Predicting Loan Default: A Comparative Analysis of Model Performance and Explainability

Hoang To (ht8758)

The University of Texas at Austin

December 01, 2025

Abstract

Credit underwriting is the lifeblood of financial institutions, yet regulatory requirements for model interpretability have constrained the adoption of state-of-the-art machine learning algorithms in credit risk assessment. This study investigates the magnitude of this interpretability-performance tradeoff by comparing four machine learning models (logistic regression, LightGBM, XGBoost, and neural networks) on a dataset of 255,347 loan records from Coursera's Loan Default Prediction Challenge. Model performance was evaluated using ROC-AUC as the primary metric, with SHAP (SHapley Additive exPlanations) employed to assess explainability across architectures. Results indicate that XGBoost achieved the highest predictive performance (ROC-AUC = 0.759), while logistic regression, the most interpretable model, attained a competitive ROC-AUC of 0.753, a performance gap of 0.82%. Notably, SHAP analysis revealed consistent feature importance rankings across all models. These findings suggest that the performance cost of regulatory compliance may be smaller than commonly assumed, and that explainable AI techniques can provide consistent interpretations even for complex methods.

1 Introduction

Since the dawn of banking, credit has been the core business through which financial institutions grow and compete with one another. The discipline operates at the intersection of risk management and customer access to credit. As a finance professional at a leading lending company, I experience first hands how our capacity to lend ties directly with our ability to underwrite consumer risk appropriately. At the same time, the Equal Credit Opportunity Act (ECOA) mandates that lenders provide specific reasons for adverse credit decisions (U.S. Congress, 1974). Regulatory guidance from the Federal Reserve and Consumer Financial Protection Bureau reinforces these requirements (Consumer Financial Protection Bureau, 2022; *Report to Congress on credit scoring and its effects on the availability and affordability of credit*, 2007). This has created a significant barrier to adopting advanced machine learning models, particularly neural networks, which are often considered "black boxes" due to their complex internal mechanisms (Hurley & Adebayo, 2016). Thus, financial institutions must predict which borrowers are likely to default while simultaneously maintaining transparency in their lending decisions. This dual challenge has intensified with the advancement of machine learning techniques that offer superior predictive accuracy but often lack the interpretability required for regulatory compliance. As a result, many financial institutions continue to rely on logistic regression and tree-based models despite potentially sacrificing predictive performance (Consumer Financial Protection Bureau, 2022; Lessmann et al., 2015).

This paper aims to understand two fundamental questions: (1) Are neural network based models outperforming more interpretable logistics and tree-based models? and (2) Can we better explain what drive prediction outcomes of more complex neural network based models to sufficiently meet regulatory requirements? Accurate credit worthiness assessments would help both the lenders and borrowers, whereas the former can assign appropriate level of risk to the borrowers and lower the risk of unexpected losses, allowing the latter to borrow more cheaply (Einav et al., 2013).

By comparing four distinct machine learning models, ranging from highly interpretable model (logistic regression) to the "black box" model (neural networks), and employing SHAP values for explainability analysis, this paper seeks to provide a repeatable framework to quantify the interpretability-performance tradeoff that financial institutions face, providing actionable insights

for practitioners navigating the balance between model performance and regulatory compliance.

2 Research Background

Credit scoring has been a key pillar of lending decisions for decades. Traditional approaches, particularly FICO scores and logistic regression models, have dominated the industry due to their interpretability and regulatory acceptance (Baesens et al., 2003; Brown & Mues, 2012; Lessmann et al., 2015). These methods rely on a limited set of financial and demographic variables to predict default probability, with coefficients that can be directly interpreted as the marginal effect of each variable.

Within the last two decades, the application of machine learning to credit risk has improved significantly. Benchmarking studies demonstrated that ensemble methods could outperform traditional logistic regression (Khandani et al., 2010) with random forests and gradient boosting emerged as powerful algorithms for credit scoring. Wang et al. (2011) found that ensemble approaches consistently outperformed single classifiers in credit risk assessment. XGBoost, a scalable tree boosting system (Chen & Guestrin, 2016), has become particularly popular due to its performance and built-in handling of missing values. Xia et al. (2017) demonstrated that boosted decision trees with Bayesian optimization could further improve credit scoring performance.

