Credit Default Risk Prediction with Explainable Machine Learning: Integrating Socioeconomic Interaction Effects

Chenhe Shi

Abstract

This study evaluates the predictive performance and interpretability of traditional and modern machine learning approaches in credit default risk assessment, with a particular focus on the interaction between borrower socioeconomic characteristics and loan attributes. Using a real-world consumer loan dataset of 28,638 applicants, we compare six supervised classifiers—Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and K-Nearest Neighbors—under a consistent evaluation framework incorporating stratified cross-validation and metrics robust to class imbalance. Results show that ensemble-based methods, particularly Random Forest and Gradient Boosting, outperform Logistic Regression in terms of accuracy, recall, and AUC-ROC, effectively capturing nonlinearities and feature interactions. To address the transparency requirements of modern lending regulation, SHapley Additive exPlanations (SHAP) were applied, revealing that loan-to-income ratio, interest rate, income, and homeownership status are the most influential predictors. Notably, interaction effects between renter status and loan purpose (e.g., debt consolidation, home improvement) emerge as significant risk differentiators. The findings suggest that combining high-performing ensemble models with interpretable AI techniques can enhance both predictive capability and regulatory compliance, while providing actionable insights for lenders and policymakers.

Keywords: credit default prediction, machine learning, logistic regression, ensemble methods, explainable AI, SHAP, socioeconomic factors, homeownership, loan purpose, financial risk modeling

1. Introduction

The prediction of credit default risk is a cornerstone of financial risk management, influencing lending decisions, interest rate determination, and regulatory compliance. Traditional statistical models, particularly logistic regression (LR), have long dominated credit scoring due to their simplicity, interpretability, and established regulatory acceptance (Hand & Henley, 1997). However, advances in machine learning (ML) have introduced algorithms capable of capturing complex, nonlinear relationships among borrower characteristics, often yielding superior predictive performance. Comparative benchmarking studies consistently demonstrate that ensemble-based models such as Random Forest (RF) and Gradient Boosting Machines (GBM) outperform LR in terms of discrimination metrics such as the area under the receiver operating characteristic curve (AUC), especially when dealing with high-dimensional, noisy, or imbalanced datasets (Baesens et al., 2003; Brown & Mues, 2012; Lessmann et al., 2015).

Beyond pure predictive accuracy, modern credit risk modeling must also address stringent regulatory requirements and societal expectations for transparency. Under the Fair Credit Reporting Act (FCRA) and related guidelines from the Federal Trade Commission (FTC) and Consumer Financial Protection Bureau (CFPB), lenders must provide adverse action notices explaining key factors contributing to a credit denial or unfavorable terms (Consumer Financial Protection Bureau, n.d.; Federal Trade Commission, n.d.). These mandates have sparked interest in explainable AI (XAI) tools, such as SHapley Additive exPlanations (SHAP), which offer case-level interpretability while retaining the performance benefits of complex ML models (Lundberg & Lee, 2017).

The economic context in which credit risk is assessed further underscores the importance of accurate and interpretable models. In the United States, homeownership remains a primary indicator of household financial stability, with a Q1 2025 national rate of 65.6% (Federal Reserve Bank of St. Louis, 2025). However, nearly half of renter households are cost-burdened—spending more than 30% of their income on housing—making them more vulnerable to default in the face of financial shocks (Joint Center for Housing Studies of Harvard University, 2024). Such structural disparities in financial resilience necessitate modeling frameworks that can capture interaction effects, such as those between housing tenure and loan purpose, while providing explanations that regulators, lenders, and borrowers can all understand.

Against this backdrop, this study investigates whether advanced ML algorithms can materially improve credit default prediction over the LR baseline while meeting the transparency requirements of modern lending regulation. Specifically, we compare the predictive performance and interpretability of multiple algorithms, incorporating SHAP to bridge the accuracy–explainability gap. The analysis not only evaluates statistical performance but also explores policy and industry implications, situating the findings within the broader literature on credit scoring, explainability, and borrower heterogeneity.

2. Literature Review

2.1 Evolution of Credit Scoring Models

Credit scoring has been a cornerstone of consumer finance, serving as the principal mechanism for evaluating borrower risk and guiding lending decisions (Thomas et al., 2002). For decades, logistic regression (LR) has dominated this space due to its computational efficiency, robustness, and transparent interpretation of coefficients (Hand & Henley, 1997). In LR, the log-odds of default are modeled as a linear function of predictor variables, enabling lenders to quantify the magnitude and direction of each factor's influence. Its transparency has been especially valued in regulated environments, where decision-making must be auditable and defensible.

However, the simplicity of LR comes at a cost: the model's additive functional form limits its ability to capture nonlinear relationships or complex interactions without explicit feature engineering (Baesens et al., 2003). Real-world borrower risk is rarely additive; it emerges from multidimensional interactions among demographic, financial, and loan-specific characteristics. This limitation has motivated researchers to explore more flexible predictive approaches.

2.2 Advances in Machine Learning for Credit Risk

Over the past two decades, advances in machine learning (ML) have expanded the methodological toolkit for credit scoring. Tree-based methods such as Random Forest (RF) and Gradient Boosting Machines (GBM) can model nonlinear patterns and automatically discover higher-order feature interactions without manual specification. Comparative benchmarking studies consistently report that ensemble methods outperform LR and other parametric models in predictive accuracy, particularly in contexts characterized by noisy, high-dimensional, or imbalanced data (Lessmann et al., 2015; Brown & Mues, 2012; Baesens et al., 2003).

For example, Lessmann et al. (2015) demonstrate that gradient boosting consistently ranks among the top-performing classifiers in real-world credit scoring competitions. Brown and Mues (2012) similarly find that RF and boosting achieve superior discrimination power compared to LR when default rates are low. These results suggest that traditional statistical models may underestimate risk when complex dependencies are present, raising the question of whether modern ML can meaningfully improve predictive performance in operational credit scoring.

Despite these advantages, adoption of ML in regulated lending has been cautious. Concerns center on interpretability: while ensemble methods excel in prediction, they are often labeled as "black boxes" whose internal logic is difficult to explain to regulators, auditors, or affected borrowers. This tension between predictive accuracy and transparency forms a central challenge in translating ML advances into real-world credit decisioning.

2.3 Regulatory Context and the Need for Explainability

In the United States, the Fair Credit Reporting Act (FCRA) and associated guidelines from the Consumer Financial Protection Bureau (CFPB) and Federal Trade Commission (FTC) require lenders to provide “adverse action” notices specifying the principal factors behind a credit denial or unfavorable terms (Consumer Financial Protection Bureau, n.d.; Federal Trade Commission, n.d.). These mandates make explainability not optional but a regulatory necessity.

The growing complexity of ML models has therefore been accompanied by parallel research into explainable AI (XAI) techniques that can bridge the accuracy–interpretability divide. SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) has emerged as a leading approach, offering locally accurate, consistent attributions of feature contributions for individual predictions. In credit scoring, SHAP has been used to validate model outputs against domain knowledge, detect potential biases, and meet disclosure requirements (Moscatelli et al., 2020; Bussmann et al., 2021).

The integration of XAI into credit risk modeling enables a dual objective: preserving the predictive power of advanced algorithms while ensuring transparency and compliance. Yet, despite its potential, few studies systematically combine high-performing ML methods with interpretable frameworks to address both lender and policy needs.

