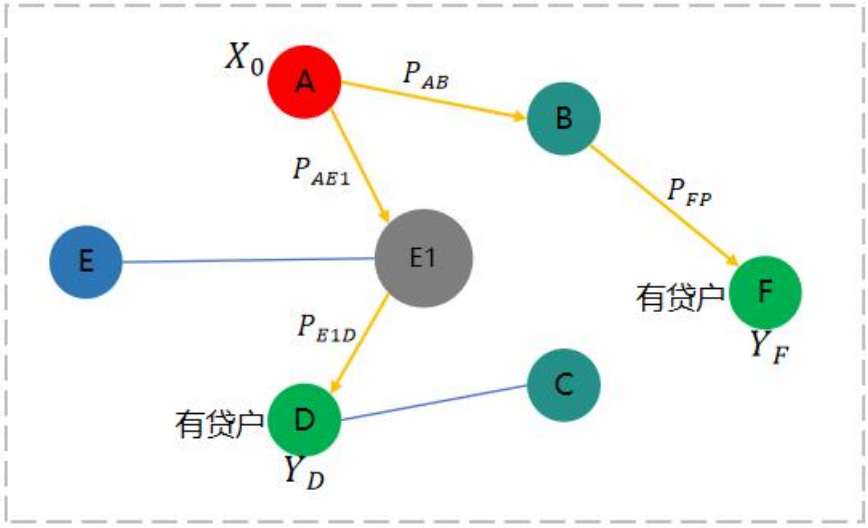
基本传导算法：K 层 BFS 传导

最大/最小风险企业计算：转化到最短路径问题或 Veterbi 算法求解

单风险源多种风险类型同时传导计算：向量计算

单风险源扩展到多路径风险源传导：并行叠加计算

如场景 1：当集团内一家企业发生风险时，银行需要对集团内的有贷户进行授信风险评估，需要分析集团内部产生风险的企业对授信主体产生的影响，并对该授信主体进行风险评估及风险信号预警。



解决方案：

构建集团内部关联企业风险传导图谱，A 为集团内部发生风险的企业，D、F 为银行授信主体（有贷户），E1 为企业集团。

根据集团内部企业风险信息及自身特征信息，量化企业风险危害值 X；

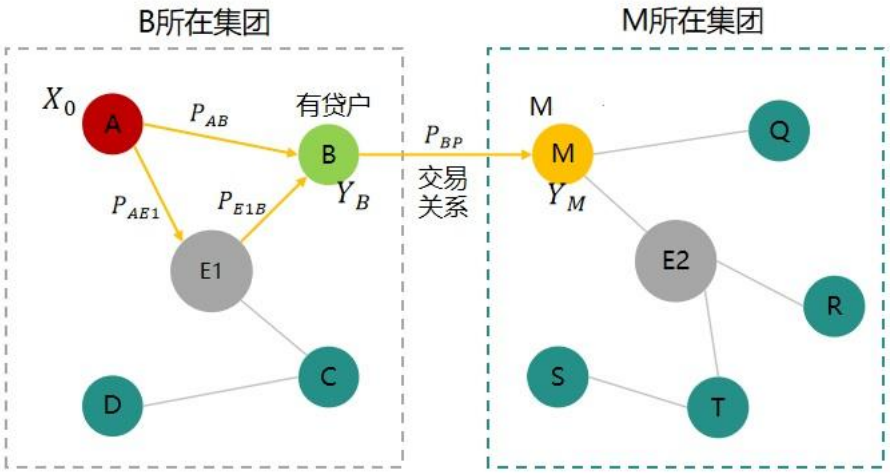
利用集团内部的所有关系信息，计算 D、F 的风险传导路径：

D 的传导路径：A → E1 → D

F 的传导路径：A → B → F

运用风险传导基本算子和多路累加运算进行 K 层 BFS 传导，计算出 D、F 的企业风险值 Y。

如场景 2：在银行信贷风控管理的业务场景中，银行需要对行内有贷户进行全面的风险传导分析，当一家企业发生风险时，对其他所有存在关联关系的有贷户的风险传导影响，以及对银行整体的风险影响分析。



解决方案：

构建集团内部关联企业风险传导图谱，A 为发生风险的企业，B 为银行授信主体（有贷户），M 为被传染企业，E1、E2 为企业集团。

根据 B 所在集团内部企业风险信息及自身特征信息，量化企业风险危害值 X；

利用 B 所在集团内部的所有关系信息，计算 A 产生风险对 B 的风险传导路径，如图有两条路径：

路径 1：A → B

路径 2：A → E1 → B

运用行内交易数据构建模型，计算 B → M 的风险传导，并输出 P 的风险值 Y。

本次项目拟采用机器学习+图算法的方式进行模型开发。

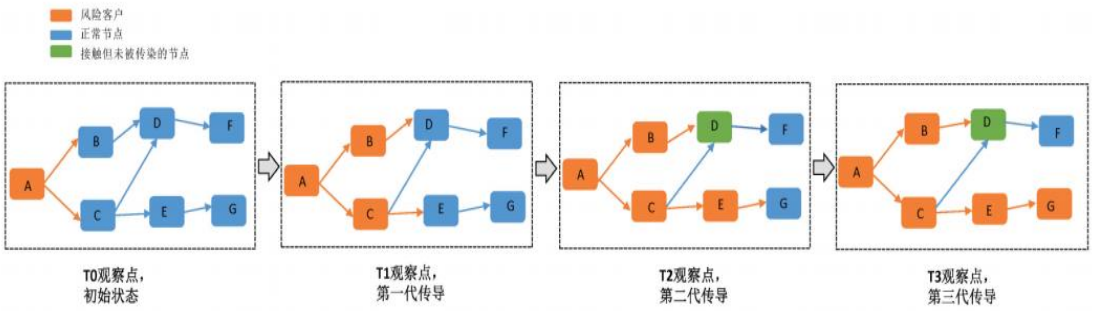
基于知识图谱技术的企业风险监测及传导分析方法，构建基于数据驱动的风险传导模型，设计 N 个特征变量作为模型的输入，同时，模型需要针对各种风险类型分别建模计算。在模型训练方面，拟采用 LightGBM 算法，它是一种具备训练效率高、内存使用低、支持直接使用类别特征的机器学习算法。同时，专家经验方法和机器学习算法两者结合，可以避免人工规则及方法的片面性与主观性，也可以弥补深度学习算法在解释性方面的不足，更方便于模型的更新迭代。

