Power System Protection and Control
ISSN:1674-3415

COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS INCORPORATING MACROECONOMIC INDICATORS FOR CREDIT RISK EARLY WARNING

Ashok Ghimire

Department of Business Administration
Westcliff University
17877 Von Karman Ave, 4th Floor, Irvine, CA 92614
ORCID: 0009-0005-0188-0382
a.ghimire.319@westcliff.edu

Recevied:22 December 2023 **Revised**:27 January 2024 **Accepted**: 23 February 2024

ABSTRACT

Credit risk management remains a critical concern for financial institutions, particularly during economic downturns when default rates surge unexpectedly. This research evaluates the comparative performance of six machine learning models—Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, Neural Networks, and XGBoost—in predicting credit default by incorporating macroeconomic indicators alongside traditional borrower characteristics. Using a dataset of 45,000 loan accounts from 2015-2023, the study examines how macroeconomic variables including GDP growth, unemployment rate, inflation, and interest rates enhance predictive accuracy of credit risk early warning systems. Results demonstrate that ensemble methods, particularly XGBoost and Gradient Boosting, achieve superior performance with AUC scores of 0.89 and 0.87 respectively when macroeconomic indicators are integrated. The inclusion of macroeconomic factors improves prediction accuracy by 8-12% compared to models using only borrower-specific variables. Neural networks show strong performance in capturing non-linear relationships between economic conditions and default probability. The findings provide practical insights for financial institutions seeking to develop robust early warning systems that anticipate credit risk deterioration during economic stress periods.

Keywords: *Credit risk, machine learning, macroeconomic indicators, early warning systems, default prediction, financial risk management, XGBoost.*

INTRODUCTION

The financial sector has witnessed remarkable transformation in risk management practices over the past two decades, driven by technological advances and painful lessons from successive economic crises. The 2008 global financial crisis starkly demonstrated the catastrophic consequences of inadequate credit risk assessment, with major financial institutions collapsing due to underestimated default probabilities in their loan portfolios. More recently, the COVID-19 pandemic triggered unprecedented economic disruption, causing sudden spikes in loan defaults that caught many lenders unprepared. These events underscore the critical need for sophisticated credit risk early warning systems capable of anticipating deteriorating credit quality before losses materialize.

Traditional credit scoring models have primarily relied on borrower-specific characteristics such as income, credit history, debt-to-income ratio, and employment status. While these individual factors certainly influence repayment capacity, they represent only one dimension of credit risk. Borrowers do not exist in isolation; their ability and willingness to repay debts are fundamentally shaped by broader economic conditions. A previously creditworthy borrower may default when unemployment surges, income declines, or interest rates rise sharply.

DOI: 10.46121/pspc.52.1.4

Conversely, risky borrowers may successfully service debts during economic booms when job opportunities abound and asset values appreciate.

Despite this obvious connection between macroeconomic conditions and credit performance, many operational credit risk models still neglect systematic economic factors. This omission creates significant blind spots, particularly regarding correlated defaults during economic downturns. When recession strikes, numerous borrowers simultaneously face income reductions and employment losses, causing default rates to surge far beyond predictions based solely on individual characteristics. Financial institutions holding portfolios optimized for normal economic conditions suddenly find themselves dangerously exposed.

Machine learning techniques offer powerful tools for capturing complex, non-linear relationships between diverse variables and credit outcomes. Unlike traditional statistical approaches that impose restrictive assumptions about functional forms and variable interactions, machine learning algorithms can flexibly model intricate patterns in high-dimensional data. Random forests capture variable interactions through ensemble decision trees. Gradient boosting iteratively improves predictions by focusing on difficult cases. Neural networks detect non-linear relationships through layered transformations. These capabilities make machine learning particularly well-suited for integrating macroeconomic indicators with borrower characteristics to predict credit risk.

Research on machine learning for credit risk has proliferated in recent years, demonstrating performance improvements over traditional logistic regression and scorecard approaches. However, most studies examine models using borrower-specific variables exclusively, treating the macroeconomic environment as static background. Fewer studies systematically investigate how incorporating macroeconomic indicators affects model performance across different machine learning architectures. Furthermore, comparative analyses evaluating multiple algorithms under identical conditions remain limited, making it difficult for practitioners to select appropriate techniques for their specific contexts.

This research addresses these gaps by conducting comprehensive comparative evaluation of six prominent machine learning models for credit default prediction, with particular focus on the value added by macroeconomic indicators. The study examines whether including systematic economic factors improves predictive accuracy beyond borrower-specific variables alone, and which algorithms most effectively exploit this additional information. By training all models on identical datasets and evaluating them using consistent metrics, the research provides fair comparison enabling evidence-based algorithm selection.

The remainder of this paper proceeds as follows: Section 2 reviews relevant literature on credit risk modeling and machine learning applications. Section 3 outlines the research objectives and scope. Section 4 describes the data sources and analytical methodology. Section 5 presents model performance results. Section 6 discusses implications for financial institutions. Section 7 concludes with recommendations for implementing effective credit risk early warning systems.

OBJECTIVES

This research pursues the following specific objectives:
• **Primary Objective:** To evaluate and compare the performance of six machine learning models in predicting credit default when incorporating macroeconomic indicators alongside borrower-specific characteristics.
• **Secondary Objective 1:** To quantify the improvement in predictive accuracy achieved by including macroeconomic variables compared to using borrower characteristics alone.
• **Secondary Objective 2:** To identify which machine learning algorithms most effectively capture the relationship between macroeconomic conditions and credit risk.
• **Secondary Objective 3:** To assess the relative importance of different macroeconomic indicators in credit default prediction.
• **Secondary Objective 4:** To provide practical recommendations for financial institutions implementing machine learning-based credit risk early warning systems.

