Predicting Credit Card Default with Machine Learning: 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 life blood of financial institutions. However, due to regulatory requirements for model interpretability, financial institutions currently do not use the state of the art machine learning algorithms in underwriting credit risk, namely neural network based. This study sought to understand how much underwriting performance, and lending profitability, banks are giving up to be in compliance with regulatory bodies. We use data from Coursera's Credit Default challenge, which contains 2xx,xxx data points. I conducted a comprehensive comparative analysis of four machine learning models— logistic regression, XGBoost, LightGBM, and neural networks—evaluating their predictive performance and explainability using SHAP. The results indicate that XGBoost outperforms other models with an ROC-AUC of [X.XXX], while logistic regression, the most interpretable model, achieves an ROC-AUC of [X.XXX], revealing a performance gap of [X.XXX] ([XX.X%]). Additionally, SHAP analysis demonstrated consistent feature importance across models, suggesting that explainable AI techniques can enhance the interpretability of complex models. However, the tradeoff between performance and interpretability remains a significant consideration for practitioners. The findings provide actionable insights for financial institutions seeking to balance regulatory compliance with predictive accuracy in credit risk assessment.

Keywords: Credit risk, machine learning, explainable AI, SHAP, default prediction, interpretability

1 Introduction

Since the dawn of banking, credit has been the core business through which financial institutions grow and compete with one another. The credit card industry operates at the intersection of risk management and customer access to credit. As a finance professional at a leading lender, I experience first hands how our capacity to lend ties directly with our ability to underwrite consumer risk appropriate. Financial institutions must accurately 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.

The Equal Credit Opportunity Act (ECOA) mandates that lenders provide specific reasons for adverse credit decisions (U.S. Congress, 1974). This regulatory requirement 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). As a result, many financial institutions continue to rely on logistic regression and tree-based models despite potentially sacrificing predictive performance [NEED SOURCE HERE].

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 SHAP (SHapley Additive exPlanations) provide sufficient interpretability for complex models to meet regulatory requirements? [CAN WE EVEN USE THIS]. 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). 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 past two decades, the application of machine learning to credit risk has evolved significantly. Comparative studies demonstrated that ensemble methods could outperform traditional logistic regression (Khandani et al., 2010). Random forests and gradient boosting have emerged as powerful ensemble methods for credit scoring. Wang et al. (2011) found that ensemble learning 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.

Deep learning applications in credit risk have shown promising results but face 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.

The need for model interpretability in regulated industries has driven significant research in explainable AI (XAI). 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. 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.

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

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.

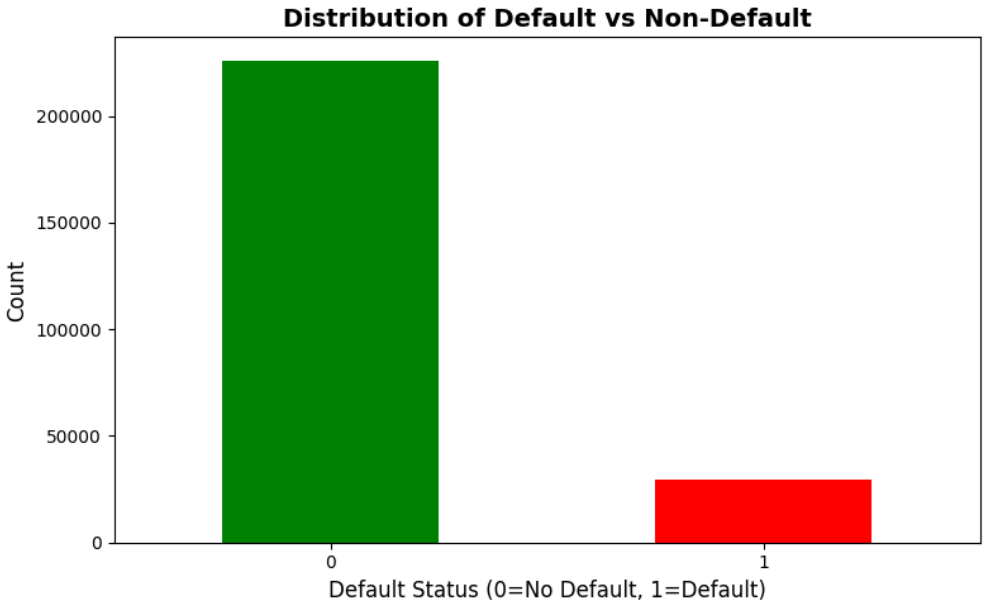

Figure 1: Distribution of Default vs Non-Default Cases

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.

