A REPORT

ON

INTELLIGENT DATA QUALITY MONITORING USING MACHINE LEARNING FOR REGULATORY REPORTING SYSTEMS

BY

Sujatha Chittiri

ID No.: 2023AC05729

AT

HSBC Global Technologies Pvt Ltd, Hyderabad

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

January, 2026

A REPORT

ON

INTELLIGENT DATA QUALITY MONITORING USING MACHINE LEARNING FOR REGULATORY REPORTING SYSTEMS

BY

Sujatha Chittiri   ID No.: 2023AC05729   Discipline: M.Tech Artificial Intelligence & Machine Learning

Prepared in partial fulfilment of the
WILP Dissertation

AT

HSBC Global Technologies Pvt Ltd, Hyderabad

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

January, 2026

Acknowledgement

I would like to express my sincere gratitude to my supervisors, **Phani Kumar PVSKP** and **Hemendra LK Akuthota**, for their continuous guidance, encouragement, and valuable technical insights throughout the course of this dissertation. I am thankful to the management and colleagues at HSBC Global Technologies Pvt Ltd, Hyderabad, for providing a supportive environment and practical exposure to regulatory reporting systems.

I also express my sincere gratitude to the faculty of BITS Pilani WILP for their academic guidance and continuous support throughout this program. I would like to specifically thank my BITS evaluator, **Prof. Sivagami R**, for her valuable feedback and evaluation of this dissertation. Finally, I express my heartfelt gratitude to my family for their constant encouragement, patience, and support throughout this academic journey.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

WILP Division

Organization: HSBC Global Technologies Pvt Ltd **Location**: Hyderabad

Duration: 3 Months **Date of Start**: 24-Oct-2025

Date of Submission: 1-Feb-2026

Title of the Project:

INTELLIGENT DATA QUALITY MONITORING USING MACHINE LEARNING FOR REGULATORY REPORTING SYSTEMS

ID No./Name of the student: 2023AC05729 / Sujatha Chittiri

Name(s) and Designation(s) of your Supervisor and Additional Examiner:

Phani Kumar PVSKP, Associate Director; Hemendra LK Akuthota, Associate Director

Name of the Faculty mentor:- Prof. Sivagami R

Key Words: Data Quality, Regulatory Reporting, Anomaly Detection, Machine Learning, Explainable AI

Project Areas: Financial Risk Analytics, Data Governance, Artificial Intelligence

Abstract:

Ensuring high data quality in regulatory reporting is a critical requirement for financial institutions, as inaccurate or inconsistent data can lead to regulatory breaches, capital misstatements, and reputational risk. Traditional data quality frameworks in banks are predominantly rule-based and rely on deterministic validation checks. While such systems are essential, they are inherently limited because they can only detect predefined error patterns and are unable to identify previously unknown or complex multivariate anomalies.

This dissertation proposes and implements an **Intelligent Data Quality Monitoring Framework** that combines rule-based validation with unsupervised machine learning and explainable artificial intelligence techniques for regulatory reporting systems. The proposed system integrates multiple anomaly detection models including Isolation Forest, Local Outlier Factor, One-Class SVM, PCA-based anomaly detection, and Autoencoders, and combines them into an ensemble scoring mechanism. A hybrid decision layer fuses rule-based and ML-based results to achieve maximum anomaly coverage while preserving regulatory governance and control.

To ensure transparency and auditability, Explainable AI (XAI) techniques using SHAP (SHapley Additive exPlanations) are incorporated to interpret the behaviour of the machine learning models and to identify the key features driving anomaly detection. An end-to-end operational pipeline is developed, from data generation and quality rule evaluation to ML scoring, unified anomaly decisioning, and business-facing visualization through an interactive dashboard.

The system is evaluated on a realistic synthetic dataset designed to resemble regulatory reporting data used in large banking environments. Experimental results demonstrate that while rule-based systems effectively capture deterministic violations, the machine learning models identify a significant number of additional anomalies that pass all rule checks. The hybrid approach achieves the best overall performance, combining governance, detection power, and explainability.

