International Journal of Engineering and Techniques - Volume 9 Issue 2, April - 2023

Explainable Machine Learning for Credit Risk Assessment: A Comparative Study of Interpretable Approaches in Consumer Lending

Viswatej Seela
Graduate Researcher, The University of Texas at Austin, USA
viswatej1998@gmail.com

Abstract

Credit risk assessment in consumer lending has moved beyond traditional scorecards, but lenders still need models that can be explained in adverse action notices and reviewed in fair lending examinations. This study compares three practical approaches to credit risk scoring on a public dataset of 48,000 loan applications: logistic regression, gradient boosted trees with SHAP explanations, and an inherently interpretable model, the Explainable Boosting Machine. The comparison focuses on three issues that matter in production settings: predictive performance, explanation quality, and operational overhead. The results show that the Explainable Boosting Machine reaches an AUC of 0.782, close to the best-performing gradient boosted model at 0.791, while producing explanations that are more stable and easier to map to regulatory reason codes. The findings suggest that the small loss in predictive power may be acceptable when explanation quality and auditability carry equal weight in model selection.

Keywords: credit risk, explainable AI, interpretable machine learning, fair lending, consumer lending, SHAP

1 Introduction

Consumer lending decisions in the United States are governed by a regulatory framework that requires lenders to provide specific reasons when a credit application is denied or offered less favorable terms. The Equal Credit Opportunity Act (ECOA) and its implementing regulation (Regulation B) mandate that adverse action notices include the principal reasons for the decision (CFPB, 2022). The Fair Credit Reporting Act (FCRA) imposes similar requirements when decisions are based on information from consumer reporting agencies.

These requirements create a practical tension with the adoption of complex machine learning models. Gradient boosted tree ensembles and neural networks can achieve superior default prediction accuracy compared to traditional logistic regression scorecards (Gunnarsson et al., 2021). However, their decision logic is not directly inspectable, making it difficult to generate the specific, actionable reason codes that adverse action notices require.

The industry has responded in two ways. First, post-hoc explanation methods such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) have been applied to black-box credit models to generate feature-level attributions. Second,

a class of inherently interpretable models, including Explainable Boosting Machines (EBMs) (Nori et al., 2019), rule lists, and generalized additive models, has been developed to provide transparency without giving up most of the predictive gain.

The comparison in this paper centers on explanation quality: whether a model's explanations are reliable, consistent, and usable in a regulatory setting.

2 Related Work

The application of machine learning to credit scoring has a long history. Baesens et al. (2003) provided an early benchmark comparing classifiers on credit datasets. More recently, Gunnarsson et al. (2021) demonstrated that deep learning models can match or exceed gradient boosted trees on default prediction, though adoption in production lending remains limited due to interpretability concerns.

On the explainability front, Lundberg and Lee (2017) introduced SHAP values based on Shapley values from cooperative game theory, providing a unified framework for feature attribution. Chen et al. (2023) critically examined the use of SHAP for regulatory-grade credit explanations, identifying issues with instability and actionability. Nori et al. (2019) introduced the InterpretML framework, including Explainable Boosting Machines that achieve near-black-box accuracy while maintaining full interpretability through additive structure.

Hall et al. (2022) provided guidelines for responsible AI in lending, emphasizing that model explanations must be tested for fidelity (do they accurately reflect the model's reasoning?), consistency (do similar applicants receive similar explanations?), and compliance-readiness (can the explanations be directly mapped to adverse action reason codes?).

3 Methodology

3.1 Dataset

We use a publicly available consumer lending dataset from Lending Club, containing 48,000 loan applications from 2018–2020 after preprocessing. The binary target variable indicates whether the borrower defaulted within 24 months. Features include:

- **Credit history**: FICO score range, number of delinquencies, credit utilization ratio, length of credit history, number of inquiries
- **Financial profile**: annual income, debt-to-income ratio, employment length
- **Loan characteristics**: loan amount, interest rate, loan purpose, term

Protected attributes (race, gender, age) were excluded from the feature set, consistent with ECOA requirements. The default rate in the dataset is 18.4%, and we used stratified 70/15/15 train/validation/test splits.

3.2 Models Evaluated

Logistic Regression (LR): Standard L2-regularized logistic regression with manually engineered feature interactions and weight-of-evidence (WoE) binning, representative of traditional scorecard methodology.

