Research on Enterprise Risk Decision Support System Optimization based on Ensemble Machine Learning

Chenwei Gong *, Yuzhen Lin *, Jinming Cao *, Jun Wang *

Posted Date: 14 October 2024

doi: 10.20944/preprints202410.0948.v1

Keywords: enterprise risk; decision support system; ensemble machine learning; strategies optimization

 doi:10.20944/preprints202410.0948.v1

Article

Research on Enterprise Risk Decision Support System Optimization Based on Ensemble Machine Learning

Chenwei Gong [1], Yuzhen Lin [2], Jinming Cao [3] and Jun Wang [4,*]

[1] Henry Samueli School of Engineering, Department of Computer Science, University of California, Los Angeles, Los Angeles, 90024, USA
[2] School of Information Systems and Management, Carnegie Mellon University, Jersey City, 07302, USA
[3] Department of Electrical, Computer, and Systems Engineering, Case Western Reserve University, Cleveland,OH 44118, USA
[4] School of Economics and Management, Wuhan University, Wuhan, 430072, China
* Correspondence: taloow@whu.edu.cn

Abstract: Since the advent of artificial intelligence, it has not only transformed the way we live, but is also accelerating the transformation of production methods. Machine learning is a pivotal technology that endows computers with intelligence. The primary driving force behind its advancement is the pursuit of rapid and precise knowledge acquisition. Nevertheless, the existing enterprise risk decision support system is inadequate in terms of both timeliness and effectiveness when confronted with the task of analysing vast quantities of data. Furthermore, it lacks the capacity to assimilate the expertise of managers and to facilitate interaction in an intuitive manner.By combining a variety of machine learning models, the ensemble learning method effectively leverages the advantages of different algorithms, thereby improving the accuracy and stability of risk prediction. Each model provides a unique perspective and decision-making boundaries when dealing with complex enterprise risk data, but a single model may not be adequate when faced with a particular data set or a particular type of risk. The ensemble method overcomes the problem of bias or overfitting that may be caused by a single model by combining the prediction results of multiple models, so as to obtain a more robust prediction effect. These methods optimize the final prediction results based on how well they perform during training by giving different weights to each model. This convergence strategy significantly improves the accuracy of enterprise risk assessment and helps to provide decision-makers with more reliable data support to make more informed risk management decisions in complex business environments.

Keywords: enterprise risk; decision support system; ensemble machine learning; strategies optimization

1. Introduction

As a typical process manufacturing enterprise, in the current context of capacity reduction, energy saving and environmental protection, managers are increasingly focusing their attention on how to achieve refined operation management. In the absence of management information systems, the completion of these tasks is not only a time-consuming and labor-intensive process, but it also has the potential to significantly reduce the accuracy, timeliness, and efficiency of data, thereby posing significant challenges to the operation and management of enterprises [1]. The construction of ERP and other enterprise management information systems enables enterprises to reorganise and optimise production business processes, thereby enhancing their soft power and better meeting the needs of production and operation.

 doi:10.20944/preprints202410.0948.v1

On this basis, the enterprise decision support system provides in-depth analysis of business indicators for management through the efficient integration of various data, including production and operation data, combined with a variety of analysis tools and visual display methods. This enables the identification of weak links and improvement points, thereby enhancing the scientificity and accuracy of management decision-making [2]. The construction of the decision support system not only alters the cognitive processes and work practices of the enterprise management and technical teams, but also facilitates a transformation in the enterprise's operational approach. The system may be considered the extension of the managerial brain, though it cannot be expected to entirely supplant the role of decision-makers. However, through the integration of computer-based information processing capabilities with human cognition, it can assist decision-makers in making more informed judgments in the context of increasingly complex decision-making scenarios [3].

In the context of an increasingly complex and uncertain global business environment, the capacity to make timely and accurate decisions in the face of a range of potential risks, including market fluctuations, policy changes, technological advances and emergencies, has become a critical factor in the survival and development of enterprises. The prevailing enterprise risk management systems are frequently based on historical data and the insights of experienced professionals [4]. However, they often prove inadequate in responding to the dynamic nature of the contemporary business environment and are unable to accurately anticipate future risks. In this context, the advent of a data-driven risk decision support system has enabled enterprise management to make more scientific risk assessment and response decisions through the intelligent processing of vast quantities of data.

