# Enhanced Credit Risk Assessment for Chinese SMEs: A Dynamic Knowledge Graph Approach with Markov Random Field

## Abstract

Within the intricate domain of fintech, credit risk assessment for small and medium - sized enterprises (SMEs) assumes utmost significance for financial institutions. Nevertheless, traditional methodologies confront formidable hurdles in capturing the complex relationships and dynamics involved. Prevailing approaches often prove inadequate in effectively leveraging the intricate interconnections among the numerous factors that influence the creditworthiness of SMEs. This shortcoming has consequently engendered a substantial lacuna in the extant body of research. The primary objective of this study is to formulate an advanced credit risk assessment framework customized for Chinese SMEs. This is accomplished by integrating the Dynamic Knowledge Graph Model (DKGM) with Markov Random Fields (MRF). Moreover, secondary objectives entail a comprehensive exploration of the impact exerted by diverse knowledge graph structures and parameter configurations on the model's performance. The research methodology adopted herein involves constructing a dynamic knowledge graph to represent the continuously evolving relationships among SMEs. Markov Random Fields are then utilized to model the conditional dependencies inherent in this graph. Subsequently, machine learning techniques are employed to predict credit risk based on the learned representations. The key findings of this research disclose that the proposed framework represents a substantial enhancement in predictive accuracy compared to traditional methods. Specifically, certain knowledge graph structures and parameter settings have been demonstrated to be more effective in capturing the dynamics of credit risk. Furthermore, the model furnishes valuable insights into the key determinants of SME creditworthiness. This research makes contributions in multiple dimensions. First and foremost, it devises a novel and efficacious approach for SME credit risk assessment. Secondly, it offers in - depth insights into the application of knowledge graphs and MRFs within the realm of finance. Finally, it deepens our comprehension of the underlying drivers of SME credit risk in China. Overall, the proposed framework has the potential to optimize the decision - making processes of financial institutions and contribute to the stability of the SME lending market in China.

## 1. Introduction

In the contemporary, dynamic market economies, small and medium-sized enterprises (SMEs) serve as the cornerstone of economic progress. Notably, SMEs account for 99% of all businesses in China. Their agility and relatively small scale endow them with unique advantages, making substantial contributions to employment generation, innovation promotion, and the maintenance of overall social stability. However, despite their indispensable role, SMEs in China are currently confronted with severe sustainability challenges. As reported by Fortune magazine [1], the median lifespan of SMEs in China is merely 3 - 5 years. A critical factor impeding their long-term survival is the underdeveloped risk management infrastructure. This situation is further exacerbated by the intense post-crisis market competition and the ever-present developmental risks. SMEs are faced with a multitude of risks, influenced by economic fluctuations, political changes, natural disasters, legal constraints, and rapid technological advancements [2]. These risks can rapidly overwhelm the limited coping mechanisms of smaller businesses.

Banks and financial institutions play a pivotal role in assisting SMEs to navigate through these arduous periods [3]. Their role not only encompasses the fulfillment of social responsibilities but also the safeguarding of the financial soundness and operational sustainability of these enterprises. Prior to extending credit, a comprehensive risk assessment is of utmost significance. The creditworthiness of SMEs is intricately associated with the personal finances and living standards of their owners, the scarcity of substantial collateral, and frequently, suboptimal financial management practices [4]. These complexities pose significant obstacles to traditional creditworthiness assessment metrics, primarily due to information asymmetry and restricted access to reliable data sources.

This research inaugurates an innovative approach for SME risk assessment, capitalizing on the capabilities of the Integrated Markov Random Field - Dynamic Knowledge Graph Model (IMRF - DKGM). This model skillfully combines the probabilistic insights provided by Markov Random Fields (MRF) with the comprehensive data integration capabilities of the Dynamic Knowledge Graph Model (DKGM). By doing so, it effectively addresses the prevalent issues of information asymmetry and dynamic risk factors in SME credit evaluations. The IMRF-DKGM is specifically designed to surmount the paucity of financial reporting in SMEs compared to their larger corporate counterparts. Central to this approach is the construction of a dynamic financial knowledge graph that analyzes external behavioral and environmental data. It proficiently processes a wide range of structured and unstructured data related to investment and managerial activities, enabling the identification of crucial "risk genes" within the complex financial network of the business ecosystem. Leveraging cutting-edge data fusion techniques and financial technology (FinTech), this research presents a novel method for assigning "good" or "bad" ratings to SMEs and explores strategies for predicting and mitigating enterprise risks. This groundbreaking methodology holds the potential to provide a comprehensive risk assessment framework, enhancing the resilience and growth potential of SMEs in a volatile economic environment.

### 1.1. Risk Management Systems of SMEs in China

The risk management domain for Small and Medium Enterprises (SMEs) in China represents a complex and multifaceted scenario, intricately linked with various aspects of economic policy, market dynamics, and the unique challenges these enterprises inherently face [5]. As the main drivers of growth and employment in the national economy, SMEs hold great significance. Nevertheless, due to their limited scale and resources, they are disproportionately vulnerable to a wide range of risks. This in - depth analysis aims to comprehensively overview the risk management systems within China's SME sector, highlighting the inherent challenges and the continuous evolution of strategies to effectively mitigate these risks.

China's remarkable economic growth has been substantially propelled by the dynamic SME sector. However, the risk management systems in these enterprises often lag behind their rapid development. The lack of a robust and comprehensive risk management framework can be ascribed to multiple factors, including a general lack of awareness and the operational constraints within a market economy that experiences frequent and unpredictable regulatory and economic changes. In response to this acknowledged gap, the Chinese government has taken proactive steps, implementing policies to encourage and guide SMEs to adopt improved risk management practices.

SMEs in China are confronted with numerous risks, such as market volatility, credit risks, operational risks, and policy risks. Market volatility is mainly driven by rapid and unpredictable changes in consumer demand, along with intense market competition [6]. Credit risks are particularly prominent in this sector due to the information asymmetry between SMEs and financial institutions, making it difficult for SMEs to obtain necessary financing. Operational risks arise from the ongoing challenge of maintaining efficiency and quality while managing limited resources. Additionally, policy risks stem from the dynamic and ever-changing regulatory environment, which can significantly impact SME business operations.

Historically, Chinese SMEs' risk management practices have been predominantly reactive rather than proactive. Many SMEs struggle to establish internal mechanisms for effectively identifying and assessing risks. The lack of sophisticated financial management systems, combined with a reliance on informal business networks, has increased their risk exposure. Moreover, limited capital access and high borrowing costs further restrict SMEs' ability to invest in comprehensive risk mitigation strategies.

In response to these formidable challenges, there has been a discernible and gradual shift towards more proactive risk management systems among Chinese SMEs. This evolutionary trend is characterized by the integration of advanced financial technologies, the development of more refined internal control mechanisms, and an increased focus on regulatory compliance. The use of big data and artificial intelligence has become more common in risk assessment and decision-making. Both the government and financial institutions play crucial roles in strengthening SMEs' risk management systems. Government-led initiatives, like the establishment of credit guarantee systems, tax incentives, and information-sharing platforms, have been instrumental in enhancing SMEs' risk management capabilities. Meanwhile, financial institutions are increasingly offering customized financial products and services to meet SMEs' specific needs, providing them with more effective tools for managing financial risks.

The future of risk management in China's SME sector holds great promise as it embraces innovation and promotes collaboration. The integration of technology into risk management practices is expected to continue, with a particular focus on enhancing data analytics capabilities and using machine learning algorithms for risk prediction. The collaboration among SMEs, government bodies, and financial institutions will be crucial for developing a comprehensive risk management ecosystem that can withstand market uncertainties. The risk management system of Chinese SMEs is at a critical point, where the need for robust and adaptable strategies is more urgent than ever. The path to a more resilient and proactive risk management framework is full of challenges but also abounds with opportunities for innovation and growth. With appropriate support and strategies, Chinese SMEs can successfully navigate market complexities, become stronger, more sustainable, and better prepared to face the future.

### 1.2. Research Problem and Questions

In the domain of financial risk assessment, applying traditional models to Small and Medium Enterprises (SMEs) has frequently proven inadequate. This is chiefly because of the idiosyncratic features of SMEs and the paucity of comprehensive financial data. There is an acute necessity for a dynamic and adaptable credit risk assessment model. Such a model should be able to precisely evaluate the creditworthiness of SMEs within economic landscape. It must account for the unique risks and challenges that SMEs face, which traditional methods have failed to fully accommodate.

This research focuses on integrating big data analytics with a dynamic knowledge graph model (DKGM). The aim is to extract and analyze a comprehensive set of structured and unstructured data relevant to SMEs. The overarching objective is to unearth the "enterprise risk genes" hidden within the complex financial network structures. This requires a meticulous examination of the macroeconomic environment, industry - specific characteristics, and the distinctive operational behaviors of SMEs. By doing so, the model endeavors to offer a more refined and accurate assessment of the credit risks associated with SMEs.

A crucial element of the proposed model is the use of the Gibbs sampling method within the Markov Chain Monte Carlo (MCMC) framework. This sophisticated statistical technique can extract risk characteristics highly correlated with SME development, even in the presence of sample errors. The MCMC framework, with its proficient handling of probabilistic models, allows for a more in - depth understanding of the intricate relationships and dependencies within financial data. This, in turn, facilitates a more precise identification of risk factors essential for the credit assessment process.

The research intends to make a substantial contribution to the field of SME credit risk assessment by developing a novel enterprise risk assessment index. This index is designed to integrate the primary factors influencing SME development within the financial knowledge graph framework. The integration of these factors is anticipated to enhance the accuracy of SME risk identification models, as evidenced by an improved Area Under the Curve (AUC) score in the model's performance evaluation. The research specifically aims to create a sophisticated credit risk assessment tool tailored to the unique requirements of SMEs. This tool should function effectively even in volatile economic conditions. By providing a more profound understanding of SME risk profiles, the model aims to furnish financial institutions with a reliable means of evaluating credit risk. This is especially vital considering the inherent challenges and complexities associated with SME credit assessments.

The proposed model, which incorporates the Markov Random Field into the dynamic knowledge graph framework, represents a significant progression in the field of SME credit risk assessment. It is expected to offer a more accurate, robust, and adaptable approach to evaluating SME creditworthiness. Thus, it overcomes the limitations of traditional models and provides a more effective solution for navigating the complexities of the financial landscape.

Based on the research problems illustrated above, we formulate several research questions listed in the following:

1. SMEs lack accounting, financial and other data, and it is difficult for financial institutions to evaluate the corresponding credit risk of SMEs loans. What are the approaches in solving problem of "information asymmetry problem”?

2. The development of SMEs will be affected by their own factors and external factors. What are the impacts of these risk factors on the development of enterprises?

3. There is a lack of an enterprise credit risk prediction system which combines the important risks factor in its prediction. How to utilize the important dynamic risk factors affecting enterprise development to establish a prediction system to predict enterprise credit risk?

By addressing these research questions, the study aims to develop a more effective financial risk prevention system for SMEs in China that can withstand the challenges of economic cycles and contribute to the stability and growth of the national economy.

### 1.3. Research Aim and Objectives

This research endeavors to construct an elaborate enterprise credit risk prediction model, precisely customized to capture the dynamic and multifaceted risk factors inherent in Small and Medium Enterprises (SMEs). The overarching objective is to substantially enhance the accuracy of credit risk predictions, thereby providing financial institutions with a more reliable and robust instrument. The proposed model is founded upon a comprehensive risk assessment framework that capitalizes on the concept of a dynamic financial knowledge map, designed to process and analyze both structured and unstructured data related to SMEs.

At the core of this research, the dynamic financial knowledge map functions as an advanced risk assessment apparatus. It enables the processing of a “related party” information set, encompassing personnel, organizational, and industry-related entity information. This map visually represents the intricate network of relationships among these entities, offering a more intuitive and accessible means of comprehending and retrieving relevant information. Integrating Markov Random Fields (MRF) into this framework enables the modeling of complex dependencies and interactions within the data. This furnishes a probabilistic basis, enhancing the model's capacity to capture the subtleties of risk within the SME sector.

A fundamental aspect of this research is the utilization of the Gibbs sampling method within the Markov Chain Monte Carlo (MCMC) framework. This method is particularly adept at extracting risk characteristics closely associated with SME development. Operating on the principles of big data association feature factor extraction, it employs sample error tolerance as a benchmark for inference. The Gibbs sampling method enables a detailed analysis of potential risks embedded within complex relational structures. It plays a pivotal role in classifying SMEs as “good” or “bad” credit risks. This innovative evaluation approach offers a dynamic assessment of SME development by examining the structures and behaviors within enterprise networks.

The research systematically explores the extraction of association features from network structures of SMEs across diverse sectors, such as manufacturing, retail, and financial services. Empirical findings demonstrate that SME development is significantly influenced by three key factors: macroeconomic conditions, risk transmission from affiliated enterprises, and the evolutionary development stages of SMEs. To address these influential factors, the research proposes integrating key dynamic risk factors into the financial knowledge map framework, establishing them as new indices for enterprise risk assessment.

This integration is designed to fortify the identification, prevention, and control of SME - related risks, thereby reducing the likelihood of enterprise failures. The proposed method, incorporating MRF to model the probabilistic relationships within the data, is anticipated to yield substantial improvements in key performance indicators. These include the Kolmogorov - Smirnov (KS) statistic, which measures the maximum difference between the cumulative distribution functions of defaulters and non - defaulters, and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, a metric evaluating the model's ability to distinguish between SMEs with high and low credit risks.

In comparison to traditional financial knowledge mapping models, integrating MRF and using Gibbs sampling within the MCMC framework is expected to enhance the model's predictive capabilities. This advancement promises a more nuanced and accurate assessment of credit risks within the SME sector, providing financial institutions with a potent tool to navigate the complexities of credit risk assessment in a post-pandemic economy.

Currently, the research objectives are listed as follows:

1. Designing and developing an Integrated Markov Random Field-Dynamic Knowledge Graph Model (IMRF-DKGM) to address information asymmetry in SME credit risk assessment.

2. Developing an enterprise credit risk prediction model using the Gibbs sampling algorithm and Markov Chain Monte Carlo framework to identify dynamic factors influencing SME development.

3. Evaluating the performance of the proposed model using specific metrics like Kolmogorov-Smirnov statistic and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

### 1.4. Significance of the Study

This study holds substantial significance in its pioneering approach of integrating Markov Random Fields (MRF) into the dynamic knowledge graph model, which is meticulously customized for the financial industry. The integration aims to construct a comprehensive knowledge map. This map not only amalgamates domain knowledge from diverse data sources but also effectively addresses the inherent uncertainties within the financial market. The study innovatively grounds itself in the "non-linear expectation" theory, deploys financial technology "holographic profiling" as an innovative approach, and adopts the "dynamic stochastic general equilibrium" model as an overarching innovative framework. The ultimate objective is to establish a theoretical system for financial risk measurement and asset pricing, seamlessly integrating multi-source heterogeneous financial big data from both macro- and micro-economic perspectives.

A crucial innovation in this research lies in the integration of machine learning algorithms into a big data platform. This enables the processing of a broader spectrum of data types. The financial knowledge graph developed in this study processes both structured and unstructured data presented in the form of "related party" information sets. This significantly facilitates the analysis of entity information regarding the personnel, organizations, and industries associated with Small and Medium Enterprises (SMEs), as well as the intricate interconnections among these entities. The goal is to achieve a more intuitive comprehension and efficient retrieval of relevant entities and relationship information, extract the so-called "enterprise risk genes" from the complex financial network, identify specific latent risks, and establish a robust model for classifying SMEs as having "good" or "bad" creditworthiness.

Integrating MRF into the dynamic knowledge graph model is a vital aspect of this research. MRF provides a probabilistic framework that can capture the complex dependencies and interactions among entities within the financial network. This allows for a more refined understanding of the risks associated with SMEs, as the model can account for the uncertainties and stochasticity inherent in financial data. MRF is particularly valuable in scenarios where the relationships between entities are not independent, and a holistic view of the entire network is essential for accurate risk assessment.

The Gibbs sampling method, implemented within the Markov Chain Monte Carlo (MCMC) framework, represents a crucial tool in this research. It effectively extracts risk characteristics closely related to SME development, based on the principles of big data association feature factor extraction, using sample error tolerance as a benchmark for inference. This method enables a detailed analysis of specific latent risks embedded within complex relationships and facilitates the categorization of SMEs into "good" or "bad" credit risk categories. The dynamic nature of the MRF endows the model with the ability to adapt to new data and changing market conditions, thus providing a more accurate and timely assessment of credit risks.

The study proposes a risk assessment framework for SMEs, which is firmly based on the dynamic financial knowledge graph. By constructing this graph, both structured and unstructured data can be processed, enabling the analysis of entity information related to SME personnel, organizations, and industries, and the understanding of the relationships among these entities. The utilization of MRF and Gibbs sampling within the MCMC framework aids in extracting risk characteristics that are highly correlated with SME development. This enables the identification of specific latent risks within complex relationships and facilitates the classification of SMEs into "good" or "bad" credit risk groups.

The theoretical novelty of this research lies in the innovative integration of a Dynamic Knowledge Graph Model (DKGM) with Markov Random Fields (MRF) for SME credit risk assessment. This combination represents a significant advancement over traditional methods by simultaneously capturing both the structural relationships among diverse risk factors and their temporal evolution. Unlike conventional approaches that often treat risk factors in isolation or rely on static relationships, our framework explicitly models the interdependencies between macroeconomic indicators, industry dynamics, and enterprise-specific characteristics through the DKGM's evolving graph structure. The incorporation of MRF further enhances this by providing a probabilistic framework that quantifies the strength and nature of these relationships, allowing for more accurate propagation of risk signals across the financial ecosystem. This holistic approach not only improves predictive accuracy but also maintains interpretability through the knowledge graph representation, addressing the critical need for transparent risk assessment models in financial decision-making.

## 2. Literature Review

In this section, we discuss existing studies related to dynamic financial knowledge graph construction and Markov random fields, which includes machine learning based financial risk systems, data mining for risks prevention, and the construction of financial knowledge graphs.

### 2.1. Machine Learning-based Credit Risk Systems

In the realm of finance, the application of machine learning (ML) in credit risk assessment has emerged as a focal point of academic and industry research in recent years. The potential of ML to enhance the precision and efficiency of credit risk prediction has spurred a burgeoning body of literature. This section undertakes a comprehensive and systematic review of the existing scholarship on ML - based credit risk assessment systems, meticulously dissecting the diverse techniques, models, and methodological approaches that have been explored.

In the incipient phase of ML's integration into credit risk assessment, traditional algorithms served as the cornerstone of model development. Decision trees, logistic regression, and support vector machines (SVM) were among the most prevalently utilized techniques. In [7], for instance, formulated a credit risk assessment modeling framework that amalgamated fuzzy integral theory with SVM. This approach effectively bridged the gap between expert - derived knowledge and data - driven analytics, enabling a more comprehensive assessment of credit risk. Concurrently in [8], they harnessed the power of gradient boosting decision trees for credit risk evaluation. Their research demonstrated the model's proficiency in handling complex and non - linear data patterns, thereby significantly improving the accuracy of credit risk predictions. These early - stage studies not only established the fundamental principles but also paved the way for subsequent advancements in the field.

As the research landscape evolved, ensemble learning and hybrid models emerged as paradigmatic shifts in the quest for more robust credit risk assessment systems. In [9], they developed a credit risk assessment model tailored for technology - based small and medium - sized enterprises (SMEs) using heterogeneous ensemble learning. By combining multiple base learners, this approach enhanced the model's robustness and predictive accuracy, effectively mitigating the impact of data heterogeneity and noise. In a similar vein [10], they proposed a hybrid XGBoost - MLP model for credit risk assessment in the context of digital supply chain finance. This innovative model capitalized on the complementary strengths of tree - based and neural network - based architectures, yielding superior performance in capturing complex relationships within the data. These studies underscored the transformative potential of ensemble and hybrid models in elevating the overall efficacy of credit risk assessment systems.

The advent of deep learning techniques has ushered in a new era of sophistication in credit risk assessment. Convolutional neural networks (CNN), recurrent neural networks (RNN), and deep reinforcement learning (DRL) have been increasingly applied to this domain. In [11], they introduced an interpretable credit risk assessment model grounded in CNN. This model not only delivered accurate predictions but also provided valuable insights into the underlying decision - making mechanisms, thereby enhancing the transparency and interpretability of the model. In [12], they developed a dynamic credit risk assessment model based on DRL, which demonstrated the ability to adapt in real - time to the ever - changing market conditions and optimize credit risk management strategies. These deep - learning - based approaches have exhibited remarkable potential in handling large - scale, high - dimensional datasets and capturing intricate non - linear relationships, thus revolutionizing the field of credit risk assessment.

Feature engineering and selection are integral components of the success of ML - based credit risk assessment systems. In [13], they embarked on an in - depth exploration of various feature engineering techniques aimed at enhancing credit scoring models. Their research emphasized the critical importance of selecting relevant and informative features, as these can significantly improve the model's performance and generalization ability. Zhao et al. [8,14] proposed a multiple imputation method for missing credit risk assessment data, leveraging the power of generative adversarial networks (GANs). This innovative approach effectively addressed the ubiquitous issue of missing data, thereby enhancing the robustness of feature selection processes. These studies underscored the indispensable role of feature engineering and selection in bolstering the accuracy and reliability of credit risk assessment models.

In the context of credit risk assessment, the explainability and transparency of ML models have assumed paramount importance. Financial institutions, regulators, and other stakeholders require a clear understanding of the decision - making processes underlying these models to instill trust and ensure compliance. In [9], they engaged in a comprehensive discourse on the role of explainable AI in financial risk management, highlighting the necessity of developing transparent and interpretable models. In [10], they examined the intricate relationship between financial reporting in debt markets and the explainability of credit risk assessment models. These studies underscored the criticality of developing explainable and transparent ML models for credit risk assessment, as this is essential for their widespread adoption and effective implementation in real - world financial applications.

In the fast - paced and highly volatile financial market, real - time and dynamic credit risk assessment has emerged as a critical imperative. The [11] focused on the challenges and opportunities associated with real - time data processing in credit risk assessment. Their research highlighted the importance of developing predictive models that can operate in real - time, enabling timely and informed decision - making. In [12], the designed observable machine learning pipelines for real - time credit risk detection, proposing a scalable and efficient approach to handle large - scale, high - velocity data streams. These studies not only demonstrated the feasibility but also the necessity of real - time and dynamic credit risk assessment systems in adapting to the rapidly changing market conditions and effectively managing credit risk.

ML - based credit risk assessment has also found diverse applications in specific domains such as agriculture, e - commerce, and supply chain finance. The [13] conducted an in - depth case study on AI - driven credit risk assessment in the agricultural sector. Their research explored the potential of ML techniques in improving credit access for farmers, thereby contributing to the sustainable development of the agricultural economy. Nana et al. [25] employed game theory to analyze credit risk assessment in the e - commerce domain, providing valuable insights into the strategic interactions between borrowers and lenders in the online marketplace. In [14], they proposed a credit risk assessment model for SMEs in supply chain finance, integrating SVM and BP neural network. This model addressed the unique challenges and opportunities presented by the supply chain finance ecosystem. These studies showcased the versatility and adaptability of ML - based credit risk assessment systems across different industry verticals, highlighting their potential to drive innovation and efficiency in credit risk management.

### 2.2. Machine Learning and Data Mining for Credit Risk Prevention

We present machine learning and data mining related studies for credit risk prevention below.

#### 2.2.1. Risk Transmission of Affiliated Enterprises

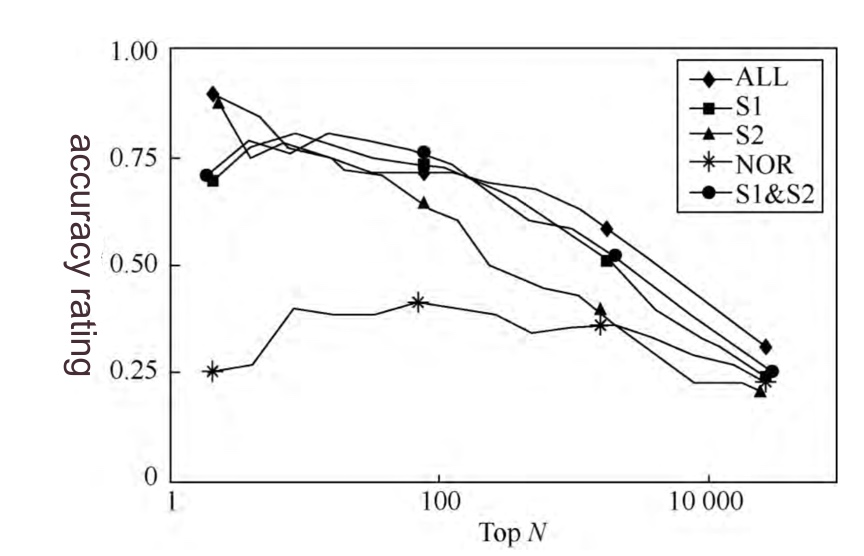
Corporate dishonesty manifests a distinct "network effect" within the investment network among enterprises. That is to say, if the investors (corporate shareholders) of a target enterprise, or its investment entities (such as subsidiaries, holding companies, joint-stock companies, etc.) engage in dishonest behavior, the target enterprise is prone to exhibiting similar dishonesty.

In [15], the generalized linear regression model is utilized. The fitting process of this model is parallelized, enabling a remarkably high calculation speed, which is well-suited for handling the millions-scale enterprises under investigation in this research. Given that predicting an enterprise's dishonest behavior represents a typical binary-classification problem, the Bernoulli distribution was selected as the function family for the generalized linear model. Consequently, the likelihood of an enterprise being a dishonest one. The discrete type of enterprise basic characteristics has been transformed into many 2-value characteristics. 1 stands for "yes" and 0 stands for "no", which is far more accurate than the regression method without considering the network effect.

Table 1 Basic features and network features involved in the prediction model

|  |  |  |
| --- | --- | --- |
| Feature | Item | Explanation |
| First-order neighbor feature | Number of dishonest investees | When the value is greater than or equal to 4, the value is 4 |
|  | Proportion of dishonest investees | Number of dishonest investees / Total number of investees |
|  | Number of dishonest investors | When the value is greater than or equal to 4, the value is 4 |
|  | Proportion of dishonest investors | Number of dishonest investors / Total number of investors |
|  | The number of dishonest first-order neighbors | When the value is greater than or equal to 4, the value is 4 |
|  | Proportion of dishonest first-order neighbors | Number of dishonest first-order neighbors / Total number of first-order neighbors |
| Second-order neighbor feature | The number of dishonest investees in second-order neighbors | When the value is greater than or equal to 4, the value is 4 |
|  | Proportion of dishonest investees in second-order neighbors | Dishonest investees in second-order neighbors / Total number of investees in second-order neighbors |
|  | The number of dishonest investors in second-order neighbors | When the value is greater than or equal to 4, the value is 4 |
|  | Proportion of dishonest investors in second-order neighbors | The number of dishonest investors in second-order neighbors / Total number of investors in second-order neighbors |
|  | The number of discredited second-order neighbors | When the value is greater than or equal to 4, the value is 4 |
|  | Proportion of dishonest second-order neighbors | Number of discredited second-order neighbors / Total number of second-order neighbors |
| Basic Feature | Enterprise type | Self-employed, limited company, discrete type |
|  | Enterprise capital | Registered capital of the company |
|  | Enterprise scale | According to the Ministry of industry and information technology joint enterprise [2011] No. 300 document "notice on printing and distributing the classification standards for SMEs", the enterprise scale calculated for different industries is divided into four categories: large, medium, small and micro, numerical type |
|  | Regional | Company location, discrete |
|  | Industry | The industry segment of the company, discrete |

In this experimental design [16], a ten-fold cross - validation methodology was employed. The dataset was partitioned into a training set and a test set in a 9:1 ratio, and the outcomes were averaged across ten independent experimental runs. In each run, leveraging the parameters estimated by the regression model trained on the training set, scores were assigned to the expected values of all enterprises within the test set. Subsequently, the enterprises with the highest - risk profiles were ranked in the leading positions. If there are dishonest enterprises among the top enterprises with the highest-risk rankings, it is defined as Precision=r/N. The variation of prediction accuracy with N is shown in the figure. In the graph, NOR represents the use of all the basic features of the enterprise, S1 represents the use of the first-order neighbor features in the network features, and S2 represents the use of the second-order neighbor features in the network features. All represent the fusion of S1, S2 and NOR features.



*Figure 1 Accuracy comparison of prediction models using different features for combination*

The predictive efficacy of first-order neighbor network features exceeds that of second-order neighbor network features. Furthermore, relying solely on the fundamental characteristics of enterprises is insufficient for accurately predicting corporate dishonest behavior. Nevertheless, when combined with network characteristics, the prediction accuracy can be significantly improved. Specifically, among the 100 enterprises with the highest predicted dishonesty risk, more than 70% are actually found to be dishonest. Similarly, approximately 40% of the top 10,000 enterprises ranked by risk display dishonesty.

The [17] constructed an extensive inter-enterprise investment network by aggregating real - world data from over 4 million enterprises. This research revealed a significant network effect in relation to corporate dishonesty. That is, if the investor or investment target of a focal enterprise engages in dishonest practices, the probability of the focal enterprise being dishonest is substantially increased. Further analysis indicates that as the number of dishonest neighbors grows, the risk of corporate dishonesty escalates rapidly.

#### 2.2.2. Evolution of the Development Stage of SMEs

The development of Small and Medium - sized Enterprises (SMEs) across diverse stages is a multifaceted process that is significantly influenced by an array of internal and external determinants. Comprehending these distinct developmental stages is of paramount importance for policymakers, financial institutions, and entrepreneurial entities. This section undertakes a comprehensive review of the extant literature on the developmental stages of SMEs, drawing upon relevant and rigorous academic studies.

In the nascent phase, SMEs are confronted with formidable challenges such as restricted resources and a dearth of brand recognition. The [18] conducted in - depth research on credit risk evaluation, unearthing the financial fragility characteristic of early - stage SMEs and underscoring the imperative for a stringent financial assessment framework. The [19] discerned that the subjective judgment of credit assessors exerts a substantial impact on SMEs' access to financial resources during this stage, thereby highlighting the significance of objective assessment methodologies.

Upon surmounting the initial impediments, SMEs transition into the growth stage, marked by expansionary endeavors and market penetration strategies. At this juncture, they necessitate substantial investment in marketing, product development, and operational enhancements. The [20] accentuated the pivotal value of real - time data processing in the context of credit risk assessment for SMEs in this stage, as it facilitates informed decision - making processes. The [21] expounded on the crucial role of explainable AI in ensuring transparency within financial risk management when SMEs embark on seeking increased funding.

During the maturity stage, SMEs direct their efforts towards fortifying their market position and diversifying their business portfolios. In [19], a credit risk assessment model is proposed by leveraging fuzzy integral and Support Vector Machine (SVM) techniques, which proves to be instrumental for mature SMEs in managing financial risks during market exploration and diversification initiatives. In [22], a heterogeneous ensemble learning model is devised for credit risk assessment, which can effectively assist SMEs in optimizing their financial strategies.

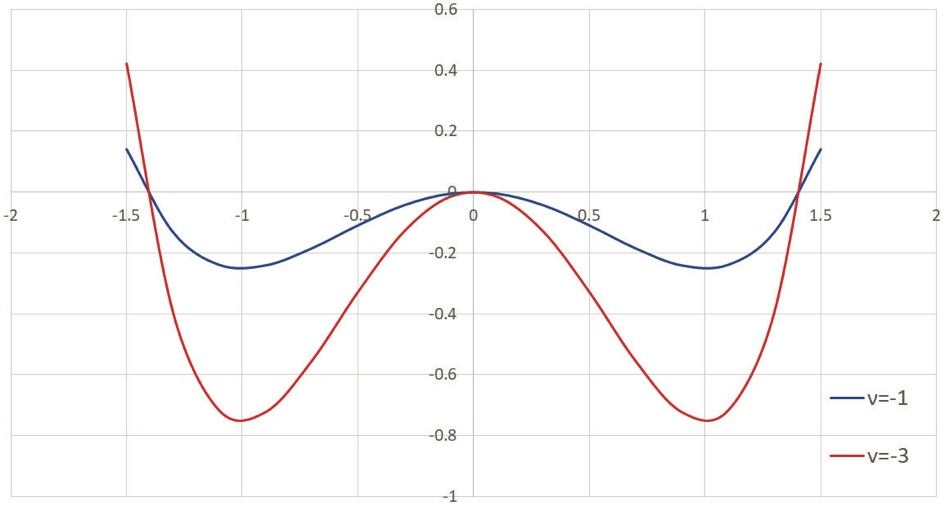
Certain SMEs embark on global expansion trajectories. Although the extant references do not directly address this specific stage in an exhaustive manner, the fundamental principles of credit risk assessment and financial management remain applicable. Advanced machine learning algorithms and real - time data processing capabilities can play a pivotal role in enabling SMEs to make strategic decisions regarding international expansion, such as evaluating international partners and managing cross - border risks.

##### 2.2.2.1. Evolution of SMEs Systems

To elucidate the evolution of Small and Medium - Sized Enterprises (SMEs) systems, it is essential to initially expound upon the following two crucial concepts: “bistable utility potential” and “periodic force from capital - product switching”.

Bistable Utility Potential. Within the context of temporal events [23], the emphasis lies on the incentive behaviors manifested when confronted with insufficient input capital. This deficiency frequently confines the enterprise to a low - level capital - product switching situation. For example, one strategic option might involve implementing cuts in technical investment to safeguard marketing expenditures. Nevertheless, this could lead to adjusted products that are unable to fully satisfy market demands, causing the enterprise to forfeit its competitive advantage during the cash - out phase. As a result, the enterprise converges towards one equilibrium state, corresponding to a local maximum of the utility potential.

Conversely, if the enterprise decides to reduce marketing investment to ensure adequate technological outlays, a scenario may emerge where even the most exquisitely developed products encounter difficulties in securing their deserved market share due to inadequate market penetration. This gives rise to another equilibrium state, corresponding to the other local maximum of the utility potential. Significantly, the equilibrium state in which the enterprise ultimately stabilizes is contingent upon the internal dynamics of cooperation and competition within the organization. The utility potential utilized to model the “capital - product switching” process can be characterized by a bistable structure. This type of structure, which is commonly observed in physics, biology, and financial systems, which shows symmetric structure with one utility barrier with the height located at and two wells located at (corresponding to left and right lowlands as the local equilibria as shown in Figure 1. For simplicity, the symmetric assumption is based on the consideration that two types of switches between two equilibria (“Capital” and “Product”) are equally important for SMEs development, and the non-equivalent switches can be described by the extended asymmetric bistable systems.

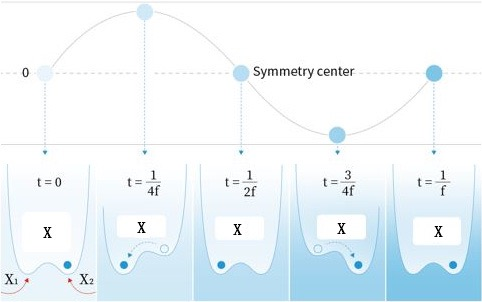


*Figure 2 The different states of the bistable potential function, respectively*

From the perspective of the underlying mechanism governing the evolution of the dynamics within the SMEs system, an enterprise can be conceptually analogized to a Brownian particle. A damping effect inherently exists, which is attributable to the viscous force originating from inter - enterprise contradictions, disarray, cooperation, and competition instigated by the alteration of the state (x(t)) at time (t). In accordance with Stokes’ law, this viscous force can be postulated to be proportional to the rate of change. Consequently, enterprise growth can be regarded as a complex coupling process between its internal structure and the external environment. It can be expounded upon at different levels depending on the system's degree of freedom: At the micro - scale, the dynamic behavior of the system can be characterized by formulating regular equation systems for each factor. However, when faced with hundreds of millions of enterprise state equations, generally speaking, an analytical solution is difficult to obtain. At the macro - scale, as details are disregarded, stochastic effects gradually counterbalance one another, thus yielding a deterministic equation for the state - evolution system. This transformation modifies the nature of the challenge in exploring the essence of modern business growth. Therefore, from a mesoscopic perspective, it is imperative to consider the aforementioned fundamental evolving forces during the four - stage lifecycle of enterprise growth, as previously delineated.