This work demonstrates that machine learning is not a replacement for rule-based controls, but a powerful complement that significantly enhances the robustness, coverage, and intelligence of regulatory data quality monitoring frameworks.

ch. Sujatha

Signature of Student:
Date: 31-01-2026

Hari Kumar

Signature of Supervisor:
Date: 31-01-2026

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background and Motivation

Financial institutions operate in a highly regulated environment where accurate, consistent, and timely reporting of risk and capital information to supervisory authorities such as the Prudential Regulation Authority (PRA), European Central Bank (ECB), and other regulators is mandatory. Regulatory reports are produced from complex data pipelines that integrate information from multiple upstream systems including trading, risk, finance, and reference data platforms. Due to the scale, heterogeneity, and continuous evolution of these systems, data quality issues such as missing values, inconsistent attributes, incorrect mappings, duplicate records, and out-of-range values are inevitable.

Traditionally, data quality in regulatory reporting has been ensured through rule-based validation frameworks, where predefined business rules and reconciliation checks are applied at various stages of the reporting pipeline. While such approaches are effective in capturing known and well-defined issues, they struggle to detect complex, multivariate, and previously unseen anomalies that arise due to unusual combinations of attributes or subtle distributional shifts. With the increasing volume and complexity of regulatory data, there is a growing need for intelligent, adaptive, and data-driven approaches to complement existing rule-based controls.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have demonstrated strong potential in anomaly detection, pattern discovery, and automated monitoring across several domains. In particular, unsupervised learning techniques are well suited for regulatory reporting environments, where labelled examples of erroneous data are scarce or unavailable. These techniques can learn the structure of "normal" data and identify records that deviate significantly from learned patterns.

1.2 Regulatory Reporting and Data Quality Challenges

Regulatory reporting data is characterized by high dimensionality, strong interdependencies between attributes, and strict consistency requirements across exposure, risk, and capital measures. Typical reports include information such as exposure amounts, risk weights, risk-weighted assets (RWA), capital requirements, maturity profiles, collateral indicators, and credit quality classifications. Errors in any of these fields can lead to incorrect regulatory submissions, supervisory findings, reputational damage, and potential capital add-ons.

A key challenge is that not all data quality issues can be expressed as deterministic rules. For example, while a missing value or an invalid currency code can be captured using simple checks, an exposure with an unusual combination of maturity, risk weight, and capital requirement may still pass all rule-based validations while being economically implausible. Detecting such cases requires learning the joint behavior of multiple attributes, which motivates the use of machine learning based techniques.

1.3 Problem Statement

The problem addressed in this dissertation is the design and implementation of an intelligent data quality monitoring framework for regulatory reporting systems that goes beyond traditional rule-based validation. The objective is to develop a hybrid system that combines deterministic data quality rules with unsupervised machine learning models to detect both known and unknown anomalies in regulatory datasets, and to provide explainable insights to support investigation and remediation by data quality and risk management teams.

1.4 Objectives of the Dissertation

The main objectives of this work are:

- To analyze the limitations of traditional rule-based data quality monitoring in regulatory reporting systems.
- To design and implement an unsupervised machine learning based anomaly detection framework using multiple models such as Isolation Forest, Local Outlier Factor, One-Class SVM, PCA, and Autoencoders.
- To develop an ensemble and hybrid scoring mechanism that combines rule-based and ML-based signals.
- To evaluate and compare the performance of individual models, the ensemble, and the hybrid system using appropriate metrics.
- To integrate Explainable AI (XAI) techniques using SHAP to provide feature-level interpretability of model decisions.
- To build an operational dashboard for monitoring, investigation, and explainability of data quality issues.

1.5 Contributions of This Work

The key contributions of this dissertation are:

- A practical hybrid architecture combining rule-based and machine learning based data quality monitoring for regulatory reporting systems.
- A comparative evaluation of multiple unsupervised anomaly detection models and an ensemble strategy.
- An explainability layer based on SHAP that provides interpretable insights into model behavior for regulatory and audit purposes.
- An end-to-end prototype implementation including data generation, rule engine, ML pipeline, evaluation framework, and operational dashboard.

1.6 Organization of the Report

The remainder of this report is organized as follows. Chapter 2 reviews related work in data quality, anomaly detection, and explainable AI. Chapter 3 describes the source dataset and feature engineering strategy. Chapter 4 presents the overall system architecture. Chapters 5 and 6 detail the rule-based engine and machine learning models respectively. Chapter 7 introduces the hybrid scoring framework. Chapter 8 presents the evaluation methodology and comparative results. Chapter 9 discusses the explainability framework based on SHAP. Chapter 10 describes the operational dashboard. Chapter 11 discusses the results and business impact. Chapter 12 outlines limitations and future work, and Chapter 13 concludes the dissertation.

