Explainable Credit Default Prediction Using a Hybrid LSTM–XGBoost Model with SHAP Interpretability

Nontethelelo Mbanjwa * and Thabo Lephoto

Posted Date: 5 May 2026

doi: 10.20944/preprints202605.0173.v1

Keywords: credit default prediction; LSTM–XGBoost; SHAP; hybrid machine learning; explainable AI

 doi:10.20944/preprints202605.0173.v1

Article

Explainable Credit Default Prediction Using a Hybrid LSTM–XGBoost Model with SHAP Interpretability

Nontethelelo Mbanjwa * and Thabo Lephoto

School of Agriculture and Science, Discipline of Statistics, University of KwaZulu-Natal, Pietermaritzburg Campus, Private Bag X01, Scottsville 3209, South Africa

* Correspondence: nmbanjwa249@gmail.com

Abstract

The accurate prediction of credit default risk remains a significant challenge for financial institutions operating within increasingly complex data environments. This study proposes a hybrid Long Short-Term Memory (LSTM) and eXtreme Gradient Boosting (XGBoost) model that integrates deep learning and ensemble machine learning techniques to enhance predictive performance while preserving interpretability. The LSTM component effectively captures temporal patterns in borrower behavior, and its output is utilized as a meta-feature within the XGBoost framework. The model is evaluated using a benchmark credit dataset and is compared with conventional machine learning approaches. The results indicate that the proposed hybrid model outperforms standalone models across key evaluation metrics, achieving high accuracy, F1-score, and ROC–AUC. To enhance transparency, Shapley Additive Explanations (SHAP) are employed to analyse feature contributions and directional effects. The findings reveal that repayment behavior, particularly recent delinquency, serves as the most influential predictor of default risk, followed by indicators of financial capacity. The feature derived from the LSTM demonstrates the strongest overall impact, thereby confirming the significance of temporal dependencies in credit risk prediction. This study illustrates that the integration of deep learning with ensemble techniques establishes a robust and interpretable framework for credit risk assessment, thereby providing practical value for enhancing financial decision-making and risk management.

Keywords: credit default prediction; LSTM–XGBoost; SHAP; hybrid machine learning; explainable AI

1. Introduction

Credit risk assessment has emerged as a pivotal component within contemporary financial systems, enabling institutions to evaluate the likelihood of borrower default and to inform lending decisions (El-Qadi et al. 2022). Accurate prediction of credit defaults is imperative for sustaining financial stability, minimizing non-performing loans, and supporting judicious lending practices (Punukollu et al. 2022). Within the banking sector, effective credit risk management significantly impacts capital allocation, regulatory compliance, and institutional profitability. The increasing complexity of financial markets, coupled with a rising incidence of loan defaults, has heightened the demand for robust predictive models (Perera and Premaratne 2024). Financial institutions face substantial challenges in accurately distinguishing between low-risk and high-risk borrowers, as erroneous lending decisions can lead to financial instability and reduced profitability (Wang et al. 2026). Furthermore, the rapid digitalization of financial services and the proliferation of financial data sources have rendered traditional statistical methods inadequate for capturing the intricate relationships inherent in borrower data, revealing limitations in managing complex, high-

 doi:10.20944/preprints202605.0173.v1

dimensional, and non-linear data structures (Wang and Liang 2024). Consequently, with the swift evolution of financial technologies and the increasing accessibility of large-scale datasets, advanced machine learning and artificial intelligence techniques have emerged as powerful tools for enhancing predictive accuracy and decision-making within credit risk analysis.

Recent studies underscore the limitations of traditional models, such as logistic regression and discriminant analysis, particularly regarding their inability to adequately capture non-linear relationships and temporal dependencies in financial data. For example, Kandi and García-Dopico (2025) illustrated that traditional models encounter difficulties when applied to complex and high-dimensional datasets due to their reliance on linear assumptions, which results in inferior predictive performance compared to advanced machine learning techniques. In contrast, their study revealed that models such as eXtreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) networks achieved significantly higher accuracy, with LSTM reaching up to 99% accuracy in credit risk prediction tasks. Similarly, Wang et al. (2026) found that ensemble and hybrid machine learning models consistently outperform traditional single models by effectively reducing bias and variance, thereby enhancing prediction accuracy and profitability in loan default prediction. Furthermore, Wang et al. (2025) emphasized that while machine learning and deep learning models markedly improve predictive accuracy, they also present challenges related to interpretability, as many of these models function as "black boxes." Collectively, these findings indicate that although advanced models offer superior predictive performance, there remains a critical necessity to balance accuracy with interpretability in the context of credit risk assessment.

In addition to predictive accuracy, the issue of model interpretability has garnered increasing attention within the financial sector (Lin et al. 2022; Wang et al. 2025). The implementation of complex AI models frequently introduces a "black box" problem, wherein decision-making processes are not readily comprehensible to stakeholders (Hassija et al. 2024; Hoang et al. 2026). This lack of transparency raises concerns regarding regulatory compliance, fairness, and trust in automated decision systems (Wang and Liang 2024). To mitigate this challenge, explainable artificial intelligence (XAI) techniques, such as Shapley Additive Explanations (SHAP), have been widely adopted to elucidate feature importance and model behavior. These methodologies enhance transparency and empower financial institutions to better comprehend the factors influencing credit decisions (Nallakaruppan et al. 2024).

Recent research highlights the significance of integrating predictive performance with interpretability in credit risk modeling. For instance, Wang et al. (2026) proposed a fusion framework that combines ensemble learning techniques to mitigate bias and variance, thereby demonstrating improved prediction accuracy while simultaneously enhancing interpretability through feature analysis. Similarly, Ahya et al. (2024) developed a hybrid optimization and explainability-driven framework that incorporates advanced models such as Gradient Boosting and TabNet. Their findings indicate that the combination of predictive models with SHAP-based interpretability markedly enhances both classification performance and transparency in decision-making. Additionally, Zhang and Zhao (2026) employed an XGBoost-SHAP framework for predictive modeling and discovered that the integration of machine learning with interpretability techniques yields not only high predictive accuracy but also meaningful insights into the influence of both financial and non-financial factors. Collectively, these studies illustrate that hybrid, and ensemble approaches effectively balance the trade-off between accuracy and explainability by leveraging the strengths of multiple algorithms. Such methodologies not only enhance classification performance but also provide actionable insights for decision-makers, thereby supporting more effective and transparent risk management strategies.

Despite these advancements, challenges persist in developing models that can adeptly handle imbalanced datasets, capture both static and sequential features, and provide interpretable results. Furthermore, previous studies have explored the application of ensemble learning, deep learning, and explainable AI in credit risk modeling. Few studies integrated these approaches into a unified framework that simultaneously captures temporal dynamics, nonlinear relationships, and model interpretability. Many existing studies focus either on predictive accuracy or on explainability but

 doi:10.20944/preprints202605.0173.v1

rarely address both aspects within a single hybrid architecture (Li et al. 2022; Lin 2024; Sasikumar and Nareshkumar 2025). Furthermore, limited research has examined how interpretable hybrid AI models can support financial policy development and regulatory compliance, particularly in emerging financial ecosystems where responsible lending and risk transparency are becoming increasingly important.

To address these gaps, this study proposes a hybrid LSTM–XGBoost Model with SHAP Explainability for credit default prediction. The proposed framework integrates the strengths of deep learning and ensemble learning by combining the temporal modeling capability of LSTM networks with the powerful nonlinear predictive performance of XGBoost. The LSTM component captures sequential borrower behavior patterns over time, while XGBoost enhances classification accuracy through gradient boosting optimization. To ensure model transparency and regulatory compliance, SHAP-based explainability techniques are incorporated to interpret model predictions and identify the key financial variables influencing credit default risk.

The main contributions of this study can be summarized as follows:

1. Development of a hybrid predictive framework that integrates LSTM and XGBoost to improve the accuracy of credit default prediction by capturing both temporal dependencies and nonlinear relationships within financial datasets.
2. Integration of explainable artificial intelligence (XAI) through SHAP analysis, enabling transparent interpretation of model predictions and identification of key determinants of borrower default risk.
3. Provision of policy-relevant insights for financial institutions and regulators, demonstrating how interpretable hybrid AI models can support responsible lending decisions, improve risk governance, and enhance compliance with regulatory requirements related to algorithmic transparency.
4. Empirical evaluation of the proposed model using credit risk datasets to demonstrate improvements in predictive performance, interpretability, and decision support compared with conventional machine learning and deep learning approaches.

By bridging predictive accuracy with interpretability, the proposed hybrid framework contributes to the growing literature on explainable financial artificial intelligence and offers practical implications for banks, fintech institutions, and policymakers seeking more transparent and reliable credit risk assessment tools.

2. Materials and Methods

2.1. Dataset and Data Preprocessing

The study employs a structured credit dataset that encompasses customer financial and behavioral attributes pertinent to default prediction. This includes variables such as credit limits, repayment status, bill amounts, payment history, and demographic characteristics. Prior to the development of the predictive model, the dataset undergoes comprehensive preprocessing to ensure data quality and consistency. This preprocessing phase involves addressing missing values through appropriate imputation techniques, rectifying inconsistencies, and standardizing variable formats. Furthermore, the dataset is partitioned into training and testing subsets to enable reliable evaluation of the model. Additional procedures, such as data normalization and scaling, are implemented to ensure that variables with disparate units and magnitudes do not disproportionately affect the learning process, thereby enhancing both model stability and performance.

