WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online) | Volume 5, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijgem.v5n2.30
Journal homepage: https://wepub.org/index.php/IJGEM/index

Machine Learning Enterprise Financial Intelligent Risk Control System Based on New Database

Hongyi Guo, Yanjun Liu, Ludong Hu, Xuanbo Zhang

Inspur Yunzhou Industrial Internet Co., Ltd., Shandong, Jinan, China

ABSTRACT

This article concentrates on the field of financial investment service technology and presents an intelligent enterprise financial risk control system based on the Milvus vector database with big data and machine learning. It employs algorithms such as random forest, Euclidean distance, and cosine similarity. Through a series of operations including meticulously designing the risk assessment system and developing the risk prediction system, it is committed to accurately analyzing and predicting risk situations such as cancellation or revocation that enterprises may face. This provides strong support for enterprise risk management and control, helps enterprises achieve sustainable development in the complex and changeable market environment, enhances enterprise stability and competitiveness, and strengthens the enterprise's ability to deal with risks, enabling them to make more effective strategic decisions and ensure long-term stability.

KEYWORDS

Milvus; Machine learning; Using random forest; Euclidean distance; Cosine similarity

1. INTRODUCTION

In modern enterprise management, establishing a robust and effective risk early warning system has become an essential component of strategic decision-making. With the complexities and uncertainties of today's business environment, enterprises are exposed to a wide range of risks that can arise from both internal and external factors. Among these, legal and financial risks are particularly critical, as they can significantly impact the stability, reputation, and long-term viability of a business. The ability to identify and assess these risks in a timely manner is paramount, as it allows decision-makers to implement targeted measures to mitigate potential threats, safeguard assets, and maintain operational continuity .

This article delves into the development of an innovative enterprise risk early warning system, leveraging the capabilities of the Milvus vector database and state-of-the-art machine learning models. The proposed system focuses on analyzing and predicting two specific risk scenarios: whether an enterprise is at risk of cancellation or revocation. Cancellation and revocation risks often serve as indicators of deeper underlying issues, such as financial distress, regulatory non-compliance, or deteriorating market conditions. Accurately forecasting these risks can provide enterprises with the foresight needed to address root causes proactively and minimize negative outcomes .

The integration of the Milvus vector database into this system is a key enabler of its advanced functionality. Milvus is designed to handle high-dimensional vector data efficiently, making it ideal for storing and retrieving the feature-rich data required for enterprise risk analysis. By combining this with machine learning models that excel in pattern recognition and prediction, the system offers a comprehensive solution that not only identifies potential risks but also quantifies their likelihood and

severity. This approach enhances the precision and reliability of risk assessments, empowering enterprises to make data-driven decisions with greater confidence [6].

Beyond risk identification, this system brings significant improvements to the efficiency and effectiveness of risk management processes. Traditional risk management methods often rely on manual analysis, which can be time-consuming, resource-intensive, and prone to errors. In contrast, the automated nature of this system allows for real-time analysis and continuous monitoring of enterprise data [2-4]. This ensures that risks are flagged promptly, giving decision-makers ample time to respond. Additionally, the system's ability to process large volumes of data enables it to scale seamlessly as the enterprise grows, ensuring its relevance and utility over the long term [5, 7].

More importantly, the system contributes to the sustainable development of enterprises by fostering resilience and adaptability. By providing early warnings of potential risks, the system helps enterprises allocate resources more effectively, optimize their operational strategies, and strengthen compliance efforts [8]. This not only reduces the likelihood of disruptions but also enhances the organization's ability to navigate challenges and seize opportunities in a competitive marketplace. Ultimately, the system serves as a vital tool for modern enterprises, enabling them to align risk management with strategic objectives and secure a foundation for sustained success..

1.1. Designing an Enterprise Risk Assessment System: Feature Engineering, Target Variable Definition, and Vector Database Integration

The design of the system necessitates the careful definition of a comprehensive set of features, which will serve as the core input data for analysis and model training. Selecting appropriate features is a critical step, as the quality and relevance of these features directly influence the prediction performance and generalization ability of the machine learning models [1, 2]. In this study, we have identified a range of features that comprehensively capture various dimensions of an enterprise's profile. These include basic information such as the legal representative, legal person title, registration status, and registration authority. Financial attributes such as registered capital, paid-in capital, new operating income, and debt ratio are also included, along with operational characteristics such as the years of establishment, business term, and the enterprise's market competition level. Additionally, we account for the enterprise's legal compliance through features like the legal litigation record [8]. Together, these features offer a holistic view of the enterprise's business status, financial health, and adherence to legal regulations, enabling a deep and multi-faceted evaluation of its risk profile.

