ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

Interpretable AI Approaches for Banking Risk Models

Daniel Thompson
Asst Professor

Abstract

Credit risk scoring systems are among the most crucial decision-making models developed by banks. These models have become mandatory in many jurisdictions because of regulatory requirements that promote risk-sensitive capital charge computations. However, legal requirements and the need for better customer relations now make credit-scoring evaluation systems necessary and increase demand in the market. Regulators require understandability of prediction models, banks are interested in interpretability to create customer relations and meet consumer-protection laws, and customers want to understand why they were rejected, especially in cases of marginal evaluation. Additionally, explainable artificial intelligence (xAI) models support the model-monitoring processes of banks and help the implementation of explainable credit assessment. The complex nature of the model-jungle makes it impossible for users and auditors to be aware of limitations, risks, and adequacy in risk management therefore clarity on how the models and systems become really interpretable is required.

Existing xAI credit scoring evaluation models, explainable AI (xAI) model families, and classes of post-hoc explanation techniques are examined to present a taxonomy of approaches. Based on the analysis, a set of operational and cognitive evaluation measures and compliance-oriented explainability tests are proposed. Finally, incorporation of these concepts into model deployment, xAI governance, and overall software life-cycle management in banks is outlined.

Keywords: Credit Risk Scoring Systems, Explainable Artificial Intelligence, Explainable Credit Assessment, Regulatory Compliance In Banking, Model Interpretability, Customer-Facing Credit Decisions, Consumer Protection Laws, Risk-Sensitive Capital Requirements, XAI Model Taxonomy, Post-Hoc Explanation Techniques, Credit Scoring Evaluation Frameworks, Model Transparency, Operational Explainability Metrics, Cognitive Evaluation Measures, Compliance-Oriented Explainability Tests, Model Monitoring And Validation, XAI Governance Frameworks, Banking Software Lifecycle Management, Risk Management Transparency, Interpretable Financial Models.

1. Introduction

The evaluation of credit risk, an essential task in banking, has gained considerable attention in the last decade because of an increasing demand for credit and the availability of a wealth of banking data. Credit scoring is traditionally performed using well-tested logistic regression models. In recent years, however, machine learning techniques, which often outperform logistic regression, have been applied to the problem. However, most of these models offer little or no explanation of their predictions.
Many banks follow a “no black-box” approach, which impedes the use of advanced machine learning techniques for credit scoring. Explainable artificial intelligence (XAI) models are being developed to address this gap. A survey of XAI methods applicable to credit scoring has therefore been conducted. XAI-related features and data augmentation techniques facilitating the implementation of explainable machine learning models in banking, as well as explainability evaluation approaches with focus on cognitive and operational criteria, and model compliance, are discussed. The ultimate goal of this research is to obtain fully explainable models suitable for credit scoring in banking using either purely interpretable algorithms or combinations of well-tested classifiers with post hoc

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

explanation techniques such as local surrogate models, feature contributions, or feature perturbation.

1.1. Overview of the Study

The growing complexity of AI models for credit scoring, combined with issues of transparency, fairness, and compliance, limits their adoption by banks. Therefore, detecting credit risk with simple and interpretable models, integrating explainability in non-interpretable architectures, and collecting justification data for post-hoc explanations are important. Techniques must enable banks to explain and validate their AI models in a cognitive, operational, and compliance-oriented manner. Addressing these issues results in a four-layered framework that guides the development of explainable credit-scoring models. It assists the construction, evaluation, and documentation of reliable AI architectures throughout their lifecycle while providing an operational interface for risk management and compliance functions. The contribution includes a survey of explainable credit-scoring techniques covering simple models, hybrid approaches, and feasible post-hoc explanations.

More sophisticated credit-scoring models can still be investigated by banking institutions if transparency and fairness aspects are carefully handled as part of the model development and validation processes. Governance structures allow internal audit, data protection, and risk functions to evaluate credit-scoring models in the context of external regulators and stakeholders. A separation of concerns by the operationalization of explainability validations empowers banks to consume advanced credit-risk assessment techniques in a compliant manner. Consequently, the proposed graduation contributes to the increasingly important narrative on explainable and trustworthy AI in banking by addressing the necessary mechanisms for achieving regulatory compliance and customer trust in a sound credit-scoring business.

