# 研究计划

##项目主题：基于主体关联性的信用风险预警研究

##项目介绍：突破单一主体局限，融合多源关联数据构建动态图谱，量化风险传导路径，开发基于复杂网络与机器学习的主动式、前瞻性信用风险预警体系。

## 一、七月份计划（7.1-7.31）

7.1-7.3 确定使用Zotero进行文献收集，并且使用微软共享word插入Zotero，对文献进行简要概括。

7.4-7.25 收集文献（大致30-40篇，每日1-2篇/人）

考虑使用关键词：“信用”、“信用风险”、“信用风险预警”、“风险预警“、“企业关联”、“供应链（关联）”、“产业链（关联）”、“股权（关联）”、“企业网络”、“金融关联”、 ‘Subject Association’ 、“Credit Risk Early Warning”、“Complex Network”、“Supply Chain Finance Risk”、“Financial Contagion”、“Panel Data Models”、等，可考虑关键词的组配

考虑使用数据库：CNKI、万方、维普、CSSCI；谷歌学术、Web of Science

7.25-7.30 文献综述撰写（具体格式？内容要求？字数要求？）

## 二、八月份计划（8.1-8.31）

8.1-8.20 根据论文中提到的数据库/集或者开放的数据库/集收集数据（需要创建的数据集种类、数量？收集的数据量？）；数据处理

8.20-8.31 实验，具体内容待定

# 论文

##论文题目使用标题3（小四字号，黑色），论文介绍使用正文（五号，黑色），“ctrl+F”可在左侧看到目录。

## 2.1 中文文献

### 1、基于机器学习的信用债流动性风险预警研究(张宗新等, 2024)

论文来源：国际金融研究

研究结论：第一，**包含一层隐含层的神经网络**对信用债流动性风险的预警能力最强，在不同类型债券和不同外部环境的样本中预警能力的稳定性较强，能准确预警市场层面的流动性枯竭事件；第二，**券龄的重要性最高**，新发行的债券会通过引发噪声交易的方式形成流动性风险，但随着券龄增加，流动性风险减少的程度是递减的；第三，流动性风险是由多类风险因子协同运动生成的，其中，**经济状况变动、货币政策改变或跨市场冲击与券龄的非线性联动**对于驱动流动性风险最为重要。

数据来源：国泰安CSMAR和同花顺iFind数据库

研究方法：**简单线性回归（OLS）**模型作为比较基准。构建**弹性网（Elastic Net，ENET）模型**等惩罚线性回归模型。基于降低风险因子维度，构建**主成分分析（Principal Components Analysis，PCA）模型**和**偏最小二乘（Partial Least Squares，PLS）模型**等特征降维模型。由于OLS、ENET、PCA和PLS作为线性模型，未考虑风险因子的非线性联动，本文加入**随机森林（RF）模型**和**极端梯度提升回归树（XGB）模型**等回归树模型，以及**神经网络（NN）模型**。采用**置换重要性**识别风险因子重要性。

### 2、基于复杂网络的信用风险传染模型研究(陈庭强等, 2014)

论文来源：中国管理科学

数据来源：无（没有提到数据从哪里来）

研究结论：（1）在不完全市场条件下，只要有信用事件，总会引发信用风险传染。当其他条件不变的情况下，个体对风险较为敏感时，个体的风险态度会加速信用风险的影响范围。换句话说，风险规避型个体会加速信用风险的传染和影响。在既定条件下，个体的风险抵御能力和金融市场的监管强度会减小信用风险的传染及其影响。而社会网络中个体之间的关联程度会增加信用风险的传染。同时，对于具有较大影响力的信用事件的发生，会增加金融市场信用风险持有者的恐惧心理和非理性行为，提高信用风险的速度和影响范围，严重时导致“蝴蝶效应”。（2）在其他条件不变的情况下，社会网络越密集，信用风险传染的影响范围越广。（3）社会网络中个体的异质性越高，越不利于信用风险在社会网络中传染和扩散（4）信用风险传染的规模是**个体之间的关联程度、风险态度和信用事件的影响力**的单调**递增**凹函数，是**金融市场的监管强度和个体的风险抵御能力**的单调**递减**凸函数。

研究方法：构建社会网络（WS\BA\指数网络）+仿真实验

### 3、基于企业关联关系的信用风险分析新思路(刘新海, 2014)

*（本篇文章发表时间距今比较久，主要是讨论了研究信用风险时考虑主体关联的必要性）*

论文来源：征信

研究结论：对征信系统中的关联关系进行数据挖掘不仅可以描述微观角度每一个信用主体的风险状况，还可以描述中观角度行业、地区和金融机构的信用风险模式，并且还能从宏观角度对全局的信贷结构、金融问题提供一些量化指标作参考。此外关联分析的可视化以及动态分析可以作为担保圈为代表的系统性信用风险预防和释缓的有力工具。