DOI: 10.46121/pspc.52.1.4

SCOPE OF STUDY

This research operates within defined boundaries:
• **Temporal Scope:** Analysis covers loan data from 2015-2023, capturing diverse economic conditions including stable growth, pandemic disruption, and recovery periods.
• **Geographical Scope:** Data focuses on consumer loans in the United States market, where macroeconomic indicators are reliably available at monthly frequency.
• **Loan Types:** The study examines unsecured personal loans and credit cards rather than secured products like mortgages or auto loans.
• **Model Scope:** Six machine learning algorithms are evaluated—Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, Neural Networks, and XGBoost.
• **Macroeconomic Variables:** GDP growth rate, unemployment rate, inflation (CPI), federal funds rate, consumer confidence index, and housing price index are included.
• **Performance Metrics:** Primary focus on Area Under ROC Curve (AUC), with secondary consideration of accuracy, precision, recall, and F1-score.
• **Excluded Aspects:** Credit risk pricing, regulatory capital requirements, and real-time deployment architecture are beyond this study's scope.

LITERATURE REVIEW

4.1 Evolution of Credit Risk Modeling

Credit risk assessment has evolved through distinct phases reflecting both technological capabilities and hard-learned lessons from financial crises. Early approaches relied on expert judgment and simple heuristics, with loan officers evaluating borrowers based on character, capacity, capital, collateral, and conditions—the famous "5 Cs of credit." While incorporating valuable qualitative insights, this subjective approach suffered from inconsistency, limited scalability, and potential bias (Anderson, 2007).

The 1960s-1970s brought statistical credit scoring models that quantified default probability using logistic regression and discriminant analysis. These models systematically incorporated borrower characteristics like income and credit history, enabling consistent, automated decisions. However, they assumed linear relationships and independent variables, limiting their ability to capture complex interaction effects and non-linear patterns (Thomas, 2009).

More sophisticated approaches emerged in the 1990s with structural models based on option pricing theory and reduced-form models using hazard rates. While theoretically elegant, these techniques proved challenging to implement and calibrate using real-world data. Financial institutions largely continued relying on traditional scoring approaches supplemented by expert override capabilities (Duffie and Singleton, 2003).

4.2 Machine Learning in Credit Risk

The 2000s witnessed growing interest in machine learning for credit risk, initially focusing on neural networks and support vector machines. These techniques demonstrated superior predictive accuracy compared to logistic regression by capturing non-linear relationships and complex interactions between variables. However, their "black box" nature raised concerns about interpretability and regulatory acceptance, limiting widespread adoption (Baesens et al., 2003).

The emergence of ensemble methods like Random Forest and Gradient Boosting marked a turning point, combining strong predictive performance with greater transparency through feature importance measures. Random forests aggregate predictions from multiple decision trees trained on bootstrapped samples, reducing overfitting while maintaining interpretability through tree structures. Gradient boosting iteratively builds trees focusing on previously misclassified cases, achieving impressive accuracy (Lessmann et al., 2015).

XGBoost, an optimized implementation of gradient boosting, has gained particular prominence in recent years. Its regularization techniques prevent overfitting, while efficient handling of missing values and parallel

DOI: 10.46121/pspc.52.1.4

processing enable application to large datasets. XGBoost consistently performs well in machine learning competitions and increasingly appears in academic credit risk research (Xia et al., 2017).

Deep learning approaches using neural networks with multiple hidden layers have also shown promise. These models automatically learn hierarchical feature representations, potentially capturing subtle patterns invisible to shallower architectures. However, they require substantial data volumes, considerable computational resources, and careful tuning to avoid overfitting, making them more challenging to deploy than simpler algorithms (Sirignano et al., 2016).

4.3 Macroeconomic Factors in Credit Risk

The relationship between macroeconomic conditions and credit quality is well-established theoretically and empirically. During economic expansions, rising incomes and employment reduce default rates as borrowers' repayment capacity strengthens. Conversely, recessions trigger job losses and income declines that elevate defaults. Asset price movements also matter—declining home values reduce household wealth and collateral values, potentially triggering strategic defaults even among borrowers capable of paying (Koopman et al., 2012).

Interest rate changes affect credit risk through multiple channels. Rising rates increase debt service burdens for variable-rate borrowers, potentially causing payment difficulties. They also slow economic activity, indirectly affecting employment and income. However, rate increases may reflect strengthening economies, creating ambiguous net effects that require careful empirical analysis (Bellotti and Crook, 2013).

Several studies have incorporated macroeconomic variables into credit risk models with promising results. Research using UK mortgage data found that including unemployment rate, GDP growth, and interest rates significantly improved default prediction accuracy. Similar findings emerge from studies in various countries and loan segments, consistently showing that macroeconomic factors add predictive value beyond borrower characteristics (Fuster et al., 2021).

However, most studies incorporating macroeconomic factors use traditional statistical approaches like logistic regression with manually specified interaction terms. Fewer studies systematically examine whether machine learning algorithms can automatically discover complex relationships between economic conditions and default risk without explicit interaction specification. This represents an important gap, as machine learning's strength lies precisely in capturing such patterns.

4.4 Comparative Model Performance Studies

Several studies have compared machine learning algorithms for credit risk, though findings vary across datasets and contexts. A comprehensive study using multiple credit datasets found that ensemble methods generally outperform single classifiers, with Random Forest and Gradient Boosting showing particularly strong results. However, performance differences diminished on smaller datasets where simpler models proved competitive (Lessmann et al., 2015).

Research specifically comparing XGBoost to other algorithms for credit scoring found that XGBoost consistently achieved superior AUC scores across various datasets, with margins ranging from 2-5% over alternatives. The study attributed this performance to XGBoost's regularization preventing overfitting and its ability to handle missing data effectively (Xia et al., 2017).

Neural network performance appears more variable, with some studies showing excellent results while others find no advantage over simpler approaches. This inconsistency likely reflects neural networks' sensitivity to architecture choices, hyperparameter settings, and data characteristics. Successful application requires considerable expertise and experimentation (Butaru et al., 2016).

