

AI-Powered Credit Scoring vs. Traditional Scoring Models:**A Comparative Study on Loan Default Prediction**Author: Gaelle Youssef¹*Affiliation:*

1. *Business School, Holy Spirit University of Kaslik, P.O. Box 446, Jounieh, Lebanon*
2. *CIRAME Research Center, Holy Spirit University of Kaslik, P.O. Box 446, Jounieh, Lebanon*

Abstract:

This study compares the predictive performance of a traditional statistical credit scoring model against a modern artificial-intelligence-based model in forecasting consumer loan default. Because access to a real bank's confidential loan portfolio was not feasible for this project, a synthetic loan-applicant dataset (5,000 records) was engineered to mirror the statistical structure of real-world credit bureau data, including a realistic 22.8% default rate and ten borrower-level features such as income, debt-to-income ratio, credit utilization, and payment history. A logistic regression model, representing the rule-based scorecards traditionally used in retail banking, was benchmarked against a gradient boosting model, representing contemporary AI-driven scoring engines. Both models were trained on an identical 75/25 train-test split and evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The AI model outperformed the traditional model on every metric, most notably raising AUC-ROC from 0.814 to 0.936 and recall (the ability to correctly flag true defaulters) from 43.9% to 68.4%. These findings suggest that AI-based scoring can materially reduce a lender's exposure to undetected default risk, though questions of interpretability, fairness, and regulatory compliance remain important considerations for real-world adoption.

Keywords:

Artificial Intelligence, Credit Scoring, Loan Default Prediction, Machine Learning, Financial Risk Management, Gradient Boosting

Introduction:

Credit scoring is the process by which lenders estimate the probability that a borrower will default on a loan, and it sits at the centre of retail banking risk management. For decades, banks have relied on statistical scorecards, most commonly logistic regression models, built on a small set of standardized variables such as income, existing debt, and repayment history. These models are valued for being transparent and easy to justify to regulators and rejected applicants alike, but their simplicity is also a limitation: logistic regression assumes a linear relationship between risk factors and default probability, which may not capture the more complex, non-linear ways that borrower behaviour actually signals risk.

Over the past decade, advances in machine learning have introduced a new generation of scoring tools, including random forests, gradient boosting machines, and neural networks, that can model complex interactions between dozens of variables without requiring those relationships to be specified in advance. Prior literature suggests these methods can improve default prediction accuracy, but they also raise concerns around interpretability and regulatory acceptance, since a model's decision logic is harder to explain to a loan applicant or an auditor than a scorecard's coefficients.

This creates a practical gap for financial institutions: is the accuracy gain from AI-based scoring large enough to justify the added complexity and reduced transparency? This study addresses that gap directly through a controlled, side-by-side comparison.

Research Question: Does an AI-based (machine learning) credit scoring model predict loan default more accurately than a traditional statistical scoring model, and if so, by how much?

Objectives:

- Build and train a traditional logistic regression scorecard and an AI-based gradient boosting model on identical loan-applicant data.
- Compare the two models on accuracy, precision, recall, F1-score, and AUC-ROC.
- Identify which borrower characteristics most influence each model's predictions.
- Discuss the practical trade-offs (accuracy vs. interpretability) relevant to real-world adoption.

Literature Review:

The academic foundation for credit scoring dates back to statistical scorecard methods, which Thomas (2000) surveyed as the dominant approach in consumer lending for most of the twentieth century, valued for their transparency and ease of regulatory validation. As computing power and data availability increased, researchers began testing whether machine learning classifiers could outperform these traditional models. Breiman (2001) introduced the random forest algorithm, an ensemble method that reduces overfitting by averaging predictions across many decision trees, while Friedman (2001) developed gradient boosting, which builds trees sequentially so that each new tree corrects the errors of the ones before it; both techniques became foundational to modern AI-based credit scoring.

Khandani, Kim, and Lo (2010) were among the first to apply machine-learning classifiers directly to consumer credit-risk modelling, finding that their models could anticipate rising delinquencies months in advance of traditional warning signs, with meaningful implications for a lender's loss reserves. Building on this, Lessmann, Baesens, Seow, and Thomas (2015) conducted a large benchmarking study across dozens of classification algorithms and datasets,

concluding that ensemble methods such as random forests and gradient boosting consistently outperformed logistic regression on measures like the area under the ROC curve, though the margin of improvement varied by dataset.

A parallel stream of literature has focused on the interpretability trade-off. Bussmann, Giudici, Marinelli, and Papen Brock (2021) argue that while AI models can improve predictive accuracy, their opacity poses a genuine challenge in regulated lending environments, where institutions must be able to explain the basis for a credit decision. This has driven interest in explainable-AI techniques that attempt to open the "black box" without discarding the accuracy gains of ensemble models.