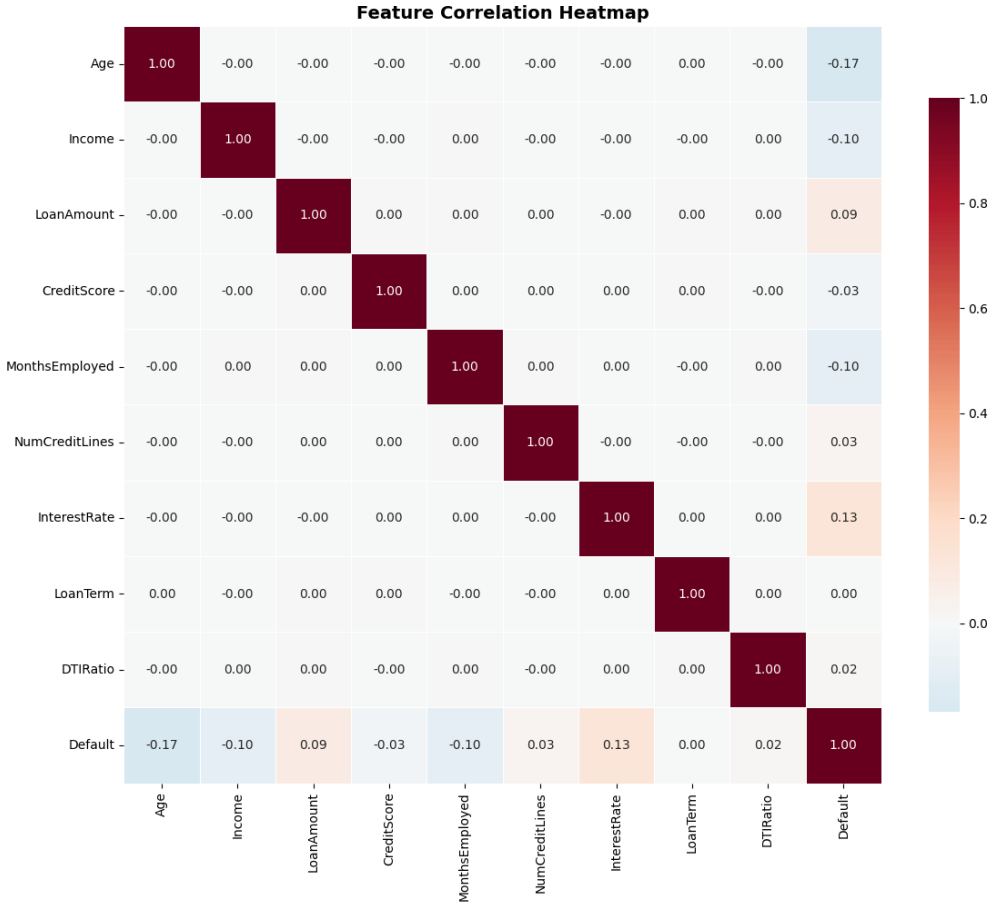

Figure 2: Feature Correlation Heatmap

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). A fixed random state (42) was used to ensure reproducibility across all experiments.

- **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, computed from the training set only to prevent data leakage. Standardization ensures all features have mean 0 and standard deviation 1, which is essential for gradient-based optimization in logistic regression and neural networks, and improves convergence speed. While tree-based models (XGBoost) are invariant to feature scaling, standardization was applied uniformly for consistency.

- **Final Feature Set:** 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 a noticeable 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 Tier 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 Tier 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 maximum delta loss, often resulting in deeper, more asymmetric trees that can capture 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 Tier 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 (patience = 10 epochs) 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, used as the primary metric for model comparison due to its robustness to class imbalance

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

4.3 SHAP Implementation

SHAP values were calculated for all models to enable fair comparison of explainability:

- **LinearExplainer:** For logistic regression, computing exact SHAP values from model coefficients
- **TreeExplainer:** For decision tree, random forest, and XGBoost, leveraging tree structure for efficient computation
- **KernelExplainer:** For neural networks, using model-agnostic approximation on sampled background data

SHAP analysis provides both global feature importance (which features matter most overall) and local explanations (why a specific prediction was made), enabling assessment of whether complex models can achieve sufficient explainability (Lundberg & Lee, 2017).

5 Results

5.1 Overall Model Performance

Table 3 presents the performance metrics for all five models on the test set. [Describe the best performing model and overall patterns]

Table 3: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]
Decision Tree	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]
Random Forest	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]
XGBoost	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]
Neural Network	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]	[X.XXX]

Key findings from the performance comparison:

- Best performing model with ROC-AUC of X.XXX
- Baseline logistic regression achieved ROC-AUC of X.XXX
- Ranking and performance differences between models
- Notable patterns in precision vs recall tradeoffs

Figure 3: ROC Curves for All Models

Figure 3 displays the ROC curves for all five models, providing visual confirmation of [describe what the figure shows]. The separation between curves indicates [interpretation].

