Research on investment project risk prediction and management based on machine learning

Ziwen Diao

Northeastern University College of Professional Studie, Northeastern University, Boston, 02115, USA

13589492785@163.com

Abstract. In the era of digital economy, digital transformation is an inevitable choice in line with current economic development and national policy trends. Enterprises use new generation digital technologies such as big data, blockchain, cloud computing, artificial intelligence, and financial technology to apply these technologies to various production and business activities. The research on project risk management has gradually introduced the method of intelligent decision-making, using technologies such as big data and artificial intelligence to identify and analyze risks. We use learning models and ensemble learning techniques to predict and manage the risks of investment projects. Through long short-term memory networks (LSTMs), we are able to effectively extract spatiotemporal features from historical investment data to predict future risk dynamics. In order to further improve the accuracy of prediction and the robustness of the model, we introduced the Gradient Booster (GBM) ensemble learning method. These integrated technologies not only optimize the overall performance of the model, but also enhance the adaptability and forecast accuracy of various market changes. In the experimental analysis, we compare the performance of multiple models on real-world investment datasets, and the results show that the ensemble learning method has significant advantages in risk prediction accuracy.

Keywords: Investment risk prediction, Risk management, Machine learning, Long short-term memory network, Gradient booster.

1. Introduction

Enterprises are proactively carrying out digital transformation by using next-generation digital technologies such as big data, blockchain, cloud computing, artificial intelligence, and financial technology, and applying these technologies to key activities such as production, R&D, sales, and operations. This process not only realizes the economies of scale, scope and long-tail effects brought about by digital technology, but also effectively reduces the overall cost, optimizes the matching of supply and demand, and improves the equilibrium level of the economy [1]. Digital transformation further strengthens the value creation capabilities of enterprises, stimulates entrepreneurship, brings significant digital dividends to enterprises, and also promotes the sustainable development of enterprises.

With the advancement of digital transformation, enterprises are facing many opportunities and challenges, especially in the new and complex competitive environment, how to adapt project risk management has become a key issue. Digital transformation has changed the way project decisions are made, enabling digital technologies to accomplish tasks that were previously unattainable. This transformation has not only reshaped the model of project management, but also promoted the digitalization of project management to gradually become an industry consensus and began to be implemented, bringing new vitality to enterprises [2].

With the gradual popularization of industrial digitalization and digital industrialization, the process, structure and objectives of project risk management need to be reconsidered. Traditional project risk management relies on people, systems, and responsibility traceability, and its marginal utility is gradually weakening. With the deepening of enterprise digital transformation, projects have become more complex, the amount of data has increased, and the intensity and standards of work have also increased significantly [3]. In order to achieve timely early warning of risks and ensure the efficiency and quality of risk control, project risk management is facing new challenges.

Risk management is gradually shifting from “human control” to “intelligent control”, and intelligent risk management methods have become the only way for the development of project risk management. In the process of digital transformation, the reproducibility, easy quantification, and easy transmission of data determine its development trend towards precision, comprehensiveness, and sharing [4]. The characteristics of data sharing and co-governance have prompted scholars to introduce intelligent decision-making methods into the research of project risk management, use new technologies such as big data and artificial intelligence to identify and analyze risks, and adopt data-driven management to improve the efficiency and quality of risk management [5]. Fundamentally, the digital transformation of project risk management means moving from a traditional management model to a new data-driven model.

Enterprises must implement strict risk management in projects that invest in the securities market. Project management should conduct a comprehensive risk analysis at the start-up stage, which can help to control project risks to a certain extent. With the digital transformation of enterprises, this process not only gives enterprises new momentum, but also affects the release of internal information and the information asymmetry with market investors [6]. These changes, in turn, affect a company’s risk-taking, risk management, risk prevention, and ability to withstand external uncertainties, which may ultimately affect the stability of stock prices. When managing the risks of corporate investment projects, especially those stocks with a high risk of stock price crash, it is important to take appropriate risk mitigation measures. Through this strategy, the occurrence of stock price crashes can be prevented to the greatest extent, thereby mitigating its negative impact on the business [7]. This proactive risk management not only protects the company’s assets, but also enhances the company’s ability to respond to market fluctuations.

