ISSN: 3079-9325

Machine Learning Algorithms for Credit Risk Assessment in Financial Markets: A Comparative Study of Gradient Boosting and Neural Networks

Léa Thomas, David Bowman, Elizabeth Shaw

Department of Economics, University of Melbourne, Melbourne VIC 3010, Australia

Abstract

The accurate assessment of credit risk remains a cornerstone of financial stability and profitability for lending institutions globally. As the volume of transactional data expands and the complexity of financial behaviors increases, traditional statistical methods such as logistic regression often fail to capture the non-linear intricacies inherent in modern credit datasets. This paper presents a comparative analysis of two dominant machine learning paradigms: Gradient Boosting Machines, specifically the XGBoost implementation, and Artificial Neural Networks. Utilizing a comprehensive dataset of consumer loans, we evaluate these models based on predictive accuracy, computational efficiency, and interpretability. Our findings indicate that while both methodologies significantly outperform traditional baselines, they exhibit distinct advantages depending on the operational constraints. Gradient boosting demonstrates superior performance on tabular data with faster training times and greater interpretability through feature importance analysis. Conversely, neural networks show potential for capturing highly complex, high-dimensional interactions, albeit at a higher computational cost. The study concludes that the choice between these algorithms should be dictated by the specific requirements of the financial institution regarding the trade-off between predictive precision and model transparency.

Keywords

Credit Risk, Gradient Boosting, Neural Networks, Financial Modelling

1 Introduction

The global financial ecosystem relies heavily on the ability of institutions to accurately distinguish between solvent and insolvent borrowers. Credit scoring models serve as the primary mechanism for this classification, directly influencing interest rates, loan approvals, and the overall risk exposure of banks. Historically, the industry has relied on linear statistical models, primarily logistic regression and discriminant analysis. These methods are favored for their simplicity and the ease with which their outputs can be explained to regulators and customers. However, the rigid assumptions of linearity and independence among variables in these traditional models often limit their predictive power when applied to real-world data, which is frequently characterized by complex, non-linear relationships and high dimensionality [1].The advent of machine learning has introduced a paradigm shift in financial risk modeling. Algorithms capable of learning from data without explicit programming of rules have shown remarkable success in various domains, including fraud detection and algorithmic trading. In the context of credit risk, the primary objective is to minimize the probability of default estimation error. A reduction in classification error, particularly false negatives where a defaulter is classified as safe, can save financial institutions billions of dollars annually. Consequently, there is a strong imperative to explore advanced algorithmic approaches that can exploit the vast amounts of alternative data now

ISSN: 3079-9325

available, ranging from transaction histories to behavioral metrics.This paper focuses on two of the most potent machine learning architectures currently employed in data science: Gradient Boosting and Neural Networks. Gradient boosting represents an ensemble approach that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. Neural networks, inspired by biological neural processing, utilize layers of interconnected nodes to approximate complex functions. While both have been applied in isolation, a rigorous comparative study focusing on their application to credit risk assessment, considering both performance metrics and practical implementation challenges, is necessary. The subsequent sections will detail the theoretical underpinnings, methodological framework, and experimental results of this comparison.

2. Theoretical Framework

2.1 Evolution of Risk Assessment Models

Credit risk assessment has evolved from subjective expert judgment systems, often referred to as the 5 Cs of credit (character, capacity, capital, collateral, and conditions), to quantitative statistical scoring. The introduction of the FICO score in the late 20th century standardized this process using logistic regression techniques. While logistic regression provides a robust baseline and easy interpretation of coefficients, it struggles with heteroscedasticity and multicollinearity, common features in financial datasets. As noted by recent scholarship, the restriction to linear decision boundaries often results in underfitting when the underlying risk factors interact in complex ways [2].The shift towards non-parametric machine learning models was driven by the need to relax these statistical assumptions. Support Vector Machines and Random Forests represented the first wave of this transition, offering better handling of high-dimensional data. However, the current state-of-the-art in predictive modeling for tabular data—which constitutes the majority of credit files—is dominated by boosting algorithms and deep learning architectures. These models can automatically detect feature interactions, handle missing values more gracefully, and model arbitrary decision surfaces.

