Predicting Credit Card Default Risk with Machine Learning: A Comparative Analysis of Model Performance and Explainability

Hoang To

ht8758

University of Texas at Austin

hto@utexas.edu

December 01, 2025

Abstract

Credit underwriting is the core area for financial institutions aiming to maximize lending profitability while managing risk. Due to regulatory requirements for model interpretability, financial institutions do not use the state of the art algorithms in underwriting credit risk. This study sought to illustrate how much incremental performance these companies are leaving on the table to be in compliance with regulatory bodies. I conducted a comprehensive comparative analysis of five machine learning models—logistic regression, decision tree, random forest, XGBoost, and neural networks—evaluating their predictive performance and explainability using SHAP. The research addresses three key questions: which model achieves the highest accuracy, what is the performance cost of interpretability, and can SHAP provide sufficient explanations for complex models? Using a publicly available credit dataset, models were trained and evaluated based on ROC-AUC, precision, recall, and F1-score.

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%]). 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

The credit card industry operates at the intersection of risk management and customer access to credit. 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.

[Expand on your personal motivation for this research, your experience in the fintech industry, and the specific problem you observed...]

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 traditional logistic regression models despite potentially sacrificing predictive performance.

This study addresses three fundamental research questions:

1. Which machine learning models achieve the highest predictive accuracy for credit card default prediction?

2. What is the quantifiable performance cost when financial institutions choose interpretable models over complex alternatives?

3. Can SHAP (SHapley Additive exPlanations) provide sufficient interpretability for complex models to meet regulatory requirements?

By comparing five distinct machine learning models—ranging from highly interpretable (logistic regression) to complex "black box" models (neural networks)—and employing SHAP values for explainability analysis, this research quantifies the interpretability-performance tradeoff that financial institutions face. The findings provide actionable insights for practitioners navigating the balance between model accuracy and regulatory compliance.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature on credit scoring, machine learning applications, and explainable AI. Section 3 describes the

methodology, including model selection and evaluation metrics. Section 4 details the dataset and preprocessing steps. Section 5 presents the empirical results. Section 6 discusses the findings and their implications. Section 7 concludes with recommendations and future research directions.

2 Literature Review

2.1 Traditional Credit Scoring Methods

Credit scoring has been a cornerstone 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.

[Expand on the history, strengths, and limitations of traditional methods...]

2.2 Machine Learning in Credit Risk Assessment

The application of machine learning to credit risk has evolved significantly over the past two decades. Early comparative studies demonstrated that ensemble methods and neural networks could outperform traditional logistic regression (Khandani et al., 2010). Lessmann et al. (2015) conducted a comprehensive benchmark of classification algorithms for credit scoring, finding that ensemble methods consistently ranked among the top performers across multiple datasets.

More recent research has explored the application of specific ML techniques to credit default prediction. Brown and Mues (2012) examined the challenges of class imbalance in credit scoring datasets, while Bellotti and Crook (2013) focused on dynamic models for credit card default prediction.

[Continue with more literature on ML applications in credit risk...]

2.3 Ensemble Methods and Deep Learning

Random forests (Breiman, 2001) and gradient boosting machines (Friedman, 2001) 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.

[Add more discussion on neural networks in finance...]

2.4 Explainable AI in Financial Services

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) extended this 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.

[Expand on XAI applications in lending and regulatory perspectives...]

2.5 Research Gap

While extensive research has compared ML models for credit scoring and explored explainability techniques separately, few studies have systematically quantified the interpretability-performance tradeoff across a spectrum of models while applying SHAP analysis consistently. This research addresses this gap by providing empirical evidence of the performance cost of interpretability and evaluating whether SHAP can make complex models sufficiently explainable for regulatory compliance.

3 Methodology

3.1 Research Design

This study employs a comparative experimental approach to evaluate five machine learning models across the interpretability spectrum. The research follows a quantitative methodology with cross-validation and standardized evaluation metrics to ensure robust and reproducible results.

3.2 Model Selection

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

3.2.1 Tier 1: Highly Interpretable Models

Logistic Regression: Selected as the baseline model due to its widespread use in the credit industry and direct coefficient interpretation. Logistic regression provides odds ratios for each feature, making it straightforward to explain predictions to regulators and customers.