Deep learning applications in credit risk have shown promising results but has faced adoption challenges. Sirignano et al. (2016) demonstrated the potential of deep learning for mortgage risk prediction, while Hamori et al. (2018) compared ensemble learning with deep learning approaches for default risk analysis. Kvamme et al. (2018) applied convolutional neural networks to mortgage default prediction, while Bellotti and Crook (2013) explored dynamic models for credit card default forecasting. However, the "black box" nature of neural networks raises concerns about interpretability and regulatory compliance (Hurley & Adebayo, 2016). This has limited their widespread usage in credit risk assessment despite their predictive power.

The need for model interpretability in regulated industries has driven research in explainable AI. Industry reports from major consulting firms have highlighted the growing importance of explainable AI in financial services (Deloitte, 2018; McKinsey & Company, 2017; PwC, 2018). FICO, a leading credit scoring company, has published guidance on implementing explainable ma-

chine learning for credit risk (Fahner, Gerald, 2018). Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations), a unified framework for interpreting model predictions based on cooperative game theory. This approach assigns each feature an importance value for a particular prediction, providing both local (individual prediction) and global (overall model) explanations. Lundberg et al. (2020) and Molnar (2020) extended their work by demonstrating how SHAP values can provide global understanding of tree-based models. Bussmann et al. (2021) specifically examined explainable machine learning in credit risk management, arguing that SHAP and similar techniques could bridge the gap between model performance and regulatory requirements. Prior to SHAP, Ribeiro et al. (2016) introduced LIME (Local Interpretable Model-agnostic Explanations), which provides local explanations for any classifier. Barredo Arrieta et al. (2019) provides a taxonomy of explanable AI methods and their applications in responsible AI. In this research, we will use SHAP to gauge the interpretability across machine learning models.

3 Data

In aimming to addresses the research questions, this paper utilizes the Loan Default Prediction dataset from Coursera's Loan Default Prediction Challenge, which contains borrower information for credit default prediction. The dataset includes 255,347 observations with 16 columns describing a borrower's characteristics such as age, income, credit score, education, etc. that will be used as features (Table 1). Last but not least, there is a binary target variable indicating default status.

Table 1: Dataset Features

Feature	Type	Description
Age	Numerical	Age of the borrower
Income	Numerical	Annual income
LoanAmount	Numerical	Amount of money borrowed
CreditScore	Numerical	Credit score (creditworthiness)
MonthsEmployed	Numerical	Months of employment
NumCreditLines	Numerical	Number of open credit lines
InterestRate	Numerical	Loan interest rate
LoanTerm	Numerical	Loan term in months
DTIRatio	Numerical	Debt-to-Income ratio
Education	Categorical	Highest education level
EmploymentType	Categorical	Employment status
MaritalStatus	Categorical	Marital status
HasMortgage	Categorical	Mortgage status (Yes/No)
HasDependents	Categorical	Has dependents (Yes/No)
LoanPurpose	Categorical	Purpose of the loan
HasCoSigner	Categorical	Co-signer status (Yes/No)

3.1 Exploratory Data Analysis

Prior to apply any algorithms, we looked through the dataset to have a gasp of the features being leveraged to answer the research questions (Table 2). The dataset composed of loans being made to a diverse range of borrowers (income ranges from $15,000 to $150,000, loan amount from $5,000 to $250,000, etc.). Fortunately, there was no missing-values among the features.

Lastly, the default rate within the dataset is around 11.6%, which raised a question of imbalanced data, which will be discussed in Section 3.3.

Table 2: Summary Statistics

	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio
mean	43.50	82,499	127,579	574.26	59.54	2.50	13.49	36.03	0.50
std	14.99	38,963	70,841	158.90	34.64	1.12	6.64	16.97	0.23
min	18.00	15,000	5,000	300.00	0.00	1.00	2.00	12.00	0.10
25%	31.00	48,826	66,156	437.00	30.00	2.00	7.77	24.00	0.30
50%	43.00	82,466	127,556	574.00	60.00	2.00	13.46	36.00	0.50
75%	56.00	116,219	188,985	712.00	90.00	3.00	19.25	48.00	0.70
max	69.00	149,999	249,999	849.00	119.00	4.00	25.00	60.00	0.90