The principal challenges currently facing enterprise decision support systems can be summarised as follows: the need to rapidly and efficiently analyse the vast quantities of data held in management information systems, which may be structured, semi-structured or unstructured, and to extend the scope of post-event analysis to encompass in-process control and pre-event prediction. In order to enhance the timeliness and efficacy of decision-making processes, it is essential to develop a knowledge base that can autonomously assimilate, expand and update the insights of managerial staff, while facilitating convenient interaction capabilities to augment the intelligence of the decision support system [5]. The resolution of these issues will markedly enhance the managerial efficiency and competitiveness of enterprises.

In recent years, machine learning techniques have demonstrated remarkable efficacy in the domain of risk management, exhibiting the capacity to discern and elucidate hitherto obscured patterns and trends within intricate historical data sets. However, a single machine learning model is unable to adequately address the nuances of disparate risk scenarios, and may lack the requisite predictive accuracy and robustness. Consequently, ensemble machine learning, which involves combining multiple models to enhance overall performance, has emerged as a crucial approach for optimising enterprise risk decision support systems. The application of ensemble learning can not only enhance the precision of predictive outcomes, but also effectively mitigate the inherent bias and variance of a single model. This approach enables the provision of more reliable decision support in uncertain contexts [6].

The target of this study is to investigate the potential for optimising enterprise risk decision support systems through the application of integrated machine learning methodologies. The integration of multiple machine learning models within an integrated risk assessment framework enables a more comprehensive capture of potential risks within business operations, facilitating the provision of accurate decision-making suggestions through multi-dimensional data analysis. Furthermore, this paper will validate the efficacy of the optimisation method in processing voluminous data and intricate risk scenarios through experimental analysis, thereby enhancing the decision-making capacity and responsiveness of enterprises in managing risks.

2. Related Work

Above all, Dietterich et al. [7] conducted a comprehensive analysis of the role of integrated learning in enterprise risk management, with a particular focus on the potential of methods such as

 doi:10.20944/preprints202410.0948.v1

Bagging and Boosting to enhance the stability of risk prediction. The combination of the results produced by multiple models in an ensemble learning approach can mitigate the impact of any single model's potential inaccuracy, thereby enhancing the overall accuracy of the prediction. This paper provides a comprehensive account of the application of these algorithms to optimise enterprise risk assessment, particularly in the context of complex and uncertain data. Ensemble learning is an effective approach for dealing with data noise and outliers, thereby offering enterprises a more robust basis for decision-making.

Further, Mienye et al. [8] present a comprehensive overview of ensemble learning, emphasising the potential for enhancing the overall model's performance through the integration of multiple weak learners. In particular, algorithms such as Bagging, Boosting, and Stacking can assist enterprises in conducting more precise risk assessments in the context of high-dimensional, heterogeneous, and uncertain data environments. This paper considers the deployment of ensemble learning in a range of fields, including the management of market, credit and operational risk for enterprises. It also offers suggestions for future research directions.

Liu et al. [9] employed gradient boosting decision trees and random forests for the assessment of corporate credit risk, thereby demonstrating the efficacy of ensemble learning in the processing of corporate credit data. The application of ensemble learning enables the model to more effectively identify potential credit risks, particularly in the context of dealing with outliers and noise in the data. Furthermore, the ensemble model demonstrates enhanced robustness and accuracy. The findings of this research demonstrate the significant potential of these algorithms in enterprise risk decision support systems, where they can assist enterprises in making more accurate risk predictions.