第一、将各类关联关系按类别分别进行特征嵌入（Embedding）从而将关系表达为结构化的数据—向量，然后再将每类关系对应的 Embedding 向量和结构化数据拼接为宽表并作为机器学习模型的输入特征进行模型训练。

第二、采用 GCN（图卷积神经网络）的方式，直接将企业在图拓扑空间中的邻居的特征（结构化数据）和自身的特征进行聚合从而得到新的特征，并以此特征进行模型训练。图卷积是一个半监督的学习方式，当坏样本数目不是很多的情况下采用此方式建模往往会取得优秀的训练结果。

第三、利用同业经验和业务经验，将机器学习算法的模型进行业务视角的特例事项的调整，如股权链中母公司至子公司的风险传导以及子公司至母公司的分析风险传导方式差异，通过业务属性中的调整因子进行体现。

2.2.5.2 模型特征变量设计



企业间的风险传导包括但不限于以下几种方式。

1. 由股权关系产生的传导。母公司发生风险后有可能处置子公司的股权，使原本经营正常的子公司出现困难，面临信誉风险。

2. 由担保关系产生的传导。被担保公司发生违约事件后，原本经营正常的担保公司为履行代为赔偿责任，资金大量流失，面临债务负担加重。

3. 由供应链关系产生的传导。下游企业财务出现困难，无法按时支付款项，如果上游企业资金长期无法回笼，可能面临资金链断裂的风险。

4. 由共同控制人产生的传导。具有隐性关联关系的企业之间，当共同控制人出现问题，原本正常经营的企业会由此受到波折，面临流动性风险。

基于以上风险分析，在指标构建阶段，我们首先通过股权关系、出资比例、担保关系、供应链关系、控股层级、企业实际控制人关系等维度设计特征变量，利用图算法等模型挖掘特征变量，利用特征变量来刻画企业间关联关系；通过机器学习模型求解特征变量权重系数，形成风险传导边参数。

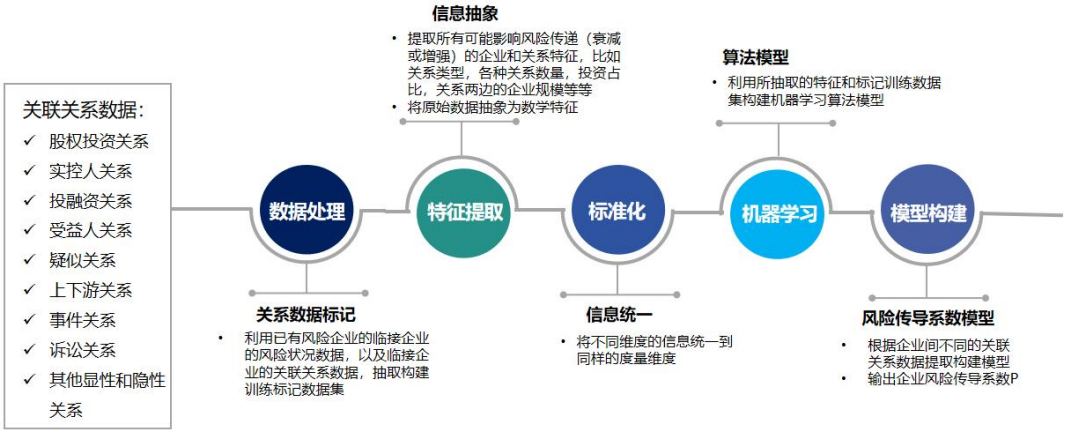
2.2.5.3 风险传导系数

通过机器学习（逻辑回归模型）求解特征变量权重系数，形成风险传导边参数。同时，结合相关性分析的结果，对高相关的风险传导边权重结合信息值占比予以调整，以确保风险传导边权重排序的意义。

风险传导系数模型

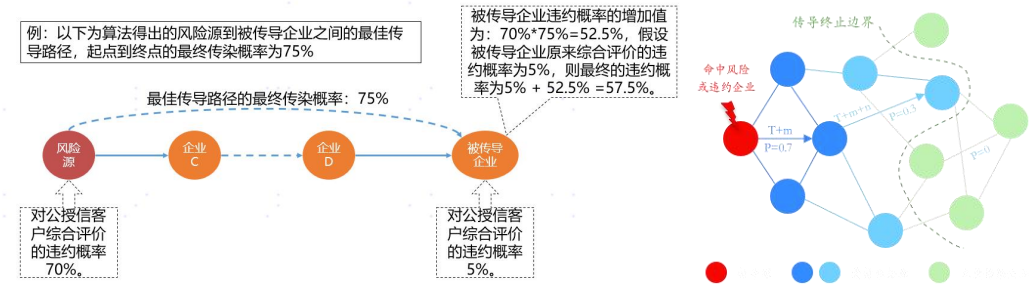
输入：企业关系特征、企业自身特征

输出：风险传导系数（P）



2.2.5.4 风险传导量化结果

依托对公授信客户综合评价结果风险源的违约概率，乘以最佳传导路径的传染概率，算出被传导对象违约概率的增加值，最终建立起综合基于客户自身风险和受传染风险的全息风险综合评价体系。