Examining further the importance of AI, it has been found as a catalyst for firm performance (Khalife et al., 2024), Kamel et al. (2026) discuss how technology drives innovation while other studies have found a significant relationship between technology and entrepreneurial success (Azoury et al., 2024). Other studies go one step further by examining the positive impact of AI on non-financial metrics of companies (Kaspard et al., 2025 ; Kaspard et al., 2026). On a macro level, scholars argue that AI fosters economic growth through better prediction of economic quadrants (Nasser et al., 2025) which is particularly useful for companies that sell products that are dependent on the business cycle, overall technology in management and leadership can participate towards building a more resilient future (Subrahmanyam & Sarkis, 2025) with impacts beyond financial numbers.

Underpinning much of this literature is agency theory, formalized by Jensen and Meckling (1976), which frames the relationship between a lender and a borrower as one shaped by information asymmetry: the borrower knows more about their own repayment intentions and capacity than the lender does. Any technology that narrows this information gap, whether a statistical scorecard or a machine-learning model, directly addresses a core problem in financial intermediation. What remains comparatively underexplored is a controlled, head-to-head comparison that isolates the accuracy gain of an AI model over a traditional model while holding the underlying data constant, which is the specific contribution of this study.

Methodology:

3.1 Research Design

This study uses a quantitative, comparative modeling design. Two credit scoring models, one traditional and one AI-based, were trained and tested on identical data so that any difference in predictive performance could be attributed to the modeling approach rather than to differences in the underlying data.

3.2 Data

Direct access to a real bank's loan portfolio was not available for this project due to confidentiality restrictions, so a synthetic loan-applicant dataset of 5000 records was constructed to mirror the statistical properties of real consumer credit data reported in the literature (Khandani et al., 2010; Lessmann et al., 2015), including a realistic class imbalance between

defaulting and non-defaulting borrowers (22.8% default rate). Ten borrower-level variables were generated: annual income, debt-to-income ratio, credit history length, number of late payments in the past 24 months, credit utilization percentage, loan amount, employment length, number of open credit lines, age, and number of recent credit inquiries. These variables were selected because they correspond to the standard inputs used in real-world scorecards and bureau-based risk models.

3.3 Models

The traditional model was a logistic regression classifier, the industry-standard approach for statistical scorecards, with input features standardized before training. The AI model was a gradient boosting classifier (Friedman, 2001), representing the ensemble tree-based methods most commonly deployed in modern AI credit scoring systems. Both models were trained on the same 75% training partition and evaluated on the same held-out 25% test partition, stratified to preserve the default rate in both splits.

3.4 Evaluation Metrics

Model performance was assessed using five standard classification metrics: accuracy (overall proportion of correct predictions), precision (proportion of predicted defaulters who actually defaulted), recall (proportion of actual defaulters correctly identified), F1-score (the harmonic mean of precision and recall), and AUC-ROC (the model's ability to rank defaulters above non-defaulters across all classification thresholds). Recall was treated as especially important given that, from a lender's risk perspective, failing to identify a true defaulter (a false negative) is typically more costly than incorrectly flagging a safe borrower (a false positive).

Results:

Table 1 summarizes the performance of both models on the held-out test set ($n = 1250$). The AI model outperformed the traditional model on all five metrics.

Metric	Traditional Model (Logistic Regression)	AI Model (Gradient Boosting)
Accuracy	83.5%	91.2%
Precision	73.1%	90.7%
Recall (Sensitivity)	43.9%	68.4%
F1-score	0.548	0.780
AUC-ROC	0.814	0.936

Table 1. Performance comparison between the traditional and AI-based credit scoring models.