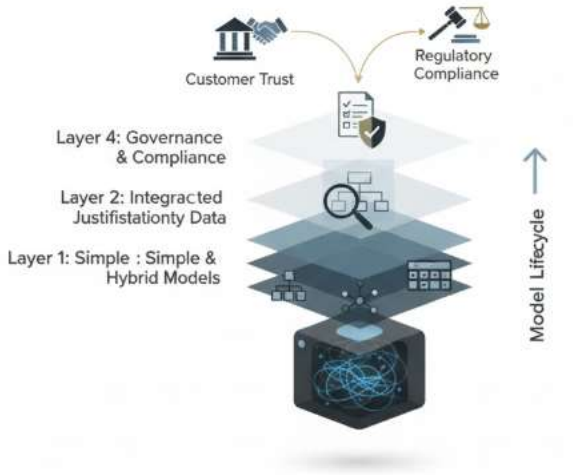


Fig 1: The Trustworthy Credit Architecture: A Four-Layered Framework for Integrating Explainable AI, Governance, and Regulatory Compliance in Banking

2. Background and Rationale

Effective and sustainable credit risk management is key to the soundness and stability of banks. Adverse selection and moral hazard problems associated with information asymmetry are present both at the credit-lending decision stage and during the monitoring phase. Credit risk models are widely used by banks to facilitate the quantification of credit risk and improve regulatory compliance. However, several considerations hinder the adoption of machine learning approaches. To assess the applicability and fitness of explainable AI solutions for credit scoring, the credit risk process and the specific needs for explainability are analyzed from a banking business perspective. Requirements, limitations, and perspectives emerge from the discussion.

Credit risk models in banks undergo numerous evaluations and validations during the model lifecycle. Various implementing functions require a wide range of performances beyond predictive accuracy. Addressing problems such as the data privacy concern, such models have, in most cases, been developed over the years using classical statistical techniques, being interpretable by

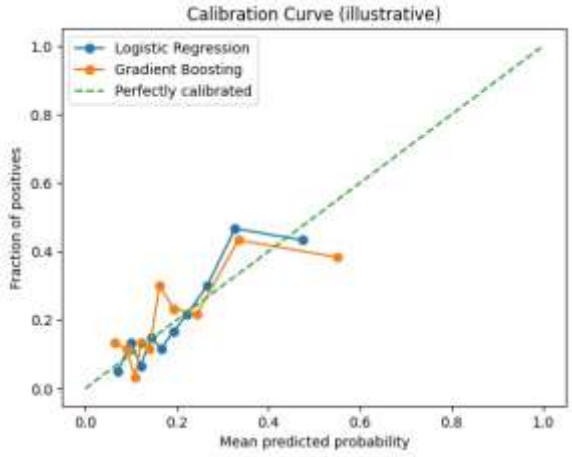
ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

human decision-makers. The more recent application of machine learning techniques, which have become mainstream in many areas, is still restrained by a myriad of requirements that explainable AI attempts to fulfill.



Equation 1: Core credit-risk quantities used in scoring

1.1 Probability of Default (PD)

Let $Y \in \{0,1\}$ be the default indicator over a horizon (e.g., 12 months).

- $Y = 1$: default, $Y = 0$: no default
- The **score model** outputs:

$$\mathrm{PD}(x) = \Pr(Y = 1 \mid X = x)$$

1.2 Expected Loss (EL)

With exposure at default EAD and loss given default $LGD \in [0,1]$:

$$\boxed{\mathrm{EL}(x) = \mathrm{PD}(x) \cdot LGD(x) \cdot EAD(x)}$$

2.1. Credit Risk Management in Banking

Credit risk management is a key concern for banks and plays a fundamental role in their profitability, growth, and operating model. Credit risk is defined as the risk that a borrower may default on its payments. A default occurs when a debtor fails to fulfill any of its contractual obligations, such as the payment of interest, the repayment of capital, or the violation of any covenant of the loan agreement. In practice, early detection of potential defaults and clear communication are crucial factors that can help avoid a formal default. However, it is inevitable that a portion of loans will inevitably go into default; hence the process of estimating potential losses on credit exposure. Credit scores have been used by banks for many years as one of the most successful tools to assess credit risk. Credit-scoring models generate quantitative estimates that guide credit decision-making and help a bank to detect high-risk customers. Despite decades of research into credit-scoring models, the banking industry is still looking for improved tools.

Previously separated by the large margin of success compared with other fields of predictive analytics, constrained by banking regulations, and protected by the competitive advantage they created, credit-scoring models, particularly logistic regression, are no longer immune to such trends. Fintech companies are aggregating and analyzing vast amounts of data, allowing customers to receive credit decisions almost instantly. This increase in automation and decrease in turnaround time put additional pressure on banks to reduce risk while still supporting deal flow to meet business goals.