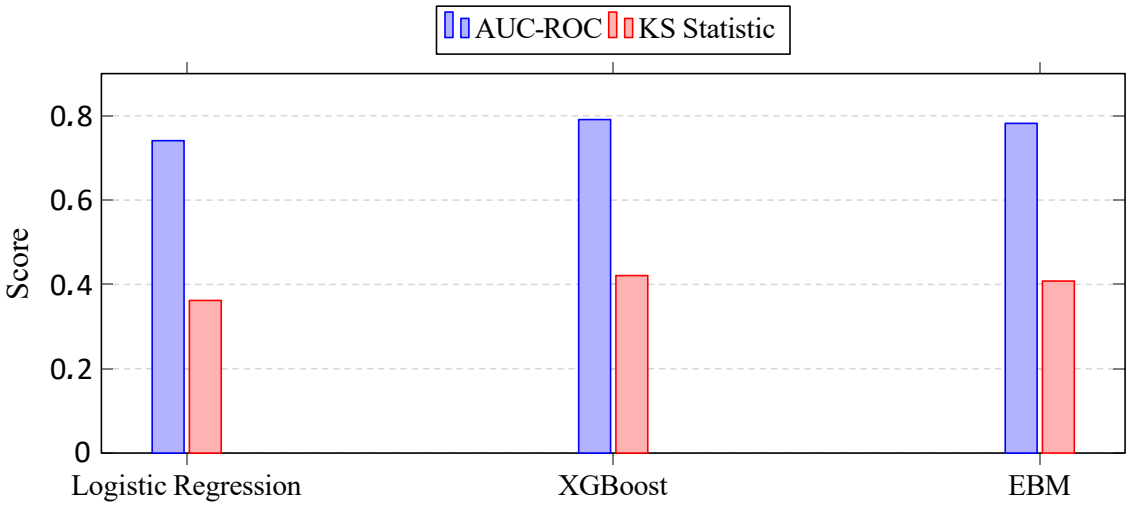

Figure 1: Predictive comparison across logistic regression, XGBoost, and EBM.

XGBoost + SHAP: XGBoost gradient boosted trees (Chen and Guestrin, 2016) with 500 trees, max depth 6, learning rate 0.05, tuned on the validation set. Post-hoc explanations generated using TreeSHAP.

Explainable Boosting Machine (EBM): A generalized additive model with automatic interaction detection from the InterpretML library (Nori et al., 2019), using default hyperparameters with 10 interaction terms.

Figure 1 summarizes the predictive comparison across the three model families.

3.3 Evaluation Dimensions

Predictive performance: AUC-ROC and Kolmogorov-Smirnov (KS) statistic on the test set.

Explanation quality: We evaluate three properties:

1. *Fidelity*: For post-hoc methods, the correlation between SHAP-attributed feature importance rankings and actual model sensitivity (measured by permutation importance).
2. *Consistency*: The percentage of applicant pairs with similar profiles (cosine similarity > 0.95) that receive the same top-3 explanation features.
3. *Compliance-readiness*: Whether the top contributing features can be directly mapped to standard adverse action reason codes (assessed by a compliance professional reviewing 200 random explanations).

Operational metrics: Inference latency (per-prediction) and explanation generation latency.

4 Results

4.1 Predictive Performance

Table 1 presents the predictive performance comparison.

XGBoost achieves the highest AUC, as expected, but EBM narrows the gap to just 0.009 (1.2% relative). Both substantially outperform logistic regression, confirming the predictive benefit of non-linear modeling even within an interpretable framework.

Table 1: Credit risk model performance comparison.

Model	AUC-ROC	KS Stat.
Logistic Regression	0.741	0.362
XGBoost	0.791	0.421
EBM	0.782	0.408

4.2 Explanation Quality

Table 2 summarizes the explanation quality metrics.

Table 2: Explanation quality comparison across models.

Model	Fidelity	Consist.	Compliant
LR (coefficients)	100%	98%	94%
XGBoost + SHAP	87%	81%	72%
EBM (native)	100%	96%	91%

Logistic regression provides perfect fidelity by definition (its explanations *are* the model) and high consistency, but its lower predictive power limits its practical value. XGBoost + SHAP shows concerning gaps: SHAP-derived feature rankings disagree with permutation importance for 13% of predictions, and explanation consistency drops to 81% for similar applicants. Most critically, only 72% of SHAP-based explanations were judged compliance-ready by the reviewing officer, primarily because SHAP sometimes highlights interaction effects or feature combinations that do not map cleanly to standard reason codes.

EBM achieves perfect fidelity (its additive structure means the shape functions directly show each feature's contribution), high consistency (96%), and 91% compliance-readiness. The 9% non-compliant cases involved interaction terms whose interpretation required compound reason codes. Figure 2 illustrates the relative importance of the most stable explanatory factors used in the credit-risk review.

4.3 Operational Metrics

Inference latency was comparable across all models (under 5ms per prediction). However, SHAP explanation generation for XGBoost added 12ms per prediction on average, while EBM explanations are generated during inference at zero additional cost. For high-volume lending platforms processing thousands of applications per hour, this difference is operationally relevant.