Chapter 2: Literature Review

2.1 Data Quality in Financial Regulatory Systems

Financial institutions are required to submit large volumes of regulatory data to supervisory authorities such as the Prudential Regulation Authority (PRA), European Central Bank (ECB), and European Banking Authority (EBA). These reports are used for capital adequacy assessment, stress testing, and systemic risk monitoring. Due to the complexity of data pipelines spanning multiple source systems, regulatory datasets are prone to a wide range of data quality issues including missing values, inconsistent classifications, invalid reference codes, duplicated records, and incorrect derivations of risk and capital measures.

Traditional approaches to data quality management in banking environments rely heavily on deterministic rule-based validation frameworks. These typically include completeness checks, domain validations, reconciliation rules, and cross-field consistency checks. While such methods are effective for capturing known and well-defined issues, they are inherently limited to scenarios that can be explicitly expressed as rules. As regulatory datasets continue to grow in size and complexity, rule-based approaches alone are increasingly insufficient to detect subtle, multivariate, and previously unseen data anomalies.

Several industry and academic studies have highlighted the need for more intelligent and adaptive data quality monitoring mechanisms that can complement existing rule frameworks with data-driven techniques.

2.2 Anomaly Detection in Financial Data

Anomaly detection refers to the task of identifying patterns or observations that deviate significantly from the expected behavior of a dataset. In financial systems, anomalies may correspond to operational errors, data integration issues, system defects, or exceptional but valid business events. In the context of regulatory reporting, undetected anomalies can lead to incorrect capital calculations, supervisory findings, and potential regulatory penalties.

Financial anomaly detection presents several challenges:

1. The data is high-dimensional and highly interdependent.

2. True anomalies are rare compared to normal records.

3. Labeled examples of erroneous data are usually unavailable or unreliable.

4. The notion of “anomaly” may evolve over time as business and regulatory conditions change.

Due to these characteristics, unsupervised and semi-supervised anomaly detection methods are widely preferred in financial data quality applications.

2.3 Unsupervised Learning for Anomaly Detection

Unsupervised anomaly detection methods operate without requiring labeled training data. Instead, they attempt to learn the structure of normal data and identify observations that deviate significantly from this learned structure. This paradigm is particularly well suited for regulatory reporting environments, where obtaining reliable labels for erroneous data is difficult and costly.

Schölkopf et al. introduced the concept of One-Class Classification for novelty detection, laying the foundation for several unsupervised approaches. Since then, a wide range of techniques have been proposed, including density-based, distance-based, isolation-based, projection-based, and reconstruction-based methods.

Recent surveys in machine learning literature consistently identify Isolation Forest, Local Outlier Factor, One-Class SVM, PCA-based methods, and Autoencoders as among the most widely used and effective unsupervised anomaly detection techniques in structured data domains.

2.4 Review of Key Models

2.4.1 Isolation Forest

Isolation Forest, proposed by Liu et al., is based on the principle that anomalies are easier to isolate than normal points. Instead of profiling normal data, the algorithm explicitly isolates observations using random partitioning. Records that require fewer splits to isolate are more likely to be anomalies. Isolation Forest is computationally efficient, scales well to large datasets, and has been widely adopted in industrial anomaly detection applications.

2.4.2 Local Outlier Factor (LOF)

Local Outlier Factor, introduced by Breunig et al., is a density-based method that measures how isolated a point is with respect to its local neighborhood. A data point is considered anomalous if its local density is significantly lower than that of its neighbors. LOF is particularly useful for detecting local anomalies in datasets where global distributions may be complex.

2.4.3 One-Class Support Vector Machine (OCSVM)

One-Class SVM, proposed by Schölkopf et al., learns a decision boundary that encloses the majority of the data and classifies points lying outside this boundary as anomalies. It is a powerful theoretical framework for novelty detection but can be sensitive to parameter tuning and feature scaling.

2.4.4 PCA-based Anomaly Detection

Principal Component Analysis (PCA) has been widely used for anomaly detection by modeling the dominant subspace of normal data. Observations with high reconstruction error or large projection residuals are considered anomalous. PCA-based methods are computationally efficient and provide a linear approximation of data structure.

2.4.5 Autoencoders for Anomaly Detection

Autoencoders are neural networks trained to reconstruct their input data through a compressed latent representation. When trained on mostly normal data, they learn to reconstruct normal patterns well but produce higher reconstruction errors for anomalous records. Autoencoders have been successfully applied to anomaly detection in finance, cybersecurity, and industrial monitoring.

