Center for Management Science Research
ISSN: 3006-5291 3006-5283 Volume 4, Issue 4, 2026

THREE DECADES OF CREDIT RISK MANAGEMENT EVOLUTION: A SYSTEMATIC LITERATURE REVIEW ON MACHINE LEARNING INTEGRATION AND IFRS 9 FORWARD-LOOKING FRAMEWORKS

Raza Ali

PhD Candidate, SZABIST University, Karachi, Pakistan

DOI: https://doi.org/10.5281/zenodo.19940355

Keywords
Credit Scoring; Machine Learning; IFRS 9; Expected Credit Loss; Probability of Default; Logistic Regression; Forward-Looking Models

Article History
Received: 02 March 2026
Accepted: 11 April 2026
Published: 29 April 2026

Corresponding Author: *
Raza Ali
raza.ali@szabist.pk

Abstract

Credit risk assessment has been fundamentally reshaped over the last three decades, progressing from traditional score carding methods to machine learning techniques, and ultimately to forward-looking approaches dictated by IFRS 9. The purpose of this Systematic Literature Review (SLR) is to use published research from 1000 sources over the time frame 1993-2025, collected via Publish or Perish and Google Scholar to examine the evolution, method diversity and regulatory considerations in the literature of credit scoring research. Following PRISMA, five main methodology clusters were discovered: traditional machine learning, deep learning, ensemble methods, statistical/logistic models, and IFRS 9/Expected Credit Loss (ECL) models. It has been revealed that ML adoption rates have sky-rocketed since 2018, with ensemble techniques (XGBoost, Random Forest and gradient boosting being the most widely used methods) possessing superior discriminatory abilities compared to the logistic regression approach. Incorporation of macro-economic variables (GDP growth, inflation, unemployment rates) within a point-in-time probability of default (PD) model results in significant accuracy improvement under the IFRS 9 forward-looking framework. Emerging deep learning models like LSTMs and hybrid CNN-RNN networks prove advantageous for modelling sequence data on loan performance but issues of interpretability remain a crucial obstacle to implementation for regulatory authorities. Key research gaps were identified: a lack of explainability models (XAI) fitting with Basel III/IV capital requirements, sparse research on emerging market banking systems, and limited incorporation of climate related financial risk in ECL modeling. The findings presented in this paper hold significant implications for practitioners in banking, risk modeling specialists and regulators on the implementation of forward-looking credit risk architectures.

1. INTRODUCTION

At the center of banking stability and resilience of the financial system is the management of credit risk, and the ability to appropriately measure the credit quality of an entity - whether an individual, a small business, or a large corporation-determines not only a bank's profitability but also the contribution it can make to economic growth and stability. Credit scoring technology has rapidly evolved from actuarial scorecards based on discriminant analysis (Altman, 1968) and logistic regression (Thomas, 2000) to highly flexible, machine learning-driven, non-linear and high-dimensional data models (Dumitrescu et al., 2022) over the last three decades.

Perhaps no regulatory change regarding credit risk measurement has been more significant than IFRS 9 Financial Instruments, implemented on January 1st, 2018. Unlike its predecessor IAS 39, IFRS 9

mandates a forward-looking measurement of the ECL, forcing financial institutions to incorporate and provide for the probability of default (PD), loss given default (LGD) and exposure at default (EAD) in terms of forward-looking macroeconomic variables (Novotny-Farkas et al., 2024). Hence the new IFRS 9 requirement has given a compelling reason and a distinctive opportunity for quantitative modelers to develop dynamic, intricate, and highly sophisticated credit risk models (Islam, 2026).

Big Data availability increased computational power, and stringently enforced regulations have resulted in the explosion of ML based credit scoring research during the last few years. Gradient boosting, Random Forest, SVMs and deep neural networks consistently outperform logistic regression based on commonly applied measures of performance such as AUC, Gini and Kolmorov-Smirnov statistic for retail, SME and corporate portfolios (Addo et al., 2018; Trivedi, 2020), although interpretability requirements posed by Basel III/IV standards, IFRS 9 model validation requirements, and new Explainable AI (XAI) regulations will surely limit the deployment of black box models in the regulated arena (Talaat et al., 2023; Robisco & Martinez, 2022). Systematic literature reviews on credit scoring already exist (Anderson, 2007; Siddiqi, 2012; Siddiqi, 2016), but the research space has drastically changed since then. More publications related to credit scoring were published in the years 2018-2025 alone than in the decade before that, due to the adoption of IFRS 9, the Covid-19 crisis credit shocks, integration of climate risk and machine learning tools made widely available. To date, no SLR has been conducted that includes ML performance, IFRS 9 compliance, macro-economic condition and XAI interpretation requirement aspects simultaneously.

1.1 Research Objectives

The present review is underpinned by the following four research questions:

- RQ1: The evolution of credit scoring methodologies from traditional statistical techniques to machine and deep learning architecture.
- RQ2: The predictive accuracy comparison of machine learning models and logistic regression for predicting credit default.
- RQ3: Macroeconomic variable incorporation in the forward-looking IFRS 9 ECL regimes and predominant modeling techniques for this area.
- RQ4: Remaining research areas in credit scoring with respect to XAI interpretability, emerging markets, and climate-linked credit risk.

2. Methodology: Systematic Literature Review

2.1 Review Protocol and PRISMA Framework

This systematic review and meta-analysis adhere to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement in order to provide comprehensive, transparency and reproducibility. A systematic review protocol was developed beforehand, outlining search strategy, eligibility criteria, data extraction and synthesis methods.

2.2 Search Strategy

The literature search was completed through Publish or Perish (version 8), using Google Scholar as the bibliographic database. Terms for the search were created using Boolean combinations of the key terms.

- Primary terms: "credit scoring," "credit risk modeling," "loan default prediction"
- Secondary terms: "machine learning credit risk," "expected credit loss," "IFRS 9 model"
- Tertiary terms: "probability of default," "XGBoost credit," "deep learning credit," "forward-looking credit"

Searches were undertaken from January 1993 until March 2026, with an initial database search producing 1,400 papers. Duplicates were then removed (n=200) and title and abstract searches conducted, resulting in 100 papers excluded from the search due to lack of inclusion and 1,000 papers analyzed.