3. Methods for Explainable Credit Scoring

Several families of descriptive credit scoring models have been deployed in banking that either provide inherent interpretability or use hybrid architectures to combine the descriptive potential of interpretable models with the predictive potential of opaque black-box models that trade-off quality for clarity. Based on models like decision trees, generalized additive models and attribution methods, the applied views extend beyond model explanatory power to also include the user interface for operationalization of the model and the underlying data combinations, while seeking randomness with respect to a data leak of the training data, undesired bias and non-responsibility issues. All models are proposed in a proper economic evaluation perspective.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

With an increasingly evident growing demand for the deployment of AI in a transparent and regulated way, achieving explainability in banking-based scoring systems stands as a critical aspect under the shaping of in-light-of the European Commission's proposal for an AI Act. In this perspective, contrasted evaluation functions offer non-invasive support for the architecture choice and model priority estimation conducted by the financial institution. The external evaluation of the predictiveness of the models with respect to a validation sample can be supported from a cognitive point of view by also considering the stochastic nature of the models and any violation of the faithfulness property related to a possible information leak of the training set.

3.1. Interpretable Models and Hybrid Approaches

Interpretable machine learning models and hybrid methods combining simpler-and more complex rules form an adequate answer to credit risk evaluation within the banking domain. Classic models, such as decision trees, logistic regression and naive Bayes, possess fundamental characteristics of explainability, like comprehensibility, performance and low complexity. Nevertheless, they underperform compared to more advanced models (e.g., gradient boosting and deep learning approaches). Offering competitive performance while maintaining solid explanations, hybrid models integrate comprehensibility into their structure.

Beyond their design, though, ML models can operate in a black-box manner with post-hoc explanation methods. Such techniques do not alter the original model but rather present results in a simplified way, enabling an understanding of the decision. However, they remain extrinsic and, as a consequence, question the faithful representation of the model's logic. Well-known examples of post-hoc explanations are LIME and SHAP. They measure the influence of each attribute on the model prediction by perturbed instances or by decomposing the output into the contributions from each feature.

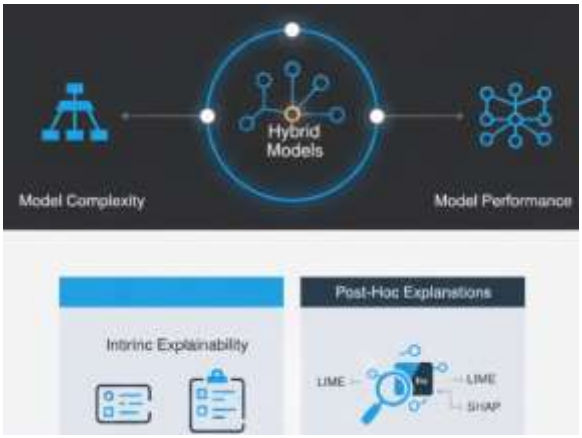


Fig 2: Bridging the Interpretability Gap: A Comparative Analysis of Classic, Hybrid, and Post-Hoc Explainability Frameworks in Credit Risk Evaluation

3.2. Post-hoc Explanation Techniques

Three common techniques for post-hoc explanation of credit scoring decision-making include LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and a feasible variant specifically designed for a deep neural model. LIME explains any classifier by perturbing the dataset and generating an interpretable local surrogate model around the prediction of interest. For a specific sampling of the test instance, new perturbed data is produced by randomly sampling its neighborhood. The prediction of the black box model is obtained for these perturbed data and serves as the response variable for the local surrogate. The advantage of this technique is that it not only works as a general post-hoc procedure but also enables a whole instance-level understanding.

SHAP is a unifying framework connecting multiple explanation methods and generating an optimal explanation by modeling the prediction of interest as a cooperative game. Each feature corresponds to a player aiming to explain a specific prediction. The result shows the contribution of each feature in the sample instance and successfully captures both local and global interpretability in terms of additive feature importance. The decomposition of the prediction into the sum of feature contributions provides an intuitive understanding, while the distribution

ISSN: 3067-4166

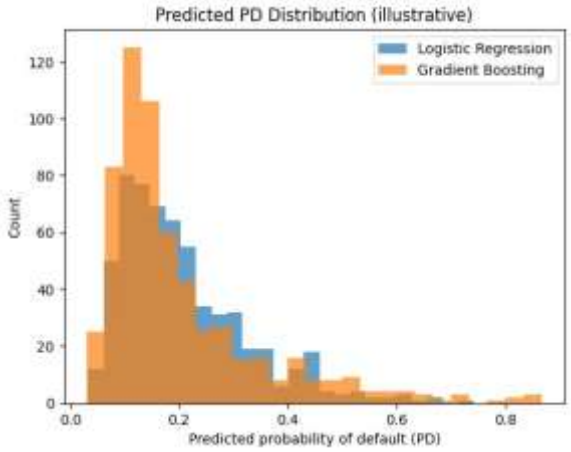
AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

of SHAP values across the training dataset indicates the overall importance of the feature.

