International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | ICAIDSCT26-Conference Proceeding
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.ICAIDSCT26-136

Applying Interpretable AI Techniques to Capital and Regulatory Reporting for Supervisory Transparency

Laxmi Naga Durga Pandrapragada
Independent Researcher, California, United States.

Abstract: Capital and regulatory reporting in large banking institutions operates under supervisory expectations that demand reporting outcomes be accurate, transparent, traceable, and defensible under audit and examination. Although artificial intelligence (AI) can improve efficiency and analytic depth, black-box models create supervisory risk when they cannot be clearly explained, reproduced, and governed. This paper proposes a governance-aligned interpretable AI framework tailored for capital and regulatory reporting. The framework integrates controlled data preparation, policy-aligned regulatory logic, interpretable analytical techniques, and end-to-end governance artifacts across the reporting lifecycle. The approach enables responsible adoption of AI by preserving auditability, supporting independent validation, and aligning analytics with supervisory intent.

Keywords: Audit defensibility, capital reporting, explainable AI, interpretable AI, model governance, regulatory reporting, supervisory transparency.

1. Introduction

This paper introduces an interpretable AI framework designed specifically for capital and regulatory reporting. The contribution is a practical architecture that (1) preserves authoritative rule-based regulatory logic, (2) uses interpretable analytics to enhance transparency and root-cause explainability, and (3) embeds governance artifacts needed for supervisory defensibility [2], [4]. The framework is intended to be applied in regulated banking environments where explainability and audit readiness are non-negotiable requirements. Capital and regulatory reporting is a core supervisory control mechanism used to evaluate the safety and soundness of financial institutions. Regulatory programs governing capital adequacy and stress testing (for example, CCAR and related prudential requirements) require institutions to demonstrate not only accurate reported outcomes but also clear evidence of how those outcomes were produced [1], [2]. In practice, supervisory review evaluates traceability from reported figures back to underlying data sources, transformation logic, assumptions, and controls, consistent with BCBS 239 principles [1]. This expectation has expanded over time as reporting architectures have become more complex and as institutions have adopted distributed data platforms.

In parallel, institutions are investing in AI and advanced analytics to improve operational efficiency, detect anomalies, and accelerate analytical insight. Yet, many machine learning approaches are not easily compatible with regulated reporting environments. When model behavior cannot be explained at an appropriate level of detail, supervisory confidence declines and the institution's ability to defend outcomes under examination is impaired [3]. Model risk management expectations, including the need for independent validation, reproducibility, and controlled change management, further constrain the use of opaque models [2].

2. Materials and Methods

The proposed approach is framed as a system architecture and operating model rather than a single predictive model. The method integrates interpretable analytics as a controlled layer within the reporting lifecycle, consistent with supervisory expectations for traceability and governance [1], [2]. This section describes the framework components, how they interact, and the governance mechanisms that ensure reproducibility and audit defensibility.

2.1. Data and Inputs

The framework assumes standard regulatory reporting inputs commonly available in large institutions: transactional records, reference data, counterparty and product attributes, risk parameters, scenario variables, and supporting metadata [1]. Inputs are treated as controlled datasets subject to access controls, data quality rules, and lineage capture. Where reporting depends on derived attributes, derivation logic is explicitly version-controlled and documented to enable reproduction across reporting cycles, in alignment with supervisory guidance [2].

2.2. Regulatory Logic Decomposition

Regulatory requirements are decomposed into logic components that map supervisory intent to implementable rules. This decomposition includes policy- aligned transformations, controlled assumptions, and calculation steps that produce reportable measures [1], [2]. The decomposition is structured to support traceability such that each output measure can be linked to specific inputs, transformation steps, and assumptions, supporting supervisory challenge responses and independent validation.

2.3. Interpretable Analytics Layer Design

Interpretable AI techniques are applied as transparent analytical overlays rather than replacements for authoritative calculations. Appropriate methods include decision trees, constrained regression, monotonic models, rule lists, and feature attribution techniques that produce human-understandable explanations [3], [4], [5]. Outputs from this layer are designed to be explainable at the feature level and reproducible under controlled data and parameter versions, supporting audit and validation requirements.

2.4. Governance and Control Artifacts

The method embeds governance artifacts across the lifecycle, including documentation of assumptions, lineage records, validation check outcomes, model cards for analytical components, change management approvals, and audit trails [2], [6]. These artifacts are treated as first-class deliverables to ensure that analytics remain defensible under examination and consistent with supervisory expectations.

3. Results and Discussion

3.1. Framework Architecture and Operational Flow

The resulting framework architecture integrates four functional layers Data Sources, Data Preparation and Controls, Regulatory Logic, and Interpretable AI culminating in Reporting Outputs. A cross-cutting Governance and Explainability layer spans all components to ensure traceability, transparency, and audit readiness, consistent with BCBS 239 and SR 11-7 expectations [1], [2].