传导路径的传染概率可按不同的关联关系量化，举例如下：

风险传导类别	关联关系与量化分析
企业与企业间的传导	股权/集团关系 子公司/被投公司->母公司/参股公司的传导概率 = 固定股权参数（需确认母公司的出资信息） 母公司/参股公司->子公司/被投公司传导概率 = 固定股权参数 * 母公司持股比例
	担保关系 担保公司->被担保公司的传导概率 = 固定担保参数 * 担保比例 被担保公司->担保公司传导概率 = 固定担保参数 * 担保金额 * 担保比例
	交易关系 上游企业->下游公司的传导概率 = 固定交易参数 * 交易(订单量) 占比 * 应付占比 下游企业->上游公司的传导概率 = 固定交易参数 * 交易(订单量) 占比 * 应收占比
人与企业间的传导	高管与企业关系 法人(实控人)->个体工商户/小微企业的传导概率 = 固定参数 * 持股比例 高管-> 企业的传导概率 = 固定参数 * 持股比例

2.2.5.5 模型训练样本选择

在模型设计阶段，针对于不同的风险类型的企业，正样本的定义也有差异。在样本选择阶段，以已经发生风险的企业作为目标客户群体，与目标客户群体有接触（或业务关联）后亦发生风险的企业客户群体作为正样本。从接触到风险暴露经历的时间定义为表现期，接触后但未发生风险的企业客户群体作为模型开发的负样本。表现期窗口长度的 2~3 倍作为观察期时间窗口，用来加工特征变量。

通过历史违约记录和关联路径违约记录等确认建模样本池，比如股权链传导：会通过股权的占比、出

资比例、控股层级选择建模样本和关键特征，如担保链传导：会利用担保总额、担保期限、担保人评级、借款人规模等选择建模样本和关键特征。同时，在正样本充足的情况下，我们考虑基于不同的风险传导方式进行分类建模，以增强模型的预测能力和模型效果的稳定性。在样本不足的情况下，可以将四类不同风险传导方式的建模样本进行合并分析，利用图算法和 GCN（图卷积神经网络）的方式训练模型，避免因样本量不足而训练不出好的模型的问题。

2.2.6 风险可视化系统开发

2.2.6.1 系统功能界面设计

2.2.6.1.1 客户风险驾驶舱

1.客户风险概览

基于风险预警模型动态跑批结果，针对不同风险等级进行预警展示，展示高风险企业数量、低风险企业数量、疑似风险企业数量、提示风险企业数量，同时对于业务重点关注的指标进行展示或可视化的图表分析，并以轮播的方式展示最新企业风险动态。

点击如“高风险”，下钻至二级页面，展示当前处于高风险预警等级的企业清单，点击企业可进入企业风险档案，追溯企业风险详情。

2.风险预警信号

以企业为单位按时间展示风险预警信号，涵盖企业自身风险和关联传导风险，帮助业务人员及时发现客户风险，实现早期预警。展示风险事件基本信息和事件传导图谱。

2.2.6.1.2 企业风险画像

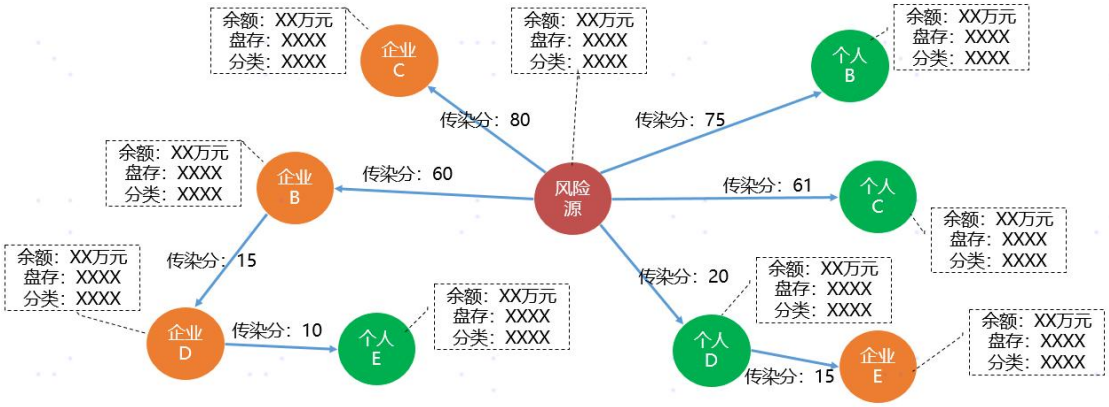
企业基本信息：展示企业基本信息，包括工商照面、近期风险、授信余额、贷款余额、敞口等信息。

风险研判：展示企业综合风险研判结果，包括违约风险（及概率）、经营风险、司法风险、舆情风险等。

2.2.6.1.3 风险传导图谱

以行内客户为主，生成一张以风险源（可扩展至所有盘存、分类非正常及其他正常类客户）为核心的关联传染得分图谱。即分别以股权关系、担保关系、资金往来关系、高管关系、雇佣关系等维度，展示各维度企业、个人的关系分布图以及各主体在我行的授信余额、盘存类别及分类等情况，每个关联线路都展示出风险源与被传染对象之间最佳传导路径的传染分（由起点到终点的最终传染概率折算而来）。