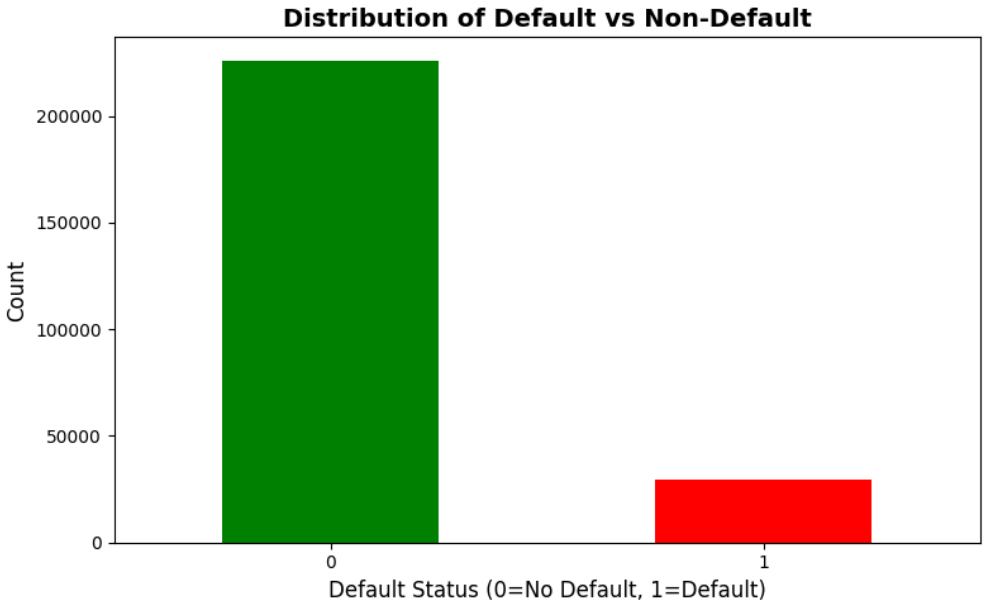


Figure 1: Distribution of Default vs Non-Default Cases

3.2 Data Preprocessing

3.2.1 Correlated Features

To identify multicollinearity among the numerical features, a correlation matrix was computed (Figure 2). Features with correlation coefficients above 0.8 would be considered highly correlated. In this dataset, we did not see any features with correlation coefficients higher than 0.2, therefore no numerical feature was dropped from the analysis.