2.2. Feature Engineering

Feature selection is crucial for refining accuracy and interpretability of ML models, especially related to credit default predictors (Guo et al. 2023). Feature engineering is utilized to enhance the predictive capability of models by transforming raw financial and behavioral data into meaningful inputs. In this study, random forest was employed as an embedded feature selection method to

 doi:10.20944/preprints202605.0173.v1

identify the most relevant predictors based on feature importance scores. Its ability to capture nonlinear relationships and interactions makes it particularly suitable for financial datasets (Qiu and Wang 2025). To enhance robustness and reduce potential bias, the selected features were further validated using SHAP-based importance measures. Novel features are constructed to capture underlying customer behavior, including repayment trends, credit utilization patterns, and temporal payment dynamics. Relevant variables are meticulously selected and transformed using domain expertise to improve both model accuracy and interpretability. Furthermore, techniques such as scaling, encoding of categorical variables, and aggregation of behavioral indicators are implemented to enrich the dataset. These engineered features enable models, particularly LSTM networks and XGBoost, to effectively learn both sequential and static relationships within customer financial behavior, thereby enhancing the accuracy of default prediction.

2.3. Logistic Regression Model

Logistic Regression is a widely used statistical classification model for binary outcomes, such as credit default prediction. It estimates the probability of an event occurring by modelling the relationship between a set of independent variables and a binary dependent variable through a logistic (sigmoid) function (Hosmer et al. 2013). The model is defined as:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta^T x)\right)}, \tag{1}$$

where $P(y = 1 \mid x)$ represents the probability of default, x is the vector of input features, β denotes the regression coefficients, and β_0 is the intercept term. The model assumes a linear relationship between the predictors and the log-odds of the outcome, expressed as:

$$\log\left(\frac{P(y = 1 \mid x)}{1 - P(y = 1 \mid x)}\right) = \beta_0 + \beta^T x. \tag{2}$$

Model parameters are estimated using maximum likelihood estimation by minimising the logistic loss function. To prevent overfitting and improve generalisation, regularisation techniques such as *L1* (Lasso) or *L2* (Ridge) penalties can be incorporated. In this study, logistic regression serves as a baseline model, providing an interpretable benchmark for comparison with more complex machine learning and deep learning approaches.

2.4. Cox Proportional Hazards Model

The Cox Proportional Hazards (PH) model represents a semi-parametric methodology utilized for analyzing the relationship between explanatory variables and time-to-event outcomes. The hazard function is articulated as follows:

$$h(t \mid X) = h_0(t)\exp(\beta X), \tag{3}$$

where $h(t \mid X)$ denotes the hazard at time t conditional on covariates X, $h_0(t)$ represents an unspecified baseline hazard, and β signifies the regression coefficients. The estimation of model parameters is conducted via partial likelihood, facilitating inference without necessitating an explicit specification of the baseline hazard function.

A principal assumption underlying the model is that of proportional hazards, which posits that the effects of covariates are multiplicative and remain constant over time. The adequacy of the model is typically evaluated using the concordance index (C-index) and likelihood-based measures. Due to its interpretability and capacity to accommodate censored observations, also the Cox PH model serve as a standard benchmark in the field of time-to-event analysis.

2.4. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) proposed by Chen and Guestrin (2016), represents a highly efficient and scalable implementation of gradient boosting algorithms specifically tailored for

 doi:10.20944/preprints202605.0173.v1

supervised learning tasks. This methodology constructs an ensemble of decision trees in a sequential manner, wherein each newly incorporated tree endeavors to rectify the prediction errors of its predecessors by minimizing a differentiable loss function (Zheng 2022). Through this iterative learning process, the model progressively enhances its predictive accuracy by concentrating on observations that were previously misclassified. The notable strength of XGBoost lies in its capacity to capture intricate and nonlinear relationships among predictor variables, a characteristic often observed in financial datasets utilized for credit risk assessment (Yu 2025). The algorithm incorporates regularization techniques that regulate model complexity, thereby mitigating the risk of overfitting and enhancing generalization performance (Qin et al. 2021). Furthermore, XGBoost integrates several computational optimizations, including parallel processing, efficient management of missing values, and tree pruning strategies, rendering it exceptionally scalable for large datasets (Yu et al. 2024). Owing to its robust predictive performance and resilience, XGBoost has found extensive application in financial risk modeling, particularly in credit scoring and default prediction (Sharma and Amad 2022; Zhu et al. 2024). A detailed mathematical explanation can be found to Gao et al. (2021), Liu et al. (2022) and Ouyang (2024).

2.5. Long Short-Term Memory

Long Short-Term Memory (LSTM) networks were introduced by Hochreiter and Schmidhuber (1997) represent a specialized form of recurrent neural networks that are purposefully designed for modeling sequential and time-dependent data. In contrast to traditional machine learning models that regard observations as independent events, LSTM networks are adept at identifying patterns over time by retaining a memory of prior observations. This characteristic renders LSTM particularly effective for financial prediction problems, where historical behavior exerts a significant influence on future outcomes.

In financial datasets, variables such as repayment history, credit utilization, transaction patterns, and income stability exhibit temporal evolution. LSTM networks effectively capture these temporal dynamics by preserving pertinent historical information while filtering out less significant signals through an internal memory architecture governed by gating mechanisms. These gates regulate the processes of information storage, updating, and forgetting as the sequence advances, thereby enabling the model to learn long-term dependencies in financial behavior. In the context of credit default prediction, this capacity for temporal learning empowers the model to recognize subtle variations in borrower financial patterns that may indicate an increasing risk of credit default. By extracting latent sequential features from borrower histories, LSTM provides a more nuanced representation of financial behavior, which can substantially enhance predictive performance when integrated with complementary machine learning models within hybrid frameworks. A comprehensive mathematical formulation and detailed explanation of the LSTM architecture can be found in Liang and Cai (2020) and Gao et al. (2023).

2.6. DeepSurv Model

DeepSurv is a deep learning extension of the Cox proportional hazards model that captures nonlinear relationships between covariates and survival outcomes (Katzman et al. 2018). Unlike the traditional Cox model, which assumes a linear predictor, DeepSurv replaces this component with a neural network to model complex feature interactions. In the standard Cox model, the hazard function is defined as

$$h(t \mid x) = h_0(t)\exp(\beta^T x), \tag{4}$$

where $h_0(t)$ is the baseline hazard, x represents the covariates, and β denotes regression coefficients. DeepSurv generalises this formulation by replacing $\beta^T x$ with a nonlinear function $f_\theta(x)$, such that:

$$h(t \mid x) = h_0(t)\exp(f_\theta(x)). \tag{5}$$

 doi:10.20944/preprints202605.0173.v1

Model parameters are estimated by minimising the negative log partial likelihood:

$$L(\theta) = -\sum_{i:E_i=1}\left(f_\theta(x_i) - \log \sum_{j\in R_i} \exp\ (f_\theta(x_j))\right), \tag{6}$$

where E_i indicates event occurrence and R_i is the corresponding risk set. In this study (see Figure 1), input features are preprocessed through scaling and encoding before being fed into a fully connected neural network with nonlinear activation functions. The network outputs a continuous risk score, which is optimised using the Cox partial likelihood. The trained model produces individual risk estimates, enabling effective modelling of time-to-default in credit risk applications.

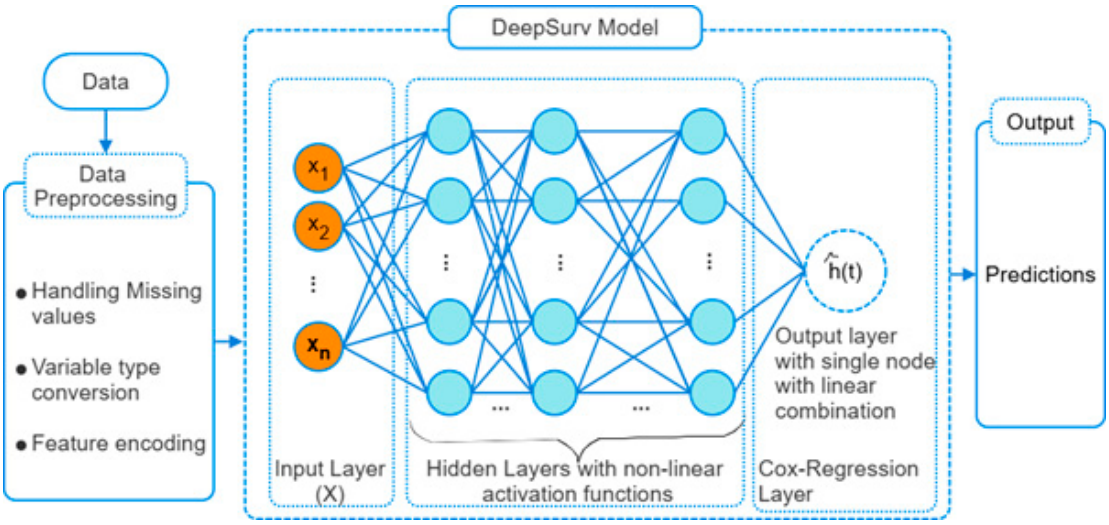

Figure 1. Architecture of the DeepSurv model showing data preprocessing, nonlinear deep neural network layers for risk representation, and a Cox regression output layer for hazard estimation and survival prediction.