Equally vital is the definition of the target variable, which serves as the dependent variable for model training and prediction. In this context, the target variable is defined to capture two specific risk outcomes: whether an enterprise is likely to be cancelled (with 1 indicating cancellation and 0 indicating non-cancellation) and whether it is at risk of revocation (with 1 indicating revocation and 0 indicating non-revocation). These target variables form the foundation for subsequent supervised learning tasks, allowing the models to predict and differentiate between these outcomes effectively.

Before feeding the data into machine learning algorithms, a robust data preprocessing stage is required to ensure its quality and consistency. This begins with processing the raw data, including the conversion of date fields into a format that represents the enterprise's years of establishment. Additionally, categorical features are transformed into numerical representations using one-hot encoding, a method that ensures compatibility with machine learning models and avoids introducing bias from ordinal assumptions. This preprocessing not only standardizes the data but also enhances its operability and ensures that it serves as a reliable input for subsequent algorithms [6, 9].

Once the preprocessing is complete, the resultant feature vectors are stored in the Milvus database. Milvus, as a high-performance vector database, plays a pivotal role in the system by enabling efficient management of large-scale feature data. Its capabilities in fast retrieval and similarity computation make it an ideal choice for the storage and analysis of feature vectors. This database acts as a powerful backend, providing the scalability and speed required for real-time risk assessments and iterative

model optimization [1, 14]. By leveraging Milvus, the system ensures robust technical support for enterprise risk assessment, setting the stage for accurate and efficient predictions.

1.2. Developing a Risk Prediction System: Feature Extraction, Random Forest Model Training, and Implementation of Enterprise Risk Assessment

To begin the model training process, we first extract the feature data from the Milvus database, ensuring that the data is complete, clean, and aligned with the defined input and target variables. Once extracted, the data is divided into a training set and a test set. This division is critical for building reliable machine learning models, as the training set is used to teach the model how to identify patterns and relationships in the data, while the test set evaluates the model's ability to generalize to new, unseen data. The target variables—whether the enterprise is at risk of cancellation or revocation—guide the supervised learning process by providing clear outcomes for the model to predict.

We employ the random forest algorithm for model training, creating two distinct models: one focused on predicting the likelihood of enterprise cancellation and the other on predicting revocation risk . The random forest algorithm is a robust and versatile choice due to its high accuracy, resilience to overfitting, and excellent generalization capabilities. This algorithm is particularly well-suited for handling high-dimensional datasets with complex nonlinear relationships, making it ideal for our feature-rich dataset. The diversity of the input features, ranging from financial metrics to operational indicators, benefits significantly from the ensemble learning approach of random forests, which combines multiple decision trees to improve prediction reliability.

During the model training process, we incorporate cross-validation techniques to further enhance the reliability of the models. Cross-validation systematically partitions the training data into subsets, allowing the model to be trained and validated on different data segments. This approach helps to mitigate overfitting, ensures the models perform consistently across various data samples, and improves their ability to handle unseen data effectively. By validating the models during training, we increase their robustness and credibility, ensuring they are suitable for deployment in real-world scenarios [4, 5].

After completing the training process, the trained models are saved locally to facilitate future use. Storing the models ensures their reusability across different applications and eliminates the need to retrain them for every prediction task. This practice enhances the system's efficiency and enables seamless integration into various workflows. The locally stored models can be accessed and utilized as needed, making them versatile tools for risk assessment in multiple contexts.

The final component of the system is the implementation of a user-friendly prediction function [10, 11]. This function allows enterprise managers and decision-makers to input relevant enterprise data, such as financial indicators or operational details, into the system. The function processes this input data, generates the corresponding feature vectors, and feeds them into the trained models to obtain predictions. The system then returns the results, indicating the enterprise's risk of cancellation or revocation.

This prediction capability empowers enterprise managers to quickly and easily assess their organization's risk status. If the results suggest a high risk of cancellation or revocation, managers can take proactive measures to address potential issues, such as restructuring the organization, refining business strategies, or strengthening compliance efforts. By providing accurate and actionable insights, the system supports timely and informed decision-making, contributing to improved business stability and long-term success.