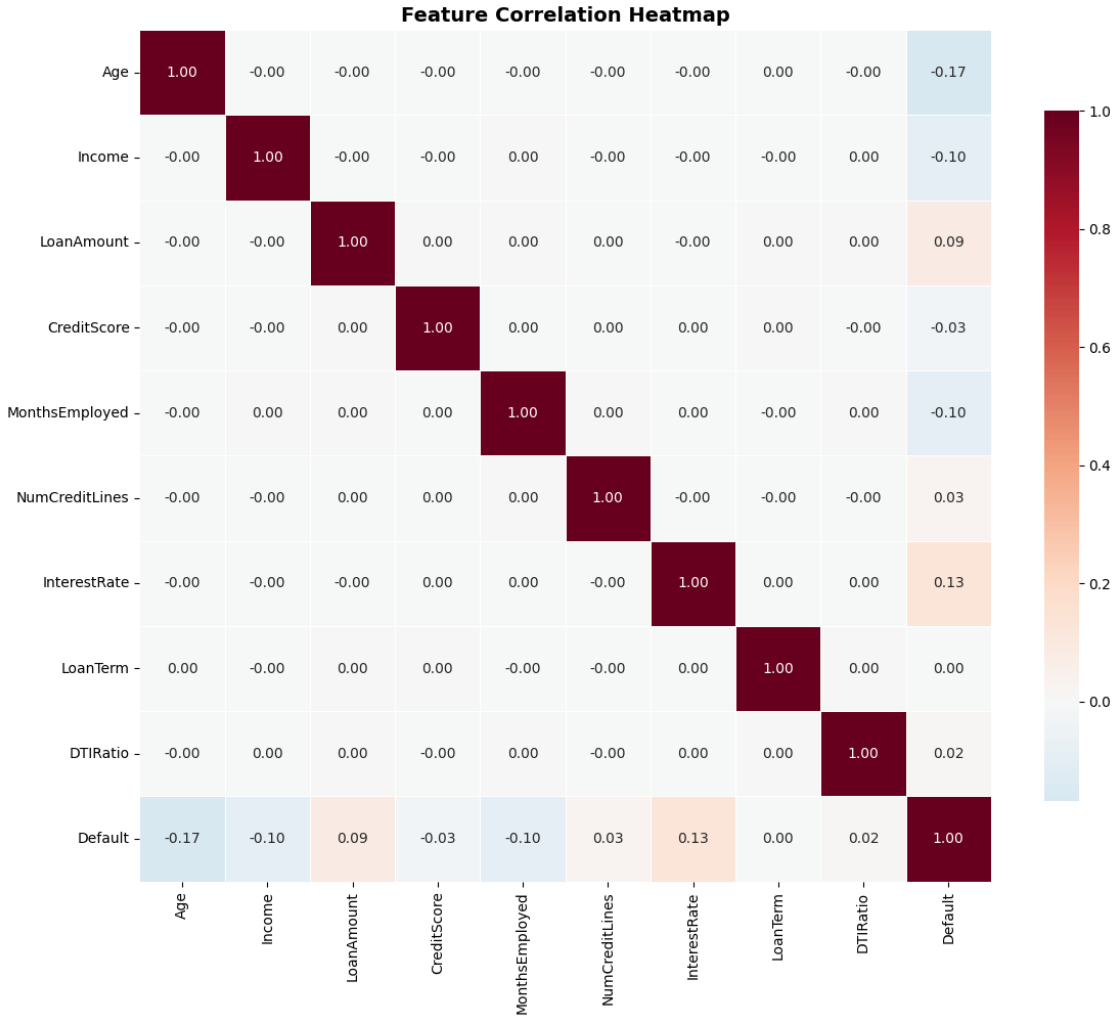


Figure 2: Feature Correlation Heatmap

3.2.2 Normalization and Encoding

Subsequently, the following steps were applied to preprocess the data:

- **Train-Test Split:** The dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to preserve the class distribution of the target variable (Default) in both sets. This is critical given the class imbalance present in the data (approximately 11.6% default rate).
- **Categorical Encoding:** One-hot encoding (dummy variable encoding) was applied to all categorical variables: Education (4 categories: High School, Bachelor's, Master's, PhD),

EmploymentType (4 categories: Full-time, Part-time, Self-employed, Unemployed), MaritalStatus (3 categories: Single, Married, Divorced), HasMortgage (2 categories: Yes, No), HasDependents (2 categories: Yes, No), LoanPurpose (5 categories: Home, Auto, Education, Business, Other), and HasCoSigner (2 categories: Yes, No). One-hot encoding was chosen over label encoding to avoid introducing ordinal relationships where none exist, which could mislead tree-based and linear models.

- **Feature Scaling/Normalization:** Numerical features (Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines, InterestRate, LoanTerm, DTIRatio) were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the feature mean and σ is the standard deviation. Standardization ensures all features have mean 0 and standard deviation 1, which is important for gradient-based optimization in logistic regression and neural networks, and improves convergence speed.

After preprocessing, the original 16 predictor variables (9 numerical, 7 categorical) were transformed into 24 features: 9 standardized numerical features plus 15 binary indicator variables from one-hot encoding.

3.3 Class Imbalance Considerations

With a default rate of 11.6%, there is an imbalance in the data. Rather than applying resampling techniques like SMOTE (Chawla et al., 2002), which can introduce artificial patterns, ROC-AUC was selected as the primary evaluation metric as it is robust to class imbalance and evaluates model performance across all classification thresholds (Fawcett, 2006).

4 Methods

4.1 Model Selection

Four machine learning models were selected to represent different levels of interpretability and modeling approaches:

4.1.1 Highly Interpretable Models

Logistic Regression: Selected as the baseline model due to its widespread adoption in the credit industry and regulatory acceptance. Logistic regression models the probability of default using the sigmoid function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+...+\beta_n X_n)}} \tag{2}$$

where each coefficient β_i can be exponentiated to obtain odds ratios, providing direct interpretation: a one-unit increase in feature X_i multiplies the odds of default by e^{β_i} (Hosmer Jr et al., 2013). This transparency makes logistic regression particularly valuable for regulatory compliance under frameworks such as the Equal Credit Opportunity Act (ECOA) and SR 11-7, which require explainable lending decisions. The model was trained with L2 regularization (C=1.0) and a maximum of 1,000 iterations to ensure convergence.

4.1.2 Moderately Interpretable Models

LightGBM: A gradient boosting framework developed by Microsoft that uses histogram-based algorithms and leaf-wise tree growth for computational efficiency (Ke et al., 2017). Unlike traditional gradient boosting methods that grow trees level-wise, LightGBM grows trees by splitting the leaf with the highest delta loss, often resulting in a deeper, more asymmetric trees that can capture more complex interactions.

XGBoost (Extreme Gradient Boosting): A regularized gradient boosting implementation that has become the de facto standard for tabular data competitions and industry applications. XGBoost minimizes a regularized objective function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3}$$

where l is a differentiable convex loss function and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2$ penalizes model complexity through the number of leaves T and L2 regularization on leaf weights w. This built-in regularization helps prevent overfitting, a common concern with high-dimensional credit data.