以图谱形式展现风险事件在行内企业和关联企业之间的传导，支持以行内企业为中心查询并展示自身风险事件和关联风险事件的传导路径，帮助业务人员发掘近期可能对行内客户造成影响的风险事件。图谱支持显示风险企业之间的关联关系、源头风险事件、目标企业的风险传导分及违约概率等信息。



2.2.6.1.4 关联关系图谱

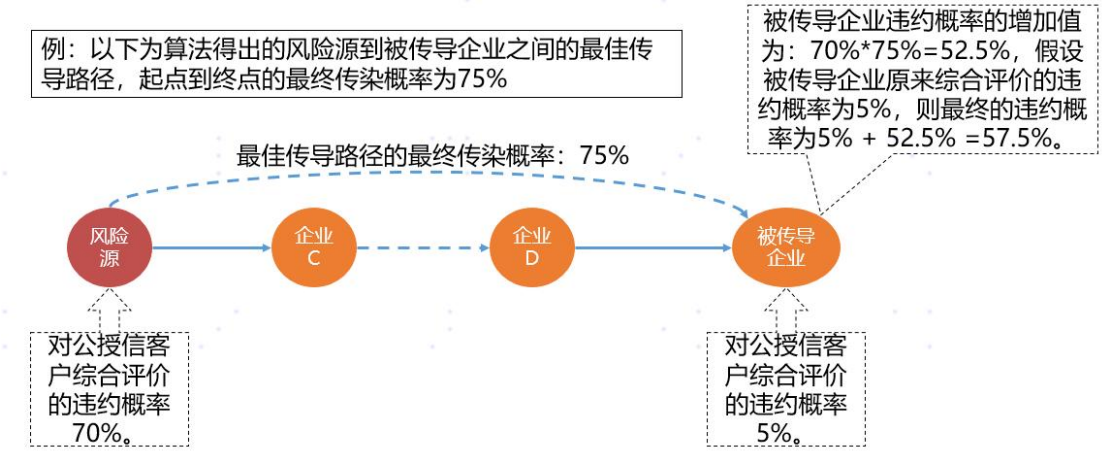
以图谱形式展示行内企业和关联企业（拓展到非我行客户）之间的关联关系，涵盖股权关系、担保关系、资金链关系等。

关联关系筛选：支持基于关联关系强弱对图谱中的企业进行筛选，例如筛选关系类型、持股比例等。

企业信息概览：选中节点支持显示企业的基本信息、风险信息。风险信息涵盖近期风险标签、预警信号和风险走势。

2.2.6.1.5 关联关系违约概率传染结果

依托对公授信客户综合评价结果风险源的违约概率，乘以最佳传导路径的传染概率，算出被传导对象违约概率的增加值，最终建立起综合基于客户自身风险和受传染风险的全息风险综合评价体系。



2.2.6.1.6 信息推送与应用

传染风险预警信号的推送等应用设计及相关界面，设置是否推送的开关，并可在界面中个性化选择相应的预警信号。

2.2.6.2 系统非功能要求说明

1.系统采用微服务容器化部署的架构，满足行方平台的接入要求，实现各组成部分灵活组装。通过分层架构的设计方式，实现数据、服务、交易和展示的有效分离，实现业务数据和业务逻辑的分离，系统模块间松耦合，功能分布合理，以适应 IT 系统、服务、产品、流程的变化。

2.系统开发遵循微服务化、模块化、参数化设计原则，保持软件系统架构的易于改造和扩展，满足新业务功能的不断扩充，不影响应用系统的各种原有功能。

3.系统结构合理、效率高、资源占用率低，避免过多的数据冗余。提供横向扩展和纵向扩展能力，以提高系统处理性能，满足业务量和数据量的快速增长。

4.系统采用开放平台设计，采用独立于中间件平台、数据库平台的开发技术，软硬件平台具备开放性、通用性、标准性和安全性的特点。

5.基于组件化设计和发布，提高代码的可重用性。具有相对独立、自包含的功能，具有良好契约的接口，可独立开发、独立测试，实现独立于编程语言，可通过组件替换进行功能升级及扩充，保证软件系统架构的易于改造和可扩展性。

第三章 其他要求和建议

3.1 数据标准及质量要求

本需求属存量系统范畴，逻辑数据模型应与行内相关要求一致，并遵循有关制度与管理办法，不涉及影响数据质量的问题。

3.2 开发方式建议

因该项目实施时间要求紧迫，在我行信科人力资源紧张的情况下，为了保证项目的顺利实施，需外购技术服务协助开发，项目计划于 2024 年 6 月投产。

1.供应商依赖性风险

在合同稿要求供应商在项目验收时提交完整的技术文档及程序源代码等相关资料。项目的知识产权归属我方，通过加强供应商对相关人员的转培训和技术传承培训，不断降低对外包商的过度依赖。

2.供应商监控管理风险

为降低项目对供应商的依赖性，尽量减少项目商在项目中途意外离场造成的影响，项目将采用由信息科技部负责需求分析、系统设计、核心流程开发、项目总体测试及投产，外围处理前置系统、前台页面、报表模块向开发商购买外包服务的方式完成项目开发。同时在合同中要求开发商在项目验收时提交详尽的设计文档及程序源码，可供后续部署到总行的集团全面风险管理门户。

3.3 风险传导模型开发技术方案

3.3.1 建模需求分析、开发设计

本次项目目标为：基于企业内外部风险信息、客户风险系统中监管内部数据以及可获取的外部数据，采用前沿的机器学习/图算法的方法开发企业风险预警模型、实现客户风险前瞻性预警；将企业的信用风险预警纳入统一模型处理并配合接入现有的智能监控平台。