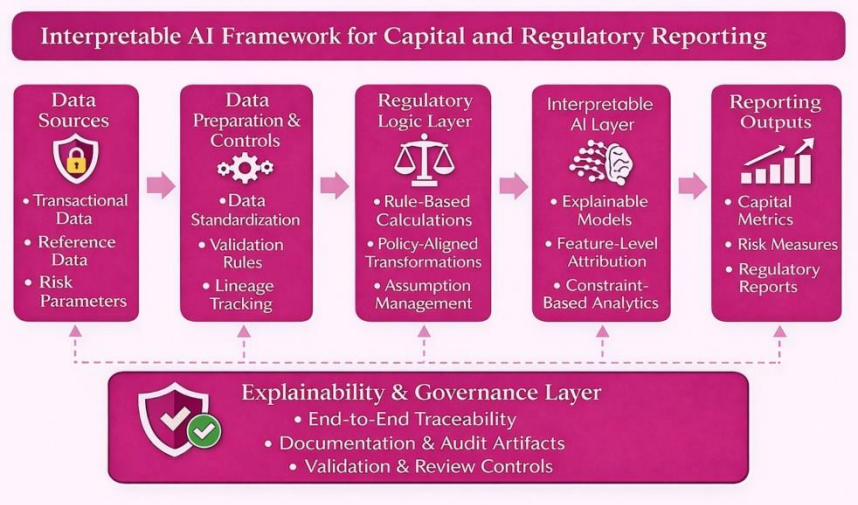

Figure 1: Interpretable AI Framework for Capital and Regulatory Reporting

As shown in Fig. 1, controlled data sources feed a preparation layer that enforces standardization, validation, and lineage capture [1]. The Regulatory Logic layer then applies policy-aligned rules, including controlled assumptions and transformations, to produce authoritative measures [2]. The Interpretable AI layer evaluates data behavior and output drivers using transparent methods, producing explanations suitable for supervisory review [3], [5]. Reporting Outputs are generated through authoritative logic, while interpretability artifacts provide defensible rationale and traceability.

3.2. Why Interpretability is Essential in Supervisory Environments

Interpretability is essential because supervisory review requires more than correlation-based explanations; it requires a defensible mapping between regulatory intent and reported outcomes [2]. When examiners challenge a variance, an outlier, or a change in reported values across cycles, institutions must explain the drivers in a reproducible manner aligned to policy expectations [1]. Interpretable AI supports this by producing explanations that can be expressed as rules, feature contributions, or constrained relationships, consistent with regulatory governance principles [3], [4].

3.3. Governance, Validation, and Audit Defensibility

A key benefit of the framework is that governance is embedded as an architectural requirement rather than an afterthought. Each analytical component is subjected to independent validation, including data integrity checks, explanation stability testing, and back-testing against historical reporting cycles [2]. Model documentation captures purpose, intended use, limitations, and change history, while audit trails record approvals and deployments, reducing remediation risk associated with undocumented model behavior [6].

Validation checkpoints can be integrated at multiple stages pre-ingestion, post- standardization, post-regulatory-logic, and post-interpretability enabling proactive identification of issues before reports are finalized and strengthening supervisory confidence [1], [2].

3.4. Practical Implementation Patterns

Implementation can follow a phased approach. Phase 1 applies interpretable analytics to variance explanation and data quality anomaly detection without altering authoritative calculations. Phase 2 integrates interpretable models to prioritize remediation and identify systemic drivers of recurring findings. Phase 3 extends the framework to controlled automation of diagnostics and exception handling, while maintaining governance and approval controls [3], [4]. Each phase preserves compliance by separating authoritative reporting logic from analytical overlays.

3.5. Limitations and Risk Considerations

The framework is not intended to justify the use of opaque models in reporting. Where predictive models are used for supporting analytics, they must remain interpretable or be paired with strong constraints and explanation artifacts acceptable under supervisory review [3]. Clear labeling, access controls, and governance approvals reduce the risk of inappropriate reliance on analytical outputs [2].

4. Conclusion

Interpretable AI provides a practical pathway for integrating advanced analytics into capital and regulatory reporting without compromising supervisory transparency. The framework presented in this paper demonstrates how interpretability, policy-aligned regulatory logic, and governance artifacts can be integrated into a layered architecture that supports traceability and audit defensibility, consistent with supervisory guidance [1], [2]. By treating explainability and governance as architectural requirements, institutions can modernize reporting operations while maintaining regulatory confidence.

Conflicts of Interest

The author declares that there is no conflict of interest concerning the publishing of this paper.

Acknowledgements

The author thanks professional reviewers and colleagues for feedback on supervisory transparency patterns and reporting governance practices.

References

1. G. Basel Committee on Banking Supervision, “Principles for effective risk data aggregation and risk reporting (BCBS 239),” Bank for International Settlements, 2013. BisPrinciples for effective risk data aggregation and risk reporting
2. Federal Reserve Board, “Supervisory Guidance on Model Risk Management (SR 11-7),” 2011. FederalreserveThe Fed - Supervisory Letter SR 11-7 on guidance on Model Risk Management -- April 4, 2011
3. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, 2019. NatureStop explaining black box machine learning models for high stakes decisions and use interpretable models instead
4. F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv preprint, 2017. ArxivTowards A Rigorous Science of Interpretable Machine Learning
5. S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, 2017. NeuripsA Unified Approach to Interpreting Model Predictions
6. European Banking Authority, “Guidelines on ICT and security risk management,” 2019. EuropaGuidelines on ICT and security risk management | European Banking Authority