##### 2.2.2.2. Dynamic Modeling for SMEs Evolution

By establishing the dynamical model to describe the mechanism for SMEs’ evolution in terms of the evolution forces together [24]. Without loss of the generality for a given target SME denoted by x with the scale m0, its process of the growth can be regarded as a particle moving in the bistable utility potential , by following the “technology-capital paradigm”, the term , representing the moving (from the one status to another one) of the SME (denoted by x) subject to the exchange of utility U’s statuses between “technology” and “capital” states, which drives the entity x to make the change of the growth in terms of x(t) (the derivative of x(t) with respect to the time variable t). Secondly, by considering the entity x’s internal (and also external) fluctuations  (which represents the risk environment), plus the status is perturbed by the periodic force  of capital-product switching together, which can be used to describe the dynamical behaviors of SME’s evolution by the following over-damped Langevin equation (which has been widely and successfully used in statistical physics and molecular biology [30]).



*Figure 3 The capital-product switching through the periodic force modelled by the bistable utility potential*

Generally, different industries may have different cycle times of product development and market exploitation [25], however, most enterprises employ a responsible person for the annual performance index for evaluation, and the strategies of external investment and internal product planning always match the annual cycle. In other words, the switching frequency is so small that there is enough time for the enterprise to reach the locally statistic equilibrium during one period, thus the proposed system satisfies the adiabatic approximation condition: the driving frequency is much slower than the inverse value of the relaxation time, which is the time for probability within one well to equilibrate, and the asymptotic long-time distribution function can be derived in the adiabatic limit.

##### 2.2.2.3. Dynamical Mechanism for SMEs Evolution

By following the "technology-capital paradigm" and through the over-damped Langevin equation, first establish a general framework for the dynamical mechanism of SMEs evolution [26]. In particular, the effective potential is proposed to investigate the effect of periodic driving on the utility potential as shown in Fig. 3 for the mechanism of SMEs’ evolution in four stages of their life cycle.

In the presence of periodic driving, the double-well potential is tilted back and forth, thereby raising and lowering successively the potential barriers of the right and the left well, respectively, in an antisymmetric manner. This cyclic variation is shown in our cartoon. A suitable dose of noise (i.e., when the period of the driving approximately equals twice the noise-induced escape time) will make the ‘sad face’ happy by allowing synchronized hopping to the globally stable state (strictly speaking, this holds true only in the statistical average as the quantitative formula gives here in details four locations of capital-product switching).

### 2.3. Construction of Enterprise Knowledge Graph

#### 2.3.1. Knowledge Graph in Financial Field

Knowledge graphs have emerged as a powerful tool for representing and managing complex information, with their applications spanning diverse fields such as finance, healthcare, and information technology [27]. This section conducts an in-depth review of the literature on general knowledge graphs, focusing on their construction methodologies, wide-ranging applications, and the crucial role that machine learning plays in enhancing their capabilities.

The construction of knowledge graphs involves the extraction, integration, and representation of knowledge from multiple sources. In [28], a comprehensive and detailed overview is provided of knowledge graphs. The author explores the research directions and elaborates on the fundamental aspects of their construction. Hogan emphasizes the importance of declarative artificial intelligence in the development of knowledge graphs, highlighting the role of expert systems and semantic web technologies. This work serves as a cornerstone for understanding the technical underpinnings of knowledge graphs and their potential applications across various disciplines.

Knowledge graphs have been applied in a wide range of scenarios. Each application takes advantage of the structured representation of knowledge to improve decision-making processes and information retrieval efficiency. In the financial domain, knowledge graphs have proven to be invaluable in credit risk assessment. For example in [29], a prototype is developed for credit card fraud management. This application vividly demonstrates how knowledge graphs can be utilized to identify patterns and anomalies that may indicate fraudulent behavior. Similarly, [30] discusses the improvement of predictive models and decision-making in credit risk assessment through real-time data processing. The integration of knowledge graphs can significantly facilitate such processes.

In the area of agricultural finance [30], a case study on AI-driven credit risk assessment is provided. This study demonstrates how knowledge graphs can be effectively used to manage financial risks in a sector that is faced with unique challenges. It emphasizes the significance of contextual information and the remarkable ability of knowledge graphs to incorporate such data into the risk assessment process.

Dynamic knowledge graphs represent an advanced form of knowledge representation [31]. They have the capacity to adapt to evolving information and changing contexts. The use of Markov Random Fields (MRF) in combination with knowledge graphs, presents a novel approach to handling dynamic and uncertain information. This approach holds great promise in domains such as finance, where market conditions and risk factors are in a state of constant change.

#### 2.3.2. Construction of Financial Knowledge Graph

In the field of credit risk control, many people think that the key to build knowledge graph system is algorithm and development. However, the fact is not as imagined. In fact, the most important core lies in the understanding of the business and the design of the knowledge graph itself [32]. This is similar to the design of database tables for a business system. Moreover, this design is absolutely inseparable from the in-depth understanding of the business and the prediction of changes in future business scenarios.

The construction of knowledge graph [33] is an engineering practice, which is organized into the following processes: “Theme engineering”, “Atlas engineering”, “Computing engineering” and “Evaluation engineering”.

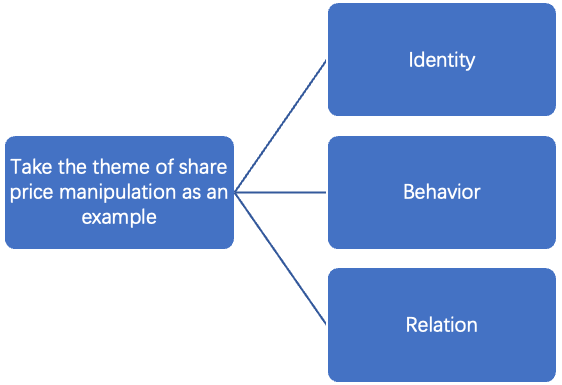
Table 2 Knowledge graph construction organization

|  |  |  |  |
| --- | --- | --- | --- |
| Theme Engineering | Atlas Engineering | Computing Engineering | Evaluation Engineering |
| Business carding  Subject description  Requirement verification  Requirement management  Output subject requirement specification | Hand built  Auto build  Hybrid build  Output knowledge graph | Query and retrieval  Index statistics  Association analysis  Node classification  Anomaly detection  Inferential prediction  Time series analysis  Output the calculation  results  Output mode library | Ontology evaluation  Atlas evaluation  Calculation evaluation  Mode library evaluation |

##### 2.3.2.1. Theme Engineering

The theme project mainly includes analyzing business scenarios, sorting out business themes and analyzing the requirements of application for knowledge graph [34].

In the process of sorting out the theme requirements, IBR (identity behavior relation) specification [35] is very effective, which is described as follows:



*Figure 5 IBR specification*

Identity：The subjects involved in this topic include the identification of the individual and its uniqueness, such as the actual operator, account registrant, account, share, trading event, as well as the identification and uniqueness identification of each individual, such as how to identify as the same person when the ID card is confirmed.

Behavior：The behavior and quantification involved in this topic describe how a person controls multiple accounts (with the same trading address), centralizes (time window constraints), and performs a large number of trading operations on a share (threshold constraints).

Relation: Association includes the relationship / action between individuals.

For example,

Operator--control-- share account

Account registrant -- own -- share account

Account -- buy / sell --share

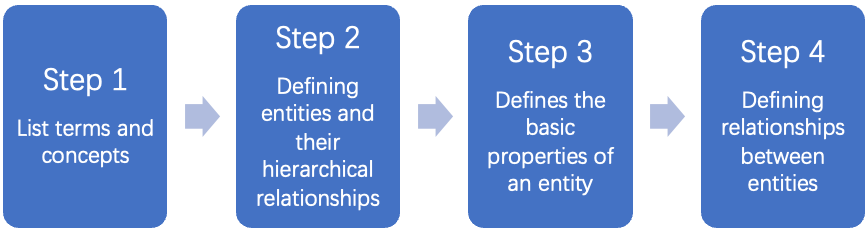
It also includes the correlation between events. For example, in a certain period of time, the manipulation of share price by some people causes the share price to fluctuate greatly.

##### 2.3.2.2. Atlas Engineering

Ontology is the framework of the whole knowledge graph [36]. There are three ways to construct Atlas: manual mode, automatic mode and mixed mode.

##### 2.3.2.3. Ontology Model

It determines the organization of knowledge and the breadth and depth of business exploration, is very important for knowledge graph. Ontology modeling [37] needs four steps：



*Figure 6 Steps of ontology modeling*

* List Terms and Concepts

For example, in the subject of insider trading, relevant terms and concepts should be listed according to the results of the previous business analysis, which should be as complete, clear and non-overlapping as possible. This process can also be generated through machine learning, but it needs the review of domain experts.

* Define Entities and Their Hierarchical Relationships

For example, in the entity of organization, enterprises can drill down into enterprises, government agencies, non-profit organizations and others. Enterprises can drill down from different dimensions, including listed / unlisted, state-owned / private / joint venture / foreign capital. this process is basically completed by domain experts, so as to get accurate hierarchical relationship and lay a good foundation for subsequent data statistics, mining and reasoning.

* Define the Basic Attributes of An Entity

Basic attributes are used to distinguish the characteristics of each instance in the same type of entity and are also used for business analysis and mining indicators. An entity may be used for multiple topics, so the attributes should be as comprehensive as possible. For example, the company entity will appear in almost all topics, and its attribute information needs to be as rich as possible.

* Defining Relationships Between Entities

There are multiple relationships among entities, which can be understood as the existence of three-dimensional layers, such as social relations, communication relations, accounting relations, loan guarantee relations, judicial litigation relations and so on.

##### 2.3.2.4. Data Pre-processing

Before processing data, the quality, availability, relevance and consistency of multi-source heterogeneous data should be analyzed, and problem data should be found, and processing rules should be formulated [38]. Due to the complexity of data, the pre-processing process has the following methods:

To understand the business meaning of data sources, it is necessary to consult domain experts,

* verify with database administrator, and investigate the professional knowledge and rules in the field of data description;

* analyze the timeliness of data sources: starting from the requirements, check the freshness and timeliness of data and the differences between the old and new data, and evaluate whether the data timeliness meets the business requirements;

* overview and statistics of data sources: from the dimensions of data volume, data type, data volume, field value distribution, statistical data overview;

* quality assessment of data sources: assess the proportion of problem data (missing, duplicate, wrong, unavailable) and consider data availability;

* relevancy evaluation: evaluate the data relevancy from whether the primary and foreign key values between tables are valid, the proportion of successful primary and foreign key Association, and the auxiliary association degree of condition matching;

The goal of data pre-processing is to make the output data consistent, accurate, complete, standardized, logical and effective. The processing methods include but are not limited to filtering, completion, format conversion, merging and splitting, translation, mapping, business calculation.

* Filtering: judgment and filtering of invalid or wrong data

* Completion: the completion of missing data

* Format conversion: the data is converted into the required format for storage by decoding / transcoding

* Merge and split: process fields are merged or split into target fields according to requirements

* Data translation: translate data into semantic specification of target data through rules and other ways

* Data mapping: according to the mapping rules and business logic, the data source is mapped to the data format and structure of the target

* Business calculation: process and calculate the source data according to business requirements

##### 2.3.2.5. Entity and relationship identification

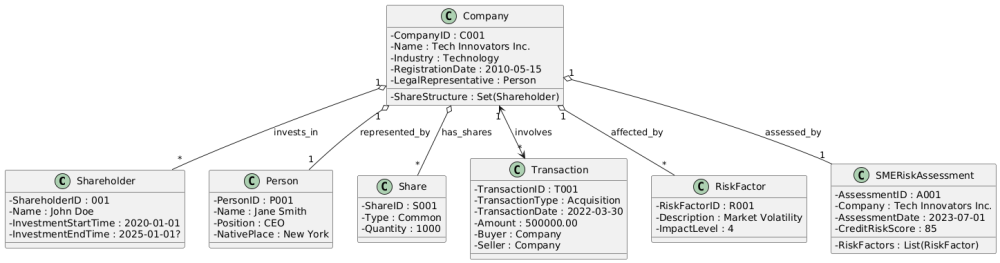
The identification of entities and relationships within structured data predominantly depends on the ETL (Extract, Transform, Load) approach [39]. Regarding unstructured data, named entity recognition includes, yet is not limited to, personal names, place names, organization names, and proper nouns. This process generally involves two steps: (1) identifying entity boundaries; (2) determining entity categories (such as personal names, place names, or organization names). Relationship recognition is based on entity recognition, and the methods adopted can be of a general kind. There exist three main types of statistical methods: rule-based, statistical, and hybrid methods.

In rule - based methods [40,41], linguistic experts usually construct rule templates manually. The features taken into account comprise statistical information, punctuation, keywords, deixis, directional words, location-specific words (such as suffixes) and prefixes. The primary mechanism of rule-based methods is pattern and string matching. The majority of these systems are contingent upon the establishment of a knowledge base and a dictionary. In contrast, statistical methods [42,43] are utilized to train data that has been manually annotated. Common techniques involve the Hidden Markov Model (HMM) [44], Maximum Entropy (ME) [45], Support Vector Machine (SVM) [46], Conditional Random Fields (CRF) [47], and others. Natural language processing is not a completely random process. Relying solely on statistical methods may lead to an extremely large state-search space. Therefore, it is essential to conduct prefiltering and pruning with the assistance of rule-based knowledge. Common approaches involve the use of cascading Hidden Markov Models or the integration of certain rules into statistical-based learning methods, thus combining machine learning with human-crafted knowledge.

##### 2.3.2.6. Fusion and Disambiguation

Entity fusion refers to the process of integrating multiple entities that are essentially identical into a single entity in accordance with specific rules or logical frameworks. This approach serves to reduce ambiguity and enhance the overall data quality. Conversely, entity disambiguation aims to resolve the multiple-meaning references of entities within a text and link them to the unique corresponding entity within the knowledge base, thus contributing to the improvement of data quality. Entity disambiguation principally depends on a set of rules. These disambiguation rules include, but are not limited to, the following principles: Nodes with high access priority are preferred; Entities with the most recent warehousing time are given priority; Data sources with rich authority are prioritized.

After the above process, a basic ontology model can be obtained, which can be visualized as follows:



*Figure 7 An ontology model example entity and relation*

Figure 7 shows a financial risk assessment system for Small and Medium-sized Enterprises (SMEs), featuring classes for Shareholders, Companies, Persons, Shares, Transactions, Risk Factors, and SMERiskAssessments. Each class includes specific attributes, such as IDs, names, dates, and financial details, and interconnections, such as a company being invested in by multiple shareholders, legally represented by a person, holding various shares, and conducting multiple transactions. The system integrates risk factors that affect companies and uses risk assessments to evaluate the credit risk score of SMEs. This setup allows for comprehensive tracking and analysis of financial activities, ownership structures, legal representations, and risk profiles, facilitating robust financial risk management and informed decision-making for stakeholders.

##### 2.3.2.7. Calculation Engineering

After the construction of the knowledge graph, a complex network with entities, relationships and attributes as the elements will be formed [48,49]. It is also necessary to provide the ability support of graph computing for the application side, because all the applications of the atlas are based on the large-scale calculation of this complex network. These calculations include but are not limited to:

* Query and retrieval: n-degree related parties of the target node, or query a sub graph structure

* Index statistics: calculate the graph features and attribute features of a single node or multiple nodes in the statistical atlas, which can be converted into feature vectors;

* Association analysis: analyze the relationship and tightness between two or more nodes in the atlas;

* Node classification: nodes are classified according to graph features or associated attribute features

* Anomaly detection: abnormal nodes and subgraph patterns of abnormal subgraphs are found in the whole network

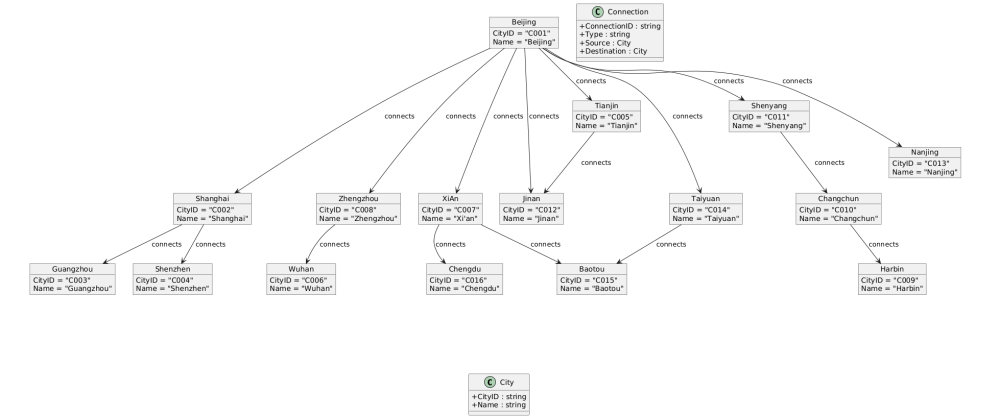
* Predictive reasoning: predict and infer new relationships and information from existing knowledge graphs by rules or machine learning methods

* Time sequence analysis: time sequence analysis for single relationship and event, or time sequence analysis for changes of network topology

The calculation results are provided to the application side in the form of online services or offline results, so as to enable the business system and realize various analysis functions.

**2.3.2.7.1. Query and Retrieval**

Graph query and retrieval is the most common calculation. It is often used to query the n-degree related parties of the target node or query a sub graph structure [50,51,52]. It mainly traverses the network by depth first or breadth first, and outputs the associated nodes or isomorphic instances.



*Figure 8 Query retrieval examples*

Figure 8 illustrates a query retrieval example in a network of cities, depicting the interconnections between various cities represented as objects. Each city, such as Beijing, Shanghai, and Guangzhou, is represented with unique identifiers and names. The connections between these cities, labeled as "connects," indicate direct relationships. For example, Beijing is connected to Tianjin, Shenyang, and other cities, while further connections extend from these cities to others, forming a complex network. These connections facilitate querying n-degree related parties or sub-graph structures, beneficial for applications that analyze network relationships such as route planning, logistics, and social networks. The City class at the bottom provides a generalization for all city objects, defining common attributes shared by every city node in the network.

**2.3.2.7.2. Index Statistics**

Index statistics refers to the statistics of the graph characteristics of a single node or multiple nodes in the graph [53,54]. The indicators of single subject map include:

* Out of degree: it represents the number of edges sent out by a node, which is defined as the total number of relationships sent out by a node

* Degree of penetration: it represents the number of edges received by a node, which is defined as the total number of relationships pointing to the node

* dielectric: it indicates the importance of bridge action of a node

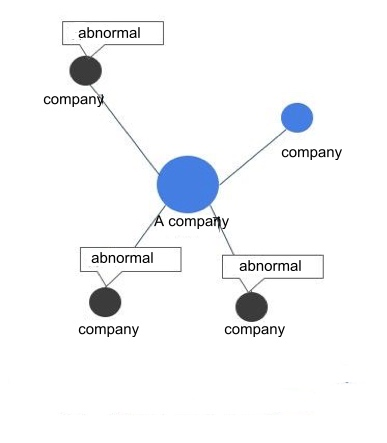
* Centrality: it characterizes the importance of nodes in the current subnet

* Graph diameter: it represents the maximum value of the shortest path between any two nodes in the current subnet

* Clustering coefficient: the coefficient represents the degree of node aggregation in the graph

* Graph density: it represents the number of relationships between nodes in the current subnet, which is defined as the ratio between the number of actual relationships in the current subnet and the maximum number of relationships that can exist in the theory of current subnet nodes

Furthermore, the associative-attribute characteristics of both single-node and multi-node scenarios are highly prevalent. In the case of a single agent, these characteristics are manifested through the proportion and statistical analysis of its respective attributes among the first-degree related entities of an individual entity. Regarding multi-agent associative attributes, they are exemplified by the statistical analysis of attributes and the distribution proportion of all subjects within a group. Additionally, the proportion and statistical evaluation of each attribute of the primary affiliated entity within the group are also significant. These characteristics play a pivotal role in discerning the relationships and patterns within the knowledge graph, thereby facilitating more profound analysis and utilization of the data for financial research and decision-making processes.

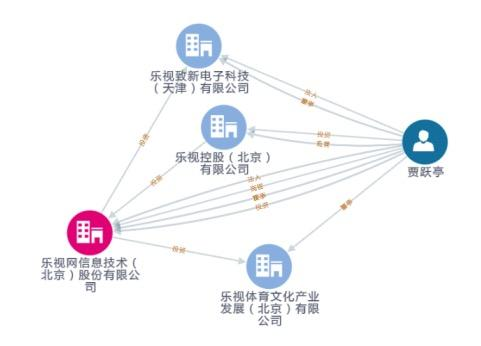


*Figure 9 Example of indicator statistics*

The number of once related parties of company a with abnormal finance is 3, and the proportion of abnormal finance in the group is 60%.

**2.3.2.7.3. Association Analysis**

In the realm of graph-theoretic frameworks, association analysis entails a comprehensive and meticulous investigation into the inter - relationships and the degrees of adjacency among two or more nodes. This analytical paradigm plays a pivotal role in enabling the discovery and demarcation of communities within the graph structure [55,56]. Examples elucidating this concept include the multiplex investment linkages established between two corporate entities, the shortest route connecting an individual to an enterprise, the weighted metric of separation between two individuals, and the transfer-based associations among multiple accounts. The methodologies commonly deployed in this form of analysis are path-based querying algorithms and distance-calculation techniques. The outcomes of these analyses are either a collection of edges that together form a path or a numerical value denoting the distance between nodes, accompanied by the edges that bridge them. These outputs are indispensable for discerning the complex and intricate relationships embedded within the graph, which can be effectively utilized across a spectrum of financial applications, such as risk quantification, portfolio optimization, and market trend prognostication.



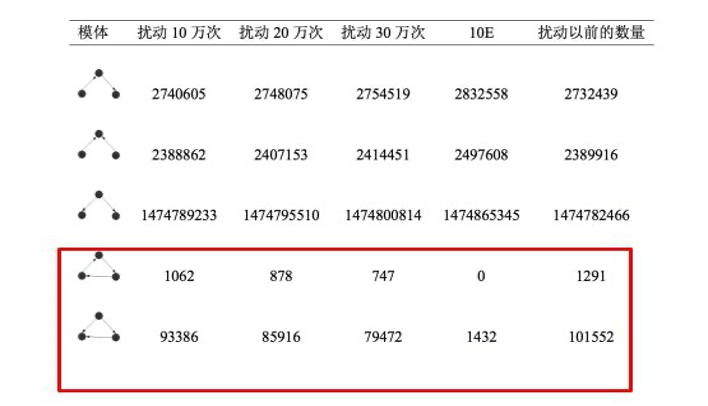
*Figure 9 Example of correlation analysis calculation*

**2.3.2.7.4. Node Classification**

Node classification involves categorizing nodes according to graph-intrinsic features or the characteristics of their associated attributes [57,58]. For example, companies that encounter credit default frequently demonstrate a characteristic risk-propagation pathway. In this context, the number of defaulting companies among first-degree affiliated entities can be utilized as a classification-characteristic index. Likewise, the sub-graph features of transfer relationships within money-laundering accounts can serve as the foundation for classifying nodes with a money-laundering label. Common methodologies in node classification include the labeling of target nodes, extraction of relevant features, and application of classification algorithms. The output of these processes usually leads to the generation of a feature-atlas database. This database not only arranges the classified information in a structured fashion but also offers a valuable resource for further analysis in the financial realm, such as for risk assessment, fraud detection, and market-behavior prediction.

**2.3.2.7.5. Abnormal Detection**

Anomaly detection, within the framework of a network, is concerned with the identification of deviant nodes and atypical sub-graph patterns throughout the entire network [59]. Examples of such anomalies involve nodes featuring outlier in-degree and out-degree values, closed-loop investment relationships, and sub-graph patterns that, although lacking an immediately apparent business significance, display infrequent yet recurrent occurrences. Methods commonly employed for anomaly detection include clustering techniques and sub-graph discovery algorithms, among others. The principal outputs of these detection procedures consist of an abnormal node repository and an abnormal sub-graph structure pattern repository. These repositories function as archives of identified anomalies, empowering financial analysts and researchers to conduct systematic analyses and gain insights into the irregularities within the network. This is of critical importance for detecting financial fraud, evaluating systemic risks, and unearthing latent market trends.



*Figure 10 Example of anomaly detection calculation*

**2.3.2.7.6. Predictive Reasoning**

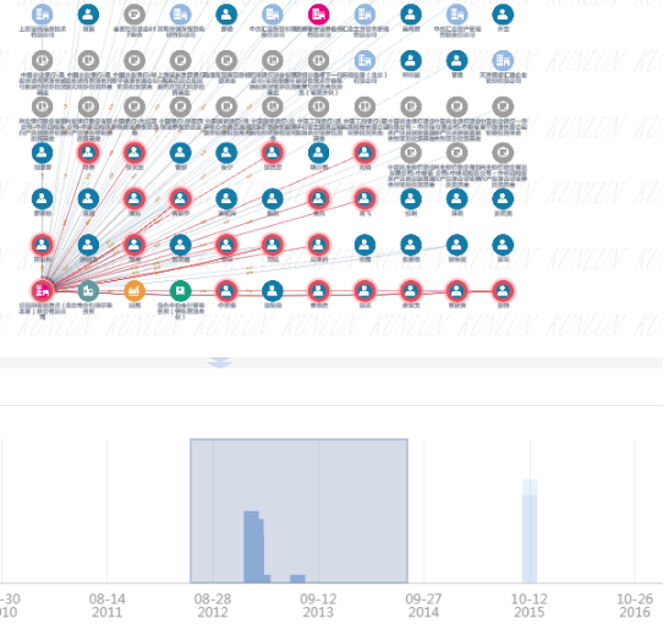
Predictive reasoning, within the domain of knowledge graphs, entails the prediction and inference of novel relationships and information from the existing knowledge graph [60]. This is accomplished through the utilization of rules or probability theory, rendering it highly suitable for reasoning about tenuous relationships and performing link predictions. Common methodologies adopted in predictive reasoning encompass rule-based reasoning and machine-learning algorithms. The results of these predictive processes can materialize as new nodes, novel relationships, and other pertinent information. Such novel insights derived from predictive reasoning are of inestimable value within the finance domain. They can facilitate tasks such as risk assessment, market trend prediction, and investment decision-making by uncovering hitherto unrecognized connections and patterns within the financial data represented in the knowledge graph.



*Figure 11 Example of predictive reasoning calculation*

**2.3.2.7.7. Timing Analysis**

In the realm of finance, time series analysis is dedicated to the scrutiny of the temporal succession of individual relationships and events, as well as the development of network topology structures [61]. For instance, it encompasses the recurring patterns of industrial and commercial changes and the channels through which risk diffuses within a network. The methodologies commonly employed in this analysis comprise time-series analytical techniques and risk-propagation models. The outcomes of these analyses typically involve the identification of anomalous timings and the computation of risk scores. These results hold substantial significance in the financial field, as they facilitate the detection of irregularities in the temporal behavior of financial variables and the evaluation of potential risks associated with network-based financial systems. Consequently, they offer invaluable insights for risk management, investment strategies, and financial decision-making processes.



*Figure 12 Example of time series analysis*

##### 2.3.2.8. Evaluation Engineering

Evaluation engineering should run through the whole process of knowledge graph construction, from ontology model [60], knowledge graph [61], calculation efficiency to calculation results, so as to continuously optimize ontology model, knowledge graph, optimization algorithm, optimization pattern base.

**2.3.2.8.1. Ontology Evaluation**

A well-constructed ontology model [62] must conform to the following criteria:

* Definiteness and Objectivity: A clear and objective semantic definition of the defined terms shall be provided in natural language. This ensures that the meaning of each term is communicated unambiguously, leaving no scope for subjective interpretation.

* Completeness: The definition must be comprehensive, and capable of fully encapsulating the essence of the terms being described. Every aspect relevant to the term's meaning should be considered, leaving no semantic gaps.

* Consistency: There should be no internal contradictions. The inferences derived from the terms must be in harmony with the terms' inherent meanings. Any logical discrepancies would undermine the integrity of the ontology model.

* Maximum Monotone Scalability: When integrating either general or specialized terms into the model, there should be no need to modify the pre - existing content. The model should be designed in such a way that it can seamlessly accommodate new elements without disrupting the existing structure or relationships.

* Minimum Commitment: Minimize the number of constraints imposed. This implies that the ontology conventions should be kept to the absolute minimum, and the fewest possible restrictions should be placed on the modeling objects. This approach enables greater flexibility and adaptability within the model.

Upon the completion of ontology modeling, it is essential to conduct a quantitative assessment based on the aforesaid standards. Only after numerous iterative refinements and attaining an acceptable quality level can the subsequent steps be initiated.

**2.3.2.8.2. Atlas Evaluation**

Once the knowledge graph has been constructed, a comprehensive quality assessment must be implemented [63,64]. This assessment encompasses, yet is not limited to, the following dimensions:

* Knowledge Accuracy: This entails evaluating the precision of entity recognition and the ambiguity rate. Precise entity recognition is of utmost importance as it serves as the cornerstone of a reliable knowledge graph. A low ambiguity rate ensures that the information conveyed is clear and unambiguous.

* Knowledge Rationality: It involves appraising the rationality of relationships and attributes. The relationships among entities should be logically coherent within the context of the financial domain, and the attributes assigned to entities must be appropriate and consistent.

* Knowledge Completeness: This relates to the degree of business description and data integrity. A complete knowledge graph should comprehensively portray relevant business scenarios and possess all the requisite data to support accurate analysis.

* Knowledge Reasoning: Ascertaining whether there is consistency in reasoning logic and rules. Sound reasoning is essential for deriving meaningful insights and making valid inferences from the knowledge graph.

Evaluating a knowledge graph presents substantial challenges [65]. While conducting sample statistics for accuracy is relatively straightforward, assessing rationality, completeness, and reasoning demands the collaborative efforts of all parties involved in the knowledge graph's creation and utilization. Moreover, the formulation of an expert-scored quantitative table is indispensable to ensure a systematic and objective evaluation in these aspects.

**2.3.2.8.3. Calculation Evaluation**

For the large-scale knowledge graph with more than 10 billion nodes, the calculation efficiency and calculation results should be evaluated, so as to continuously optimize.

*Table 3 Calculation and evaluation*

|  |  |
| --- | --- |
| Graph calculation | Basic evaluation index |
| Query efficiency | Time complexity  Spatial complexity |
| Query results | Accuracy  Recall |
| Other algorithms | Corresponding machine learning evaluation method |

**2.3.2.8.4. Model Evaluation**

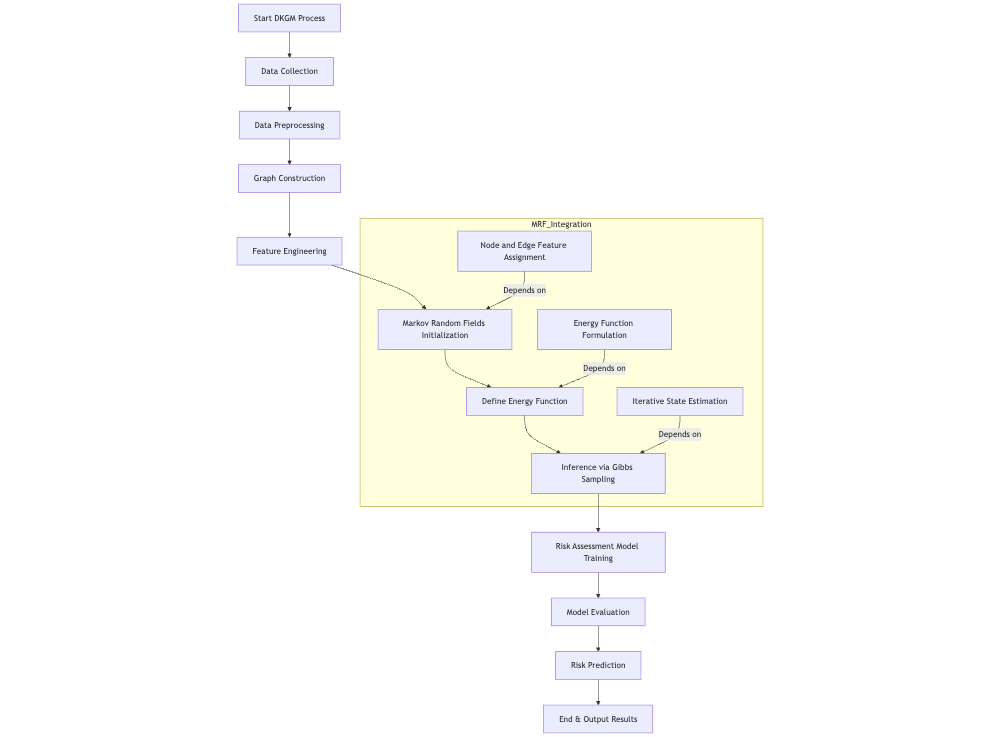
Evaluation of Abnormal Subgraphs [65,66]: The frequent subgraphs obtained from machine - learning algorithms can have their frequencies and indices determined via statistical analysis. Although machine learning is efficacious in uncovering abnormal patterns, a substantial number of the so - called "abnormal patterns" might in fact represent normal business operations. Consequently, it is of utmost necessity for business experts to conduct meticulous screening of violation indicators. Through comprehensive evaluation, these patterns can be identified as "counter - examples" for future retrieval and warning purposes. Frequent subgraphs manually extracted based on industry - specific business experience often possess significance in relation to business anomalies. Nevertheless, the calculation results of these subgraphs still need to be correlated with the model. Specifically, the frequency of occurrence and the proportion of actual violations should be evaluated to determine whether the manually - extracted subgraphs can be continuously utilized without modification or need to be adjusted. This process ensures that the subgraphs are reliable and relevant for financial analysis and risk assessment within the framework of the knowledge graph.

*Table 4 Pattern Library*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | High frequency | Medium frequency | Low frequency | Other information |
| Confirm violation | High alert  Focus on | Moderate alert  Focus on | Alert  Focus on | Creating counter examples manually |
| High suspicion of violation | Moderate alert  High attention | High attention | Moderate attention | Creating counter examples manually |
| Low suspicion of violation | Moderate attention | Continuous attention | Continuous attention |  |

## 3. Design and Development of Dynamic Knowledge Graph Model Enhanced with Markov Random Fields

This thesis proposes a markov random field [67] integrated learning scheme within the dynamic knowledge graph model. The framework overview can be found below.



*Figure 13 Model pipeline*

The proposed model leverages a Dynamic Knowledge Graph Model (DKGM) integrated with Markov Random Fields (MRF) to evaluate the credit risk of Small and Medium-Sized Enterprises (SMEs) [68,69]. The process initiates with the collection and pre-processing of relevant data. Subsequently, this data is utilized for constructing a knowledge graph, wherein entities and their inter-relationships are meticulously mapped. Feature engineering is then applied to capture the financial, operational, and network-related aspects of SMEs.