在对本次项目需求进行充分、准确理解的基础上，基于知识图谱技术的企业风险监测及传导分析方法，构建基于数据驱动的风险传导模型，专家经验方法和机器学习算法两者结合，可以避免人工规则及方法的片面性与主观性，也可以弥补深度学习算法在解释性方面的不足，更方便于模型的更新迭代。同时，模型需要针对各种风险类型分别建模计算。

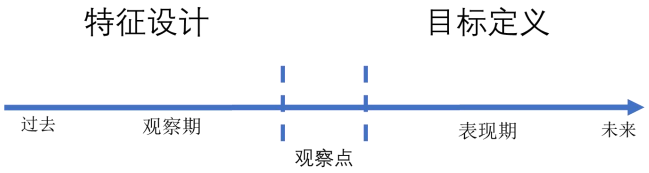
模型开发设计主要包括以下几项内容。

3.3.1.1 确定目标客户

以授信企业为本次模型开发的目标客户群体。

3.3.1.2 观察期、表现期的确定

观察期和表现期是模型设计中重要的一个环节。下图直观的展示了观察期和表现期的定义方式：



表现期：构造特征的时间窗口为观察期。本次项目模型开发中所需的特征（X）将在观察期内进行特征提取和特征衍生。综合考虑数据的时效性、预测性以及特征的完整性、本次项目中，拟将观察期定义为 1 年的时间窗口；

表现期：目标变量 Y 需要在表现期内进行定义，通常 Y 定义为在表现期内为坏的客户，因此表现期的选择往往对实现需求目标尤为重要。表现期越长，信用风险暴露将越彻底，但意味着观察期离当前越远，用以提取样本特征的历史数据将越陈旧，建模样本和未来样本的差异也越大。反之，表现期越短，风险还未暴露完全，但好处是能用到更近的样本。为了实现企业风险预警的前瞻性，本次拟采用 6 个月的时间窗口作为表现期。

3.3.1.3 目标变量定义

在确定好目标客户和表现期、观察期窗口后就可以定义目标变量 Y：

企业在表现期（6 个月）内逾期天数大于 3 天的定义为坏客户，即 Y=1；

企业在表现期从未逾期的定义为好客户，即 Y=0；

企业在表现期内逾期天数为 1 到 3 天的作为中间客户（灰色客户），表现期不足的也作为中间客户。中间客户不进入最终的模型开发样本。

3.3.1.4 样本构建

建模样本构建需要结合具体建模场景而定。此次项目场景拟以最近 2 年内大企业客户作为模型开发的样本集，特征 X 和目标变量 Y 分别在观察期、表现期时间窗口获取。同时，可采用滑窗的方式增加样本数量。另外，对于中间客户（逾期天数 1 到 3 天、表现期不足 6 个月的企业）需要从样本中排除。

3.3.1.5 宽表设计

本次项目数据来源和上述大型企业数据来源相同，主要包括企业的基本面数据、工商数据、司法数据、财务数据、企业舆情等。基于以上数据来源，我们设计了企业预警模型所需的字段，以下为部分主要字段表和字段。

授信额度	企业(机构)类型	企业税号
贷款授信额度	注册资本	企业名称
贷款余额	最后年检年度	企业类型
持有债券余额	最后一次工商变更日期	切片时间
持有股权余额	最近一段时期法人变更次数	是否制造业
其他表内信用风险资产余额	最近一段时期工商变更次数	是否销户
表外业务余额	最近一段时期股份超过28%的股东股权变更次	是否增容
现有业务余额占用授信额度	最近一段时期高管备案变更次数	是否减容
贷款余额占用贷款授信额度	是否属于银监会逾期30天以上客户清单	违约金分段
尚可使用授信额度	是否银监会多头授信超5家银行客户	违约金占比
尚可使用贷款授信额度	是否银监会授信资产比超过150%客户	是否愉窃用电
发放金额	不良客户不良类型名称	是否非他违约用电
贷款余额	大额授信客户等级名称	景早用电月份数量
五级分类	分类施策结果名称	有效户号数量
贷款类型	集团客户类型名称	月电量企业范围分段
贷款业务种类	客户准入类别名称	日电量分类范围分段
投向行业	是否敞口客户	日电量企业范国分段
币种代码	是否关联复杂	月电量分类范围分段
担保方式	是否集团内保	月电费企业范围分段
欠本余额	是否绿色信贷客户	月电贵分类范围分段
欠本天数	是否司法查询	半年度也费分段
欠息余额	是否司法冻结	半年度动的金分段
欠息天数	是否循环担保	季度电量分段
本期还款	最后一次被起诉时间	参度电费分段
还本方式	最后一次司法冻结账款时间	季度违约金分段
还息方式	最近2年行政处罚次数	半年度电量分段
下期还本日期	最近3年作为被告发生法律诉讼的次数	年度电量分段
下期还本金额	最近3年作为原告发生法律纠纷的次数	年度电贵分段
下期还息日期	近1年还款次数	年度违约金分段
下期还息金额	近1年还款金额	季度，半年度，年度的电量，电费，违约金分段间隔值(多个用_拼接)

贷款发放类型	往账账户是否是网贷平台	流动负债合计
减值准备	最近1年司法冻结次数	财务报表类型
产业结构调整类型	最近1年司法冻结的金额	财务报表日期
工业转型升级标识	征信对外保证担保金额	客户所属行业代码
战略新兴产业类型	征信对外抵押担保笔数	贷款卡号
银团贷款标识	注册资本	股东/关联企业名称
支付方式	主营业务收入/流动负债	股东/关联企业类型
关联集团代码	主营业务收入/应付债券	股东/关联企业证件类型
关联集团名称	营业利润增长率	股东/关联企业证件代码
关联关系类型	股东权益增长率	登记注册代码
风险预警信号	营业收入增长率	股东/关联企业客户代码
关注事件	资产负债率	国别代码
违约概率	财务费用前资产利润率	持股比例
信用评级结果	净利润现金比率	股东结构对应日期
资产总额	成本费用利润率	更新信息日期
负债总额	主营业务收入/总资产	实际控制人标识
税前利润	营业利润率	担保合同号
主营业务收入	成本费用率	担保合同类型
存货	短期负债占比	押品类型
应收账款	主营业务成本/存货	押品名称
其他应收款	每股净利润	押品代码
流动资产合计	速动比率	押品权属人或保证人名称

