Balancing Interpretability and Predictive Performance in Regulated Machine Learning Models: A Comparative Evaluation Framework

Jim McCormack

May 13th, 2025

Abstract

This paper investigates the tradeoff between interpretability and predictive performance in machine learning models used for regulated decision-making. Using loan approval as a representative case study, we compare interpretable models (e.g., logistic regression) with high-performing black-box models (e.g., CatBoost) across multiple evaluation criteria, including ROC AUC, F1 Score, and a composite interpretability score. We introduce a structured framework to quantify interpretability based on traceability, intrinsic transparency, sparsity, and explanation complexity. Our results show that while black-box models outperform on predictive metrics, simpler models remain viable with well-structured designs, reinforcing the importance of model transparency in high-stakes, compliance-driven environments.

1 Introduction

As machine learning becomes integral to high-stakes domains like credit approval, insurance underwriting, and healthcare triage, institutions face a growing challenge: balancing the predictive power of complex models with the interpretability required by regulatory, legal, and ethical standards. This tradeoff is now central to operational AI design, especially as regulations like the European Union's AI Act and U.S. CFPB guidelines explicitly mandate that automated decisions must be explainable and auditable.(2; 4)

Black-box models such as ensemble methods and gradient boosting machines often outperform simpler models on key metrics such as ROC AUC and F1 Score. However, their opaque decision-making processes create barriers to transparency, governance, and consumer recourse. Conversely, interpretable models, such as logistic regression or rule-based classifiers, are easy to audit but may lack the predictive performance necessary in competitive environments.

This paper investigates whether and when the performance gap between interpretable and black-box models is significant enough to justify complexity. Using loan approval modeling as a case study, we empirically evaluate four classifiers across three data configurations. We propose a structured interpretability scoring framework to quantify model transparency and

assess tradeoffs directly. Our findings offer practical guidance for teams building machine learning systems in regulated industries, where interpretability is not optional, but required.

2 Background and Related Work

Credit scoring has evolved significantly over the past few decades, transitioning from traditional statistical methods such as logistic regression and linear discriminant analysis (10), to more advanced ensemble learning techniques and deep neural networks that offer improved predictive performance (7). However, as these machine learning methods become more complex, concerns have arisen regarding their lack of transparency and interpretability. Recent research highlights the need for explainable models in high-stakes areas like lending, where decisions significantly affect individuals' financial access and legal rights. (1).

In parallel, dimensionality reduction techniques like Principal Component Analysis (PCA) have been applied in fraud detection and credit scoring to reduce noise, improve model efficiency, and address multicollinearity (8). While these techniques can enhance performance, they often obscure the original meaning of input features, thus complicating interpretability, an essential consideration for regulators and practitioners alike.

3 Dataset and Experimental Setup

3.1 Dataset Description

We used a Kaggle-sourced synthetic dataset with 58,645 samples and 12 columns. The target variable is `loan_status`, indicating loan approval (1 = approved, 0 = rejected). Class distribution is imbalanced with 14.2% approved loans.

- **Demographic:** `person_age`, `person_income`, `person_home_ownership`, `person_emp_length`
- **Loan:** `loan_intent`, `loan_grade`, `loan_amnt`, `loan_int_rate`, `loan_percent_income`
- **Credit history:** `cb_person_default_on_file`, `cb_person_cred_hist_length`

3.2 Feature Engineering Strategies

1. **All Features (Quantity Focused)**: All 11 original input features were used with minimal preprocessing. Categorical variables were encoded using integer factorization, and numerical features were standardized for compatibility with linear models.

2. **Statistically Selected Features (Quality Focused)**: After applying outlier clipping using the interquartile range (IQR) method, categorical variables were transformed using target encoding, with one-hot encoding additionally applied to those with fewer than five unique values. The original categorical columns were then dropped, and univariate feature selection was performed using ANOVA F-value scoring via `SelectKBest`. The top 10 most informative features were retained. This approach balances dimensionality reduction with improved feature relevance and interpretability.

3. **Domain-Driven Features (Expertise Focused)**: This approach incorporates expert-informed transformations commonly used in credit risk analysis. Features include engineered variables such as debt-to-income ratio, age group bins, and a default risk factor combining default history and credit length. Additionally, categorical variables were risk-encoded based on their mean target value. After removing the original categorical columns, the final feature set included 14 attributes representing a blend of raw data and derived domain insights.