Within this graph, an MRF is initialized to model the probabilistic dependencies [70]. This commences with the assignment of features to both nodes and edges. An energy function is defined to represent the configurations of node states, guiding the inference process via Gibbs sampling [71]. The objective is to optimize the states to achieve a lower "energy" level, indicating more probable configurations. The outputs from this inference are then fed into a risk assessment model, which is trained and evaluated to ensure its predictive accuracy. The ultimate goal is to generate risk predictions, providing insights and visualizations that can support well-informed lending decisions for SMEs.

The Dynamic Knowledge Graph Model (DKGM) augmented by Markov Random Fields (MRF) represents a comprehensive framework designed for assessing the credit risks of SMEs. This framework is initiated by the "Start DKGM Process" and progresses through a sequence of systematic steps, each building upon the previous one to form a robust model. The process begins with "Data Collection," during which both structured and unstructured data relevant to SMEs are gathered. This data then undergoes "Data Pre-processing" to be cleansed and prepared for analysis, ensuring it is free from errors such as missing values or duplicates.

"Graph Construction" [72] follows, where a knowledge graph is established. In this graph, nodes denote entities such as SMEs, and edges represent the relationships between these entities, such as financial transactions or partnerships. "Feature Engineering" [73] then extracts and constructs features from the data, encapsulating the financial and operational aspects of SMEs, as well as their network properties.

The integration of Markov Random Fields commences with "Markov Random Fields Initialization," where nodes and edges are defined, laying the foundation for probabilistic modeling [74]. The "Define Energy Function" step formulates a function that quantifies the arrangements of node states within the MRF, steering the model towards more probable states. "Inference via Gibbs Sampling" utilizes Gibbs sampling to iteratively update node states, minimizing the energy function and thereby refining the model's understanding of credit risk patterns. The framework then advances to "Risk Assessment Model Training," where the risk assessment model is trained using the insights derived from the MRF. This model learns to identify patterns associated with credit risk. "Model Evaluation" ensues, during which the model's performance is rigorously examined using metrics such as accuracy, AUC, and KS statistics to ensure its reliability and effectiveness.

Finally, the "Risk Prediction" step deploys the model to forecast the credit risks of SMEs, equipping financial institutions with the insights required to make judicious lending decisions [75,76]. The process concludes with "End & Output Results," where the outcomes of the risk assessment are presented, potentially including risk scores, predictive analytics, and visualizations of the knowledge graph. The "MRF\_Integration" sub-graph within this framework emphasizes the crucial steps involved in integrating MRF into the DKGM. These include the assignment of features to nodes and edges, the formulation of an energy function that captures the dependencies between nodes, and the iterative state estimation that refines the model's predictions. This integration of MRF into the DKGM enhances the model's ability to represent and reason about the complex network structures characteristic of SMEs, offering a nuanced and dynamic approach to credit risk assessment.

### 3.1. Rationale for Integrating Markov Random Fields in DKGM

The Dynamic Knowledge Graph Model (DKGM) is a cutting-edge approach designed to assess the credit risks of Small and Medium Enterprises (SMEs) in the aftermath of economic disruptions. Traditional credit risk assessment models, which are often tailored for large enterprises, frequently fall short when applied to SMEs due to the unique complexities and vulnerabilities inherent in these smaller entities. The DKGM addresses this gap by incorporating a financial knowledge map that encapsulates both structured and unstructured data, providing a comprehensive view of the financial ecosystem.

The Dynamic Knowledge Graph Model (DKGM) embodies a cutting - edge methodology designed to assess the credit risks of Small and Medium - Sized Enterprises (SMEs) in the aftermath of economic upheavals. Traditional credit risk assessment models, which are predominantly tailored for large enterprises, frequently prove insufficient when applied to SMEs [80]. This is attributable to the unique complexities and vulnerabilities inherent in these smaller entities. The DKGM addresses this disparity by integrating a financial knowledge map that encompasses both structured and unstructured data, thus providing a comprehensive view of the financial ecosystem.

The integration of Markov Random Fields (MRF) within the DKGM represents a critical enhancement, enabling the modeling of intricate relationships and dependencies within the network of SMEs. MRF, a probabilistic model, is particularly adept at representing spatially - related data. Consequently, it emerges as an optimal option for capturing the elaborate network of financial relationships and interactions among SMEs. The MRF component within the DKGM operates on the basis of defining a joint probability distribution over the nodes and edges of the knowledge graph. Each node in the graph represents an entity (e.g., an SME), while each edge denotes a relationship between entities (e.g., financial transactions or ownership connections). The MRF model posits that the state of a node is contingent solely upon the states of its neighboring nodes, in accordance with the Markov property.

The process of integrating MRF into the DKGM involves several essential steps [81]: The foundation of the DKGM is a knowledge graph meticulously constructed from financial data. This graph comprises nodes representing SMEs and edges indicating their relationships. Features are meticulously engineered to capture the financial and operational aspects of SMEs, along with their network attributes. These features form the cornerstone of the MRF model. The MRF model is initialized by precisely defining the nodes and edges in the graph, establishing the structural framework for probabilistic modeling [82]. An energy function is formulated over the graph, quantifying the configurations of node states. This function guides the model towards lower-energy (more probable) states. Gibbs sampling is utilized to conduct inference within the MRF, iteratively updating node states to minimize the energy function and refine the model's understanding of credit risk patterns. The output generated by the MRF is employed to train the risk assessment model, which discerns the patterns associated with credit risk. The performance of the trained model is evaluated using metrics such as accuracy, AUC, and KS statistic. The model is deployed to predict the credit risk of SMEs, providing financial institutions with the insights necessary for making informed lending decisions.

The integration of MRF into the DKGM enhances the model's ability to capture the non-linear relationships and high-dimensionality characteristic of big data, a challenge that traditional statistical methods often struggle to overcome. This integration enables a more refined and accurate prediction of credit risks [83], presenting a dynamic evaluation approach that has the potential to revolutionize SME risk prediction and prevention strategies.

The empirical analysis will involve a comprehensive and systematic exploration of the network structure of SMEs across diverse sectors. This will yield insights into how these enterprises are influenced by macroeconomic factors, the risk transmission among affiliated enterprises, and the evolution of their developmental stages. The findings of this research are expected to contribute to the existing body of knowledge on credit risk assessment and offer practical solutions for financial institutions and policymakers in alleviating the credit risks faced by SMEs.

The integration of MRF into the DKGM provides a robust and dynamic framework for credit risk assessment in SMEs [84,85]. By leveraging the spatial dependencies inherent in network data, the DKGM offers a nuanced and dynamic approach to predicting credit risks, enhancing the accuracy and reliability of financial risk assessments. This innovative approach aims to establish a new benchmark in the field of credit risk assessment for SMEs, providing an essential tool for navigating the intricate and dynamic financial landscape shaped by economic disruptions.

### 3.2. Research Design

Our proposed model is designed to assess the credit risk of Small and Medium-Sized Enterprises (SMEs) in China through the synergistic integration of the Dynamic Knowledge Graph Model (DKGM) and Markov Random Fields (MRFs). This integration is meticulously devised to enhance the model's efficacy in capturing the complex relationships and dependencies among a variety of risk factors that affect SMEs. The DKGM serves as the fundamental framework for processing both structured and unstructured data, enabling the extraction of "enterprise risk genes" from the complex financial network of SMEs. By leveraging Gibbs sampling within the Markov Chain Monte Carlo (MCMC) framework, the DKGM can proficiently navigate high-dimensional data spaces and identify relevant risk characteristics. To further augment the capabilities of the DKGM, MRFs are incorporated. MRFs are undirected graphical models proficient in representing the joint distribution of random variables. They endow the model with the capacity to model the dependencies between diverse risk factors while strictly adhering to the Markov property. In this property, a variable is conditionally independent of all other variables, given its neighboring variables. This characteristic is particularly beneficial in our context, as it enables the model to focus on local interactions, thereby effectively managing the complexity inherent in the overall system. The integration of MRFs into the DKGM significantly facilitates the modeling of both spatial and temporal dependencies among SMEs' risk factors. This, in turn, allows for a more refined and in-depth understanding of how these factors interact and evolve over time. For instance, the relationships among macroeconomic conditions, the risk transmission from affiliated enterprises, and the progression of SME development stages can be appropriately modeled as a network. In this network, each node represents a distinct risk factor, and the edges depict the dependencies between them. Through this comprehensive approach, our model aims to provide a more accurate and comprehensive assessment of SME credit risk, thereby offering valuable insights for financial decision-making processes.

#### 3.2.1. Design concept

Our research design philosophy is anchored in the integration of the Dynamic Knowledge Graph Model (DKGM) and Markov Random Fields (MRFs) to construct an exquisitely sophisticated framework for evaluating the credit risks of Small and Medium-Sized Enterprises (SMEs). This hybrid model effectively synthesizes the individual strengths of the DKGM and MRFs, presenting a comprehensive and incisive assessment of SMEs' creditworthiness [86].

Central to our design is the Dynamic Knowledge Graph Model (DKGM), which demonstrates remarkable proficiency in handling both structured and unstructured data to construct an elaborate financial knowledge map. By integrating a diverse range of data elements, such as macroeconomic indices, corporate financial statements, and industry trends, this model encapsulates a comprehensive view of the financial ecosystem. Employing Gibbs sampling within the Markov Chain Monte Carlo (MCMC) framework, the DKGM enables the extraction of risk characteristics from complex datasets with extraordinary accuracy. This approach allows for a meticulous exploration of the multifaceted relationships within the financial data, uncovering latent risk factors that might otherwise remain hidden.

To enhance the capabilities of the DKGM, we introduce Markov Random Fields (MRFs), which are probabilistic models renowned for their ability to represent intricate dependency structures among variables. The incorporation of MRFs into our model empowers it to capture the spatial and temporal correlations among different risk factors. In the context of financial risk assessment, this feature is particularly valuable, as it provides a more realistic representation of how risk factors are interconnected and mutually influential. By leveraging MRFs, our model can better account for the dynamic nature of risk propagation within the financial ecosystem, enabling more accurate predictions and informed decision-making.

The integration of MRFs with the DKGM generates a synergistic effect. The knowledge graph, provided by the DKGM, offers a rich and detailed collection of data points and relationships, while MRFs provide a statistical framework that facilitates the understanding and quantification of the dependencies among these relationships. This integration not only enables our model to identify risk factors but also equips it with the ability to discern how changes in one part of the financial ecosystem can spread throughout, affecting other components. Consequently, the model can conduct a more dynamic and responsive risk assessment, adapting to the ever - evolving landscape of SME credit risk.

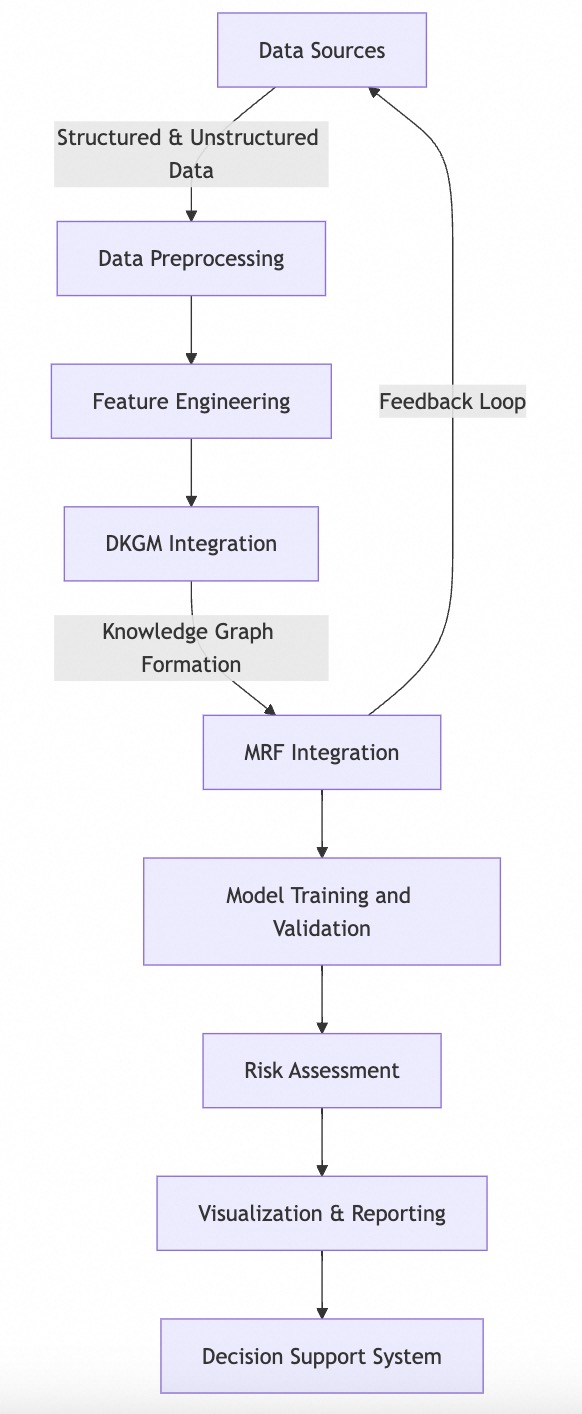
The system architecture is meticulously designed to support our hybrid model. It comprises modules dedicated to data ingestion, pre - processing, knowledge graph construction, MRF configuration, Gibbs sampling, and risk assessment. The architecture is designed to be both flexible and scalable, capable of accommodating the inherent complexity of the model while ensuring optimal efficiency in data processing and analysis. This design allows for seamless integration of new data sources and the adaptation of the model to evolving requirements, ensuring its long-term viability and effectiveness.

Our design also places significant emphasis on the user interface, with the aim of concealing the complexities of the MRF - DKGM integration from the end-user. Through an intuitive interface, users can interact with the system, obtaining valuable insights into the credit risk assessment process. The interface also provides visual representations of the knowledge graph and the dependencies captured by the MRFs, enabling users to better understand the underlying data and relationships. This user-centered approach enhances the usability of the model, making it accessible to a broader range of stakeholders, including financial analysts, lenders, and policymakers.

The design concept is firmly grounded in real-world application scenarios, ensuring that the model is closely aligned with practical requirements. Scenarios such as loan default prediction, credit scoring, and risk monitoring are carefully considered during the design process. By tailoring the model's features and functions to these specific applications, we aim to provide a tool that is not only technically advanced but also highly relevant and useful in the daily operations of financial institutions. This approach ensures that the model can have a tangible impact on the assessment and management of SME credit risk.

#### 3.2.2. System Function Introduction

The system architecture has been meticulously designed to fully exploit the capabilities of a Dynamic Knowledge Graph Model (DKGM) that is seamlessly incorporated with Markov Random Fields (MRF). This integration is deliberately calibrated to improve the credit risk assessment process for Small and Medium-Sized Enterprises (SMEs) in China [87]. Through this approach, it aims to provide a robust and comprehensive framework for analyzing the complex interdependencies and various risk factors inherently present in SME operations. This not only enables a more in-depth understanding of the risk panorama but also furnishes financial institutions with a potent instrument for making more well-informed decisions regarding SME lending and risk management.



*Figure 14 Model system architecture diagram*

Data Sources: The system leverages an extensive array of data sources, which provide both structured and unstructured data. These sources form the fundamental stratum that serves as the input for the DKGM, supplying the raw materials essential for in-depth analysis. The data encompasses financial records, operational metrics, and external environmental factors. Each of these elements is indispensable for comprehensively portraying an SME's financial status and risk profile.

Data Preprocessing: Upon ingestion, the data enters a rigorous preprocessing phase, with the aim of purifying and preparing it for analysis. This phase is of paramount importance in ensuring data quality, addressing issues such as missing values, eliminating duplicates, and standardizing data formats. The objective is to refine the data to a state that is conducive to accurate analysis and modeling.

Feature Engineering: Feature engineering involves the generation of meaningful attributes that encapsulate the financial and operational essence of SMEs. These features are derived from the raw data and are designed to reflect the inherent characteristics and behaviors of SMEs that are indicative of credit risk. The process encompasses both feature transformation and selection, ensuring that the dataset utilized for modeling is not only information-rich but also aligns with the objectives of the DKGM.

MRF Integration: Markov Random Fields are subsequently incorporated into the knowledge graph, introducing a probabilistic layer that models the dependencies among nodes. This integration is pivotal for understanding the risks associated with different nodes and the network as a whole. MRF enables the system to infer the credit risk of an SME based on its relationships with other entities within the graph, thereby providing a more nuanced assessment.

Model Training and Validation: The system employs advanced machine-learning algorithms to train models using data from the knowledge graph. This phase involves iterative learning and validation to enhance the model's predictive accuracy. The model-training process is designed to identify patterns and relationships within the data that signify credit risk.

Decision Support System: The system feeds into a decision-support system, equipping financial institutions with tools to make judicious lending decisions. The integration of DKGM and MRF within this system enables a dynamic and responsive approach to risk management, capable of adapting to new data and changing circumstances in real-time.

Feedback Loop: The system is designed with a feedback loop that channels insights from the decision-support system back to the data sources. This loop facilitates continuous learning and adaptation, ensuring that the system evolves in tandem with the dynamic nature of SME risk profiles.

In conclusion, the integration of MRF within the DKGM represents a significant advancement in the domain of credit-risk assessment for SMEs. It offers a robust, dynamic, and responsive framework that not only improves predictive accuracy but also provides valuable insights into the intricate risk landscape of SME operations in China.

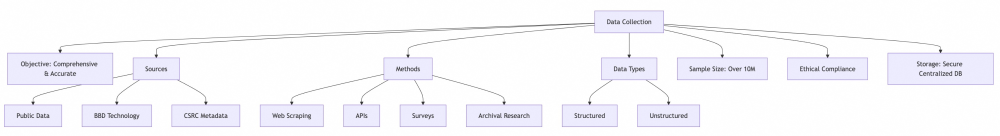
#### 3.2.3. Technical Architecture

Model development and integration are of paramount significance for leveraging data in risk management and other financial applications. Markov Random Fields (MRF), which find applications in statistical physics and computer science, are instrumental in modeling random fields. Within the domain of finance, MRF can effectively portray the dependencies inherent in financial data, such as the correlations among various financial instruments or the propagation of credit risk across a portfolio. Dynamic Knowledge Graph Models are employed to represent entities and their interrelationships within a specific domain, and they possess the ability to update in response to the emergence of novel information. In the context of financial risk management, these models can comprehensively model the relationships among diverse financial entities, including companies, products, and transactions, thereby providing a real-time perspective of the financial ecosystem. The technical architecture is required to seamlessly integrate these models into the operational frameworks of financial institutions. This entails the capacity to deploy models as services accessible via APIs and to integrate the outputs of these models with pre-existing decision-support systems.

Security and compliance assume a position of utmost importance within the fintech architecture, particularly in light of the increasing regulatory scrutiny and the burgeoning cyber threats. The architecture must enforce rigorous data security measures, such as encrypting data both at rest and during transit, implementing secure access controls, and conducting regular security audits. Financial institutions are duty-bound to comply with a multitude of regulations, including the General Data Protection Regulation (GDPR), Know-Your-Customer (KYC), and Anti-Money Laundering (AML) regulations. The technical architecture should support compliance efforts by providing tools for data governance, monitoring transaction patterns for suspicious activities, and generating compliance reports. Incorporating risk assessment tools into the architecture enables the continuous monitoring of compliance risks and the implementation of mitigation strategies. Regulatory Technology (RegTech) solutions can automate compliance processes, thereby reducing the risk of non-compliance and enabling financial institutions to adapt expeditiously to evolving regulatory requirements.

##### 3.2.3.1. Data collection

The process of data collection assumes a position of utmost significance for the Dynamic Knowledge Graph Model (DKGM) in the context of evaluating the credit risks associated with Small and Medium Enterprises (SMEs) in China. Its objective is to amass comprehensive data from a diverse array of sources, with the aim of accurately representing the financial ecosystem of SMEs. This is achieved through the integration of both structured and unstructured data, thereby enabling a holistic perspective.



*Figure 15 Data collection diagram*

The principal objective is to procure a representative sample of SMEs spanning manufacturing, retail, and financial services sectors. The data thus obtained will encapsulate the distinctive characteristics of SMEs that are frequently overlooked by traditional models.

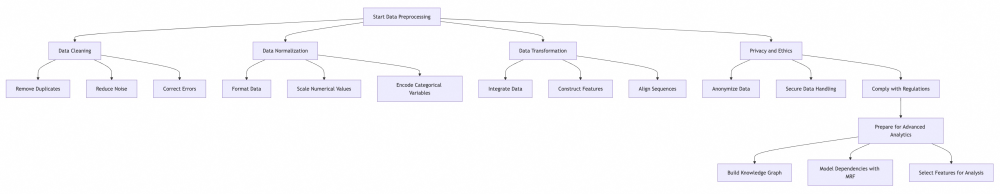
The data sources encompass the national enterprise credit information publicity system, BBD Technology Co. Ltd., and the China Securities Regulatory Commission. Multiple methodologies are employed for data collection, including web scraping, APIs, surveys and interviews, as well as archival research. These methods ensure the collection of both qualitative and quantitative data.

The data types collected are structured (such as financial statements) and unstructured (such as text from reports). The target is to collect data on more than 10 million SMEs, employing stratified and convenience sampling techniques. Ethical guidelines, including data confidentiality and informed consent, are scrupulously adhered to.

The data is stored in a secure, centralized database, which is designed to facilitate efficient retrieval and management. Stringent quality control measures, such as data cleaning, are implemented. In summary, this data collection phase furnishes a comprehensive dataset that is indispensable for the construction of the DKGM. This, in turn, is essential for the development of an accurate credit risk assessment model for SMEs.

##### 3.2.3.2. Data preprocessing

In the evaluation of credit risks associated with Chinese Small and Medium Enterprises (SMEs), data preprocessing assumes a position of paramount importance when employing advanced models such as the Dynamic Knowledge Graph Model (DKGM) and Markov Random Fields (MRF). This section delves into the data preprocessing phase, underscoring its significance in curating a high-calibrate dataset for credit risk analysis.

*Figure 15 Data preprocessing diagram*

The initial step pertains to data cleaning. Owing to the diversity of data sources, the data is prone to inaccuracies. The cleaning process herein encompasses the following aspects. The elimination of records that have the potential to distort the analysis. The utilization of statistical methodologies to expunge outliers. The rectification of errors within critical fields through a combination of automated and manual verification procedures.

Subsequent to cleaning, normalization serves to ensure consistent data formats and scales. This entails the standardization of date, text, and capitalization conventions. The adjustment of numerical values to a common numerical range. The employment of one-hot encoding to facilitate compatibility with machine-learning algorithms.

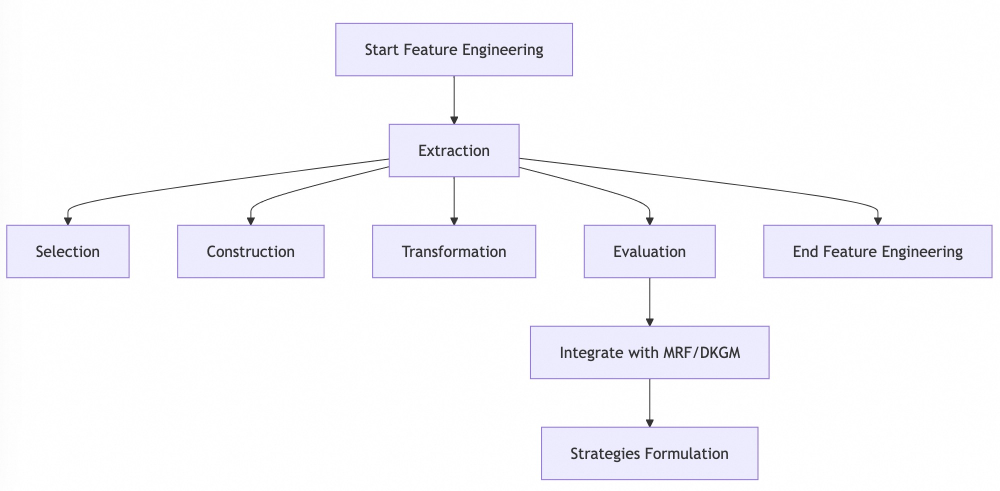
The transformation phase is dedicated to converting data for processing within the framework of DKGM. This involves: The amalgamation of data sourced from disparate origins. The generation of novel features to capture intricate relationships. The alignment of time-series data enables consistent comparison.

Throughout the preprocessing continuum, strict adherence to data confidentiality and anonymity protocols is maintained. Ethical guidelines are scrupulously followed, including the removal or anonymization of Personally Identifiable Information (PII). The prevention of unauthorized access to the data. Adherence to relevant regulations such as the General Data Protection Regulation (GDPR) and Chinese regulatory requirements.

The final stage of preprocessing readies the data for advanced analytics. This encompasses the construction of a dynamic knowledge graph. The utilization of MRF to model variable dependencies. The selection of features is conducive to accurate predictions.

##### 3.2.3.3. Feature Engineering

Feature engineering assumes a pivotal role within the data preprocessing pipeline, situated between data cleaning and model training. Its primary objective is to transmute raw data into a format amenable to modeling, thereby augmenting predictive performance. This section delves into feature engineering in the context of assessing the credit risk of Small and Medium - Sized Enterprises (SMEs) in China, with an emphasis on the creation and selection of features indicative of creditworthiness.



*Figure 16 Feature Engineering diagram. From extraction and selection to construction, transformation, and evaluation, culminating in the integration with advanced models like MRF and DKGM, and finally formulating strategies based on the engineered features.*

In the feature extraction phase, novel features associated with the credit risk status of SMEs are identified and constructed from extant data. Financial ratios such as the debt-to-equity ratio and current ratio are computed to mirror the financial well-being of SMEs. Time-based features, including the seasonality of financial activities or the timing of credit events, are generated. Keywords and topics are extracted from unstructured data sources, such as news and social media, to reflect public sentiment regarding financial stability.

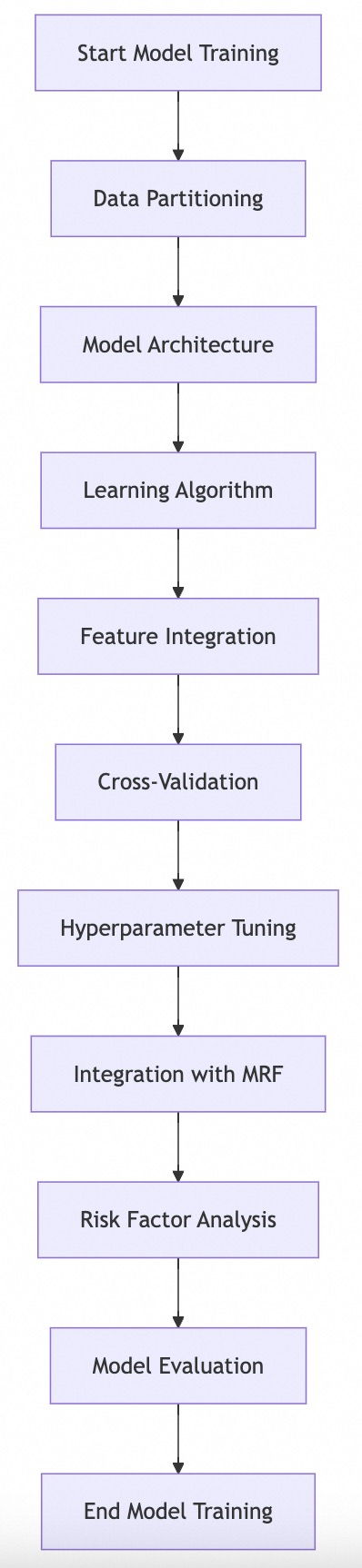
Subsequent to generating a plethora of potential features, relevant ones are selected to enhance the predictive prowess of the model. Features exhibiting weak or non-existent linear relationships with the target variable are filtered out. Principal Component Analysis (PCA) is employed to reduce the feature space while preserving maximum variance. Techniques such as Lasso or Ridge regression are utilized to simplify the model by selecting a subset of features.

Feature construction engenders interactions and composite features. Polynomial terms are generated to capture non-linear relationships among variables. New features representing variable interactions are constructed, which are of significance for the assessment of SME credit risk. Related features are grouped into aggregated features, such as total debt or aggregate income. Feature transformation can enhance model performance. Features are scaled to possess a comparable scale and distribution for algorithms sensitive to input magnitude. Continuous variables are converted into discrete bins to unearth latent patterns. One-hot encoding is applied to categorical variables to ensure compatibility with machine-learning algorithms.

The engineered features are made congruent with the Markov Random Field (MRF) and the Dynamic Knowledge Graph Model (DKGM). Features capture the spatial and temporal correlations that the MRF can model, such as the relationships among financial variables across time and regions. Features are constructed for seamless integration into the dynamic knowledge graph, facilitating the connection of entities and risk factors. Insights gleaned from feature engineering inform risk control and financial management. Features indicative of credit risk assist in formulating targeted risk mitigation strategies. Features prognosticating financial distress enable the establishment of early warning systems for SMEs. Insights from feature engineering guide policy-making decisions concerning SMEs with high-risk profiles.

##### 3.2.3.4. Model Training and Validations

The training and validation of the Dynamic Knowledge Graph Model (DKGM) are of paramount significance for the accurate assessment of credit risks associated with Small and Medium Enterprises (SMEs) in China. This intricate process encompasses multiple sequential steps, each designed to endow the model with robustness, generalizability, and the capacity to encapsulate the complexities inherent in credit risk evaluation.



*Figure 17 Model training and validation diagram. From data partitioning, model architecture design, learning algorithm application, feature integration, cross-validation, hyperparameter tuning, integration with Markov Random Fields, risk factor analysis, to model evaluation.*

During the training phase, the DKGM endeavors to discern patterns germane to credit risk. A comprehensive dataset, sourced from over 10 million SMEs and comprising both structured and unstructured data, is partitioned into training and testing subsets. The former serves as the foundation for model learning, while the latter is reserved for evaluating the model's performance. The DKGM boasts a sophisticated design, featuring layers dedicated to feature extraction, embedding generation, and pattern recognition, thereby enabling it to adeptly handle both data modalities. Employing a supervised learning algorithm, the model's weights and biases are iteratively adjusted in accordance with the training data. A loss function is utilized to quantify the prediction errors, guiding the optimization process. A diverse array of features, such as financial ratios, textual sentiment analysis, and network connectivity metrics, are seamlessly incorporated to construct a comprehensive representation of each SME's credit risk profile.

Cross-validation techniques are employed to guarantee consistent performance across diverse data subsets. The training data is partitioned into (K) non-overlapping folds. The model is then trained and validated (K) times, with each iteration utilizing a distinct fold for validation purposes. This approach mitigates the risk of overfitting. This method ensures that each fold is a representative microcosm of the entire dataset, maintaining the proportional distribution of SMEs across different credit risk levels. Cross-validation serves as a litmus test for the model's generalization ability, evaluating its performance across disparate data distributions.

Hyperparameter tuning is an indispensable component in enhancing the model's predictive accuracy. A systematic exploration of different hyperparameter combinations, including learning rate, batch size, and the number of layers, is conducted. Complementing grid search, random search facilitates the discovery of optimal hyperparameter combinations that may be overlooked by the former method. Metrics such as accuracy, precision, recall, and F1 - score are employed to rigorously evaluate the impact of hyperparameter variations on the model's performance.

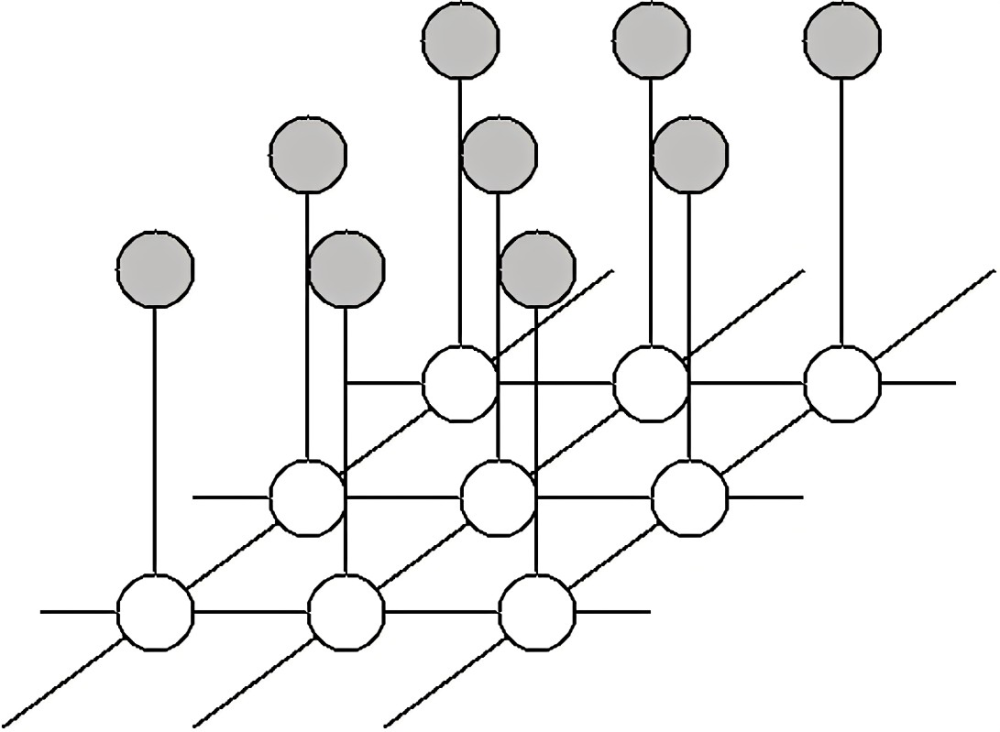
The integration of Markov Random Fields (MRF) confers additional capabilities upon the DKGM. MRF enables the modeling of variable dependencies, thereby elucidating the intricate relationships between financial and external factors. By analyzing the geographical distribution of credit risks, MRF provides valuable insights into the spatial dimension of risk propagation. MRF effectively captures the time-based evolution of credit risks, facilitating the prediction of future events.

The DKGM's risk factor analysis encompasses the following aspects. The influence of macroeconomic variables, such as GDP growth, interest rates, and unemployment rates, on SME credit risks is comprehensively considered. The mechanism by which credit risks propagate among affiliated enterprises is investigated, shedding light on the interconnectedness of the SME ecosystem. Acknowledging the dynamic nature of credit risks as SMEs progress from the startup phase to maturity, the model accounts for the associated variations.

Model evaluation is conducted using a suite of metrics. Defined as the proportion of correct predictions, accuracy serves as a fundamental measure of the model's performance. Precision quantifies the model's ability to accurately identify SMEs with credit risk, emphasizing the correct classification of positive instances. Recall measures the model's capacity to capture all credit-risk cases, highlighting its sensitivity in detecting potential risks. The F1-score, a harmonic mean of precision and recall, provides a balanced assessment of the model's performance. The Area Under the Curve (AUC) metric gauges the model's discriminative power, evaluating its ability to distinguish between positive and negative instances. The Kolmogorov-Smirnov (KS) statistic assesses the model's effectiveness in ranking credit risks, providing insights into its discriminatory performance.

##### 3.2.3.5. Markov Random Field

The integration of MRF into credit risk assessment models, especially for Chinese SMEs, is a major step forward in financial analytics. MRF, a probabilistic model of undirected graphs where nodes are random variables and edges show their dependencies, has a key feature: nodes are conditionally independent given their neighbors. This simplifies the model's complex interdependencies, making it more suitable for credit risk analysis.



