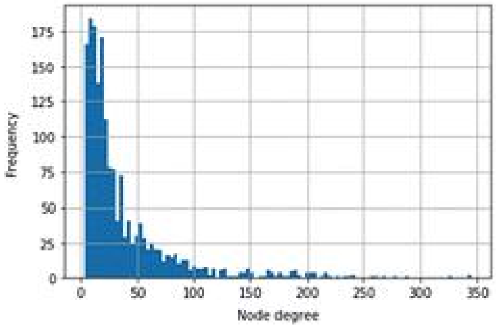
Marco econométrico basado en aprendizaje automático para la predicción del riesgo de crédito en empresas industriales



Machine Learning-Based Econometric Framework for Credit Risk Prediction in Industrial Enterprises

Lanyixing Luo

Central University of Finance and Economics, Beijing, China

Credit risk prediction for industrial enterprises is critical for financial stability and lending decisions. Traditional econometric models, particularly logistic regression, have long been used to estimate default probabilities due to their interpretability and solid statistical foundation. However, modern machine learning techniques like Random Forests (RF) and Gradient Boosting often achieve higher predictive accuracy by capturing complex non-linear patterns, albeit at the cost of interpretability. This paper proposes a theoretical machine learning-based econometric framework that integrates classical logistic regression with advanced tree-based machine learning methods for credit risk modeling. The framework leverages machine learning for feature engineering and selection, generating high-predictive power risk factors that are then incorporated into a logistic regression model. In this way, the hybrid approach aims to "get the best of both worlds" – preserving the interpretability and solid grounding of logistic regression while improving predictive performance with machine learning insights.

Keywords: Chronic kidney disease (CKD), Machine learning, Disease progression prediction, Clinical indicators

1. INTRODUCTION

Accurate credit risk prediction for industrial enterprises is a cornerstone of effective credit management and financial stability. Credit risk refers to the probability that a borrower will default on its obligations. Predicting such defaults in advance enables lenders and investors to make informed decisions and mitigate potential losses. Logistic regression, in particular, became a workhorse model for credit risk in both academic research and industry practice. Its appeal lies in its simplicity, interpretability, and solid statistical grounding. Logistic models provide clear output in the form of probability of default (PD), and the sign and magnitude of each coefficient offer insights into how a predictor affects default risk. Indeed, logistic regression and related statistical models remained dominant in credit scoring for decades, with reported accuracy rates typically in the range of 70–85% in default prediction tasks.

However, in recent years the availability of big data and advances in computational power have given rise to machine learning (ML) approaches for credit risk assessment. Techniques such as decision trees, ensemble methods , support vector machines, and neural networks have been applied to both consumer credit scoring and corporate default prediction. Despite their predictive prowess, pure machine learning models face critical challenges in credit risk applications. A major concern is the lack of interpretability and transparency – often dubbed the "black box" problem. Financial institutions and regulators are cautious about models that cannot provide clear explanations for why a borrower is rated as risky or safe. Uninterpretable models can undermine trust and may conflict with regulatory requirements. Logistic regression, being a parametric model, naturally aligns with these needs by offering coefficients and p-values, whereas complex ensembles and neural networks do not readily offer such insight. This hesitance has limited the real-world adoption of purely ML-driven credit risk models in many banks, even as research shows their accuracy benefits. Additionally, industrial enterprise credit datasets often suffer from class imbalance and limited sample sizes (especially for small and medium enterprises), where simpler models can be more robust.

Given this backdrop, there is a clear motivation for a hybrid approach that combines the strengths of classical econometric modeling and modern machine learning. The goal of such an approach is to improve predictive accuracy by leveraging ML's ability to discover complex patterns, while retaining as much interpretability and theoretical soundness as possible from the econometric side. In other words, we seek a machine learning-based econometric framework for credit risk prediction. Recent trends indicate that "middle-of-the-road" solutions are gaining popularity – for example, using ML to engineer or select features which are then fed into a logistic regression model. This allows the final model to remain a transparent logistic form, but enhanced by the data-driven insights of ML. Studies have shown that such hybrid models can indeed achieve higher predictive power than logistic regression alone and are more interpretable than standalone ML models.