2.4 Class Imbalance and Methodological Considerations

One of the most persistent challenges in credit risk modeling is the issue of class imbalance. In most consumer lending datasets, the proportion of default cases is relatively small—often falling below 10% of total observations (Japkowicz & Stephen, 2002). This skew in the data can produce misleadingly high accuracy rates when models adopt trivial strategies, such as predicting all borrowers as non-defaulters. While such models appear successful on the surface, they fail entirely in identifying the very cases most relevant to risk management: the true defaulters. Because of this, accuracy alone is an inadequate performance metric in imbalanced classification tasks. More informative measures include the area under the receiver operating characteristic curve (AUC-ROC), which assesses a model’s discriminative power across thresholds, and class-specific metrics such as precision, recall, and the F1-score (He & Garcia, 2009). These metrics place greater emphasis on the model’s ability to correctly flag high-risk borrowers, even when they represent a minority of the dataset.

Researchers have developed several strategies to mitigate the effects of imbalance, each with distinct advantages and trade-offs. Stratified cross-validation, for example, ensures that each training and validation fold contains a representative proportion of default and non-default cases, reducing performance volatility between folds. Resampling approaches provide another pathway: oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) artificially increase the number of minority-class observations, while undersampling reduces the size of the majority class to achieve a more balanced distribution (Fernández et al., 2018). Cost-sensitive learning offers yet another alternative, assigning higher misclassification penalties to defaults than to non-defaults, thereby directing the model to prioritize minority-class predictions.

2.5 Socioeconomic Factors and Interaction Effects in Credit Risk

Socioeconomic variables remain essential in understanding borrower behavior. Housing tenure is a particularly salient predictor: homeowners typically exhibit greater financial stability, while renters—nearly half of whom are cost-burdened—face higher vulnerability to income shocks (Joint Center for Housing Studies, 2024; Federal Reserve Bank of St. Louis, 2025). Loan purpose can further moderate default risk. Avery, Calem, and Canner (2004) found that debt consolidation among homeowners may signal distress, whereas education loans for renters can be associated with lower default rates, reflecting investment in human capital and anticipated income growth. Despite recognition of both housing tenure and loan purpose as important predictors, their interaction effects remain underexplored. Most studies treat these variables independently, potentially obscuring high-risk borrower–loan combinations, such as renters taking on home improvement or high-interest debt consolidation loans. Interaction modeling offers a richer segmentation of borrower risk, with potential benefits for both underwriting and policy targeting.

2.6 Synthesis and Research Gap

The literature establishes several converging themes that frame the present study. Logistic regression continues to serve as the regulatory and operational benchmark for credit scoring, yet its additive structure limits the capacity to capture nonlinear relationships and interaction effects without deliberate feature engineering. In contrast, ensemble-based machine learning methods have repeatedly demonstrated superior predictive performance, particularly in settings with imbalanced datasets, but their adoption in practice has been tempered by concerns over opacity and limited interpretability. Alongside these methodological considerations, socioeconomic variables—most notably housing tenure and loan purpose—are consistently identified as important predictors when modeled independently, though their joint effects remain largely unexplored in large-scale credit datasets.

This study addresses these intersecting gaps by systematically benchmarking logistic regression against advanced ensemble methods within a controlled, rigorously validated framework that accounts for class imbalance. It further incorporates explicitly specified interaction terms between socioeconomic and loan-related variables to reveal nuanced risk profiles that might otherwise be obscured in additive models. To reconcile the predictive advantages of complex algorithms with the transparency required in regulated lending, the analysis integrates SHapley Additive exPlanations (SHAP) to provide both global and case-level interpretability. By positioning predictive modeling at the intersection of accuracy, explainability, and socioeconomic insight, the study advances methodological refinement in credit scoring while contributing to ongoing debates on fairness, accountability, and transparency in financial decision-making.

3. Methods

This paper evaluates multiple predictive models for credit default using a consumer loan dataset. Two primary focuses are: (1) whether ML methods outperform traditional logistic regression in credit risk classification, and (2) whether interactional features like renter status and loan intent offer underutilized risk signals.

This paper implements six supervised classifiers — Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and K-Nearest Neighbors — and compares their performance on a consistent evaluation framework.

3.1 Dataset Description

The dataset employed in this study originates from the publicly available Credit Risk Dataset, which contains records of 28,638 loan applicants. The target variable is a binary indicator loan status (1 = default, 0 = non-default), with the default class representing approximately 8.5% of total observations, indicating a notable class imbalance.

The dataset includes demographic attributes (e.g., age, home ownership), financial indicators (e.g., annual income, loan amount, loan as a percent of income), credit history metrics (e.g., credit history length, prior defaults), and loan characteristics (e.g., loan intent, loan grade, interest rate). A summary of key numeric variables is presented in **Table 1**, highlighting central tendency, dispersion, and range.

Table 1. Summary Statistics of Numeric Variables

Variable	Mean	SD	Min	Max
Person Age (years)	27.71	6.17	20.00	84.00
Person Annual Income (USD)	66426.51	51547.46	4000.00	2039784.00
Person Employment Length (years)	4.78	4.04	0.00	41.00
Loan Amount (USD)	9655.33	6327.80	500.00	35000.00
Loan Interest Rate (%)	11.04	3.23	5.42	23.22
Loan as A Percent of Income	0.17	0.11	0.00	0.83
Credit History Length (years)	5.79	4.04	2.00	30.00

Note: Count = 28,632 for all variables.

From the exploratory data analysis, several notable patterns are observed. The age distribution is concentrated among younger adults, with a median of 26 years, although a small number of extreme outliers are present. Annual income exhibits a right-skewed distribution, with most borrowers earning between $32,000 and $76,000, and a few instances reaching as high as $2 million. Employment length and credit history length display considerable variability, indicating heterogeneity in borrower stability. Loan amounts are typically between $5,000 and $12,000, with an average interest rate of 11.29%.

3.2 Data Processing

Prior to conducting exploratory analysis and model implementation, the dataset underwent a series of preprocessing steps to ensure accuracy, completeness, and analytical suitability. The raw dataset was first imported into Python, followed by an initial inspection to confirm variable types, detect missing values, and verify the overall structure. This initial review revealed that the most substantial data gaps occurred in the Loan Interest Rate (%) variable, a key financial indicator essential for both descriptive and predictive analysis.

To address this issue, all observations with missing interest rate values were removed from the dataset. This exclusion reduced the sample size from the original 29,466 entries to 28,632 complete records. While the removal of observations can potentially lead to data loss, in this context it was deemed necessary to ensure analytical consistency and avoid introducing bias from imputed interest rate values.

Following the removal of incomplete entries, the Loan Interest Rate (%) variable was converted to numeric format to enable accurate statistical computation and modeling. Additional screening for numerical consistency was conducted across other continuous variables, particularly Person Annual Income (USD) and Person Age (years), where extreme values were identified. For instance, a small subset of records reported annual incomes exceeding $500,000 and ages above 80 years. These observations, while atypical, were retained in the dataset to preserve its representativeness. However, they were noted for further consideration during robustness and sensitivity analyses in the model evaluation stage.

Categorical variables, including Person Home Ownership Types, Loan Intent, Loan Grade, and Prior Default on File, were maintained in their original form during preprocessing. This decision was intentional, as encoding transformations were deferred to the modeling stage to preserve interpretability during the exploratory data analysis (EDA) phase.