3.3.2 模型开发

3.3.2.1 建模算法选型

3.3.2.1.1 随机森林

随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习方法。随机森林的“随机”主要体现在两个方面，数据的随机性

选取，以及待选特征的随机选取，来消除过拟合问题。

随机森林算法的特点：

1）在当前所有算法中，具有极好的准确率。

2）能够有效地运行在大数据集上。

3）能够处理具有高维特征的输入样本，而且不需要降维。

4）能够评估各个特征在分类问题上的重要性。

5）在生成过程中，能够获取到内部生成误差的一种无偏估计。

6）对于缺省值问题也能够获得很好得结果。

3.3.2.1.2XGBoost

Boosting 算法的其中一种。Boosting 算法的思想是将许多弱分类器集成在一起形成一个强分类器。因为 XGBoost 是一种提升树模型，所以它是将许多树模型集成在一起，形成一个很强的分类器。而所用到的树模型则是 CART 回归树模型。

该算法思想就是不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。当我们训练完成得到 k 棵树，我们要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数，最后只需要将每棵树对应的分数加起来就是该样本的预测值。

XGBoost 优点：

1）使用许多策略去防止过拟合，如：正则化项、Shrinkage 和 Column Subsampling 等。

2）目标函数优化利用了损失函数关于待求函数的二阶导数

3）支持并行化，虽然树与树之间是串行关系，但是同层级节点可并行。具体的对于某个节点，节点内选择最佳分裂点，候选分裂点计算增益用多线程并行，训练速度快。

4）添加了对稀疏数据的处理。

5）交叉验证，early stop，当预测结果已经很好的时候可以提前停止建树，加快训练速度。

6）支持设置样本权重，该权重体现在一阶导数 g 和二阶导数 h，通过调整权重可以去更加关注一些样本。

3.3.2.1.3LightGBM

LightGBM 适用于大样本和高维度的环境。

LightGBM 是个快速的，分布式的，高性能的基于决策树算法的梯度提升框架。可用于排序，分类，回归以及很多其他的机器学习任务中。

传统的 Boosting 算法（如 GBDT 和 XGBoost）已经有相当好的效率，在大样本和高维度的环境下，传统的 Boosting 似乎在效率和可扩展性上不能满足现在的需求了，主要的原因就是传统的 Boosting 算法需要对每一个特征都要扫描所有的样本点来选择最好的切分点，这是非常的耗时。为了解决这种在大样本高纬度数据的环境下耗时的问题，LightGBM 使用了如下两种解决办法：一是 GOSS（基于梯度的单边采样），不是使用所用的样本点来计算梯度，而是对样本进行采样来计算梯度；二是 EFB（互

斥特征捆绑），这里不是使用所有的特征来进行扫描获得最佳的切分点，而是将某些特征进行捆绑在一起来降低特征的维度，使寻找最佳切分点的消耗减少。这样大大的降低的处理样本的时间复杂度，但在精度上，通过大量的实验证明，在某些数据集上使用 LightGBM 并不损失精度，甚至有时还会提升精度。

3.3.2.1.4 图嵌入（Embedding）

网络分析中涉及对节点和边的预测，但是想要利用图中的信息是比较困难的，因为图本身是离散的。因此，我们需要使用一种方式将图结构转化为便于计算的表达方式。

本次项目拟采用 Node2vec，对企业之间、企业法人、股东、上下游等关系进行知识抽取，构建新的风险特征。

3.3.2.1.5GCN 基于图特征的表示学习

特征提取器拟选择 GCN（图卷积神经网络），基于图特征的表示学习对结点的向量表示既包含了图的拓扑信息（x 的邻接矩阵表达的图结构）也包含了已有的特征向量（各个维度为包含结点特征的向量，如交易金额、担保金额等信息）。本次项目拟采用 GCN，综合利用企业结构化数据和非结构化数据，构建客户全方位的特征向量。

3.3.2.2 数据准备

ETL 人员根据模型设计部分提供的宽表特征进行指标加工、宽表设计人员需参与 ETL 工作，以确保 ETL 过程中逻辑的正确性；并对最终获取的宽表进行交叉验证，以保证建模样本的正确性。

在宽表提取完成后、建模人员对宽表数据进行探索、从而检查数据的质量、数据是否存在异常值、分析训练集和预测集样本分布的差异以及特征的分布等情况。数据探索主要包括以下几个方面：