1.3. Enhancing Enterprise Stability: A Machine Learning and Vector Database-Based Risk Management System

This system provides a powerful tool for users to quickly and efficiently assess the risk status of an enterprise, offering actionable insights based on advanced machine learning algorithms and vector database technologies [12]. By analyzing a comprehensive set of features, the system generates precise predictions regarding whether an enterprise faces risks of cancellation or revocation. These predictions enable managers to take proactive measures to mitigate potential risks before they escalate into significant issues. For instance, if the system identifies a high risk of cancellation, managers can take steps to restructure the organization, streamline operations, or improve financial practices. Similarly, if there is a risk of revocation, they can enhance compliance efforts, address legal disputes, or rectify regulatory shortcomings. Such preemptive actions not only help to minimize potential financial losses but also reinforce the long-term stability and operational resilience of the enterprise.

Furthermore, the ability to manage risks effectively has broader implications for an enterprise's ecosystem. When managers respond promptly to risk insights, it not only stabilizes the internal operations of the enterprise but also fosters confidence among key stakeholders, including investors, partners, and customers. Investors are more likely to trust and support an enterprise that demonstrates a proactive approach to risk management, while business partners may feel reassured about the reliability of their collaborations. Over time, this confidence translates into stronger relationships, enhanced reputation, and better market positioning.

At the heart of this system lies the integration of the Milvus vector database and cutting-edge machine learning technology. The Milvus database allows for efficient storage and retrieval of high-dimensional feature vectors, enabling the system to process vast amounts of data with speed and precision. This capability ensures that risk assessments are not only accurate but also delivered in real time, making the system highly adaptable to dynamic business environments. Combined with machine learning algorithms, the system offers a flexible and scalable solution for enterprise risk management, capable of handling increasingly complex datasets and evolving business needs [13].

An added advantage of this approach is the system's ability to continuously learn and improve over time. As more data is collected and incorporated into the system, the underlying models can be retrained and fine-tuned to enhance their predictive accuracy. This self-improving capability ensures that the system remains relevant and effective, even as the business landscape and risk factors evolve. It provides enterprises with a sustainable tool that adapts to their growth and changing challenges.

This innovative approach to risk management opens up new opportunities for enterprises by equipping them with a forward-looking strategy. Instead of reacting to crises, enterprises can now anticipate potential challenges and take strategic actions to mitigate risks. By doing so, they not only safeguard their current operations but also create a solid foundation for future growth. In an increasingly competitive and uncertain market, such a system provides enterprises with a decisive edge, enabling them to thrive while others may struggle to adapt. Ultimately, this advanced risk management method paves the way for more resilient, confident, and successful enterprises.

2. RESEARCH CONTENT

The present invention provides a method and system for screening small and medium-sized enterprises based on training large models with vector databases.

Feature vector representation

Convert each field into a numerical feature for model training and similarity calculation;

2.1. Data Field Preparation

Legal representative (Li): represented by text embedding or one-hot encoding.

Legal person title (Ti): also represented by text embedding or one-hot encoding.

Registration status (Si): represented by one-hot encoding, such as normal, cancelled, etc.

Registration authority (Ai) represented by one-hot encoding.

Registered capital (C_{reg_i}): directly use numerical values.

Paid-in capital (C_{paid_i}): directly use numerical values.

Date of establishment (Ei): converted to the years of establishment of the enterprise

$$E_{age_i} = \text{current date} - \text{date of establishment}$$

Business term (Di): calculate the effective business term.

The following are the risk parameters for whether to cancel:

Operating income ($C_{revenue_i}$): represented by numerical values.

Debt ratio (C_{debt_i})is represented by numerical values.

Market competition degree ($C_{competition_i}$): Use numerical representation.

Legal litigation record ($L_{lawsuit_i}$): Use binary variable representation.

Target variable:

y_{cancel}: Whether to cancel (1: cancel, 0: not cancel)

$y_{revoked}$: Whether to revoke (1: revoke, 0: not revoke)

Feature vector construction

2.2. Construct Feature Vector Xi

$$X_i = [L_i, T_i, S_i, A_i, C_{reg_i}, C_{paid_i}, E_{age_i}, D_i]$$

2.3. Cosine Similarity Calculation

Use cosine similarity to evaluate the similarity of enterprises, as shown below:

$$Cosine\ Similarity(X_i, X_j) = \frac{X_i \cdot X_j}{||X_i||\ ||X_j||} = \frac{\sum_{k=1}^{n} X_{i,k} X_{j,k}}{\sqrt{\sum_{k=1}^{n} (X_{i,k})^2} \cdot \sqrt{\sum_{k=1}^{n} (X_{j,k})^2}}$$

$X_{i,k}$: The kth feature from enterprise i.

$X_{j,k}$: The kth feature from enterprise j.