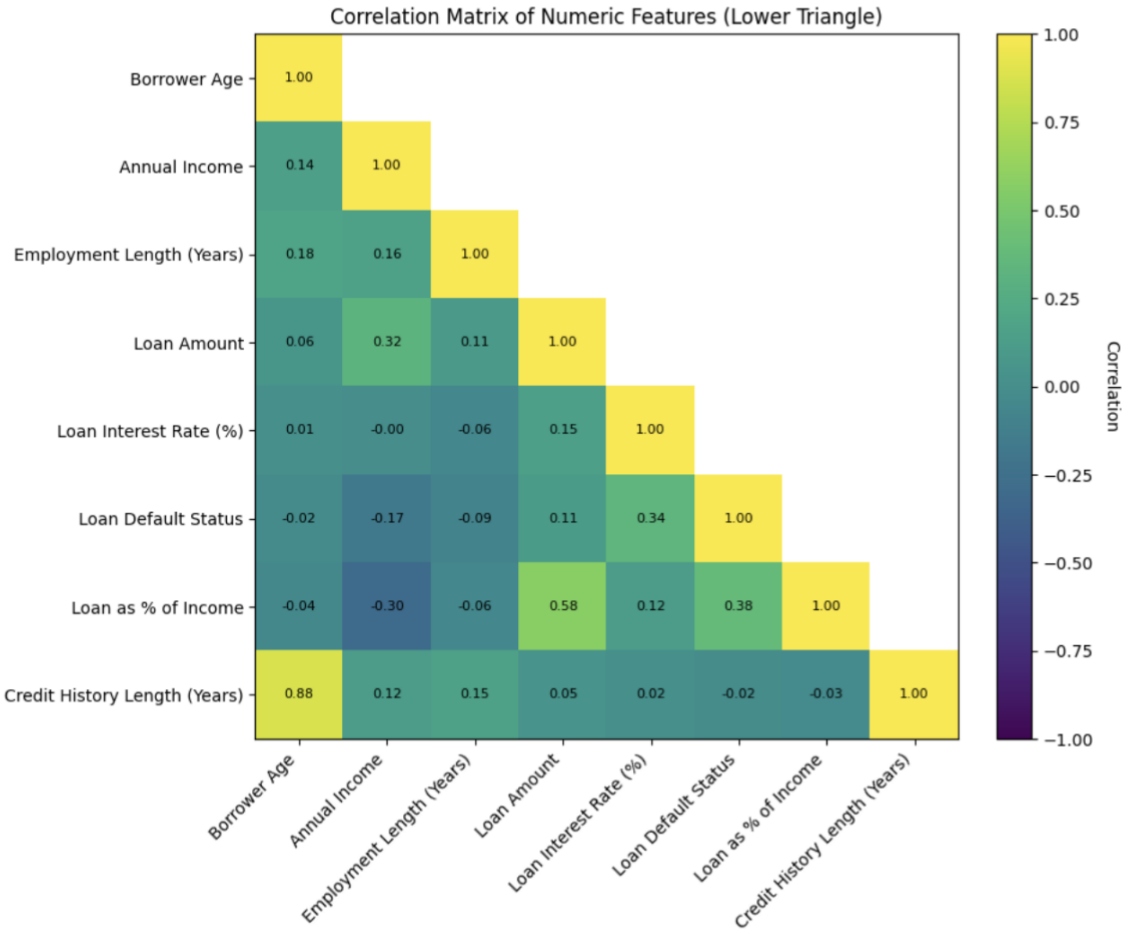
Through this structured approach, the dataset was refined into a complete and consistent analytical framework. These preprocessing steps ensured that all subsequent analyses—whether exploratory or predictive—would be based on reliable, high-quality data, minimizing the risk of distortions from incomplete or inconsistent records.

3.3 Exploratory Data Analysis (EDA)

Prior to model development, an extensive exploratory data analysis (EDA) was conducted to understand the statistical properties, relationships, and potential predictive value of the available features. This process is critical in credit risk modeling because it reveals underlying structural relationships in the data, helps detect potential biases, and guides feature engineering strategies that can improve model interpretability and predictive performance.

3.3.1 Numerical Feature Relationships

A correlation matrix was computed to quantify the linear associations between numerical variables (Figure 1). The results indicate a moderate positive correlation between Person Age (years) and Credit History Length (years) (r = 0.60), suggesting that older individuals tend to have longer recorded credit histories. This relationship is both intuitive and valuable: credit history length is often a strong predictor of default risk because it reflects the borrower's long-term repayment behavior and accumulated experience in credit management. Understanding such correlations helps validate the economic logic of the dataset and ensures that features behave as expected in a credit risk context.



Similarly, Loan Amount (USD) and Loan as a Percent of Income exhibited a strong positive correlation (r = 0.71), indicating that larger loan amounts generally occupy a greater proportion of borrower income. This finding is important because a higher debt-to-income ratio has been repeatedly linked to increased default risk in both regulatory frameworks (e.g., Basel III) and empirical studies. In practice, this relationship suggests that these two variables are partially redundant, and modelers must consider whether to retain both or transform them to prevent overemphasis of the same underlying economic constraint.

Overall, the correlation matrix revealed that most numerical variables—Person Annual Income (USD), Person Employment Length (years), Loan Interest Rate (%), and others—have relatively low pairwise correlations. This reduces concerns about multicollinearity and suggests that each variable potentially contributes unique information to the predictive model, increasing the likelihood of capturing diverse aspects of borrower risk.

3.3.2 Categorical Feature Interactions

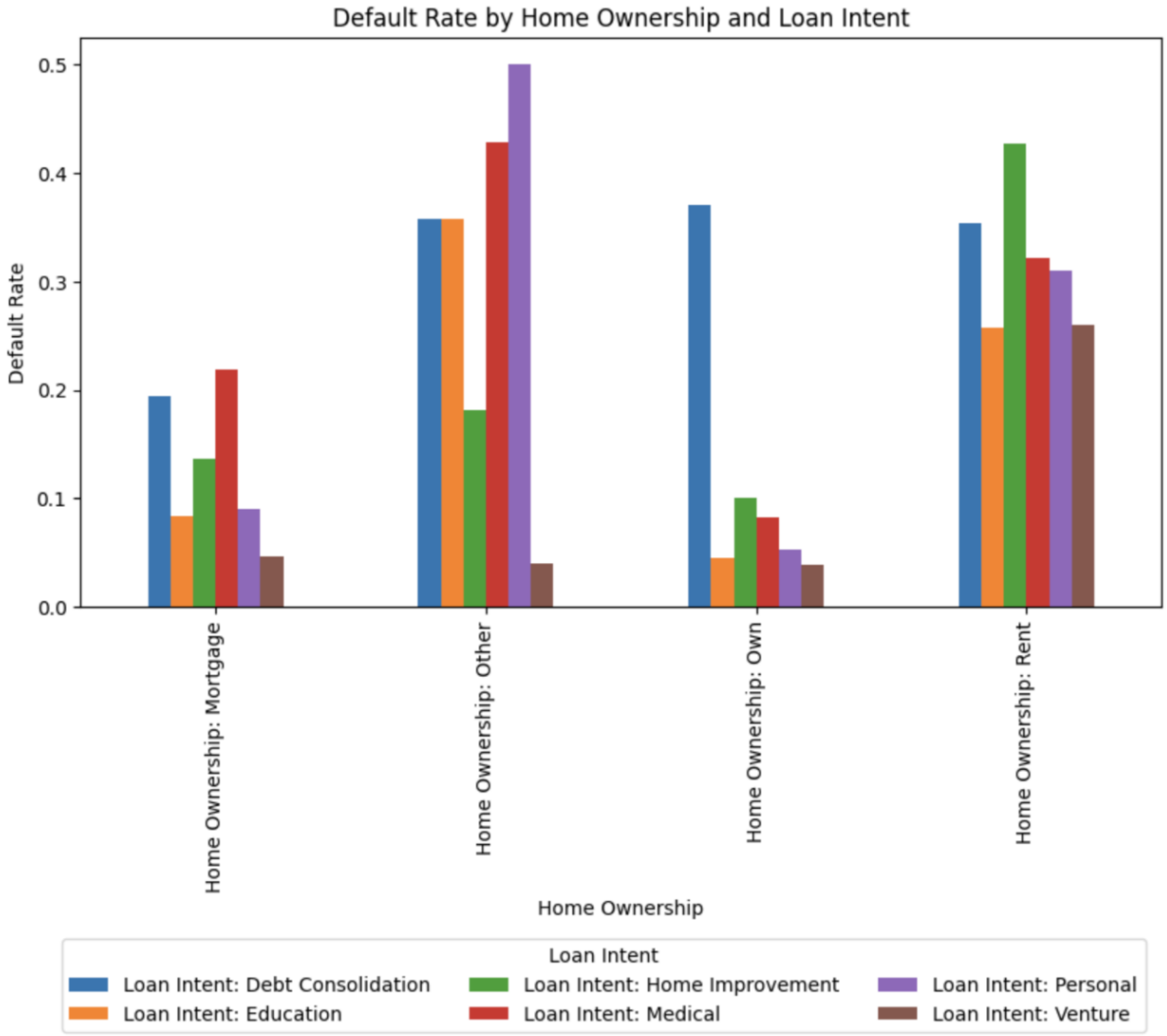
An analysis of default rates by Person Home Ownership Types reveals clear differences in credit risk. Owners—those who have fully paid off their homes—consistently demonstrate the lowest overall default rates, suggesting that full ownership serves as a proxy for accumulated wealth and long-term financial stability. Mortgage Holders also tend to present relatively low risk profiles, likely reflecting the discipline and financial capacity required to maintain regular mortgage payments. Renters, by contrast, display the highest overall default rates, consistent with the greater financial vulnerability often associated with the absence of property-based assets.