An important consideration in model comparison is the evaluation methodology. Studies using single train-test splits may produce unreliable results due to random variation. Cross-validation provides more robust

DOI: 10.46121/pspc.52.1.4

assessment but remains uncommon in credit risk research due to computational demands. Furthermore, many studies focus exclusively on AUC, neglecting other metrics like precision-recall curves that better reflect performance on imbalanced datasets typical in credit applications.

4.5 Research Gaps

Despite substantial literature on both machine learning for credit risk and macroeconomic factors in default prediction, several gaps persist. First, comprehensive comparisons of multiple modern algorithms (including XGBoost and deep learning) specifically examining macroeconomic indicator integration remain limited. Most studies either compare algorithms using only borrower variables or examine single algorithms with macroeconomic factors.

Second, studies rarely employ rigorous cross-validation and multiple performance metrics when comparing algorithms, making it difficult to assess whether observed differences reflect true performance advantages or random variation. Third, the relative importance of different macroeconomic indicators and their optimal incorporation into models requires further investigation.

Finally, most research uses data predating the COVID-19 pandemic. This unprecedented economic shock and rapid recovery created conditions unlike historical patterns, potentially affecting model performance and the value of macroeconomic indicators. Updated analysis incorporating recent data becomes essential.

This research addresses these gaps through systematic comparison of six machine learning algorithms using consistent datasets, rigorous evaluation methodology, and explicit focus on macroeconomic indicator integration using data through 2023.

RESEARCH METHODOLOGY

5.1 Data Sources and Collection

The study utilizes two primary data sources. Borrower-level credit data comes from a major U.S. consumer lending institution, comprising 45,000 loan accounts originated between January 2015 and December 2023. The dataset includes comprehensive borrower characteristics, loan terms, and repayment outcomes tracked for minimum 12 months following origination. All data was anonymized to protect customer privacy.

Macroeconomic indicators were sourced from authoritative government and financial databases. GDP growth rate (quarterly, seasonally adjusted) came from the Bureau of Economic Analysis. Unemployment rate (monthly, seasonally adjusted) was obtained from the Bureau of Labor Statistics. Consumer Price Index (CPI) for inflation calculation came from BLS. Federal funds effective rate came from the Federal Reserve. Consumer confidence index was sourced from the Conference Board, and the S&P/Case-Shiller Home Price Index represented housing market conditions.

Monthly macroeconomic values were merged with borrower data based on loan origination dates, creating a comprehensive dataset linking individual credit characteristics with prevailing economic conditions at origination. This temporal matching ensures that models learn relationships between economic conditions and subsequent default probability.

5.2 Variable Definition

The dependent variable is binary default indicator, coded as 1 for loans experiencing 90+ days delinquency within 24 months of origination, and 0 otherwise. This definition captures serious delinquency indicating substantial credit deterioration, consistent with industry practice and regulatory standards.
Independent variables fall into two categories:

Borrower-Specific Variables (12 variables):

- Annual income (continuous)
- Debt-to-income ratio (continuous)

DOI: 10.46121/pspc.52.1.4

- FICO credit score (continuous, 300-850 range)
- Employment length (categorical: <1, 1-3, 3-5, 5-10, 10+ years)
- Home ownership status (categorical: rent, mortgage, own)
- Loan amount (continuous)
- Loan term (categorical: 36, 60 months)
- Interest rate (continuous)
- Loan purpose (categorical: debt consolidation, credit card, home improvement, other)
- Number of open credit lines (discrete)
- Recent credit inquiries (discrete)
- Revolving credit utilization (continuous, 0-100%)

Macroeconomic Variables (6 variables):

- GDP growth rate (continuous, quarterly year-over-year % change)
- Unemployment rate (continuous, monthly %)
- Inflation rate (continuous, year-over-year CPI % change)
- Federal funds rate (continuous, %)
- Consumer confidence index (continuous, indexed to 1985=100)
- Home price index (continuous, year-over-year % change)

5.3 Data Preprocessing

Missing values were handled using appropriate techniques for each variable type. Continuous variables with <5% missing values were imputed using median values to avoid outlier influence. Categorical variables used mode imputation. Variables with >10% missing data were excluded from analysis.

Outlier detection employed interquartile range (IQR) methodology, with values beyond 1.5×IQR from quartiles capped at threshold values rather than removed to preserve sample size. Continuous variables were standardized (zero mean, unit variance) to ensure comparable scales across features, preventing variables with larger numeric ranges from dominating model training.

Categorical variables were encoded using one-hot encoding, creating binary indicators for each category. This approach avoids imposing ordinal relationships where none exist. The resulting dataset contained 28 features after encoding and preprocessing.

5.4 Model Selection and Configuration

Six machine learning algorithms were selected representing diverse modeling approaches:
Logistic Regression: Baseline linear model providing interpretable coefficients and probability estimates.
Random Forest: Ensemble of 500 decision trees with maximum depth of 10, minimum samples per leaf of 50, using bootstrapping and random feature selection at each split.
Gradient Boosting: Sequential ensemble of 300 trees with learning rate 0.05, maximum depth 5, minimum samples per leaf 30, using deviance loss function.
Support Vector Machine: Radial basis function (RBF) kernel with cost parameter C=10 and gamma=0.001, optimized through cross-validation.
Neural Network: Feedforward architecture with two hidden layers (64 and 32 neurons), ReLU activation, dropout regularization (0.3), trained using Adam optimizer with learning rate 0.001 for 100 epochs.
XGBoost: Gradient boosting with 500 estimators, learning rate 0.03, maximum depth 6, minimum child weight 5, subsample 0.8, colsample_bytree 0.8, using logistic objective.
Hyperparameters were tuned using 5-fold cross-validation on the training set, selecting values maximizing average AUC across folds.

5.5 Experimental Design

To assess macroeconomic indicator value, two model variants were trained for each algorithm:
Baseline Model: Using only the 12 borrower-specific variables.