While we evaluate three distinct feature engineering strategies, our objective is not to determine which approach produces the best features per se. Instead, these scenarios provide controlled variations in data quality and dimensionality, allowing us to evaluate whether the relative tradeoff between model interpretability and predictive performance holds consistently across different input conditions. This ensures that observed differences between models are robust to feature set design and not artifacts of a specific preprocessing choice.

3.3 Models Compared

- Logistic Regression (L1 and L2)
- Random Forest Classifier
- CatBoost

3.4 Rationale for Excluding PCA

While Principal Component Analysis (PCA) is commonly used for dimensionality reduction in Loan scoring and fraud detection (5), we intentionally chose not to incorporate it for this exercise. We emphasized maximizing interpretability, particularly in regulated financial decision-making where transparency is essential.

Unlike PCA, which generates orthogonal components that are difficult to trace back to original features, regularization methods such as Lasso (L1) and Ridge (L2) regression retain the interpretability of feature importance by preserving the input variable space. These models help reduce multicollinearity and allow us to observe how feature variations influence model predictions, a key requirement when decisions must be explained to stakeholders and regulators (9; 1).

Mathematically, PCA transforms the original feature space $\mathbf{X} \in R^{n \times p}$ into a set of linearly uncorrelated components $\mathbf{Z} = \mathbf{XW}$, where $\mathbf{W}$ is a $p \times k$ matrix of eigenvectors of the covariance matrix $\mathbf{X}^\top \mathbf{X}$. Each principal component is a linear combination:

$$Z_j = \sum_{i=1}^{p} w_{ij} X_i \tag{1}$$

These components Z_j do not correspond to single original features, making it difficult to interpret which variables are driving predictions.

By contrast, Lasso and Ridge operate directly in the original feature space and optimize:

$$\min_{\mathbf{w}} [\mathcal{L}(\mathbf{w}) + \lambda |\mathbf{w}| 1] \quad \text{(Lasso)} \tag{2}$$

$$\min \mathbf{w} \left[\mathcal{L}(\mathbf{w}) + \lambda |\mathbf{w}|_2^2 \right] \quad \text{(Ridge)} \tag{3}$$

This retains direct access to feature weights w_i, enabling a transparent explanation of how each original input affects the outcome.

Therefore, although PCA may offer performance or efficiency gains, it introduces latent variables that obscure traceability and accountability, two factors critical in financial and regulatory environments where interpretability is not optional.

3.5 Evaluation Metrics

- **ROC AUC**: Model's ability to rank positive over negative samples across all thresholds
- **F1 Score**: Balances precision and recall
- **Accuracy**: Overall correctness
- **Interpretability**: Qualitative assessment based on model structure and feature transparency

3.6 Why ROC AUC Matters Most

Due to class imbalance and the need for threshold-independent evaluation, ROC AUC is the most appropriate metric in loan approval modeling. Unlike accuracy, it measures rank-order performance, which aligns with risk-based decision-making, where thresholds can shift. A model with high ROC AUC consistently ranks proper approvals higher than rejections, even if the cutoff changes. This robustness is critical in financial contexts where minimizing false approvals and denials has a monetary and legal impact.

Additionally, while ROC AUC is essential for assessing threshold-independent ranking, the F1 Score becomes particularly important for operational decision-making teams. It balances precision and recall, crucial in credit environments where false approvals (risk) and denials (missed revenue) carry significant costs.