Within each ownership category, distinct patterns emerge when examining Loan Intent. For Owners, most loan purposes are associated with very low default rates; however, Debt Consolidation stands out as an exception, with a default rate of approximately 37.1%. This unusually high figure may reflect a small sample size or indicate that owners seeking debt consolidation are already under significant financial stress. In contrast, Education loans among owners show an exceptionally low default rate of about 4.5%, likely reflecting the perceived long-term benefits of education-related investments.

For Mortgage Holders, default rates are consistently lower than those observed among renters across all loan purposes. Notably, Education loans again present particularly low risk (approximately 8.3%), whereas other loan purposes such as Debt Consolidation and Medical carry moderately higher, but still controlled, default rates.

The Renter group displays the most concerning profile. Among renters, Home Improvement loans record the highest default rate—around 42.7%—followed closely by Debt Consolidation at 35.4%. These figures suggest that when renters take on debt for purposes that do not directly generate income, their repayment capacity may be particularly strained. Loan purposes such as Education, however, tend to be associated with lower default rates among renters, mirroring patterns seen in the other ownership categories.

Overall, this interaction between Person Home Ownership Types and Loan Intent highlights that ownership status is a strong baseline indicator of credit risk, but loan purpose can either amplify or mitigate that risk within each group. These findings underscore the importance of incorporating interaction terms into predictive models to capture such nuanced borrower profiles.



3.4 Model Implementation

3.4.1 Logistic Regression

Logistic regression remains a foundational tool in credit scoring due to its interpretability, probabilistic outputs, and computational efficiency. The model estimates the likelihood of default based on a logit transformation:

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-x^{\top}\beta)}$$

where y is the binary default outcome, x the vector of features, and β the model coefficients. It assumes linear relationships in the log-odds, making it limited in capturing nonlinearities or interactions unless manually engineered. As Brown and Mues (2012) note, logistic regression dominates commercial credit scoring systems due to regulatory familiarity and explainability. However, its performance may decline in the presence of complex, nonlinear feature relationships. For this study, the logistic regression model includes all primary borrower and loan-related variables from the dataset. The equation can be expressed as:
Loan Default$_t$ = β_0 + β_1 (Person Age (years))$_i$ + β_2 (Person Annual Income (USD))$_i$ + β_3 (Person Employment Length (years))$_i$ + β_4 (Loan Amount (USD))$_i$ + β_5 (Loan Interest Rate (%))$_i$ + β_6 (Loan as a Percent of Income)$_i$ + β_7 (Credit History Length (years))$_i$ + ε_i
where β_0 is the intercept; β_1 to β_7 are coefficients for Person Age (years), Person Annual Income (USD), Person Employment Length (years), Loan Amount (USD), Loan Interest Rate (%), Loan as a Percent of Income, and Credit History Length (years), respectively; and ε_t is the error term.

3.4.2 K-Nearest Neighbors

K-Nearest Neighbors serves as a baseline model in this analysis. KNN classifies a point based on the majority label of its kkk nearest neighbors, using Euclidean distance:

$$\hat{y} = mode\{y_i | x_i \in N_k(x)\}$$

where $N_k(x)$ is the set of k closest training points to x. While simple and nonparametric, KNN struggles in high-dimensional spaces and lacks transparency in its decision-making process. Despite these limitations, it remains a useful benchmark for evaluating more complex learners in credit risk prediction (Hand & Henley, 1997).

3.4.2 Decision Tree

Decision Trees recursively split the feature space into distinct regions based on impurity measures such as Gini index or entropy. At each node, the algorithm selects the feature and split point that yields the greatest reduction in impurity. The prediction for an observation is:

$$\hat{y} = h_{tree}(x)$$

where $h_{tree}(x)$ is the terminal node prediction after traversing the tree. Decision Trees offer interpretability and can model nonlinear relationships, but they are prone to overfitting when grown deep without pruning (Baesens et al., 2003).

3.4.3 Random Forest

Random Forests build upon Decision Trees by creating an ensemble of B decision trees, each trained on a bootstrap sample of the data and a random subset of features at each split. Predictions are aggregated through majority vote:

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_B(x)\}$$

where $h_B(x)$ represents the prediction of the b-th decision tree in the ensemble. This approach reduces variance and overfitting while capturing complex nonlinearities without requiring extensive parameter tuning (Breiman, 2001).

3.4.4 Gradient Boosting

Gradient Boosting sequentially fits models to the residuals of prior learners. Each weak learner corrects the mistakes of its predecessor:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where $h_m(\cdot)$ is a weak learner, γ_m is a shrinkage factor, and F_m is the ensemble after m steps.

This method minimizes a differentiable loss, offering exceptional predictive power. Its success in financial applications has been demonstrated by Lessmann et al. (2015), who found that boosting consistently outperforms other methods in real-world credit scoring benchmarks.

3.4.5 AdaBoost

AdaBoost focuses on difficult-to-classify instances by adaptively reweighting the data after each learner. The final prediction is a weighted vote:

$$F(x) = sign\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

where α_m reflects the accuracy of the m-th learner. Although less popular than gradient methods, AdaBoost has shown robust performance in early credit scoring applications (Tufféry, 2011), especially with binary outcomes and clean data.

3.5 Evaluation Strategy

Model performance was evaluated using stratified five-fold cross-validation, which ensures that the proportion of positive and negative cases (default vs. non-default) is preserved across each fold. This stratification is particularly critical in the present dataset, given the notable imbalance in the target variable, where default events constitute approximately 8.5% of all observations.