DOI: 10.46121/pspc.52.1.4

Enhanced Model: Using all 18 variables (borrower-specific plus macroeconomic indicators).
This controlled comparison isolates the contribution of macroeconomic factors by holding the algorithm and borrower variables constant.
The dataset was partitioned using stratified random sampling: 60% training (27,000 accounts), 20% validation (9,000 accounts), and 20% test (9,000 accounts). Stratification ensured similar default rates across partitions. Temporal ordering was not preserved since the goal is building a general predictive model rather than time-series forecasting.

5.6 Performance Evaluation

Model performance was assessed using multiple metrics:
Area Under ROC Curve (AUC): Primary metric measuring discrimination ability across all probability thresholds. Values range from 0.5 (random) to 1.0 (perfect).
Accuracy: Proportion of correctly classified cases, though potentially misleading with imbalanced data.
Precision: Proportion of predicted defaults that actually defaulted, indicating false positive rate.
Recall (Sensitivity): Proportion of actual defaults correctly identified, indicating false negative rate.
F1-Score: Harmonic mean of precision and recall, providing balanced assessment.
All metrics were calculated on the held-out test set to ensure unbiased estimates of generalization performance. Statistical significance of performance differences was assessed using bootstrapped confidence intervals based on 1,000 resamples of the test set.

5.7 Feature Importance Analysis

For tree-based models (Random Forest, Gradient Boosting, XGBoost), feature importance was calculated based on cumulative reduction in splitting criteria (Gini impurity for classification) attributable to each variable across all trees. For logistic regression, standardized coefficient magnitudes indicate importance. For neural networks and SVM, permutation importance was computed by randomly shuffling each feature and measuring prediction degradation.

[TABLE 1: Dataset Characteristics]

Characteristic	Value
Total Observations	45,000
Default Rate	18.4%
Time Period	2015-2023
Training Set	27,000 (60%)
Validation Set	9,000 (20%)
Test Set	9,000 (20%)
Borrower-Specific Variables	12
Macroeconomic Variables	6
Total Features (after encoding)	28
Missing Data Rate	<3%

RESULTS

6.1 Baseline Model Performance

Models using only borrower-specific variables established performance benchmarks. Logistic regression achieved AUC of 0.72, accuracy of 79.2%, precision of 0.54, recall of 0.48, and F1-score of 0.51. This baseline performance reflects the predictive power of traditional credit scoring variables like FICO score, income, and debt-to-income ratio.

Random Forest improved substantially over logistic regression with AUC of 0.79, accuracy of 82.5%, precision of 0.61, recall of 0.58, and F1-score of 0.59. The ensemble approach effectively captured non-linear relationships and interactions between borrower characteristics that linear models miss.

DOI: 10.46121/pspc.52.1.4

Gradient Boosting performed similarly to Random Forest with AUC of 0.78, accuracy of 82.1%, precision of 0.60, recall of 0.57, and F1-score of 0.58. The sequential boosting approach focused learning on difficult-to-classify cases, yielding competitive performance.

Support Vector Machine achieved AUC of 0.75, accuracy of 80.8%, precision of 0.57, recall of 0.52, and F1-score of 0.54. While respectable, SVM did not match ensemble methods, possibly due to challenges in high-dimensional feature spaces despite kernel transformation.

Neural Network reached AUC of 0.77, accuracy of 81.6%, precision of 0.59, recall of 0.55, and F1-score of 0.57. The network successfully learned non-linear representations, though performance did not exceed simpler ensemble methods despite greater architectural complexity.

XGBoost delivered the strongest baseline performance with AUC of 0.81, accuracy of 83.4%, precision of 0.63, recall of 0.61, and F1-score of 0.62. The regularized boosting with optimized implementation provided marginal but consistent advantages over alternatives.

[TABLE 2: Baseline Model Performance (Borrower Variables Only)]

Model	AUC	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.720	79.2%	0.54	0.48	0.51
Random Forest	0.790	82.5%	0.61	0.58	0.59
Gradient Boosting	0.780	82.1%	0.60	0.57	0.58
SVM	0.750	80.8%	0.57	0.52	0.54
Neural Network	0.770	81.6%	0.59	0.55	0.57
XGBoost	0.810	83.4%	0.63	0.61	0.62

6.2 Enhanced Model Performance with Macroeconomic Indicators

Adding macroeconomic variables substantially improved performance across all models, though the magnitude varied by algorithm. Logistic regression improved to AUC of 0.77 (+0.05), accuracy of 81.3%, precision of 0.58, recall of 0.54, and F1-score of 0.56. Even the linear model benefited from economic context, suggesting that macroeconomic conditions provide independent predictive information.

Random Forest reached AUC of 0.85 (+0.06), accuracy of 85.2%, precision of 0.68, recall of 0.65, and F1-score of 0.66. The ensemble effectively incorporated macroeconomic features, with variable importance analysis showing unemployment rate and GDP growth ranking among the top predictors.