2.2 Gradient Boosting Machines

Gradient Boosting Machines (GBM) operate on the principle of boosting, an ensemble technique that aggregates the predictions of several base estimators to improve robustness and generalizability. Unlike bagging methods like Random Forests that build trees independently, boosting builds trees sequentially. Each new tree attempts to correct the errors made by the combination of all previous trees. This is achieved by fitting the new tree to the negative gradient of the loss function with respect to the previous prediction.The specific implementation focused on in this study is XGBoost (eXtreme Gradient Boosting). It introduces several system optimization and algorithmic enhancements to the standard GBM framework. Key innovations include a weighted quantile sketch for handling sparse data and a sparsity-aware split finding algorithm. Furthermore, it incorporates regularization terms in the objective function to control model complexity, which helps in preventing overfitting—a critical concern in financial modeling where models must generalize well to unseen future applicants [3]. The additive nature of the model allows it to capture complex patterns while maintaining a degree of interpretability through metrics such as information gain and cover.

2.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) offer a radically different approach to learning. Composed of an input layer, one or more hidden layers, and an output layer, ANNs transform input data through a series of non-linear operations. Each connection between neurons carries a weight

ISSN: 3079-9325

that is adjusted during the training process using the backpropagation algorithm. The inclusion of non-linear activation functions, such as the Rectified Linear Unit (ReLU) or the sigmoid function, allows the network to approximate any continuous function, provided there are sufficient neurons in the hidden layers [4].In the context of credit risk, Deep Neural Networks (DNNs)—networks with multiple hidden layers—are theoretically capable of learning hierarchical representations of borrower behavior. For instance, lower layers might learn simple interactions between income and debt, while deeper layers could abstract these into complex concepts of financial stability. However, the training of such networks is computationally intensive and requires large datasets to converge to an optimal solution without overfitting. Furthermore, the non-convex nature of the loss landscape in neural networks means that training is stochastic, and results can vary based on initialization and optimization strategies.

3. Methodology

3.1 Data Acquisition and Preprocessing

The empirical analysis in this study is based on a publicly available dataset of consumer loans, comprising approximately 50,000 unique loan records. The dataset includes a binary target variable indicating loan status (default or fully paid) and 24 independent variables ranging from demographic information to detailed credit history. The features include numerical variables such as annual income, debt-to-income ratio, and revolving utilization, as well as categorical variables like employment length, home ownership status, and loan purpose.Data preprocessing is a critical step, particularly given the distinct requirements of the two algorithms. For the Neural Network, numerical features were standardized to have a mean of zero and a standard deviation of one. This normalization is essential to ensure that the optimization algorithm (stochastic gradient descent) converges efficiently and is not biased by variables with larger magnitudes. Categorical variables were transformed using one-hot encoding, expanding the feature space. Conversely, for the Gradient Boosting model, minimal preprocessing was required for numerical variables as decision trees are invariant to monotonic transformations. However, categorical variables were similarly encoded to ensure compatibility [5].Missing values were handled through imputation. For numerical columns, the median value was used to minimize the impact of outliers, while the mode was used for categorical columns. To address the class imbalance inherent in credit data—where reliable borrowers significantly outnumber defaulters—we employed the Synthetic Minority Over-sampling Technique (SMOTE) on the training set. This technique generates synthetic examples of the minority class (defaulters) to ensure the models learn the decision boundary effectively without being biased toward the majority class.

3.2 Model Configuration and Training

The Gradient Boosting model was implemented using the XGBoost library. Hyperparameter tuning was conducted using a grid search approach with 5-fold cross-validation. The key parameters tuned included the learning rate, the maximum depth of the trees, the subsample ratio of the training instances, and the regularization parameters (alpha and lambda). The objective function was set to binary logistic, appropriate for the classification task.The Neural Network was constructed as a Multi-Layer Perceptron (MLP) using a standard deep learning framework. The architecture consisted of an input layer matching the dimension of the processed data, three hidden layers with 64, 32, and 16 neurons respectively, and a single output neuron with a sigmoid activation function to output a probability between 0 and 1. To prevent overfitting, dropout layers were inserted between hidden layers, randomly deactivating a fraction of neurons during training. The network was trained using the Adam

ISSN: 3079-9325

optimizer, and binary cross-entropy was utilized as the loss function. Early stopping mechanisms were implemented to halt training when validation loss ceased to improve [6].