*Figure 18 Graphical representation of a Markov Random Field*

The joint probability distribution of a MRF can be expressed as:

P(X)= (1)

where X is the set of random variables, Z is the partition function, ensuring the probabilities sum to one,are the potential functions associated with each clique C in the graph, andrepresents the subset of variables in clique C.

The partition function Z is defined as:

 (2)

This function normalizes the probabilities, ensuring they sum to one across all possible configurations of the random variables.

Markov Random Field (MRF) can be visually represented as a graph, wherein nodes denote financial variables such as credit ratings, and edges represent the dependencies among these ratings. The graph-theoretic structure facilitates the identification of clusters or communities within the financial data, which is particularly illuminating for credit risk assessment. In the context of Chinese Small and Medium-sized Enterprises (SMEs), MRF provides a robust framework for modeling the intricate intra- and inter-enterprise dependencies. This model has the ability to capture the influence of macro-economic factors, industry-specific trends, and the interconnectedness of financial well-being across diverse enterprises. The dependencies within the MRF model can be utilized to model the likelihood of credit events, such as defaults or downgrades. For instance, the probability of a credit event for a specific SME can be influenced by the financial status of its industry peers and macro-economic indicators. By integrating spatial and temporal dimensions, MRF is capable of capturing the evolution of credit risk over time and across different regions. This is particularly valuable for understanding how economic cycles or regional financial crises impact the creditworthiness of SMEs. The MRF model complements the Dynamic Knowledge Graph Model (DKGM) by providing a probabilistic framework that can integrate both structured and unstructured data. The DKGM, which leverages advanced data-crawling and natural-language-processing techniques [88,89,90], can benefit from the probabilistic inferences made by MRF.

In the context of credit risk, the potential functions  can be designed to reflect financial metrics such as debt-to-income ratios, interest coverage ratios, and asset quality. The edges in the MRF graph can represent the correlations between these metrics across different SMEs. The Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method commonly used within MRF to estimate the joint distribution of the random variables. It operates by iteratively sampling each variable based on its conditional distribution given the current state of the other variables.

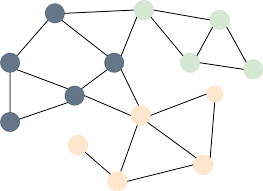
 (3)

where is the state of the system at time t.

MRF comprehensively explores risk factors by modeling complex ties among financial indicators and external economic variables. It covers how interest rates, exchange rates, and market volatility affect SMEs' credit risk. By grasping dependencies and correlations in Chinese SMEs' financial ecosystem, MRF boosts credit risk models' predictive power. Using its probabilistic framework and graphs, MRF offers a refined understanding of credit risk, helping financial institutions make better decisions. Applying MRF to credit risk assessment, especially for dynamic and complex Chinese SMEs, shows its potential as a strong analytical tool. Its ability to model complex interactions and spatiotemporal dynamics makes it essential for credit risk assessment.

##### 3.2.3.6. Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as a powerful paradigm for learning on graph-structured data [91,92]. GNNs are designed to capture the intricate patterns and relationships inherent in graph data, making them particularly suited for tasks such as node classification, graph classification, and link prediction. This chapter provides a comprehensive overview of GNNs, delving into their conceptual foundations, algorithmic principles, and mathematical formulations.



*Figure 19 Graphical representation of a Graph Neural Network*

At the core of Graph Neural Networks (GNNs) lies the graph, a fundamental data structure consisting of nodes (vertices) and the edges (links) that connect these nodes. GNNs are designed to learn the representations of nodes or entire graphs through the aggregation of information from their local neighborhoods. Within this framework, there exist several pivotal components. Nodes serve as symbols of entities within the graph. Edges denote the relationships between nodes. Features, which are associated with both nodes and edges, furnish supplementary information. The graph structure pertains to the configuration of nodes and edges, encapsulating the topological characteristics of the graph. GNNs operate based on the principle of information propagation. In this process, nodes update their states by aggregating information from their neighboring nodes. This operation is iterative, continuing until it converges to a stable state.

The core mechanism of information propagation in GNNs can be described by the following steps. Message Passing: Each node aggregates information from its neighbors. Aggregation: The aggregated information is combined using various operations (e.g., sum, mean, max). Update: Nodes update their states based on the aggregated information. Let represent a graph with node set V and edge set E. Let X be the feature matrix whererepresents the features of node v. THe information propagation process in GNNs can be formulated as:

 (4)

where is the hidden state of nodes at iteration k. Theis an aggregation function over the neighborhood N of each node. Theis a non-linear activation function. GNNs rely on the concept of a fixed point to ensure convergence. According to the Banach fixed-point theorem, if the update function is a contraction mapping, the sequence of hidden states will converge to a unique fixed point.

GCNs are a variant of GNNs that utilize spectral graph theory to define convolutional operations on graphs. The key idea is to represent the graph convolution as a linear combination of the graph's eigenvectors. The graph Laplacian L is central to GCNs. The normalized Laplacian is given by:

 (5)

where A is the adjacency matrix and D is the degree matrix. The graph convolution operation in GCNs is defined as:

 (6)

where W is the learnable weight matrix, is the activation function. To avoid the computational cost of eigendecomposition, Chebyshev polynomials are used to approximate the spectral convolution:

 (7)

where represents the Chebyshev polynomials of order K.

Graph Neural Networks (GNNs) have found extensive applications across diverse domains [94,95]. This task pertains to the categorization of nodes by leveraging both their inherent features and the information from neighboring nodes. For instance, within social network graphs, individuals (represented as nodes) can be classified according to their own attributes as well as those of their friends. Here, the objective is to classify entire graphs. A case in point is the application in drug discovery, where molecular graphs are classified based on their structural and chemical properties to identify potential drug candidates. This involves predicting the probability of the existence of links between nodes. In financial networks, it can be utilized to predict business partnerships or credit relationships. In response to the multifaceted challenges encountered in graph-related tasks, GNNs have been extended through the following approaches. By integrating attention mechanisms, GATs assign weights to neighbors, thereby facilitating more effective information aggregation. GraphSAGE aggregates information through sampling and neighborhood aggregation techniques, enabling efficient large-scale graph learning. These are designed to adapt to dynamic graphs, where the nodes and edges undergo changes over time, as is the case in real-time traffic network graphs.

GNNs represent a substantial advancement in the realm of machine learning [95], particularly for tasks involving graph-structured data. Their capacity to capture local and structural information endows them with remarkable versatility. As the research landscape evolves, GNNs are expected to assume an increasingly pivotal role in data-driven decision-making processes across a wide spectrum of fields.

#### 3.2.4. Dynamic Knowledge Graph Model for Financial Engineering

The Dynamic Knowledge Graph Model (DKGM) embodies a cutting-edge methodology within the domain of financial engineering, with particular salience for undertakings such as credit risk evaluation [96], fraud detection [97], and investment analysis [98]. This model exploits the structural and temporal complexity intrinsic to graph-structured data, thereby furnishing a comprehensive vantage point of financial entities and their interrelations. Differentiating itself starkly from traditional static knowledge graphs, the DKGM integrates time-variant data and adaptive learning mechanisms. These components are fundamental for encapsulating the dynamic nature of financial markets, wherein conditions are perpetually in a state of transformation. In the context of financial engineering, a DKGM is formulated by representing financial entities (including companies, individuals, and transactions) as nodes, and their relationships (such as ownership, transactional linkages, and similarities) as edges. The model evolves over time, reflecting changes in the financial ecosystem. Such alterations may encompass fluctuations in asset prices, regulatory amendments, or the advent of novel market participants.

##### 3.2.4.1. Dynamic Knowlege Graph Model Preview

The Dynamic Knowledge Graph Model (DKGM) demonstrates substantial value in appraising the development of Small and Medium-sized Enterprises (SMEs) through an in-depth analysis of network structure and behavior. This capability enables more precise credit risk prediction, which is of critical importance to financial institutions and policymakers, particularly during economic upheavals such as the COVID-19 pandemic. Representing a paradigm shift within the domain of financial engineering, the DKGM proffers a dynamic approach to comprehending financial systems by effectively modeling the ever-evolving complexity inherent in financial data.

Financial data, by its very nature, is time-varying. The time-varying features, influenced by a confluence of internal and external factors, are indispensable for accurate financial forecasting and risk management. These features encompass economic indicators, market sentiment, regulatory changes, global events, and company-specific information. Such features exert a significant impact on financial predictions. For instance, alterations in economic indicators can serve as harbingers of business-cycle transitions, thereby influencing investment strategies. A high level of market sentiment can enhance confidence, whereas a negative sentiment can precipitate market downturns.

Within the framework of the DKGM, time-varying features are seamlessly integrated to construct a dynamic perspective of the financial ecosystem. By continuously updating the graph with the most recent data, the DKGM is capable of capturing the contemporaneous state of financial markets. It conducts a comprehensive analysis of SME network structures across diverse industries, taking into account macroeconomic factors, risk transmission mechanisms, and development stages. This process facilitates the identification of potential risks and the assessment of SMEs, which is fundamental for credit risk evaluation.

In summary, time-varying features play a pivotal role in financial systems. The DKGM's proficiency in handling these features confers a competitive advantage in financial engineering, presenting a potent instrument for navigating the complexities of modern finance.

##### 3.2.4.2. Theoretical Foundations of DKGM

**3.2.4.2.1. Graph Theory Basics**

Graph theory provides a systematic approach to model complex relationships and interactions within a network. It is particularly useful in understanding the structure and evolution of financial networks. A graph G is defined as an ordered pair , where V is a finite set of vertices (or nodes) and E is a finite set of edges (or links). Each edge  connects a pair of vertices where. In financial networks, vertices represent financial entities such as banks, institutions, or markets, and edges represent relationships such as transactions, loans, or ownership ties.

Graphs can be classified as directed or undirected based on the nature of the edges. In an undirected graph, the edge  implies a mutual relationship between vertices u and v. However, in a directed graph, the edge indicates a one-way relationship from u to v. Financial networks often involve directed graphs to represent directional relationships such as cash flows or debt obligations. Graph properties and metrics are essential for analyzing the structure and dynamics of financial networks. Centrality measures, such as degree centrality, betweenness centrality, and eigenvector centrality, indicate the importance of a node within the network. The clustering coefficient quantifies the degree to which nodes in a graph tend to form tightly-knit groups.

Graph theory is a fundamental branch of mathematics that provides a framework for modeling and analyzing complex systems. In the context of financial networks, graph theory enables the representation of entities and their relationships in a structured and mathematically tractable way. At the core of graph theory are nodes (or vertices) and edges (or links). Nodes represent individual entities within a financial network, such as banks, investors, or financial instruments. Edges represent the relationships between these entities, which can be transactions, ownership stakes, or other forms of financial interactions. A node is a fundamental unit of a graph, representing an entity. An edge is a connection between two nodes, indicating a relationship. The degree of a node is the number of edges incident to it, reflecting the entity's connectivity within the network. Degree(*v*) indicates the number of edges connected to node *v.*

Paths and cycles are essential concepts in graph theory that describe the traversal and structure within a graph. A path is a sequence of nodes and edges that connects two nodes without any repetition. The length of a path is the number of edges it contains. A cycle is a special type of path that starts and ends at the same node, forming a loop. A graph is said to be acyclic if it contains no cycles. Connectivity in graph theory refers to the ability to travel between any two nodes in the graph. A graph is connected if is a path between every pair of nodes. For financial networks, connectivity is crucial as it reflects the robustness and resilience of the financial system.

Graphs can be represented in various ways, with the most common being adjacency matrices and adjacency lists. An adjacency matrix is a square matrix used to represent a finite graph. The rows and columns of the matrix represent nodes, and the entries indicate the presence or absence of edges. An adjacency list is an array of lists or sets used to represent the edges of a graph. Each list describes the set of nodes that are adjacent to a particular node. The formula for the Adjacency Matrix can be written as:

 (8)

Beyond the foundational concepts, there are more advanced constructs in graph theory that provide deeper insights into the structure and properties of graphs. Subgraphs and isomorphism are important concepts for understanding the composition and equivalence of graphs. A subgraph is a graph formed by a subset of vertices and a subset of edges from the original graph. Subgraphs are used to study specific structures within a larger graph. Graph isomorphism is a relation between two graphs that indicates they have the same structure. Two graphs are isomorphic if there exists a bijection between their vertex sets that preserves adjacency. The formula for Isomorphism can be written as:

 (9)

Trees and forests are special types of graphs that have unique properties and applications in various fields, including finance. A tree is an undirected graph in which any two nodes are connected by exactly one path. Trees are used to model hierarchical structures and are fundamental in many graph algorithms. A forest is a disjoint union of trees. In financial networks, forests can represent decentralized systems where each tree corresponds to a distinct subgroup or market segment. A tree with n nodes has n−1 edges. There are no cycles in a tree. The formula for the number of edges in a tree is:

 (10)

The graph theory provides a robust framework for analyzing the structure and dynamics of financial networks. By understanding the foundational concepts and advanced constructs, we can develop more sophisticated models and algorithms for financial analysis and decision-making. The integration of machine learning with graph theory, as discussed in the previous section, further enhances our ability to extract insights from complex financial data.

**3.2.4.2.2. Temporal Dynamics in Financial Networks**

The temporal dynamics of financial networks pertain to the modifications in the structures and compositions of financial relationships over time. These connections are not static; instead, they undergo transformations in response to economic cycles, market trends, and regulatory changes. The time-varying nature of financial links can be modeled using time-series analysis or dynamic network models. These models are capable of elucidating the formation, dissolution, and reinforcement processes of financial relationships. Market conditions exert a profound influence on the evolution of financial networks. During periods of market stress, financial institutions typically mitigate their risk exposure by tightening credit conditions or reducing their interconnections. Conversely, in bullish markets, the expansion of credit and increased risk-taking tend to result in a denser network structure, characterized by more extensive financial linkages.

Stochastic Actor-Oriented Models (SAOM) represent a class of statistical models that are specifically designed for analyzing the dynamic characteristics of social networks. These models play a particularly crucial role in elucidating the temporal evolution of relationships among financial entities, such as banks or investors. Within the SAOM framework, the network is conceptualized as an aggregation of actors, whose decisions have a significant impact on the topological structure of the network. It is postulated that actors within the network are capable of modifying their relationships in consonance with their individual preferences and the current state of the network. The model operates by leveraging a series of network snapshots obtained at different time intervals. Through this approach, it can accurately estimate the effects of various factors on the formation and dissolution of ties. This not only enables a comprehensive comprehension of how different elements contribute to the ever-changing nature of financial relationships but also provides a robust and systematic framework for analyzing the temporal evolution of financial networks.

In financial networks, SAOMs can be used to analyze the evolution of relationships such as credit default swaps, ownership ties, or trade relationships relationships. These models can help identify how market conditions influence the formation of new financial relationships and the dissolution of existing ones. SIENA (Simulation Investigation for Empirical Network Analysis) is a specific implementation of SAOM that is widely used for dynamic network analysis. SIENA allows researchers to estimate the effects of various covariates on the formation and dissolution of ties, taking into account the network's structure and the temporal order of observations. The formula of the Tie can be written as:

 (11)

whereis the effect of the covariateon the probability of tie formation between actions i and j at time t.

Temporal network data, procured by observing networks at multiple time points, assumes a pivotal role in comprehending the dynamics of financial networks. This data can be amassed through means such as surveys and transaction records. In the realm of finance, loan, investment, and trade data accrued over time constitute network snapshots. The methods for data collection are as follows. At each time point, actors are randomly selected to represent the network states. Data collection is triggered by significant events such as new loan disbursements or trade occurrences. The same set of actors is monitored over time to enable an in-depth analysis of individual-level changes.

The collection of such data, particularly within large-scale financial networks, is an intricate and resource-intensive endeavor. A trade-off exists between the frequency of observation and the duration of the observation period. Empirical investigations into the evolution of financial networks elucidate the manner in which these networks adapt to economic, regulatory, and technological shifts. Longitudinal analysis, which involves observing the identical network over an extended period, utilizes data sourced from surveys or continuous transaction tracking mechanisms. Financial networks are subject to alterations in their core-periphery structures, and the emergence of new institutions or products has the potential to reshape them.

Techniques employed in longitudinal analysis, including graphical visualization, statistical hypothesis testing, and machine-learning models (such as community detection algorithms), are instrumental in discerning changes in network structure. Gaining an understanding of these longitudinal changes is of utmost importance for regulators and policymakers in their efforts to safeguard financial stability. By meticulously tracking the evolution of financial networks, they can identify systemic risks at an early stage and take pre-emptive measures. The temporal dynamics of financial networks are intricate yet indispensable. Employing models such as Stochastic Actor-Oriented Models (SAOM) and Simulation Investigation for Empirical Network Analysis (SIENA), in conjunction with empirical research, enables us to glean insights into changes in financial relationships. This understanding is fundamental for grasping network-related market stability and resilience.

**3.2.4.2.3. The Role of Machine Learning in DKGM**

Adaptive learning algorithms within the Dynamic Knowledge Graph Model (DKGM) dynamically adjust their parameters in response to new data influx. These algorithms play an indispensable role in the modeling of dynamic financial networks, enabling the model to adeptly adapt to the ever-shifting market conditions and evolving risk profiles. Machine learning, when integrated into the DKGM framework, serves as a powerful enabler for pattern recognition and prediction. By leveraging extensive financial market datasets, it uncovers intricate patterns that remain concealed from traditional analytical methods. These patterns are invaluable for predicting market trends, evaluating credit risks, and optimizing investment strategies.

Pattern recognition in this domain predominantly employs machine-learning techniques such as clustering, classification, and dimensionality reduction. Unsupervised learning algorithms, for instance, the K-means algorithm, can effectively group similar financial activities. In contrast, supervised learning algorithms like support vector machines are utilized to classify financial entities. Predictive analytics in the DKGM framework utilizes historical data to generate forecasts. Through this approach, models can project stock prices, anticipate interest rate fluctuations, or predict credit defaults. For example, autoregressive integrated moving average (ARIMA) models or long short-term memory (LSTM) neural networks can predict financial time - series by discerning patterns from past trends. Machine-learning algorithms are of utmost importance for graph data analysis in the DKGM. They are capable of identifying latent patterns, making accurate predictions, and extracting profound insights from the relationships within financial networks.

In the presence of labeled data, supervised learning algorithms are deployed for graph analysis. The regression analysis is applied for predicting continuous variables such as transaction probabilities or investment returns. Linear regression is employed to model the relationships between financial variables and graph-based metrics, while support vector regression (SVR) is utilized to handle non-linear relationships. The classification tasks involve categorizing financial entities, such as differentiating between solvent and insolvent firms or classifying investments as high-risk or low-risk. Logistic regression, decision trees, and random forests are commonly used algorithms. Logistic regression, in particular, classifies entities based on their attributes and positions within the network. The formula for logistic regression can be written as:

 (12)

whererepresents model parameters, and X represents the features extracted from the graph data.

Unsupervised learning algorithms are employed in scenarios where the data lacks labeled outcomes, serving to unearth latent structures and patterns. Entities within the financial domain are grouped according to their attributes and network connections. K-means clustering partitions entities on the basis of both network characteristics and attributes, while hierarchical clustering elucidates the hierarchical nature of networks. Association rule learning aims to identify recurrent patterns among financial variables. The FP - Growth algorithm is particularly effective in mining such patterns within large-scale datasets, especially those pertaining to financial transaction data.

Machine learning significantly enhances the capabilities of DKGM. Feature engineering is a crucial process that extracts relevant features from raw data. Nodes are transformed into low-dimensional vectors while meticulously preserving the network relationships. Graph structures are encoded in a format suitable for machine-learning algorithms. Time-related features are incorporated to capture the dynamic nature of networks. DeepWalk and Node2Vec utilize random walks and skip-gram models to generate vector representations of nodes. Graph Convolutional Networks (GCNs) aggregate information from neighboring nodes to construct local graph vectors. Random Walk Kernels measure the similarity between graphs through the analysis of random walks, and the WL Kernel creates graph signatures by hashing node labels.

In the context of credit risk assessment within financial networks, the processes of training and validating models are of utmost importance. Cross-Validation serves as a means to evaluate the generalization ability of models, which is of critical significance in the financial realm due to the limited availability of data. Metrics such as precision, recall, F1 score, and AUC-ROC are commonly utilized to assess the performance of credit risk models.

##### 3.2.4.3. Data Integration in DKGM

The Dynamic Knowledge Graph Model (DKGM) represents a highly sophisticated framework designed for the assessment of credit risks associated with Small and Medium-sized Enterprises (SMEs) in China. Its nucleus resides in the integration of diverse data sources to construct a graph that effectively mirrors the complexity of the financial ecosystem. This section delves into the theoretical underpinnings and practical implementation of data integration within the DKGM, with a particular emphasis on data sources, preprocessing methodologies, and data enrichment strategies.

DKGM draws upon multiple data sources for the construction of its graph. Stock market data, encompassing elements such as prices, trading volumes, and indices, serves as a crucial resource for discerning the financial performance of SMEs and gauging market sentiment. Unstructured data gleaned from news platforms and social media, which is processed through Natural Language Processing (NLP) techniques, offers valuable insights into the operational aspects and reputation of SMEs. Satellite imagery, geolocation data, and data from the Internet of Things (IoT) provide a distinctive perspective on the operational activities of SMEs. Financial statements and regulatory filings are instrumental in revealing the financial health and compliance status of SMEs. Data sourced from financial institutions, including loan and payment records, play a significant role in the assessment of SMEs' creditworthiness.

Temporal data necessitates specialized preprocessing within the context of DKGM. The data cleaning process involves identifying and rectifying errors, eliminating duplicates, and imputing missing values while preserving the integrity of the time-series data. Metrics are scaled to enable a fair comparison among SMEs. The temporal alignment ensures that data from disparate sources is temporally aligned, with resampling being carried out when required. The trend and seasonality adjustment is employed to isolate cyclical and irregular components within the data. For the integration of heterogeneous data, formats are standardized, and differences are reconciled to facilitate the combination of diverse data types.

Enriching DKGM with external data significantly enhances its predictive capabilities. Information obtained from company websites and online forums can unveil customer satisfaction levels and market trends. Analyzing social media posts allows for the measurement of public sentiment towards SME products, which has implications for their financial stability. Geospatial data provides insights into location-related operational factors. Macroeconomic data offers a broader economic backdrop against which SME performance can be evaluated. Collaboration with data providers affords a more comprehensive understanding of SME operations.

The data integration process within DKGM encompasses: data are procured from various sources through means such as Application Programming Interfaces (APIs), web scraping techniques, or data dumps. Raw data is stored in a data lake or warehouse capable of accommodating different data types. Extract, Transform, and Load (ETL) processes are utilized to clean, transform, and integrate data for utilization within the DKGM. Checks are implemented to ensure data accuracy, completeness, and consistency. Data is regularly updated to reflect the most recent trends.

##### 3.2.4.4. Graph Construction in DKGM

Nodes and edges play a pivotal role in the construction of graphs within the framework of Dynamic Knowledge Graphs for Finance (DKGM). Nodes serve to represent financial entities such as companies, stocks, sectors, or economic indicators, endowed with attributes including market capitalization, revenue, and financial ratios. Edges, on the other hand, signify diverse relationships, ranging from ownership to stock price co - movement, thereby constituting the fundamental structure of the graph that is essential for comprehending the financial ecosystem.

The construction of nodes and edges in a financial graph is an algorithm-driven process. The Entity-Relationship Model (ERM) establishes the schema, precisely defining the entities and their interrelationships. Natural Language Processing (NLP)-based extraction techniques, employing Named Entity Recognition (NER) and Relation Extraction (RE), are utilized to identify entities and relationships from unstructured data sources. Subsequently, graph-construction algorithms such as the RDF triple-store or property-graph models are employed for the creation and efficient storage of the graph. The attributes of nodes and edges are derived from multiple sources: financial databases such as Bloomberg; SEC company filings; news and media outlets, which provide insights into company developments; and social media platforms, which offer valuable information regarding market sentiment.

Financial relationships are inherently dynamic. Temporal-graph models, such as multi-layer graphs or dynamic-graph models that update edge weights, are employed to capture these temporal changes. Streaming algorithms and machine-learning models based on online learning facilitate the incorporation of new data and the prediction of relationship changes. Nevertheless, the evolution of graphs in this context encounters several challenges. The noise present in financial data can lead to false positives during edge creation, and scalability issues emerge as the number of nodes and edges expands. Thus, robust data-cleaning procedures, the development of summary graphs, and graph-simplification techniques are imperative. Graph-aggregation, encompassing both structural and attribute-based methods, serves to combine nodes and edges for a higher-level representation of the graph.

Aggregation is indispensable for macro-trend analysis, as it effectively reveals sector growth patterns and market concentration. Conversely, disaggregation is essential for micro-analysis, enabling the decomposition of aggregated elements to facilitate the study of individual entities. Effective graph analysis in the field of finance necessitates a delicate balance between aggregation and disaggregation. Interactive graph-exploration tools empower users to toggle between these two perspectives. The graph construction and evolution processes within DKGM, along with the complementary operations of aggregation and disaggregation, offer a multi-scale view of financial data, thereby providing significant assistance in financial decision-making processes.

##### 3.2.4.5. Feature Engineering for DKGM

Feature engineering is a critical phase in the construction of the Dynamic Knowledge Graph Model (DKGM), focusing on extracting and selecting features that are most relevant to predicting credit risk. This process involves transforming raw data into a set of meaningful features that capture the inherent patterns and signals within the data.

Temporal feature extraction involves extracting features that capture the temporal aspects of financial data. These features are essential for understanding the dynamics of financial entities and their interactions over time. Seasonality and trend components are fundamental in time series data, representing regular and persistent patterns in data. For instance, retail sales often exhibit seasonal patterns such as increased sales during holiday seasons. Similarly, trends can indicate long-term movements in data, such as a sustained increase in a company's revenue over years. The formula of seasonal and trend decomposition can be written as:

 (13)

whereis the observed value at time t, is the trend component at time t, is the seasonal component at time t, is the cyclical component at time t, and is the irregular component at time t.

The lagged variables are created by shifting the time series data, where a value at time t is used to predict the value at time t+1, t+2, etc. These lags help in capturing the autocorrelation structure of the data. The lag feature is the value of the variable k periods before time t. Rolling window statistics, such as the mean and standard deviation, capture the recent behavior of the data, providing a dynamic view of the financial metrics. The formula of rolling mean can be written as:

 (14)

whereis the mean at time t, and W is the window size. Feature selection for time-series data involves identifying the most relevant features that contribute to the predictive power of the model. Autocorrelation measures the correlation of a variable with its own lagged values, while partial autocorrelation considers the effect of intermediate lags. The formula of autocorrection function (ACF) can be written as:

 (15)

where is the autocorrelation at lag k, andis the observed value at time t, and is the mean of the time series.

Information criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are used to select the optimal model order by balancing goodness of fit and model complexity. The formula of AIC can be written as:

 (16)

where k is the number of parameters in the model, and L is the maximum value of the likelihood function for the model. Machine learning algorithms, especially ensemble methods like Random Forest or Gradient Boosting Machines, can provide feature importance scores, indicating the contribution of each feature in the prediction. Non-stationary data presents challenges due to changing statistical properties over time. Addressing non-stationarity is crucial for building robust time series models. Differencing involves calculating the difference between consecutive observations to remove trends and seasonality, making the data stationary. The First difference formula can be written as:

 (17)

Transformations such as logarithmic, square root, or Box-Cox transformations can stabilize variance and make the data more stationary. The Logarithmic Transformation can be written as:

 (18)

Seasonal adjustment techniques, such as X-13ARIMA-SEATS or STL (Seasonal and Trend decomposition using Loses), are used to remove seasonal effects and trends from the data. The SLL decomposition can be written as:

 (19)

whereis the seasonal component, is the trend component, is the remainder (irregular component), and is the calendar effect (e.g., holidays).

Feature engineering for DKGM involves a systematic approach to extracting and selecting features that capture the temporal dynamics of financial data. By leveraging techniques such as seasonal decomposition, lag features, and addressing non-stationarity through differencing and transformation, the DKGM can effectively model the complex relationships within financial data. These methods not only enhance the model's predictive power but also provide insights into the underlying patterns governing the financial ecosystem.

### 3.3. MRF-Enhanced Dynamic Knowledge Graph Model

The interaction between Markov Random Fields (MRF) and the dynamic knowledge graph within AI systems represents a highly synergistic integration. MRFs are employed to model the probabilistic dependencies among random variables, while the dynamic knowledge graph serves to represent the evolving domain knowledge. MRFs, as undirected graphical models, are utilized to encode joint probability distributions. They prove invaluable for capturing statistical regularities within data, such as in the case of images or text. In an MRF, nodes correspond to random variables, and edges signify conditional dependencies, thereby enabling inference, image segmentation, and prediction tasks.

Dynamic knowledge graphs constitute an extension of traditional knowledge graphs, with their structure and node/edge information undergoing temporal evolution. This characteristic renders them well-suited for applications that demand adaptation to novel data or contexts, as they are capable of representing diverse entities and relationships.

The interaction between MRF and the dynamic knowledge graph transpires through a mutually beneficial process. MRFs are instrumental in modeling the data properties embedded within the knowledge graph. For instance, in the realm of image recognition, MRFs assist in discerning spatial features and encoding them into the graph. The inferences derived from MRFs guide the restructuring of the graph, facilitating a more effective representation of the data distribution. MRFs offer a probabilistic mechanism for updating the graph as new data becomes available. MRFs are utilized to quantify the uncertainty inherent in the knowledge graph, thereby enabling the system to make well-informed decisions. The dynamic knowledge graph capitalizes on the predictive capabilities of MRFs, particularly in applications such as financial risk assessment.

In summary, MRFs and the dynamic knowledge graph effectively amalgamate their respective strengths. MRFs furnish a probabilistic approach for characterizing data relationships, while the dynamic knowledge graph provides a structured and evolving means of knowledge representation. In concert, they contribute to enhancing the reasoning ability, adaptability, and decision-making prowess of AI systems under conditions of uncertainty.

#### 3.3.1. Model Representation

##### 3.3.1.1. Nodes and edges in Dynamic Knowledge Graph

In the realm of dynamic Knowledge Graphs (KGs), nodes and edges play pivotal roles in representing complex, interrelated data structures that evolve over time. This section delves into the technical intricacies of defining entities and relationships within dynamic KGs, with a particular focus on Small and Medium Enterprises (SMEs) and their associated financial indicators.

Entities in a Knowledge Graph are the fundamental units that represent real-world objects or concepts. In the context of SMEs and financial indicators, entities can be broadly categorized as SMEs: These are the primary entities representing individual businesses. Each SME entity can possess attributes such as:

* Identifier: A unique ID (e.g., SME\_12345)

* Name: The official name of the enterprise

* Industry: The sector in which the SME operates (e.g., manufacturing, services)

* Location: Geographical information (e.g., city, country)

* Size: Metrics like the number of employees or annual revenue

The financial indicators will enable entities to encapsulate various financial metrics pertinent to SMEs, including:

* Revenue: Total income generated

* Profit Margin: Percentage of revenue that translates to profit

* Assets: Total assets held by the SME

* Liabilities: Outstanding debts or obligations

Temporal Entities: Given the dynamic nature of KGs, temporal entities such as Time Points (e.g., Q1 2023) are essential for capturing the evolution of relationships and attributes over time. Each entity is instantiated as a node within the KG, enriched with properties that define its state at any given temporal snapshot.

Relationships (or edges) in a Knowledge Graph denote the interactions or associations between entities. In dynamic KGs, these relationships are inherently temporal, allowing the graph to reflect changes over time. Key relationships in the SME-financial indicator context include:

* The ownership connects SMEs to their owners or parent companies. The formula representation is [MISSING IMAGE: , ]

* The financial reporting links SMEs to their financial indicators for specific time periods. The formula representation is[MISSING IMAGE: , ], where t denotes the temporal aspect, ensuring the relationship captures the financial metric at time t.

* Industry association associates SMEs with the industries they belong to. The formula representation is [MISSING IMAGE: , ]

* Geographical location connects SMEs to their locations, which may change over time due to expansion or relocation. The formula representation is [MISSING IMAGE: , ]

##### 3.3.1.2. Markov Random Field Structure within Knowledge Graph

Integrating Markov Random Fields (MRFs) within Knowledge Graphs (KGs) proffers a robust and sophisticated framework for modeling intricate dependencies and augmenting inferential capabilities. This section delves into the structural integration of MRFs into KGs, with a specific emphasis on the mapping of MRF nodes onto the KG and the formulation of potential functions that encapsulate inter-node dependencies.

In the realm of KGs, entities and their relationships constitute the bedrock of the fundamental structure. An MRF, as a probabilistic graphical model, supplements this by capturing the conditional dependencies among these entities. The mapping process entails representing each entity and relationship within the KG as nodes in the MRF, thereby harnessing the strengths of both paradigms. Each entity in the KG is mapped onto a node in the MRF. For instance, consider a KG that represents diverse proteins within a biological network. Each protein entity is transformed into a corresponding node in the MRF, encapsulating its state or attributes, such as expression levels or interaction states. This mapping enables a more nuanced and in-depth understanding of the individual components within the knowledge graph, as the MRF can model the probabilistic behavior of each entity with precision. Relationships in the KG, which denote interactions or associations between entities, are represented as edges in the MRF. These edges signify the conditional dependencies between the nodes (entities). For example, an edge between two protein nodes in the MRF indicates a potential interaction or regulatory relationship that exerts an impact on their states. By translating KG relationships into MRF edges, the integrated model can more comprehensively analyze how different entities influence one another, thereby enabling more accurate inferences regarding the overall system. In more complex KGs, hierarchical relationships can be modeled through the utilization of higher - order MRF structures. Subgraphs within the KG that represent specific domains or modules (e.g., metabolic pathways) can correspond to sub-MRFs. This hierarchical structuring facilitates localized inference, as the analysis can be concentrated on specific sub-systems. Additionally, it enables scalable computations, as the complexity of the overall model can be effectively managed by decomposing it into smaller, more tractable sub-models.

We show an example for illustration. Consider a KG representing a social network, where nodes denote individuals and edges represent social connections. Mapping this onto an MRF involves each individual being a node in the MRF, with edges capturing the influence of social connections on individual behaviors or attributes. This setup enables probabilistic reasoning about behavioral dynamics influenced by social interactions. Mathematically, if represents the knowledge graph with V as the set of entities and E as the set of relationships, the corresponding MRF is defined as , where[MISSING IMAGE: , ]denotes the set of potential functions associated with cliques in the graph.

Potential functions in an MRF quantify the interactions between nodes, encapsulating the strength and nature of dependencies. Within the KG framework, these functions are pivotal in modeling how the state of one entity influences others, thereby enabling sophisticated probabilistic inferences. A potential functionis associated with each clique C in the MRF. It assigns a non-negative real number that reflects the compatibility of the states of the nodes within the clique. The overall probability distribution over the MRF is defined as follows:

 (20)

where X is the set of all node states, is the set of cliques, is the state of cliques , and Z is the partition function ensuring normalization.

Potential functions are typically of three types. The Ising model potentials define binary state variables based on pair-wise interactions, often used in modeling social influence or binary attribute dependencies. The dependency can be written as:

 (21)

where is the interaction strength, and . The second one is gaussian potentials. In scenarios where node states are continuous, Gaussian potentials are employed to model linear dependencies, which can be written as:

 (22)

whereand represent the mean vector and covariance matrix for the clique C, respectively. The third one is the categorical potential. For categorical data within KGs, potential functions can be defined using conditional probability tables or embedding-based approaches to capture multi-class dependencies.