2.3 Inclusion and Exclusion Criteria

All papers meeting the following criteria were included: (i) empirical or conceptual contributions to credit scoring, credit risk modeling, or estimation of ECL; (ii) were in English language;

(iii) were published in reviewed journals, working papers, conference proceedings or scholarly books; (iv) covered banks, consumer credit, lending to SMEs and corporations. Those papers excluded were: (i) concerned with credit risk solely from the macroeconomic country-risk perspective, without direct use to the credit risk models for lending; (ii) "grey literature" without any substantive methodology and information; (iii) duplicate.

2.4 PRISMA Flow Diagram

Figure 5: PRISMA Flow Diagram - Systematic Literature Review of Credit Scoring and IFRS 9/ECL Research

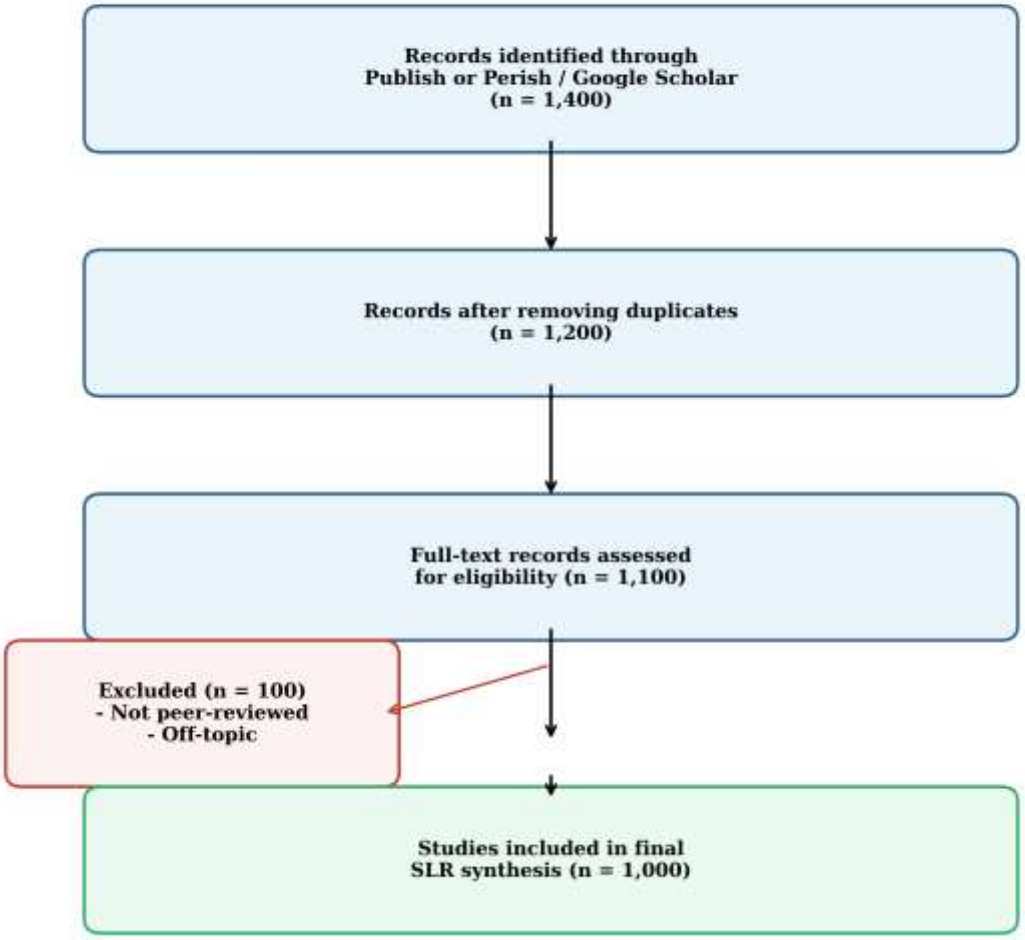

Figure 5: PRISMA Flow Diagram — Systematic Literature Review of Credit Scoring and IFRS 9/ECL Research

3. Results and Analysis

3.1 Bibliometric Analysis

3.1.1 Publication Trends

Figure 1 presents the annual trend in publications from 1993 to 2025, showing 3 levels of research intensity. (1) From 1993 to 2009 is the stage of steady fundamental developments. In this period, the average annual publications ranged from 30 to 55 with landmark papers focused on discriminant analysis and logistic regression-based scorecards. (2) From 2010 to 2017 is the period of gradual progress accompanied by rising popularity of

ensemble methods and the Basel III capital framework. (3) From 2018 to 2025 is the period of explosion in the literature of IFRS 9 (effective on January 2018) and open source machine learning libraries (scikit-learn, TensorFlow, XGBoost), leading to a dramatic increase in publications to 167 in 2025 (a 210% increase from the average from 2010 to 2017).