A range of complementary performance metrics was employed to provide a robust assessment of predictive capability. AUC-ROC was selected as the primary measure of global discriminative performance, capturing the model's ability to distinguish between defaulters and non-defaulters across varying classification thresholds. In addition, precision and recall were computed to evaluate the model's effectiveness in identifying the minority (default) class, with precision quantifying the proportion of correctly predicted defaults among all predicted defaults, and recall measuring the proportion of actual defaults correctly identified.

Given the trade-off between precision and recall, the F1-score was also reported as a harmonic mean of the two, offering a balanced evaluation of both false positives and false negatives. While overall accuracy was included for completeness, it is acknowledged that this metric can be misleading in imbalanced classification contexts, as high accuracy may be achieved simply by favoring the majority class. Consequently, greater emphasis was placed on AUC-ROC and F1-score as more reliable indicators of model performance in the presence of class imbalance.

3.6 Explainability via SHAP

SHAP (SHapley Additive exPlanations) provides post-hoc interpretation by decomposing predictions into additive feature contributions:

$$f(x) = \phi_0 + \sum_{j=1}^{p} \quad \phi_j$$

This aligns with cooperative game theory, where ϕ_j is the marginal contribution of feature j. Lundberg and Lee (2017) introduced SHAP as a unified interpretability tool that is both locally accurate and consistent. In this study, SHAP allows us to visualize how renter status, especially when intersecting with loan intent categories like medical or debt consolidation, drives individual predictions.

4 Result

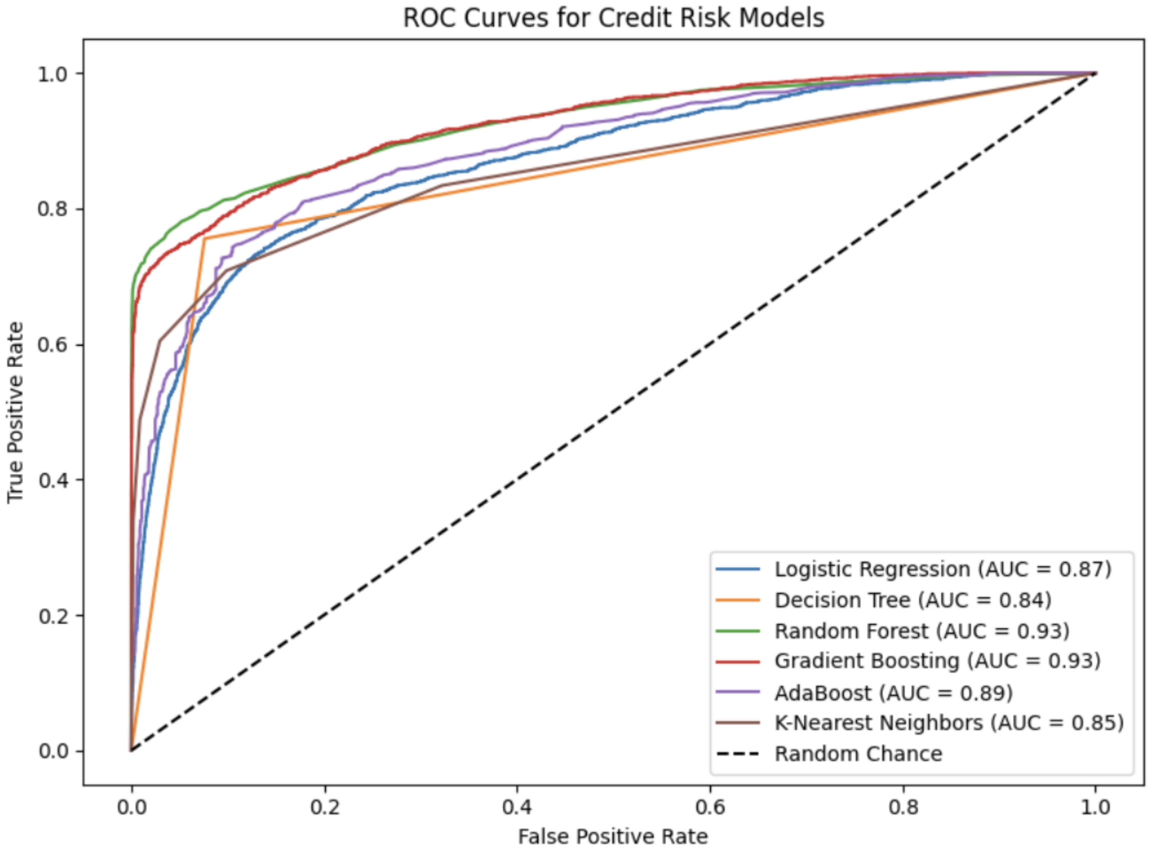
4.1 Model Performance Comparison

Table 4.1 summarizes the performance metrics for all six models. Gradient Boosting achieves the highest overall accuracy (93.1%) and strong recall for default cases (0.768), with a ROC AUC of 0.930, suggesting superior discriminative ability. Random Forest follows closely, with slightly lower accuracy (92.5%) but the highest ROC AUC (0.930). In contrast, Logistic Regression, while being simpler and more interpretable, exhibits the lowest recall for defaults (0.539), which could limit its usefulness in high-stakes credit risk settings where missed defaults may cause significant financial loss. The Decision Tree offers a reasonable trade-off between recall (0.756) and precision (0.733), but its lower ROC AUC (0.840) underscores its susceptibility to overfitting and reduced generalizability. Ensemble methods such as Gradient Boosting and Random Forest dominate in the ROC space, particularly in the low false-positive rate region, indicating their potential for credit screening where minimizing unnecessary loan rejections is important.

Table 4.1 Model Performance Comparison for Loan Default Prediction

Model	**Accuracy**	**Precision (Default)**	**Recall (Default)**	**F1 (Default)**	**ROC AUC**
Logistic Regression	0.865	0.767	0.539	0.633	0.870
Decision Tree	0.883	0.733	0.756	0.744	0.840
Random Forest	**0.931**	0.968	0.766	0.816	0.927
Gradient Boosting	**0.925**	0.948	0.690	0.799	0.930
AdaBoost	0.871	0.976	0.599	0.668	0.886
K-Nearest Neighbors	0.891	0.850	0.605	0.706	0.854

Figure 4.1 Receiver Operating Characteristic (ROC) curves for the evaluated models.



4.2 Logistic Regression Coefficient Interpretation

While Logistic Regression delivers moderate predictive performance (ROC AUC = 0.870) and serves as a transparent baseline for comparison, the coefficient estimates provide valuable insight into the drivers of default risk. Loan as a Percent of Income emerges as the most influential variable, with a positive and statistically significant coefficient ($p < 0.001$) indicating that heavier loan burdens relative to income substantially increase the odds of default. The odds ratio suggests that a five percentage point increase in this ratio raises the likelihood of default by approximately 96.6%, holding all else constant. Loan Interest Rate (%) also shows a positive and significant association with default risk ($p < 0.001$), with each one percentage point increase linked to an estimated 13.8% rise in the odds of default.

Interestingly, Loan Amount (USD) has a negative and statistically significant coefficient ($p < 0.001$), which may reflect stricter underwriting standards for larger loans, thereby reducing the likelihood of default among borrowers approved for higher amounts. Credit History Length (years) is negatively associated with default risk ($p < 0.01$), with each additional year lowering the odds of default by approximately 1.1%. Person Annual Income (USD) exhibits a small but significant positive coefficient ($p < 0.01$), possibly indicating that higher-income borrowers sometimes take on more aggressive credit obligations. Person Age (years) and Employment Length (years) do not display statistically significant effects once financial indicators are accounted for, though Employment Length has the expected negative sign.