4 Results

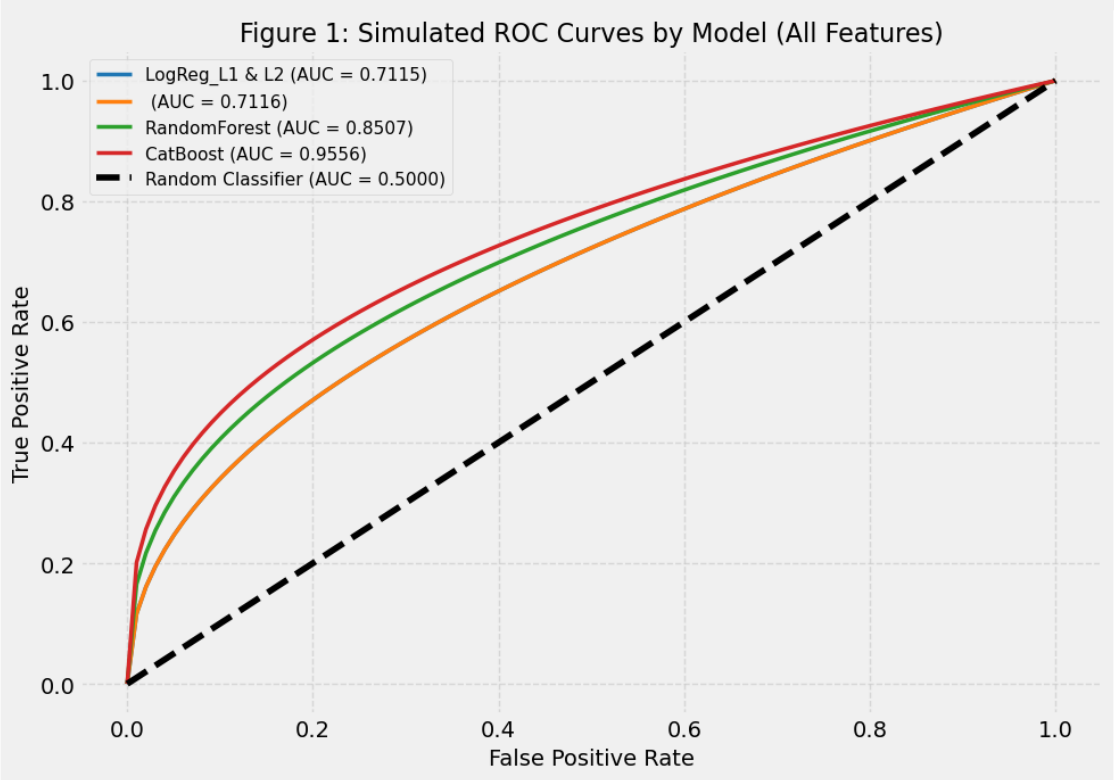

Figure 1: Simulated ROC curves for all models using all features. CatBoost dominates in true positive rate across thresholds, followed by Random Forest.

4.1 Model Performance

4.2 Discussion of Model Performance

The results in Table 1 highlight clear performance trends across different classifiers and feature engineering strategies. CatBoost consistently outperforms other models, achieving the highest ROC AUC (0.9556) and F1 Score (0.8156) when using all features. This is expected given CatBoost's ability to capture complex, non-linear interactions and its built-in handling of categorical variables without extensive preprocessing. Its gradient boosting framework enables strong predictive performance, particularly in imbalanced classification tasks such as Loan approval.

Random Forest also performs well, especially in accuracy, but slightly lags behind CatBoost in capturing minority class patterns, as reflected in its lower F1 and ROC AUC scores. While ensemble methods like Random Forest offer robustness and partial interpretability via feature importance, they are generally less expressive than gradient boosting regarding handling interactions and subtle variable dependencies.

Logistic regression with L1 (Lasso) and L2 (Ridge) regularization demonstrates the lowest ROC AUC values across all feature sets. This is due primarily to its linear nature, which

Table 1: Performance Across Feature Engineering Strategies

Model	Feature Set	Accuracy	F1 Score	ROC AUC
LogReg (L1)	All Features	0.8992	0.5592	0.7115
LogReg (L2)	All Features	0.8993	0.5595	0.7116
RandomForest	All Features	0.9510	0.8049	0.8507
CatBoost	All Features	0.9530	0.8156	**0.9556**
LogReg (L1)	Selected	0.9103	0.6230	0.7477
LogReg (L2)	Selected	0.9103	0.6228	0.7476
RandomForest	Selected	0.9329	0.7318	0.8121
CatBoost	Selected	0.9415	0.7668	0.8308
LogReg (L1)	Domain-Driven	0.9115	0.6287	0.7510
LogReg (L2)	Domain-Driven	0.9115	0.6290	0.7512
RandomForest	Domain-Driven	0.9499	0.7987	0.8449
CatBoost	Domain-Driven	0.9523	0.8138	0.8608

restricts its ability to model complex relationships in the data. However, these models remain valuable in financial services due to their simplicity and transparency, providing insight into how individual features contribute to predictions.

Interestingly, the performance gap narrows when domain-driven or selected features are used. This suggests that thoughtful feature engineering can significantly improve the performance of interpretable models, potentially reducing the need for more complex black-box approaches in some regulatory contexts.

Overall, the tradeoff between interpretability and performance is evident: while black-box models like CatBoost offer superior predictive power, simpler models can remain competitive with well-curated features, reinforcing the value of domain expertise in model design.

