

Artificial Intelligence in Credit Risk Assessment:**A Comparative Study of Machine Learning Models versus Traditional Credit Scoring in Retail Lending**Author: **Melanie El Khoury¹***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 evaluates the transition from traditional logistic-regression scorecards to advanced machine learning (ML) models in retail credit risk assessment. As financial institutions increasingly adopt data-driven cultures to navigate macroeconomic uncertainty, a central question emerges: do ML algorithms provide sufficient predictive and inclusive advantages to justify their ethical and regulatory complexities? Utilizing a comparative secondary-research methodology, this paper synthesizes evidence from a massive academic benchmark (Lessmann et al., 2015), real-world regulatory reporting on Upstart's AI underwriting model (CFPB), and a comprehensive analysis of 26.2 million credit-bureau records (Albanesi & Vamossy, 2024). The findings demonstrate that ensemble methods, particularly random forests, significantly outperform traditional models in discriminatory accuracy, achieving up to a 0.171 AUC advantage. In applied settings, AI-driven underwriting expanded credit access by 27% to 43.4% while substantially lowering average interest rates. Furthermore, ML models proved highly effective in reducing misclassification (by 41%), with the most significant accuracy gains observed among low-income, young, and minority borrowers. However, integrating these predictive "black box" frameworks into corporate business culture introduces severe challenges regarding algorithmic transparency, ethical governance, and trust. The study concludes that while ML offers transformative improvements in credit access and risk prediction, financial institutions must balance these technological advantages with rigorous internal controls, deep business expertise, and transparent regulatory compliance to ensure equitable and responsible lending.

Keywords:

Artificial Intelligence (AI), Machine Learning, Credit Risk, Credit Scoring, Retail Lending, Financial Inclusion, Random Forest, Logistic Regression

Introduction:

Retail credit scoring is one of the most established applications of statistical modeling in business. For decades, logistic regression has been widely used in credit-scoring scorecards to estimate the probability that a borrower will default and to support decisions concerning approval and pricing. These models remain important because they are relatively transparent, interpretable, and well established within financial institutions.

However, modern lending increasingly involves large digital datasets, alternative information sources, and borrowers with limited traditional credit histories. Applicants with thin or limited credit files may be difficult to evaluate using conventional credit information alone. At the same time, machine-learning classifiers such as random forests, gradient boosting methods, and neural networks can identify nonlinear relationships and interactions among large numbers of variables. This development raises an important research question: does replacing or supplementing traditional logistic-regression credit scoring with machine-learning models produce measurable improvements in predictive accuracy and credit access, and are those improvements sufficient to justify the additional complexity and regulatory considerations associated with these models?

To answer this question, the study pursues four objectives:

1. To Compare the discriminatory accuracy of machine-learning classifiers with logistic regression using evidence from a major academic benchmark.
2. To Examine whether differences in model performance are also observed in a real-world lending application involving Upstart's AI-based underwriting system.
3. To Examine differences in predictive performance across borrower groups using the large-sample study of Albanesi and Vamossy (2024).
4. To Assess the governance, fairness, interpretability, and regulatory issues associated with the use of machine learning in credit assessment.

Literature Review:

Credit scoring has been described as one of the most successful applications of statistics in business, and researchers have spent decades examining whether alternative statistical and machine-learning techniques can improve its performance. An important early benchmarking study was conducted by Baesens et al. (2003). The study compared 17 classification algorithms across eight real-world credit datasets and found that several more flexible techniques, including neural networks, could achieve competitive or superior discriminatory performance relative to traditional approaches. This research provided an important foundation for later comparisons of credit-scoring methods. A major update was provided by Lessmann, Baesens, Seow, and Thomas (2015). Their study benchmarked 41 classification algorithms across eight real-world credit-scoring datasets. The algorithms included logistic regression, random forests, boosted trees, and other ensemble and heterogeneous classification methods. The datasets varied considerably in size

and characteristics, allowing the authors to compare model performance across different lending environments.