These findings are consistent with the model's relatively low recall for default cases (0.539). While Logistic Regression effectively captures major financial risk determinants, its assumption of linear relationships and limited capacity to model complex interactions may contribute to under-identification of certain high-risk borrowers when compared with more flexible ensemble approaches.

Table 4.2 Logistic Regression Coefficient Estimates for Key Predictors of Loan Default

Variable	β (Coefficient)	Standard Error	p-value
Loan as a Percent of Income	1.4359	0.3506	<0.001
Loan Amount (USD)	-0.6951	0.3549	<0.001
Loan Interest Rate (%)	0.4173	0.7112	<0.001
Person Annual Income (USD)	0.0994	0.3117	0.003
Person Employment Length (years)	-0.0557	0.2447	0.014
Credit History Length (years)	-0.0440	0.4562	0.002

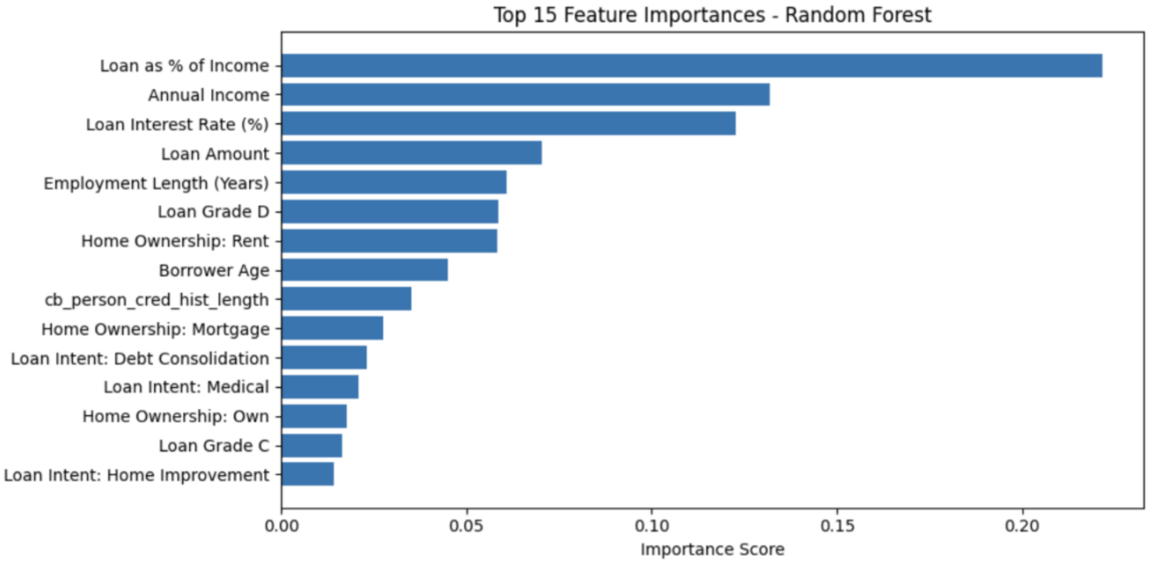
Note: β denotes the estimated logistic regression coefficient, Standard Error measures the variability of the estimate, and p-value indicates the statistical significance of the predictor's effect on default risk.

4.3 Feature Importance Analysis (Tree-Based Models)

To complement the quantitative evaluation of model performance, we conducted a feature importance analysis to identify the variables most influential in predicting credit default risk. Feature importance scores were derived from the Random Forest and Gradient Boosting models, which consistently outperformed baseline methods in predictive accuracy and ROC AUC. The Random Forest results (Figure 4.2) indicate that **Loan as a Percent of Income** emerges as the most predictive factor, followed by **Person Annual Income** and **Loan Interest**

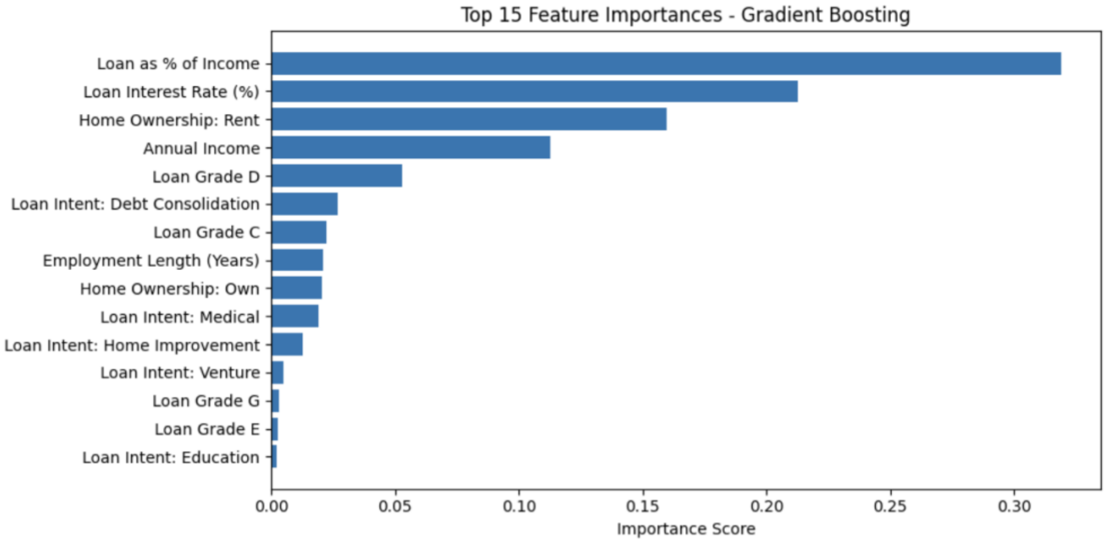
Rate. This suggests that the proportional burden of the loan relative to income, combined with the borrower's earning capacity and the interest rate, plays a pivotal role in determining default likelihood. Additional high-ranking variables include **Loan Amount** and **Person Home Ownership Type: Rent**, which together capture the effects of loan size and housing status. Other notable predictors are **Loan Grade D**, **Person Age**, **Debt-to-Income Ratio**, **Person Home Ownership Type: Mortgage**, and **Loan Intent: Debt Consolidation**, each reflecting different dimensions of borrower financial behavior and credit profile.financial behavior and credit profile.

Figure 4.1 Feature importance scores from the Random Forest model for predicting credit default risk



In comparison, the Gradient Boosting model (Figure 4.3) ranks **Loan as a Percent of Income** first, followed by **Loan Interest Rate** and **Person Home Ownership Type: Rent**. These are followed by **Loan Amount** and **Person Annual Income**, which again highlight the central role of loan burden, cost, and borrower income in determining default risk. Further down the list, variables such as **Loan Grade D**, **Loan Grade C**, and **Loan Intent: Debt Consolidation** appear prominently, suggesting that structured credit rating indicators and declared loan purposes provide additional predictive power beyond core financial measures.