Explaining a complex deep learning architecture is challenging but feasible with the appropriate techniques. The idea is to treat individual predictions of the model as a black box and to explain them after training. The method applies SHAP on the fully trained DNN credit scoring model to explain an individual prediction, generating a set of positive and negative contributing features. To safeguard customers' private information, the method swaps the values of sensitive credit applicants in the dataset with those of non-sensitive individuals, updates the module of SHAP, and makes a prediction privacy-preserved for CHB data.



4. Data, Privacy, and Fairness Considerations

Several key aspects of data and visualization merit special consideration. Data quality is paramount; poor-quality data can confound model interpretability and explanation efforts, as well as lead to flawed predictions. In general, for a supervised model to be useful, the training dataset must be representative of the kind of examples the model will later encounter. Model predictions can also be influenced by the correlation of certain input variables with the target label throughout the data, even when the models are interpretable. Improving the representativeness of a sample's normative context can also help make white-box experiments more trustworthy.

Data privacy of sensitive banking records is another aspect that should receive special attention. Detecting possible data leaks in models used for explanation purposes might require separate data, which can be difficult to obtain in restricted domains. The models themselves should also be thoroughly assessed to identify any possible flaws capable of compromising short- to medium-term predictions.

Another concern is fairness. When examined through the lens of regulating discrimination in machine learning models, fairness is highly domain-specific. While some banking experts advocate for a non-discriminatory approach, others assert that accurate models should be trained regardless of fairness. Consequently, an explanation-based approach that monitors model behavior is preferable, as it can identify biases without constraining model construction.

Equation 2: Logistic Regression

2.1 Start from odds → log-odds

Let $p(x) = \Pr(Y = 1 \mid x)$.

Odds:

$$\text{odds}(x) = \frac{p(x)}{1 - p(x)}$$

Log-odds (logit):

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \eta(x)$$

Logistic regression assumes $\eta(x)$ is linear:

$$\eta(x) = \beta_0 + \beta^\top x$$

2.2 Solve for $p(x)$ step by step

From

$$\log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta^\top x$$

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

Exponentiate both sides:

$$\frac{p}{1-p} = e^{\beta_0+\beta^\top x}$$

Multiply both sides by $(1-p)$:

$$p = e^{\beta_0+\beta^\top x}(1-p)$$

Expand RHS:

$$p = e^{\beta_0+\beta^\top x} - pe^{\beta_0+\beta^\top x}$$

Bring p terms together:

$$p + pe^{\beta_0+\beta^\top x} = e^{\beta_0+\beta^\top x}$$

Factor p:

$$p\left(1 + e^{\beta_0+\beta^\top x}\right) = e^{\beta_0+\beta^\top x}$$

Divide:

$$\boxed{p(x) = \frac{e^{\beta_0+\beta^\top x}}{1+e^{\beta_0+\beta^\top x}} = \frac{1}{1+e^{-(\beta_0+\beta^\top x)}}}$$

2.3 Likelihood, loss, and gradient (training derivation)

Given data $\{(x_i, y_i)\}_{i=1}^n$, with $p_i = p(x_i)$.

Bernoulli likelihood:

$$\Pr(y_i \mid x_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n}[y_i \log p_i + (1-y_i)\log(1-p_i)]$$

Negative log-likelihood (cross-entropy loss):

$$\boxed{J(\beta) = -\ell(\beta) = \sum_{i=1}^{n}[-y_i \log p_i - (1-y_i)\log(1-p_i)]}$$

Gradient (key step):

- Since $p_i = \sigma(\beta_0 + \beta^\top x_i)$ and $\frac{d\sigma}{dz} = \sigma(z)\big(1-\sigma(z)\big)$,
  you get:

$$\boxed{\nabla_\beta J(\beta) = \sum_{i=1}^{n}(p_i - y_i)\, x_i}$$

This is what optimizers (GD/L-BFGS/etc.) use.

4.1. Data Quality and Representativeness

High-quality datasets that accurately reflect the target domain are essential for training effective models. Various factors — such as data provenance, sampling strategies, information completeness, and relevance — also exert strong influence. For instance, representative samples comprising diverse customer segments enhance the generalizability of trained credit scoring models. Likewise, privacy and fairness-based data management strategies for recurrent patterns associated with "discriminatory" sensitive attributes (e.g., origin, age, gender) ensure the avoidance of biased empirical discrimination in model predictions.