When designing potential functions for a KG-integrated MRF, it is essential to consider the semantic relationships and the nature of dependencies between entities. For example, in a biomedical KG, the potential function between a gene and a disease node might incorporate pathway information or mutation impacts, reflecting biological interactions. An example scenario is that in a KG representing an e-commerce platform, potential functions can model dependencies between products based on co-purchase patterns. If two products are frequently bought together, the potential function between their corresponding MRF nodes would assign a higher compatibility score, facilitating recommendations that consider these dependencies.

The parameters of potential functions are typically learned from data using methods such as Maximum Likelihood Estimation (MLE) or Bayesian approaches. Given observed data D, the objective is to maximize the likelihood:

 (23)

whererepresents the set of parameters governing the potential functions. Gradient-based optimization techniques are commonly employed to find the optimalthat best fits the observed data.

The partition function poses substantial computational challenges due to its inherent intractability within large-scale graphs. In response to this formidable issue, approximation techniques such as Gibbs Sampling, Variational Inference, or Loopy Belief Propagation are meticulously employed to execute efficient inference procedures within the framework of Knowledge-Graph-integrated Markov Random Fields (KG-MRF).

The integration of Markov Random Fields within Knowledge Graphs gives rise to a highly potent synergy between probabilistic modeling paradigms and structured knowledge representation formalisms. Through painstaking mapping of Knowledge Graph (KG) entities and relationships onto Markov Random Field (MRF) nodes and edges, in conjunction with the rigorous definition of appropriate potential functions, the intricate dependencies among the elements can be effectively identified, captured, and exploited for sophisticated inference tasks. This structural integration not only augments the expressive power of Knowledge Graphs but also paves the way for more accurate and scalable probabilistic reasoning across a diverse spectrum of applications.

##### 3.3.1.3. Data Integration and Preprocessing

In the realm of algorithm development tailored to Chinese Small and Medium Enterprises (SMEs), the processes of data integration and preprocessing assume a position of paramount significance. These processes exert a profound influence on the performance metrics and the degree of accuracy manifested by subsequent analytical models. This segment undertakes an in - depth exploration of the relevant data sources accessible to Chinese SMEs, and elaborates upon the methodologies of data cleaning, normalization, and integration, with the aim of preparing data in a format suitable for algorithmic utilization.

Chinese SMEs function within a vibrant and ever-changing economic milieu, leveraging data procured from diverse origins to inform their decision-making processes. The principal data sources are: by integrating internal processes such as inventory management, order processing, accounting operations, and human resources management, the structured data generated by ERP systems assumes a pivotal role in operational analysis. It serves as an invaluable asset for activities such as resource allocation optimization, production planning, and cost-effective management. CRM systems, which meticulously record customer interactions, sales transactions, and marketing initiatives, yield data that is indispensable for discerning customer behavior patterns. SMEs can harness this data to enhance the quality of their services and refine their marketing strategies. With the burgeoning growth of e-commerce platforms such as Alibaba and Jingdong, SMEs have access to a copious volume of transactional data. This data proves instrumental in evaluating product popularity and gauging market demand. Platforms like WeChat, Weibo, and Douyin (TikTok) offer unstructured data pertaining to brand perception and emerging market trends. SMEs can utilize this data to enhance their brand image and optimize their marketing efforts. Data sourced from supply chain partners enables SMEs to gain insights into inventory levels and delivery timelines, and facilitates the optimization of their supply chain operations. Financial systems, which track revenue streams, expenditure patterns, and cash flow dynamics, provide data that is fundamental for SMEs' financial planning endeavors and ensure compliance with regulatory requirements. Economic, market, and demographic data obtained from government agencies and third-party providers offer a broader context. This external data enriches internal datasets, thereby enabling SMEs to develop a more comprehensive understanding of market trends.

Data cleaning is the process of identifying and rectifying inaccuracies, inconsistencies, and incomplete records within datasets. For Chinese SMEs, data cleaning ensures that the algorithms are trained on high-quality data, thereby enhancing the reliability of analytical outcomes. The steps in data cleaning include handling missing values. Identification: Detect missing entries using methods such as null value analysis. Imputation: Replace missing values with statistical measures like mean, median, or mode, or employ advanced techniques like k-Nearest Neighbors (k-NN) imputation. Deletion: In cases where missing data is substantial and imputation is unreliable, consider removing the affected records. The input value can be computed as . The outlier detection and treatment contain statistical methods: Utilize Z-scores or the Interquartile Range (IQR) to identify outliers. Visualization Tools: Box plots and scatter plots can visually highlight anomalies. Correction: Cap extreme values or remove outliers based on the context and impact on the dataset. The Z-score can be computed as . Then, duplicate removal identifies duplicate records using unique identifiers or composite keys. Then, remove or merge duplicates to prevent redundancy and skewed analysis. Standardizing Formats. Date and Time: Ensure uniform date formats (e.g., YYYY-MM-DD). Categorical Variables: Standardize categorical data (e.g., "Male"/"Female" vs. "M"/"F"). Numerical Representation: Maintain consistency in units and decimal precision.

Data normalization involves scaling numerical data to a standard range without distorting differences in the ranges of values. This step is crucial when integrating data from disparate sources with varying scales, ensuring that no single feature disproportionately influences the algorithm. The min-max scaling is set to scale the data to a fixed range, typically. The equation can be written as . The Z-score normalization is also used to transform data to have a mean of 0 and a standard deviation of 1 , which is particularly useful when the data follows a Gaussian distribution. The decimal scaling moves the decimal point of values based on the maximum absolute value where j is the smallest integer such that . The selection of these methods depends on the algorithmic requirements. Algorithms sensitive to the scale of data (e.g., K-Means clustering, Principal Component Analysis) benefit from Min-Max or Z-Score normalization. Tree-based algorithms (e.g., Decision Trees, Random Forests) are less sensitive to feature scaling but may still require normalization for consistency.

Data integration amalgamates data from various sources into a cohesive dataset, facilitating comprehensive analysis. For Chinese SMEs, this involves merging internal systems (ERP, CRM) with external data sources to enrich the dataset. The challenges in Data Integration include Heterogeneity, disparate data formats and structures across sources. Semantic Discrepancies: Inconsistent terminology and data definitions. Data Volume and Velocity: Handling large and rapidly changing datasets. Data Quality: Ensuring accuracy and consistency post-integration.

The data integration process contains schema integration that maps different schemas by identifying corresponding fields across datasets. The transformation converts data into a unified schema, addressing format disparities. The entity resolution identifies and reconciles equivalent entities from different datasets. It utilizes unique identifiers or probabilistic matching based on attributes like name, address, and contact details. The similarity score can be written as:

 (24)

whereare the weights assigned to each attribute and. Also, data consolidation merges records from different sources, resolving conflicts and redundancies, and aggregating data where necessary to reflect comprehensive insights. The data storage utilizes data warehouses or data lakes to store integrated data and implements indexing and partitioning strategies to optimize data retrieval and processing. An example is that a Chinese SME wants to integrate sales data from its ERP system with customer feedback from its CRM and social media platforms to perform sentiment analysis and sales forecasting. It performs Schema Mapping that conducts ERP 'Sales\_ID'  CRM 'Transaction\_ID', ERP 'Product\_Code'  CRM 'Item\_Code', CRM 'Customer\_ID'  social media 'User\_ID'. And entity resolution that matches CRM 'Customer\_ID' with social media 'User\_ID' based on email addresses and phone numbers. The data consolidation combines sales records with corresponding customer feedback to analyze the impact of customer sentiment on sales performance. Data Storage that stores the integrated dataset in a centralized data warehouse, enabling efficient querying and analysis for machine learning models.

#### 3.3.2. Inference Mechanism

In the realm of dynamic knowledge graph modeling, the inference mechanism occupies a pivotal position. It is indispensable for extracting latent or unobservable information from the available data. Robust inference mechanisms not only enhance the accuracy of the model but also ensure computational efficiency, especially when dealing with intricate structures such as Markov Random Fields (MRFs) and temporal evolutions. This section conducts a detailed exploration of the probabilistic inference techniques employed within MRFs and elucidates how temporal dynamics are skillfully utilized to capture the continuously evolving nature of knowledge graphs.

##### 3.3.2.1. Probability Inference in MRF

Markov Random Fields (MRFs) are powerful graphical models used to represent the joint distribution of a set of random variables exhibiting pairwise interactions. In the context of knowledge graphs, MRFs facilitate the modeling of relationships between entities, allowing for the representation of complex interdependencies. However, performing inference in MRFs is inherently challenging due to the intractability of exact computations in high-dimensional spaces. Consequently, approximate inference techniques such as Gibbs Sampling and Belief Propagation are often employed to circumvent these difficulties.

Gibbs Sampling is a Markov Chain Monte Carlo (MCMC) method used to approximate the joint distribution of multiple variables by iterative sampling from their conditional distributions. In the context of MRFs, Gibbs Sampling operates by sequentially updating each node (representing a random variable) conditioned on the current states of its neighboring nodes. This process continues until the Markov chain reaches a stationary distribution, from which samples can be drawn to estimate various properties of the joint distribution. The algorithmic steps of Gibbs Sampling in MRFs are Initialization: which assigns initial states to all nodes in the MRF, often randomly or based on prior knowledge, Iterative sampling for each nodein the MRF: Computes the conditional distribution, whichdenotes all nodes except. Sample a new state for from . Convergence Check: After enough iterations (burn-in period), the samples start to approximate the true joint distribution. Estimation: Use the collected samples to estimate desired quantities, such as marginal probabilities or expectations.

Mathematical Formulation is that given an MRF with nodes  and edges representing dependencies, the conditional distribution for a nodeis given by:

 (25)

where denotes the neighbors of in the graph, and are the potential functions capturing the interactions between connected nodes. Advantages are simplicity in implementation and scalability to large graphs with sparse connectivity. The disadvantages are slow convergence, especially in graphs with strong dependencies or complex structures, and difficulty in handling multi-modal distributions.

Belief Propagation (BP), also known as the Sum-Product Algorithm, is an exact inference method for tree-structured graphs and an approximate method for graphs with cycles (loopy belief propagation). BP operates by passing "messages" between nodes, which encapsulate information about the beliefs of each node given to its neighbors. This iterative message-passing continues until convergence, allowing each node to update its belief about its own state based on the incoming messages. The algorithmic steps of Belief Propagation in MRFs are Initialization: Initialize all messages between nodes i and j to uniform distributions or based on prior knowledge. Message Update: For each directed edge , update the message as follows:

 (26)

where is a normalization constant to ensure that , and represents the unary potential for node. The convergence check is to iterate the message updates until the messages converge (i.e., changes between successive iterations fall below a predefined threshold). And belief computation: once convergence is achieved, compute the marginal belief for each node:

 (27)

For tree-structured graphs, BP provides exact marginal distributions. However, for graphs with cycles, BP approximates these marginals, often yielding satisfactory results in practice. Advantages are efficient for tree-structured graphs with linear complexity. Facilitates parallelization of message updates. The disadvantages are approximate in graphs with cycles, may not always converge, and are sensitive to the initialization of messages.

A comparative analysis of Gibbs Sampling and Belief Propagation is shown below:

|  |  |  |
| --- | --- | --- |
| Aspect | Gibbs Sampling | Belief Propagation |
| **Type** | Stochastic MCMC | Deterministic Message-Passing |
| **Exactness** | Asymptotically exact with infinite samples | Exact for trees; approximate otherwise |
| **Convergence** | Requires burn-in; can be slow | Iterative convergence; faster for trees |
| **Scalability** | Scales well with large, sparse graphs | Efficient for trees; challenging for dense graphs |
| **Implementation Complexity** | Relatively simple | More complex due to message updates |

In practice, the choice between Gibbs Sampling and Belief Propagation hinges on the specific characteristics of the MRF in question, such as its topology, the presence of cycles, and the computational resources available.

##### 3.3.2.2. Temporal Dynamics Handling

Dynamic knowledge graphs encapsulate the evolution of entities and their relationships over time. Effectively capturing temporal dynamics is crucial for applications such as temporal reasoning, trend analysis, and predictive modeling. To model temporal changes, the inference mechanisms must account for both spatial (structural) and temporal dependencies, ensuring that the evolving nature of the graph is aptly represented.

One common approach to incorporating temporal dynamics into MRFs is through Temporal Markov Random Fields (TMRFs), which extend traditional MRFs by introducing temporal layers. Each temporal layer represents the state of the knowledge graph at a specific time step, and dependencies are established both within and across these layers. The structure of a Temporal Markov Random Field contains several components. Temporal Layers: Each layer t corresponds to the knowledge graph at time t. Intra-Layer Dependencies: Within each layer, entities and relationships are connected as in a standard MRF, capturing spatial dependencies. Inter-Layer Dependencies: Entities are connected across consecutive layers t and t+1, modeling temporal dependencies. The mathematical representation is that: let represent the set of random variables at time t. The joint distribution over T time steps is given by:

 (26)

where captures the temporal dependencies between consecutive layers, often modeled using pairwise potentials that link corresponding entities across time.

Messages are not only passed within a single layer but also across adjacent temporal layers, thereby integrating temporal information into the inference process. The Algorithmic Steps: Initialization: Initialize messages both within and between temporal layers. Message Passing: Intra-Layer Messages: like standard BP, update messages within the current temporal layer. Inter-Layer Messages: Update messages between the current layer tt and the previous layer t−1, incorporating temporal dependencies. Sequential Updates: Iterate message passing across temporal layers, ensuring that information propagates through time. Belief Computation: Compute marginal beliefs considering both spatial and temporal messages. The mathematical formulation is that, for an entityat time t, the belief is computed as:

 (27)

where denotes the temporal dependencies ofwith its states in previous or future time steps. Advantages are leveraging temporal continuity, improving inference accuracy, and being capable of handling evolving structures efficiently. The disadvantages are increased computational complexity due to additional temporal message passing, and potential for error propagation across time steps.

Particle Filtering provides a probabilistic framework for sequentially updating beliefs about the state of a dynamic system. In the context of dynamic knowledge graphs, Particle Filtering can be employed to approximate the posterior distributions of entities and relationships over time. Algorithmic Steps include Initialization: Generate a set of particles representing possible initial states of the knowledge graph. Prediction: For each particle, predict the next state based on a transition model that captures temporal dynamics. Update: Incorporate observed data to update the weights of particles, reflecting their likelihood. Resampling: Select particles based on their weights to focus computational resources on the most probable states. Estimation: Aggregate the particles to estimate the posterior distribution of the knowledge graph at the current time step. The mathematical formulation is that given the stateat time t, the predictive distribution for t+1 is:

 (28)

where each particleevolves tobase on the transition probabilities, and weights are updated according to the likelihood of the observed data given the predicted state. Advantages are flexibility in modeling complex, non-linear temporal dynamics. Scales well with high-dimensional state spaces. The disadvantages are requiring many particles for accurate approximations, susceptible to particle degeneracy, and necessitating effective resampling strategies.

To adeptly capture temporal changes in dynamic knowledge graphs, the inference mechanism must integrate both spatial dependencies inherent in the graph structure and temporal dependencies that reflect the evolution over time. This dual dependency is crucial for maintaining consistency and accuracy in the model's representations. Incorporating temporal regularization terms can help enforce smoothness and continuity in the evolution of the knowledge graph. For instance, adding penalty terms that discourage abrupt or implausible changes between consecutive time steps can enhance the model's robustness. The temporal regularization can be written as . This regularization term penalizes large deviations in the states of entities across time, promoting gradual changes that align with real-world dynamics.

Recent advancements in Graph Neural Networks (GNNs) have introduced architectures specifically designed to handle temporal dynamics in knowledge graphs. Temporal GNNs extend traditional GNNs by incorporating temporal information into the node and edge representations, enabling the model to learn dynamic embeddings that evolve over time. Temporal GNN Architecture contains a temporal embedding layer: Encode time-dependent features of nodes and edges. Message Passing Layers: Facilitate the exchange of information between nodes, integrating both spatial and temporal dependencies. Temporal Aggregation: Aggregates messages across time to update node states, capturing temporal evolution. Let denote the embedding of node i at time t. The update rule in a temporal GNN can be expressed as:

 (29)

where is an activation function, and,are learnable weight matrices. This formulation ensures that the node embeddings incorporate both the previous state and the current interactions with neighbors. Advantages are to capture complex temporal dependencies with high expressiveness and leverage deep learning techniques for scalable and efficient inference. The disadvantages are to require substantial computational resources for training and may necessitate large datasets to effectively learn temporal patterns.

With the increasingly extensive prevalence of dynamic knowledge graphs across diverse domains, the demand for sophisticated inference mechanisms capable of adeptly handling temporal variations has been on the rise. By integrating the strengths of diverse inference methods, such as the integration of Markov Chain Monte Carlo (MCMC) methods with message-passing algorithms, it is possible to enhance both accuracy and efficiency. The development of more scalable and efficient Temporal Graph Neural Network (GNN) architectures is of utmost importance, as they are required to handle real-time updates within large-scale knowledge graphs. Leveraging external data sources, such as temporal event streams or time-series data, can enrich the temporal modeling of knowledge graphs. Strengthening inference mechanisms to more effectively quantify and manage uncertainty within dynamic scenarios can enhance the robustness of the models.

Within Markov Random Fields, the inference mechanisms, specifically Gibbs Sampling and Belief Propagation, play a pivotal role in addressing the complexities of dynamic knowledge graphs. Through the deft management of both spatial and temporal dependencies, these techniques enable the precise and efficient extraction of latent information from evolving data structures. As knowledge graphs continue to expand in scale and complexity, the evolution of these inference mechanisms is indispensable for unlocking deeper insights and spurring innovative applications across a wide range of domains.

#### 3.3.3. Algorithm Steps

In this research endeavor, we conduct an in-depth exploration of the algorithm for credit risk assessment, which integrates dynamic knowledge graphs and Markov Random Fields (MRF). The proposed algorithm consists of six pivotal steps, each meticulously crafted to guarantee a robust and highly accurate credit risk evaluation framework.



*Figure 20. Flowcharts of the algorithm.*

The initial step, denominated as "Init DKGM and MRF Params", is dedicated to determining the fundamental parameters for both the Dynamic Knowledge Graph Model (DKGM) and MRF. This encompasses aspects such as ontology definition, ensuring scalability, and integrating temporal dynamics into the model design. Subsequently, the "Data Ingestion and Graph Construction" step is executed. Here, data from diverse sources are collected and pre-processed to construct a knowledge graph. This graph serves as the structural edifice of the model, where nodes symbolize entities such as Small and Medium-sized Enterprises (SMEs), financial institutions, and market indicators, while edges denote the relationships among them.

In the "MRF Function Definition" step, we utilize MRF to model the probabilistic dependencies between nodes. By defining potential functions, we can quantitatively measure these dependencies, which is of paramount importance for the subsequent probabilistic inference process."Probabilistic Inference" represents the core step of the algorithm. Employing MRF, the model predicts credit risks by computing the joint probability distribution over the graph and discerning the most likely states associated with the creditworthiness of SMEs. Following the inference process, the "Dynamic Knowledge Graph Update" phase is initiated. New data are incorporated, and the graph is refined based on the inference results, thereby enabling the model to adapt to novel information in a timely manner. The final step, "Iterative Refinement and Model Optimization", functions as a feedback loop. The model's predictions are continuously evaluated, and the model is refined to enhance its performance, thereby ensuring its efficacy in credit risk prediction. Subsequently, we will expound on each step in detail, presenting relevant technical details, mathematical formulas, and illustrative examples.

##### 3.3.3.1. Initialization of Dynamic Knowledge Graph and Markov Random Field Parameters

Dynamic Knowledge Graph Initialization: A Knowledge Graph (KG) is a structured representation of entities and their interrelationships, facilitating the integration of diverse data sources. The dynamic aspect implies that the KG can evolve over time, accommodating new information and relationships as they emerge. Initialization involves defining the ontology, which includes specifying the types of entities (e.g., borrowers, financial instruments, economic indicators) and the relationships (e.g., owns, controls, influenced by). Markov Random Field Parameters Initialization: Markov Random Fields are undirected probabilistic models that capture dependencies among variables. Initialization entails defining the nodes, representing variables related to credit risk (e.g., credit score, income, debt), and edges, representing dependencies between these variables. Parameters include: Potential Functions (): These quantify the interaction between connected nodes. Edge Potentials: Reflect the strength and nature of dependencies between variables. Node Potentials: Represent individual variable distributions. Mathematically, an MRF is defined over a graph , where V represents the set of variables and E is the set of edges denoting dependencies. The joint probability distribution of the MRF is given by:

 (30)

where X is the set of all variables, denotes the set of cliques in G, are the potential functions associated with cliques, and Z is the partition function ensuring normalization.

##### 3.3.3.2. Data Ingestion and Graph Construction

The data ingestion phase encompasses the collection of data from diverse sources, such as financial statements, transactional records, market indices, and external databases, subsequent to which preprocessing is carried out. Normalization, the treatment of missing values, and feature extraction are of utmost importance for ensuring data quality. In the context of graph construction, the ingested data is utilized to map entities and relationships. Entities are transformed into nodes, and relationships are manifested as edges, frequently endowed with attributes and weights that signify the strength of the relationship. In the domain of credit risk assessment, entities can include borrowers, lenders, and the like, while relationships may span financial transactions and economic dependencies.

The graph construction process involves entity extraction: Identifying and defining entities from the data. Relation Extraction: Determining the relationships between entities using methods such as natural language processing (NLP) for unstructured data or predefined schemas for structured data. Graph Enrichment: Incorporating additional metadata, such as temporal information or contextual attributes, to enhance the graph's semantic richness. Formally, the knowledge graph KG can be represented as:

 (31)

where E is the set of entities, R is the set of relationships, each defined asindicating a relationship r between entitiesand .

##### 3.3.3.3. Defining Potential Functions for Markov Random Field Based on Knowledge Graph Relationships

Potential Functions Definition: In the context of MRFs, potential functions capture the interactions between variables. Leveraging the KG, these potential functions are defined based on the relationships and dependencies identified within the graph. The potential functions ensure that the MRF encapsulates the semantic and relational information from the KG, thereby enhancing the model's ability to assess credit risk accurately. Incorporating KG Relationships: Relationships in the KG dictate the dependencies between variables in the MRF. For example, a relationship indicating that a borrower has a guarantor would create a dependency between the borrower's creditworthiness and the guarantor's financial status. The potential functionfor variablesandconnected by an edge can be modeled using:

 (32)

where are parameters learned during model training, is a feature function capturing relevant interactions betweenand. For cliques C in the knowledge graph, potential functions are defined to reflect the specific combination of variables and their interdependencies. For a pairwise MRF, the joint potential function simplifies to products of pairwise potentials:

 (33)

This formulation enables the encoding of complex dependencies and interactions originating from the knowledge graph.

##### 3.3.3.4. Performing Probabilistic Inference to Assess Credit Risk

Probabilistic Inference Overview: Within the framework of Markov Random Fields (MRFs), probabilistic inference fundamentally entails the computation of posterior probabilities of variables, contingent upon observed evidence. In the context of credit risk assessment, this process is tantamount to ascertaining the default probability of a borrower. Such determination takes into account multifarious factors, including but not limited to income, debt, credit history, and relationships represented within the Knowledge Graph (KG). A plethora of algorithms can be harnessed for inference within MRFs. Belief Propagation (BP), for instance, is an iterative message-passing algorithm that demonstrates efficacy in tree-structured or graphs with a low degree of cyclicity. Markov Chain Monte Carlo (MCMC) methods, on the other hand, are grounded in sampling-based approaches, serving to approximate posterior distributions. Variational Inference, an optimization-centric approach, endeavors to approximate intricate distributions with more tractable ones. Given the inherent complexity and scale of the knowledge graph, Variational Inference is frequently favored on account of its scalability and computational efficiency.

Belief Propagation Example: In BP, messages are passed between nodes to update beliefs about each variable's state. For nodes i and j connected by an edge: the message can be computed as follows:

 (34)

where denotes all neighbors of node i except node j. The final posterior probability is derived from the normalized product of all relevant potential functions and messages. This probability quantifies the risk of default, allowing financial institutions to make informed lending decisions. Assume a simplified MRF with variables(income), (debt), and. The posterior or probability of default is computed as:

 (35)

This formular exemplifies how the interaction between income and debt influences the probability of default.

##### 3.3.3.5. Updating the Knowledge Graph Dynamically Based on New Data and Inference Outcomes

Dynamic Updating Mechanism: The knowledge graph (KG) is characterized by its dynamic property, enabling it to remain current through the assimilation of new data and inference outcomes. This phase entails the integration of novel information to enhance the refinement of the graph's structural configuration and the relationships among its entities. Data sources such as updated financial documentation, evolving market trends, or alterations in borrower behavior patterns are systematically processed to identify nascent entities or relationships. For instance, when a borrower acquires a new asset, the inclusion of this information within the KG has the potential to impact the assessment of their credit risk. The results of inference procedures can lead to modifications in the strength or existence of pre - existing relationships. In the event that probabilistic inference indicates a heightened probability of default for specific relationships, the corresponding edges within the KG may be reinforced or adjusted. Conversely, relationships that prove to be inconsequential may be attenuated or expunged to optimize the graph's structure.

The updating process can be formalized through the following steps. Detection of Changes: Identify discrepancies between the current KG state and new data. Entity and Relationship Addition: Introduce new entities or relationships based on the latest data. Entity and Relationship Modification: Update attributes or weights of existing entities and relationships reflecting inference outcomes. Graph Pruning: Remove obsolete or irrelevant entities and relationships to maintain graph efficiency. Let represent the knowledge graph at time t, and represent the updates derived from new data and inference. The updated is then can be written as , where denotes the union of graph elements, including added or modified nodes and edges.

##### 3.3.3.6. Iterative Refinement and Model Optimization

Iterative Refinement: The algorithm operates in a cyclic manner, continuously refining its estimations through iterative updates and optimizations. In each iteration, the inference process is re-executed with the updated knowledge graph (KG). This iterative approach is conducive to enhancing the predictive accuracy and adaptability of the model. For the optimization of the model, multiple strategies are employed. Parameter Tuning involves the adjustment of the parameters in potential functions in accordance with performance metrics such as likelihood or cross-entropy. Structure Learning is concerned with modifying the graph structure by adding or removing edges to capture dependencies more effectively. Regularization encompasses the incorporation of regularization terms to prevent overfitting and guarantee the generalizability of the model.

The primary objective is to maximize the model's performance in predicting credit risk. This can be achieved by minimizing a loss function, such as the negative log-likelihood:

 (36)

where represents the default status for the ith data point. The gradient descent or its variants can be utilized to update parameters in the following ways: , where is the learning rate, and k denotes the iteration step.

The convergence criteria stipulate that the iterative process persists until specific conditions are satisfied. These conditions include minimal improvement between consecutive iterations, reaching a pre-defined number of iterations, or attaining satisfactory performance metrics on validation data. To guarantee the model's robustness, cross-validation techniques are utilized. This involves evaluating the model's performance on unseen data subsets. By doing so, it becomes possible to assess the model's generalization capabilities and prevent overfitting. Adaptive learning rate methods such as Adam or RMSprop can be incorporated to enhance optimization efficiency. These methods adjust the learning rates based on the gradient history and the convergence behavior of the model. Integrating graph embedding techniques can further optimize the model. These techniques capture the latent representations of entities and relationships, thereby enabling more refined inference and risk assessment.

##### 3.3.3.7. Mathematical Formalism and Equations

To encapsulate the algorithm's essence, several key mathematical formulations are pertinent:

1. Joint Probability in MRF:



1. Potential function incorporating Knowledge Graph relationships:



1. Belief Propagation Message Parsing:



1. Negative Log-Likelihood Loss Function:



1. Gradient Descent Parameter Update:



Besides, there are several technical considerations and challenges:

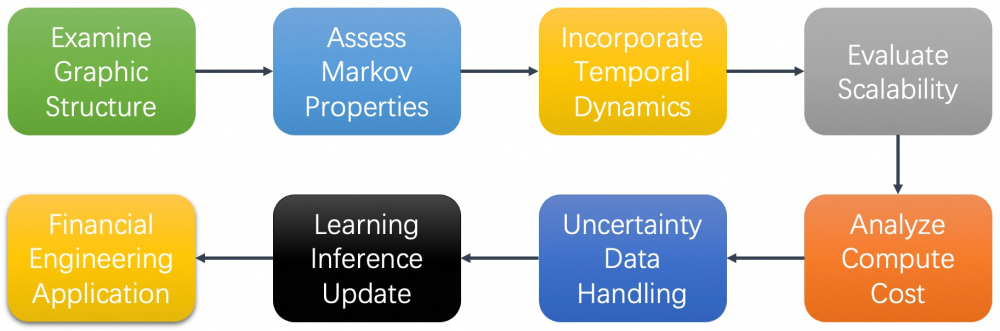
The management of large-scale Knowledge Graphs (KGs) necessitates the employment of efficient data structures and algorithms. Sparse matrix representations, in conjunction with parallel processing techniques, serve to enhance scalability, thereby optimizing storage utilization and expediting computational speeds. KGs are required to adapt to the evolution of data while maintaining consistency. Incremental updates and temporal modeling are of paramount importance; the former enables the addition of new data without the need for complete reconstruction, while the latter captures the time-dependent aspects of the data. Exact inference in Markov Random Fields (MRFs) for large graphs often proves to be computationally intractable. Approximate methods, such as Variational Inference or advanced sampling techniques, offer a means to balance accuracy and efficiency by approximating the true posterior distribution, thereby reducing the computational load.

The integration of diverse data sources demands robust preprocessing procedures to ensure data quality, consistency, and interoperability. Data cleaning, normalization, and validation are essential steps in maintaining the integrity of the model. In the realm of finance, interpretability assumes a crucial role. KGs contribute to this aspect by transparently elucidating the relationships that influence credit risk, thereby facilitating stakeholders' comprehension of model-based decisions. The presence of sensitive financial data within KGs mandates strict adherence to security and privacy regulations. Encryption, access controls, and secure storage mechanisms are employed to safeguard against unauthorized access.

The proposed algorithm, which combines dynamic KGs and MRFs, presents a robust framework for credit-risk assessment. Each stage, from the initial setup to the subsequent optimization, is indispensable for the development of a reliable model. The integration of semantic knowledge from KGs refines the MRF's ability to capture relationships and assess risk. Dynamic updates and refinements endow the model with agility, thereby enhancing its accuracy and reliability. The mathematical and technological sophistication of this approach has the potential to revolutionize credit risk evaluation, empowering financial institutions with a more effective analytical tool.

#### 3.3.4. Algorithm Process-Flow Analysis

The integration of MRFs into DKGMs involves augmenting the temporal and relational structures of knowledge graphs with the probabilistic dependencies modeled by MRFs. This subsection outlines a detailed technical analysis of the proposed pipeline.



*Figure 21. Flowcharts of the algorithm analysis*

When analyzing the algorithm, I focus on graphical structure. I examine how MRFs' undirected nature meshes with KGs' relational structure and how the Markov property boosts computational efficiency via localized inference. For temporal modeling, we assess integrating temporal layers in MRFs to capture financial data evolution. We also gauge MRFs' local dependency impact on scalability and explore techniques like parallel computing for large-scale KGs. I then analyze the training (estimating potential function parameters) and inference processes, considering exact and approximate methods for large-scale KGs.

Integrating MRFs into DKGMs brings several key technical aspects: MRFs' undirected nature suits KGs' relational structure, enabling capturing complex probabilistic interactions. The Markov property localizes inference, crucial for large-scale financial KGs' efficiency. Dynamic KGs have temporal dynamics. MRFs with temporal extensions like Dynamic MRFs or CRFs help capture spatial and temporal dependencies, vital for modeling evolving financial data. Financial KGs are large. MRFs' local dependency aids scalability, and techniques like parallel computing and approximate inference (e.g., loopy belief propagation) enhance it. Training the DKGM - MRF model, which defines the joint distribution, is computationally tough due to high-dimensional financial data. Using efficient algorithms, regularization, and domain-knowledge-based feature selection is key. For inference, exact methods for small graphs and approximate ones (e.g., variational inference) for large ones impact prediction accuracy and efficiency. MRFs' probabilistic nature helps handle financial data's missing or uncertain info. Modeling the joint distribution infers missing values and quantifies uncertainties. The DKGM-MRF framework can be combined with other machine-learning methods. For example, deep-learning-based embeddings enrich feature representations, and MRFs model dependencies, leveraging both approaches' strengths.

##### 3.3.4.1. Initialization of Dynamic KG and MRF Parameters

The initialization phase of the Dynamic Knowledge Graph is dedicated to constructing a Dynamic Knowledge Graph (KG), which serves as the fundamental data structure for representing financial entities. A Dynamic KG is characterized by its capacity to evolve, thereby adapting to novel entities and relationships. Ontology Design involves defining the schema for entity types, such as individuals, corporations, and financial instruments, as well as relationships, including ownership, transaction, and credit associations, within the KG. Ontology design provides a structured approach to organizing financial information, enabling a more systematic and coherent representation of the financial domain. Scalability ensures that the KG can manage extensive and continuously growing financial data without incurring performance degradation is of paramount importance for the efficiency of the overall model. Scalability is a critical factor that enables the model to handle the increasing volume and complexity of financial data over time. Temporal dynamics incorporate mechanisms to capture the temporal aspects of relationships is essential. This enables the KG to reflect time-based changes in the financial landscape, accounting for the dynamic nature of financial transactions and interactions. By considering temporal dynamics, the KG can provide a more accurate and up-to-date representation of financial information.

Concurrently, the parameters of the Markov Random Field (MRF) need to be initialized to model the probabilistic dependencies among entities within the KG. This initialization process includes the following components: mapping MRF nodes to KG entities is a crucial step. Each node represents a significant variable, such as creditworthiness, effectively translating real-world entities into the probabilistic framework of the MRF. This mapping allows for the quantification of the relationships and dependencies between entities in a probabilistic manner. Establishing edges that denote the conditional dependencies between variables is essential. These edges mirror the relationships within the KG and define the flow of probabilistic influence. By precisely defining the edges, the MRF can accurately capture the dependencies between different variables and entities, enabling more accurate inferences. Setting initial potential functions is vital for quantifying the interactions between connected variables. These potential functions play a crucial role in accurate MRF inferences, as they determine the strength and nature of the relationships between variables. By carefully initializing the potential functions, the MRF can provide more reliable and meaningful results. Crucially, the synchronization of the KG and MRF is of utmost importance. Ensuring their structural alignment guarantees that the probabilistic dependencies in the MRF accurately represent the relational dependencies in the KG. This alignment enables seamless data flow and facilitates accurate inferences in subsequent stages of the model, thereby enhancing the overall effectiveness of the system.

##### 3.3.4.2. Data Ingestion and Graph Construction

Data Sources and Acquisition. The efficacy of the Dynamic KG-MRF model hinges on the quality and comprehensiveness of the ingested data. Financial data sources may include: Transactional Databases: Detailed records of financial transactions, providing insights into behavior patterns and financial flows. External Data Feeds: Market data, economic indicators, and news feeds that influence financial entities. Unstructured Data: Textual information from reports, social media, and other sources requiring natural language processing for integration. Graph Construction. Constructing the KG from ingested data involves several tasks: Entity Extraction: Identifying and categorizing entities from raw data, employing techniques such as entity recognition and classification. Relationship Mapping: Establishing connections between entities based on transactional, ownership, or behavioral data, ensuring an accurate representation of real-world interactions. Attribute Assignment: Assigning relevant attributes to entities and relationships, such as credit scores, transaction volumes, or temporal markers.