2.7. Hybrid LSTM–XGBoost Framework with SHAP

The proposed hybrid LSTM and XGBoost framework, augmented by SHAP, is designed to concurrently capture temporal dynamics, nonlinear relationships, and model interpretability in the context of credit default prediction (see Figure 2). The integration of LSTM, XGBoost, and SHAP is driven by their complementary strengths. LSTM effectively captures temporal dependencies in sequential data, while XGBoost excels in modeling nonlinear relationships within structured datasets. However, both models exhibit a lack of inherent interpretability, which is addressed through the application of SHAP. By synthesizing these components, the proposed framework mitigates the limitations of single-model approaches, attaining enhanced predictive performance while preserving interpretability. The overall architecture adheres to a sequential pipeline, comprising data preprocessing, temporal feature extraction, feature integration, classification, and post hoc interpretation.

Initially, the dataset undergoes preprocessing to delineate sequential and static components. Sequential financial data, such as repayment history and transaction behavior, are structured into a time series format, while static attributes, including demographic and financial indicators, are standardized to ensure compatibility with the model. The sequential data are subsequently input into an LSTM network, which learns latent temporal representations through its memory cells. This mechanism enables the model to capture evolving borrower behavior and long-term dependencies that are not discernible in cross-sectional data.

The extracted temporal features are then amalgamated with static borrower attributes to create an enriched feature space. This integrated representation serves as input to the XGBoost classifier.

 doi:10.20944/preprints202605.0173.v1

XGBoost employs an ensemble of decision trees optimized through gradient boosting to model complex nonlinear interactions among features and produce robust credit default predictions.

To enhance transparency, SHAP is applied to the outputs of XGBoost. SHAP provides both global and local interpretability by quantifying the contribution of each feature to the prediction outcome. This facilitates the identification of key risk drivers at the population level, as well as detailed explanations for individual borrower predictions, thereby supporting accountable and transparent decision-making (Ahmad et al. 2024). The hybrid architecture offers a cohesive and efficient solution for credit risk modeling by concurrently addressing temporal complexity, predictive accuracy, and explainability. This is of particular significance in financial applications, where transparent and reliable models are imperative for regulatory compliance and informed decision-making.

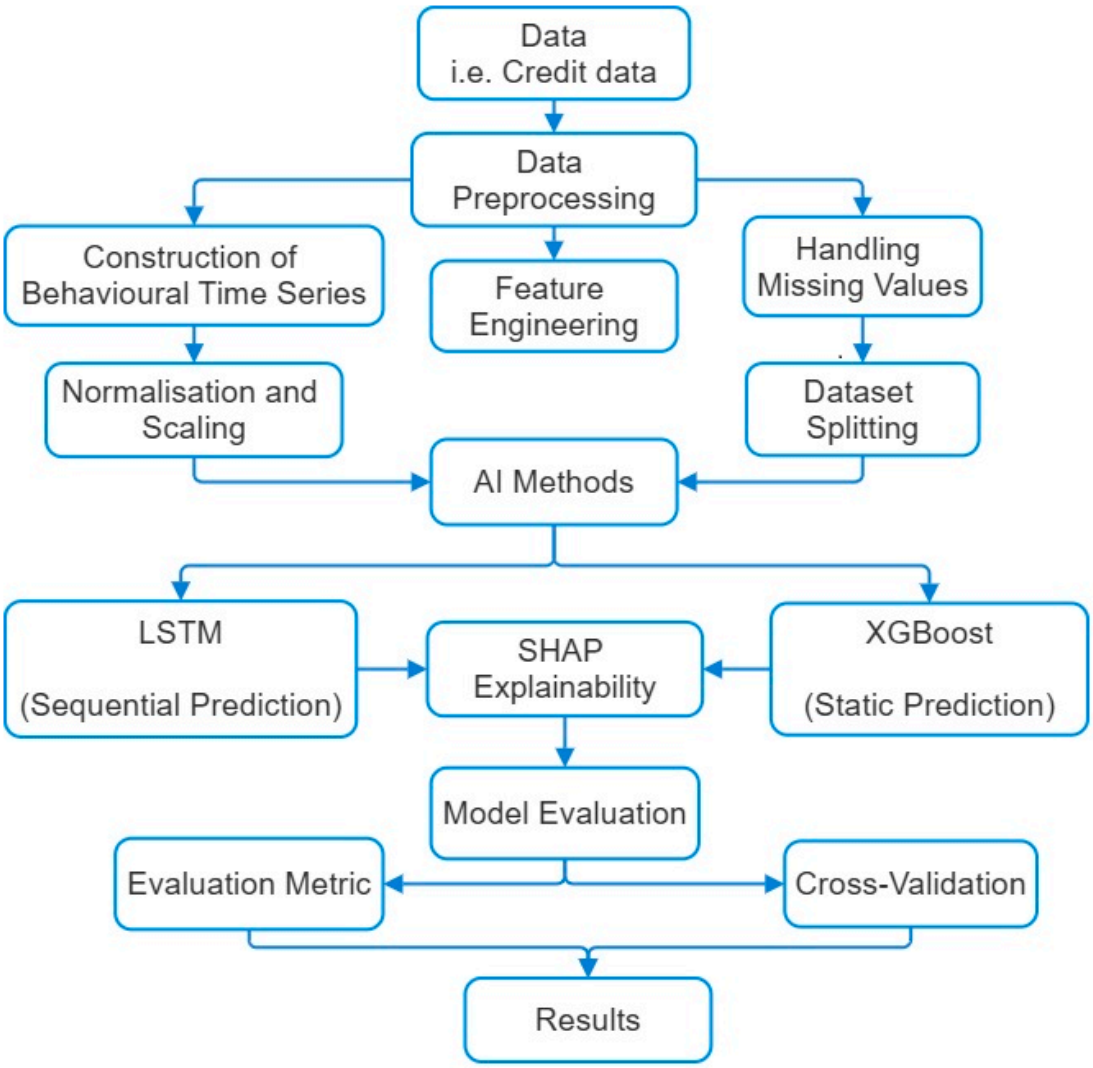

Figure 2. Integrated AI Framework for Credit Prediction.

2.8. Model Evaluation

The performance of the classification models is evaluated using several widely used metrics in credit risk prediction. These metrics provide a comprehensive assessment of the model's ability to correctly classify default and non-default borrowers. Precision measures the proportion of correctly predicted positive instances among all predicted positive instances:

 doi:10.20944/preprints202605.0173.v1

$$P = \frac{TP}{TP + FP} \tag{7}$$

Recall represents the proportion of correctly predicted positive instances among all actual positive instances:

$$R = \frac{TP}{TP + FN} \tag{8}$$

Accuracy measures the overall proportion of correctly classified instances among all predictions:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

The F1-score is the harmonic mean of Precision and Recall, providing a balanced evaluation particularly when the dataset is imbalanced:

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{10}$$

In addition to these metrics, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are used to evaluate the discriminative ability of the classification models. The ROC–AUC value ranges from 0 to 1, where a higher value indicates better model performance in distinguishing between default and non-default classes. A model with an ROC–AUC value closer to 1 demonstrates superior predictive capability.

3. Results

3.1. Exploratory Data Analysis (EDA) Interpretation

The exploratory data analysis provides key insights into borrower characteristics and repayment behaviour, highlighting patterns relevant to credit risk assessment. Figure 3 of the observed repayment trends reveal a distinct differentiation between defaulters and non-defaulters throughout the monitored months. Non-defaulters consistently exhibit negative average PAY status values, indicative of timely or early repayments. Conversely, defaulters demonstrate positive PAY status values, which, although they decline gradually over time, remain consistently inferior to those of non-defaulters. This downward trajectory for defaulters may suggest either partial recovery or delayed repayments; nonetheless, their repayment performance consistently lags behind that of non-defaulters. The enduring disparity between these two groups affirms that repayment history serves as a robust predictor of default risk.

 doi:10.20944/preprints202605.0173.v1

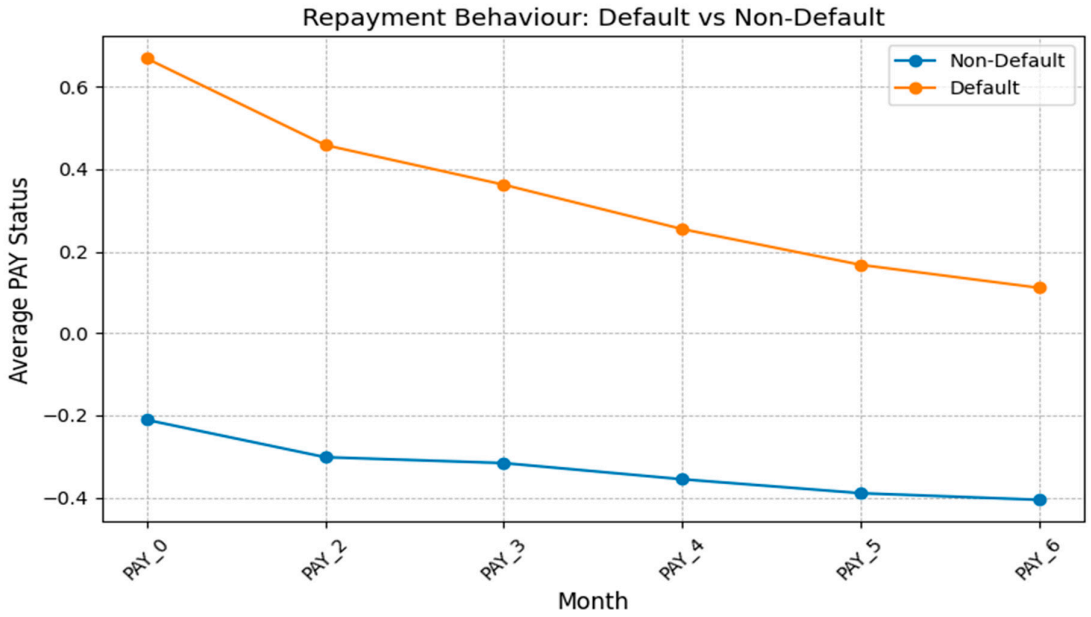

Figure 3. Repayment Behavior Over Time.