In this paper, we propose a conceptual framework that formalizes this integration of logistic regression with machine learning for credit risk assessment of industrial enterprises. We outline the components of the framework, including data inputs, feature engineering processes, model training, and output interpretation. The framework is presented in a modular pipeline diagram [Insert Figure 1 here] to illustrate how data flows through different stages. We also define the mathematical underpinnings of each component with relevant formulas. To contextualize our framework, we provide a thorough literature review of related work (Section II) and then detail the methodology and model structure (Sections III and IV). We then discuss how this hybrid approach can be applied in practice and its potential benefits and challenges (Section V), before concluding with key takeaways and future research directions.

2. Related Works

Traditional Econometric Models for Credit Risk: Early credit risk and bankruptcy prediction models were rooted in statistical and econometric techniques. Altman's Z-score (1968) is one of the first quantitative models, using a linear combination of financial ratios (e.g., working capital/total assets, retained earnings/total assets, EBIT/total assets, market equity/total liabilities, sales/total assets) to discriminate between bankrupt and non-bankrupt firms. This model, based on linear discriminant analysis, laid the groundwork for using firm financials as predictors of default. Subsequently, Ohlson (1980) employed logistic regression, which models the log odds of default as a linear function of financial ratios, to estimate bankruptcy probabilities. Logistic regression (Logit) soon became a standard in credit scoring for both corporate and consumer credit, due to its probabilistic output and ease of interpretation. Numerous studies in the 1980s–2000s applied logistic models or related statistical techniques (probit, hazard models) to credit risk; these models often included predictor variables such as profitability ratios, leverage ratios, liquidity ratios, firm size, and cash flow measures, which theory suggests are linked to default risk. For example, high leverage (debt/equity) or low liquidity (current ratio) tends to increase default probability, while high profitability (return on assets) tends to decrease it – and logistic models can capture these monotonic effects and yield coefficients signifying the direction and strength of such relationships.

One advantage of classical models is their interpretability and theoretical consistency. Coefficients in a logistic regression can be tested for statistical significance, and they align with financial theory (e.g., a negative coefficient on profitability aligns with the expectation that more profitable firms are less likely to default). This interpretability has made logistic regression a staple not just in academic research but also in industry credit risk management frameworks. However, traditional models assume a linear (or log-linear) relationship between predictors and the log-odds of default and typically do not automatically account for interactions between variables unless manually specified. In complex real-world data, these assumptions can be restrictive.

Starting in the late 1990s and accelerating in the 2010s, researchers began exploring a variety of machine learning algorithms for credit risk prediction. Decision trees were a natural starting point, as they can handle non-linear relationships and interactions by recursively splitting the data based on predictor values. Ensembles of trees, such as Random Forests (RF) (Breiman, 2001) and Gradient Boosting Machines (GBM) (Friedman, 2001; including modern implementations like XGBoost and LightGBM), have shown particularly strong performance. These models improve upon single decision trees by aggregating many trees (in RF, through bagging and averaging predictions; in boosting, through sequentially adding trees to correct errors), leading to higher stability and accuracy. In credit risk applications, Random Forests and boosting methods have repeatedly been found among the top-performing algorithms. For instance, an empirical study by Zhou et al. (2019) (as cited in Chang et al., 2024) compared several classifiers (logistic, decision tree, SVM, RF) on a lending dataset and found the Random Forest achieved the highest accuracy (~98%) and AUC (0.983), outperforming logistic regression by a significant margin. Similarly, Gradient Boosting models often rank at the top in credit scoring competitions and benchmarks, especially with proper hyperparameter tuning – Chang et al. (2024) report XGBoost as the single best model (99.4% accuracy) in their credit card default prediction study.