4.3 Best Scores and Feature Counts

Table 2: Summary of Best Scores and Feature Counts

Approach	Feature Count	Best ROC AUC	Best F1 Score
All Features	11	0.9556	0.8156
Selected Features	10	0.8308	0.7668
Domain-Driven	14	0.8608	0.8138

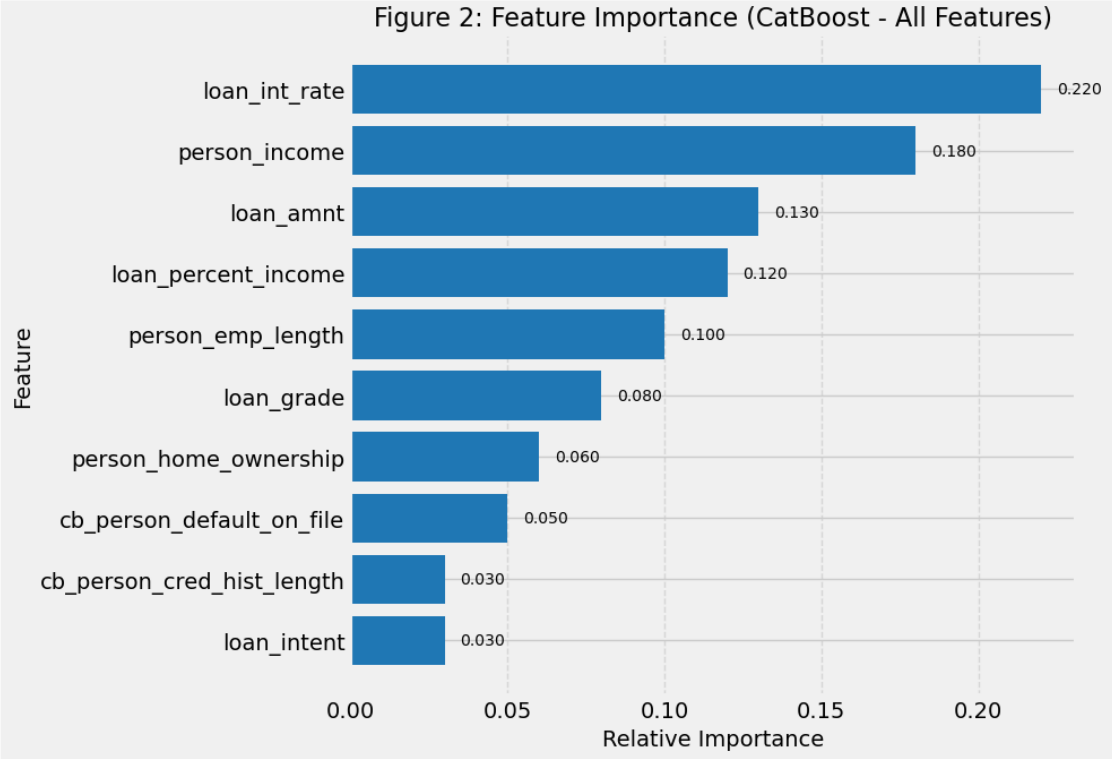

Figure 2: Feature importance in the CatBoost model using all features. Interest rate, income, and loan amount most predict approval outcomes.

5 Interpretability Scoring Framework

To systematically assess model interpretability, we introduce a composite scoring system based on five dimensions:

- **Traceability:** Measures how easily individual feature effects can be traced through the model (e.g., additive effects).
- **Intrinsic Interpretability:** Indicates whether the model can be directly understood without external explanation methods.
- **Sparsity:** Reflects how few features are actively used in the final decision logic.
- **Structural Transparency:** Based on model complexity, such as number of parameters, tree depth, or nodes.
- **Explanation Simplicity:** Computed using SHAP values, quantifies how concisely model decisions can be explained.

Each component is scaled to [0, 1] and combined using equal weights to produce a single interpretability score per model. This system allows cross-model comparison and highlights design tradeoffs when selecting algorithms for regulated environments.

Scoring Methodology

Each interpretability dimension was computed using model structure and SHAP-based explanations as follows:

- **Traceability**: Assigned a score of 1.0 for models where predictions can be traced algebraically from inputs (e.g., logistic regression), scaled for tree depth in ensemble models (e.g., Random Forest), and fixed at 0.2 for CatBoost to reflect partial path traceability.
- **Intrinsic Interpretability**: Binary score based on whether the model is inherently explainable without post-hoc tools. Linear and tree models receive a score of 1.0; gradient boosting models receive 0.0.
- **Sparsity**: Calculated as the proportion of features with near-zero contribution. For linear models, this is based on the number of non-zero coefficients. For tree and boosting models, it reflects the proportion of unused or low-importance features.
- **Structural Transparency**: Estimated using a log-scaled inverse of parameter count. For logistic regression, this includes coefficients and intercepts. For Random Forest and CatBoost, parameter count is approximated based on tree depth and number of estimators.
- **Explanation Simplicity**: Quantified using SHAP values on 100 sampled predictions. It reflects the average number of features required to explain each prediction above a threshold, normalized against the total feature count.