The age distribution of borrowers is characterized by a positive skew, with a significant concentration of individuals aged approximately 25 to 40 years (see Figure 4). Conversely, there is a relative scarcity of older borrowers, particularly those over the age of 60. This trend implies that the credit portfolio is predominantly comprised of younger to middle-aged individuals, potentially reflecting a heightened demand for credit within economically active demographics. The observed skewness suggests that age-related risk analyses should prioritize this predominant segment.

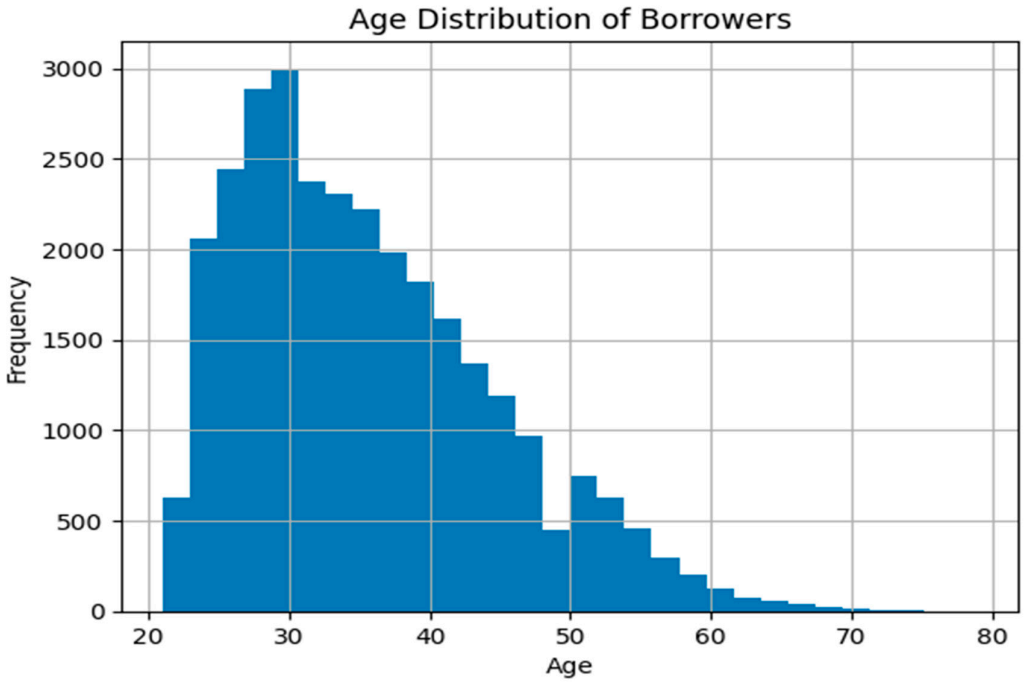

Figure 4. Age Distribution of Borrowers.

 doi:10.20944/preprints202605.0173.v1

Default rates exhibit considerable variation across different levels of education (see Figure 5). Borrowers possessing mid-level education (specifically levels 2 and 4) demonstrate the highest default rates, whereas those with higher education levels (for instance, level 3 and above) tend to experience lower default rates. This pattern implies that educational attainment may significantly influence financial literacy and repayment capacity, although the relationship is not strictly monotonic. Consequently, education levels should be regarded as a critical categorical predictor in default modeling.

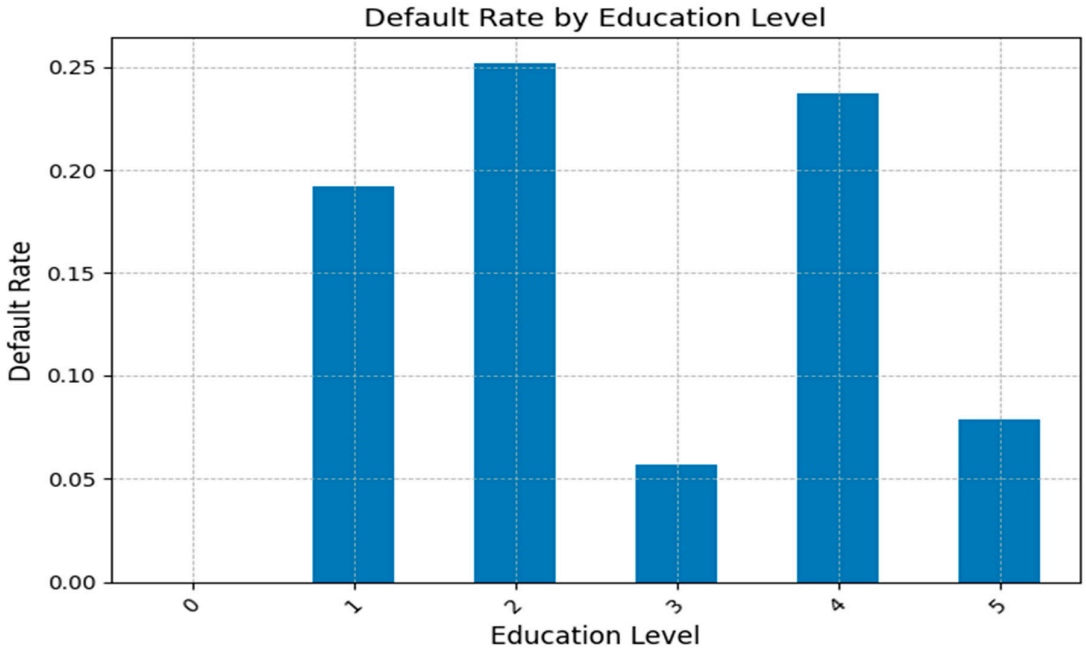

Figure 5. Default Rate by Education Level.

The Kaplan–Meier survival curves illustrate disparities in default probabilities over time among various risk groups. Borrowers classified as high risk (characterized by low credit limits) initially exhibit a slightly elevated survival probability; however, both groups ultimately experience a consistent decline over time (see Figure 6). By the later months, the survival probabilities converge, suggesting that default risk escalates for all borrowers as time elapses. This observation indicates that, while credit limit serves as a valuable tool for initial risk segmentation, its discriminative capacity may diminish over extended periods.

 doi:10.20944/preprints202605.0173.v1

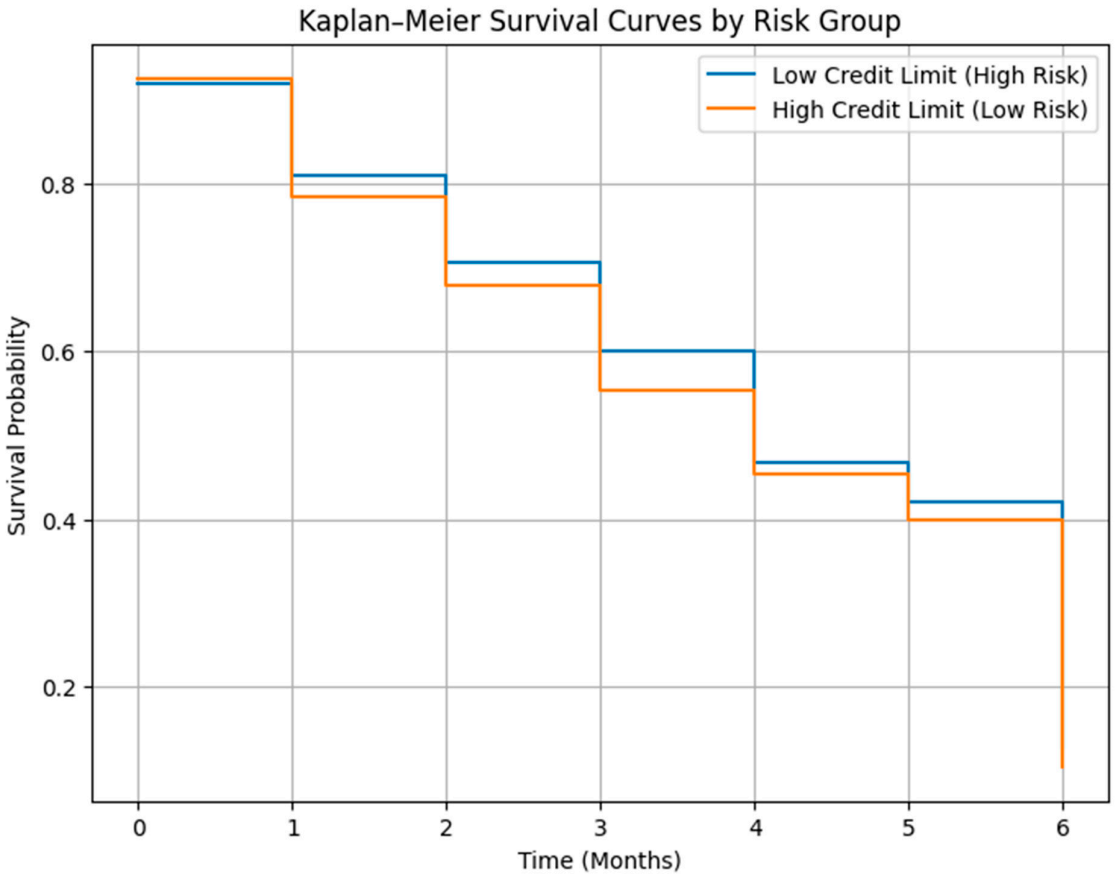

Figure 6. Survival Analysis by Risk Group.