The study evaluated the classification methods using multiple performance measures, including AUC, percentage correctly classified, Brier score, H-measure, partial Gini, and Kolmogorov-Smirnov statistics. The results showed that ensemble methods performed particularly well across the benchmark, although no single model was universally superior across every dataset and every performance measure. Parallel to the academic literature, a real-world case involving Upstart Network provided evidence concerning the potential effects of AI-based underwriting. Upstart sought a No-Action Letter from the U.S. Consumer Financial Protection Bureau in 2017 in connection with the use of alternative data and machine-learning techniques in underwriting. The CFPB subsequently reported results comparing Upstart's AI-based underwriting model with a traditional credit-score-only approach. The CFPB reported that the Upstart model was associated with 27% more approved borrowers and 16% lower average APRs than the traditional credit-score-only model. The CFPB also reported that these results were observed across the demographic groups tested under the company's fair-lending compliance plan. These findings provide useful evidence about the possible effects of AI underwriting on credit access and pricing, although the reported results should not be interpreted as an independent replication of Upstart's internal analysis. A third strand of research examines the relationship between credit-score performance and borrower characteristics. Albanesi and Vamossy (2024) used 26.2 million Experian credit-bureau records covering 2006–2016 and developed a machine-learning model to predict default. They compared its performance with a conventional credit score.

The authors found that the conventional credit score misclassified 41% of consumers into the wrong risk category, with the rate increasing to 47% among Subprime borrowers and 70% among Near Prime borrowers. They also found that the difference in predictive performance between the conventional score and the machine-learning model was larger for several borrower groups, including low-income, young, and minority borrowers. Together, these studies provide evidence from different perspectives. Lessmann et al. provide a broad academic benchmark, the CFPB/Upstart case provides evidence from an actual lending application, and Albanesi and Vamossy provide evidence from a very large credit-bureau dataset. This study brings these sources together to examine what the evidence suggests about the potential advantages and limitations of machine-learning credit assessment.

Macroeconomic Dynamics, Ensemble Learning, and Data-Driven Culture

The transition from traditional logistic regression scorecards to advanced machine learning models represents a fundamental shift toward a data-driven organizational culture, which requires robust big data analytics and knowledge management capabilities to improve overall firm performance (Azoury et al., 2024; Subrahmanyam et al., 2024). The superior discriminatory power of ensemble machine learning frameworks observed in credit scoring mirrors their broader success across

financial markets, where they are actively used to predict complex economic cycles and analyze deep-learning shifts in asset returns and investor sentiment (Khalife et al., 2025; Nasser et al., 2025). For retail lenders operating in vulnerable economic environments, navigating macroeconomic interrelatedness and financial uncertainty requires highly adept analytical tools (Athari et al., 2023; Bejjani et al., 2024). In these contexts, machine learning models function as smart decision support systems, far surpassing traditional vulnerability metrics like the Altman Z-score in forecasting retail and corporate financial distress amid Industry 4.0 disruptions (Elia et al., 2021; Subrahmanyam & Beainy, 2027).

Trust, Ethics, and the Governance of AI in Lending

Despite the predictive advantages of machine learning, deploying "black box" algorithms for consumer credit introduces severe ethical challenges regarding transparency, bias, and accountability within business culture (Diab & El Hajj, 2024). As evidenced in related financial sectors like auditing, the widespread adoption of artificial intelligence is heavily dictated by the perceived trust, operational benefits, and behavioral profiles of the institutions deploying these tools (Jabbour Al Maalouf et al., 2026; Sawaya et al., 2026). To satisfy regulatory bodies such as the CFPB and ensure equitable credit access, financial institutions cannot rely on algorithmic accuracy alone; they must supplement predictive models with deep business experience (Chaanine et al., 2023) and rigorous internal controls to maintain absolute financial transparency (Saber & Sassine, 2022). For smaller financial enterprises and SMEs adopting these technologies, navigating equity financing and securing consumer data may eventually rely on sustainable, blockchain-enabled circular business models to ensure data integrity and compliance without sacrificing predictive power (Khalife et al., 2024; Rizk & Sassine, 2023).