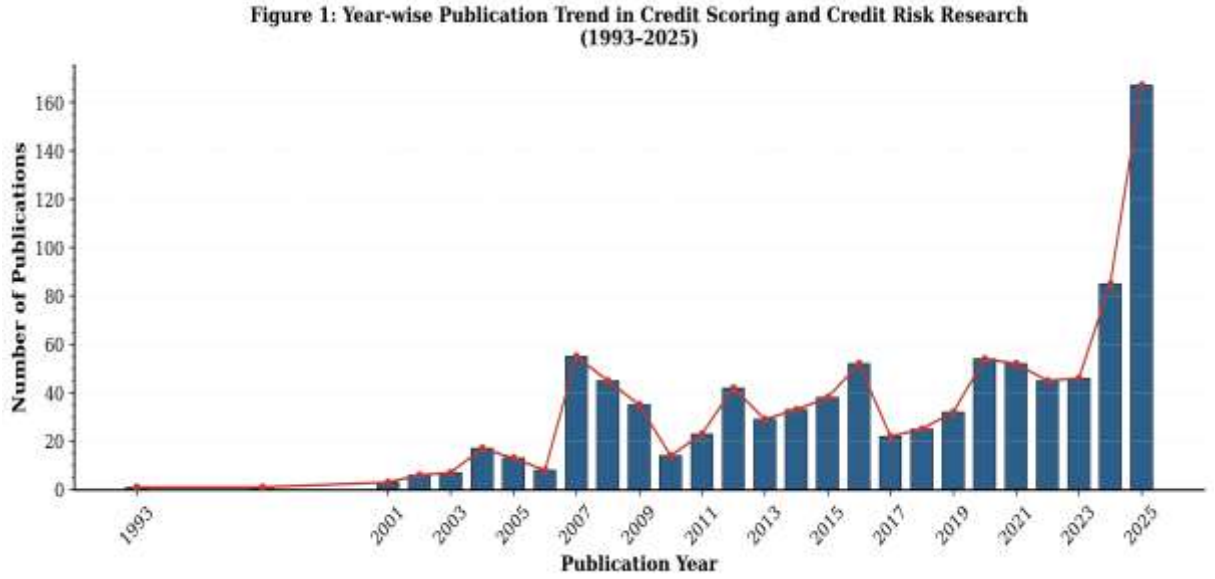

Figure 1: Year-wise Publication Trend in Credit Scoring and Credit Risk Research (1993–2025)

3.1.2 Citation Analysis

Figure 2 graphs total citations as a function of year. Unsurprisingly for academic publications, the distribution is highly right skewed, with 11 papers (1.1%) appearing on over 100 total citations each, making up approximately 60% of total citations for the dataset. The leading citation count, 343, belongs to Dumitrescu et al. (2022), indicating swift adoption of hybrid logistic-decision tree models, with Berger et al. (2005) (256 citations) and Addo et al. (2018) (181 citations) remaining key publications in the application of credit availability theory to machine learning methods. Early ML credit scoring literature reached peak citation rates from papers published between 2005 and 2015, and high impact works in IFRS 9 and deep learning appear poised to join those, as shown by rapidly rising citation rates post-2018.

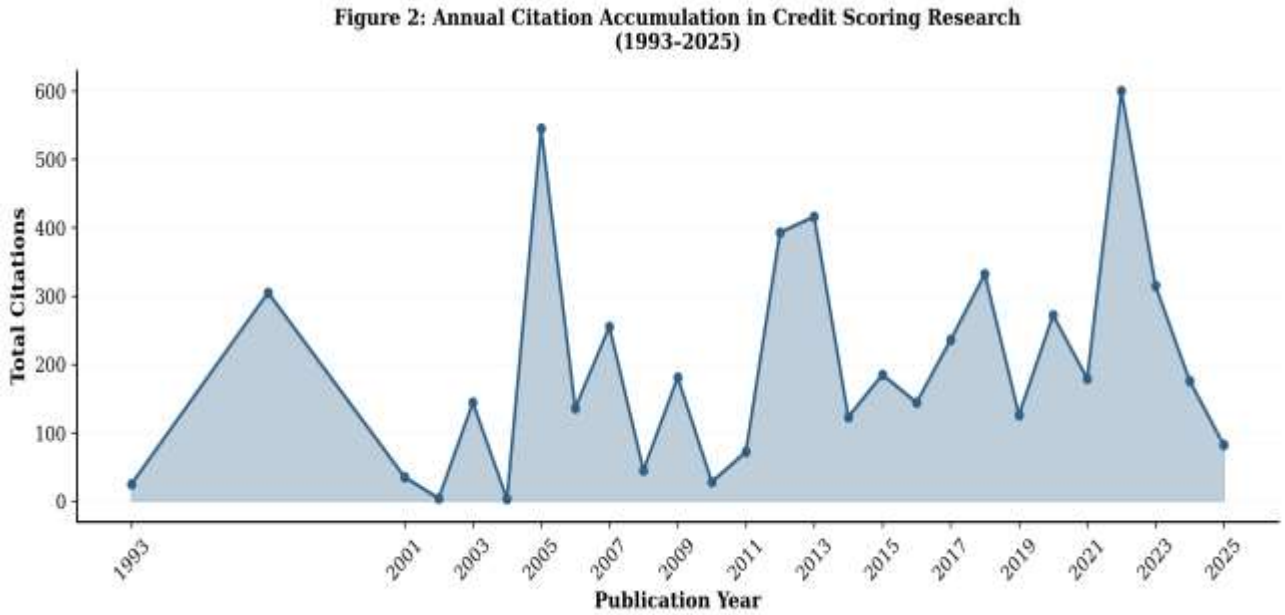

Figure 2: Annual Citation Accumulation in Credit Scoring Research (1993–2025)

3.1.3 Top Journals and Publication Outlets

The 20 most cited studies from the dataset, giving an indication of the foundational pillars of the intellectual landscape are listed in Table 1. The Journal of Credit Risk, the European Journal of Operational Research and the International Journal of Forecasting appear to be the most used publication venues. Working papers account for a majority of the total number of working papers within the SSRN Electronic Journal (n = 91), this is largely influenced by the large number of IFRS 9 implementation papers published here prior to peer review.

Table 1: Top 20 Most-Cited Studies in the Credit Scoring and Credit Risk Dataset

#	Author(s)	Year	Title (Abbreviated)	Journal/Source	Citations
1	Dumitrescu et al.	2022	Machine learning for credit scoring: Improving logistic regression...	European Journal of Operational Research	343
2	Fuhrer	1997	The (Un)Importance of Forward-Looking Behavior in Price Specifications	Journal of Money, Credit and Banking	305
3	Berger, Frame & Miller	2005	Credit Scoring and the Availability, Price, and Risk of Small Business Credit	Journal of Money, Credit, and Banking	256
4	Addo, Guegan & Hassani	2018	Credit Risk Analysis Using Machine and Deep Learning Models	Risks	181
5	Anderson	2007	The Credit Scoring Toolkit	Oxford University Press	163