2.4. Calculation of Euclidean Distance

Calculate Euclidean distance to identify similar enterprises

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (X_{i,k} - X_{j,k})^2}$$

$X_{i,k}$: The kth feature from enterprise i.

$X_{j,k}$: The kth feature from enterprise j.

2.5. Machine Learning Model Training

Use random forest or other models for training, and the output of the model predicts risk:

$$F(X_i) = \frac{1}{M} \sum_{m=1}^{M} h_m(X_i)$$

$F(X_i)$:Risk prediction for enterprise i.

M: Number of trees in the model.

$h_m(X_i)$: Prediction of enterprise i by the mth tree

2.6. Risk Prediction

Finally predict whether there is a risk of revocation:

$$Risk(i) = \begin{cases} 1 & if\ F(X_i) > \text{threshold} \\ 0 & \text{otherwise} \end{cases}$$

If the model output exceeds the set threshold, it is predicted that the enterprise is at risk of revocation.

3. IMPLEMENTATION PROCESS

Training data and model: We input 1,000 real enterprise data and use the random forest algorithm to train the prediction of enterprise cancellation risk based on these data. The training data includes registered capital, paid-in capital, years of enterprise establishment, debt ratio, and market competition.

Feature importance analysis: The feature importance scores extracted from the model are as follows:

- Registered capital: 0.251
- Paid-in capital: 0.250
- Years of establishment: 0.177
- Debt ratio: 0.255
- Market competition: 0.067

Prediction accuracy: The overall prediction accuracy of the model is 50.5%, indicating that the model's performance in deregistration prediction is relatively basic and may require more data or features to optimize.

Classification report:

The precision is 0.47 and 0.55

The recall is 0.54 and 0.47

The F1 score shows the balanced performance of the model for the classification task, with a score close to 0.50.

Confusion matrix: The distribution of the model's performance in predicting non-deregistration and deregistration of enterprises is shown in the figure, which helps to intuitively understand the prediction errors and correct classifications.

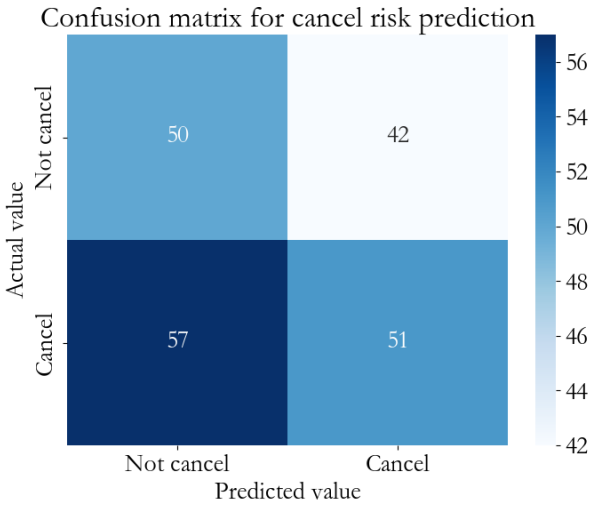

Figure 1. Predicted Value

REFERENCES

[1] Altman, E. I. "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy." The Journal of Finance, 1968.

[2] Huang, Z., et al. "Credit scoring with a data mining approach based on support vector machines." Expert Systems with Applications, 2007.

[3] Grover, A., & Leskovec, J. "Node2vec: Scalable Feature Learning for Networks." KDD, 2016.

[4] L. Breiman. "Random Forests for Classification and Regression." Machine Learning, 2001.

[5] Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, 2018.

[6] Bahdanau, D., Cho, K., & Bengio, Y. "Neural Machine Translation by Jointly Learning to Align and Translate." arXiv, 2014.

[7] Mikolov, T., et al. "Efficient Estimation of Word Representations in Vector Space." arXiv, 2013.

[8] Schlichtkrull, M., et al. "Modeling Relational Data with Graph Convolutional Networks." arXiv, 2017.

[9] Goodfellow, I., et al. "Generative Adversarial Nets (GAN)." NeurIPS, 2014.

[10] Wang, J., T.-T. Quoc, et al. "Efficient similarity search for large-scale datasets using FAISS." 2017.

[11] Zaharia, M., et al. "Apache Spark: A unified engine for big data processing." Communications of the ACM, 2016.

[12] Kreps, J. "Kafka: A Distributed Messaging System for Log Processing." LinkedIn, 2014.

[13] Dean, J., & Ghemawat, S. "MapReduce: Simplified Data Processing on Large Clusters." OSDI, 2004.

[14] Chang, F., et al. "Bigtable: A Distributed Storage System for Structured Data." OSDI, 2006.