Decision Tree: Chosen for its visual interpretability and rule-based decision structure. Decision trees can be directly translated into if-then rules, offering complete transparency in the decision-making process.

3.2.2 Tier 2: Moderately Interpretable Models

Random Forest: An ensemble method that aggregates predictions from multiple decision trees. While individual tree predictions are interpretable, the ensemble nature introduces complexity. However, feature importance measures provide global interpretability.

XGBoost (Gradient Boosting): A state-of-the-art gradient boosting implementation known for high performance in structured data tasks. XGBoost provides feature importance scores and supports SHAP analysis, offering a balance between performance and explainability.

3.2.3 Tier 3: Low Interpretability Models

Neural Network: A feedforward neural network with multiple hidden layers represents the "black box" end of the spectrum. Neural networks can capture complex non-linear relationships but lack inherent interpretability without external explanation methods like SHAP.

3.3 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).

3.4 Hyperparameter Tuning

For each model (except logistic regression), hyperparameter optimization was performed using GridSearchCV with 5-fold stratified cross-validation. Stratified folds maintain class distribution in each fold, which is crucial for imbalanced datasets. The optimization metric was ROC-AUC to identify configurations that maximize discriminative ability.

3.5 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).

4 Data and Preprocessing

4.1 Dataset Description

This study utilizes the Loan Default Prediction dataset from Kaggle, which contains borrower information for credit default prediction. The dataset includes [insert sample size] observations with [insert features] features and 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)

4.2 Exploratory Data Analysis

The dataset exhibits [describe class distribution, e.g., "a default rate of X%, indicating moderate class imbalance"]. Key observations from exploratory analysis include:

- Default rate: [X%]

feature distributions

ns between features

y data quality issues

4.3 Data Preprocessing

The following preprocessing steps were applied:

1. **Train-Test Split:** The data was split 70% for training and 30% for testing prior to any preprocessing to prevent data leakage.

2. **Categorical Encoding:** One-hot encoding was applied to all categorical variables (Education, EmploymentType, MaritalStatus, HasMortgage, HasDependents, LoanPurpose, HasCoSigner), creating binary indicator variables while dropping the first category to avoid multicollinearity.

3. **Feature Scaling:** Numerical features were standardized using StandardScaler to have mean 0 and standard deviation 1. This is particularly important for logistic regression and neural networks, which are sensitive to feature scales.

4. **Final Feature Set:** After one-hot encoding, the final feature set consisted of [X] features.

4.4 Class Imbalance Considerations

With a default rate of [X%], the dataset exhibits [mild/moderate/severe] class imbalance. 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.

5 Results

5.1 Overall Model Performance

Table 2 presents the performance metrics for all five models on the test set. [Describe the best performing model and overall patterns]

Table 2: 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:

OC-AUC of X.XXX

OC-AUC of X.XXX

ces between models

n vs recall tradeoffs

Figure 1: ROC Curves for All Models

Figure 1 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 3 quantifies this tradeoff:

Table 3: 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 2: Model Performance vs Interpretability Tradeoff

Figure 2 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 3: 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 Discussion

6.1 Interpretation of Findings

6.1.1 Research Question 1: Best Performing Model

[Model Name] achieved the highest ROC-AUC of [X.XXX], outperforming the baseline logistic regression by [X.X%]. This result [aligns with/contradicts] previous findings in the literature (Lessmann et al., 2015). The superior performance can be attributed to [discuss why this model performed best - ability to capture non-linear relationships, ensemble benefits, etc.].

6.1.2 Research Question 2: Cost of Interpretability

The quantified performance gap of [X.XXX] ([XX.X%]) between the best overall model and the best interpretable model represents the tangible cost that financial institutions pay for regulatory

compliance. From a business perspective, this translates to:

Estimated number of additional missed defaults per [time period]

Financial impact in potential losses

- Conversely, fewer false alarms leading to [customer experience benefits]

Whether this cost is acceptable depends on institutional priorities and regulatory constraints. For institutions facing strict regulatory scrutiny or serving populations with fairness concerns, the interpretability benefits may justify the performance sacrifice.