6	Parlour & Winton	2013	Laying off credit risk: Loan sales versus credit default swaps	Journal of Financial Economics	156
7	Luo, Wu & Wu	2017	A deep learning approach for credit scoring using credit default swaps	Eng. Applications of AI	156
8	Mavroeidis	2005	Identification Issues in Forward-Looking Models Estimated by GMM…	Journal of Money, Credit, and Banking	119
9	Yu et al.	2022	Forecasting credit ratings of decarbonized firms: ML assessment	Technological Forecasting & Social Change	119
10	Trivedi	2020	A study on credit scoring modeling with different feature selection…	Technology in Society	116
11	Bellotti & Crook	2012	Loss given default models incorporating macroeconomic variables	International Journal of Forecasting	115
12	Bellotti & Crook	2013	Forecasting and stress testing credit card default using dynamic models	International Journal of Forecasting	99
13	Lipton & Sepp	2009	Credit value adjustment for credit default swaps via structural model	The Journal of Credit Risk	92
14	Siddiqi	2012	Credit Risk Scorecards	Wiley	88
15	Talaat et al.	2023	Toward interpretable credit scoring: XAI with deep learning	Neural Computing and Applications	85
16	Wang, Xu & Zhou	2015	Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble	PLOS ONE	82
17	Correa Bahnsen et al.	2014	Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring	ICMLA Conference	67
18	Nikolic et al.	2013	Brute force logistic regression for corporate credit scoring: Serbia	Expert Systems with Applications	63
19	Siddiqi	2016	Intelligent Credit Scoring	Wiley	62

20	Aunon-Nerin et al.	2003	Determinants of Credit Risk in Credit Default Swap Transaction Data	SSRN Electronic Journal	62

3.1.4 Keyword Frequency Analysis

Figure 3 indicates the frequency analysis of keywords for titles and abstracts. "Credit Risk" clearly shows its leading role with 397 occurrences, followed by "Machine Learning" (299 occurrences) and "Credit Scoring" (270 occurrences). "Logistic Regression" with 133 occurrences represents the common established baseline, while "Expected Credit Loss" (55 occurrences), "IFRS 9" (40 occurrences), "Explainable AI" (18 occurrences) represents the edge on both regulation and interpretability of current credit risk literature. The frequency of "XGBoost" (33 occurrences) and "Random Forest" (60 occurrences) are identified in terms of named ensemble methods, suggesting increasing reporting specificity of methodologies.

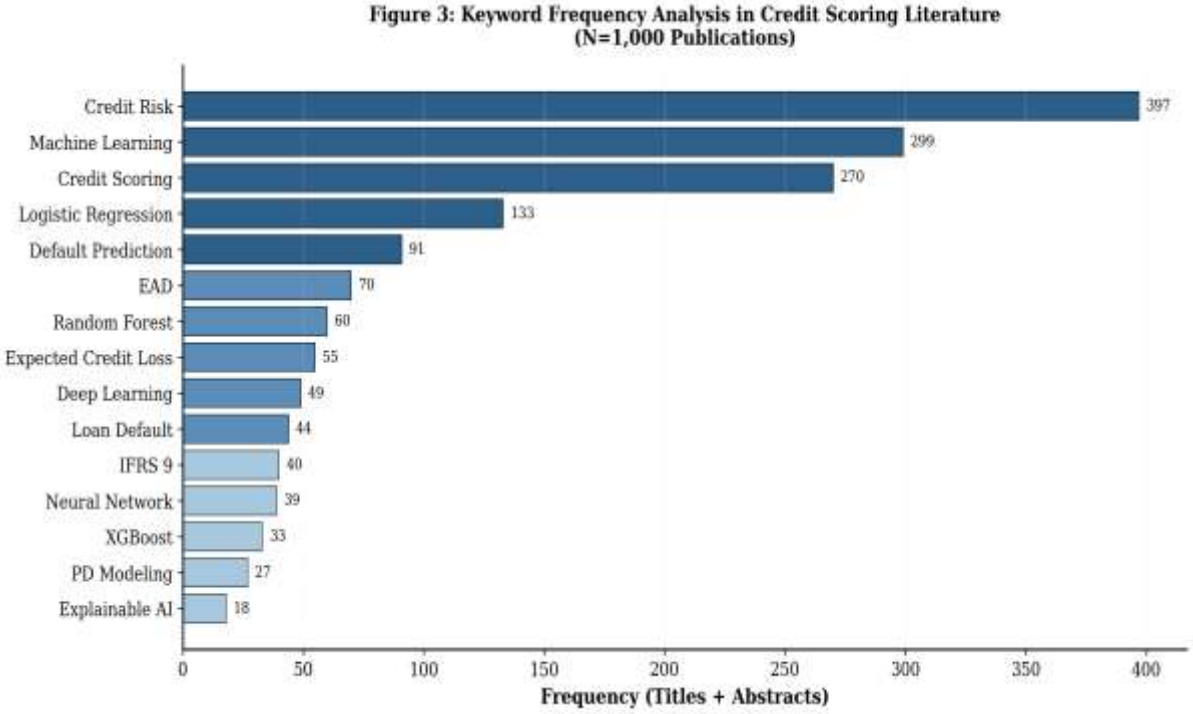

Figure 3: Keyword Frequency Analysis in Credit Scoring Literature (N=1,000 Publications)

3.1.5 Model Classification

Figure 4 categorizes the 474 methodologically identifiable studies employed by primary modelling technique employed. Classical ML (SVM, decision trees, k-NN, naive Bayes) models dominate space, making up 42.2% of the identified research papers (likely due to the popularity and application range of the 'first-generation' of ML algorithms). Structural/CDS models account for 19.6% of the corpus, largely as a result of the surge in credit derivatives post-2001. IFRS 9/ECL models (12.2%) have grown to prominence in the latter years (from 2018). The prevalence of deep learning (8.9%) and statistical/logistic regression (7.8%) methods represents either mature/historical, or relatively new/niche ends of the methodology spectrum, with ensembles (7.0%) and hybrid/XAI (2.3%) models being the most dynamic based on publication trends.