（1）检查数据质量，确保数据一致性，同时确保特征类型不会出现异常。

（2）比较训练集和预测集的异同，包括特征数量、特征类型、记录条数等等。

（3）考察训练集中标签特征的分布状况，以确定后续是否进行不均衡抽样。

（4）分析挖掘训练集中每个特征的分布情况、取值情况、与标签特征的相关性等等，可以使用作图、显著性检验、相关系数等方法。

（5）考察训练集特征的缺失状况、零值状况，以便进行后续的处理。

数据预处理

（1）无用特征：某些特征如客户 ID、客户具体地址信息、客户名称、手机号码等等，明显对分析挖掘没有意义，可以考虑进行删除操作。

（2）唯一取值特征处理：唯一属性通常是一些 id 属性，这些属性并不能刻画样本自身的分布规律，所以简单地删除这些属性即可。

（3）缺失特征处理

造成数据缺失的原因是多方面的，主要可能有以下几种：

1）有些信息暂时无法获取，致使一部分属性值空缺出来。

2）有些信息因为一些人为因素而丢失了。

3）有些对象的某个或某些属性是不可用的。如一个未婚者的配偶姓名。

4）获取这些信息的代价太大，从而未获取数据。

缺失值处理的一般原则：

1）信息损失最少；

2）信息增益最大；

3）尽量不改变原有的数据分布。

通过考察缺失比例的大小，综合数据的整体状况，可以删除缺失比例超过某一阈值的特征；针对缺失比例较少的特征，进行填充处理：

（4）零值特征处理

考察零值特征主要指的是数值型特征，分为整型和浮点型。

零值产生的原因:

1）数据本身是零值。

2）在数据加工的时候，必须把空值置为零值，方可进行某些运算。

3）有些异常值无法识别，比如乱码、非数值型，需要进行零值转化。

零值处理方式：1）使用缺失值的处理方式；2）其他处理方式。如果零值率超过 50%，进行特征衍生，即零值置为 0，非零值置为 1，这样数值型特征衍生为类别型特征。数值型特征，如果零值率小于 50%，不做处理。

（5）异常值处理

异常值，即在数据集中存在不合理的值，又称离群点。比如年龄为-1，笔记本电脑重量为 1 吨等，都属于异常值的范围。从集合角度来看，异常值即离群点，如下图所示：

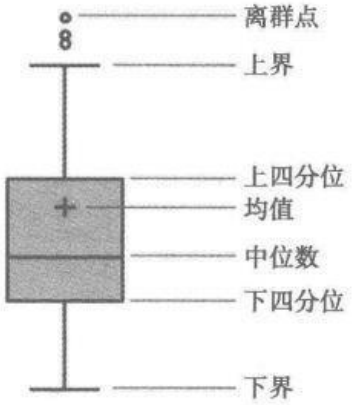
异常值判别方法

1）简单统计分析：对属性值进行一个描述性的统计，从而查看哪些值是不合理的。比如对年龄这个属性进行规约：年龄的区间在[0:200]，如果样本中的年龄值不再该区间范围内，则表示该样本的年龄属性属于异常值。

2）3δ原则：根据正态分布的定义可知，距离平均值 3δ之外的概率为 P(x-μ>3δ)<=0.003，这属于极小概率事件，在默认情况下我们可以认定，距离超过平均值 3δ的样本是不存在的。因此，当样本距离平均值大于 3δ，则认定该样本为异常值。

3）箱型图分析

箱型图提供了一个识别异常值的标准，即大于或小于箱型图设定的上下界的数值即为异常值，箱型图如下图所示：



首先我们定义下上四分位和下四分位。上四分位我们设为 U，表示的是所有样本中只有 1/4 的数值大于 U。同理，下四分位我们设为 L，表示的是所有样本中只有 1/4 的数值小于 L。那么，上下界又是什么呢？我们设上四分位与下四分位的插值为 IQR，即：IQR=U-L。那么，上界为 U+1.5IQR，下界为：L-1.5IQR。

箱型图选取异常值比较客观，在识别异常值方面有一定的优越性。

异常值处理方法：检测到了异常值，我们需要对其进行一定的处理。而一般异常值的处理方法可大致分为以下几种：

1）删除含有异常值的记录：直接将含有异常值的记录删除；

2）视为缺失值：将异常值视为缺失值，利用缺失值处理的方法进行处理；

3）平均值修正：可用前后两个观测值的平均值修正该异常值；

4）不处理：直接在具有异常值的数据集上进行数据挖掘.

（6）特征衍生：特征衍生是指对原始数据进行特征学习得到新特征。特征衍生有两种原因：数据自身的变化，使数据中出现很多原来没有的特征；进行特征学习时，算法根据特征之间的某种关系，产生了衍生特征，有时衍生特征更能反应数据特征之间的关系。

在具体项目中，我们可以将日期特征衍生为天数，还可以根据当前月份的数据生成上月、下月、同比、环比或差分类的数据。

3.3.2.3 特征选择

在用机器学习模型方开发预警规则之前，需要对自变量进行筛选。比如我们有 400 个候选自变量，通常情况下，不会直接把 400 个变量直接放到模型中去进行拟合训练，而是会用一些方法，从这 400 个自变量中挑选一些出来，放进机器学习模型，形成入模变量列表。特征选择的目的，就是将有价值的特征提取出来，同时特征的减少，可以很大程度上避免特征之间的相互干扰，从而保证机器学习给出的预测结果更为稳定。

特征选择和机器学习算法两者存在紧密的联系，根据特征选择中子集评价标准和后续学习算法的结合方式可分为嵌入式(Embedded)、过滤式(Filter)和封装式(Wrapper)式三种。

3.3.2.4 模型开发

经过前面的数据分析探索、数据预处理，得到可以用来进行模型开发的数据。标准建模过程包括以下

4 个步骤：

分割训练数据。一部分作为算法训练、另一部分作为测试算法的效果。

确保验证数据的处理与训练数据标准是一致的。

项目的最终目的是把客户分为目标客户（Y=1）和非目标客户（Y=0）两类，算法采用上述技术要点部分的各种算法进行学习、择最优模型为最终模型。

3.3.2.5 模型评估

3.5.1 模型区分度：KS、AUC 及 Gini 系数均为区分度的衡量

K-S 检验即 Kolmogorov-Smirnov 检验，是基于累积分布函数，用以检验一个经验分布是否符合某种理论分布或比较两个经验分布是否有显著性差异。这里我们采用的是单样本 K-S 检验，即检验一个数据的观测经验分布是否是已知的理论分布，当两者间的差距很小时，推断该样本取自已知的理论分布。将模型结果对于样本进行评分，计算样本变坏概率，并对好样本和坏样本分别进行升序排列，K-S 统计量即为二者累计密度函数之间的最大差值。该差值越大，评分卡的预测能力越强。根据经验：