Methodology:

This paper uses a comparative secondary-research design rather than developing a new predictive model. This approach is appropriate because the objective is to compare evidence from established studies and a real-world lending application rather than to produce new statistical estimates from a small classroom dataset.

The analysis draws on three main sources of evidence.

The first source is the benchmarking study by Lessmann et al. (2015). The study compared 41 classification algorithms across eight real-world credit datasets using multiple performance measures. For the main comparison presented in this paper, particular attention is given to logistic regression and random forest and to their AUC results. AUC measures discriminatory ability, with 0.5 representing performance equivalent to random classification and 1.0 representing perfect discrimination.

The eight datasets in the Lessmann et al. benchmark include Australian Credit, German Credit, Th02, Bene 1, Bene 2, UK bank, PAKDD, and Give Me Some Credit. The "Give Me Some Credit" dataset contains approximately 150,000 observations and is the largest dataset in the benchmark.

The second source is the Upstart case reported by the CFPB and supplemented by Upstart's subsequent public disclosures. The CFPB reported comparisons between Upstart's AI-based underwriting approach and a traditional credit-score-only model. The analysis in this paper uses the CFPB's reported 27% increase in approvals and 16% reduction in average APR, while separately reporting Upstart's later disclosed figures of 43.4% more approvals and 43.2% lower average APR.

The third source is Albanesi and Vamossy's (2024) study based on 26.2 million Experian credit-bureau records. The study compares a conventional credit score with a machine-learning default model and examines differences across borrower groups. These results are used to investigate whether differences in predictive accuracy vary systematically across income, age, and minority status.

Combining these sources allows the study to examine machine-learning credit assessment from three perspectives: academic model benchmarking, a real-world lending application, and large-scale borrower-level evidence.

Results:

Table 1 reports the AUC results for logistic regression and random forest across the seven datasets selected for the comparison below. The complete Lessmann et al. benchmark contained eight datasets, including Th02.

Table 1: AUC of Logistic Regression vs. Random Forest

Dataset	Cases	Logistic Regression (AUC)	Random Forest (AUC)	RF Advantage
Australian Credit	690	0.931	0.931	Statistical tie
German Credit	1,000	0.784	0.789	+0.005
Bene 1 (Benelux bank)	3,123	0.779	0.796	+0.017
Bene 2 (Benelux bank)	7,190	0.792	0.805	+0.013
UK bank	30,000	0.720	0.742	+0.022
PAKDD (bank)	50,000	0.626	0.643	+0.017
Give Me Some Credit	150,000	0.693	0.864	+0.171

Source: Lessmann, Baensens, Seow & Thomas (2015). AUC values reported from the study's benchmark results. Higher AUC indicates better discriminatory performance.

The results show that random forest matched logistic regression on the Australian Credit dataset and achieved a higher AUC on the other six datasets shown in the table. The largest difference occurred in the 150,000-observation Give Me Some Credit dataset, where random forest achieved an AUC of 0.864 compared with 0.693 for logistic regression.

The 0.171 AUC difference in the Give Me Some Credit dataset is substantially larger than the differences observed in the other datasets. This demonstrates that the magnitude of any machine-

learning advantage can depend heavily on the dataset and lending environment.

Lessmann et al. also found that ensemble methods performed strongly when evaluated across the broader set of classification algorithms and performance measures. Therefore, the evidence does not support the claim that one machine-learning model is universally superior, but it does demonstrate that several machine-learning approaches can outperform logistic regression in particular credit-scoring applications.

Real-World Lending Evidence: Upstart

The real-world evidence provides a different perspective from the academic benchmark.

According to the CFPB's 2019 report concerning Upstart's No-Action Letter, the company's AI-based underwriting model was associated with 27% more approved borrowers and 16% lower average APRs than a traditional credit-score-only model.

The CFPB also reported that the results were observed across the race, ethnicity, and sex groups examined under Upstart's fair-lending compliance framework. These results should be interpreted within the scope of the CFPB's reporting framework rather than as proof that all AI underwriting systems eliminate discrimination.