2.5 Ensemble and Hybrid Approaches in Literature

No single anomaly detection algorithm performs best across all datasets and anomaly types. As a result, ensemble approaches that combine multiple detectors have gained significant attention in recent research. By aggregating the outputs of diverse models, ensemble methods improve robustness and reduce sensitivity to the weaknesses of individual techniques.

In parallel, several studies in financial systems advocate hybrid frameworks that combine rule-based controls with machine learning models. Rules capture known regulatory and business constraints, while machine learning models discover unknown or emergent patterns. Such hybrid systems provide both reliability and adaptiveness, making them particularly suitable for regulatory environments.

2.6 Explainable AI in Financial Systems

A major limitation of many machine learning models, especially in regulated industries, is their lack of interpretability. Regulatory and audit functions require not only anomaly detection but also a clear explanation of why a record was flagged.

Explainable AI (XAI) techniques address this requirement by providing feature-level attributions for model outputs. SHAP (SHapley Additive exPlanations) is one of the most widely accepted methods for model-agnostic explainability, based on cooperative game theory. SHAP has been increasingly adopted in financial risk modeling, fraud detection, and regulatory analytics due to its strong theoretical foundation and consistent explanations.

For complex models such as ensembles and neural networks, surrogate modeling approaches are often used to approximate model behavior and apply SHAP-based explanations.

2.7 Summary and Research Gap

The literature clearly demonstrates:

1. The importance of robust data quality management in regulatory systems.
2. The effectiveness of unsupervised anomaly detection for high-dimensional financial data.
3. The complementary strengths of Isolation Forest, LOF, OCSVM, PCA, and Autoencoders.
4. The growing adoption of ensemble and hybrid approaches.
5. The critical role of explainability in regulated environments.

However, most existing works either focus purely on rule-based systems or purely on machine learning models. There is a relative lack of practical, end-to-end frameworks that integrate:

1. Rule-based validation,
2. Multiple unsupervised models,
3. Ensemble scoring,
4. Hybrid decision logic,
5. And explainable AI,
   within a single operational monitoring system for regulatory reporting.

This dissertation addresses this gap by proposing and implementing a unified, explainable, hybrid data quality monitoring framework tailored for regulatory reporting systems.

Chapter 3: Source Dataset and Feature Engineering

3.1 Nature of Regulatory Data in Banks

Large financial institutions such as HSBC maintain extensive regulatory data repositories that consolidate information from multiple upstream systems including trading platforms, credit risk engines, collateral management systems, and finance ledgers. These datasets are used to generate regulatory returns such as COREP, FINREP, and other supervisory submissions to authorities like the PRA, ECB, and EBA.

A typical regulatory risk dataset contains hundreds of attributes per exposure or contract. These include identifiers, counterparty details, product classifications, exposure measures, risk parameters, capital metrics, accounting attributes, and reporting-specific metadata. In addition, several derived fields are computed through complex regulatory formulas and aggregation pipelines.

In real production environments, the full dataset is significantly larger and more complex than what can be used in an academic prototype due to confidentiality, data volume, and access restrictions. Therefore, this dissertation uses a synthetic dataset that is designed to be structurally and statistically representative of real regulatory reporting data used in large banks.

3.2 Typical Fields in Regulatory Risk Data (HSBC-Inspired)

Based on the author's professional experience in regulatory reporting systems and industry standards, a real-world dataset typically contains the following categories of fields:

1. Identification and Metadata Fields
Account ID, Contract ID, Counterparty ID, Legal Entity, Booking Entity, Report Date, Source System, Snapshot Version.

2. Counterparty and Product Attributes
Counterparty Type, Sector, Rating Grade, Product Type, Exposure Class, Portfolio, Country Code, Currency.

3. Exposure and Risk Measures
Exposure Amount, Drawn Amount, Undrawn Amount, Risk Weight, Credit Conversion Factor, Probability of Default (PD), Loss Given Default (LGD), Maturity, Collateral Indicator.

4. Capital and Regulatory Measures
Risk-Weighted Assets (RWA), Capital Requirement, Capital Floor Adjustments, Exposure-to-Capital Ratio, Leverage Exposure.

5. Contract and Structural Attributes
Maturity Bucket, Interest Type, Seniority, Netting Flag, Collateral Type, Guarantee Indicator.

In total, such datasets typically contain several dozens to several hundreds of columns.