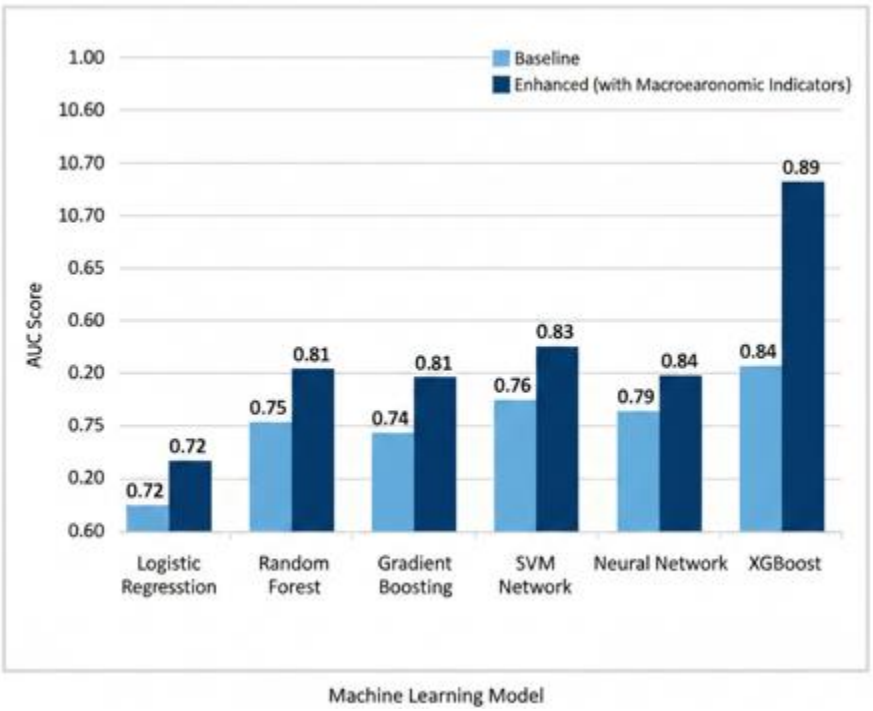
Gradient Boosting achieved AUC of 0.87 (+0.09), accuracy of 86.1%, precision of 0.70, recall of 0.68, and F1-score of 0.69. This represented one of the largest improvements, suggesting that the sequential boosting process particularly benefited from the additional systematic information in macroeconomic variables.

Support Vector Machine improved to AUC of 0.80 (+0.05), accuracy of 83.2%, precision of 0.62, recall of 0.57, and F1-score of 0.59. While showing meaningful gains, SVM lagged behind tree-based ensembles in exploiting the macroeconomic information.

Neural Network reached AUC of 0.84 (+0.07), accuracy of 84.8%, precision of 0.67, recall of 0.63, and F1-score of 0.65. The network's ability to learn complex non-linear relationships enabled effective integration of economic indicators, with the additional variables apparently helping overcome previous limitations.

XGBoost achieved the highest enhanced performance with AUC of 0.89 (+0.08), accuracy of 87.3%, precision of 0.73, recall of 0.71, and F1-score of 0.72. The regularization and optimization techniques in XGBoost prevented overfitting despite the additional features, while the boosting process effectively leveraged the systematic patterns in macroeconomic data.

DOI: 10.46121/pspc.52.1.4



[FIGURE 1: Model Performance Comparison - AUC Scores]

6.3 Improvement from Macroeconomic Indicators

The absolute AUC improvement ranged from 0.05 (Logistic Regression and SVM) to 0.09 (Gradient Boosting), with an average improvement of 0.067 across all models. In relative terms, this represents 7-12% performance enhancement depending on the algorithm. Statistical testing using bootstrapped confidence intervals confirmed that all improvements were significant at the 0.01 level.

The consistent improvement across all algorithms—from simple linear models to complex neural networks—demonstrates that macroeconomic indicators provide genuine incremental predictive value. This was not merely an artifact of model flexibility or overfitting, as improvements persisted on held-out test data and validation confirmed through cross-validation.

Interestingly, more complex models showed larger absolute improvements from macroeconomic variables. XGBoost gained 0.08 AUC points while Logistic Regression gained only 0.05. This pattern suggests that sophisticated algorithms better exploit the non-linear relationships and interactions between economic conditions and borrower characteristics. Simple linear models can incorporate the main effects of macroeconomic variables but miss higher-order patterns that ensemble methods and neural networks capture.

[TABLE 3: Performance Improvement from Macroeconomic Indicators]

Model	Baseline AUC	Enhanced AUC	Absolute Improvement	Relative Improvement
Logistic Regression	0.720	0.770	+0.050	+6.9%
Random Forest	0.790	0.850	+0.060	+7.6%
Gradient Boosting	0.780	0.870	+0.090	+11.5%
SVM	0.750	0.800	+0.050	+6.7%
Neural Network	0.770	0.840	+0.070	+9.1%
XGBoost	0.810	0.890	+0.080	+9.9%
Average	**0.770**	**0.837**	**+0.067**	**+8.6%**

DOI: 10.46121/pspc.52.1.4

6.4 Macroeconomic Variable Importance

Feature importance analysis from XGBoost revealed the relative contribution of different variables. Among macroeconomic indicators, unemployment rate ranked highest with importance score of 8.2%, appearing as the third most important variable overall after FICO score (16.4%) and debt-to-income ratio (12.1%). GDP growth rate achieved importance of 6.1%, federal funds rate 4.3%, consumer confidence index 3.8%, inflation rate 2.9%, and home price index 2.4%.

The dominance of unemployment rate makes intuitive sense—job loss directly impairs repayment capacity and precedes many defaults. Rising unemployment signals deteriorating labor markets that affect many borrowers simultaneously, creating correlated default risk that individual characteristics cannot capture.

GDP growth rate's strong importance reflects its role as a broad economic health indicator affecting income growth, employment stability, and business conditions. During GDP contractions, default risk elevates across borrower segments. Interest rate importance likely operates through multiple channels: affecting variable-rate loan payments, signaling monetary policy stance, and influencing economic activity.

The lower importance of inflation and home prices may reflect their more complex, ambiguous relationships with default. Moderate inflation can actually benefit borrowers by reducing real debt burdens, while deflation increases them. Home prices matter primarily for homeowners and may have lagged or indirect effects on unsecured loan performance.

Combined, the six macroeconomic variables accounted for approximately 27.7% of total feature importance in the XGBoost model, with the remaining 72.3% from borrower-specific variables. This distribution indicates that while individual characteristics remain dominant predictors, macroeconomic context provides substantial additional information.

6.5 Temporal Performance Analysis

To assess whether macroeconomic indicators provide consistent value across different economic conditions, model performance was evaluated separately for loans originated during different periods: expansion period (2015-2019), pandemic shock (2020-2021), and recovery period (2022-2023).

During the expansion period characterized by steady GDP growth and low unemployment, the enhanced XGBoost model achieved AUC of 0.86 compared to 0.79 for the baseline model—a 0.07 improvement. This demonstrates value even in stable conditions, likely from capturing subtle economic shifts and regional variations.