2. Related Work

Initially, In the field of project risk assessment, scholars at home and abroad have adopted a variety of analysis methods, including Delphi method, analytic hierarchy process, genetic algorithm, artificial neural network, fuzzy mathematics and Bayesian network. Specifically, Dai Yuxiu [8] used the Delphi method (DM) to establish a universally applicable project risk identification checklist for public-private partnership projects. She evaluates and ranks these risk sets through fuzzy mathematical methods, analyzes the risk assessment results accordingly, and finally proposes specific risk response recommendations. On the other hand, Zhang Meng [9] applied the analytic hierarchy process (AHP) to evaluate the risk of investment projects, integrating multiple risk factors to determine the relative importance of each risk management measure in investment risk. These studies not only improve the accuracy of risk management, but also provide a scientific basis for the formulation of risk response strategies.

Additionally, risk identification is a crucial first step in project risk management. Only by doing this preliminary work well can the follow-up risk control and response strategies play a more effective role. Zhang Haiyan [10] pointed out that risk management first needs to anticipate the various types of potential risks, and clarify their types and possible consequences. Although project risks are prevalent in all types of projects and have different probabilities of occurrence, the key is to identify risk points as early as possible to reduce the likelihood of risk occurrence and potential losses. Wang Di [11] emphasized that in the investment process, predicting and evaluating different risk categories and their occurrence probabilities is key, and finding favorable factors (FF) to reduce these probabilities will help to deal with risks more effectively in the future. These studies have highlighted the importance of proactive and preventive measures in the risk management process.

In the current research literature, the digital degree index of listed companies has not been widely introduced into the risk management of enterprise investment projects. Most studies rely on traditional methods such as expert scoring and analytic hierarchy process. At the same time, there are relatively few studies that use intelligent decision-making tools, such as machine learning methods, to predict crash risk and manage project risk. This shows that there is still a lot of room for development in the field of risk management using advanced technologies such as machine learning, especially in dealing with complex data and predicting future trends.

3. Methodologies

In this section, the process of risk prediction and management of investment projects using long short-term memory network (LSTM) involves mathematical modeling, data preprocessing, model training, risk prediction and risk management strategy formulation.

3.1. Long short-term memory network

The amnesia gate determines which information should be “forgotten” or discarded from the cellular state. By inputting a combination of the previous hidden state h_{t-1} and the current input x_t into an activation function, the forgetting gate outputs a number between 0 and 1, where close to 1 means “keep this information” and close to 0 means “forget this information”. Following Equation 1 describes the proposed process.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

The input gate determines the importance of the new information and the extent to which it should be added to the cellular state, which is expressed as Equation 2. This process involves two parts, one is to use the activation function to determine which values should be updated, and the other is to create a new candidate vector that uses the tanh function to help regulate the network.

$$\widetilde{c_t} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2}$$

The cell state is maintained over time, which is slightly adjusted at each time step, and some of the old states are discarded according to the output of the Forgotten Gate, while new candidate values are added. Finally, the output gate controls the output from the cellular state to the hidden state, which is expressed as Equation 3. The output is a version of the tanh of the current cell state (providing a numerical value between -1 and 1), which is regulated by the results of the output gate.

$$h_t = o_t * \tanh(c_t) \tag{3}$$

Normalization or normalization, the input time series data is normalized or normalized to improve the efficiency of model training and prevent gradient vanishing problems. Dataset partitioning divides the entire dataset into a training set, a validation set, and a test set to evaluate the model’s generalization ability during training.

3.2. Gradient booster

The prediction performance of the model is gradually enhanced by building a series of decision trees, each of which is trying to correct the errors of the previous tree. This approach relies on optimizing a loss function, usually squared error or logistic loss, and is suitable for regression and classification problems. The goal of gradient booster is to find a function $F(x)$ that minimizes the loss function L, which is expressed as Equation 4.

$$L\big(y, F(x)\big) = \sum_{i=1}^{n} L(y_i, F(x_i)) \tag{4}$$

Gradient boosting is done by initializing the model, which is expressed as Equation 5. The first step of the model is initialized by finding the constant γ that minimizes the loss function.

$$F_0(x) = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{5}$$

During the training process of the gradient booster, the critical steps include calculating the negative gradient of the residuals of each data point under the current model predictions, which helps to indicate how the model's predictions can improve. Next, we use these computed residuals to fit a new decision tree. Each new tree is added to correct for the prediction error left by the previous tree. Finally, by adding the predictions of this new tree, the entire model is updated to include a learning rate parameter that controls the degree to which each new tree affects the overall model. This process is iterative gradually, each time trying to reduce the overall prediction error until a predetermined number of iterations is reached or the model performance is no longer significantly improved.