研究方法：**复杂网络**：复杂网络可以描述出企业信用主体的复杂关联关系，企业信用主体为网络中节点，企业之间关联关系为网络中的连接，关联关系的强弱体现为网络中连接的强弱，以担保圈为例构建了复杂网络。

## 2.2 英文文献

### 1.A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method(Emmanuel等, 2024)

*（该篇文章更偏使用机器学习的方法对信用进行预测的模型间的性能比较，与主体关联性无关）*

论文来源：Journal of Big Data

研究结论：堆叠方法的结果优于通过单个估计器和其他现有算法获得的结果。

数据来源：加利福尼亚大学尔湾分校（UCI）机器学习存储库

研究方法：使用了堆叠模型，堆叠模型包括以下基本估计器：随机森林（RF），梯度提升（GB）和极端梯度提升（XGB）。此外，堆叠体系结构中的估计器依次连接以提取最佳性能。本研究中使用的基于过滤器的FS方法基于信息增益（IG）理论。使用准确度，F1得分和曲线下的面积（AUC）评估所提出的算法。此外，将堆叠的算法与以下方法进行了比较：人工神经网络（ANN），决策树（DT）和K-Nearest邻居（KNN）。

### 2.Construction and Evaluation of Credit Risk Early Warning Indicator System of Internet Financial Enterprises Based On AI and Knowledge Graph Theory(Peng, 2024)

（使用CNN做二分类）本论文以互联网金融企业为研究对象，构建了一个融合**财务与非财务指标**的信用风险预警体系，利用**主成分分析**与**灰色关联法**优化指标集（确定指标），并引入卷积神经网络（CNN）模型进行风险预测，设计双子网络分别处理动态与静态数据。为增强模型性能，论文采用ADASYN方法平衡样本，并通过多组实验优化超参数。结果显示，所构建模型在准确率、F1值和AUC等指标上优于传统方法，具有较强的预警能力与应用价值。同时，论文强调“AI+知识图谱”在金融风险管理中的重要支撑作用（没有细说），指出将知识图谱用于构建企业主体关联和语义推理，有助于提升风险识别的全面性与解释性。（数据来源不明）垃圾论文

### 3. DebtStreamness: An Ecological Approach to Credit Flows in Inter-Firm Networks(Rodríguez-Martínez等, 2025)

（图类比生态圈 提出DebtStreamness来量化企业在信贷链条中的相对位置）240家企业 乌干达 这个指标意味着深度 离银行的远近 **网络重建** 基于最大熵 与稀疏性假设 还做了一些分布的分析 展望：冲击分析 将信贷网络与供应链网络做层级分析 （指标有意思 也是垃圾论文）

### 4.Credit risk prediction for small and medium enterprises utilizing adjacent enterprise data and a relational graph attention network(Wang J等, 2024)

**高度相似！！！** 可以查看一下综述 综述做的不错

关联的类型有：涵盖关系 交易 供应链 SCF的发展说明空间很大 新的方法（RGAT）考虑竞争对手 GNN—RGAT（有详细的图设计）

数据来源：中国的某上市银行 已经发邮件询问了

### 5. Research on Personal Credit Risk Assessment Model Based on Instance-Based Transfer Learning(Wang M等, 2021)

*（主要是介绍了一个新的模型方法，数据量不够的话可以看看）*

论文来源：Intelligence Science III

数据来源：Prosper Online P2P贷款网站和银行

研究结论：Tradaboost（XGBoost）模型充分利用源域信息，以成功完成目标域信息的培训，并解决了由于缺乏样本和大量缺失值而无法培训数据集的困境。本文实现了个人信用风险领域样本的转移，该样本对金融领域具有一定的参考价值。基于本文提出的基于传统样本转移的模型增加了XGBoost集合学习算法，该算法提高了模型的准确性，增强了模型的性能，并且具有良好的概括能力。

研究方法：**基于实例的转移模型**，可以用于数据量不足的时候模型的训练问题。在每次迭代中，根据基学习器在目标域样本上的分类误差来调整权重。如果目标域样本被错误分类，则增加其权重；如果源域样本被错误分类，则减少其权重。

##转移学习具体原理：

实例权重调整：基于实例的转移学习模型会对源域中的每个实例赋予一个权重，这个权重反映了该实例对目标域任务的重要性。权重的调整通常基于实例在目标域上的表现，如果某个源域实例与目标域数据相似度高，且对目标域任务的预测有帮助，则其权重会被增加；反之，如果某个源域实例与目标域数据差异较大，或者对目标域任务的预测产生干扰，则其权重会被减少。