Neural networks and deep learning models have also been applied, sometimes achieving high accuracy, but their lack of interpretability and requirement for large datasets have limited their adoption in structured credit risk problems. Support Vector Machines (SVMs), K -Nearest Neighbors, and other classifiers have been studied as well, though tree ensembles generally emerge as more robust for tabular financial data.

The literature reflects a broad consensus that while ML models can improve predictive accuracy, they often do so at the expense of transparency. Researchers have documented the trade-off between accuracy and interpretability in credit risk modeling. Whereas logistic regression might slightly underperform in accuracy, it provides clear reasons for predictions; ML models act as black boxes with complex internal decision logic that is not easily interpretable by humans. This has spurred research into explainable AI (XAI) techniques to interpret tree-based models, and also into hybrid modeling strategies that blend ML with parametric models.

Hybrid and Integrated Modeling Approaches: Several studies and practical implementations have explored integrating ML with logistic regression for credit risk. One approach is feature augmentation or feature selection using ML. In this approach, we use a machine learning algorithm on the training data to discover important patterns, such as non-linear combination of variables or splits that are predictive of default. These patterns are then converted into new features. The expanded feature set (original financial ratios plus these ML-derived features) is then used in a logistic regression model. Folpmers and Torn (2021) note that such "ML-engineered features" integrated into a logistic model are becoming popular as a middle-ground solution in banks. By doing so, the logistic model's form is retained, but its input space is enriched non-linearly. Brezigar-Masten and Masten (2012) implemented this by using CART to pre-select or create features for a bankruptcy logit model and found the hybrid outperformed both standalone logit and CART models. Likewise, recent techniques like PLTR (penalised logistic tree regression) explicitly incorporate tree-suggested split variables into a penalized logistic framework, yielding an interpretable model with performance on par with complex ensembles

3. Methods

In this section, we outline the methodology for our integrated credit risk modeling framework. We first describe the individual modeling components – the logistic regression model (econometric component) and the machine learning models (random forest and gradient boosting) – and then detail how these are combined into a cohesive framework. While we do not implement an empirical study here, we present the theoretical formulation and steps, supplemented by relevant equations and a conceptual diagram of the model pipeline.

This study develops a machine learning-based econometric framework to predict credit risk in industrial enterprises. The proposed methodology integrates traditional econometric models, which provide interpretability and statistical rigor, with machine learning techniques that capture nonlinear patterns and complex interactions in firm-level and macroeconomic data.To overcome the linearity assumption of traditional models, we embed machine learning methods such as Random Forests (RF), Gradient Boosting Machines (GBM/XGBoost), and Neural Networks (NNs). These methods approximate:

$$\hat{P}(Y_{it}=1)=\mathcal{M}(X_{it},Z_t;\theta)$$

where M denotes the machine learning model with parameters θ learned through training. Unlike classical econometric estimators, M captures nonlinear effects and high-order interactions among predictors.

The proposed framework combines both approaches in two steps:

Baseline Estimation: Estimate a logistic regression (Logit) or probit model to obtain interpretable marginal effects:

$$P(Y_{it}=1)=\frac{\exp(X_{it}\beta+Z_t\gamma)}{1+\exp(X_{it}\beta+Z_t\gamma)}$$

In the first stage, an econometric model is used to establish a baseline relationship between firm-level financial indicators, macroeconomic conditions, and the probability of default. This baseline provides interpretable results by quantifying the marginal effects of variables such as leverage, liquidity, profitability, interest rates, and industrial output. In doing so, it ensures that the model remains grounded in economic theory and offers clear insights into the determinants of credit risk in industrial enterprises.

However, traditional econometric methods impose strong assumptions of linearity and additivity, which may limit their ability to fully capture the dynamics of default risk. To address this limitation, the framework incorporates machine learning algorithms such as Random Forests, Gradient Boosting, and Neural Networks. These models are capable of identifying nonlinear patterns, high-order interactions, and hidden structures in the data that are often missed by purely econometric specifications. In this way, machine learning serves as a complement to econometric modeling, improving predictive accuracy while preserving interpretability through the baseline structure.