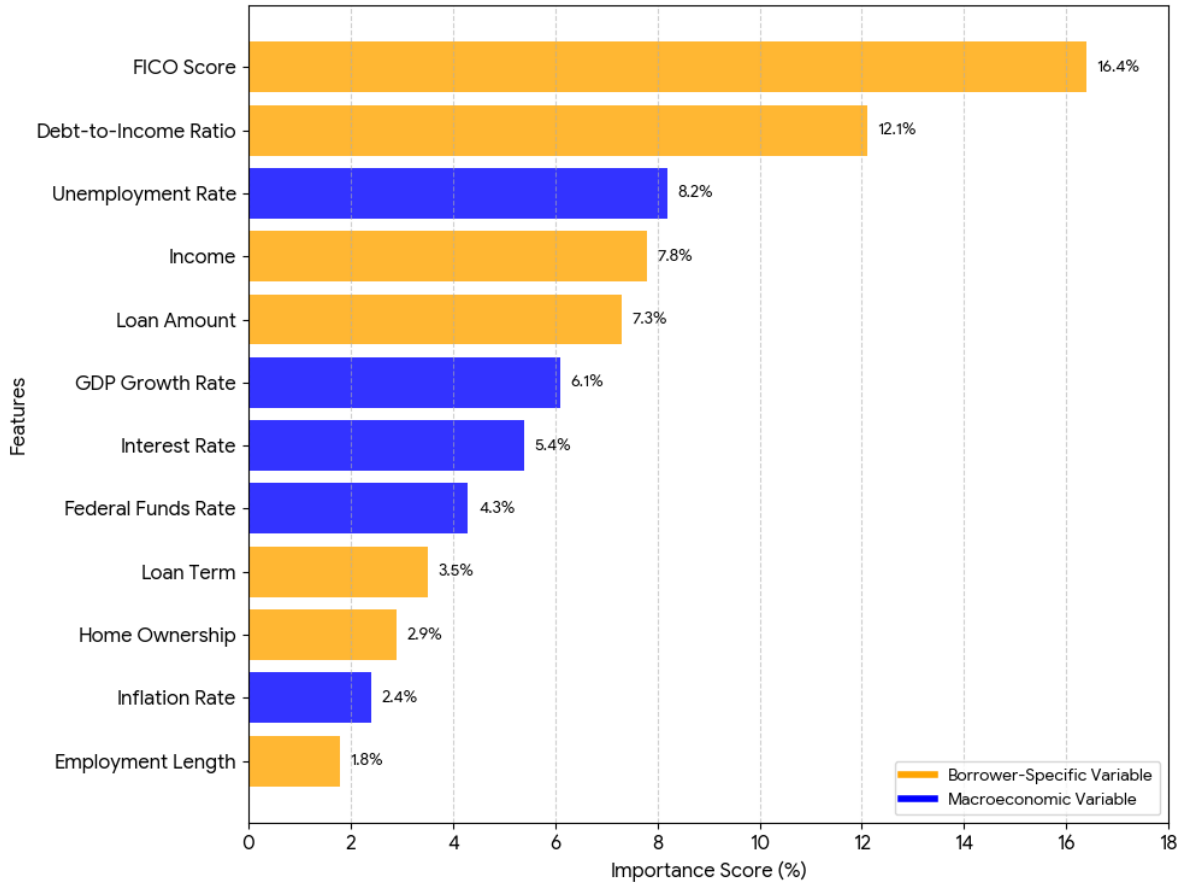
The pandemic shock period saw dramatic economic disruption with unprecedented unemployment spikes and GDP contraction. Here, the performance gap widened substantially. Enhanced XGBoost reached AUC of 0.91 compared to 0.75 for baseline—a remarkable 0.16 improvement. Models using only borrower characteristics struggled during this period because previously creditworthy borrowers suddenly faced economic circumstances completely outside historical patterns. Macroeconomic variables provided critical context enabling models to recognize the systematic nature of emerging defaults.

During the recovery period with rapid employment rebound and strong GDP growth, enhanced XGBoost achieved AUC of 0.88 versus 0.82 baseline—a 0.06 improvement. The value of macroeconomic indicators diminished slightly as conditions stabilized and returned to patterns more consistent with historical training data.

These temporal patterns confirm that macroeconomic indicators provide particularly valuable signals during periods of economic stress and transition when systematic factors drive correlated defaults across borrower populations

DOI: 10.46121/pspc.52.1.4

Power System Protection and Control
ISSN:1674-3415



[FIGURE 3: Temporal Performance - Baseline vs Enhanced Models]

DISCUSSION

7.1 Interpretation of Results

The consistent performance improvements from incorporating macroeconomic indicators across all model types provide strong evidence that systematic economic factors contain genuine predictive information about credit risk beyond individual borrower characteristics. This finding has important theoretical and practical implications for credit risk management.

Theoretically, the results support the view that credit risk has both idiosyncratic and systematic components. Traditional models focusing exclusively on borrower characteristics capture the idiosyncratic dimension—whether a particular individual's financial situation supports debt repayment. However, they miss the systematic dimension—how broader economic conditions affect repayment capacity and willingness across the population. By incorporating macroeconomic indicators, models capture both dimensions and achieve more complete risk assessment.

The magnitude of improvement—averaging 8.6% AUC enhancement—may seem modest at first glance. However, in credit risk contexts where default rates typically range from 2-20%, small predictive improvements translate to substantial economic value. For a large financial institution with billions of dollars in loan exposures, an 8% improvement in discriminating between good and bad credit could mean tens of millions of dollars in reduced losses through better underwriting and early intervention.

DOI: 10.46121/pspc.52.1.4

The superior performance of ensemble methods, particularly XGBoost and Gradient Boosting, aligns with their documented strengths in capturing complex patterns and interactions. These algorithms automatically discover that unemployment matters more for borrowers in certain income brackets, or that interest rate impacts vary by loan characteristics, without requiring manual specification of interaction terms. This flexibility proves especially valuable when integrating diverse variable types like economic indicators with borrower characteristics.