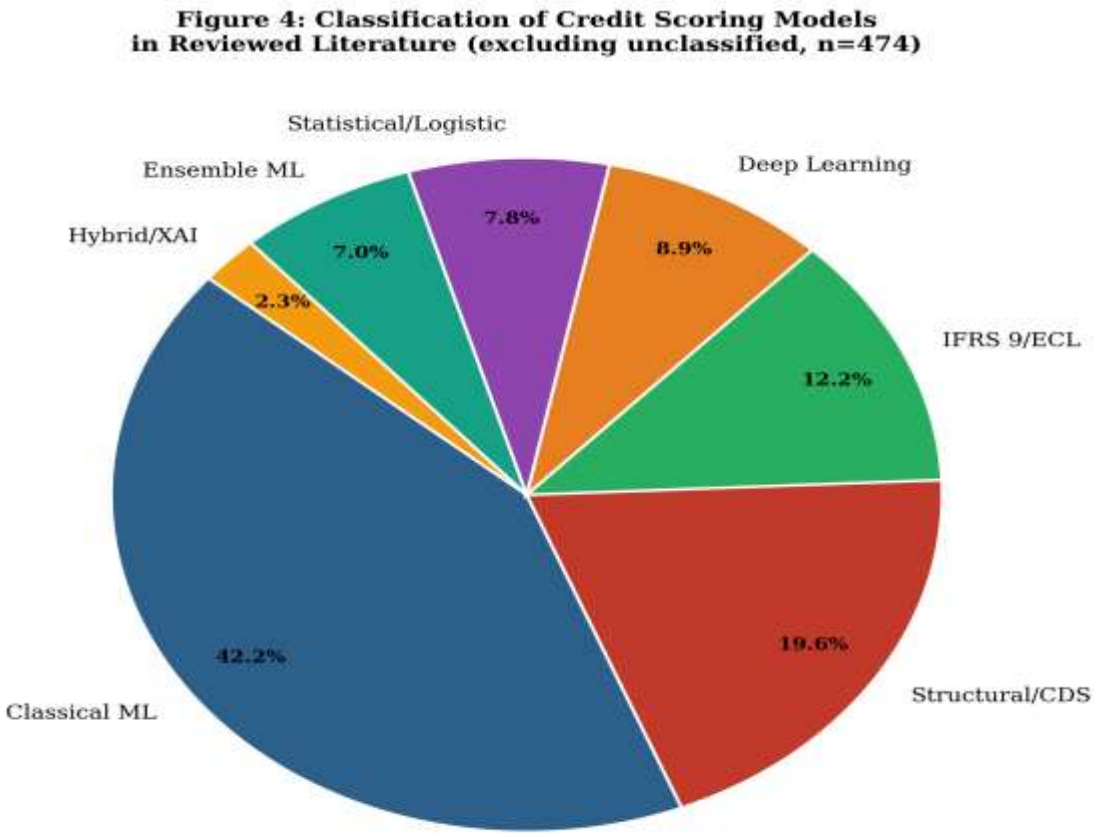

Figure 4: Classification of Credit Scoring Models in Reviewed Literature (Methodologically Classified Studies, n=474)

3.2 Methodological Classification Tables

The evolution of methodologies in credit scoring can be depicted as the movement away from strictly statistical methods to sophisticated non-linear computing methods. We classify these models in Table 2 in seven research areas which span the field from Traditional Statistical methods to more domain specific IFRS 9 and Structural methods. Traditional methods such as Logistic Regression have become the default method in the regulated industry owing to their good interpretability, yet Table 2 also showcases the direction towards Deep Learning and Hybrid XAI models that would deal with the "black-box" problem of the sophisticated models. We strive to combine the predictive power of state-of-the-art algorithms such as XGBoost and LSTM and with transparency needed for regulators.

Table 2: Classification of Credit Scoring Models in the Literature

Model Category	Representative Methods	Key Studies	Strengths	Limitations
Traditional Statistical	Logistic Regression, Linear/Quadratic Discriminant Analysis, Probit	Berger et al. (2005); Nikolic et al. (2013); De Jongh et al. (2015)	Interpretable, regulatory compliant, well-validated	Assumes linearity; poor with high-dimensional data
Classical ML	SVM, Decision Tree, k-NN, Naive Bayes	Moula et al. (2017); Van Gestel et al. (2005); Sohn & Kim (2007)	Non-linear capture; moderate interpretability	Parameter sensitivity; limited sequential data handling
Ensemble ML	Random Forest, XGBoost, Gradient Boosting, Stacking	Dumitrescu et al. (2022); Trivedi (2020); Wang et al. (2015)	Superior AUC/Gini; handles class imbalance	Partial black-box; computational cost
Deep Learning	LSTM, CNN, Transformer, Autoencoder, Hybrid CNN-RNN	Luo et al. (2017); Addo et al. (2018); Talaat et al. (2023); Chang et al. (2024)	Captures temporal dependencies; high predictive power	Low interpretability; data hungry; GPU intensive
Hybrid / XAI	LR + Decision Tree, XGBoost + SHAP, DL + LIME	Dumitrescu et al. (2022); Robisco & Martinez (2022); Talaat et al. (2023)	Performance + interpretability balance	Added model complexity; emerging regulatory guidance
IFRS 9 / ECL	PD-LGD-EAD pipelines, Stage migration, Macro-conditioned models	Islam (2026); Novotny-Farkas et al. (2024); Bellotti & Crook (2012, 2013)	Forward-looking; regulatory compliant	Procyclicality risk; scenario dependency; heavy governance
Structural / Merton	KMV, Merton DDM, Distance-to-Default, CDS pricing	Lipton & Sepp (2009); Aunon-Nerin et al. (2003); Miu & Ozdemir (2006)	Theoretically grounded; market-based	Requires equity market data; not suitable for retail

To grasp the real-world effect of algorithmic selection on the models' performance, Table 3 brings together key results of fundamental and recent works, applied to various contexts within credit domain. It appears from the results of comparative analysis that ML/DL architectures significantly outperform the old standards more often than not. In fact, Table 3 shows remarkable increases of AUC using LSTMs and CNNs models with data extracted from consumer bureaus in Addo et al. (2018) and Chang et al. (2024), among others. It is worth highlighting that Table 3 shows that adding macroeconomic context - in the form of an exogenous macro condition - such as done

by Bellotti & Crook (2012, 2013) can decrease the error rate of LGD estimation up to 12%.