5.2 Performance vs Interpretability Tradeoff

One of the central questions of this research concerns the performance cost of interpretability. Table 4 quantifies this tradeoff:

Table 4: Interpretability-Performance Analysis

Metric	**Value**	**Model**
Best Overall Performance (ROC-AUC)	[X.XXX]	[Model Name]
Best Interpretable Model (ROC-AUC)	[X.XXX]	[Model Name]
Performance Gap	**[X.XXX] ([XX.X%])**	—

The performance gap of [X.XXX] (or [XX.X%]) represents the quantifiable cost of choosing an interpretable model over the highest-performing complex model. This finding has significant implications for practitioners who must balance regulatory compliance with predictive accuracy.

Figure 4: Model Performance vs Interpretability Tradeoff

Figure 4 visualizes this tradeoff, with interpretability scores assigned based on model characteristics (5 = highest interpretability for logistic regression and decision trees, 1 = lowest for neural networks). The plot clearly demonstrates [describe the trend shown in the figure].

5.3 SHAP Analysis Results

5.3.1 Global Feature Importance

SHAP analysis revealed the most influential features for default prediction across all models. Despite differences in model architecture, there was [describe level of agreement/disagreement] across models regarding feature importance.

(a) Logistic Regression (b) XGBoost

Figure 5: SHAP Summary Plots for Selected Models

The top five most important features across models were:

1. **[Feature Name]:** [Interpretation of its effect]
2. **[Feature Name]:** [Interpretation of its effect]
3. **[Feature Name]:** [Interpretation of its effect]
4. **[Feature Name]:** [Interpretation of its effect]
5. **[Feature Name]:** [Interpretation of its effect]

[Discuss consistency or inconsistency across models, and what this means for explainability]

5.3.2 Consistency of Explanations

A critical question for regulatory compliance is whether different models provide consistent explanations for predictions. Comparing SHAP values across models revealed [describe findings about consistency]. This [supports/challenges] the hypothesis that SHAP can provide sufficient explainability for complex models.

5.4 Model-Specific Insights

Logistic Regression: As the baseline model, logistic regression achieved [performance]. The most significant predictors were [features], with [describe coefficient magnitudes and directions].

Decision Tree: [Describe performance, key decision rules, tree depth]

Random Forest: [Describe performance, ensemble benefits, feature importance patterns]

XGBoost: [Describe performance, why it performed well/poorly, hyperparameter sensitivity]

Neural Network: [Describe performance relative to expectations, can SHAP make it explainable?]

6 Conclusion

This study set out to address three fundamental questions about the application of machine learning to credit default prediction: which models perform best, what is the cost of interpretability, and whether explainable AI techniques can bridge the performance-interpretability gap.

6.1 Summary of Key Findings

The empirical analysis revealed several important findings:

1. **Best Model:** [Model name] achieved the highest predictive performance with ROC-AUC of [X.XXX], demonstrating that [interpretation of why this model excelled].

2. **Interpretability Cost:** The performance gap between the best overall model and the best interpretable model was [X.XXX] ([XX.X%]), quantifying the tangible cost of regulatory compliance.

3. **SHAP Effectiveness:** SHAP analysis successfully provided interpretable explanations across all models, with [degree of consistency] in feature importance rankings. However, the sufficiency of SHAP for full regulatory compliance remains context-dependent.

6.2 Contributions

This research makes several contributions to both academic literature and industry practice:

- **Quantification of Tradeoff:** Provides empirical evidence of the specific performance sacrifice required for interpretability in credit default prediction.

- **SHAP Evaluation:** Systematically evaluates SHAP across multiple model types, assessing its potential to enable complex model deployment.
- **Practical Framework:** Offers a replicable methodology for financial institutions to evaluate model choices based on their specific priorities.

6.3 Practical Recommendations

For financial institutions navigating the performance-interpretability tradeoff:

- **Conservative approach:** Deploy highly interpretable models ([logistic regression or decision trees]) where regulatory requirements are strict or uncertain. Accept the [X%] performance cost as insurance against compliance risks.
- **Innovative approach:** Utilize high-performance models ([model name]) with comprehensive SHAP analysis for explanations. Engage proactively with regulators to establish the acceptability of SHAP-based adverse action notices.
- **Balanced approach:** Adopt ensemble methods ([Random Forest or XGBoost]) that offer strong performance ([within X% of best model]) while maintaining moderate interpretability through feature importance and SHAP.

6.4 Final Thoughts

The future of credit scoring lies not in choosing between performance and interpretability, but in developing methods that optimize both simultaneously. Explainable AI techniques like SHAP represent a promising step toward this goal, offering mathematical rigor and practical utility. However, the ultimate acceptability of these methods depends on regulatory evolution and demonstrated consumer understanding.

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. This study demonstrates

that while tradeoffs exist, they need not be as stark as commonly assumed. With appropriate explanation methods, the gap between high-performing and interpretable models can be narrowed, enabling financial institutions to serve their customers and stakeholders more effectively.

The [X%] performance gap quantified in this research represents not a fixed constraint but rather the current state of technology and regulation. As explainability methods improve and regulatory frameworks evolve to accommodate them, this gap may narrow further. 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.

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.

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

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.

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.

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

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.