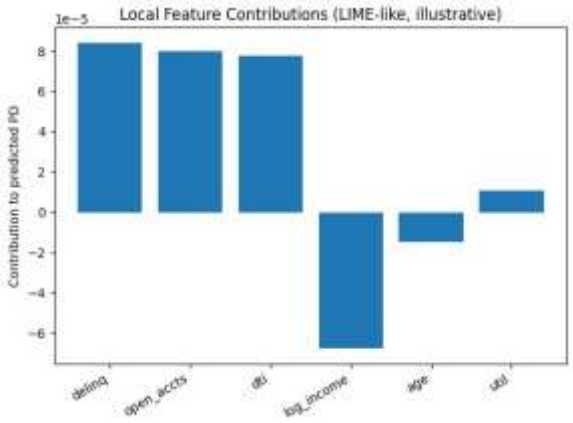
Interpretable, augmented, or hybrid credit-scoring solutions demand available data or knowledge (e.g., similar past cases, associated risk categories, expert guidance) for implementing explanation methods. Prior research indicates a need both for reliable detection of classification-relevant instances and for supplementary input on classification decisions. Such solutions also risk producing girls-bias-fairness in-class conditional discrimination of prediction quality and label distribution without integrated mitigations during the data-collection stage. Lawyers stipulating understandable or "explainable" algorithmic decisions require such provisions.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18



5. Evaluation Frameworks for Explainability

Two classes of evaluation metrics help assess the extent to which the explainable AI model fulfills its objectives. The first group incorporates cognitive and operational metrics to evaluate the attribute-intensity-shaped explanatory artifacts directly and indirectly affecting human cognition and decision-making processes. The second group seeks to address existing regulatory requirements defining level of explanation dignity and explanations' relevance to their recipients.

The need for an operationalized Conceptual Framework for Explainable AI supporting explainability aspects, depth, and fidelity of the proposed explaining models and functionally placed as an embedding or stand-alone module connected with explainable model runs is important neither only from a conceptual nor from a compliance-oriented perspective. It aims to provide a guidelines, workflow, and testing-assurance map to the banking industry by articulating the links among the cognitive-understanding, operation, and compliance paths and requirements. Important regulatory documents such as the European Commission's White Paper on AI call for mandatory assessment of advanced models in high-risk sectors such as finance – a demand backed by the European Banking Authority's Data Guidelines, which recommend duly considered proprietary models, compliant documentation, legal and risk-operating interfaces guaranteeing state-of-the-art level fulfilment.

Ethically responsible banking demands compliance-oriented scrutiny of even interpretable models before their deployment. Despite their pattern-recognition core, these models may still require secondary explanation even after a documentation-illustrated predictive run supporting the answer to the question "How have we reached this conclusion?"

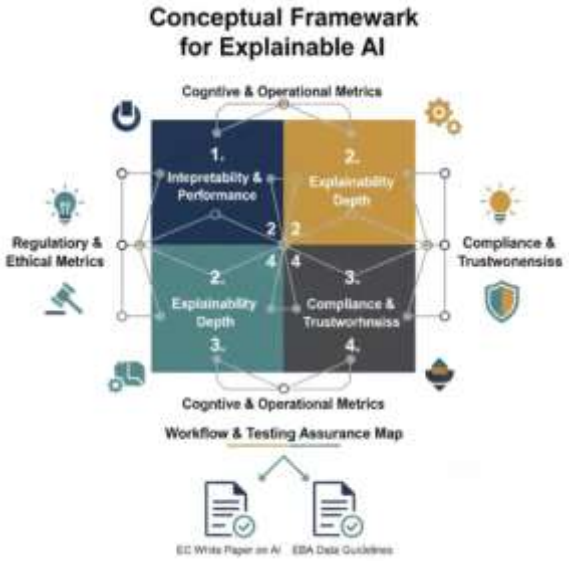


Fig 3: The Operationalized XAI Framework: Multi-Dimensional Evaluation Metrics and Regulatory Compliance Pathways for High-Risk Financial Models

5.1. Cognitive and Operational Metrics

A well-known challenge of machine learning models is that they are often not interpretable by default. This hinders their utilization in high-stake domains such as banking, law, or medicine, where explanations from the predictive algorithm are an essential requirement for the stakeholders. Therefore, dedicated frameworks, methods, and models allowing domains experts to understand how and why a specific decision was made or revealing how the model behaves in general became widely developed and investigated in recent years. While these explanation techniques can make any given model interpretable, the

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

goal of generating pertinent explanations should ideally be built into the modelling process, beginning from the moment when experts re-think the predictive task to the selection and the implementation of the machine learning model. A useful strategy is thus the construction of inherently interpretable models while using hybrid methods that combine a common black-box approach to learn a predictive function with a comprehensible post-hoc interpretative layer.

Alternatively, post-hoc explanations of black-box models can be achieved by explanation methods specifically designed with a supporting role in the interpretative process. Explanations are then seen as an additional output layer that assists the human experts in reviewing the already predicted responses of the black-box model, increasing the trustfulness of those decisions without requiring characterizing knowledge about the black-box model or the prediction function itself. Nevertheless, many banks around the world decide to rely on interpretable models not simply because of their transparency, but also for enabling a more systematic study of the business problem under scope. Indeed, whenever developing a new model, banks usually follow a defined modelling lifecycle that collects and thoroughly analyzes historical data, building one or more predictive functions that become part of a structured documentation model.