KS 范围	区分能力	建议事项
KS<0.3	区分能力较差	不建议使用
0.3<KS<0.5	区分能力一般	可正常使用，需加强持续监控
0.5<KS<0.7	区分能力强	正常使用
KS>0.7	区分能力极强	谨慎使用，需分析有无指标泄露风险

模型的排序性：LIFT 曲线

Lift 曲线它衡量的是，与不利用模型相比，模型的预测能力“变好”了多少，lift(提升指数)越大，模型的运行效果越好。实质上它强调的是投入与产出比，大部分模型在最终应用时只是利用到排序性，其主要是为了观察曲线的斜率。

模型的稳定性：群体稳定性指标（PSI）

PSI 反映了验证样本在各分数段的分布与建模样本分布的稳定性。在建模中，我们常用来筛选特征变量、评估模型稳定性。在建模时通常以训练样本作为预期分布，而验证样本（样本外 OutofSample，OOS）、测试样本（跨时间样本 OutofTime，OOT）通常作为实际分布。PSI 数值越小，两个分布之间的差异就越小，代表越稳定。根据经验：

PSI 范围	稳定性	建议事项
PSI<=0.1	稳定性良好	没有变化或很少变化
0.1<PS<=0.25	稳定性一般	有变化，需要监控后续变化
PSI>0.25	稳定性差	发生大变化，需要进行特征项分析

3.3.3 模型验证

模型验证范围包括模型及其支持体系的验证。对模型进行验证时，应当重点关注模型数据、模型方法、重要假设和参数、模型开发过程和模型结果应用等方面的审查。验证范围应当包括但不限于模型使用政策和流程、数据、信息系统、模型应用和用户反馈信息，以及相关文档记录等方面。

验证工作应当关注模型结果在应用部门或团队的表现和使用情况，验证结果和其他反馈信息应当及时提供给高级管理层和模型应用部门或团队，以推动模型的持续完善和深入应用。

验证工作是一个持续、循环进行的过程。本行应当对模型进行投入使用前全面验证（以下简称投产前验证）、持续监控和投入使用后全面验证（以下简称投产后验证），明确验证范围和内容，选择合适方法，合理安排各项工作的顺序与频率，确保验证工作按计划进行。

3.3.4 需求投产

需求投产前需要进行产前验证。

验证包括对模型开发工作的验证，重点验证模型方法的合理性、关键定义的合规性及可操作性、数据的真实完整性和风险量化的有效性等。验证还应涵盖模型和相关政策、流程、数据、信息系统和文档记录等方面，确保对模型和支持体系的稳健性、可靠性和合规性作出全面评估。验证内容包括不限于：

关键定义验证。包括对违约定义、损失定义和主标尺定义的合规性设计验证及合规性执行验证。

建模样本数据验证。包括对建模岩本数据的完整性、全面性、一致性、准确性和合规性进行检验。

模型验证。包括对各模型细分依据的核查、对模型方法论的验证、对模型参数和假设的验证、对模型建模过程合理性的核查、对模型区分能力的验证及对模型结果的合理性分析等。

支持体系验证。包括治理结构、政策和流程、数据管理、IT 系统、评级应用、文档管理等。

3.3.5 模型定期持续监控

模型定期持续监控是指持续监测内部模型体系的运行状况，及时了解模型的表现，对所使用模型进行必要的局部修正、提升和完善等工作。模型的持续监控包括但不限于：

模型系统运行情况；

模型结果和相关政策执行情况；

数据质量、存储和管理、维护情况；

模型特征的稳定性和预测性；

模型的风险区分能力(AUC、K-S)、风险校准能力和稳定性(PSI)等相关衡量指标表现；

模型应用的外部条件（包括业务变动、客群变动、市场环境）是否发生重大变化。

3.3.6 模型优化

3.3.6.1 拓展新的数据维度

模型只是对数据的无限接近、数据才是最终决定模型效能的因素。因此模型优化的一个重要的方面就是尽量拓展可用数据源，增加模型训练的数据维度。

3.3.6.2 特征筛选

持续跟踪特征的稳定性（PSI）和预测能力（IV）、以及特征之间的相关性等。筛选、保留稳定性高、预测能力衰减慢且相关性不高的特征，这样有利于保持模型预测的鲁棒性。

3.3.6.3 模型算法的优化

在坏样本量不足的情况下，项目前期有可能只能采用半监督的模型进行建模、但随着时间的推移，坏样本的数量逐步增加，这种情况下就可以用监督模型替换掉半监督模型，模型的预测能力将得到大幅

提升。

3.3.6.4 超参数优化

如果时间和算力允许，风控模型的参数直接使用暴力点的网络搜索来选择全局最优的超参也是很好的。否则的话，就使用以下的超参数优化方法：

基于贝叶斯优化的超参数优化 BayesianOptimization

基于进化算法的超参数优化 EvolutionaryAlgorithms

基于随机搜索的超参数优化 RandomSearch

基于元学习的超参数优化 MetaLearning

基于迁移学习的超参数优化 TransferLearning

超参的优化可以对模型有一定的优化能力，但是这个过程中，需要注意模型是否会被过拟合，特别是在样本数量不多的情况下。