4.1.3 Low Interpretability Models

Neural Network: A feedforward multilayer perceptron (MLP) represents the "black box" end of the interpretability spectrum. The architecture consists of an input layer matching the 24 preprocessed features, followed by three hidden layers with 128, 64, and 32 neurons respectively, each using ReLU activation functions:

$$\text{ReLU}(x) = \max(0, x) \tag{4}$$

Dropout regularization (rate = 0.3) was applied between hidden layers to prevent overfitting, and the output layer uses sigmoid activation for binary classification. The network was trained using the Adam optimizer with binary cross-entropy loss and early stopping monitoring validation loss.

4.2 Evaluation Metrics

Model performance was assessed using multiple classification metrics:

- **Accuracy:** Overall proportion of correct predictions
- **Precision:** Ability to minimize false positives (incorrectly predicting default)
- **Recall:** Ability to identify actual defaults
- **F1-Score:** Harmonic mean of precision and recall
- **ROC-AUC:** Area under the receiver operating characteristic curve

ROC-AUC was selected as the primary metric because it evaluates model performance across all classification thresholds and, as mentioned, is less sensitive to class imbalance than accuracy (He & Garcia, 2009).

5 Results

5.1 Model Performance

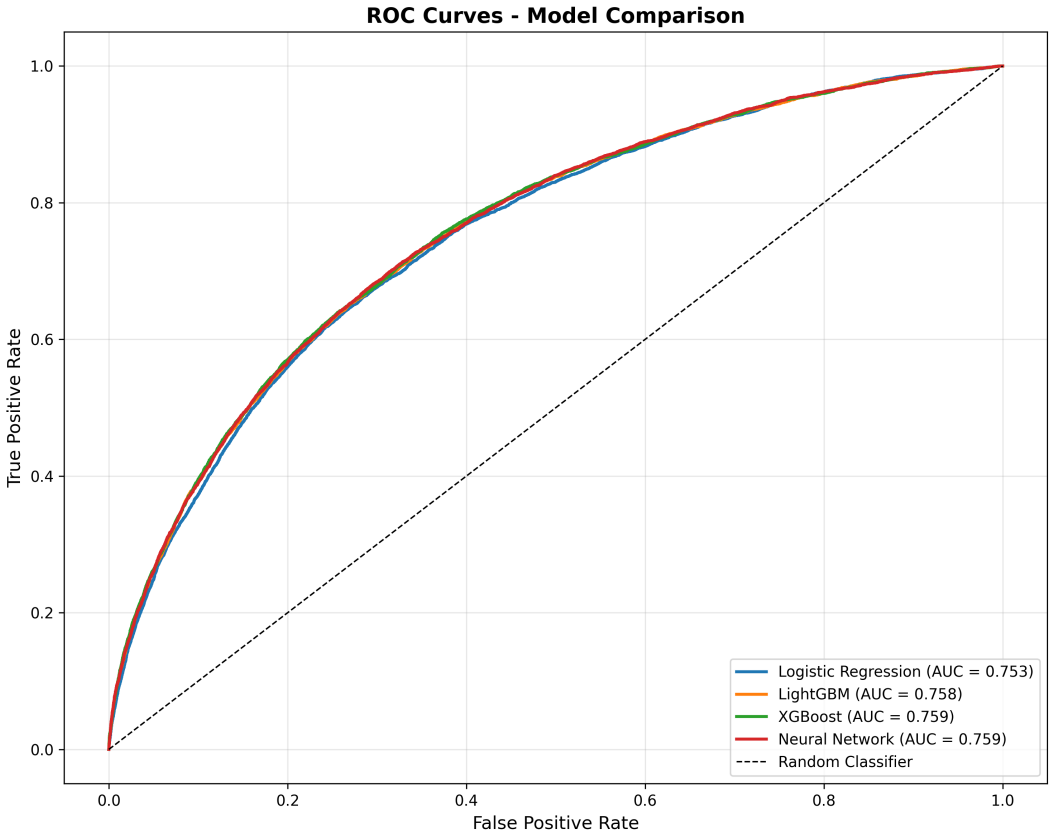
Table 3 below presents the performance metrics for all four models. As discussed, ROC-AUC is the primary metric used to rank the models' performance.