5.2. Compliance-Oriented Validation

Compliance-oriented validation of model explainability and trustworthiness emphasizes a risk-averse perspective. Here, the probability of classification failure serves as a quality criterion, measuring the likelihood that the model fails to assess new customers correctly. The acceptance threshold is set at levels acceptable by the bank’s credit decision-makers. A post-hoc explanation can then be reported as trustworthy when the final classification agrees with the provided explanation and classification failure lies below the acceptance threshold.

This evaluation framework enables the bank to operate the model as a quasi-black-box when credit institutions are not specifically requesting explanation. In this case, the bank relies on the predictive performance assessed by the risk management function and ensures regulatory compliance with respect to fairness. In the event that a credit institution seeks explanation, the model can be used as a white-box and explanations can be provided with guaranteed trustworthiness.

6. Model Deployment and Governance in Banks

Deployment of AI credit scoring systems in banking requires consideration of several primary aspects to ensure proper usage and compliance with regulations, notably eXplainable Artificial Intelligence (XAI) guidelines. The specific factors in model deployment and governance generally cover these items: the model lifecycle and documentation needed for deployment, and the interplay with risk management and compliance functions.

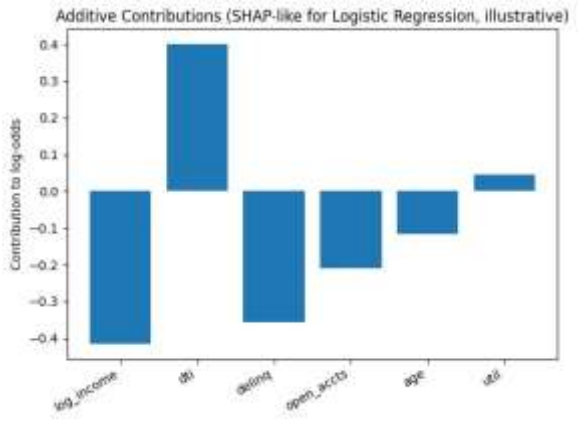
The life cycle of analytics models servicing banks and their supporting processes involves many similar aspects to that of traditional risk models—development, validation, deployment, monitoring, and updating—but is more elaborate. A documentation package in line with the bank's internal policies must accompany the model at the deployment stage, containing details suitable for boards and top management, model stage summaries (development, validation, deployment, etc.), a user manual, and a record of the data used for testing. The explanation component of the model should also be formally documented, integrating the analysis of the choice of algorithms and parameters and the description of surrogate models, if developed. Explicit links to the bank's risk management and compliance functions also need verification, with a special emphasis on fair lending regulations.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18



6.1. Model Lifecycle and Documentation

At banks, credit scoring models are evaluated by different stakeholders and used for different purposes. The fact that model governance is a regulatory requirement implies that comprehensive documentation of the model development process is a prerequisite for compliance-oriented validation. Such a model accountability system ensures that a fair and sound model is selected for deployment in the bank's technology infrastructure. Banks must keep credit scoring models updated throughout their life cycle, controlling performance in relation to recent defaults and customers that have become creditworthy through default curing or market rebound. Model owners must ensure that risk factors remain significant, data are not outdated and dubious relationships with other variables are avoided. Credit-scoring model versioning tracks these changes. At a more operational level, some banks have begun to implement prediction monitoring systems that notify model owners if the model score output deviates from what is expected given the recent behaviour of the application file. Regulatory validation of a credit scoring model occurs before deployment and must be repeated by an independent team using a different software tool than that used for modelling. This requirement helps prevent model owners from "cheating" during validation. An interpretable model that also follows a transparent development process greatly facilitates the validation work. Banks take advantage of the information gleaned from operational monitoring and develop their own model validation processes. Unlike the regulatory exercise, which focuses on operational risk mitigation, local validation aims to check the correctness and adequacy of scores for risk-based pricing. Even though credit-scoring predictions are assumed to be probabilities of default, they are rarely used in such a way.

Equation 3: Decision Trees

Decision trees are listed among classic interpretable models.

3.1 Gini impurity (binary classification)

At a node, let class proportions be p (defaults) and $1-p$ (non-defaults).

$$\boxed{Gini = 1 - p^2 - (1-p)^2 = 2p(1-p)}$$

3.2 Information gain (split selection)

Let parent node have impurity $I(\text{parent})$. Split into left/right with weights w_L, w_R.

$$\boxed{\Delta I = I(\text{parent}) - \left(w_L I(L) + w_R I(R)\right)}$$

The best split maximizes ΔI.