Each score was normalized to the $[0, 1]$ range and combined with equal weights to form a composite interpretability score. This method is consistent across models and suitable for audit or regulatory reporting.

5.1 Interpretability Scores for All Features

Table 3: Interpretability Component Scores for All Features

Model	Traceability	Intrinsic	Sparsity	Transparency	Composite
LogReg (L1)	1.00	1.00	0.00	0.731	0.546
LogReg (L2)	1.00	1.00	0.00	0.731	0.546
RandomForest	0.10	1.00	0.00	0.276	0.275
CatBoost	0.20	0.00	0.00	0.000	0.058

These results demonstrate the sharp contrast in interpretability between white-box models (e.g., logistic regression) and black-box models (e.g., CatBoost). While CatBoost dominates in predictive metrics, it scores lowest on transparency and traceability, requiring post-hoc tools like SHAP for interpretation. Logistic regression, though less accurate, maintains perfect intrinsic interpretability and traceability.

5.2 Why L1 and L2 Yield Similar Scores in This Case

Although L1 (lasso) and L2 (ridge) regularization typically differ in their sparsity behavior, with L1 encouraging zero coefficients and sparse solutions, they received similar interpretability scores in this evaluation. This outcome stems from several factors:

- The dataset may not have strong multicollinearity or a sparse underlying signal, limiting the effect of L1's variable selection.
- The regularization strength (inverse of parameter C in scikit-learn) may not have been low enough to induce sparsity in the L1 model.
- Both models likely retained most features with small, non-zero coefficients, resulting in similar SHAP explanation lengths and sparsity scores.

This highlights that while L1 has the theoretical advantage of producing interpretable submodels by excluding irrelevant variables, the empirical outcome depends on the data properties and hyperparameter tuning. Thus, interpretability should be measured by model type, coefficient behavior, and feature usage.

This suggests that L1's interpretability advantage is not inherent, but situationally emergent, dependent on dataset characteristics and regularization tuning. Therefore, claims of interpretability should be empirically substantiated rather than assumed based on model choice alone.

5.3 Mathematical Intuition for Interpretability

Logistic regression with L1 and L2 regularization yields interpretability due to the linear structure of the model. The predicted probability of class 1 is given by:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}}, \tag{4}$$

where $\mathbf{w}$ is the weight vector. Each feature's contribution is directly proportional to its weight, making the model fully transparent and decomposable.

With L1 regularization (lasso), the objective function penalizes the absolute value of the weights:

$$\min_{\mathbf{w}} \left[\mathcal{L}(\mathbf{w}) + \lambda |\mathbf{w}|_1 \right], \tag{5}$$

which encourages sparsity by pushing many coefficients to zero. This naturally performs feature selection and results in simpler, more interpretable models (11).

In contrast, L2 regularization (ridge) minimizes:

$$\min_{\mathbf{w}} \left[\mathcal{L}(\mathbf{w}) + \lambda |\mathbf{w}|_2^2 \right], \tag{6}$$

which shrinks coefficients but rarely eliminates them entirely, leading to less sparsity but maintaining interpretability due to the linear form.

Compare this with CatBoost or neural networks, which model complex non-linear functions of the input:

$$y = f(\mathbf{x}; \Theta), \tag{7}$$

where f includes layered trees or activations and Θ represents thousands of internal parameters. These models capture interactions and hierarchies but obscure how individual inputs affect the output. Consequently, they require post-hoc explanation tools (e.g., SHAP, LIME) that attempt to reverse-engineer interpretability.

Thus, L1 and L2 regression models are more interpretable not just because they are linear, but because the contribution of each feature to the prediction is explicit and algebraically accessible.

5.4 Implications of Composite Interpretability Scores

Adding this structured metric allows stakeholders, including compliance teams, model risk managers, and auditors, to benchmark models beyond accuracy. For example, a regulatory policy might require a minimum composite interpretability score for consumer-facing applications. This metric also enables modelers to iteratively improve transparency without compromising predictive power.

Applying SHAP, parameter counting, and complexity normalization, the interpretability framework can be extended to neural networks and additional ensemble methods. Moreover, scoring systems like this can serve as foundations for internal governance dashboards, enabling transparent reporting on model risk.