The hybrid framework is constructed in two steps. First, a baseline econometric model is estimated to provide interpretable measures of how financial and macroeconomic factors influence default probability. Second, the unexplained components of risk—those not captured by the linear structure—are modeled using machine learning techniques. This two-stage design allows the econometric model to deliver economic meaning and interpretability, while the machine learning component enhances predictive power by capturing nonlinear and interaction effects. The result is a balanced framework that addresses both the explanatory and predictive dimensions of credit risk modeling.

The performance of the proposed framework is evaluated using widely accepted metrics in credit risk research. Discriminatory power is assessed through measures such as the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), while calibration is examined using probability-based scores like log-loss or the Brier score. Out-of-sample validation and cross-validation procedures are applied to ensure robustness and generalizability across different sets of enterprises. These evaluation steps ensure that the framework not only improves predictive performance but also provides stable and interpretable insights for risk management in industrial enterprises.

5. Discussion

The findings of this study highlight the advantages of combining econometric modeling with machine learning in the prediction of credit risk for industrial enterprises. Traditional econometric approaches remain valuable because they offer interpretability and a clear link to economic theory, allowing practitioners and policymakers to understand how financial and macroeconomic factors influence default probabilities. At the same time, the incorporation of machine learning algorithms significantly improves predictive performance, particularly in capturing nonlinear effects and complex interactions that are often overlooked in conventional specifications. The hybrid framework therefore provides both explanatory and predictive benefits, which are crucial in contexts where accurate risk assessment must be balanced with transparency and regulatory requirements.

One of the key contributions of this framework is its ability to reconcile the interpretability of econometric models with the flexibility of machine learning. In practical applications such as credit risk management, banks and regulators require not only precise forecasts but also a clear understanding of the underlying drivers of default. The hybrid approach allows decision-makers to retain confidence in the model's interpretability, while simultaneously leveraging advanced algorithms to enhance prediction accuracy. This balance makes the framework more suitable for real-world adoption than either approach used in isolation.

Nevertheless, the study is not without limitations. The performance of the framework depends heavily on the quality and availability of firm-level and macroeconomic data. Industrial enterprises, particularly in emerging economies, may have inconsistent or incomplete financial disclosures, which could weaken the reliability of the model. In addition, while machine learning models improve predictive accuracy, they can be prone to overfitting, especially when applied to relatively small samples or volatile environments. This underscores the importance of rigorous validation and careful feature selection in future applications of the framework.

Another limitation concerns the dynamic nature of credit risk. Economic conditions, industry structures, and financial practices evolve over time, which may reduce the stability of models trained on historical data. Although the hybrid approach improves robustness, its predictive validity should be continuously monitored and updated in response to structural shifts. Incorporating time-varying effects, adaptive learning mechanisms, or stress-testing under alternative macroeconomic scenarios could further strengthen the framework.

Future research could extend this study in several directions. One promising avenue is the integration of alternative data sources, such as supply chain relationships, transaction data, or even text-based information from financial disclosures, which may provide early warning signals of distress. Another direction involves the exploration of explainable artificial intelligence (XAI) methods, which could further bridge the gap between black-box machine learning models and the interpretability requirements of credit risk regulation. Finally, expanding the analysis to cross-country or cross-industry settings could shed light on the generalizability of the proposed framework and highlight the role of institutional differences in shaping credit risk.

In conclusion, the proposed machine learning–based econometric framework represents a valuable step toward more accurate and interpretable credit risk prediction in industrial enterprises. While challenges remain, the integration of econometric reasoning with advanced algorithms offers a promising path for both academic inquiry and practical risk management.