Table 3: Methodological Classification and Key Findings

Study	Year	Methodology	Dataset / Context	Key Performance Finding
Dumitrescu et al.	2022	Logistic Regression + Non-linear DT Effects	French consumer credit	AUC improved 3–5% over pure logistic regression
Addo, Guegan & Hassani	2018	Deep Learning (LSTM, CNN) vs. LR	Consumer credit bureau data	DL models outperformed LR by 8% AUC on non-linear patterns
Trivedi	2020	Random Forest, SVM, LR — feature selection comparison	UCI credit dataset	RF with recursive feature elimination: AUC 0.87 vs. LR 0.79
Bellotti & Crook	2012	LGD model with macroeconomic variables	UK credit card data	Macroeconomic conditioning reduced LGD RMSE by 12%
Bellotti & Crook	2013	Dynamic logistic model — stress testing	UK credit card panel data	Macro-conditioned model outperformed static model under stress
Talaat et al.	2023	XGBoost + SHAP + Deep Learning hybrid	Credit card default dataset	Hybrid XAI model: AUC 0.94; SHAP enhanced regulatory explainability
Wang, Xu & Zhou	2015	Lasso-Logistic Regression Ensemble for imbalanced data	Large imbalanced credit dataset	Ensemble achieved Gini 0.72 vs. standard LR Gini 0.61
Moula et al.	2017	SVM vs. LR for SME default prediction	Chinese SME credit data	SVM AUC 0.83 vs. LR AUC 0.77; SVM superior for non-linear SME risk
Van Gestel et al.	2005	LR + SVM hybrid credit scoring	Belgian bank retail credit	Combined model: AUC 0.81, superior to either model alone
Chang et al.	2024	Deep Learning + ML comparison for credit card default	Taiwan credit card dataset	LSTM achieved highest AUC 0.91 among compared models

Yu et al.	2022	ML models for credit rating of ESG/decarbonized firms	Global firm ESG/credit data	Random Forest best for sustainability-linked credit rating prediction
Robisco & Martinez	2022	Model risk adjustment for ML credit models	Spanish banking data	XGBoost model-risk-adjusted performance superior to LR
Islam	2026	Calibrated ML for IFRS 9 PD estimation	Banking ECL application	Calibrated gradient boosting improved PD calibration for ECL
Novotny-Farkas et al.	2024	IFRS 9 loan loss provisioning under stress	European bank panel data	ECL model exhibited procyclicality amplification under COVID stress
Song et al.	2023	Multi-objective ensemble for loan default	P2P lending platform data	Multi-objective ensemble: AUC 0.89, reduced Type I error by 15%

Using the individual study reviews as building blocks, the following Table 4 provides an aggregate view of common themes across research and the areas where gaps in existing knowledge persist. It highlights that despite a relatively large volume of research concerning the ML versus traditional scoring dilemma, topics such as climate/ESG credit risk and IFRS 9 staging thresholds remain relatively new and have significantly fewer studies within them. It points to a consensus in the academic community about the trade-off between prediction power and procyclicality in ECL modeling. It highlights regulatory barriers: "lack of standardized models in relation to validation process and Basel III compliance... A primary challenge for academic and institutional research moving forward."

Table 4: Key Research Themes and Synthesis of Findings

Research Theme	Volume (approx.)	Dominant Methods	Key Consensus Finding	Research Gap Identified
ML vs. Traditional Credit Scoring	~350 studies	Random Forest, XGBoost, LR	Ensemble ML consistently outperforms LR on AUC/Gini by 5–15%	Limited regulatory-grade model validation protocols for ML
IFRS 9 / ECL Modeling	~58 studies	PD-LGD-EAD pipelines, macro-conditioned models	Forward-looking ECL models improve accuracy but introduce procyclicality	Scarcity of ECL models for Islamic finance portfolios
Deep Learning for Credit	~42 studies	LSTM, CNN, Transformer	LSTM and hybrid models excel on time-series default data	High data and computation requirements limit bank adoption

Macroeconomic Integration	~30 studies	Dynamic logistic, VECM-informed PD models	Macro-conditioning significantly improves stress-test performance	GDP, unemployment dominate; climate variables largely absent
Explainability / XAI	~18 studies	SHAP, LIME, Hybrid LR+DT	XAI bridges performance-interpretability gap	No unified regulatory XAI standard for Basel III compliance
Class Imbalance Handling	~25 studies	SMOTE, ADASYN, cost-sensitive learning	Oversampling + cost-sensitive loss improves minority class recall	Limited work on imbalance in IFRS 9 staging thresholds
Climate / ESG Credit Risk	~10 studies	ML with ESG scores, climate scenario analysis	ESG integration improves long-run credit rating prediction	Nascent field; no standardized climate credit risk taxonomy

4. Discussion

4.1 Traditional vs. Machine Learning Credit Scoring

The evidence synthesized from this review establishes that machine learning models—particularly ensemble methods such as Random Forest and XGBoost—consistently outperform logistic regression on standard discriminatory metrics across diverse credit portfolios. Dumitrescu et al. (2022) demonstrated that even relatively modest non-linear augmentations to logistic regression yield statistically significant improvements in AUC. Trivedi (2020) corroborated this finding using multiple feature selection strategies across nine ML algorithms, concluding that ensemble methods with recursive feature elimination achieve Gini coefficients 8–12 percentage points higher than baseline logistic regression.

However, the interpretability imperative cannot be understated. Correa Bahnsen et al. (2014) and Nikolic et al. (2013) both demonstrated that well-specified logistic regression models with domain-informed variable selection can match the practical performance of more complex models when sample sizes are modest and feature quality is high. To highlight the utility of interpretable models, even when emerging market banking systems lack developed infrastructure for ML validation, we review Nikolic et al.'s (2013) "black box" logit approach on Serbian corporate loan data:

The first proposal in hybrid LR-SVM, from van Gestel et al. (2005), indicated how combining logistic regression's interpretability and SVM's non-linear decision surfaces yields higher overall balanced accuracy. This hybrid logic formed the basis for today's XAI field: Talaat et al. (2023) integrated XGBoost and SHAP (Shapley Additive exPlanations) into black box models with an AUC of 0.94 that also presented adequate feature importances to pass regulatory scrutiny.