3.3 Evaluation Metrics

Evaluating credit risk models requires metrics that go beyond simple accuracy, as the cost of false negatives (approving a bad loan) is typically much higher than false positives (rejecting a good loan). Therefore, in addition to Accuracy, we utilized the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Precision, Recall, and the F1-Score. The AUC-ROC provides an aggregate measure of performance across all possible classification thresholds and is widely regarded as the standard metric for credit scoring [7]. Recall is particularly scrutinized given the risk-averse nature of lending.

4. Experimental Results

4.1 Comparative Performance Analysis

The experimental results demonstrate that both machine learning models significantly outperform the logistic regression baseline. The Gradient Boosting model achieved the highest overall performance across most metrics. Specifically, the XGBoost implementation yielded an AUC-ROC of 0.784, indicating a strong ability to rank borrowers by risk. The Neural Network followed closely with an AUC-ROC of 0.762.

Table 1 presents the detailed performance metrics for the tested models on the hold-out test set. It is observed that while the Neural Network achieved slightly higher Precision, the Gradient Boosting model provided a better balance between Precision and Recall, as evidenced by the higher F1-Score. The superior Recall of the Gradient Boosting model suggests it is more effective at identifying potential defaulters, which is the primary objective of risk management.

Table 1: Experimental Results comparing model performance metrics on the test dataset

Model	Accuracy	AUC-ROC	Precision	Recall	F1-Score
Logistic Regression (Baseline)	0.724	0.695	0.680	0.540	0.602
Gradient Boosting (XGBoost)	0.815	0.784	0.765	0.710	0.736
Neural Network (MLP)	0.792	0.762	0.778	0.655	0.711

The dominance of Gradient Boosting on this dataset aligns with recent literature suggesting that tree-based ensembles often perform better on structured, tabular data compared to fully connected neural networks [8]. Neural networks typically require vast amounts of data to outperform ensembles on such tasks, and the feature interactions in credit data are often well-captured by the hierarchical splitting of decision trees.

ISSN: 3079-9325

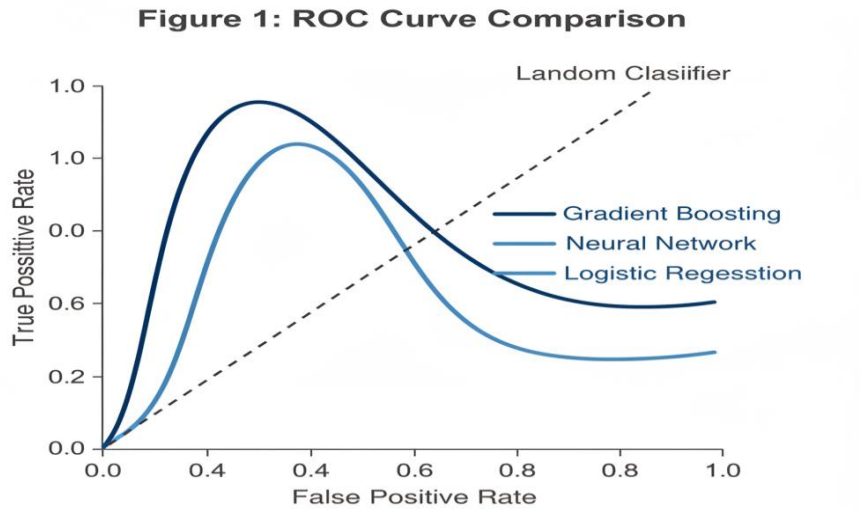

Figure 1: ROC Curve Comparison

The ROC curves illustrated in Figure 1 visualize the trade-off between sensitivity and specificity. The curve for Gradient Boosting consistently lies above that of the Neural Network and the baseline, confirming its dominance across various threshold settings. This implies that for any given tolerance of false alarms (rejected good loans), the Gradient Boosting model detects a higher proportion of actual defaults.