Table 3: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	0.886	0.618	0.063	0.114	0.759
Neural Network	0.884	0.699	0.016	0.032	0.758
LightGBM	0.886	0.620	0.063	0.115	0.757
Logistic Regression	0.885	0.608	0.034	0.064	0.753

The XGBoost model demonstrated the best overall performance with a ROC-AUC of 0.7593, closely followed by the Neural Network (0.7586) and LightGBM (0.7580), showing minimal separation among the top three models. The baseline Logistic Regression model achieved a competitive ROC-AUC of 0.7531, indicating the more complex ensemble methods provided only marginal gains in the area under the curve. Regarding the precision versus recall tradeoff, all models exhibited very high precision (XGBoost at 0.6188, Neural Network highest at 0.6993) coupled with extremely low recall (ranging from 0.0169 to 0.0634). This notable pattern confirms the severe class imbalance of the default data and suggests the models are highly selective and accurate when predicting a positive default, but they miss the vast majority of actual positive cases (high number of false negatives), which explains why the F1-Scores remain low across the board. Figure 3 displays the ROC curves for all five models.

Figure 3: ROC Curves for All Models



5.2 Model Explainability using SHAP

Shapley values explain how each feature contributes to the prediction by evaluating how various feature value combinations would affect the prediction outcome compared to the average prediction output (Lundberg & Lee, 2017). It would reveal the most important features for default prediction across all models. To ensure a fair comparison of model explainability, SHAP values were calculated for all classifiers using the most appropriate and efficient method for each architecture: LinearExplainer was used for the Logistic Regression model to compute exact SHAP values directly from its coefficients; TreeExplainer was applied to the tree-based models (LightGBM and XGBoost), taking advantage of their underlying structure for efficient computation; and KernelExplainer provided a model-agnostic, sampled approximation of the SHAP values for the Neural Network. This comprehensive SHAP analysis facilitates both global feature importance assessment (determining which features are most impactful overall) to assess the interpretability

of complex models.

The images below (Figure 4) show the SHAP summary plots for each model, illustrating the impact of the top 20 features on model output. The color represents the feature value (red = high, blue = low), while the position on the x-axis indicates the SHAP value (positive values increase predicted default risk, negative values decrease it).

The top five most important features across models were:

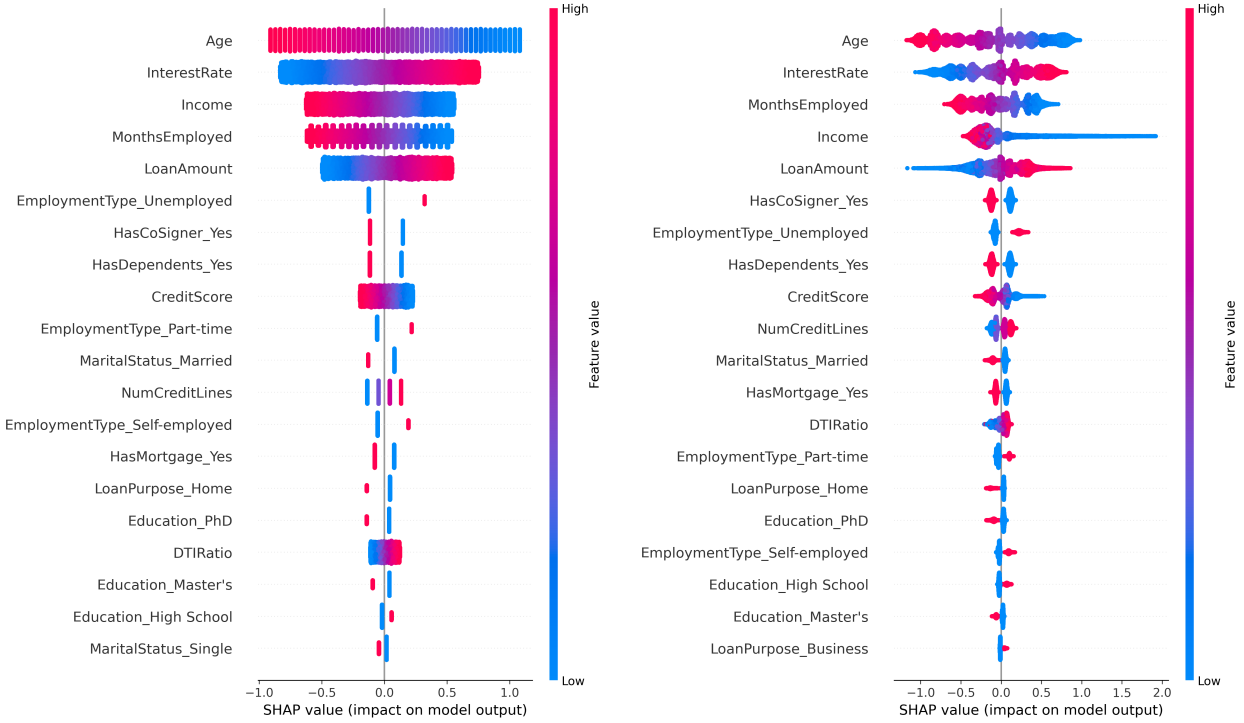
- **Age** (The age of the borrower) is the most importance feature across all models. We see that younger borrowers (lower age values) tend to have higher SHAP values, indicating a higher predicted risk of default. This suggests that younger individuals may be perceived as riskier borrowers, possibly due to less established credit histories or financial stability.