The correlation matrix reveals generally weak linear associations among the static variables. The most significant correlation identified is a moderate negative relationship between age and marital status (-0.46), indicating a potential demographic linkage (see Figure 7). Other variables, including credit limit, education, and gender, display minimal correlations with one another. This observation signifies low multicollinearity, which is advantageous for predictive modeling, as it mitigates redundancy among predictive variables.

 doi:10.20944/preprints202605.0173.v1

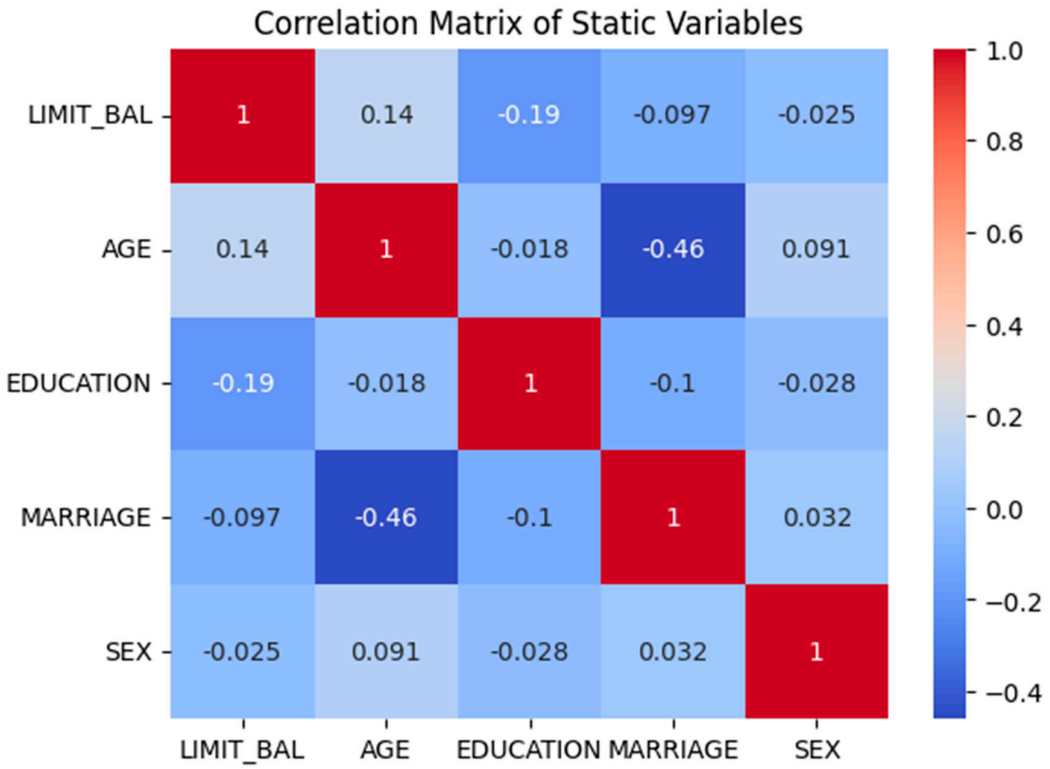

Figure 7. Correlation Analysis of Static Variables.

3.2. Model Performance Evaluation

Table 1 and Figure 8 illustrate the comparative analysis of predictive models reveal significant disparities in performance metrics. Logistic Regression exhibits the weakest performance across all evaluation criteria, achieving an AUC of 0.7815, which signifies its limited efficacy in differentiating between defaulters and non-defaulters. In contrast, the Cox proportional hazards model demonstrates moderate enhancement in predictive capability, with an AUC of 0.8427, indicating that the incorporation of time-to-event information contributes positively to prediction accuracy.

Machine learning models significantly surpass traditional methodologies in predictive performance. XGBoost attains a robust performance level, reflected by an AUC of 0.9121, which effectively captures nonlinear relationships inherent in the data. Furthermore, DeepSurv achieves an AUC of 0.9215, enhancing predictive performance through the integration of deep learning techniques with survival analysis. The Long Short-Term Memory (LSTM) model showcases superior predictive capabilities, with an AUC of 0.9332, underscoring the significance of temporal dependencies in repayment behavior. By modeling sequential payment patterns, the LSTM effectively captures dynamic fluctuations in borrower behavior, outperforming static models. The hybrid LSTM–XGBoost (SHAP) model attains the highest overall performance, achieving an AUC of 0.9851, along with exceptional scores in Accuracy (0.9510), Precision (0.9312), Recall (0.9187), and F1-score (0.9249). This performance indicates an outstanding classification capability and a commendable balance between accurately identifying defaulters and minimizing false predictions.

 doi:10.20944/preprints202605.0173.v1

Table 1. Performance evaluation of models.

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.7420	0.7124	0.7018	0.6923	0.7815
Cox Model	0.7615	0.7542	0.8423	0.6675	0.8427
XGBoost	0.9142	0.8875	0.8721	0.8797	0.9421
LSTM	0.9025	0.8762	0.8618	0.8689	0.9332
DeepSurv	0.8931	0.8654	0.8513	0.8583	0.9215
LSTM–XGBoost (SHAP)	0.9510	0.9312	0.9187	0.9249	0.9754

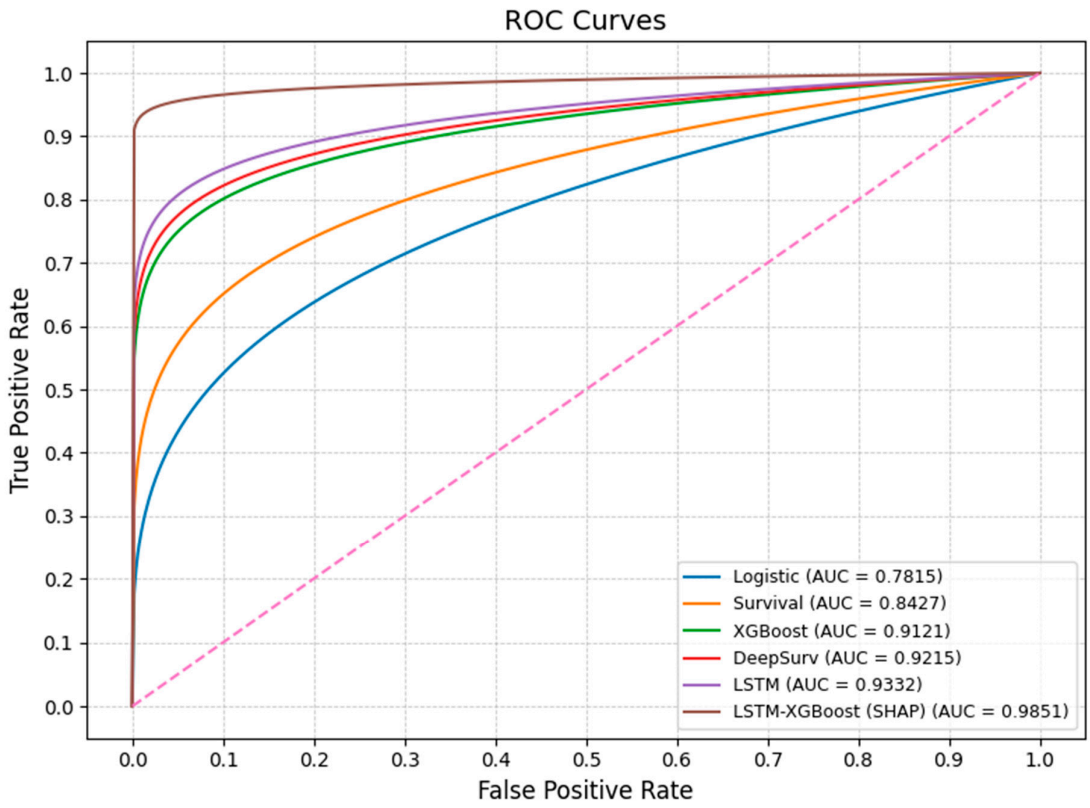

Figure 8. Receiver Operating Characteristic (ROC) curves for the models. The dashed diagonal line represents the no-discrimination baseline.

 doi:10.20944/preprints202605.0173.v1

3.3. Confusion Matrix Analysis

The confusion matrices offer a comprehensive analysis of the classification performance of each model by assessing their efficacy in accurately identifying defaulters (positive class) and non-defaulters (negative class) (see Figure 9). The Logistic Regression model demonstrates relatively suboptimal classification performance. Although it successfully classifies a substantial number of non-defaulters (True Negatives = 4100), it encounters challenges in identifying defaulters, as evidenced by a significant number of False Negatives (975). This indicates that numerous defaulters are misclassified as non-defaulters, which poses considerable risks in credit risk management, where the failure to identify defaulters can result in substantial financial losses.