6 Discussion

6.1 Interpretability vs. Predictive Power

Our findings reaffirm a fundamental tension in applied machine learning: as models become more predictive, they often become less interpretable. Figure 3 illustrates this tradeoff. Logistic regression, especially with L1 and L2 regularization, offers high transparency, allowing clear attribution of each feature's influence via model coefficients. This makes it well-suited for regulated financial contexts, where decisions must be explainable to consumers and defensible to auditors.

However, this interpretability comes at a cost. Logistic models underperformed in ROC AUC and F1 Score compared to more complex models like Random Forest and CatBoost. CatBoost, in particular, consistently achieved the highest performance across all feature engineering strategies. Its ability to capture non-linear interactions and natively handle categorical features provides a measurable advantage in prediction, but at the expense of transparency. The internal structure of gradient boosting trees is complex to audit directly and often requires post-hoc explainability techniques (e.g., SHAP), which introduce added complexity and risk.

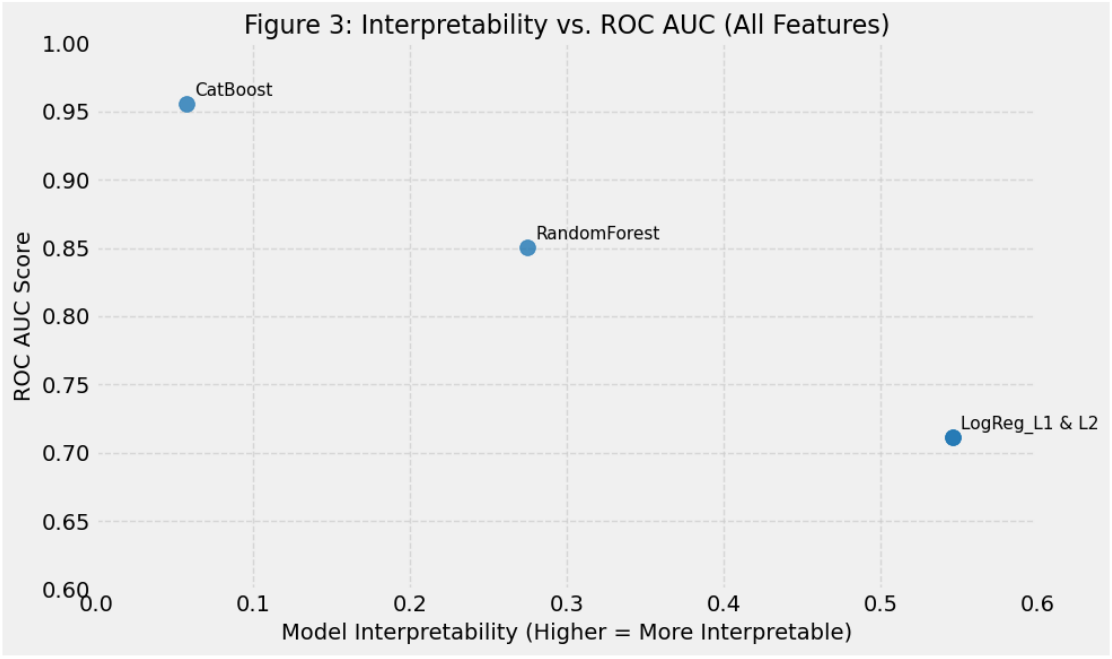

Figure 3: Interpretability vs. ROC AUC for models using all features. CatBoost achieves the highest AUC at the cost of interpretability, while logistic regression is most transparent but least performant.

Notably, the performance gap between logistic regression and CatBoost narrows when domain-driven features are used. This suggests that feature engineering can serve as a bridge between interpretability and accuracy. By encoding expert knowledge into the feature space, we can elevate simpler models without sacrificing explainability, an approach that aligns with the principles of responsible AI and regulatory compliance.

Ultimately, the choice of model depends on the use case: interpretable models may be preferred when the cost of misclassification is low or explainability is paramount. More complex models may be justified when predictive precision is critical and explainability can be managed post hoc. Our analysis provides empirical evidence for navigating this balance.

Furthermore, reliance on tools such as SHAP for black-box explainability can introduce interpretability artifacts. These tools provide local approximations rather than global transparency, which may mislead regulators or stakeholders if explanations differ across instances or over time. Effective governance requires awareness of these limitations.