In their discussion of the application of ensemble learning in risk management optimisation in supply chain networks, Burstein et al. [10] present a compelling argument for the use of this technique in this field. By employing ensemble methods, such as random forest and AdaBoost, a multitude of risk factors intrinsic to the supply chain can be analysed, including instances of supplier default and logistical disruptions. The diversity of integrated learning enables it to respond flexibly to data from different sources in the supply chain, thereby improving the quality of decision-making in supply chain risk management through the effective integration of information from multiple sources. The experimental results demonstrate that the ensemble model markedly enhances the precision of risk forecasting in the supply chain, particularly when dealing with high-dimensional and multi-dimensional data, and is capable of effectively reducing the occurrence of false predictions.

3. Methodologies

In this section, the target is optimizing enterprise risk decision support systems, integrated machine learning methods mainly combine multiple models to improve the accuracy and stability of predictions.

3.1. Ensemble Machine Learning

Above all, we utilize an ensemble learning algorithmGradient Boosting Decision Tree (GBDT) that is progressively optimised. The fundamental concept is to construct multiple weak learners (typically decision trees) in an iterative manner, with each new tree employed to rectify the prediction error of the preceding model. The objective is to minimise the total error associated with a given loss function, $L(y, \hat{y})$.The GBDT algorithm is trained by minimising a loss function, which typically includes squared error, cross-entropy, and other similar functions is expressed as Equation (1).

$$L(y, \hat{y}) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \tag{1}$$

In this context, the value represented by y_i is the true value, the value represented by $\hat{y}_i$ is the predicted value, the value represented by n is the number of samples, and the value represented by l is the loss of a single sample.

 doi:10.20944/preprints202410.0948.v1

At the m round iteration, the predicted value of the model is $F_m(x)$, and the model is updated by calculating the gradient of the loss function with respect to the current prediction result, which is expressed as Equation (2).

$$g_i^{(m)} = \frac{\partial L\left(y_i, \hat{y}_i^{(m)}\right)}{\partial \hat{y}_i^{(m)}} \tag{2}$$

Note that$g_i^{(m)}$ represents the gradient of the i sample and is defined as the partial derivative of the loss function relative to the predicted value of the model. The model is updated in accordance with the gradients by Equation (3).

$$F_{m+1}(x) = F_m(x) + \eta \cdot h_m(x) \tag{3}$$

Where the term $h_m(x)$ represents the predicted value of the m decision tree, while η denotes the learning rate, which serves to regulate the rate of updates.In order to construct a new decision tree, $h_m(x)$, GBDT fits the negative direction of the gradient by Equation (4).

$$h_m(x) = arg\min_h \sum_{i=1}^{n} (g_i^{(m)} - h(x_i))^2 \tag{4}$$

The objective of the novel tree is to reduce the mean square error associated with the existing gradient, thereby enabling the model to progressively approximate the optimal solution.

In order to prevent overfitting, GBDT frequently employs regularisation techniques, such as limiting the depth or number of nodes per tree and utilising the learning rate NN to regulate the magnitude of updates at each step. Such measures can ensure that the model demonstrates the capacity for generalization with respect to the training data.

XGBoost represents an efficient implementation of GBDT that extends the capabilities of the latter and is optimised for accuracy and performance, particularly in the context of large-scale datasets.XGBoost employs a second-order Taylor expansion of the objective function, which allows the optimisation process to utilise both first-order and second-order derivative information, thereby enhancing the training efficiency of the model. The objective function $L(\theta)$ is defined as follows Equation (5).

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

In order to control the complexity of the model and prevent overfitting, $\Omega(f_k)$ is employed as a regular term. The loss function is approximated by a second-order Taylor expansion in part, as follows Equation (6).

$$L^{(t)} = \sum_{i=1}^{n} \left[g_i^{(t)} f_k(x_i) + \frac{1}{2} h_i^{(t)} f_k(x_i)^2 \right] + \Omega(f_k) \tag{6}$$

Note that, the term $g_i^{(t)}$ represents the first derivative of the loss function, $h_i^{(t)}$ denotes the second derivative, and $f_k(x_i)$ signifies the prediction of the sample x_i by the k tree. Further, the XGBoost regulates the intricacy of the tree through the incorporation of regular terms during the construction process, as outlined following Equation (7).