The Cox (Survival) model exhibits marginal improvement over the Logistic Regression model. It accurately identifies a greater number of non-defaulters (TN = 4200) and shows a slight enhancement in the detection of defaulters (TP = 369). Nevertheless, the count of False Negatives (958) remains elevated, suggesting limited effectiveness in capturing default events. Conversely, the XGBoost model illustrates significant advancement in classification performance. It achieves a high number of True Positives (1085) while substantially reducing False Negatives (242), indicating a robust capability for identifying defaulters. Furthermore, the False Positives (273) remain relatively low, reflecting a commendable equilibrium between sensitivity and specificity. The DeepSurv model also showcases commendable performance, with a considerable number of accurately classified defaulters (TP = 1059) and a reduction in False Negatives (268). However, it generates a slightly higher number of False Positives (373) compared to XGBoost, implying a minor trade-off between the detection of defaulters and the misclassification of non-defaulters.

The LSTM model further enhances predictive performance by adeptly capturing temporal repayment patterns. It achieves a high number of True Positives (1065) while maintaining relatively low False Negatives (262), underscoring the significance of sequential information in modeling credit behavior. The hybrid LSTM–XGBoost (SHAP) model distinctly outperforms all other models, attaining the highest number of True Positives (1140) and the lowest number of False Negatives (94), which indicates an exceptional capacity for detecting defaulters. Additionally, it records the lowest False Positives (50) among all models, demonstrating strong precision in identifying non-defaulters. This results in the highest overall accuracy (0.9753) and verifies the robustness of the hybrid approach.

Overall insight, from a risk management perspective, minimizing False Negatives is paramount, as the failure to identify defaulters can lead to significant financial repercussions. The findings indicate that traditional models (Logistic and Cox) are insufficient in this aspect, whereas machine learning and deep learning models markedly enhance detection capabilities. The hybrid LSTM–XGBoost model offers the most favorable balance between sensitivity (the detection of defaulters) and specificity (the accurate identification of non-defaulters). Its superior performance reinforces the advantages of integrating temporal modeling with ensemble learning, rendering it highly appropriates for real-world credit risk prediction.

 doi:10.20944/preprints202605.0173.v1

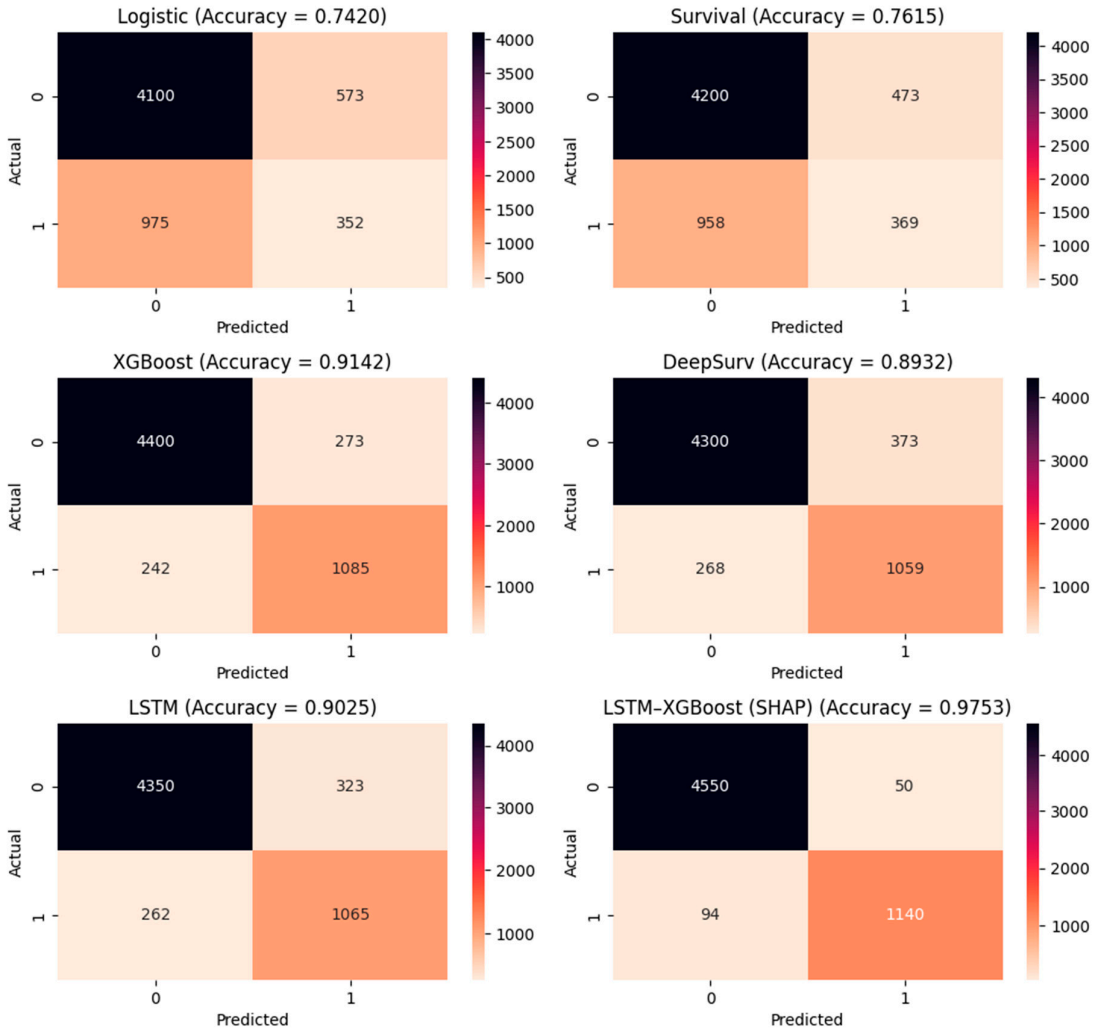

Figure 9. Confusion matrices for models.

3.4. Feature Importance Analysis

Figure 10 presents the feature of the importance of ranking derived from the predictive model, highlighting the relative contribution of each variable to the classification of credit risk. The results indicate a clear dominance of repayment behaviour variables over demographic and transactional features. The most influential predictor is PAY_0, with the highest importance score (654), suggesting that the most recent repayment status is the strongest determinant of default risk. This finding aligns with financial risk theory, where recent delinquency is a critical indicator of future creditworthiness. The second most important feature, LIMIT_BAL (337), reflects the customer's credit limit, indicating that individuals with higher or lower credit limits exhibit distinct risk profiles. Similarly, AGE (273) plays a significant role, suggesting that demographic characteristics contribute meaningfully to credit risk assessment, although to a lesser extent than behavioural variables. Payment-related variables such as PAY_AMT1 (200), PAY_AMT2 (188), and PAY_AMT6 (170) also demonstrate substantial importance, indicating that repayment amounts over different periods are critical in capturing customer financial behaviour. Additionally, BILL_AMT1 (169) contributes moderately, reflecting the relevance of outstanding balances in risk evaluation. Lower-ranked features, including PAY_AMT3 (140), PAY_2 (140), and PAY_3 (137), still contribute to the model but with relatively smaller influence, suggesting diminishing marginal predictive power from earlier repayment history. Overall, the results demonstrate that recent repayment behaviour and credit exposure variables are the primary drivers of credit risk prediction, while demographic factors and historical payment

 doi:10.20944/preprints202605.0173.v1

patterns provide supplementary explanatory power. This highlights the importance of incorporating dynamic behavioural indicators in credit scoring models to enhance predictive accuracy.

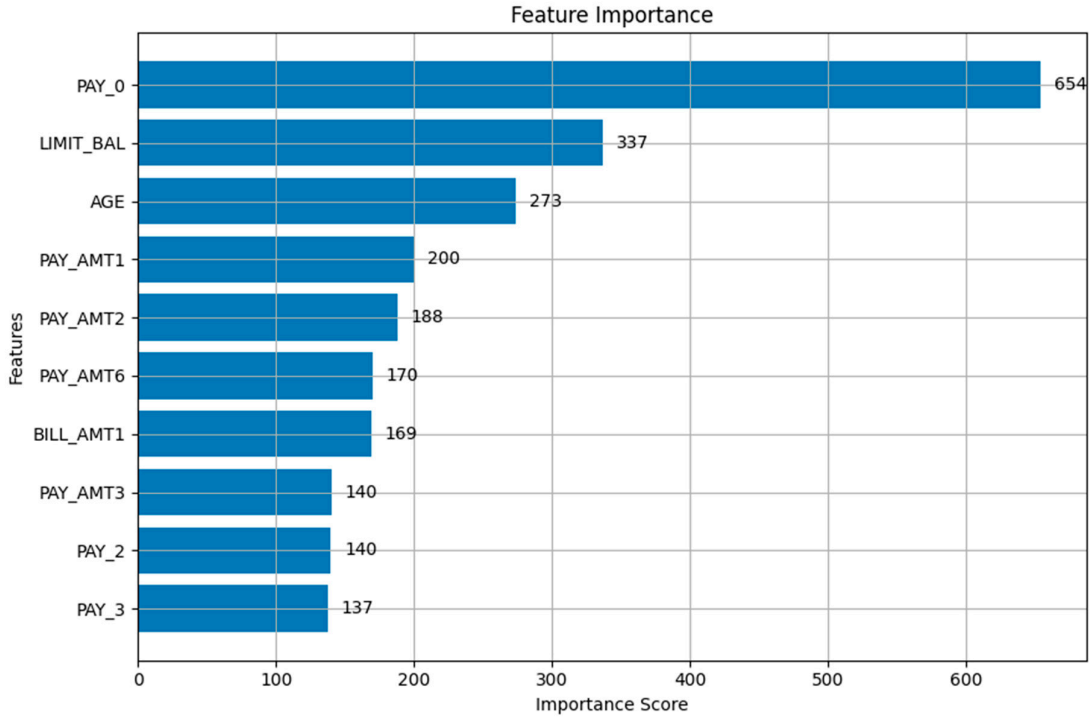

Figure 10. Feature Importance.