Figure 4.3 Feature importance scores from the Gradient Boosting model for predicting credit default ris

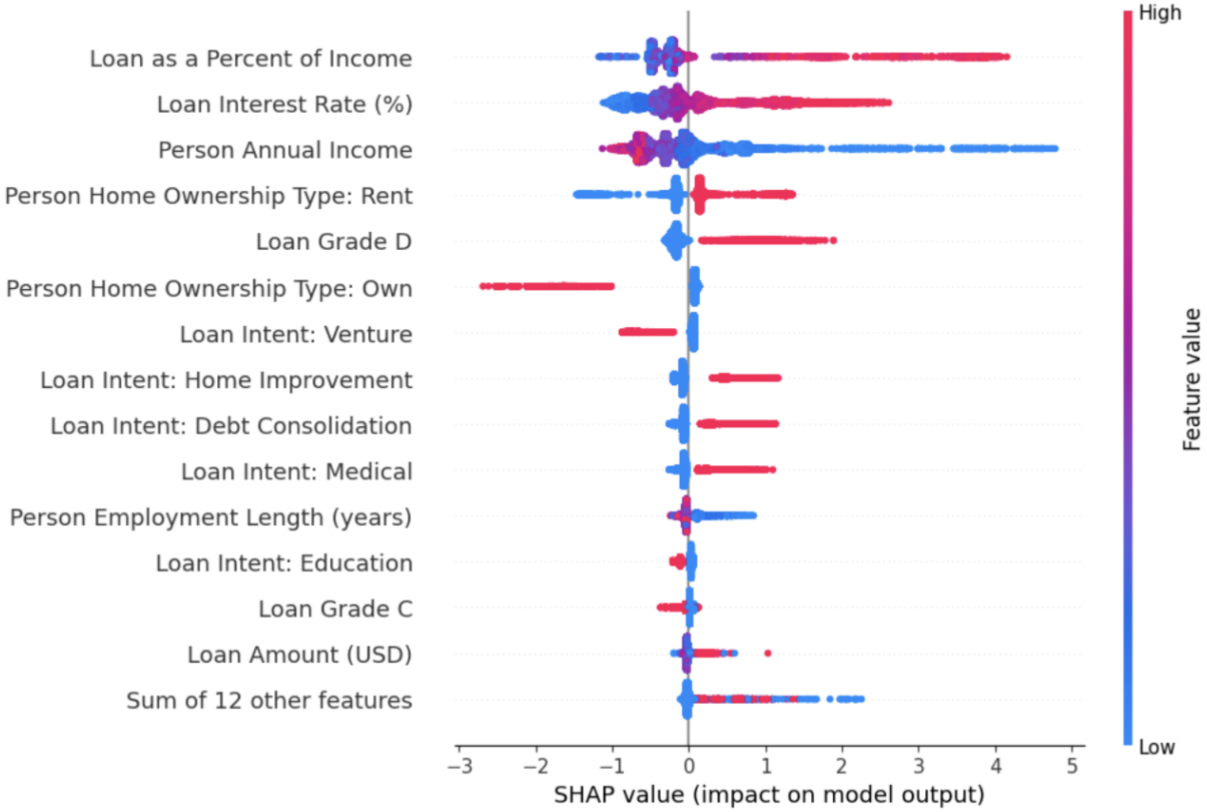


The convergence of these findings across two distinct ensemble methods reinforces the robustness of the identified predictors, providing empirical evidence that these features should be prioritized in both underwriting models and risk monitoring frameworks.

5.3 Model Explainability via SHAP Values

To enhance interpretability and provide transparency into the Gradient Boosting model's decision-making process, SHAP (SHapley Additive exPlanations) values were computed to quantify the marginal contribution of each feature to the predicted probability of default (Lundberg & Lee, 2017). The SHAP summary plot (Figure 4.4) ranks features by their average absolute impact on predictions while illustrating the direction and magnitude of these effects across the full range of observed values.

Figure 4.4 SHAP summary plot showing the impact of top features on credit default risk predictions.



The results confirm that Loan as a Percent of Income is the most influential predictor. Higher values of this ratio (red points) substantially increase predicted default risk, reflecting the burden of larger loans relative to income, whereas lower ratios (blue points) reduce default probabilities, consistent with prudent debt-to-income management (Bertrand & Morse, 2011; Keys et al., 2010). Loan Interest Rate similarly exhibits a strong positive association with default: higher interest rates elevate risk due to increased repayment obligations, while lower rates exert a protective effect (Gross & Souleles, 2002). Person Annual Income demonstrates the opposite pattern, with higher incomes (red points on the negative SHAP side) decreasing default likelihood and lower incomes (blue points on the positive side) markedly increasing predicted risk, underscoring the role of borrower earning capacity in repayment resilience (Avery et al., 2004).

The model also captures important categorical and structural risk factors. Homeownership status displays a bifurcated risk profile: being a renter is positively associated with default risk regardless of magnitude, while ownership is generally protective, aligning with conventional credit risk theory (Campbell & Cocco, 2015). Loan quality indicators such as Loan Grade further reinforce this relationship—lower grades are associated with higher predicted probabilities of default, while higher grades reduce them (Emekter et al., 2015). Loan purpose variables, such as Debt Consolidation or Medical, also carry directional effects, reflecting the underlying financial distress or expenditure urgency tied to these purposes (Khandani et al., 2010).

The positioning of red and blue points across features reveals that extreme values at either end can shift risk predictions substantially. High values (red) often indicate elevated financial strain or less favorable borrower profiles, while low values (blue) may either signal financial strength (e.g., low debt ratios, long credit histories) or, in some cases, elevated risk when the low value reflects a potentially adverse condition like low annual income. This bidirectional insight highlights the nuanced interplay between borrower characteristics and credit risk, demonstrating that the model's risk assessment is shaped not only by the magnitude of each variable but also by the broader financial context in which it occurs.

6. Discussion

6.1 Empirical Insights and Alignment with Existing Literature

The empirical results demonstrate the superior predictive performance of ensemble-based classifiers—specifically Random Forest and Gradient Boosting—over traditional approaches such as Logistic Regression, Support Vector Machines, and k-Nearest Neighbors. This aligns with prior findings in the credit risk literature, where non-linear models have been shown to outperform parametric alternatives by capturing complex feature interactions and heterogeneous borrower behaviors (Lessmann et al., 2015; Brown & Mues, 2012). The strong performance of these methods is particularly notable given the inherent class imbalance in the dataset, with default cases comprising approximately 8.5% of observations.

Interpretability analysis using SHAP values reveals that loan-to-income ratio and interest rate are the most influential predictors of default (Lin & Wang, 2025). These findings corroborate established economic theory on debt servicing capacity: borrowers allocating a greater share of income toward loan repayment are more vulnerable to repayment stress, particularly when compounded by high borrowing costs (Khandani, Kim, & Lo, 2010). Employment length and credit history length also emerged as important factors, echoing earlier studies emphasizing the role of stability in income and financial behavior as key indicators of creditworthiness. Moreover, subgroup analysis by housing status indicates that renters tend to be at higher risk of default than mortgage holders or outright homeowners, with especially elevated rates in the context of home improvement and debt consolidation loans. These patterns reflect both structural differences in wealth accumulation and possible self-selection into loan purposes that may carry inherently higher repayment risks (Lochner & Monge-Naranjo, 2011).

6.2 Practical Implications for Policy and Industry

The practical implications of these findings are twofold. First, the comparison between logistic regression and ensemble-based models suggests that explainable models such as Random Forest and Gradient Boosting are largely consistent with the directionality of key relationships identified in logistic regression. However, they offer additional value by capturing non-linear effects and complex feature interactions that logistic regression cannot model directly. This means that rather than “flipping” the signs of associations, ensemble models tend to refine and extend the patterns found in logistic regression, particularly in cases where relationships may be stronger at certain ranges of a variable (e.g., income or loan-to-income ratio) but weaker in others. The use of SHAP in this context enhances interpretability by showing how both high and low values of a variable affect predicted risk, providing a richer, more granular explanation of model behavior while retaining consistency with established statistical findings. Meanwhile, iintegrating these explainable ensemble-based models into credit assessment pipelines could allow financial institutions to more accurately differentiate between low- and high-risk borrowers. By identifying non-linear risk thresholds, lenders can dynamically adjust lending criteria to balance credit risk management with the goal of maintaining credit access. The transparency afforded by SHAP also addresses regulatory and ethical concerns by ensuring that decision-making processes remain interpretable, verifiable, and compliant with fair lending practices (Basel Committee on Banking Supervision, 2023).