知识迁移：通过调整后的权重，源域中的实例被重新加权组合，形成一个新的数据集，这个数据集在特征空间上与目标域数据更加接近，从而使得在源域上训练得到的模型能够更好地适应目标域的任务。

### 6. A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network(Liu等, 2022)

论文来源：Expert Systems with Applications

研究方法：本研究提出了一个**两阶段的混合信用风险预测模型**来***解决特征工程***。（1）首先，基于**树的模型**擅长捕获预测因子和目标之间的非线性关系，该预测因子与目标之间具有良好的特征线性化能力。因此，我们采用众所周知的基于**树的模型XGBoost**来进行特征线性化过程的第一阶段。（2）然后，借助2020年Kong和Yu提出的Lovenet模型，为了完全挖掘特征之间的依赖关系，我们通过调查具有XGBOOST的特征之间的树拆分过程来构建特征图。节点代表一个功能，边缘表示特征之间的定向关系。（3）最后，我们将特征图嵌入到深度神经网络中，以构建信用风险预测模型。

研究结论：

# 数据类型与数据源

此处的数据问题考虑 **什么是主体关联性** 什么是 **可以量化的信用风险**

## 3.1什么是可以量化的信用风险

### 1. 硬性违约事件 (Hard Default Events) - 最直接的标签

这是最明确、最无可争议的信用风险事件。

* **定义：**
  + **债券违约 (Bond Default):** 上市公司或发债企业未能按时支付债券的本金或利息。这是公开市场事件，数据最透明。
  + **贷款违约 (Loan Default):** 企业未能按时偿还银行或其他金融机构的贷款。此数据通常属于银行内部，较难获取，但对于银行风控模型至关重要。
  + **破产清算 (Bankruptcy / Liquidation):** 企业通过法律程序宣布破产。这是最严重的事件。
* **数据来源：**
  + **Wind (万得数据库):** 有非常全面的中国债券违约数据库。
  + **CSMAR (国泰安数据库):** 同样提供债券违约信息。（国内多数论文书记来源）
  + **公开信息：** 上海/深圳证券交易所公告、全国企业破产重整案件信息网。
* **特点：** 定义清晰，是风险的最终体现。但缺点是**样本稀疏**（违约毕竟是小概率事件），可能会造成模型训练时严重的类别不平衡问题。**（我们需要考虑多种违约形式）**

### 2. 财务困境代理变量 (Financial Distress Proxies) - 更常用的标签

由于硬性违约事件较少，在学术研究中，更常用一些“代理变量”来标识企业陷入了严重的财务困境。这些事件通常是硬性违约的前兆。

* **定义：**
  + **被ST或\*ST处理 (Special Treatment):** 这是中国A股市场独有的制度。当上市公司出现连续亏损或其他财务异常时，其股票名称前会被加上“ST”或“\*ST”标记，作为退市风险警示。这是一个非常强烈的负面信号。
  + **主体信用评级下调 (Credit Rating Downgrade):** 由联合资信、中诚信等信用评级机构下调企业的主体信用评级（例如从AAA下调至AA+）。
  + **股价崩盘风险 (Stock Price Crash Risk):** 指企业股价在短期内发生极端的大幅下跌。通常用未来一段时间内周收益率的负偏态系数（NCSKEW）或收益率的上下波动比（DUVOL）来度量。
* **数据来源：**
  + \**ST/ST、评级：* Wind、CSMAR数据库都有直接的字段。
  + **股价崩盘风险：** 需要利用Wind或CSMAR的股价交易数据自行计算。
* **特点：** 样本量比硬性违约多，更容易建模。它们被广泛证实与未来的实际违约高度相关，是极佳的预测标签。**对于上市公司研究，使用“未来一年内是否会被ST”是一个非常经典且有效的做法。**

## 3.2 可用主体关联性有什么

1.股权与控制关系：直接持股 间接持股 共同股东

来源：Wind、CSMAR、企查查、天眼查

2.担保关系（这是我们最可能用到的吧）：A为B直接担保 风险传染性最强

来源：Wind、CSMAR（上市公司关联交易库中非常详尽）。

3.经营往来关系 供应链

来源：**市公司年报（会披露前五大客户/供应商）**、CSMAR关联交易数据库。供应链数据较难精确获取，是研究的难点。（可能有关系 但是实现难度大 解决方法：考虑到研究主体缩小范围）

4．人员关联 董事会兼任

来源：Wind、CSMAR公司治理数据库

5.市场关联 （有趣）股民用脚投票

需要自己通过Wind、CSMAR的日/周交易数据，计算相关性矩阵来构建。

# 技术实现思路