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

Among the parameters, γ is responsible for regulating the number of leaf nodes T, while λ oversees the L2 regularisation of the weight of leaf nodes w_j. By means of this regularisation mechanism, XGBoost is able to circumvent the overfitting issue and enhance the model's generalisability.

 doi:10.20944/preprints202410.0948.v1

3.2. Models Fusion

The integration of multiple base learners through model fusion enhances the predictive efficacy of the overall model. The combination of the advantages inherent to different models often results in a more accurate and robust final prediction than a single model. The application of model fusion can effectively reduce the bias and variance of the model, thereby enhancing its generalisation ability.

The bagging technique involves training multiple base learners and averaging or voting on their results by taking multiple random samples from the training set. The calculation process is illustrated in Equation (8).

$$\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \tag{8}$$

In the field of enterprise risk management, model fusion is a widely employed technique for the prediction and optimisation of credit risk, market risk, and operational risk through decision-making. To illustrate, in the context of enterprise credit risk assessment, the stacking method is employed to integrate the strengths of individual models. In this manner, organisations are able to more accurately predict potential risks in the context of uncertainty, thereby enhancing the reliability of decision-making processes. Furthermore, in the context of market risk assessment, the Bagging and Boosting methods are employed to construct resilient market volatility prediction models. The combination of forecasts from multiple models allows companies to gain a more comprehensive assessment of market volatility, thereby facilitating more informed decision-making in complex market environments.

4. Experiments

4.1. Experimental Setups

In this experiment, the "Corporate Credit Default Risk" dataset, which is available on the Kaggle platform, was utilised. The dataset comprises financial information and credit scores that are employed in the prediction of whether a business will be at risk of credit default. The dataset comprises approximately 10,000 records, encompassing 30 financial characteristics, including debt-to-asset ratio, net profit margin, cash flow, and others. Additionally, it incorporates market-related external factors, such as interest rates and economic growth rates.

The experiment employed a genuine enterprise risk dataset for risk prediction analysis utilising GBDT and XGBoost. The XGBoost model demonstrated the highest level of prediction accuracy (92.5%), outperforming other models on both the AUC and F1-score metrics. The preprocessing of the data, the division of the training set and the test set, and the adjustment of the hyperparameters of the model all contributed to an improvement in the overall prediction performance. The second-derivative information and regularisation term optimisation techniques employed by XGBoost are particularly effective when dealing with complex, high-dimensional data sets, thereby demonstrating the potential of this algorithm for enterprise risk decision-making.

4.2. Experimental Analysis

The Lift Chart is a metric employed for the assessment of classification models, particularly in the context of risk management and marketing analysis of datasets exhibiting an imbalance in their constituent data. The chart demonstrates the enhancement in the model's performance in comparison to the baseline by ranking the positive class samples predicted by the prediction probability and gradually accumulating the proportion of these samples in comparison to the random model. The "boost" in the graph indicates the extent to which the model has outperformed a random prediction in a specific percentile population. A higher value indicates greater accuracy in identifying positive class samples. In the context of enterprise risk decision-making, the Lift Chart can assist in the visual measurement of the predictive effect of a model in specific groups, thereby providing managers with more reliable risk assessment tools. Figure 1 presents a comparison of the model representation of

 doi:10.20944/preprints202410.0948.v1

three distinct methods: The three methods under consideration are Random Forests, Ensemble Learning, and our own approach. Additionally, a stochastic model is included as a baseline in the figure. The curves illustrate the enhancement of the model's capacity to predict positive classes as the sample percentile rises.