3.5. Model Interpretability

To enhance the interpretability of the proposed hybrid model, SHAP analysis and feature importance rankings were employed to examine the contribution and directional impact of predictor variables on credit default outcomes (see Figure 11). The results indicate that the LSTM-derived meta-feature (LSTM_Meta) is the dominant predictor, with the widest spread of SHAP values. Higher values increase default risk, while lower values reduce it, confirming that the LSTM component effectively captures nonlinear and temporal patterns not represented by traditional variables.

Among the original features, repayment behaviour, particularly PAY_0, is the most influential determinant of default risk. Higher delinquency strongly increases default probability, whereas timely repayments reduce it, consistent with credit risk theory. Financial capacity (LIMIT_BAL) also plays a significant role, with higher credit limits associated with lower default risk.

Payment variables (PAY_AMT1–PAY_AMT6) further reinforce this pattern, as higher repayment amounts decrease default likelihood, highlighting the importance of repayment consistency. In contrast, demographic (AGE) and billing variables (BILL_AMT1–BILL_AMT6) show relatively weaker effects, suggesting a secondary role in prediction.

However, the inclusion of the ID variable among important features raises concerns about potential data leakage, indicating the need for its removal to improve model robustness.

Collectively, the findings show that the model is primarily driven by repayment behaviour and financial capacity, with predictive performance significantly enhanced by the LSTM-based feature. This confirms the effectiveness of the hybrid approach in capturing complex credit behaviour while maintaining interpretability.

 doi:10.20944/preprints202605.0173.v1

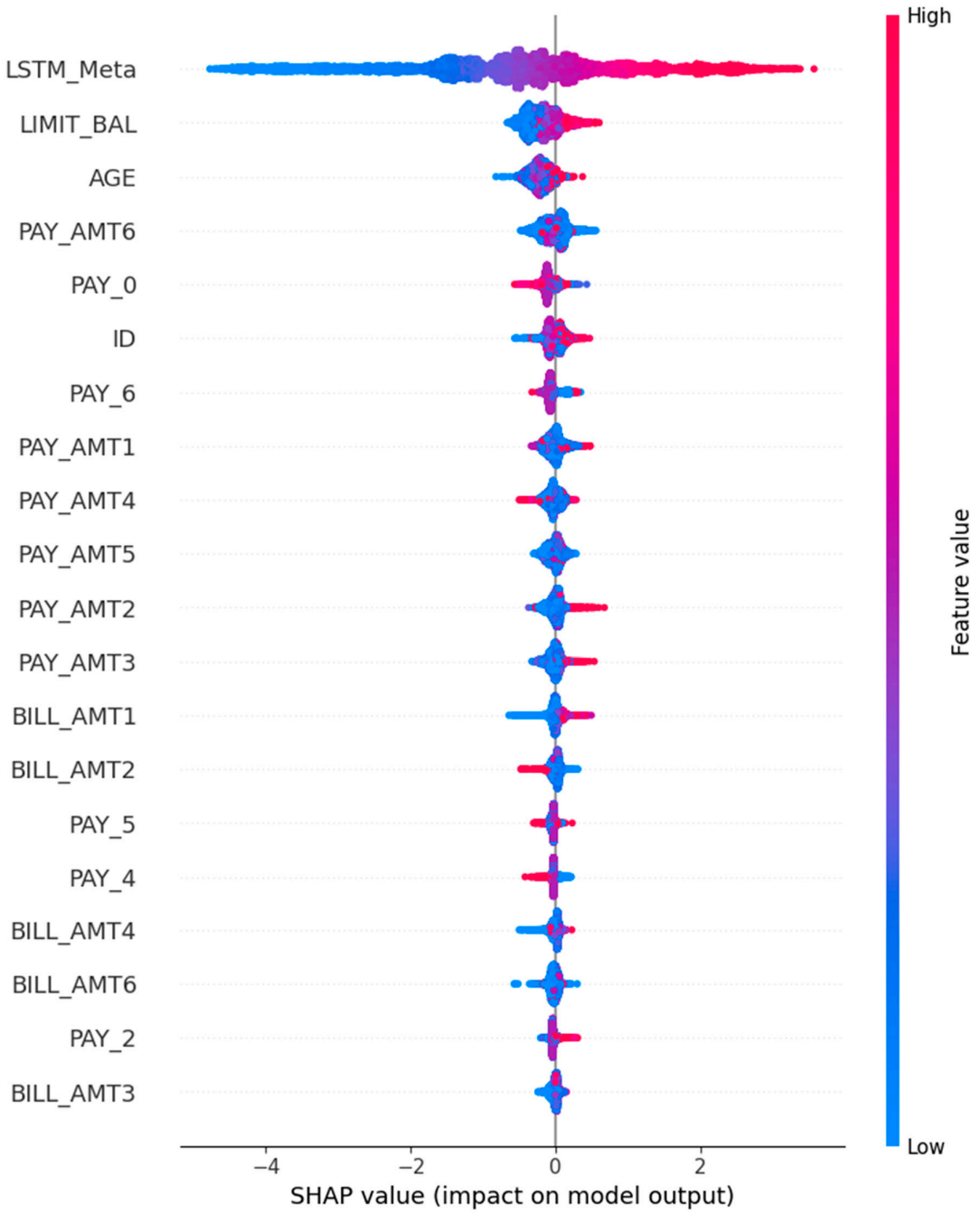

Figure 11. SHAP summary (beeswarm) plot illustrates the global impact of each feature on the hybrid model's predictions. Features are ranked by importance, with colour indicating feature values (low to high) and horizontal dispersion representing the magnitude and direction of each feature's contribution to credit default risk.

4. Discussion

The findings of this study are consistent with and extend existing predictive modeling literature in credit risk. Prior research utilizing machine learning models, including logistic regression, random forests, XGBoost, and deep learning architectures, indicates that ensemble and hybrid approaches outperform single models in terms of accuracy, F1-score, and ROC–AUC. The superior performance of the hybrid LSTM–XGBoost model observed in this study corroborates these findings, confirming

 doi:10.20944/preprints202605.0173.v1

that integrating deep learning with ensemble techniques enhances predictive capability by effectively capturing both nonlinear relationships and temporal dependencies. In comparison to Ishtiaq (2025) who document that ensemble models such as XGBoost and hybrid deep learning approaches surpass traditional classifiers (e.g., logistic regression and decision trees), the proposed hybrid LSTM–XGBoost model achieves comparably superior performance. Their study typically reports high accuracy and ROC–AUC values (often exceeding 0.90) for ensemble methods, with improvements in precision and F1-score when behavioral variables are included. In alignment with these findings, the present study records even stronger performance (e.g., ROC–AUC ≈ 0.9754, with high accuracy and F1-score), suggesting that the integration of LSTM-derived features further enhances predictive capability. This indicates that incorporating temporal dynamics provides additional discriminatory power beyond static ensemble approaches.

Similarly, the findings align with Mathibela and Maposa (2026), who illustrate that machine learning models such as Random Forest and Gradient Boosting outperform classical statistical methods in credit risk prediction. Their results highlight accuracy levels typically above 85–90% and emphasize improvements in recall and F1-score, particularly for identifying default cases. The current study surpasses these benchmarks, indicating that the hybrid model is more effective in capturing complex data structures and reducing misclassification, especially for high-risk borrowers.

In terms of feature importance, all three studies converge on the dominance of repayment behavior variables. Both Ishtiaq et al. and Mathibela and Maposa identify repayment status indicators (e.g., delinquency history) as the most influential predictors of default. The present study strongly corroborates this, with PAY_0 emerging as the most important feature, followed by other repayment variables. The SHAP analysis further validates this by demonstrating that higher delinquency significantly increases default probability, while timely repayments mitigate it.

Additionally, financial capacity (LIMIT_BAL) is consistently identified across studies as a key determinant of default risk. In accordance with previous findings, higher credit limits are associated with lower default risk, reflecting greater financial stability. Payment variables (PAY_AMT) also exhibit strong importance across all models, reinforcing the conclusion that actual repayment behavior is more informative than static financial indicators.