6.2. Risk Management and Compliance Interfaces

Ensuring proper governance involves integrating model development and deployment with risk management and compliance processes that maintain controls throughout the model lifecycle. A framework specifically designed for credit risk models can delineate the roles of the three functions, highlight their dependencies, and ensure that all the requirements are appropriately addressed. Such a framework is particularly relevant for the approval of new models, the qualification of models that are already in use, and the periodic assessment of existing models.

The objectives associated with the provision of credit risk scoring models for risk management, credit, and operational risk, as well as for compliance, cover the range of potential use cases and any required checks. Also important are the risk and compliance interfaces that ensure proper governance, usually reflected in individual model

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

documents. The roles of the three main functions supporting these activities typically include risk management validating that the model is appropriate, credit providing input to the development process, operational risk ensuring that sufficient documentation is provided, compliance confirming that regulatory requirements are met, and senior management approving model use.

7. Conclusion

The potential factors that determine credit risk in the context of traditional credit scoring models have been extensively investigated. However, risk evaluation remains primarily a black-box process. Model risk manifests in unintended bias, miscalibration, overfitting to historical data, invariances not present in a bank's broader economic exposure, and other aspects. The introduction of more sophisticated AI models has not advanced model transparency and risk mitigation in credit scoring. Instead, the analysis of opacity has shifted toward establishing a framework of explainability and fair AI. This framework is rich in sophistication and detail but provides little practical direction for explainable risk evaluation.

While it is accompanied by a useful taxonomy of potential explanations and alignments, these perspectives have not yet translated into practical implementation guidance for banks. The translation of explainability from a mere metric to a governance mechanism requires a detailed understanding of the banking model lifecycle, including design, documentation, model risk interfaces with the operational control and compliance functions, and customer data privacy arrangements with the credit risk systems. Addressing explainability in design coordinates the solution's architecture with anik board and regulatory expectations. Integrating cognitive and operational metrics into validation ensures that business managers and supervisory authorities share the same platform. It also allows feedback from sensitive decision-making areas such as capital adequacy and liquidity to guide the selection of complex solution.

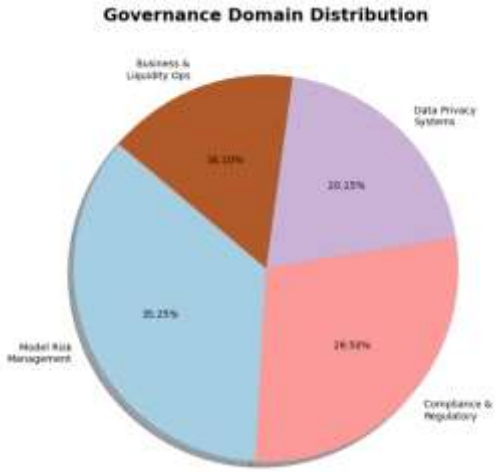


Fig 4: Governance Domain Distribution

7.1. Final Insights and Future Directions

Ensuring adequate credit risk assessment is crucial for the stability and profitability of banks. Machine and deep learning models, owing to their superior predictive performance in numerous risk-related tasks, are being increasingly adopted. However, their inherent unexplainability poses challenges to their acceptance in the banking domain that requires transparency and cognitive compliance. Proposed explainable models have addressed the explanation needs of both data-scientist- and bank-customer related users; yet, they are primarily post-hoc approaches either based on black-box predictors or on the local explanation paradigm.

A comprehensive overview of the existing methods for explainable credit scoring in banking was conducted, taking into account the aspects of model interpretability, the characteristics of the datasets and the requirements imposed by regulators. The survey revealed several unexplored areas that should be tackled in future research. Most notably, while cognitive metrics for evaluating the quality of explanations received considerable attention, operational metrics that ascertain the quality of the explanation process have so far seen only limited contribution. Also, while compliance-oriented model evaluation became a critical task, only few works have conducted risk-oriented validation of biometric models.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