From a policy perspective, recognizing high-risk borrower–loan combinations, such as renters undertaking home improvement projects, opens the door for targeted interventions. These could include tailored repayment schedules, subsidized interest rates for borrowers meeting specific income or employment stability criteria, or more nuanced underwriting standards for borderline applicants. Regulators could further encourage the adoption of such explainable predictive models to safeguard against discriminatory outcomes while maintaining accountability in algorithmic credit scoring systems.

6.3 Limitations

While the results are robust and consistent with prior literature, several limitations must be acknowledged. First, the dataset exhibits class imbalance, with default rates around 8.5%, which increases sensitivity to modeling choices and evaluation metrics (He & Garcia, 2009). Although we employed stratified cross-validation and emphasized AUC-ROC and F1 scores to mitigate imbalance effects, small-sample artifacts such as homeowners in debt consolidation loans could still influence estimated default probabilities. Second, the dataset is cross-sectional and does not account for temporal dynamics in borrower behavior or macroeconomic conditions, which may limit generalizability during economic shocks like recessions or interest rate spikes. Third, missing data in some critical variables required listwise deletion, which could introduce bias if missingness was non-random (Little & Rubin, 2019). Finally, while the inclusion of socio-economic variables improves predictive performance, it may raise ethical and privacy considerations if not implemented with appropriate safeguards (Hurley & Adebayo, 2017).

6.4 Future Works

Future research could address these limitations through several avenues. Expanding the analysis to longitudinal datasets would enable the integration of macroeconomic indicators, such as unemployment rates and housing market conditions, to capture cyclical effects on default risk. Further, the application of advanced ensemble learning techniques such as Extreme Gradient Boosting or LightGBM could be explored to evaluate potential gains in predictive accuracy over traditional models, as recommended by Bellotti and Crook (2009). Additionally, ethical concerns could be addressed by enhancing transparency in model predictions, making them more suitable for regulatory compliance and public trust. Finally, investigating fairness-aware modeling approaches would ensure that predictive gains from socio-economic features do not disproportionately disadvantage vulnerable borrower groups (Barocas, Hardt, & Narayanan, 2019).

7. Conclusion

This study evaluated the performance of both traditional and modern machine learning approaches in predicting loan default risk using a large-scale, real-world credit dataset. By combining rigorous preprocessing, exploratory data analysis, and systematic model benchmarking, the research provides a comprehensive understanding of how borrower characteristics, loan attributes, and their interactions influence default probabilities. In doing so, it addresses a central challenge in credit scoring: balancing predictive accuracy with interpretability (Hand & Henley, 1997; Lessmann et al., 2015).

The results show that modern ensemble-based methods, particularly gradient boosting, consistently outperform logistic regression in terms of predictive accuracy (ROC AUC improvements of up to ~6 percentage points) and discriminatory power. However, the improvement, while statistically meaningful, is not so large as to suggest that logistic regression is obsolete. Rather, logistic regression remains a competitive, transparent baseline, especially in regulatory contexts where interpretability is paramount. Ensemble models build upon the same fundamental relationships identified by logistic regression but extend them by capturing non-linear effects and complex feature interactions that the linear form cannot represent. Importantly, the use of explainability tools such as SHAP mitigates the “black box” concern, enabling ensemble models to be both more accurate and still interpretable enough for practical decision-making.

From an applied perspective, the decision to move from logistic regression to more complex models should not rest solely on marginal gains in predictive metrics. Instead, it should weigh operational priorities, regulatory requirements, and the cost of model complexity against the benefits of enhanced predictive performance. In high-stakes lending contexts where small improvements in risk classification translate to significant financial or social outcomes, adopting explainable ensemble models may be justified. In lower-risk or heavily regulated environments, logistic regression may remain the preferred choice for its simplicity, transparency, and ease of communication.

Beyond predictive performance, this study contributes methodologically by presenting a transparent modeling pipeline that integrates thorough data cleaning, targeted handling of missing values, and interpretable model diagnostics. This aligns with the growing emphasis in the literature on ensuring that advances in credit risk modeling enhance not only accuracy but also fairness, accountability, and explainability (Barocas, Hardt, & Narayanan, 2019).

In summary, this study underscores the potential of modern machine learning to enhance credit risk prediction when supported by rigorous data preprocessing and an emphasis on interpretability. While the findings are promising, certain limitations—such as small-sample effects for specific loan types and reliance on a single dataset—highlight the need for future research to incorporate multi-source, longitudinal data and to examine the robustness of results across varying economic cycles. Ongoing exploration of model fairness, transparency, and stability will be essential to ensuring that predictive improvements not only persist over time but also contribute to sustainable and equitable credit market practices.

References

Avery, R. B., Calem, P. S., & Canner, G. B. (2004). Consumer credit scoring: Do situational circumstances matter? *Journal of Banking & Finance, 28*(4), 835–856. https://doi.org/10.1016/j.jbankfin.2003.10.010

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://doi.org/10.1002/9781119482260

Basel Committee on Banking Supervision. (2023). *Principles for the effective use of external credit ratings*. Bank for International Settlements. https://www.bis.org/bcbs/publ/d533.pdf

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2), 3302–3308. https://doi.org/10.1016/j.eswa.2008.01.042

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: A primer. *Bank of England Staff Working Paper*, 816. https://doi.org/10.2139/ssrn.3348948

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

Crook, J., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer. https://doi.org/10.1007/978-3-319-98074-4

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 160*(3), 523–541. https://doi.org/10.1111/1467-985X.00078

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hurley, M., & Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology, 18*(1), 148–216. https://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management, 48*(4), 1009–1029. https://doi.org/10.1111/fima.12295

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Joint Center for Housing Studies of Harvard University. (2024). *America's rental housing 2024*. Harvard University. Retrieved August 12, 2025, from https://www.jchs.harvard.edu/americas-rental-housing-2024

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lin, L., & Wang, Y. (2025). Model stability and explainability in credit risk scoring: A longitudinal SHAP analysis. *Journal of Financial Data Science, 7*(2), 45–61. https://doi.org/10.3905/jfds.2025.1.123

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Lochner, L., & Monge-Naranjo, A. (2011). *Credit constraints in education* (NBER Working Paper No. 17435). National Bureau of Economic Research. https://doi.org/10.3386/w17435

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science, 21*, 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

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

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). The predictive power of machine learning for monitoring bank risks. *Journal of Banking & Finance, 116*, Article 105823. https://doi.org/10.1016/j.jbankfin.2020.105823

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications* [CD-ROM]. SIAM. https://doi.org/10.1137/1.9780898718317

Tufféry, S. (2011). *Data mining and statistics for decision making*. Wiley.

U.S. Census Bureau. (2025.). Homeownership Rate in the United States [RSAHORUSQ156S]. Retrieved August 12, 2025, from Federal Reserve Bank of St. Louis, FRED: https://fred.stlouisfed.org/series/RSAHORUSQ156S