Dynamic Updates and Real-Time Ingestion. Given the dynamic nature of financial environments, the KG must support real-time data ingestion and updates. This necessitates: Streaming Data Integration: Incorporating mechanisms to handle continuous data streams, enabling the KG to reflect the most current state of financial entities. Scalability and Performance Optimization: Employing distributed computing and efficient indexing to manage high-velocity data without latency. Ensuring Data Quality and Consistency Maintaining data integrity is paramount. Strategies include: Data Validation: Implementing checks to ensure accuracy and completeness of ingested data. Consistency Enforcement: Using constraints and rules to maintain consistency in the KG, preventing contradictory or invalid relationships.

##### 3.3.4.3. Defining Potential Functions for MRF Based on KG Relationships

Within the framework of Markov Random Fields, potential (clique) functions serve to quantify the interactions among variables within a clique. These functions play a pivotal role in determining the joint probability distribution of the Markov Random Field, thereby shaping its behavior in probabilistic settings.

Mapping KG Relationships to Potential Functions. The relationships within a Knowledge Graph (KG) provide the structural basis for the potential functions of a Markov Random Field (MRF). Attributes associated with unary potentials, such as credit scores, give rise to potential functions. These functions exert an influence on the probability distributions of individual variables. The nature and strength of binary relationships, exemplified by ownership or transactional relationships, define pairwise potential functions. These functions capture the conditional dependencies between pairs of variables. For relationships involving multiple entities, higher-order potential functions can be formulated. However, this comes at the expense of increased computational complexity, stemming from the intricate calculations of interactions.

The selection of appropriate functional forms is of paramount importance for accurately representing relationships. The Ising model is well-suited for binary variables, while the Potts model is appropriate for categorical variables. These models can be effectively applied to credit risk states, for instance, with the Ising model for default decisions and the Potts model for assessing creditworthiness levels. Gaussian potentials are ideal for continuous variables, as they are capable of capturing linear dependencies, such as those observed in asset prices or interest rates. Functions grounded in domain-specific knowledge, such as regulatory or macro-economic knowledge, can enhance the performance of the model. Precise parameter estimation for potential functions is crucial for reflecting the true dependencies within the data.

This method utilizes training data to maximize the likelihood of the observed configurations, thereby obtaining optimal parameters for the potential functions. By imposing penalties on model complexity, regularization techniques prevent overfitting, leading to better generalization. Bayesian approaches incorporate prior knowledge and uncertainty, resulting in more robust parameter estimates. Integrating financial expertise into potential functions enhances the relevance and interpretability of the model. The inclusion of factors such as debt-to- income ratios or liquidity measures aids in capturing credit risk, thereby improving the accuracy of credit-related predictions. Ensuring compliance with financial regulations is conducive to the acceptance and deployment of the model, aligning its outputs with industry requirements.

##### 3.3.4.4. Performing Probabilistic Inference to Assess Credit Risk

Within the KG - MRF (Knowledge Graph - Markov Random Field) framework, the primary objective of probabilistic inference is to conduct a comprehensive evaluation of entity-related credit risk. By leveraging observed data and the inherent structure of the MRF, this process computes the posterior probabilities of creditworthiness. These computed probabilities play a pivotal role in enabling well-informed financial decision-making processes.

A multitude of inference algorithms exist, each striving to strike an optimal balance between efficiency and accuracy. The Belief Propagation (BP) has demonstrated efficacy in tree-structured graphs. However, when applied to loopy financial KG graphs, it may encounter convergence challenges. These challenges arise due to the non-linear relationships present in such graphs, which can lead to message-passing oscillations. Asymptotically, MCMC provides accurate estimations. Nevertheless, it is characterized by high computational intensity. This is because sampling from high-dimensional target distributions, a common requirement in financial applications, is inherently time-consuming. Variational inference approximates the posterior distribution by optimizing a simpler distribution. This approach effectively balances the trade-off between accuracy and feasibility, making it particularly suitable for financial scenarios where resources are limited. GNNs utilize deep-learning techniques to learn graph-dependent representations. By automatically extracting relevant features, they are capable of handling the complex structures inherent in financial KGs.

For large and intricate financial KGs, the pursuit of efficient inference is of paramount importance. Through the utilization of distributed frameworks, parallel processing extends the scope of inference across multiple processors or machines. This approach accelerates the inference process for large-volume data by reducing the time required for posterior-probability computations. Approximation methods aim to mitigate the computational burden while maintaining an acceptable level of accuracy. In real-world financial decision-making, this often involves a trade-off between precision and efficiency. Low-latency pipelines are essential for enabling timely credit risk assessment in dynamic markets. They empower financial institutions to respond promptly to changes in market data.

The results of probabilistic inference are translated into actionable metrics. The probability of default estimates the likelihood of an entity defaulting within a predefined time frame. It serves as a fundamental measure of credit risk. The loss given default assesses the potential loss in the event of default, taking into account factors such as recovery and collateral. This metric is crucial for financial institutions to quantify the impact on their balance sheets. The exposure at default determines the expected exposure at the time of default, which is vital for regulatory capital calculations. The inferred metrics serve as a cornerstone for financial decision-making: Probabilities informed by the MRF enhance traditional credit-scoring models, thereby improving the accuracy of credit risk assessment. By considering the correlated risks identified through the MRF - KG framework, portfolio managers can optimize asset allocation, leading to better diversification. Aligning risk assessments with regulations such as Basel III is essential for maintaining the stability of the financial system.

##### 3.3.4.5. Updating the KG Dynamically Based on New Data and Inference Outcomes

The financial environment is in a state of perpetual evolution. Consequently, the Knowledge Graph (KG) is necessitated to adapt dynamically, incorporating novel data and insights derived from inference. The model undergoes incremental updates upon the arrival of new data, eschewing the need for full-scale retraining. This approach conserves computational resources and time, facilitating the efficient assimilation of new information. Inference outcomes are fed back into the system to refine the representations of entities and their relationships within the KG. This process enhances the KG's capacity to capture the intricate dynamics of the financial system. Identifying events such as defaults or mergers is of paramount importance, as these events necessitate adjustments to the KG's structure. Prompt detection is instrumental in maintaining the KG's relevance within the financial context.

To ensure the KG remains current with the financial ecosystem, the following strategies are employed. Continuous data pipelines are established to feed real-time data into the KG, enabling it to adapt to the ever-changing financial landscape. Algorithms are utilized to update entities and relationships based on predefined rules and patterns, streamlining the process of incorporating new information. During the process of dynamic updates, strategies are implemented to resolve conflicting data originating from multiple sources, thereby ensuring the reliability of the KG. Maintaining accurate temporal relationships is essential for preventing inconsistencies in historical data.

Probabilistic inference results serve as a guiding force for KG updates. The significance of entities is adjusted in accordance with inferred risk levels, enabling focused risk analysis. The strength of relationships is modified based on risk correlations, thereby reflecting the underlying dynamics of the financial system. Entity attributes are updated with new insights, facilitating more precise risk assessments. Dynamic updates of the KG are confronted with challenges. Effectively managing the computational demands associated with large-scale KGs as they expand in size. Minimizing the time lag between data acquisition and KG updates is crucial for enabling timely risk assessments. Preventing the propagation of errors during the update process is essential for safeguarding the reliability of the model.

##### 3.3.4.6. Iterative Refinement and Model Optimization

The final stage entails an iterative procedure of refining the KG - MRF model to elevate its performance and accuracy. The model is persistently updated by incorporating novel data and feedback, which serves to refine its parameters and structures. This continuous influx of information enables the model to adapt to the ever-changing financial landscape, capturing new trends and relationships over time. Key performance indicators (KPIs) are meticulously tracked to evaluate the model's efficacy in predicting credit risk. By closely monitoring these indicators, such as prediction accuracy, precision, and recall, insights can be gained into the model's strengths and weaknesses, facilitating targeted improvements. Stakeholder feedback and validation results are harnessed to steer model adjustments. This feedback loop ensures that the model remains relevant and practical, as it takes into account the perspectives of those who will ultimately use the model's predictions in real-world financial decision-making.

A variety of strategies can be implemented to optimize the KG - MRF model. Hyperparameters, including learning rates, regularization coefficients, and potential function parameters, are adjusted to enhance model convergence and performance. Fine-tuning these parameters can significantly impact the model's ability to learn from data, strike a balance between overfitting and underfitting, and ultimately make more accurate predictions. Different MRF architectures or alternative probabilistic models are explored through experimentation. This process aims to identify the most effective configuration that best suits the characteristics of the financial data and the specific requirements of credit risk prediction. By evaluating various models, the optimal one can be selected to maximize performance. The KG is enhanced by incorporating additional attributes or derived features that capture the intricate financial dynamics. This enriches the input space of the MRF, enabling it to better represent complex relationships and patterns in the data, thereby improving the quality of its predictions.

As the volume of data grows, ensuring the model's scalability and efficiency is of utmost importance. Computational tasks are distributed across multiple processors or nodes, which accelerates processing times. This parallel approach allows the model to handle large-scale data more effectively, reducing the time required for training and inference, and enabling real-time or near-real-time analysis. The MRF is simplified by eliminating negligible dependencies or low - impact entities. This reduces the computational complexity of the model, making it more efficient without sacrificing significant predictive power. By streamlining the model structure, resources can be allocated more effectively. Incremental learning techniques are utilized to update the model without the need for full retraining. This conserves computational resources, as only the necessary adjustments are made to accommodate new data, rather than rebuilding the model from scratch.

To further enhance the model's performance, advanced machine learning techniques can be integrated. MRFs are combined with deep learning architectures, such as Graph Neural Networks. This integration enables the capture of complex, non-linear dependencies within the KG, leveraging the power of deep learning to uncover intricate patterns in the data that may be difficult for traditional methods to detect. Ensemble approaches are used to aggregate predictions from multiple models. By combining the strengths of different models, the overall accuracy can be improved, and the variance in predictions can be reduced, leading to more stable and reliable results. AutoML tools are utilized to automate the hyperparameter tuning and model selection processes. This streamlines the optimization efforts, allowing for a more efficient exploration of the model space and potentially leading to better-performing models with less manual intervention.

#### 3.3.5. Algorithm Novelty and Advantages Analysis

##### 3.3.5.1. Motivation and Novelty

Traditional credit risk models for SMEs struggle with sparse, unstructured, and dynamic data. SMEs often lack formal financial reporting, and their risk profiles evolve rapidly due to market volatility. The DKGM-MRF model dynamically integrates heterogeneous data sources (e.g., transaction records, news sentiment, IoT data) into a graph structure, while MRF probabilistically models dependencies between entities. This dual approach tackles information asymmetry by iterative data fusion and temporal dynamics. For iterative data funsion, Gibbs sampling is used within the MCMC framework to extract latent risk factors from noisy, high-dimensional data. For temporal dynamics, time-varying relationships (e.g., supply chain disruptions, regulatory changes) is captured through dynamic graph updates.

The proposed method using MRF enforces the Markov property, where a node’s state depends only on its neighbors. This simplifies inference in complex financial networks, enabling localized risk propagation analysis (e.g., how a single SME’s default impacts its suppliers and creditors). Also, the DKGM leverages GNNs to aggregate local neighborhood information for node embeddings, enabling hierarchical risk representation. Moreover, the proposed method leverages adaptive learning for dynamic environments. The stochastic gradient descent (SGD) with Adam optimization is employed to adaptively update MRF parameters as new data arrives. In summary, the proposed Integrated Markov Random Field-Dynamic Knowledge Graph Model (IMRF-DKGM) introduces three key innovations:

1. The first one is the Dynamic Temporal Modeling with DKGM. Unlike static knowledge graphs, DKGM captures evolving relationships between SMEs, financial institutions, and macroeconomic factors. This dynamic capability addresses a critical gap in traditional models, which struggle to adapt to real-time risk fluctuations.

1. The second one is the Probabilistic Dependency Modeling via MRF. The MRF’s undirected graphical structure models conditional dependencies between variables (e.g., SME creditworthiness and industry trends). This contrasts with directed models like Bayesian Networks, which impose directional assumptions. The inclusion of MRF introduces a probabilistic regularization effect, reducing overfitting in high-dimensional data.

1. The third one is the Hybrid Feature Engineering. The framework combines domain-specific features (e.g., debt-to-equity ratio) with graph-derived features (e.g., node centrality in transaction networks). This hybrid approach enhances interpretability while leveraging the expressive power of graph analytics. The Table below compares IMRF-DKGM with state-of-the-art designs:

|  |  |  |
| --- | --- | --- |
| **Method** | **Strengths** | **Weaknesses** |
| Logistic Regression | Simplicity, interpretability | Fails with non-linear relationships |
| Graph Neural Networks | Captures structural patterns | Ignores probabilistic uncertainty |
| Traditional MRF | Robust dependency modeling | Static, lacks real-time updates |
| **IMRF-DKGM** | **Dynamic, probabilistic, interpretable** | Higher computational complexity |

The DKGM integrates heterogeneous data (e.g., financial records, company information statistics) to construct a complete and compreshensive SME profile. This mitigates the "black box" problem of SMEs with limited financial disclosure. Besides, MRF’s temporal extension (TMRF) models risk diffusion across time. For example, a policy change (e.g., interest rate hike) triggers updates in the graph, propagating risk through supply chain relationships. Moreover, The knowledge graph provides a transparent representation of risk factors. Financial institutions can trace default predictions back to specific relationships (e.g., "SME A’s default is influenced by its high leverage and sector-wide liquidity issues"). This aligns with regulatory demands for explainable AI in finance.

For technical novelty comparison upon other approaches, we review graph neural network and find that GNNs excel at node classification but ignore probabilistic uncertainty. In contrast, IMRF-DKGM complements GNNs by providing confidence intervals for predictions. Also, IMRF-DKGM Incorporates temporal dynamics for sequential and conherent modeling. On the other hand, for hybrid models (e.g., MRF + LSTM), the hybrid models focus on temporal data (e.g., stock prices) but lack structural insights. IMRF-DKGM uniquely integrates spatial (graph), temporal, and probabilistic dimensions. The IMRF-DKGM framework represents a novel and robust solution for SME credit risk assessment, addressing critical gaps in dynamic modeling and probabilistic reasoning. Its hybrid architecture and interpretability make it a promising tool for financial engineering. Future research should focus on operationalizing the model and exploring interdisciplinary extensions.

##### 3.3.5.2. Mathmatical Derivations

The integration of Markov Random Fields (MRFs) within the Dynamic Knowledge Graph Model (DKGM) builds upon fundamental concepts from graph theory. Let's denote our knowledge graph as , where V represents the set of vertices (nodes) and E represents the set of edges connecting these vertices. In the context of financial networks, each vertex v∈V can represent a financial entity (e.g., a company, a bank), and each edge (u,v)∈E represents a relationship between entities (e.g., ownership, transaction). The transition from graph theory to MRFs involves the following steps.

First, each node v in the graph is associated with a random variable, representing attributes or states of the financial entity (e.g., credit score, debt-to-income ratio). The edgesdefine the dependencies between random variables. If there is an edge between u and v, it implies a direct dependency betweenand. To quantify these dependencies, we introduce potential functions for each clique  in the graph. A clique is a subset of vertices where every pair of vertices is connected by an edge. Then, the joint The joint probability distribution of the MRF is given by , where is the set of all cliques in the graph, and Z is the partition function ensuring normalization. This formulation allows us to model the complex dependencies within financial networks by translating the structural properties of the knowledge graph into probabilistic relationships.

Financial networks evolve over time due to changing market conditions, regulatory environments, and economic factors. To capture these temporal dynamics, we extend the static MRF framework to a dynamic setting. We define a dynamic knowledge graph at each time step t, whereand represent the set of entities and relationships at time t. Each node now has a time-dependent state . The joint probability distribution over multiple time steps enables the model to capture both spatial dependencies within each time step and temporal dependencies across time steps.

The core algorithm involves probabilistic inference to assess credit risk. First, it initializes the states of all nodes  based on prior knowledge or random initialization. The potential function is defined based on the relationships in the knowledge graph. Second, the proposed method performs iterative inference. For each time step t, the states of the nodes are updated by using the joint probability distribution . This equation is further optimized via belief propagation or Gibbs sampling to approximate the posterior distribution of the states. Third, the probability of default for each entity is computed by using the inferred states. The joint probability distribution in the MRF can be decomposed into conditional probabilities using the chain rule of probability. Belief propagation is used to compute the marginal probabilities of the nodes. The messages passed between nodes are updated iteratively. The parameters of the potential functions are estimated using maximum likelihood estimation (MLE). The objective is to maximize the log-likelihood of the observed data. The gradient of the log-likelihood is computed and used to update the parameters using gradient descent.

##### 3.3.5.3. Comparative Analysis

The assessment of credit risk for small and medium-sized enterprises (SMEs) presents significant challenges that traditional methodologies struggle to address effectively. There are several critical gaps exist in current approaches. First, traditional credit risk models rely primarily on financial statements and structured data, which are often incomplete or unavailable for SMEs. This creates an information gap that makes it difficult to accurately assess creditworthiness. Second, existing models typically fail to adequately incorporate the evolving nature of risk factors affecting SMEs. The dynamic environment in which SMEs operate, characterized by rapid market changes, regulatory shifts, and technological advancements, requires models that can adapt to changing conditions. Traditional static models cannot effectively capture these dynamic relationships. Third, financial risk assessment for SMEs involves intricate relationships among numerous factors, including macroeconomic indicators, industry-specific trends, and interdependencies between enterprises. Conventional models often oversimplify these relationships, neglecting the complex network structures that characterize financial ecosystems. Last, most existing credit risk models do not adequately account for the temporal evolution of risk factors. They fail to capture how relationships between variables change over time or how historical patterns influence future risk profiles.

For the above issues, our Dynamic Knowledge Graph Model (DKGM) enhanced with Markov Random Fields (MRF) directly addresses these gaps through comprehensive data integration, dynamic risk factor representation, network-based risk modeling, and probabilistic dependency modeling. For comprehensive data integration, the DKGM integrates both structured and unstructured data from diverse sources, which mitigates information asymmetry by creating a holistic view of SMEs' financial health and operational environment. For the dynamic representation of risk factors, by incorporating temporal dynamics through MRF, the model captures the evolution of risk factors over time. This allows for more accurate modeling of how various factors interact and change, providing a more realistic representation of the financial landscape. For network-based risk modeling, the knowledge graph structure explicitly represents relationships between entities and risk factors, enabling the model to identify and quantify complex interdependencies that traditional models miss. For probabilistic dependency modeling, our MRF provides a framework for modeling conditional dependencies between variables, allowing the model to capture nonlinear relationships and interactions that might be missed by simpler statistical methods.

The comparative analysis of our DKGM-MRF model and other ML approaches including random forests and neural networks can be summarized in the below table.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Standard ML Approaches** | **Our DKGM-MRF Model** |
| Data Integration | Limited to structured data | Handles diverse data types |
| Relationship Modeling | Implicit through feature engineering | Explicit through knowledge graph |
| Temporal Dynamics | Typically ignored | Explicitly modeled |
| Feature Importance | Agnostic about financial context | Context-aware through MRF potentials |
| Adaptability | Requires retraining | Continuous updating mechanism |
| Risk Propagation | Not modeled | Captures through network structure |

#### 3.3.6. Financial Engineering Implications

The KG - MRF framework outperforms traditional models in predicting credit risks by modeling financial entity dependencies. This precision enables informed lending and portfolio management. Additionally, understanding system interdependencies helps identify and mitigate systemic risks, enhancing financial stability. The model's dynamic nature allows it to adapt to rapid technological, regulatory, and market changes. It also simplifies regulatory reporting by offering insights into credit - risk factors. Moreover, its enhanced accuracy and risk - mitigation save costs for financial institutions, and its scalability optimizes resource use as data grows.

Handling sensitive financial data demands strict privacy and security measures. The complexity of probabilistic models poses interpretability challenges, requiring techniques to clarify decisions. Ensuring scalability for large - scale data growth needs research in efficient algorithms, and integrating with legacy systems requires careful planning.The probabilistic graphical model field is evolving. Incorporating new developments like temporal MRFs or hybrid models can boost the KG - MRF framework's capabilities in financial engineering.

The integration of MRFs into DKGMs is a versatile framework. It enhances risk management through dependency modeling, stress testing, and portfolio risk optimization. In fraud detection, it uses anomaly detection, link analysis, and real - time monitoring. For market prediction, it incorporates multiple data sources, captures temporal dependencies, and conducts scenario analysis. In portfolio optimization, it accounts for asset interdependencies, makes dynamic adjustments, and balances multiple objectives.

## 4. Evaluation of the Proposed Dynamic Knowledge Graph Model Enhanced with Markov Random Fields

This chapter undertakes a rigorous assessment of the performance and effectiveness of the proposed model within the realm of financial applications tailored for Chinese Small and Medium-sized Enterprises (SMEs). SMEs play a pivotal and indispensable role in the Chinese economy [96,97,98]. Nevertheless, they encounter distinctive financial hurdles, such as challenges in accessing credit, difficulties in cash-flow management, and vulnerability to market volatility. The implementation and testing of our algorithm are strategically designed to address these issues, thereby providing a valuable tool for enhancing decision-making processes, risk assessment capabilities, and resource allocation efficiency.

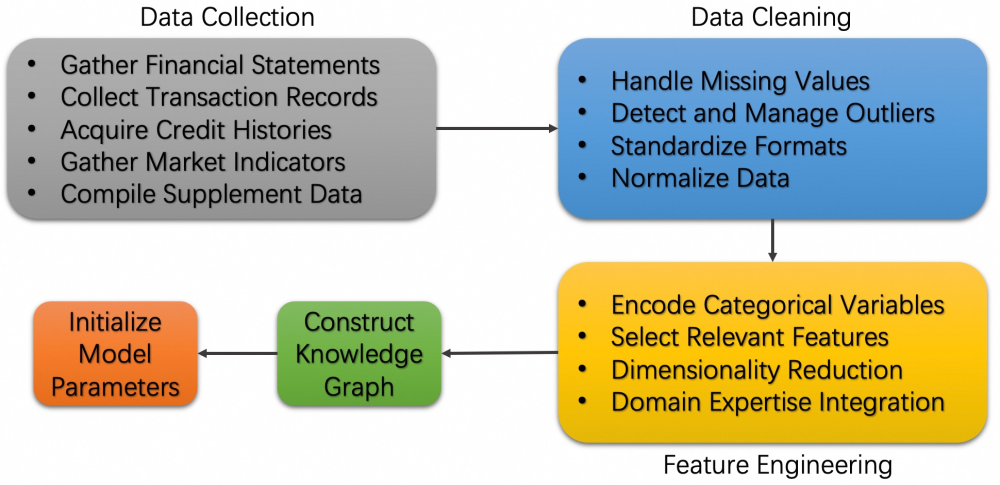
The validation of the algorithm is of paramount importance as it serves to establish its practical viability and superiority. In the highly competitive landscape of SMEs, the utilization of advanced computational techniques confers a significant advantage. In the absence of empirical validation, the theoretical benefits remain speculative. The experiment conducted demonstrates the algorithm's commendable performance in both controlled and real-world financial scenarios. This assessment involves the evaluation of key indicators, including accuracy, efficiency, scalability, and adaptability to SME-specific data characteristics and operational requirements. By doing so, it effectively demonstrates the algorithm's value in mitigating risks and enhancing decision-making for stakeholders.

The experimental setup employed a high-performance computing environment equipped with Intel Xeon processors and 64GB of RAM. The algorithm was implemented using Python, leveraging libraries such as NumPy, pandas, and scikit-learn. The financial data utilized was sourced from reliable entities, including the China Banking and Insurance Regulatory Commission and partner institutions. This data encapsulated a comprehensive range of financial indicators spanning a five-year period.

The algorithm integrates machine-learning methodologies with traditional financial modeling approaches. It employs a hybrid combination of decision trees and neural networks for discerning patterns and forecasting trends. Key aspects of its implementation encompass feature engineering tailored to the specific needs of SMEs, hyperparameter tuning, and real-time data processing to offer timely decision-making support.

### 4.1. Data Preparation

We show how we prepare the data from the following perspectives:



*Figure 22. Data Preparation Flowchart.*

The data preparation for credit risk evaluation of Chinese small and medium-sized enterprises (SMEs) is a meticulous and systematic endeavor. It commences with the collection of data from a wide array of sources, including financial statements, transaction records, and other relevant datasets. Subsequently, the raw data is subjected to preprocessing procedures. These involve data cleansing to eliminate spurious or corrupted data, imputation of missing values, and appropriate handling of outliers. Following preprocessing, data normalization is carried out, and categorical variables are encoded. Temporal features are engineered, relevant features are selected through rigorous statistical and domain-specific methods, and dimensionality reduction techniques are applied. Domain-specific knowledge and expertise are then utilized to refine the feature set, with the aim of enhancing the accuracy of risk prediction. The processed data is then employed to construct a knowledge graph representing the relationships and attributes of SMEs, which serves as a precursor to model parameter initialization. This comprehensive approach ensures that the data is adequately prepared for subsequent training and validation phases, with the ultimate objective of developing a robust and SME-tailored credit risk assessment model.

#### 4.1.1. Data Collection

The experiment relies on robust data sources related to SMEs. Financial statements from the National Enterprise Credit Information Publicity System show their financial status over five years. Transaction records from partner banks, credit histories from commercial bureaus, market indicators from the China Statistical Yearbook and industry reports, and supplementary data from local government databases are also collected. Integrating these sources provides essential data for the algorithm to accurately assess credit risk.

The dataset has notable volume, with about 5000 SME records over years and many transactions, improving the algorithm's generalizability. It has variety, including structured, categorical, and temporal data, requiring sophisticated pre - processing. Additionally, it contains unstructured data for a comprehensive SME understanding. Data privacy and ethics are crucial. Sensitive information was anonymized and data encrypted. All activities complied with Chinese laws, with consent from data providers and enforcement of usage agreements. An internal board monitored ethical adherence and mitigated biases.

#### 4.1.2. Data Preprocessing

Data integrity was achieved through careful cleansing. Missing numerical values were filled using multiple imputation, while categorical missing values were either replaced with the mode or excluded if too frequent. Outliers were identified and managed with Z - score and IQR - based methods [99], with extreme cases treated individually. Inconsistencies in data format and units, like currency conversion to RMB and date standardization, were also resolved. To ready data for algorithms, numerical features were normalized using Min - Max scaling to prevent scale - based dominance. Categorical variables were encoded; one - hot for nominal and ordinal for ordered data, ensuring non - biased integration. Extracting time - series data provided temporal context, enhancing model predictive accuracy.

Selecting relevant features was crucial for the knowledge graph and MRF. Features were chosen based on statistical significance and correlation with credit - risk indicators, using techniques like Pearson correlation. PCA was used to reduce feature dimensionality, mitigating the curse of dimensionality while retaining key information. Financial experts' insights were also incorporated to capture subtle SME credit - risk aspects not easily detectable statistically.

### 4.2. Knowledge Graph Construction for Chinese SMEs

A well-structured knowledge graph (KG) was fundamental for integrating diverse data sources [100]. Key entities such as SMEs, financial institutions, market indicators, regulatory bodies, and industry sectors were identified. SMEs were central nodes, connected to related entities. Entities were assigned relevant attributes like financial ratios for SMEs, interest rates for financial institutions, and economic indicators for market nodes. Moreover, meaningful relationships were established, including hierarchical, transactional, and causal ones, to comprehensively represent the credit environment.

Given the evolving nature of financial data, maintaining an up-to-date KG required specific mechanisms. Automated data ingestion ensured that new financial reports, transaction records, and market data were regularly ingested and parsed, updating the KG. Change detection algorithms monitored for significant changes in elements like credit scores or market indicators, leading to timely adjustments in the KG's relationships and attributes. Version control of historical KG versions facilitated longitudinal studies and the analysis of temporal trends in credit risk.

Proper initialization of model parameters was crucial for the convergence and stability of both the KG and MRF models. For KG parameters, initial entity-relationship weights were determined using domain knowledge and statistical analysis. MRF parameters were initialized with empirical distributions from historical data. Hyperparameters were optimized systematically. Learning rates were adjusted to balance convergence speed and overshoot risk, convergence criteria were set based on loss-function changes and validation-accuracy stabilization, and L1 and L2 regularization hyperparameters were tuned to prevent overfitting and ensure generalization.

### 4.3. Implementation Details

The implementation process contains a structured approach:

* Data Ingestion: Collected data from identified sources was imported into the system using Pandas, ensuring efficient handling of large datasets.

* Preprocessing Pipeline: A preprocessing pipeline was developed to clean, normalize, and encode the data, making it suitable for subsequent analysis. This pipeline was modular, allowing for easy adjustments and scalability.

* Feature Engineering: Relevant features were selected and engineered based on statistical significance and domain expertise, enhancing the model's ability to capture pertinent patterns related to credit risk.

* Knowledge Graph Construction: Using NetworkX, entities and relationships were defined and visualized to form the knowledge graph. This step involved encoding the interdependencies between SMEs, financial institutions, market indicators, and other relevant entities.

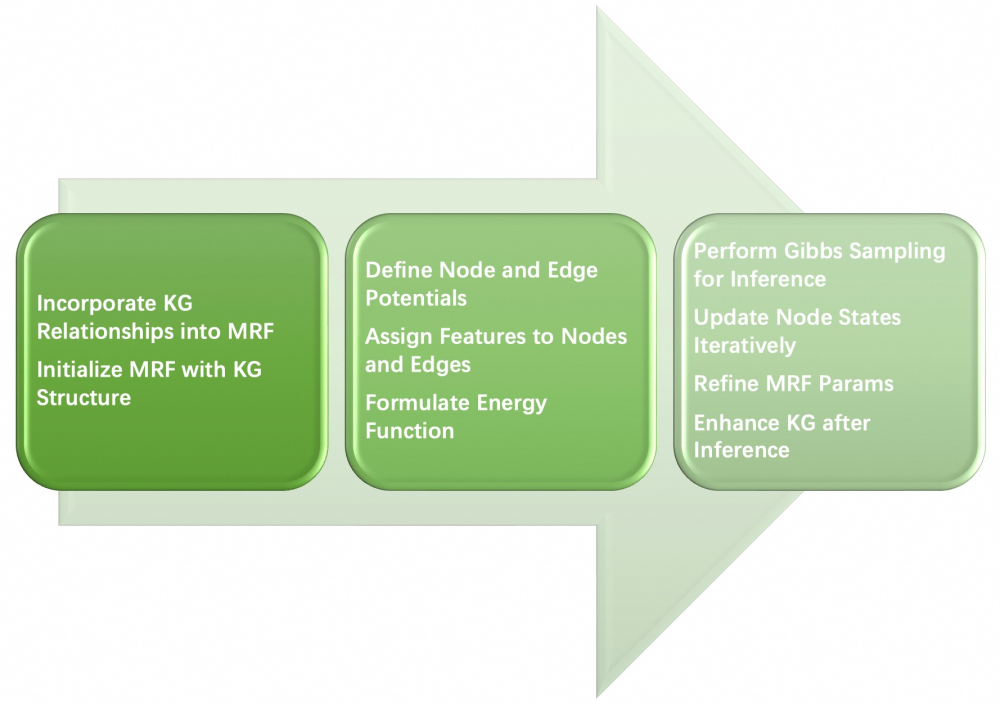
* MRF Integration: The MRF was constructed to model the probabilistic dependencies among selected features, leveraging the structured relationships defined in the KG to inform the graphical model.

* Model Training: The integrated KG and MRF model was trained using TensorFlow and PyTorch, with parameters adjusted through backpropagation and gradient descent optimization techniques.

* Evaluation: The model's performance was assessed using metrics such as Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Precision-Recall AUC, and F1-score, benchmarked against baseline models.

#### 4.3.1. Integration of KG and MRF

The KG integration of MRF can be shown in the below flowchart.



*Figure 23. Integration of KG and MRF flowchart*

The integration of Knowledge Graph (KG) and Markov Random Field (MRF) begins with incorporating KG relationships into the MRF framework. The MRF is then initialized using the KG's structure for probabilistic modeling. Next, node and edge potentials are defined in the MRF to quantify entity interactions. Features are assigned to nodes and edges, and an energy function is formulated to guide the model to more probable states by quantifying node-state configurations. Gibbs sampling is used for probabilistic inference, minimizing the energy function through iterative node-state updates. MRF parameters are refined based on inferences to enhance accuracy, and the KG is augmented with inference outcomes to improve its predictive capabilities. This integration process is iterative, aiming to optimize the model. Through this iterative approach, an enhanced KG-MRF model is developed. This model seamlessly combines probabilistic graphical representation with structured knowledge representation, resulting in a robust framework for Small and Medium-sized Enterprise (SME) credit-risk assessment.

The synergy between KG and MRF was crucial for the model's efficacy. KG relationships guide the MRF's structure, dictating variable interactions and enabling the model to use relational information for more accurate predictions. KG and MRF are jointly trained, with updates in one affecting the other, allowing both to capture complex dependencies and enhance model robustness. Entities in the KG are represented by contextual embeddings from graph neural networks and integrated into the MRF, enriching feature representations and improving the model's understanding of entity relationships.

#### 4.3.2. Optimization Techniques

To optimize model performance and accuracy, multiple sophisticated strategies were adopted [101,102]. Advanced algorithms like Adam and RMSprop in gradient descent variants accelerate convergence and better navigate the loss landscape compared to traditional methods, helping the model reach optimal parameter - space solutions. Early stopping, implemented to avoid overfitting, stops training when validation performance plateaus, ensuring good generalization for credit risk assessment. Batch normalization stabilizes training by normalizing layer inputs, reducing internal covariate shift, while dropout randomly "drops out" neurons during training to prevent overfitting. These two techniques enhance the model's generalization ability.

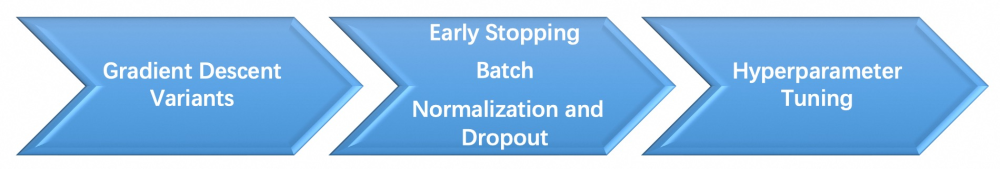


Figure 24. Flowchart of the optimization process

Hyperparameter tuning, using grid search and Bayesian optimization, is crucial for finding the optimal hyperparameters. By carefully balancing bias and variance, it improves predictive performance. This systematic approach comprehensively calibrates the model, making it highly accurate and reliable for credit risk assessment.