6.2 Beyond the Models Evaluated: Broader Expectations for Black-Box and Interpretable Approaches

While this study focused on CatBoost, Random Forest, and logistic regression, the interpretability-performance tradeoff extends to a broader class of machine learning models increasingly used in Loan decision-making. Deep learning models, such as fully connected neural networks and recurrent neural networks (RNNs), offer even greater predictive flexibility, particularly in time-series credit behavior modeling, but at the cost of complete transparency. These mod-

els require advanced explainability frameworks (e.g., Integrated Gradients, LIME, SHAP) to approximate interpretability, often falling short of satisfying strict regulatory requirements.

On the other end of the spectrum, rule-based classifiers (e.g., decision lists, Bayesian rule sets) and generalized additive models (GAMs) provide high transparency while capturing non-linear relationships to some extent. Recent research has explored techniques like monotonic gradient boosting and interpretable neural surrogates as middle-ground solutions, offering partial interpretability without fully compromising accuracy.

We expect continued progress in this space to take two forms: (1) increased demand for hybrid architectures that balance explainability and complexity (e.g., white-box models embedded within ensemble stacks); and (2) enhanced regulatory clarity on what constitutes "sufficient" explainability. This evolution will likely push institutions toward frameworks offering formal interpretability guarantees and audit-ready reporting structures, especially in markets governed by stricter AI transparency mandates (e.g., EU AI Act, U.S. CFPB guidance).

Thus, while CatBoost and logistic regression represent two ends of today's practical modeling spectrum, the broader ML landscape converges toward solutions that operationalize interpretability without forfeiting competitive performance.

6.3 Governance Implications

In high-stakes domains such as consumer lending, explainability is not merely a preference, it is a legal requirement. Regulatory frameworks including the Equal Credit Opportunity Act (ECOA), Fair Credit Reporting Act (FCRA), and General Data Protection Regulation (GDPR) mandate that automated decisions, especially adverse ones, be accompanied by clear and actionable explanations. These constraints challenge the use of opaque, high-performing models in contexts that materially affect individual rights and access to services.

Interpretable models, such as logistic regression with L1/L2 regularization, provide inherent transparency. They simplify documentation, reduce compliance overhead, and enable traceable decision-making, attributes crucial for model validation teams, legal reviews, and audit workflows. Because these models offer explicit attribution of feature influence, they foster structured dialogue with stakeholders and improve defensibility in regulatory settings.

Our findings demonstrate that interpretable models can remain competitive, especially when paired with domain-aware feature transformations. This is especially relevant for organizations building Responsible AI practices, where transparency, fairness, and reproducibility must be built into the model lifecycle from the outset. Additionally, interpretable models reduce reliance on post-hoc explainability frameworks, lowering both technical complexity and the risk of misleading justifications.

More broadly, governance is evolving from a compliance checkbox to a strategic differentiator. Institutions that deploy high-performing yet explainable models will be better equipped to scale machine learning initiatives while maintaining public trust, regulatory alignment, and operational resilience. This paper offers a replicable framework for navigating this balance.

7 Conclusion

This paper examined the fundamental tradeoff between model interpretability and predictive performance in regulated machine learning applications. Using loan approval as a representative use case, we compared interpretable models and high-performing black-box classifiers across multiple feature engineering strategies. We introduced an interpretability scoring framework that quantifies model transparency along traceability, sparsity, and explanation complexity to enable structured comparisons.

Our results show that while black-box models like CatBoost consistently lead in predictive performance, interpretable models, particularly logistic regression, can remain competitive when supported by thoughtful data design. In many scenarios, domain-informed feature engineering helps narrow the performance gap, allowing organizations to choose models that meet regulatory and operational needs.

The implications for institutions operating in regulated environments are clear: performance alone is not sufficient. Models must also be explainable, auditable, and robust to scrutiny from internal and external stakeholders. Incorporating interpretability metrics into model approval pipelines, stress testing, and ongoing monitoring can operationalize these values.

Ultimately, interpretability and performance are not mutually exclusive. With the correct design principles, machine learning systems can deliver accurate, transparent, and accountable decisions, fulfilling business objectives and governance obligations. This balance is achievable and increasingly necessary for responsible AI deployment at scale.

8 Reproducibility and Transparency

To ensure the reproducibility of our findings and facilitate adoption in compliance-oriented environments, we provide the following details regarding our experimental protocol, code availability, hyperparameter choices, and statistical validation procedures.