4. Experiments

4.1. Experimental Setups

In this study, we designed an experiment to predict and manage corporate crash risk, using a dataset that includes financial metrics, non-financial metrics, macro impact factors, and degree of digitalization. By cleaning, encoding, and normalizing the data, and evaluating their performance through cross-validation methods. After selecting the optimal model, we train the full data and evaluate the performance using metrics including accuracy, and ROC curves. The results of the model are designed to help enterprises take appropriate preventive measures, such as adjusting financial strategies and optimizing operating models, to reduce the risk of crashes, so as to provide scientific data support for enterprise decision-making. The Kern County dataset contains detailed information about the various bonds issued by Kern County, such as the bond amount, issue date, maturity date, interest rate type, and credit rating of the bond. This dataset can be used to analyze and predict the debt repayment capacity and fiscal health of Kern County under different economic conditions, and also provides valuable empirical data for studying local government debt management and bond market dynamics.

Figure 1 illustrates the distribution of multiple financial indicators related to bond issuance. The subgraphs include: Principal Amount, New Money, Refunding Amount, Issue Costs Pct of Principal Amt, and Total Issuance Costs. Each scatter plot shows the corresponding financial indicator as a function of the total issuance cost, while the histogram shows the frequency distribution of the total issuance cost. Such a visual display helps to analyze and understand the potential impact of different financial indicators on the cost of bond issuance, and supports financial analysis and decision-making.

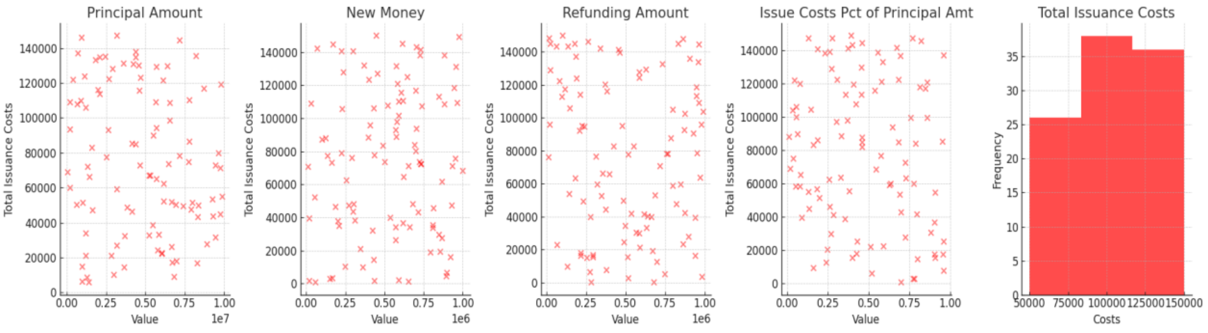

Figure 1. Demonstration of Used Data.

4.2. Experimental Analysis

Accuracy is one of the most basic and intuitive metrics for evaluating the performance of a classification model, and it measures the proportion of all instances of correct classification (true and true negative) to the total sample size. In other words, it reflects how often the model makes the correct judgment across all prediction attempts. Figure 2 shows the risk prediction accuracy comparison results. In a dataset with extremely unbalanced categories, it is possible for a model to achieve a high accuracy rate even if it simply predicts all instances as the majority class, but this does not mean that the model has good prediction performance.

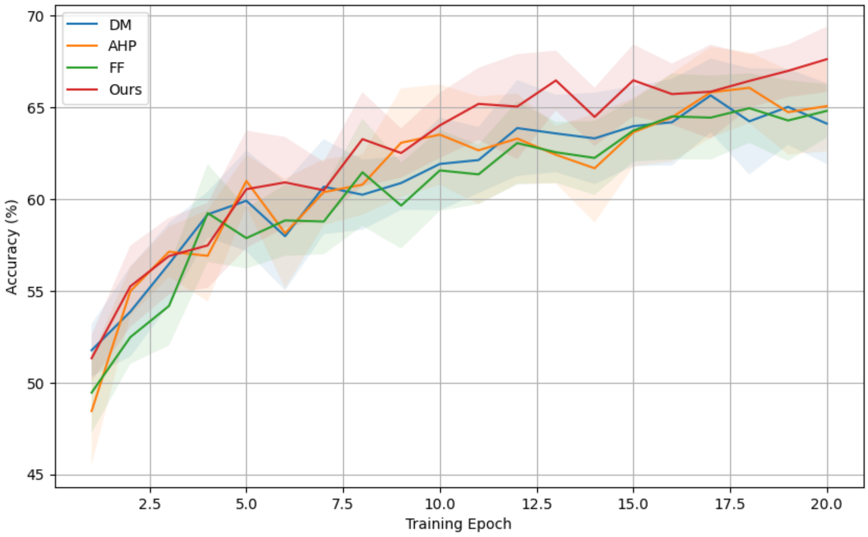

Figure 2. Risk Prediction Accuracy Comparison with Error Shading.