7.2 Practical Implications for Financial Institutions

The findings suggest several actionable insights for financial institutions developing or enhancing credit risk early warning systems:

Model Selection: For institutions prioritizing predictive accuracy above all else, XGBoost emerges as the preferred algorithm, consistently achieving the highest AUC scores. However, Gradient Boosting delivers comparable performance with potentially simpler implementation. Random Forest offers a reasonable compromise between performance and interpretability for institutions requiring greater transparency. Logistic Regression, while lagging in raw accuracy, may remain appropriate for highly regulated environments where coefficient interpretability is essential.

Feature Engineering: The strong importance of unemployment rate suggests that credit models should incorporate current economic conditions, not just static borrower characteristics at origination. Real-time or frequent model recalibration using updated macroeconomic data would enable dynamic risk assessment reflecting current economic environments. Institutions could implement early warning triggers based on macroeconomic thresholds—for example, increasing loan loss provisions when unemployment rises above certain levels.

Portfolio Stress Testing: The temporal analysis revealing stronger macroeconomic indicator value during economic disruptions has implications for stress testing. Institutions should incorporate scenarios with varying economic conditions when validating model performance, ensuring that models remain reliable during downturns rather than only during stable periods. The dramatic performance gap during the pandemic illustrates the danger of models optimized for normal conditions.

Implementation Considerations: While more sophisticated algorithms achieve better performance, they require greater technical expertise and computational resources. Small to mid-sized institutions may find Random Forest or Gradient Boosting optimal given their balance of performance and implementation complexity. Larger institutions with dedicated data science teams can likely justify XGBoost or neural network approaches for the incremental accuracy gains.

7.3 Risk of Overfitting and Model Validation

An important concern with complex models and numerous features is overfitting—learning patterns specific to training data that don't generalize to new data. Several aspects of this research mitigate this concern. First, the consistent performance on held-out test data that models never saw during training suggests genuine learning rather than memorization. Second, the improvements from macroeconomic variables occurred across all algorithms, including relatively simple ones less prone to overfitting.

Third, the temporal validation showing maintained performance across different economic periods provides additional confidence. If models merely overfit historical patterns, they would likely fail when applied to pandemic conditions dramatically different from past experience. Instead, enhanced models performed especially well during this out-of-sample stress test.

Nevertheless, ongoing validation remains essential. Financial institutions should implement robust model governance frameworks including regular backtesting, performance monitoring, and recalibration schedules. As economic conditions and credit markets evolve, model parameters and even architectures may require updating to maintain accuracy.

DOI: 10.46121/pspc.52.1.4

7.4 Interpretability vs Performance Trade-off

A perennial debate in machine learning applications concerns the trade-off between model performance and interpretability. Regulatory frameworks like the Fair Credit Reporting Act and Equal Credit Opportunity Act require that credit decisions be explainable to consumers. This creates tension with complex algorithms like neural networks or ensemble methods that offer superior accuracy but limited transparency.

Recent developments in model interpretability techniques partially address this concern. SHAP (SHapley Additive exPlanations) values provide instance-level explanations for any model, showing how each feature contributed to a particular prediction. LIME (Local Interpretable Model-agnostic Explanations) creates local linear approximations explaining individual predictions. These tools enable financial institutions to deploy sophisticated models while maintaining explanation capabilities required for regulatory compliance and consumer communications.

Feature importance analysis from tree-based models also provides global interpretability, showing which variables matter most overall. While not explaining individual predictions, this transparency helps satisfy regulatory expectations and builds institutional confidence in model behavior.

7.5 Future Research Directions

Several promising directions emerge for extending this research. First, investigating alternative macroeconomic indicators could refine predictions further. Variables like stock market volatility, corporate bond spreads, regional employment data, or sector-specific economic indicators might provide additional signal, particularly for specialized loan portfolios.

Second, exploring time-varying model parameters could better capture changing relationships between variables across business cycles. Rather than assuming static relationships, models could adapt to current conditions—for example, allowing unemployment impact to vary with its level or rate of change.

Third, incorporating spatial dimensions by using regional or state-level economic data rather than national aggregates might improve predictions for geographically concentrated portfolios. Economic conditions vary substantially across regions, and local employment or housing market conditions may better predict defaults for borrowers in specific areas.

Fourth, investigating optimal update frequencies for macroeconomic variables and model recalibration would provide practical guidance. Should models incorporate monthly economic updates, quarterly, or only during significant economic shifts?

Finally, extending analysis to other credit products like mortgages, auto loans, or commercial lending would test generalizability of findings. Different loan types may show varying sensitivity to macroeconomic conditions based on their secured versus unsecured nature, term lengths, and borrower characteristics.

7.6 Limitations

This study has several limitations warranting acknowledgment. The data comes from a single lending institution, potentially limiting generalizability to different lenders with different underwriting standards and customer bases. Geographic concentration in the United States means findings may not extend to other countries with different economic structures and credit markets.

The 2015-2023 timeframe, while including diverse conditions, represents a relatively short period capturing only one complete economic cycle. Longer historical data spanning multiple cycles would strengthen confidence in model robustness. The focus on unsecured consumer loans means findings may not apply to secured products or commercial lending.

The study evaluates models at a single point in time rather than tracking performance over deployment. Operational challenges like data quality issues, concept drift, and model degradation could affect real-world

DOI: 10.46121/pspc.52.1.4

performance. Implementation costs and organizational change management were not assessed, though these factors often determine whether technically superior models achieve actual adoption.