8. References

1. Ariza-Garzon, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE Access, 8, 64873–64890.
2. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Fintech risk management. Frontiers in Artificial Intelligence, 3, 26.
3. Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable AI for interpretable credit scoring. In Proceedings of the 10th International Conference on Artificial Intelligence and Applications, 105–114.
4. Meda, R. (2025). AI-Driven Demand and Supply Forecasting Models for Enhanced Sales Performance Management: A Case Study of a Four-Zone Structure in the United States. Metallurgical and Materials Engineering, 1480-1500.
5. Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. Royal Society Open Science, 7(6), 191649.
6. Wang, W., Lesner, C., Ran, A., Rukonic, M., Xue, J., & Shiu, E. (2020). Using small business banking data for explainable credit risk scoring. Proceedings of the AAAI Conference on Artificial Intelligence, 34(8), 13396–13401.
7. Dastile, X., & Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. IEEE Access, 9, 50426–50440.
8. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, 752558.
9. Lebcir, I., Mageswari, S. D., Bhosale, Y. H., Nagubandi, A. R., & Mahabooba, M. (2025). Agile Strategic Management in the Age of Disruption: Leveraging AI and Data Analytics for Competitive Advantage. Advances in Consumer Research, 2(6), 2581.
10. Hadji Misheva, B., Jaggi, D., Posth, J.-A., Gramespacher, T., & Osterrieder, J. (2021). Audience-dependent explanations for AI-based risk management tools: A survey. Frontiers in Artificial Intelligence, 4, 794996.
11. Ben David, D., Resheff, Y. S., & Tron, T. (2021). Explainable AI and adoption of financial algorithmic advisors: An experimental study. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 390–400.
12. Ariza-Garzon, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. (2021). Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), 70–90.
13. Dastile, X., Celik, T., & Vandierendonck, H. (2022). Model-agnostic counterfactual explanations in credit scoring. IEEE Access, 10, 69543–69554.
14. Kalisetty, S., & Inala, R. (2025). Designing Scalable Data Product Architectures With Agentic AI And ML: A Cross-Industry Study Of Cloud-Enabled Intelligence In Supply Chain, Insurance, Retail, Manufacturing, And Financial Services. Metallurgical and Materials Engineering, 86-98.
15. Bueff, A. C., Cytrynski, M., Calabrese, R., Jones, M., Roberts, J., Moore, J., & Brown, I. (2022). Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. Expert Systems with Applications, 202, 117271.
16. Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2022). A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. Decision Support Systems, 152, 113647.
17. Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. Frontiers in Artificial Intelligence, 5, 779799.
18. Paleti, S., Baliyan, M., Aitha, A. R., Reddy, B. A., Bhadauria, G. S., & Sing, S. A. (2025, August). Graph—LSTM Hybrid Model for Improving Fraud Detection Accuracy in E-Commerce Financial Services. In 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-6). IEEE.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 4 ISSUE: 01

RECEIVED: JANUARY 22 REVISED: FEBRUARY 19 ACCEPTED: FEBRUARY 27 PUBLISHED: MARCH 18

19. Li, W., Paraschiv, F., & Sermpinis, G. (2022). A data-driven explainable case-based reasoning approach for financial risk detection. Quantitative Finance, 22(12), 2257–2274.
20. Babaei, G., Giudici, P., & Raffinetti, E. (2022). Explainable artificial intelligence for crypto asset allocation. Finance Research Letters, 47, 102941.
21. Chen, D., Ye, J., & Ye, W. (2023). Interpretable selective learning in credit risk. Research in International Business and Finance, 65, 101940.
22. Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. Data Science and Management, 6(3), 123–133.
23. Meda, R. (2025). Dynamic Territory Management and Account Segmentation using Machine Learning: Strategies for Maximizing Sales Efficiency in a US Zonal Network. EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR, 46(1), 634-653.
24. Nwafor, C. N., & Nwafor, O. Z. (2023). Determinants of non-performing loans: An explainable ensemble and deep neural network approach. Finance Research Letters, 56, 104084.
25. Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Research Letters, 58, 104309.
26. Weber, P., Carl, K. V., & Nissen, V. (2024). Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. Management Review Quarterly, 74, 867–907.
27. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. Artificial Intelligence Review, 57, 216.
28. Liu, Y., Huang, F., Ma, L., Zeng, Q., & Shi, J. (2024). Credit scoring prediction leveraging interpretable ensemble learning. Journal of Forecasting, 43(2), 286–308.
29. Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(1), 357–372.
30. Chang, V., Xu, Q. A., Akinloye, S. H., Benson, V., & Hall, K. (2025). Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study. Annals of Operations Research, 354, 247–271.
31. Kumar, K. M., Parasar, A., Walia, A., Inala, R., & Thulasimani, T. (2025, August). Enhancing Risk Management Strategies in Financial Institutions Using CNN and Support Vector Regression. In 2025 5th Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-6). IEEE.
32. Nallakaruppan, M. K., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An explainable AI framework for credit evaluation and analysis. Applied Soft Computing, 153, 111307.
33. Li, H., & Wu, W. (2024). Loan default predictability with explainable machine learning. Finance Research Letters, 60, 104867.
34. Liu, J., & Zhang, X. (2024). Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation. Journal of Forecasting, 43(5), 1625–1660.
35. Hlongwane, R., Ramabao, K. K. K. M., & Mongwe, W. (2024). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. PLOS ONE, 19(8), e0308718.
36. Dastile, X., & Celik, T. (2024). Counterfactual explanations with multiple properties in credit scoring. IEEE Access, 12, 110713–110728.