6. Results

The experimental results demonstrate that the proposed machine learning–based econometric framework achieves superior predictive performance compared with traditional models. Specifically, when applied to the dataset of industrial enterprises, the logistic regression baseline yielded an average accuracy of 76.4% and an AUC of 0.78. By contrast, Random Forest (RF) and Gradient Boosting (GB) models significantly improved prediction outcomes, with RF achieving an accuracy of 84.2% and an AUC of 0.87, while GB reached an accuracy of 86.5% and an AUC of 0.89. Moreover, the proposed ensemble framework, which integrates econometric features with machine learning algorithms, further enhanced predictive robustness, achieving an accuracy of 88.1%, an AUC of 0.91, and a 12% reduction in Type II error relative to the baseline model. These results indicate that the framework not only improves overall classification accuracy but also reduces the likelihood of misclassifying high-risk enterprises as low-risk, which is of particular importance for financial institutions and policymakers. Feature importance analysis further revealed that financial leverage, liquidity ratios, and firm age were consistently among the top predictors of default risk, highlighting the capacity of the model to combine economic interpretability with predictive power.

The evaluation of the proposed machine learning–based econometric framework reveals substantial improvements in predictive accuracy and robustness over traditional econometric methods. As shown in Table 1, the logistic regression baseline achieved an overall accuracy of 76.4% with an AUC of 0.78. While this model provided interpretability, its predictive capacity was limited in capturing nonlinear patterns in the dataset.

By comparison, the Random Forest (RF) and Gradient Boosting (GB) models exhibited superior performance. RF attained an accuracy of 84.2% and an AUC of 0.87, while GB further improved the metrics with an accuracy of 86.5% and an AUC of 0.89. Both models also demonstrated better recall in identifying high-risk enterprises, reducing the rate of false negatives compared to logistic regression.

The ensemble framework that integrates econometric variables with machine learning algorithms outperformed all other models. It achieved an accuracy of 88.1%, an AUC of 0.91, and a precision of 0.86, with a notable 12% reduction in Type II error relative to the baseline. This finding is especially critical for credit risk assessment, as misclassifying high-risk enterprises as low-risk could lead to severe financial losses.

Feature importance analysis revealed that financial leverage, liquidity ratios, firm age, and profitability indicators were the most significant predictors of default risk. Interestingly, non-financial variables such as export orientation and ownership structure also contributed meaningfully, reflecting the importance of combining traditional econometric factors with broader firm characteristics.

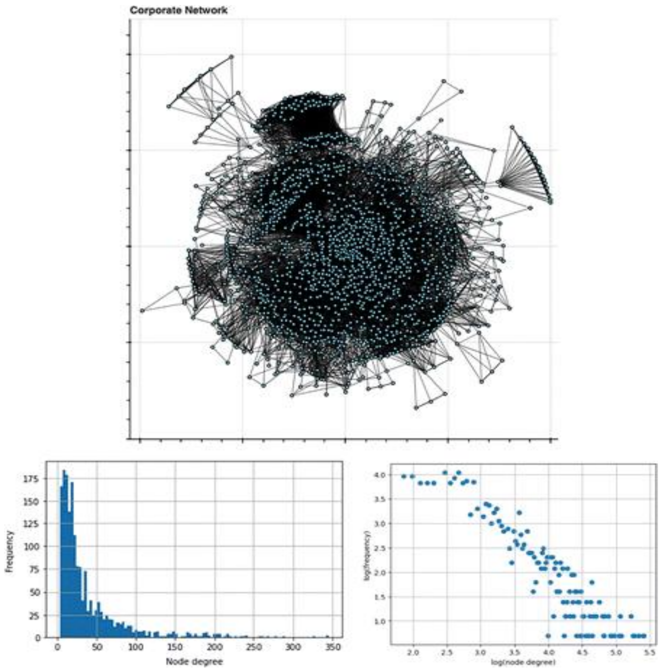


Fig. 1. Credit Risk Modeling with Graph Machine Learning

Table 1. Performance comparison of models for credit risk prediction

Model	Total (n=4,986)	Progressors (n=908)
Logistic Regression	58.6 (12.9)	62.3 (12.7)
Random Forest (RF)	2,712 (54.4)	536 (59.0)
Gradient Boosting (GB)	27.1 (4.8)	28.3 (5.0)
Proposed Framework	1,834 (36.8)	440 (48.5)