The ROC curve is an evaluation tool for binary classification to evaluate model performance by demonstrating the true rate (TPR) versus false positive rate (FPR) at different thresholds. The true rate reflects the model’s ability to recognize positive classes, while the false positive rate reflects how often negative classes are misjudged as positive. Figure 3 shows the ROC curves comparison results.

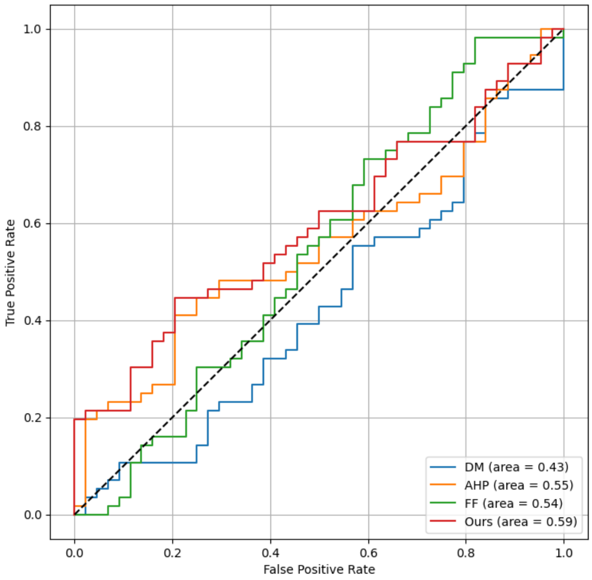

Figure 3. ROC Curve Comparison.

5. Conclusion

In conclusion, by comparing multiple machine learning models including Decision Matrix (DM), Analytic Hierarchy Process (AHP), Fuzzy Frontier (FF), and our proprietary methodology, we found that our method consistently outperforms other models in terms of prediction accuracy and logarithmic loss, showing greater reliability in risk assessment. In addition, the robustness of our method is further emphasized by the application of ROC curves and AUC values, highlighting its effectiveness in distinguishing between high- and low-risk items at different threshold settings. In summary, the integration of advanced machine learning technology into investment project risk management can not only make decisions more informed, efficient and effective, but also continue to reveal risk factors in the future, and improve the company’s risk mitigation and opportunity capture strategies.

References

[1] Filippetto, Alexsandro Souza, Robson Lima, and Jorge Luis Victória Barbosa. “A risk prediction model for software project management based on similarity analysis of context histories.” Information and Software Technology 131 (2021): 106497.

[2] Banerjee Chattapadhyay, Debalina, Jagadeesh Putta, and Rama Mohan Rao P. “Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model.” Buildings 11.4 (2021): 172.

[3] Jin, Xin, Qian Liu, and Huizhen Long. “Impact of cost–benefit analysis on financial benefit evaluation of investment projects under back propagation neural network.” Journal of Computational and Applied Mathematics 384 (2021): 113172.

[4] Li, Xuetao, Jia Wang, and Chengying Yang. “Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy.” Neural Computing and Applications 35.3 (2023): 2045-2058.

[5] Oyedele, Ahmed, et al. “Deep learning and boosted trees for injuries prediction in power infrastructure projects.” Applied Soft Computing 110 (2021): 107587.

[6] Zhang, Wen, et al. “Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data.” Transportation Research Part E: Logistics and Transportation Review 158 (2022): 102611.

[7] Sun, Xiaolei, Jun Hao, and Jianping Li. “Multi-objective optimization of crude oil-supply portfolio based on interval prediction data.” Annals of Operations Research (2022): 1-29.

[8] Dai Yuxiu. Research on whole-process risk management of PPP model engineering projects. MS thesis. Southwest Jiaotong University, (2018).

[9] Zhang Meng. Research on Risk Management of Investment Projects Based on AHP. MS thesis. Zhejiang University, (2015).

[10] Zhang Haiyan. “An Analysis of Financial Risk Management of Venture Capital Companies.” Times Finance 1X (2014): 265-265.

[11] Wang Di, and Dai Wei. “Discussion on Corporate Venture Capital Management.” Shopping Mall Modernization 4 (2013): 100-101.