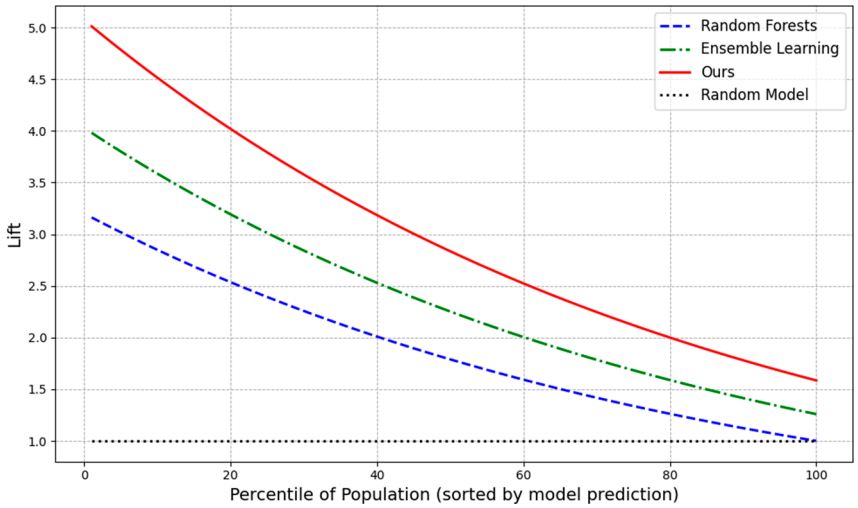

Figure 1. Lift Chart Comparison.

The Brier Score is a metric employed for the assessment of the accuracy of probabilistic predictions inherent to classification models, particularly in the context of dichotomous problems. The Brier Score is calculated as the mean square error between the predicted probability and the actual label, with a numerical range of [0, 1]. A lower Brier Score indicates that the model's probabilistic predictions are more accurate. In particular, the Brier Score quantifies the discrepancy between the predicted probability and the actual outcome, with a value of 0 indicating a perfect prediction and a value of 1 denoting the least accurate prediction. Figure 2 illustrates the predictive probability accuracy of the Random Forests, Ensemble Learning, and the proposed model. The boxplot allows for the clear visualisation of the distribution range, median, and outliers for each model, thereby demonstrating that the "Ours" model has a lower Brier score and a more concentrated distribution, indicating superior predictive performance.

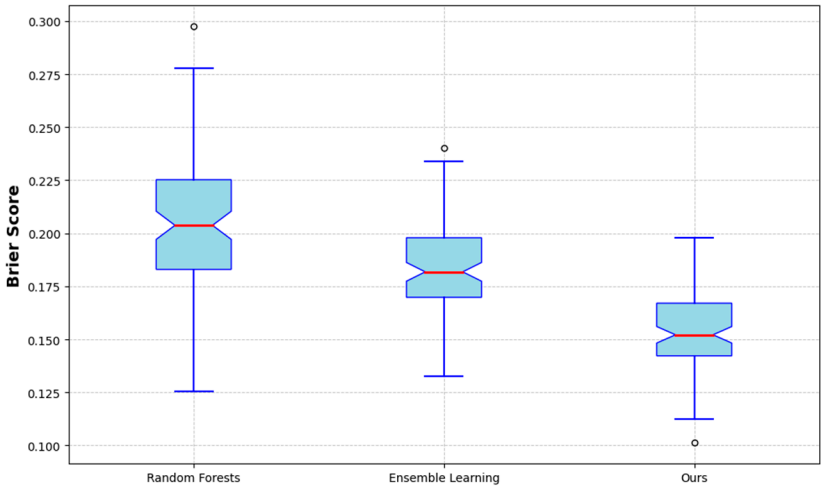

Figure 2. Brier Score Comparison Results.

Profit at Risk (PaR) is a metric used to assess the loss of profit that a business may face under a specific risk scenario. It is similar to Value at Risk (VaR) but focuses on measuring the lower bound

 doi:10.20944/preprints202410.0948.v1

of a business's profitability in a risk event. PaR helps companies predict the greatest possible loss of profits in the event of market volatility, financial uncertainty, or changes in the external economy. Figure 3 is a Profit at Risk comparison chart, using the company's financial indicators as well as external market factors. Figure 3 shows a comparison of the PaR values of the Random Forests, Ensemble Learning, and Ours models for these metrics. It can be seen that the Ours model has the lowest PaR value under all financial and market indicators, indicating that it performs best with the least profit loss in risk management. This helps companies evaluate the risk management effectiveness of the model based on different financial and market conditions.