Upstart subsequently reported larger figures in its 2022 Form 10-K. According to the company's disclosure, an internal study as of December 31, 2021 found that its AI model approved 43.4% more borrowers and produced a 43.2% lower average APR compared with a traditional credit-score-only model.

These two sets of figures should be distinguished because they come from different reporting contexts and dates. The CFPB figures relate to its 2019 reporting, whereas the 43.4% and 43.2% figures were subsequently disclosed by Upstart in its SEC filing.

Evidence from 26.2 Million Credit-Bureau Records

Albanesi and Vamossy (2024) provide another important source of evidence. Their study uses 26.2 million Experian credit-bureau records and compares a conventional credit score with a machine-learning default model. The overall AUC for the conventional credit score was 0.856, compared with 0.906 for the machine-learning model, producing a difference of 0.050.

The differences were larger for several borrower groups:

Table 2: AUC of Conventional Credit Score vs. Machine-Learning Model

Borrower Group	Conventional Credit Score (AUC)	ML Model (AUC)	Gap
All borrowers	0.856	0.906	+0.050
Bottom income quintile	0.735	0.855	+0.120
Age under 30	0.824	0.881	+0.057
Minority borrowers	0.802	0.881	+0.079

Source: Albanesi & Vamossy (2024), based on 26.2 million Experian credit-bureau records covering 2006Q1–2016Q2.

The largest difference in the table occurs among borrowers in the bottom income quintile, where the AUC gap is 0.120. The gap is 0.057 for borrowers under age 30 and 0.079 for minority borrowers.

The study also reports that the conventional credit score misclassified 41% of consumers into the wrong risk category. The corresponding figure was 47% for Subprime borrowers and 70% for Near Prime borrowers. These findings suggest that the relative performance of conventional and machine-learning approaches can vary across borrower groups. They do not, however, establish that machine learning will produce the same results in every lending market or under every model specification.

Discussion and Conclusion:

The three sources examined in this paper provide complementary evidence concerning the use of machine learning in credit risk assessment. The Lessmann et al. (2015) benchmark demonstrates that machine-learning and ensemble methods can achieve higher discriminatory performance than logistic regression on several real-world credit datasets. The Upstart case provides evidence that differences in underwriting methodology can be associated with changes in approval rates and pricing. Finally, Albanesi and Vamossy (2024) show that differences in predictive accuracy can be particularly large for certain borrower groups.

The academic evidence is particularly important because it demonstrates that the performance advantage is not limited to a single model or dataset. In the Lessmann benchmark, random forest outperformed or matched logistic regression on the datasets presented in this paper, with the largest difference occurring in the Give Me Some Credit dataset. At the same time, the differences varied substantially across datasets, which indicates that model performance depends on the characteristics of the data and the lending environment.

The Upstart evidence illustrates how model differences can potentially translate into practical lending outcomes. The CFPB reported 27% more approvals and 16% lower average APRs in its 2019 analysis of Upstart's model compared with a traditional credit-score-only approach. Upstart later disclosed larger results from an internal study. These findings indicate that machine-learning underwriting can have effects beyond statistical accuracy, including potential effects on credit access and loan pricing.

The Albanesi and Vamossy study adds an important distributional dimension. Their results show that the difference in AUC between the conventional credit score and the machine-learning model was larger for the bottom income quintile, borrowers under 30, and minority borrowers than for the overall population. This suggests that improvements in predictive accuracy may not be evenly distributed across borrowers.

However, higher predictive performance does not eliminate the challenges associated with machine-learning credit models. Machine-learning systems can be more difficult to interpret than traditional logistic-regression scorecards. This creates challenges for model validation, governance, consumer explanations, and regulatory compliance.

Another concern is the use of alternative data. Variables that are not themselves protected characteristics can potentially be correlated with protected characteristics and therefore create fairness concerns. For this reason, financial institutions need appropriate model validation, monitoring, documentation, and fair-lending testing when deploying machine-learning systems. The evidence therefore supports a more nuanced conclusion than simply stating that artificial intelligence is always superior to traditional credit scoring. Several machine-learning approaches demonstrate measurable predictive advantages over traditional scoring methods in particular datasets and applications, but the magnitude of those advantages varies and must be evaluated alongside interpretability, fairness, data quality, and regulatory requirements.