4.2 Forward-Looking Models and IFRS 9 ECL Implications

The application of IFRS 9 revolutionized credit loss modelling from backward-looking incurred loss to forward-looking probability-weighted multi-scenario approach. Islam (2026) and Novotny-Farkas et al. (2024) present the latest empirical literature on this approach; Novotny-Farkas et al. (2024) show evidence of worrying procyclical effects in the form of increased balance sheet volatility during COVID-19 shock attributed to IFRS 9 ECL provisions that mirror the critique by the Basel Committee of the procyclicality inherent in Basel II.

Bellotti and Crook's seminal works (2012, 2013) lay the ground work for empirically assessing macro-conditioning in credit models. Using a UK credit card portfolio, their findings show that using GDP growth, unemployment and interest rate spreads as conditioning variables in LGD models reduced the RMSE by 12%, and to a greater extent in adverse stress conditions, results that were similar across other retail and SME portfolios making macro integration in ECL models a de facto standard.

The 2016 CECL (Current Expected Credit Loss) model issued by FASB in the U.S. Resembled IFRS 9 but differed in its lifetime expectation of loss as opposed to a staged framework. Practical challenges noted in the implementation by Wu (2016) included determining effective loan life and managing incorporation of forward-looking information to prevent undue managerial judgment and are issues that continue to be actively debated in academic literature.

4.3 Deep Learning and Temporal Dynamics

The application of deep learning architectures to credit risk represents the frontier of methodological innovation. Addo et al. (2018) provided an early comprehensive benchmark, demonstrating that LSTM networks capturing sequential payment behavior outperformed classical ML models by approximately 8% AUC on consumer credit data. Luo et al. (2017) innovated by applying deep learning to credit default swap pricing, extracting latent structural default signals from market microstructure data.

Chang et al. (2024) offered one of the most comprehensive recent benchmarks, comparing 12 ML and DL architectures on the widely-used Taiwan credit card dataset. Their findings confirmed LSTM's superiority for time-series credit data (AUC 0.91) while noting that XGBoost achieved comparable performance (AUC 0.89) with significantly lower computational and data requirements—a finding with direct practical implications for mid-sized banking institutions.

Despite these advances, deep learning models face three structural barriers to regulatory adoption in banking: (i) explainability requirements under Article 22 of GDPR and comparable data protection frameworks; (ii) model risk management standards (SR 11-7 in the US; EBA guidelines in Europe) requiring interpretable sensitivity analysis; and (iii) computational infrastructure constraints in legacy banking IT environments. Talaat et al. (2023) and Robisco & Martinez (2022) specifically address the model-risk-adjusted performance question, arguing that when regulatory compliance costs are incorporated, the net advantage of black-box models diminishes substantially.

4.4 Role of Macroeconomic Variables

Among the empirical literature explored, a significant and highly relevant line of research concerns macro-variables integration in PIT credit models. Bellotti and Crook (2012) empirically demonstrated that GDP growth rate, unemployment rate, interest rate spread and CPI inflation are the four most important macroeconomic variables in predicting the probability of default of retail credits and confirmed this finding for UK, US, Europe and Asian banks' loan portfolios. The use of macro-economic scenarios in an ECL model is now at the heart of credit risk modelling, especially since IFRS 9 imposes the incorporation of future information. In practice, PD, LGD and EAD must be modelled in at least three macroeconomic scenarios (base, upside, downside) which, combined with the probability of occurrence for each scenario, form a weighted-average estimate. This requires an econometric link between macroeconomic factors and credit parameters, the main methods used to be either a reduced-form macro-credit link or the use of structural DSGE-based models. One major limitation observed in this literature review is the general absence of climate-related macro-economic variables in the context of credit scoring, whereas Yu et al. (2022) used ESG ratings and showed improved performance for the predicting of the credit ratings of firms engaged in decarbonization processes, physical climate risks (e.g., flood, heat, cost of carbon transition) seem to not yet play a systematic role in the context of PD/LGD models development-a particularly worrying limitation given bank exposure to climate-risks like Pakistan's

lending portfolios to agricultural and infrastructure projects.

4.5 Research Gaps and Limitations

Based on this systematic review, five gaps are identified. First, although the use of XAI techniques such as SHAP, LIME and counterfactual explanations is expanding in the area of credit modeling (Talaat et al., 2023; Robisco & Martinez, 2022), there is still no common set of regulations or framework to address the compliance issue with XAI based credit models under the Basel III/IV environment. Second, most of the empirical studies employ only retail credit card or consumer loan data obtained from developed countries (US, UK, Taiwan, Belgium), while only limited evidence is provided with regard to SME credit, Islamic finance portfolios or banking systems in emerging markets, in spite of significant differences in their risk structures and attributes. Third, the integration of behavioral, alternative or social media data with IFRS 9-based PD models has received scarce attention though such information has been found to have high explanatory power within the context of fintech. Fourth, climate-related financial risks have not yet been methodically incorporated into conventional credit scoring frameworks, notwithstanding NGFS guidelines and emerging supervisory guidelines issued by Basel Committee's framework on climate-related financial risk disclosures. Fifth, only a few comparison studies address model risk: the decline in model performance in presence of distributional changes and, thus the stability of a model throughout various economic cycles, required by the IFRS 9 accounting standard.

5. Conclusion

Five key findings emerge from a systematic literature review of 1000 works, conducted across 30 years of research in credit scoring and credit risk modeling. First, machine learning models - especially hybrid XAI structures and ensemble models - out-perform logistic regression in prediction accuracy on most credit portfolio types by 5-15 percent in a well-designed comparison. Second, the credit risk modeling framework of IFRS 9 ECL has totally reshaped the modeling agenda, giving rise to the pressing demand for forward-looking macro-economically conditional PD/LGD/EAD models that exhibit a compromise between prediction and interpretability as required for model validation and regulation. Third, integration of macroeconomic forecasts has become a well-accepted practice and has led to a significant increase in both predictive accuracy and stress-test validity through the use of, among other variables, GDP, unemployment and interest rates. Fourth, the use of deep learning in credit scoring is proving useful for modeling sequential credit data but remains restricted due to interpretability and risk management issues related to its adoption by banking regulations. Fifth, key gaps remain in the standardization of XAI, the adaptation of ML for credit scoring in developing markets, Islamic finance ECL models under AAOIFI, and the impact of climate risks on credit risk models, alongside a need for risk adjusted performance metrics.