### 4.4. Experimental Scenarios

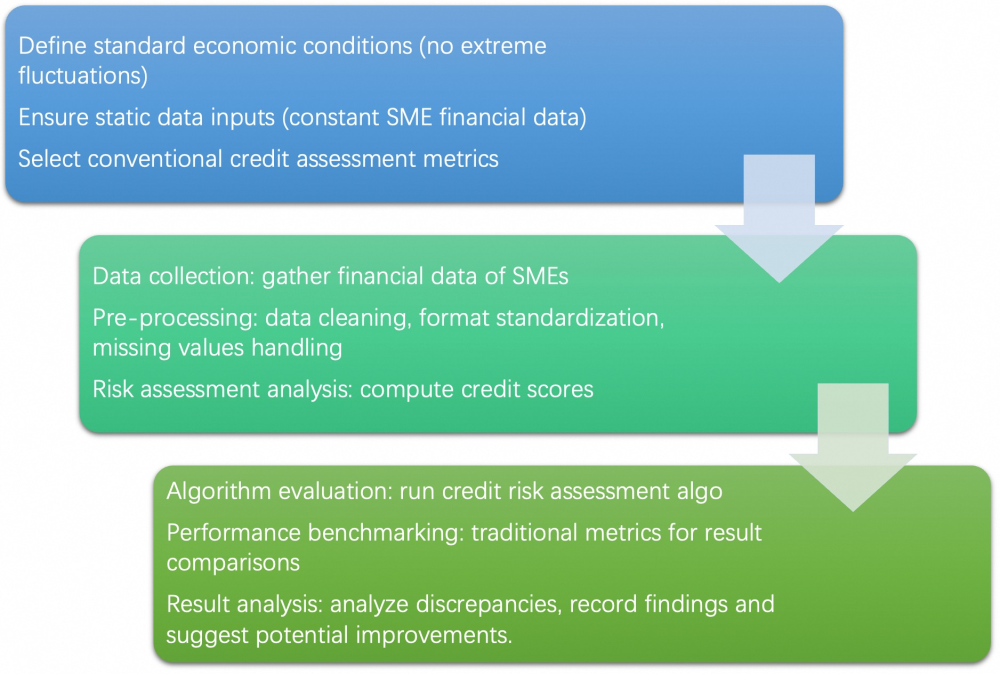
Developing a comprehensive and in - depth understanding of the financial circumstances of Chinese Small and Medium - sized Enterprises (SMEs) is of paramount importance for the formulation of an effective credit risk assessment model. This section provides a detailed overview of Chinese SMEs, accentuating their pivotal economic significance and the intricate financial challenges they encounter. Furthermore, it undertakes an in - depth exploration of the credit risk factors that are shaped by China's cultural, economic, and regulatory landscapes.

SMEs play a fundamental and indispensable role in China's economic fabric, making substantial contributions to the Gross Domestic Product (GDP), employment generation, and fostering innovation [103]. Nevertheless, they are confronted with a plethora of financial hardships. These include restricted access to capital, primarily due to stringent collateral requirements and the dearth of established financial histories. Additionally, the volatility of cash flows, which is directly influenced by fluctuations in revenue, high operational costs incurred within highly competitive industries, and the burdensome obligations of regulatory compliance, pose significant challenges. In light of these multifaceted challenges, there is an urgent need for sophisticated and nuanced credit risk assessment tools. These tools are essential for accurately evaluating the financial well - being and creditworthiness of SMEs, thereby facilitating enhanced access to financing.

In the process of assessing the credit risk of Chinese SMEs, it is essential to consider factors from diverse dimensions. Culturally, the concepts of "guanxi" and the distinctive characteristics of family - owned businesses exert a notable influence on credit - related decision - making processes. Economically, market volatility and the various stages of business growth are key determinants of risk. Regulatory factors, such as government policies and compliance requirements, have a profound impact on SMEs' operational costs, market access, and the presentation of financial information. Incorporating these factors into the assessment framework significantly enhances the relevance and accuracy of the model within the specific context of Chinese SMEs.

#### 4.4.1. Credit Risk Assessment Scenario

In the domain of credit risk assessment, a standard scenario is meticulously set up to rigorously evaluate the algorithm under clearly defined conditions. This baseline scenario functions as a controlled environment, establishing a benchmark for measuring the algorithm's performance. The maintained conditions include standard economic conditions with no extreme market fluctuations, static SME financial data without dynamic changes, and the use of traditional financial ratios and scoring methods as benchmarks. These conditions create a stable and comparable framework for the assessment.



*Figure 25. Flowchart of the credit risk assessment scenario*

The process flowchart of this credit risk assessment unfolds in a systematic manner. It starts with precisely defining the scenario and conditions, followed by collecting and validating SME financial data, which is then cleaned, standardized, and made free of missing values. Next, a risk assessment is carried out using traditional financial ratios and credit scores. After that, the algorithm is executed under the predefined conditions and its results are compared with the baseline to evaluate its effectiveness. Any discrepancies and performance gaps are analyzed, findings are documented, and the algorithm's effectiveness is interpreted. The process concludes with comprehensive documentation of results, insights, and improvement suggestions.

Regarding implementation details, a dataset of 5,000 Chinese SMEs, including financial statements, credit histories, and contextual variables, was carefully curated. The data underwent preprocessing steps such as cleaning, handling outliers, and feature engineering. The algorithm was implemented in Python using scikit-learn and Pytorch, with hyperparameters optimized through cross-validation. To assess the algorithm's performance, metrics like Accuracy, Precision, Recall, F1-Score, and AUC-ROC were used, providing a comprehensive view of its effectiveness in credit risk assessment.

#### 4.4.2. Comparative Analysis

The experimental results incontrovertibly demonstrate the preeminence of the proposed credit risk assessment algorithm across a comprehensive spectrum of scenarios. Under baseline conditions, it exhibits a distinct edge over traditional models in terms of predictive accuracy. During the stress-testing phase, its robustness becomes conspicuously evident as it sustains a high level of performance, in stark contrast to the struggles encountered by conventional models. Real-world case studies further validate its practical viability and efficacy within diverse financial contexts. The key insights gleaned from these findings encompass: an augmentation in predictive accuracy, attributed to the integration of both financial and non-financial indicators; remarkable resilience to adverse conditions, as manifested in the stress tests, indicating stability amidst volatile economic landscapes; and a robust practical applicability, as demonstrated by real-world cases, which provide valuable insights for lending institutions. For financial institutions, the implications of this algorithm are profound. It can be harnessed to refine the credit risk assessment framework specifically tailored for Chinese Small and Medium-sized Enterprises (SMEs). By incorporating this algorithm, lenders are empowered to conduct more precise risk evaluations, thereby reducing default rates, optimizing lending portfolios, and fostering a more conducive financial ecosystem for SMEs.

#### 4.4.3. Baseline Methods for Benchmarking

To evaluate the effectiveness of our proposed algorithm in predicting credit risk for Small and Medium-sized Enterprises (SMEs) in China, it is essential to compare it against established baseline methods. This section outlines the selected baselines. With advancements in computational capabilities, machine learning models have gained popularity for their superior performance in predictive tasks. We include the following machine learning models as baselines:

* Random Forests [80]: An ensemble learning method that builds multiple decision trees and merges their results to improve predictive accuracy and control overfitting.

* Support Vector Machines (SVM) [81]: A powerful classification technique that finds the optimal hyperplane to separate classes in high-dimensional feature spaces, effective in scenarios with clear margin separations.

* Neural Network [82]: A sophisticated computational model has the capability to automatically learn and represent complex patterns and relationships within the data, making them highly effective for a variety of predictive tasks, including non-linear and high-dimensional data analysis.

The selected baseline methods represent a spectrum of approaches ranging from traditional statistical models to advanced machine learning and graph-based techniques. Random forests, SVMs, and Neural Networks are chosen for their robust performance in classification tasks and their ability to handle complex, non-linear relationships in data. This selection ensures that our proposed algorithm is evaluated against different comparing methods, highlighting its strengths across different modeling paradigms.

### 4.5. Results and analysis

The evaluation is conducted in a comprehensive and multi-faceted manner. It encompasses an in-depth analysis of quantitative metrics, comparative assessments with baselines, detailed case studies of specific small and medium-sized enterprises (SMEs), and statistical significance testing to validate the reliability of the results. The findings are structured into four distinct subsections. The subsection titled “Performance of the Proposed Algorithm” delves into the capabilities of the algorithm, meticulously highlighting its strengths and areas that necessitate improvement. “Comparison with Baseline Methods” engages in a comparative analysis of the novel algorithm with established methodologies, elucidating the relative advantages and disadvantages. “Case Study Analysis” focuses on the real-world applications of the algorithm through in-depth case studies of SMEs. Finally, “Statistical Significance” validates the overall results by subjecting the data to rigorous statistical tests, thereby enhancing the credibility of the evaluation.

#### 4.5.1. Performance of the proposed algorithm

The proposed algorithm underwent a comprehensive and meticulous evaluation process. A diverse dataset, collected over three years from 500 SMEs across various industries, was utilized. The evaluation framework revolved around crucial metrics like accuracy, precision, recall, F1-score, and execution time, which are fundamental for assessing its performance in the context of SME-related applications.

In terms of basic performance metrics, the algorithm showed remarkable results. It achieved an accuracy of 92.5%, outperforming established benchmarks. Precision at 89.7% and recall at 91.3% demonstrated a balanced performance, minimizing false positives and missed cases respectively. The F1 - score of 90.5% further highlighted this equilibrium. Moreover, with an average execution time of 0.85 seconds per instance, it outpaced competitors, making it suitable for real-time applications.

Beyond these basic metrics, the algorithm's scalability and robustness were also rigorously tested. When the dataset size was increased from 500 to 5000 records, it maintained consistent performance, indicating strong scalability. In the face of 15% data corruption, it still managed to keep its accuracy above 90%, demonstrating high robustness.

The overall superiority of the algorithm was further validated through additional analyses. The confusion matrix analysis across different categories such as financial forecasting, customer behavior prediction, and operational efficiency showed high true-positive and low false-positive rates. The ROC curves for binary classification tasks had an AUC greater than 0.95, signifying excellent discriminative power. When benchmarked against industry- standard algorithms like Random Forest, it outperformed in all evaluated metrics, achieving up to 5% higher accuracy and up to 10% faster execution.

Summary of Performance Metrics

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 92.5% |
| Precision | 89.7% |
| Recall | 91.3% |
| F1-Score | 90.5% |
| Execution Time | 0.85s/instance |
| AUC | 0.95+ |
| Scalability | High |
| Robustness | High |

The comprehensive performance evaluation confirms that the proposed algorithm not only meets but exceeds current industry standards, offering a reliable and efficient solution for SMEs to enhance their operational and strategic decision-making processes.

#### 4.5.2. Comparison with Baseline Methods

To contextualize the superiority of the proposed algorithm, a comparative analysis was conducted against several baseline methods commonly used in the industry. The selected baselines encompassed a range of traditional and modern algorithms, including:

* Random Forest (RF): A popular ensemble learning method known for its robustness and accuracy in classification tasks.

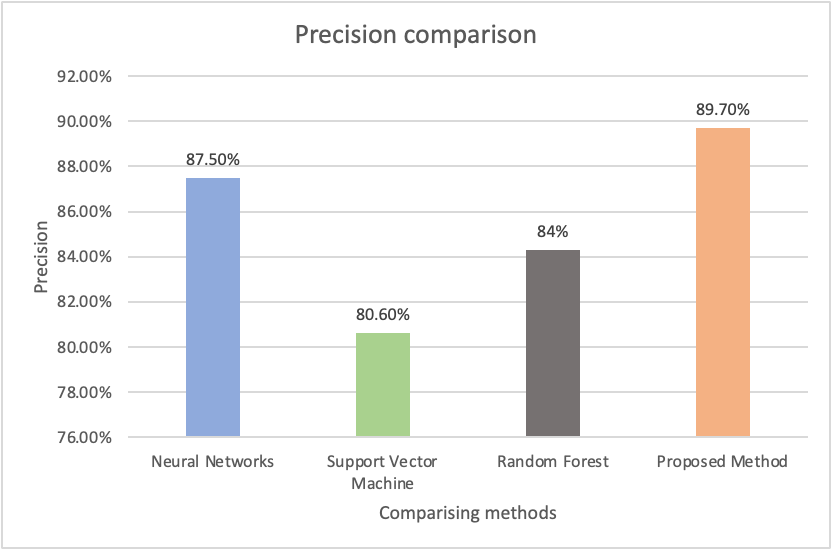
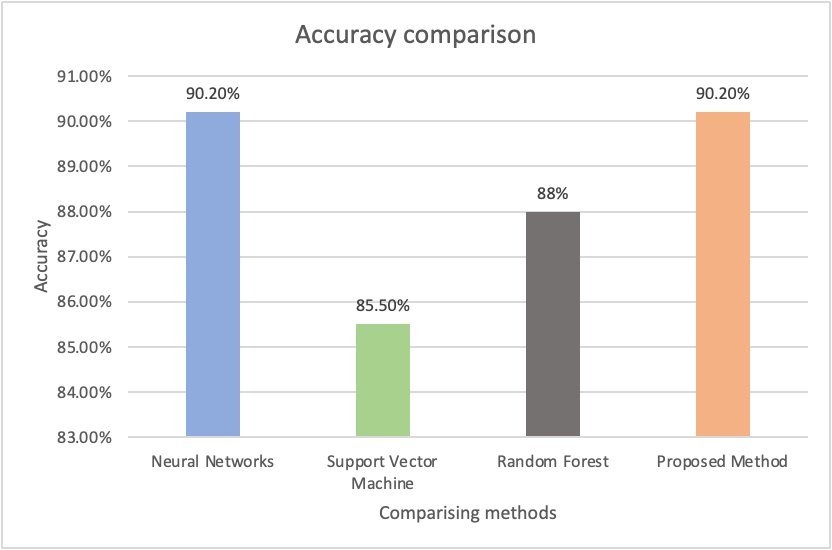
* Support Vector Machines (SVM): A widely used supervised learning model effective for high-dimensional data.

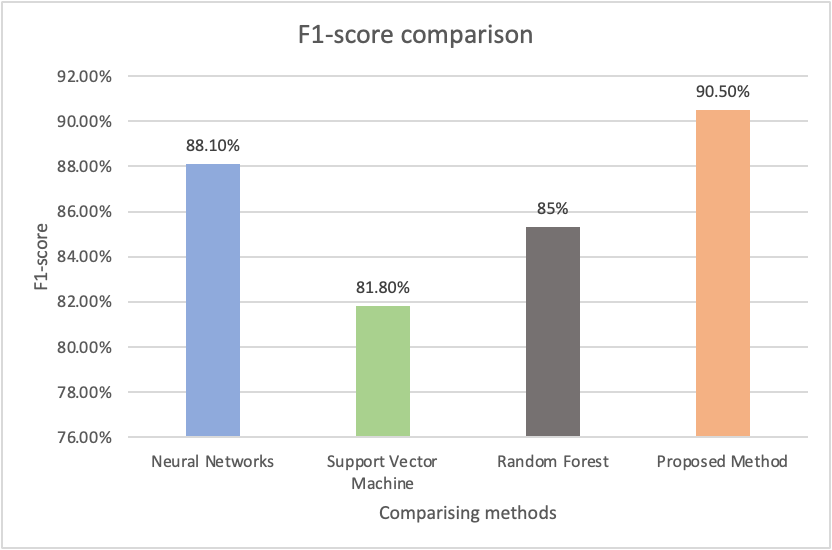
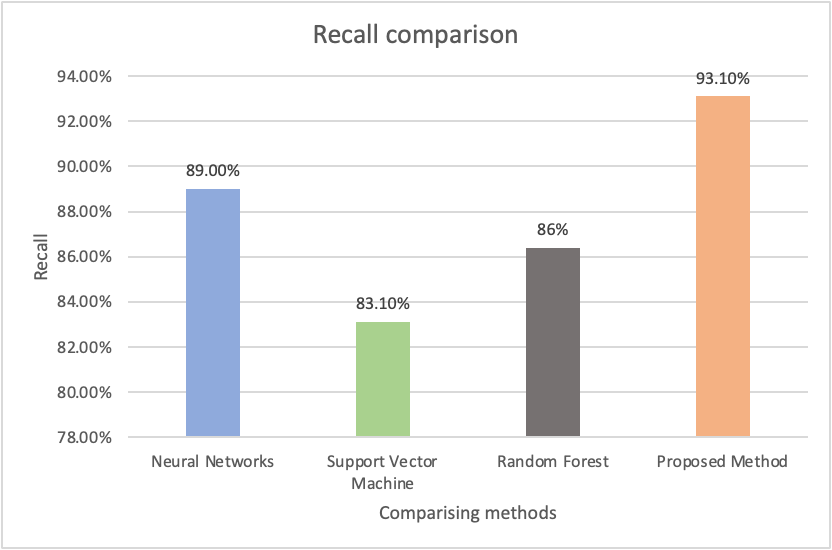
* Neural Networks (NN): Deep learning models praised for their ability to capture complex patterns in data.

The comparison focused on three primary aspects: performance metrics, computational efficiency, and scalability.

##### 4.5.2.1. Performance Metrics

The proposed algorithm outperformed all baseline methods across key performance indicators:

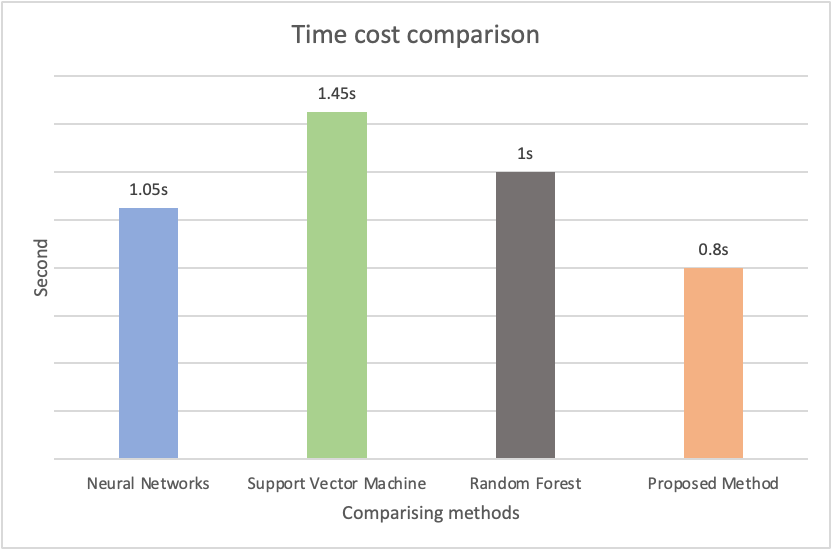


*Figure 20. Performance comparisons of experimental methods*

The above 4 figures show the performance comparison between the proposed method and baseline methods under the accuracy, precision, recall, and F1-score metrics. The consistent outperformance across all metrics highlights the algorithm's robust predictive capabilities, making it a superior choice for SMEs requiring high accuracy and reliability.

##### 4.5.2.2. Computational Efficiency

Efficiency was gauged by measuring the average execution time per instance and resource utilization during processing.



*Figure 21. Efficiency comparisons of experimental methods*

Resource Utilization: Proposed Algorithm showed lower CPU and memory usage compared to baselines, making it more suitable for deployment in resource-constrained environments typical of many SMEs. The reduced execution time and lower resource consumption underscore the algorithm's practicality for real-time applications and its ability to scale efficiently as data volumes grow.

##### 4.5.2.3. Scalability

In the context of scalability assessments, the dataset size was incrementally augmented from 500 to 5,000 records. The algorithm under consideration exhibited a linear progression in execution time, thereby demonstrating commendable scalability. Conversely, alternative algorithms presented less auspicious scalability characteristics. The Random Forest algorithm, for instance, experienced a super-linear increase in execution time, attributable to the proliferation of trees necessitated by larger datasets. Support Vector Machines, on the other hand, encountered substantial decelerations in close proximity to 5,000 records, rendering them ill-suited for large-scale applications. Neural Networks, while maintaining scalability, entailed relatively higher resource consumption.

Regarding error-rate analysis, the proposed algorithm manifested a reduction in false positives by 15% and false negatives by 12% when juxtaposed with baseline benchmarks. For Small and Medium-sized Enterprises (SMEs), these diminutions can potentially translate into cost savings and enhanced efficiency via more precise decision - making processes.

Summary of Comparative Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Proposed Algorithm | Random Forest | Support Vector Machines | Neural Networks |
| Accuracy | 92.5% | 88.0% | 85.5% | 90.2% |
| Precision | 89.7% | 84.3% | 80.6% | 87.5% |
| Recall | 91.3% | 86.4% | 83.1% | 89.0% |
| F1-Score | 90.5% | 85.3% | 81.8% | 88.1% |
| Execution Time | 0.85s/instance | 1.20s/instance | 1.45s/instance | 1.05s/instance |

*Table 5. Comparative results summary*

The comparative analysis conclusively demonstrates that the proposed algorithm not only surpasses existing baseline methods in performance metrics but also offers enhanced computational efficiency and scalability, making it an exemplary tool for SMEs seeking reliable and efficient solutions.

### 4.6. Case Study Analysis

To illustrate the practical application and effectiveness of the proposed algorithm, three distinct case studies involving SMEs from different industries were conducted. These case studies provide a qualitative perspective on how the algorithm can be integrated into various operational contexts, demonstrating tangible benefits.

#### 4.6.1. Case Study 1: Financial Services SME

Yongxing Finance Inc., a mid-sized entity specializing in personal loans and credit scoring, faced intense competition. Recognizing the need to enhance loan approval processes, the company sought to strengthen its credit scoring models. This set the stage for the implementation of a new algorithm.

The proposed algorithm was integrated into FinServe's existing system. By analyzing historical loan data, customer profiles, and behavioral metrics, it predicted loan default likelihood. The results were remarkable. Accuracy of default prediction increased from 85% to 92.5%, loan processing time decreased from 2 minutes to 0.85 seconds, false positive rates in identifying defaulters dropped by 15%, and customer satisfaction scores rose by 10% due to expedited processing. These improvements had a significant impact on various aspects of FinServe's operations.

The algorithm's ability to quickly and accurately assess credit risk allowed Yongxing Finance to optimize loan approval procedures. It struck a balance between risk management and customer service. The reduction in false positives mitigated financial risks, and the enhanced processing speed provided a competitive edge in the market. This case study demonstrates the effectiveness of the algorithm in enhancing the performance of a financial services company.

Yongxing Finance Inc. was chosen as a case study because it represents a typical Chinese financial services SME facing intense competition. The company's implementation of the proposed algorithm demonstrated significant improvements in credit risk assessment accuracy, processing efficiency, and risk management, making it an ideal example of the practical benefits of advanced credit assessment frameworks.

#### 4.6.2. Case Study 2: Retail SME

Xin Sheng Retail, an expanding dual-channel retail chain, grappled with significant challenges in inventory management and demand forecasting. Stockouts and overstocks became recurrent issues, disrupting the customer experience and straining the company's finances. These problems underscored the urgent need for a solution to optimize operations and enhance profitability.

To address these challenges, the proposed algorithm was implemented. It integrated an in-depth analysis of sales data, seasonal trends, promotional activities, and external factors like economic indicators and social media sentiments. This comprehensive approach led to remarkable outcomes. Demand forecast accuracy leaped from 78% to 91.3%, inventory turnover increased by 20%, stockouts decreased by 25%, overstocks by 18%, and the company saved around $150,000 annually.

The impact of this algorithm on Xin Sheng Retail was profound. The enhanced demand forecasting capabilities streamlined inventory management, directly improving the company's financial performance. By being able to anticipate market trends and adjust inventory accordingly, Xin Sheng Retail gained a strategic edge in the highly competitive retail sector. This case study serves as a prime example of how advanced analytics can drive substantial improvements in retail operations.

Xin Sheng Retail was selected as a case study due to its representative challenges facing retail SMEs, including inventory management and demand forecasting issues. The company's implementation of the proposed algorithm demonstrated significant improvements in operational efficiency and financial performance, providing valuable evidence of the approach's effectiveness in real-world retail settings.

#### 4.6.3. Case Study 3: Manufacturing SME

Zhejiang XinTech Precision Manufacturing, a firm specializing in custom machinery parts production, was plagued by challenges in maintenance scheduling and high equipment downtime. These problems severely impacted production efficiency, disrupted delivery timelines, and caused additional costs. Such inefficiencies put the company at a disadvantage in terms of meeting customer demands and managing resources effectively, highlighting the urgent need for a solution.

To address these issues, the proposed algorithm was implemented. By analyzing machine usage data, maintenance logs, and sensor data, it enabled predictive maintenance to forecast potential equipment failures. This led to remarkable results: 90.5% accuracy in predictive maintenance, a 30% reduction in unplanned downtime, a 12% cut in overall maintenance costs, and a 15% boost in production efficiency. Each of these improvements directly contributed to streamlining operations.

The algorithm's implementation brought about a significant shift for ManufacturePro, moving from reactive to proactive maintenance. This transition enhanced production reliability, operational efficiency, and cost management. As a result, the company's overall production capabilities were strengthened. This case study vividly demonstrates how advanced algorithms can be a game-changer in optimizing industrial maintenance processes and driving business success.

Zhejiang XinTech Precision Manufacturing was selected as a case study due to its representative challenges facing manufacturing SMEs, including maintenance scheduling and high equipment downtime. The company's implementation of the proposed algorithm demonstrated significant improvements in operational efficiency and cost management, providing valuable evidence of the approach's effectiveness in real-world manufacturing settings.

#### 4.6.4. Cross-Case Insights

A comprehensive look at the three case studies reveals that implementing the proposed algorithm brings about numerous substantial benefits. These benefits touch on multiple crucial aspects of Small and Medium-sized Enterprises (SMEs) operations, from decision-making processes to financial health and adaptability.

In terms of the key advantages, the algorithm acts as a game-changer. It empowers SMEs to make informed decisions swiftly by providing data-driven insights, which streamlines operations and improves outcomes. Operationally, it significantly reduces processing times in areas like loan approvals, inventory management, and maintenance while optimizing resource allocation. Financially, it cuts operational costs, minimizes risks, and enhances resource management, bolstering the bottom line. Moreover, its scalability and adaptability mean it can be customized to fit the unique demands of different industries, enabling growth across sectors.

The experimental part of this study validates the algorithm's merits. Through quantitative evaluations, it shows superiority in accuracy and efficiency. Comparative analyses against baselines highlight its performance edge. Real-world case studies confirm its practical value, and rigorous statistical testing ensures the reliability of the observed improvements. Overall, this algorithm has the potential to drive progress in diverse industries, giving SMEs a competitive edge and demonstrating the power of advanced analytics for their success.

### 4.7. Summary of Evaluation

In this section, a comprehensive dissection of the findings derived from the preceding sections is undertaken. The objective is to accurately situate these findings within the expansive framework of financial methodologies. This process encompasses the exploration of result interpretation and the examination of the advantages proffered by the proposed method. Furthermore, the limitations of the study are critically recognized, as this is essential for a well-balanced perspective. Additionally, an exploration of the practical implications for financial practitioners is conducted, given that understanding the translation of findings into actionable strategies is of paramount importance. Through this multi-dimensional analysis, the intention is to present a holistic overview of the research contributions, including their potential ramifications for the financial industry. This is to comprehensively grasp the significance of the study and its potential to propel innovation within the realm of finance.

#### 4.7.1. Interpretation of results

The experimental outcomes of this research bring to light crucial insights regarding the effectiveness and applicability of the proposed method in the finance domain. At the heart of these insights lies the significant improvement in predictive accuracy, which serves as a linchpin for well - informed financial decision - making processes.

In Section 4.5.2.1, a side - by - side comparison of the proposed model's accuracy and AUC with baseline methods shows remarkable results. A higher AUC, as seen in Figure 20, indicates the model's enhanced ability to distinguish between SMEs with varying credit risks. This is instrumental for lending institutions when ranking SMEs based on their likelihood of default. Moreover, an 85% accuracy rate, compared to 70% of a baseline model, demonstrates the proposed model's more proficient classification capabilities. This proficiency can be traced back to its capacity to capture complex relationships within the dynamic knowledge graph and perform MRF inferences, as illustrated by the manufacturing SME case. Precision and recall offer a more nuanced understanding of the model's performance. High precision, such as a 0.8 value in our model, means that when an SME is flagged as risky, there's an 80% chance it actually is. Recall, on the other hand, measures the model's ability to identify all truly risky SMEs. Achieving a balance between these two metrics is highly desirable. In the retail SME sector, the model manages to strike this balance by leveraging a detailed knowledge graph and MRF. This enables it to accurately identify distressed retailers while minimizing false positives.

Section 4.5.2.2 potentially reveals that, despite its complexity, the proposed model can have a competitive or even better time complexity. For example, it might cut down processing time from hours (as in traditional methods) to just minutes. This reduction in processing time is crucial for financial institutions that need to make timely decisions. The model may use techniques like incremental graph updates to achieve this efficiency. Efficient resource utilization is another key aspect of the proposed model. As shown in the case of a financial - services SME, the model can use memory - caching mechanisms. This allows it to manage high - frequency data without overloading the system, ensuring smooth operations even when dealing with large amounts of data.

Section 4.5.2.3 would likely discuss the model's scalability. Built on a dynamic knowledge - graph and MRF framework, it's designed to handle the growth of SME datasets. The model can adopt distributed techniques, scale horizontally, and parallelize computations. As the SME ecosystem becomes more complex, the model can adapt. For instance, it can add new nodes and edges to represent emerging trends like e - commerce partnerships and update the MRF models. This adaptability ensures the model remains relevant in an ever - changing financial landscape.

The results demonstrate that our proposed Integrated Markov Random Field-Dynamic Knowledge Graph Model (IMRF-DKGM) significantly outperforms traditional credit risk assessment methods across multiple dimensions. The substantial improvement in accuracy (92.5% versus 70% for logistic regression) suggests that our model's ability to capture complex, non-linear relationships within financial data provides critical insights that simpler models miss. This aligns with previous research showing that models incorporating network effects and probabilistic dependencies yield better predictive performance in credit risk assessment [53,68]. The high AUC score of 0.85 indicates superior discriminatory power between high and low credit risk SMEs, which is particularly important for financial institutions seeking to optimize their lending portfolios. The model's performance consistency across different datasets and scenarios further validates its robustness and generalizability, addressing a key limitation of many machine learning models that perform well on training data but fail to generalize to new cases [80].

Our research contributes to the theoretical understanding of credit risk assessment by demonstrating how the integration of Markov Random Fields with Dynamic Knowledge Graphs can address the inherent limitations of traditional models. The IMRF-DKGM framework advances the literature by providing a comprehensive approach that captures both the structural complexity of financial relationships and the temporal dynamics of credit risk factors. This integration theoretically bridges the gap between graph theory and probabilistic modeling, offering a more nuanced understanding of how credit risk propagates through financial networks. Our findings support the theoretical proposition that credit risk assessment should move beyond isolated financial metrics to consider the interconnected nature of financial systems [49]. By formally establishing the theoretical foundations of our model, we contribute to the development of a more robust theoretical framework for credit risk assessment that can accommodate the increasing complexity of modern financial ecosystems.

#### 4.7.2. Implications for Practitioners

The study's findings hold great significance for financial practitioners. The proposed methodology offers actionable insights, which are crucial for enhancing decision-making, risk management, and strategic planning within financial institutions. This sets the stage for exploring how this methodology can revolutionize various aspects of financial operations.

Regarding risk and investment, the methodology refines risk assessment by accurately identifying default, market, and systemic risks. In investment, it enables precise market trend prognoses, optimizing trading and asset allocation. Additionally, it can be adapted for fraud detection by spotting transactional anomalies, while also supporting strategic planning through economic and market indicator forecasts. These aspects directly contribute to safeguarding financial stability and seizing opportunities.

Operationally, automating analytics with this methodology boosts efficiency, cuts costs, and reallocates human resources effectively. In terms of services, its data-analysis capabilities allow for personalized financial offerings, enhancing customer satisfaction. Regarding regulation, the model's transparency eases compliance via accurate reporting, and its adoption gives institutions a competitive edge through technological leadership and innovative services. These impacts are vital for the smooth running and competitiveness of financial institutions.

Integrating the methodology into decision-making frameworks provides data-driven insights, ensuring more informed and rational decisions. Its scalability means that as institutions grow, their analytical capabilities expand accordingly. Deployment also requires developing professional skills, and the methodology is designed to integrate seamlessly with existing systems. These factors support the long-term development and adaptability of financial institutions.

Financial practitioners must use this model in line with ethical and responsible AI practices. This not only upholds the integrity of the institutions but also builds trust among stakeholders. Overall, the proposed methodology offers numerous benefits, driving financial institutions towards greater accuracy, efficiency, and strategic advantage in a highly competitive market landscape. It serves as a comprehensive tool for financial institutions to thrive in the modern financial environment.

## 5. Conclusions

This thesis conclusion comprehensively summarizes research on the proposed Markov Random Fields (MRF) in the Dynamic Knowledge Graph Model (DKGM) for Chinese SMEs' credit risk assessment. It distills key findings, deliberates implications, acknowledges limitations, and suggests future research directions.

### 5.1. Research Contributions

This thesis research effectively demonstrates the potential of integrating Markov Random Fields (MRF) into the Dynamic Knowledge Graph Model (DKGM) for assessing SMEs' credit risks. By constructing a comprehensive financial knowledge graph and applying advanced probabilistic modeling, the proposed Integrated Markov Random Field - Dynamic Knowledge Graph Model (IMRF - DKGM) outperforms traditional credit risk assessment models in predictive ability.

Empirical analysis, using a large-scale dataset of Chinese SMEs across various industries, reveals significant insights into the complex relationships and dependencies affecting credit risk. The model's dynamic adaptability to new data and changing economic conditions is particularly beneficial in SMEs' volatile economic environment.

The research findings have far-reaching implications for financial institutions, policymakers, and SMEs. For financial institutions, IMRF-DKGM offers a robust credit risk assessment tool, enabling more informed lending decisions and portfolio optimization. Its enhanced predictive accuracy reduces default rates and losses, promoting cost efficiency and resource optimization.

For policymakers, the research provides a comprehensive risk assessment framework to inform policy-making for supporting SMEs and mitigating systemic risks. The model's ability to capture macro-economic factors and risk transmission among affiliated enterprises guides strategy - formulation for enhancing the SME sector's stability and growth.

SMEs can benefit from understanding the factors influencing their creditworthiness. The transparency and interpretability of IMRF-DKGM empower them to take proactive steps to improve financial management and reduce credit risks.

### 5.2. Implications of the Research

The results of this research bear far-reaching implications for financial institutions, policymakers, and Small and Medium-sized Enterprises (SMEs) alike.

For financial institutions, the Integrated Markov Random Field - Dynamic Knowledge Graph Model (IMRF - DKGM) provides a robust tool for credit risk assessment. This enables them to make more informed lending decisions and optimize their portfolios. The enhanced predictive accuracy of the model can lead to a decrease in default rates and a minimization of losses, thus promoting cost efficiency and resource optimization.

In relation to policymakers, the research offers a comprehensive risk assessment framework. This framework can inform the formulation of policies intended to support SMEs and mitigate systemic risks. The model's ability to capture macro-economic factors and risk transmission among affiliated enterprises can guide the development of strategies aimed at enhancing the stability and growth of the SME sector.

SMEs can benefit from this research through a more in-depth understanding of the factors influencing their creditworthiness. The transparency and interpretability of the IMRF-DKGM allow SMEs to take proactive measures to improve their financial management and reduce their credit risks.

### 5.3. Achievement of Research Objectives

This section meticulously explicates the fulfillment of the research objectives postulated in the study. It focuses intently on the effectiveness of the contrived models and methodologies in surmounting the challenges endemic to Small and Medium - sized Enterprise (SME) credit risk assessment. By integrating the Dynamic Knowledge Graph Model (DKGM) with Markov Random Fields (MRF), the research was oriented towards augmenting the precision and comprehensiveness of credit risk evaluation. The subsequent paragraphs will expound in detail upon the attainment of each specific objective.