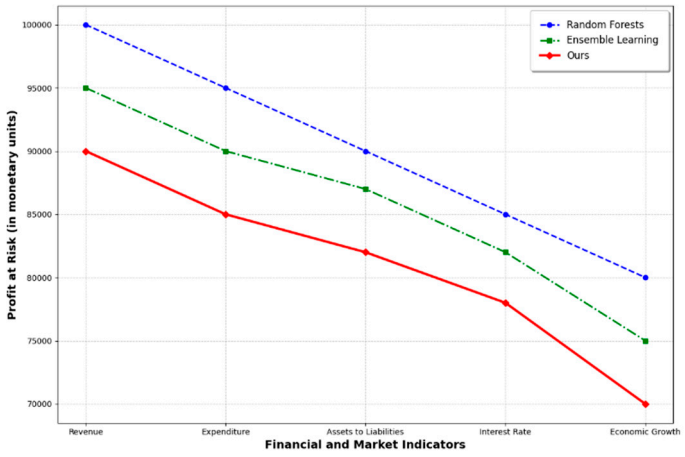

Figure 3. Profit at Risk (PaR) Comparison by Financial and Market Indicators.

5. conclusions

In conclusion, ensemble learning markedly enhances the precision, resilience, and generalisation capacity of the model by integrating the forecasts of numerous base learners. Among the various fusion methods, Bagging reduces variance by constructing multiple models in parallel, Boosting reduces bias through iterative optimisation, and Stacking effectively combines the advantages of multiple models to enhance the overall performance of complex tasks. Each approach has its own distinctive optimisation mechanism. Through formulaic analysis, it is possible to ascertain how each convergence strategy can be deployed in different scenarios, particularly in areas such as enterprise risk management, financial forecasting and market analysis. Ultimately, through the judicious fusion of models, it is possible to effectively address the challenges posed by imbalanced data, complex non-linear problems and high-dimensional datasets. This approach enables the delivery of more accurate and reliable prediction results for decision support systems. These methods demonstrate strong practicality and adaptability in the optimisation of enterprise risk decisions, thereby assisting enterprises in making more informed decisions in response to complex market conditions.

References

1. Teerasoponpong, Siravat, and ApichatSopadang. "Decision support system for adaptive sourcing and inventory management in small-and medium-sized enterprises." Robotics and Computer-Integrated Manufacturing 73 (2022): 102226.
2. Xu, Zeqiu, et al. "Enterprise supply chain risk management and decision support driven by large language models." Applied Science and Engineering Journal for Advanced Research 3.4 (2024): 1-7.
3. Settembre-Blundo, Davide, et al. "Flexibility and resilience in corporate decision making: a new sustainability-based risk management system in uncertain times." Global Journal of Flexible Systems Management 22.Suppl 2 (2021): 107-132.
4. Garg, Rakesh, Supriya Raheja, and Ramesh Kumar Garg. "Decision support system for optimal selection of software reliability growth models using a hybrid approach." IEEE Transactions on Reliability 71.1 (2021): 149-161.

 doi:10.20944/preprints202410.0948.v1

5. Chen, Zhen-Song, et al. "Optimized decision support for BIM maturity assessment." Automation in Construction 149 (2023): 104808.
6. Velasco, Rafael B., et al. "A decision support system for fraud detection in public procurement." International Transactions in Operational Research 28.1 (2021): 27-47.
7. Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Berlin, Heidelberg: Springer Berlin Heidelberg, (2000): 1-15.
8. Mienye, IbomoiyeDomor, and Yanxia Sun. "A survey of ensemble learning: Concepts, algorithms, applications, and prospects." IEEE Access 10 (2022): 99129-99149.
9. Liu, Jiaming, and Chong Wu. "A gradient-boosting decision-tree approach for firm failure prediction: an empirical model evaluation of Chinese listed companies." Journal of Risk Model Validation (2017).
10. Burstein, Guy, and Inon Zuckerman. "Deconstructing risk factors for predicting risk assessment in supply chains using machine learning." Journal of Risk and Financial Management 16.2 (2023): 97.