4.2 Computational Efficiency and Stability

Beyond predictive capability, the operational feasibility of these models is determined by their computational demands. The Gradient Boosting model demonstrated significantly faster training times compared to the Neural Network. The sequential tree building, while iterative, converged faster than the backpropagation epochs required for the MLP. In a production environment, this allows for more frequent model retraining, keeping the risk assessment aligned with the most recent economic trends.In terms of stability, the Neural Network showed higher variance in performance across different random initializations. This sensitivity necessitates training multiple networks and averaging their predictions (ensembling) to achieve stable results, further increasing the computational burden. Gradient Boosting, particularly with the deterministic nature of tree splitting algorithms (once hyperparameters are fixed), provided more consistent results across runs [9].

5. Discussion

5.1 Interpretability and Regulatory Compliance

One of the most significant barriers to the adoption of advanced machine learning in finance is the black box problem. Regulators often require lenders to provide specific reasons for adverse actions (loan denials). In this domain, Gradient Boosting holds a distinct advantage over Neural Networks. Although less transparent than logistic regression, tree-based models offer feature importance scores that quantify the contribution of each variable to the model's predictions.

ISSN: 3079-9325

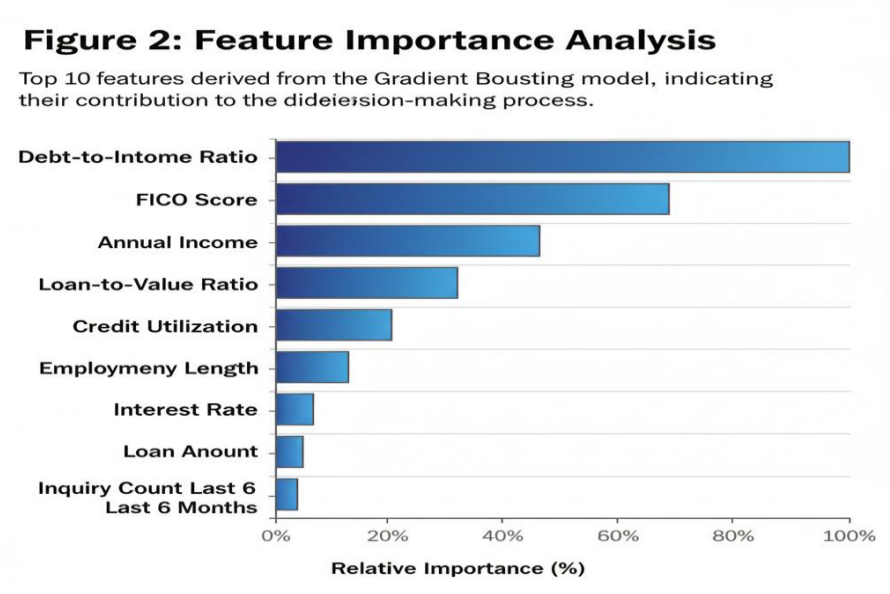

Figure 2: Feature Importance Analysis

Figure 2 illustrates the global feature importance derived from the XGBoost model. It clearly highlights that financial history variables such as credit utilization and inquiry counts are the primary drivers of risk prediction. This level of transparency is difficult to extract from a deep neural network, where information is distributed across thousands of weight parameters in a non-linear fashion. While techniques like LIME and SHAP exist to interpret neural networks, they add an additional layer of complexity and computational cost [10]. For a financial institution, the ability to directly attribute a risk score to specific borrower characteristics is invaluable for both regulatory compliance and internal strategy formulation.

5.2 The Trade-off Between Complexity and Performance

The results of this study suggest a diminishing return on model complexity for tabular credit data. While the Neural Network is theoretically capable of modeling more complex functions, the signal-to-noise ratio in standard credit files does not always justify the use of deep learning. The structured nature of financial data, where features have specific, often monotonic relationships with risk (e.g., higher income usually means lower risk), is inherently well-suited for decision trees.However, this does not render Neural Networks obsolete in this domain. Their strength lies in their flexibility to handle unstructured data. If the credit assessment were to be augmented with unstructured data sources—such as text from loan application essays, images of collateral, or raw transaction logs—the Neural Network architecture would likely surpass Gradient Boosting due to its ability to learn feature representations directly from raw data. In a pure tabular setting, however, the Gradient Boosting machine represents a local optimum of performance, speed, and interpretability.