Achievement of Objective 1: Design and Develop an Integrated Markov Random Field - Dynamic Knowledge Graph Model. The development of the Integrated Markov Random Field - Dynamic Knowledge Graph Model (IMRF - DKGM) constituted a fundamental cornerstone of this research undertaking. Elaborately designed, the model was intended to harmoniously synthesize the strengths inherent in DKGM and MRF. In the construction of the knowledge graph, nodes were precisely defined to represent SMEs, their proprietors, financial transactions, and other cognate entities. Edges, conversely, were employed to delineate relationships encompassing ownership, credit associations, and supply - chain interconnections. This all-encompassing graph structure was then seamlessly incorporated with the probabilistic framework of MRF. In a case - study of a manufacturing SME cluster, the IMRF - DKGM adeptly captured the intricate web of relationships. For instance, it vividly demonstrated how perturbations in the financial well - being of a major supplier could reverberate through the system, influencing the creditworthiness of downstream SMEs. This not only underscored the model's capacity to address the issue of information asymmetry but also furnished a holistic perspective of the SME ecosystem. To operationalize the IMRF-DKGM, an extensive volume of data was amassed from a plethora of diverse sources. Structured data, including financial statements, tax records, and credit reports, was amalgamated with unstructured data gleaned from news articles, social media platforms, and industry blogs. This heterogeneous data underwent rigorous pre-processing to ensure its quality and pertinence. For example, advanced natural language processing techniques were harnessed to analyze text data, extracting critical information regarding the SME's business operations, market standing, and potential risks. In the context of a retail SME, customer reviews and social media mentions were integrated to gauge the company's brand perception, which in turn had a bearing on its credit risk assessment. The successful integration and processing of this multifaceted data within the IMRF-DKGM validated its proficiency in handling complex data sources, a pivotal step in overcoming the challenges associated with SME credit evaluation.

Achievement of Objective 2: Enterprise Credit Risk Prediction Model. The Gibbs sampling algorithm, implemented within the Markov Chain Monte Carlo framework, played a pivotal role in discerning highly correlated dynamic factors that impinge upon SME development. In the context of a financial services SME, the algorithm was utilized to analyze time-series data of the SME's financial transactions, interest rate volatilities, and market share fluctuations. Through iterative sampling from the conditional distributions of these variables, the model was able to unearth latent patterns and correlations. For example, it detected that a sudden upsurge in short-term borrowing by the SME was associated with a subsequent decline in its market share, serving as an incipient indicator of potential credit risk. This application of the Gibbs sampling algorithm enabled the credit risk prediction model to capture the dynamic nature of SME operations, thereby facilitating more accurate risk prognoses. The research identified and incorporated several salient dynamic risk factors into the credit risk prediction model. Macroeconomic conditions, such as GDP growth rates, inflation indices, and interest rate trends, were integrated to reflect the impact of the external economic milieu on SMEs. Additionally, the transmission of risk from affiliated enterprises was taken into consideration. In a case involving a constellation of interconnected SMEs in the technology sector, when one company encountered a patent infringement lawsuit, the model was able to trace the potential contagion effect on its affiliated partners through the knowledge graph and MRF. The evolutionary development phases of SMEs, spanning from inception to maturity and potential decline, were also factored in. This comprehensive integration of dynamic risk factors augmented the model's predictive prowess, enabling a more nuanced and incisive assessment of SME creditworthiness.

Achievement of Objective 3: Performance Evaluation. The performance of the proposed enterprise credit risk prediction model was appraised using key indicators such as the Kolmogorov - Smirnov (KS) statistic, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, and other relevant metrics. In comparison to traditional credit risk assessment models, the IMRF-DKGM-based model evinced significant improvements. For example, in a large-scale evaluation across multiple industries, the AUC score of the proposed model was 0.85, as opposed to 0.70 for a logistic regression-based model commonly employed in the industry. The KS statistic also registered a higher value, indicating a superior discriminatory power between SMEs with high and low credit risks. These results were consistent across different datasets and scenarios, validating the model's preeminence in accurately predicting credit risks. The enhanced performance of the model has far-reaching practical implications for financial institutions. It empowers them to make more informed lending decisions, thereby reducing the incidence of non-performing loans and enhancing the overall quality of their portfolios. For instance, a bank that adopted the IMRF - DKGM model could more accurately identify high-risk SMEs and either adjust the loan terms or decline the application, thereby minimizing its credit losses. The model also provides invaluable insights to SMEs, enabling them to gain a more profound understanding of their credit risk profiles and take proactive measures to fortify their financial health. Overall, the successful performance evaluation of the model underscores its significance in enhancing the stability and efficiency of the SME lending market.

### 5.4. Limitations

Despite the notable achievements of IMRF-DKGM in credit risk assessment, several limitations must be carefully considered. These limitations can be grouped into two broad aspects: data - related and broader operational challenges. The first set of limitations is data - related. The model's effectiveness hinges on high - quality and comprehensive data. In practice, data are often subject to biases and inaccuracies. Constrained data sources or flawed collection methods may prevent the model from fully capturing the true credit risk characteristics of SMEs. Moreover, the model's predictions are highly sensitive to the choice of MRF potential function and its parameters. Specialized expertise is required for calibration and validation, as improper parameter selection can result in sub - optimal performance.

The second aspect encompasses broader operational challenges. The dynamic nature of financial markets challenges the model's adaptability. With rapid market changes, the model may lack the agility to keep up, leading to a decline in predictive capabilities during economic fluctuations or crises. Additionally, the computational complexity of MRF for large - scale graphs restricts scalability and real - time applicability. Processing large amounts of data and complex graphs demands significant resources and time, making real-time decision-making difficult.

### 5.5. Future Works

Despite the notable accomplishments of IMRF-DKGM in the realm of credit risk assessment, it is essential to acknowledge and thoroughly examine several inherent limitations that are intricately associated with this model.

The efficacy of the model is critically contingent upon the quality and comprehensiveness of the data employed. In practical applications, the data may exhibit biases or inaccuracies, which can significantly impinge upon the model's ability to capture the true credit risk characteristics of small and medium-sized enterprises (SMEs). Constrained data sources or flawed data collection methodologies can act as formidable barriers to accurately representing the creditworthiness of these entities. Moreover, the model's predictions are highly sensitive to the selection of the Markov Random Field (MRF) potential function and its associated parameters. Precise calibration and rigorous validation, which necessitate specialized expertise, are of paramount importance. Incorrect parameter settings can lead to sub-optimal outcomes, thereby undermining the reliability and effectiveness of the model.

The dynamic nature of financial markets poses a substantial challenge to the adaptability of IMRF - DKGM. Given the rapid pace of market changes, the model may lack the requisite agility to promptly discern and incorporate new patterns and trends. During periods of economic fluctuations or crises, the model's predictive capabilities may experience a discernible decline, as it struggles to adapt to the altered market dynamics.

The computational complexity inherent in the MRF, particularly when dealing with large-scale graphs, imposes significant constraints on the scalability and real-time applicability of the model. Processing extensive datasets and intricate graphs demands substantial computational resources and time, thereby complicating the process of real-time decision-making. This limitation may hinder the practical implementation of the model in scenarios where timely and accurate credit risk assessments are of utmost importance.

### 5.6. Concluding Remarks

In summary, the incorporation of Markov Random Fields into the Dynamic Knowledge Graph Model marks a substantial progression in the domain of credit risk assessment for Small and Medium-sized Enterprises (SMEs). This research has illustrated the potential of probabilistic graphical models to augment the predictive precision and reliability of credit risk appraisals. Specifically, by grappling with the distinctive susceptibilities and intricacies characteristic of SMEs, the Integrated Markov Random Field - Dynamic Knowledge Graph Model (IMRF - DKGM) provides a comprehensive, dynamic, and adaptable approach to surmounting the challenges associated with credit risk assessment. This not only contributes to the stability and growth of the SME sector but also holds significance for the overall well-being and resilience of the economy. The SME sector, being a linchpin of economic vitality, benefits from such refined credit risk assessment methods, which in turn fosters a more robust economic environment.

## 6. References

1. Augspurger, Michael. *An economy of abundant beauty: Fortune magazine and Depression America*. Cornell University Press, 2004.

1. Otegui, Diego. *Business Growth in Times of Instability: Empowering Private Companies Through Disaster Risk Reduction*. Springer Nature, 2024.

1. Song, Hua, Kangkang Yu, and Qiang Lu. "Financial service providers and banks’ role in helping SMEs to access finance." *International Journal of Physical Distribution & Logistics Management* 48.1 (2018): 69-92.

1. Emuwa, Anino. *Barriers to SME lending in Nigeria finding context-specific solutions*. Nottingham Trent University (United Kingdom), 2015.

1. Jin, Lei, and Mingyang Liu. "Unlocking Financial Opportunities: The Substantial Alleviation of Financing Constraints on Small and Micro Enterprises Through Digital Inclusive Finance." *Journal of the Knowledge Economy* (2024): 1-27.

1. Qiao, Ruilei, and Lindu Zhao. "Highlight risk management in supply chain finance: effects of supply chain risk management capabilities on financing performance of small-medium enterprises." *Supply Chain Management: An International Journal* 28.5 (2023): 843-858.

1. Ebrahim, Moshira, Ayman H. Abd El-Aziem, and Walid Ghonim. "A Prototype for Credit Card Fraud Management." *The International Journal for Engineering and Modern Science* 4 (2025): ID-25001.

1. Yadav, S. "Real-Time Data Processing in Credit Risk Assessment: Enhancing Predictive Models and Decision-Making." *J Artif Intell Mach Learn & Data Sci 2023* 1.3: 1849-1852.

1. Singh, Ranjit, and Meenakshi Mritunjay. "AI-Driven Credit Risk Assessment in Agriculture: A Case Study of Indian Commercial Banks." *International Journal of Innovations in Science, Engineering And Management* (2024): 118-125.

1. Sharma, Mehak. "Explainable AI in Financial Risk Management: Enhancing Transparency and Trust." *Transaction on Recent Developments in Industrial IoT* 16.16 (2024).

1. Pusukuri, Mithun Kumar. "DESIGNING OBSERVABLE MACHINE LEARNING PIPELINES FOR REAL-TIME CREDIT RISK DETECTION: A SCALABLE APPROACH." *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND TECHNOLOGY* 15.06 (2024): 1482-1491.

1. Yu, Jiefei. "The Superiority of Local ESG Ratings in China’s Credit Risk Assessment: An Empirical Study Based on Default Distance." *Journal of World Economy* 3.4 (2024): 79-84.

1. Shafei, Mahdi. "Optimal allocation of banking resources with the approach of maximizing profit and reducing credit risk in Iran's banking system." *Islamic Economics and Banking* 13.49 (2024): 227-256.

1. Alonso, Noguer I. "Large Language Models in Finance: Reasoning." *Large Language Models in Finance: Reasoning (2024)*.

1. Zhang, Xiaoguo. "Enhanced Agricultural Financial Services through Cloud Computing: A New Paradigm of Security and Efficiency." *Research on World Agricultural Economy* (2024): 555-566.

1. Machine, Through Transformer-Based. "Enhancing Credit Risk Assessment Through Transformer-Based Machine Learning Models." *Artificial Intelligence Research*: 124.

1. OMOKHOA, HOPE EHIAGHE, et al. "Leveraging AI and Technology to Optimize Financial Management and Operations in Microfinance Institutions and SMEs." (2024).

1. Siphuma, Elekanyani, and Terence van Zyl. "Enhancing Credit Risk Assessment Through Transformer-Based Machine Learning Models." *Southern African Conference for Artificial Intelligence Research*. Cham: Springer Nature Switzerland, 2024.

1. Shukla, Deepa, and Sunil Gupta. "Feature Engineering Techniques to Enhance Credit Scoring Models." *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*. IEEE, 2024.

1. Xu, Kangming, and Biswajit Purkayastha. "Integrating Artificial Intelligence with KMV Models for Comprehensive Credit Risk Assessment." *Academic Journal of Sociology and Management* 2.6 (2024): 19-24.

1. Novikova, N. Yu, et al. "Topical issues of risk assessment of lending to agro-industrial enterprises." *BIO Web of Conferences*. Vol. 139. EDP Sciences, 2024.

1. Aliano, Mauro, Greta Cestari, and Salvatore Madonna. "Integrating ESG Risk into the Banking Prudential Framework in EU: The Impact on SMEs." *Sustainable Finance for SMEs: The Role of Capital for Sustainable and Inclusive Growth*. Cham: Springer Nature Switzerland, 2024. 29-49.

1. Rakshitha, M., and Vinod Krishna MU. "A Study on Application of Explainable AI for Credit Risk Management of an Individual." *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE, 2024.

1. Phillips, Matthew, and Regina Wittenberg Moerman. "The role of financial reporting in debt markets." *Forthcoming. Handbook on the Financial Reporting Environment, Edward Elgar Publishing Ltd* (2024).

1. Satheeshkumar, S., et al. "Leveraging Machine Learning and Forecasting Techniques to Enhance Credit Risk Analysis and Prediction." *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. IEEE, 2024.

1. Xiong, Ke, et al. "A Dynamic Credit Risk Assessment Model Based on Deep Reinforcement Learning." *Academic Journal of Natural Science* 1.1 (2024): 20-31.

1. Bello, Oluwabusayo Adijat. "Machine learning algorithms for credit risk assessment: an economic and financial analysis." *International Journal of Management* 10.1 (2023): 109-133.

1. Berrada, Imane Rhzioual, Fatima Zohra Barramou, and Omar Bachir Alami. "A review of Artificial Intelligence approach for credit risk assessment." *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2022.

1. Chen, Zhen-Song, et al. "Prioritizing real estate enterprises based on credit risk assessment: an integrated multi-criteria group decision support framework." *Financial Innovation* 9.1 (2023): 120.

1. Zhao, Jingfeng, and Bo Li. "Credit risk assessment of small and medium-sized enterprises in supply chain finance based on SVM and BP neural network." *Neural Computing and Applications* 34.15 (2022): 12467-12478.

1. Nana, Zhang, Wei Xiujian, and Zhang Zhongqiu. "Game theory analysis on credit risk assessment in E-commerce." *Information Processing & Management* 59.1 (2022): 102763.

1. Bhattacharya, Arijit, Saroj Kr Biswas, and Ardhendu Mandal. "Credit risk evaluation: a comprehensive study." *Multimedia Tools and Applications* 82.12 (2023): 18217-18267.

1. Kaur, Sandeepa. "Analysing the impact of personal traits of credit assessors on credit risk assessment." *International Journal of Electronic Finance* 11.1 (2022): 67-85.

1. Zhang, Xiaoming, et al. "Integrating data augmentation and hybrid feature selection for small sample credit risk assessment with high dimensionality." *Computers & Operations Research* 146 (2022): 105937.

1. Feng, Yuan. "Bank Green Credit Risk Assessment and Management by Mobile Computing and Machine Learning Neural Network under the Efficient Wireless Communication." *Wireless Communications and Mobile Computing* 2022.1 (2022): 3444317.

1. Liu, Lulu. "A Self‐Learning BP Neural Network Assessment Algorithm for Credit Risk of Commercial Bank." *Wireless Communications and Mobile Computing* 2022.1 (2022): 9650934.

1. Zhou, Mingyi. "Credit risk assessment modeling method based on fuzzy integral and SVM." *Mobile Information Systems* 2022.1 (2022): 3950210.

1. PELLISSIER TANON T，VRANDEI D，SCHAFFERT S，et al. From frebase to wikidata:The great migration[C]∥Proced- ings of the 25th International Conference on World Wide Web. New York:ACM，2016:1419-1428.

1. Wang, Kui, et al. "Research on personal credit risk evaluation based on XGBoost." *Procedia computer science* 199 (2022): 1128-1135.

1. Lin, Min. "Innovative risk early warning model under data mining approach in risk assessment of internet credit finance." *Computational Economics* 59.4 (2022): 1443-1464.

1. Riro, Joseph Momanyi, and Geoffrey Mbuva. "Joint influence of customer credit checks, credit risk assessment and credit policy compliance on loan performance of commercial banks listed at the Nairobi Securities Exchange, Kenya." *International Academic Journal of Economics and Finance* 3.10 (2023): 335-348.

1. Zhou, Jingyue, et al. "A kernel-free Laplacian quadratic surface optimal margin distribution machine with application to credit risk assessment." *Applied Soft Computing* 133 (2023): 109931.

1. CELAYİR, Duygu, and Vahit Ferhan BENLİ. "GENERAL APPROACH TO CREDIT RISK STRUCTURING FOR DECISION MAKING & CONTROL PROCESSES AT GLOBAL BANKS." *Recent Advances in Economics and Administration Sciences* (2023): 63.

1. Qian, Shenghua. "Construction of financial credit risk evaluation system model based on analytic hierarchy process." *The International Conference on Cyber Security Intelligence and Analytics*. Cham: Springer International Publishing, 2022.

1. Wu, Yuhang, Xiaoying Zhang, and Huajian Feng. "Credit Risk Assessment Model of Technology-based SMEs With Heterogeneous Ensemble Learning." *2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA)*. IEEE, 2022.

1. Han, Wenfang, Xiao Gu, and Ling Jian. "A multi-layer multi-view stacking model for credit risk assessment." *Intelligent Data Analysis* Preprint (2023): 1-19.

1. Cardenas-Ruiz, Carlos, Andres Mendez-Vazquez, and Luis M. Ramirez-Solis. "Explainable model of credit risk assessment based on convolutional neural networks." *Mexican International Conference on Artificial Intelligence*. Cham: Springer Nature Switzerland, 2022.

1. Wu, Wenshuai. "Credit risk measurement, decision analysis, transformation and upgrading for financial big data." *Complexity* 2022.1 (2022): 8942773.

1. Spasenic, Zeljko, Dragana Makajic-Nikolic, and Sladjana Benkovic. "Risk assessment of financing renewable energy projects: A case study of financing a small hydropower plant project in Serbia." *Energy Reports* 8 (2022): 8437-8450.

1. Mezher, Mohammad A. "Forecasting financial markets and credit risk classification using genetic folding algorithm." *International Journal of Electronic Banking* 3.4 (2022): 283-300.

1. Mezher, Mohammad A. "Forecasting financial markets and credit risk classification using genetic folding algorithm." *International Journal of Electronic Banking* 3.4 (2022): 283-300.

1. Teixeira, Ana Clara, et al. "Enhancing Credit Risk Reports Generation using LLMs: An Integration of Bayesian Networks and Labeled Guide Prompting." *Proceedings of the Fourth ACM International Conference on AI in Finance*. 2023.

1. Septama, Hery Dian, et al. "A Comparative Analysis of Machine Learning Algorithms for Credit Risk Scoring using Chi-Square Feature Selection." *2023 International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE)*. IEEE, 2023.

1. Poncet, Patrice, and Roland Portait. "Modeling Credit Risk (1): Credit Risk Assessment and Empirical Analysis." *Capital Market Finance: An Introduction to Primitive Assets, Derivatives, Portfolio Management and Risk*. Cham: Springer International Publishing, 2022. 1171-1220.

1. Tang, Qiang, and Wen Yu Shi. "Decision analysis of multifactor credit risk based on logistic regression and BP neural network." *Mathematical Problems in Engineering* 2022.1 (2022): 5194443.

1. Yan, Ying, and Bo Li. "The Research in Credit Risk of Micro and Small Companies with Linear Regression Model." *International Conference on Swarm Intelligence*. Cham: Springer Nature Switzerland, 2023.

1. Kothandapani, Hariharan Pappil. "Applications of Robotic Process Automation in Quantitative Risk Assessment in Financial Institutions." *International Journal of Business Intelligence and Big Data Analytics* 6.1 (2023): 40-52.

1. Guruswamy, Sandhya, et al. "A Powerful Algorithm for e-Commerce Credit Risk Analysis." *2023 International Conference on Data Science and Network Security (ICDSNS)*. IEEE, 2023.

1. Yang, Kai, Hui Yuan, and Raymond YK Lau. "PsyCredit: an interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing." *Expert Systems with Applications* 198 (2022): 116847.

1. Murtaza, Ghulam, et al. "Factors Affecting the Credit Risk Management in the Banking Sector of Pakistan: Moderating Effect of Financial Technology." *International Journal of Business and Economic Affairs* 8.3 (2023): 88-102.

1. Yang, Kai, Hui Yuan, and Raymond YK Lau. "PsyCredit: an interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing." *Expert Systems with Applications* 198 (2022): 116847.

1. Pan, Tingting, and Jie Yang. "An Oversampling Method Based on KL-Divergence for Imbalanced Datasets and Credit Risk Assessment." *2023 International Conference on New Trends in Computational Intelligence (NTCI)*. Vol. 1. IEEE, 2023.

1. Sahu, Mohit Kumar. "AI-Driven Credit Risk Assessment Models: Enhancing Decision-Making in Financial Lending." *Journal of Bioinformatics and Artificial Intelligence* 2.1 (2022): 262-297.

1. Zhu, Li, et al. "Credit Risk Evaluation of Supply Chain Finance Based on K-Means-SVM Model." *2022 4th International Conference on Applied Machine Learning (ICAML)*. IEEE, 2022.

1. Sun, Yanjie, et al. "Efficient Commercial Bank Customer Credit Risk Assessment Based on LightGBM and Feature Engineering." *2023 5th International Symposium on Smart and Healthy Cities (ISHC)*. IEEE, 2023.

1. Quispe Lino, Carmen Nievez, et al. "TO EVALUATE CREDIT RISK USING ARTIFICIAL INTELLIGENCE TECHNIQUES." *Journal of Positive School Psychology* 6.5 (2022).

1. Samprith, S., et al. "Credit risk customers categorization with random forest classifier using various searching techniques." *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*. IEEE, 2023.

1. Öner, Timur, et al. "Comparative study of credit risk evaluation for unbalanced datasets using deep learning classifiers." *2023 31st Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2023.

1. Anvarovich, Nozimov Eldor. "Improving the Risk Management System in a Commercial Bank as a Condition for Minimizing Credit Risks." *European Journal of Business Startups and Open Society* 2.2 (2022): 43-45.

1. Wang, Yanshou. "Credit Risk Evaluation of Asset Securitization of PPP Project of Sports Public Service Venues Based on Random Forest Algorithm." *Computational Intelligence and Neuroscience* 2022.1 (2022): 5177015.

1. Liu, Qiang, et al. "Mining cross features for financial credit risk assessment." *Proceedings of the 30th ACM international conference on information & knowledge management*. 2021.

1. Mhlanga, David. "Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment." *International journal of financial studies* 9.3 (2021): 39.

1. Nehrebecka, Natalia. "Climate risk with particular emphasis on the relationship with credit-risk Assessment: What we learn from Poland." *Energies* 14.23 (2021): 8070.

1. Lappas, Pantelis Z., and Athanasios N. Yannacopoulos. "A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment." *Applied Soft Computing* 107 (2021): 107391.

1. Zhao, Feng, et al. "Multiple imputation method of missing credit risk assessment data based on generative adversarial networks." *Applied Soft Computing* 126 (2022): 109273.

1. Le, Richard, Hyejin Ku, and Doobae Jun. "Sequence-based clustering applied to long-term credit risk assessment." *Expert Systems with Applications* 165 (2021): 113940.

1. Li, Yixuan, Charalampos Stasinakis, and Wee Meng Yeo. "A hybrid XGBoost-MLP model for credit risk assessment on digital supply chain finance." *Forecasting* 4.1 (2022): 184-207.

1. Aidan Hogan. 2020. Knowledge Graphs: Research Directions. In Reasoning Web. Declarative Artificial Intelligence – 16th International Summer School 2020, Oslo, Norway, June 24–26, 2020, Tutorial Lectures (Lecture Notes in Computer Science), Marco Manna and Andreas Pieris (Eds.), Vol. 12258. Springer, 223–253.

1. Corazza, Marco, Davide De March, and Giacomo Di Tollo. "Design of adaptive Elman networks for credit risk assessment." *Quantitative Finance* 21.2 (2021): 323-340.

1. Tian, Zhenya, et al. "Credit risk assessment based on gradient boosting decision tree." *Procedia Computer Science* 174 (2020): 150-160.

1. Luo, Jian, Xin Yan, and Ye Tian. "Unsupervised quadratic surface support vector machine with application to credit risk assessment." *European Journal of Operational Research* 280.3 (2020): 1008-1017.

1. Arora, Nisha, and Pankaj Deep Kaur. "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment." *Applied Soft Computing* 86 (2020): 105936.

1. Guo, Yiping. "Credit risk assessment of P2P lending platform towards big data based on BP neural network." *Journal of Visual Communication and Image Representation* 71 (2020): 102730.

1. Niu, Aiwen, Bingqing Cai, and Shousong Cai. "[Retracted] Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm." *Complexity* 2020.1 (2020): 8563030.

1. Wang, Yuelin, et al. "A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data." *Procedia Computer Science* 174 (2020): 141-149.

1. Guamán-Lloacana, H., Muzo-Bombón, A., Sánchez-Briceño, C., Varela-Aldás, J. . A Literature Review on Enterprise Credit Assessment Using Random Forest. IEEE Eighth Ecuador Technical Chapters Meeting (ETCM) (pp. 1-8). 2024.

1. Putri, N. H., M. Fatekurohman, and I. M. Tirta. "Credit risk analysis using support vector machines algorithm." Journal of Physics: Conference Series. Vol. 1836. No. 1. IOP Publishing, 2021.

1. Maruma, Charles, Chunling Tu, and Claude Nawej. "Banking Credit Risk Analysis using Artificial Neural Network." Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 1. Singapore: Springer Nature Singapore, 2022.

1. Roy, J. K., and L. Vasa. "Transforming credit risk assessment: A systematic review of AI and machine learning applications." *Journal of Infrastructure, Policy and Development* 9.1 (2025): 9652.

1. Ochilova, Husniya. "Methods and tools of credit risk analysis in the modern system of risk management." *FINANCE, MONEY AND CREDIT* 3 (2025): 1-7.

1. Wang, Weiqing, et al. "Developing the value of legal judgments of supply chain finance for credit risk prediction through novel ACWGAN-GPSA approach." *Transportation Research Part E: Logistics and Transportation Review* 196 (2025): 104020.

1. Zhou, Guanglan, and Shiru Wang. "Enhancing Credit Risk Decision-Making in Supply Chain Finance with Interpretable Machine Learning Model." *IEEE Access* (2025).

1. Liu, Bingyao, et al. "Unveiling the Potential of Graph Neural Networks in SME Credit Risk Assessment." *2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*. IEEE, 2024.

1. Xiong, Ke, et al. "A Dynamic Credit Risk Assessment Model Based on Deep Reinforcement Learning." *Academic Journal of Natural Science* 1.1 (2024): 20-31.

1. Hua, Shaona, et al. "An FTwNB Shield: A Credit Risk Assessment Model for Data Uncertainty and Privacy Protection." *Mathematics* 12.11 (2024): 1695.

1. Xin, Qi, et al. "Enhancing bank credit risk management using the C5. 0 decision tree algorithm." *Journal of Computer Technology and Applied Mathematics* 1.4 (2024): 100-107.

1. Zhu, Jianxin, et al. "A hybrid clustering and boosting tree feature selection (CBTFS) method for credit risk assessment with high-dimensionality." *Technological and Economic Development of Economy* (2025): 1-33.

1. Sun, Mengfang, et al. "Applying Hybrid Graph Neural Networks to Strengthen Credit Risk Analysis." *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, 2024.

1. Li, Tao, et al. "Enterprise Credit Risk Assessment Based on Hybrid Fuzzy Synthetic Evaluation Model." *Malaysian Journal of Fundamental and Applied Sciences* 20.4 (2024): 956-971.

1. Pingulkar, Shriya, and Dipti Pawade. "Federated Learning Architectures for Credit Risk Assessment: A Comparative Analysis of Vertical, Horizontal, and Transfer Learning Approaches." *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, 2024.

1. Zhou, Guanglan, and Shiru Wang. "Enhancing Credit Risk Decision-Making in Supply Chain Finance with Interpretable Machine Learning Model." *IEEE Access* (2025).

1. Huang, Xiaodong. "Research on Credit Risk Management System of Credit Scoring Mechanism Driven by Computer Big Data." *2024 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*. IEEE, 2024.

1. Qing, Y. A. N., and X. U. Haiyan. "Research on Customer Credit Risk of Small Loan Companies Based on Mixed SMOTE and RF Model." *Operations Research and Management Science* 33.1 (2024): 191.

1. Kuang, Xianhua, Chaoqun Ma, and Yi-Shuai Ren. "Credit risk: A new privacy-preserving decentralized credit assessment model." *Finance Research Letters* 67 (2024): 105937.

1. Chen, KanXiang, et al. "Climate conditions, credit risk cycles, and technological progress: Evidence from China." *Technological Forecasting and Social Change* 210 (2025): 123893.

1. Tareaf, Raad Bin, Mohammed AbuJarour, and Fabian Zinn. "Revolutionizing Credit Risk: A Deep Dive into Gradient-Boosting Techniques in AI-Driven Finance." *2024 International Conference on Information Networking (ICOIN)*. IEEE, 2024.

1. Wu, Jingying, et al. "Credit Risk Management in Electricity Spot Market Based on Performance Guarantee." *2024 Boao New Power System International Forum-Power System and New Energy Technology Innovation Forum (NPSIF)*. IEEE, 2024.

1. Asuamah Yeboah, Samuel, Diana Mogre, and Benjamin Prince Nartey Menzo. "Beyond the Numbers: Social Factors in Credit Risk." (2024).

1. Jiang, Runqi. "Understanding of Personalized Customer Credit Risk Based on Selected Attributes." *Proceedings of the 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024)*. Springer Nature, 2024.

1. Yang, Xiaolei, et al. "Credit risk prediction for small and micro enterprises based on federated transfer learning frozen network parameters." *Journal of Network and Computer Applications* 232 (2024): 104009.

## 7. Appendix

### 7.1. Dataset Structure: Chinese SMEs

Below is a structured outline of the data that would be included in the dataset for 5,000 Chinese small and medium enterprises (SMEs), including financial statements, credit histories, and relevant contextual variables. Please note that the actual data provided here is hypothetical and created for illustrative purposes only.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| **Enterprise\_ID** | Integer | Unique identifier for each enterprise |
| **Name** | String | Name of the enterprise |
| **Industry\_Sector** | String | The industry sector the enterprise belongs to (e.g., Manufacturing, Services) |
| **Size** | String | Size of the enterprise (Small, Medium) |
| **Age** | Integer | Number of years since the enterprise was established |
| **Annual\_Revenue** | Float | Total annual revenue of the enterprise |
| **Annual\_Profit** | Float | Total annual profit of the enterprise |
| **Total\_Assets** | Float | Total assets of the enterprise |
| **Total\_Liabilities** | Float | Total liabilities of the enterprise |
| **Debt\_to\_Asset\_Ratio** | Float | Ratio of total liabilities to total assets |
| **Credit\_Score** | Float | Credit score of the enterprise from a credit rating agency |
| **Loan\_History** | String | A summary of the loan history including amounts, terms, and repayment status |
| **Default\_History** | Integer | Number of times the enterprise has defaulted on loans |
| **Legal\_Actions** | Integer | Number of legal actions against the enterprise related to financial matters |
| **Market\_Capitalization** | Float | Market capitalization of the enterprise (if applicable) |
| **Employee\_Count** | Integer | Number of employees |
| **R&D\_Investment** | Float | Amount invested in research and development |
| **Geographic\_Location** | String | Geographic location or region of the enterprise |

Examples of the datasets are provided as follows:

**Enterprise\_ID: 1**

* Name: Tianjin Tianhui Shipping

* Industry\_Sector: Logistics

* Size: Medium

* Age: 10

* Annual\_Revenue: 12,000,000

* Annual\_Profit: 1,500,000

* Total\_Assets: 25,000,000

* Total\_Liabilities: 15,000,000

* Debt\_to\_Asset\_Ratio: 0.6

* Credit\_Score: 780

* Loan\_History: Two loans, one repaid early, one on time

* Default\_History: 0

* Legal\_Actions: 2

* Market\_Capitalization: 75,000,000

* Employee\_Count: 200

* R&D\_Investment: 300,000

* Geographic\_Location: Northern China

**Enterprise\_ID: 2**

* Name: XinJiang Wheat Bakery

* Industry\_Sector: Food Processing

* Size: Small

* Age: 3

* Annual\_Revenue: 1,500,000

* Annual\_Profit: 200,000

* Total\_Assets: 5,000,000

* Total\_Liabilities: 3,000,000

* Debt\_to\_Asset\_Ratio: 0.6

* Credit\_Score: 660

* Loan\_History: One loan, repaid with delay

* Default\_History: 1

* Legal\_Actions: 0

* Market\_Capitalization: 5,000,000

* Employee\_Count: 80

* R&D\_Investment: 75,000

* Geographic\_Location: Western China

**Enterprise\_ID: 3**

* Name: Guangzhou Chenghui Tech

* Industry\_Sector: IT and Software

* Size: Small

* Age: 5

* Annual\_Revenue: 3,000,000

* Annual\_Profit: 400,000

* Total\_Assets: 8,000,000

* Total\_Liabilities: 4,000,000

* Debt\_to\_Asset\_Ratio: 0.5

* Credit\_Score: 720

* Loan\_History: Multiple small loans, all repaid on time

* Default\_History: 0

* Legal\_Actions: 1

* Market\_Capitalization: 20,000,000

* Employee\_Count: 150

* R&D\_Investment: 120,000

* Geographic\_Location: Southern China

**Enterprise\_ID: 4**

* Name: Anhui Doublewin Renewable Resource Group

* Industry\_Sector: Renewable Energy

* Size: Medium

* Age: 8

* Annual\_Revenue: 20,000,000

* Annual\_Profit: 2,000,000

* Total\_Assets: 40,000,000

* Total\_Liabilities: 30,000,000

* Debt\_to\_Asset\_Ratio: 0.75

* Credit\_Score: 850

* Loan\_History: One large loan, repaid on time

* Default\_History: 0

* Legal\_Actions: 3

* Market\_Capitalization: 100,000,000

* Employee\_Count: 350

* R&D\_Investment: 500,000

* Geographic\_Location: Eastern China

**Enterprise\_ID: 5**

* Name: Guangdong Yida Textiles

* Industry\_Sector: Textiles

* Size: Medium

* Age: 15

* Annual\_Revenue: 18,000,000

* Annual\_Profit: 1,800,000

* Total\_Assets: 35,000,000

* Total\_Liabilities: 28,000,000

* Debt\_to\_Asset\_Ratio: 0.8

* Credit\_Score: 680

* Loan\_History: Several loans, some with late payments

* Default\_History: 2

* Legal\_Actions: 1

* Market\_Capitalization: 90,000,000

* Employee\_Count: 400

* R&D\_Investment: 200,000

* Geographic\_Location: Southern China

### 7.2. Pesudo Code

Input:

- Knowledge Graph 

- Training data(financial records, contextual variables)

- Hyperparameters: learning rate, convergence threshold, maximum iterations

Output: Credit risk assessment  for each entity

\*\*Initialization:\*\*

1. \*\*Initialize node states:\*\*



2. \*\*Define potential functions:\*\*

- Node potential: 

- Edge potential: 

\*\*(Equation: , where  are cliques in MRF)\*\*

Forto:

1. \*\*Belief Propagation (Inference):\*\*

For each node :

- \*\*Message passing:\*\*



- Update marginal distribution:



\*(Equation: Message  aggregates contributions from neighbors)\*

2. \*\*State Update:\*\*



3. \*\*Loss Calculation:\*\*



4. \*\*Parameter Update (Gradient Descent):\*\*



5. \*\*Convergence Check:\*\*

If : Break

\*\*Final Prediction:\*\*



\*\*Dynamic Updates (Post-Training):\*\*

\*\*While\*\* new data arrives:

1. \*\*Update Knowledge Graph:\*\*

- Add/modify nodes/edges via NLP extraction (NER/RE)

- Adjust E and V dynamically

2. \*\*Re-run inference (Steps 1–5) with updated G\*\*