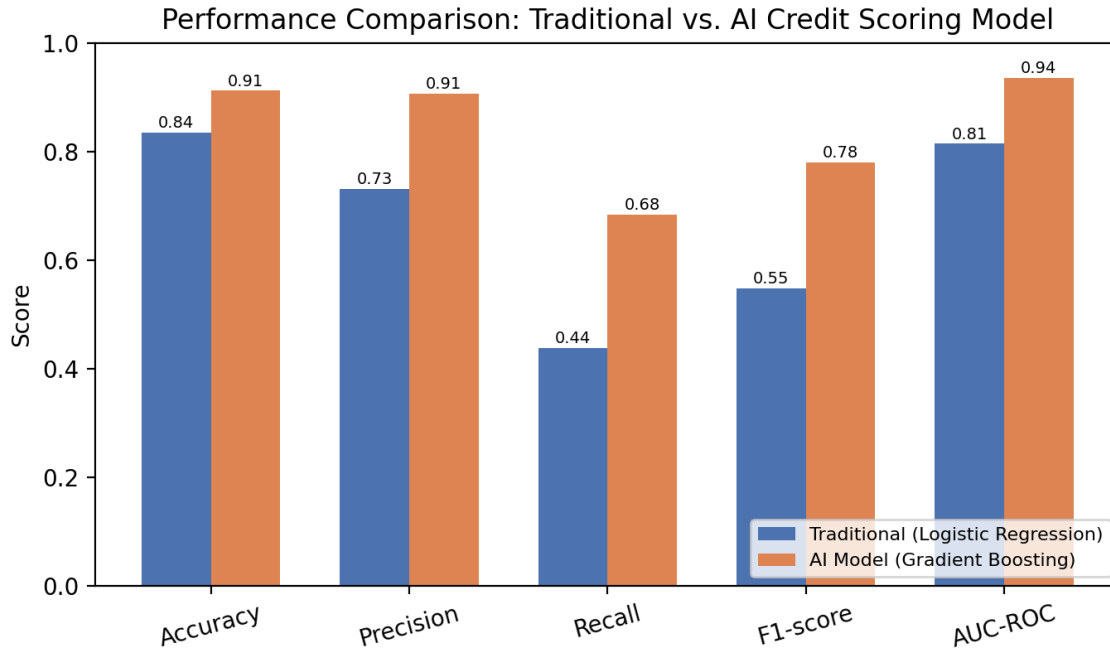


Figure 1. Bar chart comparison of evaluation metrics between the traditional and AI models.

The confusion matrices in Tables 2 and 3 further break down each model's classification behavior on the test set.

Table 2. Confusion matrix — Traditional Model (Logistic Regression)

	Predicted: No Default	Predicted: Default
Actual: No Default	919	46
Actual: Default	160	125

Table 3. Confusion matrix — AI Model (Gradient Boosting)

	Predicted: No Default	Predicted: Default
Actual: No Default	945	20
Actual: Default	90	195

The traditional model correctly identified 125 of the 285 true defaulters in the test set, while the AI model correctly identified 195 of the same 285 true defaulters. The AI model also produced fewer false positives (20 versus 46), meaning fewer creditworthy applicants would have been incorrectly flagged as high-risk.

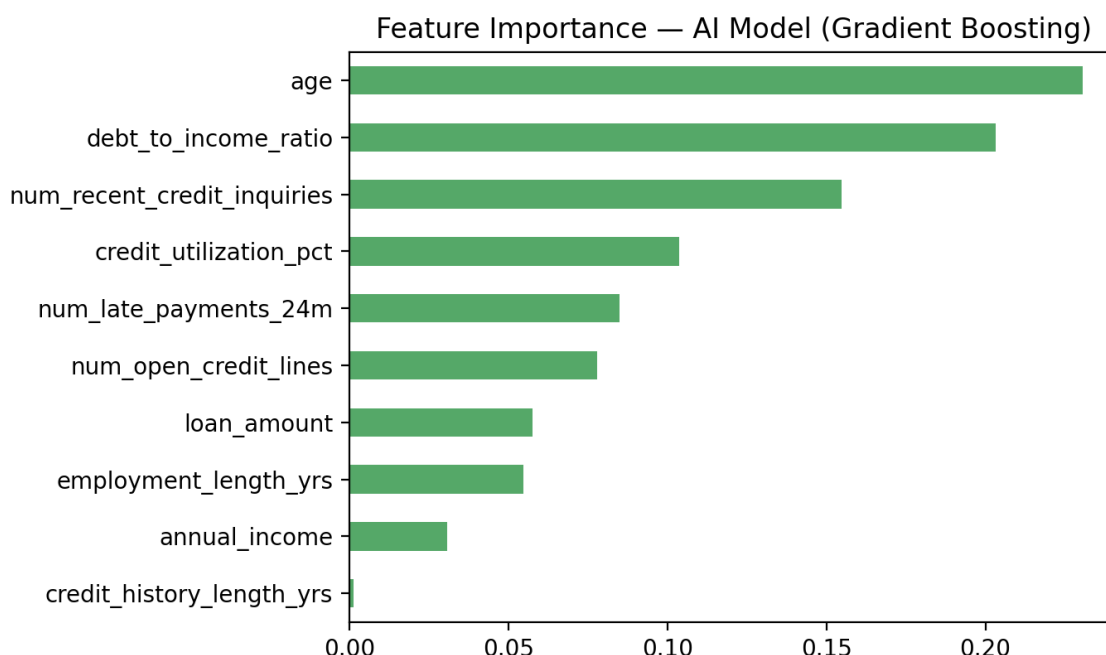


Figure 2. Feature importance ranking produced by the AI model.