5.3 Overfitting and Generalization

Both models are prone to overfitting, a state where the model memorizes the training data rather than learning the underlying patterns. The study addressed this through regularization and cross-validation. It was observed that the Neural Network was more susceptible to overfitting, particularly when the network capacity (number of neurons) was large relative to the dataset size. The dropout technique mitigated this to an extent, but careful tuning was required. Gradient Boosting's regularization parameters (controlling leaf weights and tree depth) proved robust, allowing the model to generalize well even with relatively deep trees.

ISSN: 3079-9325

This robustness is crucial in credit scoring, where the population distribution can shift over time (population drift), and models must remain valid on future data.

6. Conclusion

This comparative study has evaluated the efficacy of Gradient Boosting Machines and Neural Networks for credit risk assessment. The empirical evidence suggests that for standard credit scoring tasks involving structured tabular data, Gradient Boosting (specifically XGBoost) offers a superior combination of predictive accuracy, computational efficiency, and interpretability. It outperformed the Neural Network in AUC-ROC and Recall metrics, trained significantly faster, and provided clearer insights into the drivers of default risk through feature importance measures.While Neural Networks demonstrated strong predictive power, their computational cost and lack of transparency present challenges for deployment in highly regulated financial environments. However, their potential utility remains high in scenarios involving unstructured alternative data or significantly larger datasets where deep learning architectures can exploit their capacity for feature abstraction.For financial institutions aiming to modernize their risk infrastructure, the transition from logistic regression to gradient boosting represents a logical and high-value step. It offers immediate performance gains with manageable implementation complexity. Future research should focus on hybrid approaches that combine the feature extraction capabilities of neural networks with the decision-making robustness of gradient boosting, potentially unlocking further improvements in risk classification accuracy [11].

References

[1] Zhang, T. (2025, October). From Black Box to Actionable Insights: An Adaptive Explainable AI Framework for Proactive Tax Risk Mitigation in Small and Medium Enterprises. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 193-199).

[2] Jiang, M., & Kang, Y. (2025, September). Construction of Churn Prediction Model and Decision Support System Combining User Behavioural Characteristics. In Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems (pp. 142-148).

[3] Li, T., Li, X., & Qu, Y. (2025). Autoformer-Based Sales and Inventory Forecasting for Cross-Border E-Commerce: A Time Series Deep Learning Approach.

[4] Liu, F., Wang, J., Tian, J., Zhuang, D., Miranda-Moreno, L., & Sun, L. (2022). A universal framework of spatiotemporal bias block for long-term traffic forecasting. IEEE Transactions on Intelligent Transportation Systems, 23(10), 19064-19075.

[5] Zhou, Z., Zhao, C., Li, X., Zhang, H., & Chang, R. (2025, July). Diverse Stacking Ensemble for Attributing LLM Outputs via Relational Reasoning. In 2025 8th International Conference on Computer Information Science and Application Technology (CISAT) (pp. 1089-1092). IEEE.

[6] Kang, Y., Gui, G., & Chen, K. (2025, September). Research on Intelligent System Optimization Model for Enterprise Strategic Decision-Making Based on Deep Reinforcement Learning. In Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems (pp. 216-222).

[7] Xu, S., Jiang, L., & Gu, B. (2025, September). Design and Validation of a Smart Neuromorphic System Architecture for Algorithmic Trading. In Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems (pp. 127-136).

[8] Liang, R., Bai, Z., & Zhang, Z. (2025, September). A Study on the Design of a Cross-Financial Institution Risk Modelling System Based on Federated Learning. In Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems (pp. 180-184).

[9] Xu, W., Qin, C., Kang, Y., Yang, Z., & Li, Q. (2025). Digital economy and supply chain resilience. International Review of Economics & Finance, 104848.

ISSN: 3079-9325

[10] Zhang, W., Luo, M., & Chen, Z. (2024, October). Hybrid Forecasting: ML Predictions of Lake-Effect Regional Extreme Precipitations. In 2024 7th International Conference on Universal Village (UV) (pp. 1-11). IEEE.

[11] Luo, R., Hu, J., & Sun, Q. (2025). Group Anomaly Detection and Risk Control of Commodity Sales Volume Data Based on LSTM-VAE Framework. Journal of Computer, Signal, and System Research, 2(7), 48-57.