- **InterestRate** (The interest rate for the loan) is the second most important feature. Higher interest rates (red points) are associated with higher SHAP values, indicating that loans with higher interest rates contribute to an increased predicted risk of default. This aligns with the intuition that higher borrowing costs may strain a borrower's ability to repay.

- **Income** (The annual income of the borrower) shows that lower income levels (blue points) correspond to higher SHAP values, suggesting that borrowers with lower incomes are more likely to default. This reflects the financial vulnerability of lower-income individuals in meeting debt obligations.

- **MonthsEmployed** (The number of months the borrower has been employed) indicates that borrowers with shorter employment durations (blue points) have higher SHAP values, implying a higher risk of default. This may be due to job instability or lack of steady income.

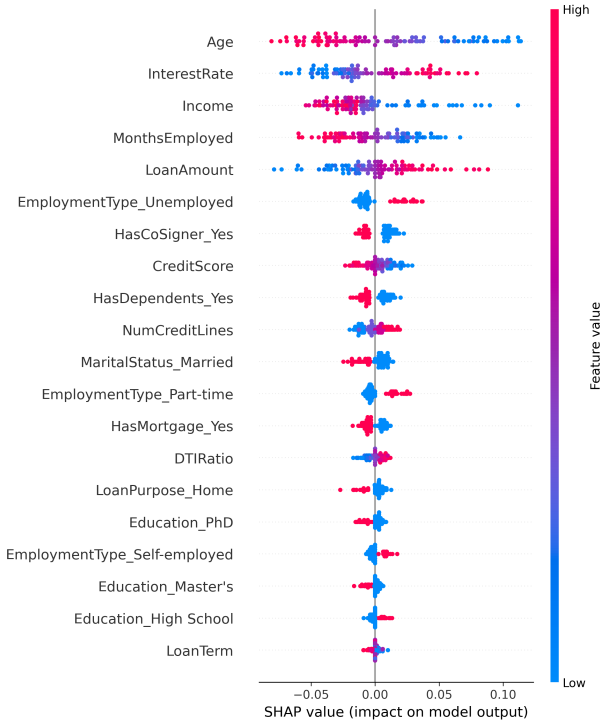
- **LoanAmount** (The amount of money being borrowed) shows that larger loan amounts (red points) are associated with higher SHAP values, indicating that borrowing larger sums increases the predicted risk of default. This could be due to the increased financial burden associated with repaying larger loans.



(a) Logistic Regression

(b) XGBoost

(c) LightGBM



(d) Neural Network

Figure 4: SHAP Summary Plots for Tested Models

While the top features are rather straightforward. It is the consistency of them showing up across models that shows the promising utility of SHAP in providing sufficient explainability for complex models. Although the magnitude of their SHAP values varied. For instance, Age had a more pronounced impact in the Logistic Regression model compared to the Neural Network, where its influence was more moderate. This variation suggests that while certain features are universally important, their relative contributions can differ based on the model architecture and complexity.

5.3 Limitations and Future Work

Despite the methodology employed, the results of our findings are subject to two primary limitations. Firstly, the study relied on a public dataset comprising approximately **255,000 observations**, which is relatively small for effectively training and optimizing complex, high-capacity models like Neural Networks, especially given the inherent class imbalance in credit default prediction. This limited data size may have constrained the true predictive potential of the more complex models, leading to the observed close performance clustering near the simpler Logistic Regression baseline. Secondly, while several state-of-the-art models were employed, our neural network architecture and hyperparameter tuning were not exhaustively optimized compared to the state of the art tree-based models being used (LightGBM and XGBoost). The modest performance of the Neural Network (ROC-AUC 0.7586) suggests that with more extensive search (e.g., deeper layers, different activation functions, or longer training epochs), its performance might improve significantly, potentially widening the performance gap between tree-based methods and deep learning. Lastly, we noted that the feature sizes are rather limited (around 20), which may not fully mirror real-world production-scale datasets where there may be hundred's of features being used. In that instance, neural network models could perform better and the SHAP values could start diverging across models. Future research should address these limitations by validating the framework on larger, production-scale datasets and incorporating more advanced and fine-tuned machine learning techniques to ensure the models are fully optimized to their best performance.