Future research should extend this comparison to additional areas of credit risk, including loss-given-default prediction, macroeconomic stress testing, model stability over time, and the performance of newer artificial-intelligence approaches. Future studies should also examine whether the results observed in large institutions can be replicated by smaller lenders that may have fewer resources for model validation, monitoring, and regulatory compliance.

The evidence therefore suggest that financial institutions should not assume that machine learning automatically replaces traditional credit scoring, but should evaluate each model using rigorous evidence concerning predictive performance, fairness, interpretability, stability, and regulatory requirements.

References

- Albanesi, S., & Vamossy, D. F. (2024). Credit Scores: Performance and Equity. NBER Working Paper No. 32917. National Bureau of Economic Research.
- Athari, S. A., Saliba, C., Khalife, D., & Salameh-Ayanian, M. (2023). The role of country governance in achieving the banking sector's sustainability in vulnerable environments: New insight from emerging economies. *Sustainability*, 15(13), 10538.
- 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. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 15(2), 123-134.
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635.
- Bejjani, G., Toros, E., Elia, J., & Jabbour Al Maalouf, N. (2024). Technology analytic adeptness for examining the interrelatedness between the main macroeconomic variables in Lebanon employing a vector autoregression (VAR). *Journal of Decision Systems*, 1-30.
- Chaanine, N., Mhanna, R. Y., Verdie, J. F., & Beainy, R. (2023). Leveraging business experience and knowledge to boost entrepreneurial achievement. *International Journal of Work Innovation*, 4(2), 104-115.

- Consumer Financial Protection Bureau. (2019). An Update on Credit Access and the Bureau's First No-Action Letter. Washington, DC: Consumer Financial Protection Bureau.
- Diab, B., & El Hajj, M. (2024). Ethics in the age of artificial intelligence: unveiling challenges and opportunities in business culture. *Cogent Business & Management*, 11(1). <https://doi.org/10.1080/23311975.2024.2408440>
- Elia, J., Toros, E., Sawaya, C., & Balouza, M. (2021). Using Altman Z"-score to predict financial distress: Evidence from Lebanese alpha banks. *Management Studies and Economic Systems*, 6(1/2), 47-57.
- Jabbour Al Maalouf, N., Ficani, A., & Elia, J. (2026). Drivers of AI adoption in auditing: a quantitative exploration of perceived benefits and trust across auditor profiles. *Cogent Business & Management*, 13(1). <https://doi.org/10.1080/23311975.2026.2709916>
- 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. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 15(2), 240-250.
- Khalife, D., Yammine, J., El Bazi, T., Zaki, C., & Jabbour Al Maalouf, N. (2025). Analyzing the impact of investor sentiment on S&P 500 prices using deep learning models. *Journal of Financial Reporting and Accounting*, 23(4), 1365-1386.
- 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*.
- Rizk, R., & Sassine, M. (2023). Equity financing and SME growth: evidence from the Eurozone. *International Journal of Work Innovation*, 4(3), 227-239.
- Saber, A. M., & Sassine, M. E. (2022). Reflections on internal control and internal audit roles for ensuring financial transparency toward Lebanese NGOs financial donors. *EuroMed Journal of Management*, 4(3), 186-198.
- Sawaya, C., Khalil, F. C., Jabbour Al Maalouf, N., & Kaspard, J. (2026). Revolutionizing audit practice in emerging economies: the role of Artificial Intelligence in enhancing the audit of consolidated financial statements. *Cogent Business & Management*, 13(1), 2625974.
- Subrahmanyam, S., & Beainy, R. H. (2027). Smart Decision Support for Uncertainty and Disruptions in Industry 4.0 Operations. In *Smart Operational Research Solutions in the Era of Industry 4.0 and Sustainability* (pp. 167-198). IGI Global Scientific Publishing.
- 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.

Upstart Holdings, Inc. (2023). Annual Report (Form 10-K), fiscal year ended December 31, 2022. U.S. Securities and Exchange Commission.