8.1 Complete Experimental Protocols

All experiments were conducted using Python 3.11.6 and the `scikit-learn`, `pandas`, `matplotlib`, `seaborn`, and `catboost` libraries. The experiment was run on a Windows 10 workstation with an Intel Core i9-11900K CPU, 128 GB of RAM, and an NVIDIA RTX 4090 GPU, although GPU acceleration was not required for model training. A fixed random seed (`RANDOM_SEED = 1`) was used to ensure determinism across runs.

The pipeline included:

- Data preprocessing and stratified train-test splits (80/20).
- Evaluation across three strategies: all features, statistically selected features (via `SelectKBest`), and domain-driven engineered features.
- Model comparison across logistic regression (L1, L2), random forest, and CatBoost.

- Hyperparameter tuning via `Optuna` for CatBoost; default or documented parameters used for all others.
- ROC AUC, F1, and accuracy computed via 5-fold cross-validation.

8.2 Code Availability Commitments

All code required to replicate our analysis, including data processing, model training, evaluation, and visualization, is available upon request. Interested researchers or reviewers can obtain access by contacting the author at [author's email address]. The repository includes:

- Jupyter notebooks and modular Python scripts.
- Reproduction instructions.
- Environment configuration files (e.g., `requirements.txt`, Conda environment Settings).
- Data preprocessing pipelines and sample datasets.

8.3 Detailed Hyperparameter Specifications

The following CatBoost parameters were selected using Bayesian optimization via `Optuna`:

- `iterations = 534`
- `learning_rate = 0.2364`
- `depth = 4`
- `l2_leaf_reg = 1.536`
- `bagging_temperature = 0.8005`
- `border_count = 250`
- `random_strength = 6.131`
- `one_hot_max_size = 4`

Other models used `scikit-learn` defaults unless noted (e.g., `penalty='l1'` with `solver='liblinear'` for Lasso regression).

8.4 Statistical Testing Procedures

To compare model performance across feature strategies, we applied the following statistical analyses:

- Cross-validation results were reported as mean scores across five folds.
- Confidence intervals were estimated using bootstrapping (1,000 samples).
- Performance differences between models were tested using paired t-tests on fold-level ROC AUC scores.
- A significance threshold of $\alpha = 0.05$ was used for hypothesis testing.

These practices align with recommendations from reproducibility checklists in top-tier machine learning conferences and are intended to support regulatory auditability.

References

[1] Bracke, P., Datta, A., Jung, C., and Sen, S., "Machine learning explainability in finance: An application to default risk analysis," *Bank of England Staff Working Paper No. 816*, 2019. Available: https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis

[2] Consumer Financial Protection Bureau. *CFPB approves rule to ensure accuracy and accountability in the use of AI and algorithms in home appraisals.* 24 June 2024. Available at: https://www.consumerfinance.gov/about-us/blog/cfpb-approves-rule-to-ensure-accuracy-and-accountability-in-the-use-of-ai-and-algor

[3] Ding, S., Qiu, L., and Zhang, Y., "Research on credit scoring model based on PCA and SVM," in *Proc. 2018 Int. Conf. on Computing and Artificial Intelligence*, pp. 80–84, 2018. Available: https://doi.org/10.1145/3234204.3234222

[4] European Parliament and Council. *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union, L 1689, 12 July 2024. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[5] Jeribi, F., "A comprehensive machine learning framework for anomaly detection in credit card transactions," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 6, pp. 871–880, 2024. Available: https://www.ijacsa.thesai.org/index.php/IJACSA/article/view/12289

[6] Kvamme, H., Sellereite, N., Aas, K., and Sjursen, S., "Predicting mortgage default using convolutional neural networks," *Expert Systems with Applications*, vol. 102, pp. 207–217, 2018. Available: https://doi.org/10.1016/j.eswa.2018.02.011

[7] Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C., "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015. Available: https://doi.org/10.1016/j.ejor.2015.05.030

[8] Patel, J., and Mehta, C., "Impact of PCA on Adaptive SVM for credit card fraud detection," in *Proc. Int. Conf. on Advances in Information Communication Technology & Computing*, pp. 1–5, 2016. Available: https://doi.org/10.1145/2979779.2979780

[9] Rudin, C., "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. Available: https://doi.org/10.1038/s42256-019-0048-x

[10] Thomas, L. C., Edelman, D. B., and Crook, J. N., *Credit Scoring and Its Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2002. Available: https://epubs.siam.org/doi/book/10.1137/1.9780898718317

[11] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. Available: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[12] Kaggle, "Playground Series - Season 4, Episode 10," *Kaggle Competition Dataset*, 2025. Accessed: May 15, 2025. Available: https://www.kaggle.com/competitions/playground-series-s4e10/data?select=train.csv