6 Conclusion

This study set out to address fundamental questions about the application of machine learning to credit default prediction: which models perform best, is there performance gain being left on the table by financial institutions due to regulatory requirements, and whether explainable AI techniques can be the bridge to understand 'black-box' models better.

Our analysis revealed several findings, starting with the model performance ranking. The XGBoost model achieved the highest predictive performance with an ROC-AUC of 0.759. Surprisingly, our neural-network model had an ROC-AUC of 0.758, less than XGBoost and indicates that tree-based models may still provide the best performance compared to more advanced models. Furthermore, SHAP analysis successfully provided interpretable explanations across all models, with a high degree of consistency in the ranking of the most important features (e.g., age, income, interest rate). However, the ultimate sufficiency of SHAP for full regulatory compliance remains context-dependent and requires further industry clarification.

This research makes several contributions to both academic literature and industry practice. Primarily, it explored the performance sacrifice required for interpretability in this domain by quantifying the minor ROC-AUC difference between complex and simple models. Secondly, the study evaluates SHAP across multiple model types (including Linear, Tree, and Neural Network), assessing its potential to enable the safe deployment of otherwise black-box models in a regulated financial environment.

For financial institutions navigating the performance-interpretability tradeoff, flexibility in approach is recommended. For a conservative approach, firms should deploy the highly interpretable models, such as Logistic Regression or Decision Trees (XGBoost), in scenarios where regulatory requirements are strict or uncertain. Conversely, an innovative approach involves utilizing more advanced complex models (neural network) coupled with SHAP analysis for generating explanations; this would require a strong partnership with regulators to establish the acceptability of SHAP-based adverse action (rejection) notices. The future of credit scoring lies not in choosing between performance and interpretability, but in developing methods that optimize both simultaneously. As machine learning continues to advance, the credit industry must balance three competing imperatives: maximizing predictive accuracy to manage risk, maintaining transparency to satisfy

regulations, and ensuring fairness to serve all consumers equitably. The path forward requires continued collaboration between researchers, practitioners, and regulators to develop credit scoring systems that are simultaneously accurate, interpretable, and fair.

References

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.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2019). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*.

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, *29*(4), 563–574.

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.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*(1), 203–216.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Consumer Financial Protection Bureau. (2022). *Cfpb acts to protect the public from black-box credit models using complex algorithms* (tech. rep.). Consumer Financial Protection Bureau. https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/

Deloitte. (2018). *Ai and risk management: Innovating with confidence* (tech. rep.). Deloitte Insights. https://www.deloitte.com/global/en/Industries/financial-services/perspectives/gx-ai-and-risk-management.html

Einav, L., Jenkins, M., & Levin, J. (2013). The impact of credit scoring on consumer lending. *RAND Journal of Economics*, *44*(2), 249–274.

Fahner, Gerald. (2018). *Developing transparent credit risk scorecards more effectively: An explainable artificial intelligence approach* (tech. rep.) (White Paper). FICO. https://www.thinkmind.org/index.php?view=article&articleid=data_analytics_2018_1_30_60077

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble learning or deep learning? application to default risk analysis. *Journal of Risk and Financial Management*, *11*(1), 12.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd). John Wiley & Sons.

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, *18*, 148–216.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

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.

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, *102*, 207–217.

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.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

McKinsey & Company. (2017). *The future of risk management in the digital era* (tech. rep.). McKinsey Global Institute. https://www.mckinsey.com

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Self-published. https://christophm.github.io/interpretable-ml-book/

PwC. (2018). *Explainable ai: Driving business value through greater understanding* (tech. rep.). PwC. https://www.pwc.com

Report to congress on credit scoring and its effects on the availability and affordability of credit (tech. rep.). (2007). Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*.

U.S. Congress. (1974). Equal credit opportunity act, 15 u.s.c. § 1691 [Regulation B - Equal Credit Opportunity]. https://www.govinfo.gov

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241.