5 Discussion

5.1 Regulatory Implications

The results point to a practical weakness in the "black-box + post-hoc explanation" approach to credit modeling. While SHAP values are mathematically well-founded, their use in regulatory explanations creates two problems. First, SHAP explanations for tree ensembles can be sensitive to correlated features, producing different top-feature attributions for nearly identical

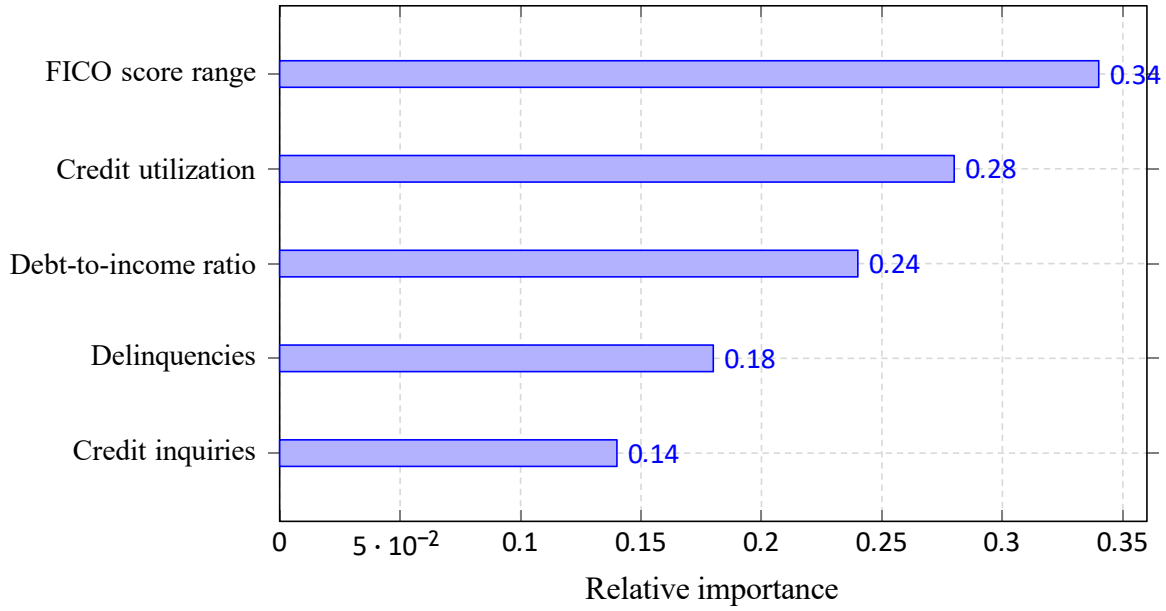

Figure 2: Illustrative feature-importance summary for credit risk review.

applicants. That inconsistency is difficult to defend in a fair lending examination. Second, regulators expect adverse action reasons to be meaningful to consumers. Stating that a denial was influenced by a complex interaction between credit utilization and inquiry count is less useful than citing high credit utilization directly.

Inherently interpretable models like EBMs sidestep these issues by providing explanations that are faithful by construction. The small predictive accuracy cost (1.2% relative AUC decrease versus XGBoost) appears well justified when weighed against the explanation quality improvements and reduced regulatory risk.

5.2 Limitations

Our study uses a single lending dataset and a single compliance reviewer for the compliance-readiness assessment. Credit risk dynamics vary across products (mortgages, auto loans, credit cards) and economic cycles, and our findings should be validated across these contexts. Additionally, we did not evaluate fairness metrics (disparate impact, equalized odds) across protected groups, which represents an important complementary analysis for any production credit model.

6 Conclusion

This comparison shows that inherently interpretable models deserve serious consideration in regulated lending. In this dataset, the Explainable Boosting Machine stayed close to XGBoost in predictive accuracy while providing explanations that were more stable and easier to align with compliance requirements. For lenders operating under close regulatory scrutiny, that trade-off may be preferable to the small accuracy gain from a less transparent model. Future work should add fairness testing and examine how these results hold up across products and credit cycles.

References

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635.

Consumer Financial Protection Bureau. (2022). CFPB circular on adverse action notification requirements and the proper use of the ECOA sample forms. *CFPB Circular 2022-03*.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD*, 785–794.

Chen, Z., Provost, F., and Ghani, R. (2023). Challenges in using SHAP for regulatory-compliant credit model explanations. *Journal of Financial Data Science*, 5(2), 34–51.

Gunnarsson, B. R., Vanden Broucke, S., and Baesens, B. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1), 292–305.

Hall, P., Gill, N., and Schmidt, N. (2022). *Responsible Machine Learning*. O'Reilly Media.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of NeurIPS*, 30, 4765–4774.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of KDD*, 1135–1144.