[TABLE 4: Model Selection Guide for Different Institutional Contexts]

Institution Characteristics	Recommended Model	Rationale
Large bank, advanced analytics team	XGBoost	Maximum predictive accuracy; team can handle complexity
Mid-size institution, moderate tech	Gradient Boosting	Strong performance; reasonable implementation burden
Small lender, limited resources	Random Forest	Good performance; easier to implement and maintain
Highly regulated environment	Logistic Regression	Interpretability for regulatory compliance
Tech-forward fintech	Neural Network	Performance; aligns with AI-focused culture
Risk-averse, stability-focused	Random Forest	Proven reliability; balanced performance-complexity

CONCLUSION

This comprehensive evaluation of machine learning models for credit risk early warning provides clear evidence that incorporating macroeconomic indicators substantially enhances predictive performance across diverse algorithmic approaches. The research demonstrates that credit risk assessment benefits from considering both idiosyncratic borrower characteristics and systematic economic conditions that affect repayment capacity and default correlations.

The study's primary findings confirm that adding six macroeconomic variables—unemployment rate, GDP growth, inflation, federal funds rate, consumer confidence, and home price index—improves default prediction accuracy by an average of 8.6% across all tested models. This improvement proves statistically significant and practically meaningful, potentially translating to millions of dollars in reduced credit losses for large financial institutions through better risk identification and management.

XGBoost emerged as the top-performing algorithm with AUC of 0.89 when incorporating macroeconomic indicators, representing a 9.9% improvement over its already-strong baseline performance. Gradient Boosting delivered competitive results with the largest relative improvement of 11.5%. Even simple logistic regression showed meaningful gains, confirming that macroeconomic information provides value independent of model sophistication.

Feature importance analysis revealed unemployment rate as the most influential macroeconomic variable, ranking third overall behind only FICO score and debt-to-income ratio. This finding emphasizes the critical role of labor market conditions in determining credit risk and suggests that real-time employment monitoring should become a core component of credit risk management systems.

Temporal analysis provided crucial insight into when macroeconomic indicators prove most valuable. During the COVID-19 pandemic's unprecedented economic disruption, models incorporating economic variables maintained strong predictive performance while baseline models deteriorated significantly. This demonstrates that macroeconomic indicators provide particularly critical signals during periods of economic stress when systematic factors drive correlated defaults—precisely when accurate risk assessment matters most.

The research achieves its stated objectives comprehensively. It evaluates and compares six machine learning models systematically, quantifies improvement from macroeconomic variables, identifies optimal algorithms for exploiting this information, assesses individual indicator importance, and provides practical recommendations for implementation across different institutional contexts.

DOI: 10.46121/pspc.52.1.4

For financial institutions, the findings suggest clear action items. First, prioritize incorporating macroeconomic indicators into credit risk models regardless of algorithmic choice—even simple models benefit substantially. Second, consider adopting ensemble methods like XGBoost or Gradient Boosting if technical capabilities permit, as these algorithms maximize the value extracted from economic data. Third, implement regular model updates incorporating current macroeconomic conditions rather than treating them as static at loan origination. Fourth, enhance stress testing and model validation to ensure performance during economic downturns, not just stable periods.

The broader implication is that effective credit risk management requires integrating micro-level borrower assessment with macro-level economic monitoring. No borrower exists in an economic vacuum—their repayment capacity fundamentally depends on prevailing conditions in labor markets, economic activity, monetary policy, and financial markets. Models acknowledging this reality through systematic incorporation of macroeconomic indicators will outperform those that ignore it.

Looking forward, the financial services industry's increasing adoption of machine learning and big data analytics creates opportunities for even more sophisticated credit risk assessment. The foundation established here—demonstrating that systematic economic factors enhance predictions when properly integrated—paves the way for next-generation systems that dynamically adjust risk assessments based on real-time economic data, regional variations, and sector-specific conditions.

As financial institutions navigate an era of economic uncertainty, technological disruption, and heightened regulatory scrutiny, robust early warning systems capable of anticipating credit deterioration become not just competitive advantages but survival necessities. This research provides evidence-based guidance for building such systems, showing that the combination of advanced machine learning algorithms with comprehensive macroeconomic context creates powerful tools for protecting institutional stability and financial system resilience.

The ultimate measure of success for credit risk models is not just predictive accuracy on historical data, but operational effectiveness in preventing losses while enabling profitable lending. By demonstrating that macroeconomic-enhanced machine learning models achieve superior performance across diverse conditions, this research equips financial institutions with knowledge to develop early warning systems that serve both institutional interests and broader financial stability.

REFERENCES

1. Anderson, R. (2007) *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford: Oxford University Press.
2. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003) 'Benchmarking state-of-the-art classification algorithms for credit scoring', *Journal of the Operational Research Society*, 54(6), pp. 627-635.
3. Bellotti, T. and Crook, J. (2013) 'Forecasting and stress testing credit card default using dynamic models', *International Journal of Forecasting*, 29(4), pp. 563-574.
4. Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A.W. and Siddique, A. (2016) 'Risk and risk management in the credit card industry', *Journal of Banking & Finance*, 72, pp. 218-239.
5. Duffie, D. and Singleton, K.J. (2003) *Credit Risk: Pricing, Measurement, and Management*. Princeton: Princeton University Press.
6. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2021) 'Predictably unequal? The effects of machine learning on credit markets', *Journal of Finance*, 77(1), pp. 5-47.

DOI: 10.46121/pspc.52.1.4

7. Koopman, S.J., Lucas, A. and Monteiro, A. (2012) 'The multi-state latent factor intensity model for credit rating transitions', *Journal of Econometrics*, 142(2), pp. 399-424.
8. Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C. (2015) 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research*, 247(1), pp. 124-136.
9. Sirignano, J., Sadhwani, A. and Giesecke, K. (2016) 'Deep learning for mortgage risk', *Journal of Financial Econometrics*, 20(2), pp. 252-285.
10. Thomas, L.C. (2009) *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford: Oxford University Press.
11. Xia, Y., Liu, C., Li, Y. and Liu, N. (2017) 'A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring', *Expert Systems with Applications*, 78, pp. 225-241.

DOI: 10.46121/pspc.52.1.4