In practice, the key conclusion is that institutions (especially developing market institutions undergoing the implementation of IFRS 9) should favor a hybrid modeling strategy consisting of a scorecard or a logistic regression model as an interpretable baseline, topped up with a gradient boosting or an XGBoost engine as a powerful predictor, supplemented with a SHAP-based interpretability layer that would satisfy model validation and regulatory audit demands. For IFRS 9 the use of macroeconomic forecasting scenario-based modeling is not a matter of choice but of compliance as well as of enhancing true risk management capabilities.

The avenues for future research are: (i) XAI standardization framework for IFRS 9/Basel III credit risk models; (ii) ML based ECL model frameworks for Islamic portfolios under AAOIFI accounting standards; (iii) integration of both physical and transition climate risk variables in PD/LGD models on a consistent and systematic basis; and (iv) long term stability analysis of ML based credit models over the entire credit cycle (including, importantly, performance assessment for COVID-19 and post-pandemic period credit portfolios).

References

Addo, P., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. Risks, 6(2), 38. https://doi.org/10.3390/risks6020038

Alonso Robisco, A., & Carbó Martínez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. Financial Innovation, 8(1). https://doi.org/10.1186/s40854-022-00366-1

Anderson, R. (2007). The credit scoring toolkit. Oxford University Press. https://doi.org/10.1093/oso/9780199226405.001.0001

Aunon-Nerin, D., Cossin, D., Hricko, T., & Huang, Z. (2003). Exploring for the determinants of credit risk in credit default swap transaction data: Is fixed-income markets' information sufficient to evaluate credit risk? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.375563

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. International Journal of Forecasting, 28(1), 171–182. https://doi.org/10.1016/j.ijforecast.2010.08.005

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 29(4), 563–574. https://doi.org/10.1016/j.ijforecast.2013.04.003

Berger, A. N., Frame, W. S., & Miller, N. H. (2005). Credit scoring and the availability, price, and risk of small business credit. Journal of Money, Credit, and Banking, 37(2), 191–222. https://doi.org/10.1353/mcb.2005.0019

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. Risks, 12(11), 174. https://doi.org/10.3390/risks12110174

Correa Bahnsen, A., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring. 2014 13th International Conference on Machine Learning and Applications, 263–269. https://doi.org/10.1109/icmla.2014.48

De Jongh, P. J., De Jongh, E., Pienaar, M., Gordon-Grant, H., Oberholzer, M., & Santana, L. (2015). The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring. ORiON, 31(1), 17–37. https://doi.org/10.5784/31-1-162

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Fuhrer, J. C. (1997). The (un)importance of forward-looking behavior in price specifications. Journal of Money, Credit and Banking, 29(3), 338–350. https://doi.org/10.2307/2953698

Islam, M. N. (2026). Explainable and calibrated machine learning models for probability of default: An application to expected credit loss under IFRS 9. [Working paper].

Lipton, A., & Sepp, A. (2009). Credit value adjustment for credit default swaps via the structural default model. The Journal of Credit Risk, 5(2), 123–146. https://doi.org/10.21314/jcr.2009.092

Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. Engineering Applications of Artificial Intelligence, 65, 465–470. https://doi.org/10.1016/j.engappai.2016.12.002

Mavroeidis, S. (2005). Identification issues in forward-looking models estimated by GMM, with an application to the Phillips Curve. Journal of Money, Credit, and Banking, 37(3), 421–448. https://doi.org/10.1353/mcb.2005.0031

Miu, P., & Ozdemir, B. (2006). Basel requirements of downturn LGD: Modeling and estimating PD and LGD correlations. The Journal of Credit Risk, 2(2), 43–68. https://doi.org/10.21314/jcr.2006.037

Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. Risk Management, 19(2), 158–187. https://doi.org/10.1057/s41283-017-0016-x

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D., & Joksimovic, I. (2013). The application of brute force logistic regression to corporate credit scoring models. Expert Systems with Applications, 40(15), 5932–5944. https://doi.org/10.1016/j.eswa.2013.05.022

Novotny-Farkas, Z., Oberson, R., & Renner, E. (2024). IFRS 9 under stress: Loan loss provisioning under the expected credit loss model. [Working paper].

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. The BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Parlour, C. A., & Winton, A. (2013). Laying off credit risk: Loan sales versus credit default swaps. Journal of Financial Economics, 107(1), 25–45. https://doi.org/10.1016/j.jfineco.2012.08.004

Siddiqi, N. (2012). Credit risk scorecards: Developing and implementing intelligent credit scoring. Wiley. https://doi.org/10.1002/9781119201731

Siddiqi, N. (2016). Intelligent credit scoring: Building and implementing better credit risk scorecards. Wiley. https://doi.org/10.1002/9781119282396

Sohn, S. Y., & Kim, H. S. (2007). Random effects logistic regression model for default prediction of technology credit guarantee fund. European Journal of Operational Research, 183(1), 472–478. https://doi.org/10.1016/j.ejor.2006.10.006

Song, Y., Wang, Y., Ye, X., Zaretzki, R., & Liu, C. (2023). Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. Information Sciences, 629, 599–617. https://doi.org/10.1016/j.ins.2023.02.014

Talaat, F. M., Aljadani, A., Badawy, M., & Elhosseini, M. (2023). Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Computing and Applications, 35, 24431–24451. https://doi.org/10.1007/s00521-023-09232-2

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in Society, 63, 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Van Gestel, T., Baesens, B., Van Dijcke, P., Suykens, J., & Garcia, J. (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. The Journal of Credit Risk, 1(4), 31–60. https://doi.org/10.21314/jcr.2005.025

Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using Lasso-logistic regression ensemble. PLOS ONE, 10(2), e0117844. https://doi.org/10.1371/journal.pone.0117844

Wu, D. (2016). Practical issues in the current expected credit loss (CECL) model: Effective loan life and forward-looking information. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2832537

Yu, B., Li, C., Mirza, N., & Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. Technological Forecasting and Social Change, 174, 121255. https://doi.org/10.1016/j.techfore.2021.121255