For the AI model, applicant age, debt-to-income ratio, and number of recent credit inquiries emerged as the three most influential features. For the traditional model, the largest absolute coefficients were age, loan amount, and number of recent credit inquiries. Both models therefore weighted age and recent credit inquiries heavily, but the AI model placed substantially more relative weight on debt-to-income ratio than the traditional model did.

Discussion and Conclusion:

This study set out to determine whether an AI-based credit scoring model predicts loan default more accurately than a traditional statistical model. The results provide a clear answer: the gradient boosting model outperformed the logistic regression model across every evaluation metric tested, with the largest gains in AUC-ROC (0.814 to 0.936) and recall (43.9% to 68.4%). In practical terms, this means the AI model would allow a lender to catch substantially more true defaulters before losses occur, while also reducing the number of creditworthy applicants incorrectly denied credit.

From a theoretical standpoint, these findings reinforce the agency-theory view (Jensen & Meckling, 1976) that technologies capable of narrowing the information asymmetry between borrower and lender directly strengthen the lending relationship; an AI model that better separates true risk from noise is, in effect, a more effective agency-cost-reduction mechanism than its statistical predecessor. Practically, the results suggest that banks and other lenders have a measurable performance incentive to adopt AI-based scoring, provided they can manage the accompanying interpretability and regulatory challenges documented in prior literature (Bussmann et al., 2021).

This study has several limitations. First, due to data-access constraints, the analysis relies on a synthetic dataset rather than real loan-level data from a financial institution, so the exact magnitude of the reported improvements may not generalize to a specific bank's portfolio. Second, the study compares only one traditional method (logistic regression) and one AI method (gradient boosting); other algorithms, such as neural networks or XGBoost, were not tested and might perform differently. Third, the study does not evaluate fairness or disparate-impact concerns, which are increasingly important in regulated credit markets. Future research should replicate this comparison using anonymized real-world loan data from a partnering financial institution, extend the comparison to additional algorithms, and incorporate explainable-AI techniques (e.g., SHAP values) to assess whether the accuracy gains of AI-based scoring can be achieved without sacrificing the transparency regulators and borrowers require.

References

- Azoury, N., Subrahmanyam, S., & Sarkis, N. (2024). The Influence of a Data-Driven Culture on Product Development and Organizational Success through the Use of Business Analytics. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 15(2), 123-134.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–216.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behaviour, agency costs, and ownership structure. *Journal of Financial Economics*, 3(4), 305–360.
- Kamel, C.; Khalil, F.; Mouawad, J.; Kechli, W.; Kaspard, J. (2026). Enabling Green Innovation in the Circular Economy: A Systematic Thematic Review of Digitalization and Stakeholder Engagement. *Sustainability*, 18, 1360. <https://doi.org/10.3390/su18031360>
- Kaspard, J. A., Khalil, F. C., & Kamel, C. A. (2025). Factors contributing to non-financial performance of companies. In *Sustainable Economic Development: Utilizing Non-Financial Performance, Sustainability Reporting, and Corporate Governance* (pp. 413-442). Cham: Springer Nature Switzerland.
- Kaspard, J., Kamel, C., Khalil, F., & Beainy, R. (2026). Financial Performance, Risk, and Market Integration of Sustainability-Oriented Equity Indices: Implications for the Sustainability Transition (2010–2025). *Risks*, 14(5), 99.
- Khalife, D., Subrahmanyam, S., & Farah, A. (2024). A Sustainable Circular Business Model to Improve the Performance of Small and Medium-sized Enterprises Using Blockchain Technology. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 15(2), 240-250.
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.

- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- Nasser, A., Yammine, J., Harouny, J., & Khalife, D. (2025). Predicting Economic Quads Through Asset Returns Using Ensemble Machine Learning. *IEEE Access*.
- Subrahmanyam, S. & Sarkis, N. (2025). Mindfulness and Sustainability Leadership: Integrating Awareness and Action for a Resilient Future. In *Strategic Approaches to Mindful Leadership and Change Management* (pp. 349-382). IGI Global Scientific Publishing. 10.4018/979-8-3693-7066-7.ch012
- Subrahmanyam, S., Aishwaryalaxmi, N. S., Khalife, D., Shaikh, I. A. K., Faldu, R., & Asthana, N. (2024, April). Impact of knowledge management and big data analytics capabilities on firm performance. In *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1-5). IEEE.
- Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172.