6.1.3 Research Question 3: SHAP Effectiveness

The SHAP analysis demonstrated that complex models can be made substantially more explainable. Key findings include:

- SHAP values provided consistent feature importance rankings across [most/all] models
- Individual prediction explanations identified the specific features driving each decision

nsistencies observed

However, the question of whether SHAP explanations satisfy regulatory requirements remains complex. While SHAP provides mathematically rigorous explanations based on cooperative game theory (Lundberg & Lee, 2017), regulatory agencies may still prefer the direct interpretability of logistic regression coefficients.

6.2 Comparison to Literature

These findings [align with/extend/contradict] existing research in several ways:

- Lessmann et al. (2015) found that [comparison point]
- Bussmann et al. (2021) argued that [comparison point]
- This study's novel contribution is [what makes your findings unique]

6.3 Business and Regulatory Implications

6.3.1 Recommendations for Lenders

Based on these findings, financial institutions should consider the following strategic approaches:

If interpretability is paramount: Use [most interpretable model] with SHAP analysis for additional insights. Accept the [X%] performance cost as the price of regulatory certainty.

If performance is critical: Deploy [best performing model] with comprehensive SHAP analysis for adverse action explanations. Engage proactively with regulators to establish acceptability of SHAP-based explanations.

Balanced approach: [Middle-ground model like Random Forest] offers [performance characteristics] while maintaining [interpretability characteristics].

6.3.2 Regulatory Considerations

Regulators face the challenge of balancing consumer protection with innovation. This research suggests that SHAP-enabled complex models could satisfy the spirit of adverse action requirements while enabling better risk assessment. Policy recommendations include:

- Develop guidelines for acceptable explanation methods beyond coefficient interpretation
- Require lenders to validate that SHAP explanations are accessible to consumers
- Consider performance-interpretability tradeoffs in model approval processes

6.4 Limitations

This study has several limitations that should be considered when interpreting results:

1. **Single Dataset:** Results are based on one credit dataset. Findings may not generalize to all credit portfolios or lending contexts.
2. **Hyperparameter Optimization:** Due to computational constraints, hyperparameter search spaces were limited. More extensive tuning might improve model performance.

3. **Class Imbalance:** While ROC-AUC is robust to imbalance, real-world default rates may differ from the dataset, affecting model performance.

4. **SHAP Computational Cost:** Calculating SHAP values for large datasets or complex models can be computationally expensive, potentially limiting real-time application.

5. **Model Selection:** Other models (e.g., Support Vector Machines, LightGBM) were not evaluated and might offer different performance-interpretability tradeoffs.

6. **Feature Engineering:** Limited feature engineering was performed. Domain-specific features or interaction terms might improve performance.

7. **Temporal Validation:** The study uses a static train-test split. Performance may degrade over time as borrower behavior evolves.

6.5 Future Research Directions

Several avenues for future research emerge from this study:

1. **Multi-Dataset Validation:** Replicate the analysis across multiple credit datasets from different industries and geographies to assess generalizability.

2. **Deep Learning Architectures:** Explore more sophisticated neural network architectures (e.g., attention mechanisms, graph neural networks) and their explainability.

3. **Alternative Data Sources:** Incorporate alternative credit data (e.g., utility payments, rent history) and evaluate its impact on the interpretability-performance tradeoff.

4. **Temporal Analysis:** Conduct longitudinal studies to understand how model performance and explanations evolve over time.

5. **Fairness Analysis:** Examine whether different models exhibit different levels of bias across demographic groups and how SHAP can support fairness auditing.

6. **Cost-Sensitive Learning:** Incorporate business-specific costs of false positives and false negatives into model optimization.

7. **Human Studies:** Conduct user studies with regulators and consumers to evaluate whether SHAP explanations are actually understandable and satisfactory.

8. **Hybrid Approaches:** Develop models that explicitly optimize for both performance and interpretability simultaneously.

7 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.

7.1 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.

7.2 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.

7.3 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.

7.4 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.

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

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

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.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

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.

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

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.

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.

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.

A Hyperparameter Configurations

[Include the final hyperparameters selected through GridSearch for each model]

B Additional Visualizations

C Code Availability

The complete code for this analysis is available at: [GitHub repository link or statement about availability upon request]

All analysis was conducted using Python 3.11 with the following key libraries: scikit-learn 1.2.0, XGBoost 1.7.0, TensorFlow 2.10.0, SHAP 0.41.0, and standard data science packages (pandas, numpy, matplotlib).