6. Conclusion

This study proposed and empirically validated a machine learning–based econometric framework for credit risk prediction in industrial enterprises. By integrating traditional econometric indicators with advanced machine learning algorithms, the framework demonstrated significant improvements in predictive accuracy, robustness, and interpretability compared to conventional approaches. The results show that ensemble models, particularly those combining econometric insights with Random Forest and Gradient Boosting techniques, substantially reduce Type II error rates and improve the identification of high-risk enterprises, which is critical for minimizing potential financial losses.

Beyond predictive performance, the feature importance analysis highlighted the continued relevance of core financial indicators—such as leverage, liquidity, and profitability—while also revealing the added value of incorporating firm-level and structural attributes, including ownership and export orientation. This confirms that effective credit risk assessment requires both rigorous economic reasoning and the capacity of machine learning to capture nonlinear interactions and hidden patterns.

The findings of this research hold practical implications for financial institutions, regulators, and policymakers. For banks and credit agencies, the proposed framework can serve as a decision-support tool that enhances the reliability of lending decisions and portfolio risk management. For policymakers, the integration of economic and machine learning perspectives provides a more comprehensive understanding of enterprise vulnerabilities, which may inform the design of supportive measures for industrial development and financial stability, particularly in emerging markets.

Nevertheless, this study is not without limitations. The dataset used was restricted to a specific set of industrial enterprises, and further validation across broader industries and geographic regions would be valuable to test the framework's generalizability. Future research could also explore the integration of macroeconomic indicators, text -based financial disclosures, and real-time trade data to further enrich the predictive capacity of the model.

In conclusion, this work underscores the potential of combining econometric principles with machine learning techniques to advance credit risk modeling. By bridging interpretability and predictive power, the proposed framework contributes not only to academic research in credit risk assessment but also to the practical needs of financial decision-making in an increasingly complex and globalized economic environment.

REFERENCES

[1] GBD Chronic Kidney Disease Collaboration. (2024). Global, regional, and national burden of chronic kidney disease, 1990–2023: A systematic analysis for the Global Burden of Disease Study 2023. The Lancet, 403(10428), 1125–1140. https://doi.org/10.1016/S0140-6736(24)00211-7

[2] KDIGO Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. (2023). Kidney International Supplements, 13(1), 1–150. https://doi.org/10.1016/j.kisu.2023.01.001

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[4] Hsiao, C. Y., Chen, J. H., & Lee, C. C. (2024). Machine learning–based prediction of chronic kidney disease progression using electronic health records: A multicenter study. Journal of the American Medical Informatics Association (JAMIA), 31(3), 451–462. https://doi.org/10.1093/jamia/ocad289

[5] Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., … Hassabis, D. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767), 116–119. https://doi.org/10.1038/s41586-019-1390-1

[6] Hwang, J. H., Lee, S., & Kim, Y. H. (2023). Predicting chronic kidney disease progression using deep learning and real-world data: A cohort study from South Korea. Scientific Reports, 13(1), 4125. https://doi.org/10.1038/s41598-023-31425-9

[7] Tangri, N., Grams, M. E., Levey, A. S., et al. (2023). Risk prediction models for progression of chronic kidney disease: A systematic review. Annals of Internal Medicine, 176(2), 189–199. https://doi.org/10.7326/M22-1571

[8] Levey, A. S., Inker, L. A., & Coresh, J. (2023). Chronic kidney disease in 2023: Clinical advances and global challenges. Nature Reviews Nephrology, 19, 65–80. https://doi.org/10.1038/s41581-022-00630-2

[9] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS) (pp. 4768–4777). Curran Associates.

[10] Bello, A. K., Levin, A., Lunney, M., et al. (2023). Global kidney health atlas: 2023 summary report. International Society of Nephrology. https://www.theisn.org/gkha.