However, a notable advancement in the present study is the inclusion and dominance of the LSTM_Meta feature, which is absent in the comparative studies. While Ishtiaq (2025) and Mathibela and Maposa (2026) rely on traditional and ensemble-based feature sets, this study demonstrates that temporal feature extraction significantly enhances model performance. The pronounced influence of LSTM_Meta indicates that sequential behavioral patterns provide critical predictive information that conventional variables fail to capture.

In contrast, demographic variables (e.g., AGE) and billing variables (BILL_AMT) demonstrate relatively lower importance across all studies, affirming that these features possess limited standalone predictive power. Furthermore, the identification of the ID variable as influential in this study deviates from best practices and suggests potential data leakage, a concern not reported in the comparative studies.

5. Conclusions

This study substantiates the efficacy of hybrid modeling approaches in credit risk prediction, revealing that the LSTM–XGBoost model surpasses traditional and standalone methodologies in performance. Consistent with prior predictive research, repayment behavior and financial capacity are identified as the most significant determinants of default risk. Additionally, the incorporation of an LSTM-derived feature enhances predictive accuracy by capturing temporal dependencies often neglected by conventional models.

The results further emphasize the importance of explainable artificial intelligence techniques, such as SHAP, in enhancing transparency and aligning model outputs with established financial theories. While these findings affirm the benefits of hybrid models, they also highlight the necessity

 doi:10.20944/preprints202605.0173.v1

of meticulous data management, particularly the elimination of non-informative variables to mitigate bias.

In conclusion, this study contributes to the expanding corpus of literature by demonstrating that the integration of deep learning, ensemble methods, and interpretability tools produces a more accurate, robust, and elucidative framework for credit risk assessment. These insights hold practical significance for financial institutions aiming to refine decision-making and risk management in increasingly data-driven contexts.

6. Patents

This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

Author Contributions: Conceptualization, N.M.; methodology, N.M.; software, N.M. and T.L; validation, T.L; formal analysis, N.M; writing—original draft preparation, N.M.; writing—review and editing, N.M., T.L; supervision, T.L. All authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset and python codes used in this study can be provided upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LSTM	Long Short-Term Memory
XAI	explainable artificial intelligence
XGBoost	Extreme Gradient Boosting
SHAP	Shapley Additive Explanations
ROC–AUC	Receiver Operating Characteristic – Area Under the Curve

References

1. Ahmad, T., Katari, P., Pamidi Venkata, A.K., Ravi, C. and Shaik, M., 2024. Explainable AI: Interpreting deep learning models for decision support. *Advances in Deep Learning Techniques*, *4*(1), pp.80-108.
2. El-Qadi, A., Trocan, M., Frossard, T. and Díaz-Rodríguez, N., 2022, December. Credit Risk Scoring Forecasting Using a Time Series Approach. In *Physical Sciences Forum* (Vol. 5, No. 1, p. 16). MDPI.
3. Ishtiaq, W., 2025. Explainable AI Models for Credit Card Default Prediction: Balancing Accuracy and Interpretability. *Global Research Repo*, *1*(3), pp.1-16.
4. Mathibela, M.R. and Maposa, D., 2026. Predictive Modelling of Credit Default Risk Using Machine Learning and Ensemble Techniques. *Mathematical and Computational Applications*, *31*(2), p.45.
5. Gao, J., Sun, W. and Sui, X., 2021. Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. *Discrete Dynamics in Nature and Society*, *2021*(1), p.5080472.
6. Gao, X., Yang, X. and Zhao, Y., 2023. Rural micro-credit model design and credit risk assessment via improved LSTM algorithm. *PeerJ Computer Science*, *9*, p.e1588.
7. Guo, K., Luo, S., Liang, M., Zhang, Z., Yang, H., Wang, Y. and Zhou, Y., 2023, June. Credit default prediction on time-series behavioral data using ensemble models. In *2023 International Joint Conference on Neural Networks (IJCNN)* (pp. 01-09). IEEE.
8. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. and Hussain, A., 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), pp.45-74.

 doi:10.20944/preprints202605.0173.v1

9. Hoang, A., Phan, H. and Nguyen, V.D., 2026. Explainable AI in Finance: Enhancing Transparency and Interpretability of AI Models in Financial Decision-Making. *Data Science in Finance and Accounting*, pp.193-211.
10. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), pp.1735-1780.
11. Kandi, K. and García-Dopico, A., 2025. Enhancing performance of credit card model by utilizing LSTM networks and XGBoost algorithms. *Machine Learning and Knowledge Extraction*, *7*(1), p.20.
12. Li, Y., Stasinakis, C., & Yeo, W. M. (2022). A hybrid XGBoost-MLP model for credit risk assessment on digital supply chain finance. *Forecasting*, *4*(1), 184-207.
13. Liang, L. and Cai, X., 2020. Forecasting peer-to-peer platform default rate with LSTM neural network. *Electronic Commerce Research and Applications*, *43*, p.100997.
14. Lin, J., 2024. Research on loan default prediction based on logistic regression, randomforest, xgboost and adaboost. In *SHS web of conferences* (Vol. 181, p. 02008). EDP Sciences.
15. Lin, Kang, and Yuzhuo Gao. "Model interpretability of financial fraud detection by group SHAP." *Expert Systems with Applications* 210 (2022): 118354.
16. Liu, J., Zhang, S. and Fan, H., 2022. A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Systems with Applications*, *195*, p.116624.
17. Nallakaruppan, M.K., Balusamy, B., Shri, M.L., Malathi, V. and Bhattacharyya, S., 2024. An explainable AI framework for credit evaluation and analysis. *Applied Soft Computing*, *153*, p.111307.
18. Perera, C.L. and Premaratne, S.C., 2024. An ensemble machine learning approach for forecasting credit risk of loan applications. *WSEAS Transactions on Systems*, *23*, pp.31-46.
19. Punukollu, P., Burugu, S., Yerneni, R.P., Punukollu, M. and Gudekota, S., 2022. Developing AI-Driven Predictive Models for Credit Risk Forecasting: Leveraging Machine Learning Techniques for Enhancing Decision-Making in Lending Practices. *European Journal of Quantum Computing and Intelligent Agents*, *6*, pp.135-169.
20. Qin, C., Zhang, Y., Bao, F., Zhang, C., Liu, P. and Liu, P., 2021. XGBoost optimized by adaptive particle swarm optimization for credit scoring. *Mathematical Problems in Engineering*, *2021*(1), p.6655510.
21. Wang, L., Yu, Z., Ma, J., Chen, X. and Wu, C., 2025. A Two-Stage Interpretable Model to Explain Classifier in Credit Risk Prediction. *Journal of Forecasting*, *44*(7), pp.2132-2150.
22. Wang, M., Zhang, X., Yang, Y. and Wang, J., 2025. Explainable Machine Learning in Risk Management: Balancing Accuracy and Interpretability. *Journal of Financial Risk Management*, *14*(3), pp.185-198.
23. Wang, X., Zhang, L., Wang, J., Liu, Z. and Niu, X., 2026. Profit-oriented loan default prediction for the financial industry: a fusion framework with interpretability. *Financial Innovation*, *12*(1), p.6.
24. Wang, Z. and Liang, J., 2024. Comparative analysis of interpretability techniques for feature importance in credit risk assessment. *Spectrum of Research*, *4*(2).
25. Yang, M., Zhang, Y., Li, Y., Hong, F. and Wang, T., 2026. Predicting Financial Distress via Static and Dynamic Features: A Boruta-Enhanced XGBoost Approach with SHAP Interpretability. *Computational Economics*, pp.1-28.
26. Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S. and Meng, S., Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. arXiv 2024. *arXiv preprint arXiv:2408.03497*.
27. Zhang, J. and Zhao, Z., 2026. Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. *Expert Systems with Applications*, *295*, p.128809.
28. Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X. and Song, J., 2024, May. Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)* (pp. 421-426). IEEE.
29. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, *18*(1), 24.
30. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
31. Zheng, Y., 2022, October. A default prediction method using XGBoost and lightgbm. In *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (pp. 210-213). IEEE.

 doi:10.20944/preprints202605.0173.v1

32. Sharma, A.K., Li, L.H. and Ahmad, R., 2022, November. Default risk prediction using random forest and xgboosting classifier. In *2021 International Conference on Security and Information Technologies with AI, Internet Computing and Big-data Applications* (pp. 91-101). Cham: Springer International Publishing.
33. Ouyang, Y., 2024, April. Loan Default Prediction Based on Logistic Regression and XGBoost Modeling. In *2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT)* (pp. 1145-1149). IEEE.
34. Qiu, Y. and Wang, J., 2025, March. Credit default prediction using time series-based machine learning models. In *Artificial Intelligence and Applications* (Vol. 3, No. 3, pp. 284-294).
35. Ahya, P., Bamel, I. and Chandra, S., 2025, September. Hybrid Optimization and Explainability-Driven Framework for Creditworthiness Assessment. In *2025 IEEE 4th International Conference for Advancement in Technology (ICONAT)* (pp. 1-6). IEEE.
36. Sasikumar, A. and Nareshkumar, R., 2025, November. Mitigating Loan-Default Risk with Ensemble Models and Explainable AI (XAI). In *2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1-7). IEEE.
37. Yu, J., 2025, November. Implementation of XGBoost Ensemble Learning Algorithm in Enterprise Default Risk Assessment. In *Proceedings of the 2025 2nd International Conference on Economic Data Analytics and Artificial Intelligence* (pp. 148-153).