3.3 Feature Categories Used in This Work

For the purposes of this dissertation, a focused subset of features was selected to represent the most critical drivers of regulatory risk and capital computation. The features used fall into the following categories:

1. Exposure and Size Indicators

1. Exposure_Amount
2. RWA
3. Exposure_to_Capital_Ratio

2. Risk and Credit Quality Indicators

1. Risk_Weight
2. Credit_Quality_Step

3. Capital Measures

1. Capital_Requirement

4. Contract and Structural Attributes

1. Maturity_Months
2. Is_Collateralized

5. Categorical Descriptors

1. Counterparty_Type
2. Product_Type
3. Exposure_Class
4. Country_Code
5. Currency

These were combined with derived and normalized numerical representations during preprocessing.

3.4 Feature Selection Criteria

The selection of features from the much larger conceptual regulatory dataset was based on the following criteria:

1. Regulatory Relevance
Only fields that directly influence or explain capital computation, risk measurement, or regulatory classification were considered.

2. Cross-Field Dependency Potential
Features that participate in multi-variable relationships (e.g., Exposure, Risk Weight, Capital Requirement, RWA, Maturity) were prioritized because such relationships are where rule-based systems typically fail.

3. Anomaly Sensitivity
Features whose abnormal combinations can create economically implausible but rule-compliant records were preferred.

4. Availability and Explainability
Features were chosen such that detected anomalies can be meaningfully interpreted and explained to business and regulatory stakeholders.

5. Practical Dimensionality Control
Using all available fields in a real regulatory dataset would introduce excessive noise, sparsity, and multicollinearity. Therefore, a curated subset was selected to ensure stable and interpretable model behavior.

3.5 Final Feature Set Used in This Work

Based on the above criteria, the final feature set used for machine learning modeling consists of:

1. Exposure_Amount
2. Risk_Weight
3. RWA
4. Capital_Requirement
5. Exposure_to_Capital_Ratio
6. Credit_Quality_Step
7. Maturity_Months
8. Is_Collateralized

9. Counterparty_Type (encoded)

10. Product_Type (encoded)

11. Exposure_Class (encoded)

12. Country_Code (encoded)

13. Currency (encoded)

These features jointly capture size, risk, capital, structural, and classification aspects of each regulatory record.

3.6 Synthetic Data Generation and Bias Control

Due to confidentiality and regulatory constraints, real production data cannot be used in an academic setting. Therefore, a synthetic data generator was implemented to simulate realistic regulatory datasets. The generator creates:

1. Statistically consistent normal records

2. Deterministic rule violations (e.g., invalid codes, missing values, capital formula breaches)

3. Subtle multivariate anomalies that do not violate explicit rules but deviate from learned patterns

To avoid biasing the machine learning models, the injection of ML-only anomalies is designed to affect multivariate relationships rather than individual fields in isolation. This ensures that machine learning models are not simply rediscovering the same patterns as rule-based checks.

3.7 Feature Justification Using Explainable AI (SHAP)

Feature importance analysis using SHAP was applied to both the Isolation Forest and the Autoencoder surrogate model. The results show that features such as:

1. Exposure_to_Capital_Ratio

2. Risk_Weight

3. Credit_Quality_Step

4. Exposure_Amount

5. RWA

6. Maturity_Months

consistently contribute the most to anomaly decisions. This empirically validates the feature selection strategy and confirms that the models rely on economically and regulatorily meaningful attributes rather than noise variables.

3.8 Summary

This chapter demonstrated that although real regulatory datasets contain hundreds of attributes, a carefully selected subset of economically and regulatorily meaningful features is sufficient to build effective, interpretable, and robust anomaly detection models. The chosen feature set reflects both industry practice and empirical evidence from explainability analysis, thereby addressing concerns regarding feature justification and model credibility.

Chapter 4: System Architecture and End-to-End Pipeline

4.1 Overall Architecture

The proposed system is designed as a modular, end-to-end data quality monitoring framework for regulatory reporting environments. The architecture follows a layered pipeline approach that integrates deterministic rule-based validation with machine learning based anomaly detection, and augments both with explainability and operational visualization.

At a high level, the system consists of the following major components:

1. Synthetic Regulatory Data Generator
2. Rule-Based Data Quality Engine
3. Machine Learning Scoring Engine (Multi-model + Ensemble)
4. Hybrid Scoring and Decision Layer
5. Model Evaluation and Explainability Layer
6. Operational Monitoring Dashboard

Each component is designed to be independently testable, configurable, and extensible, reflecting a production-style data quality monitoring architecture.

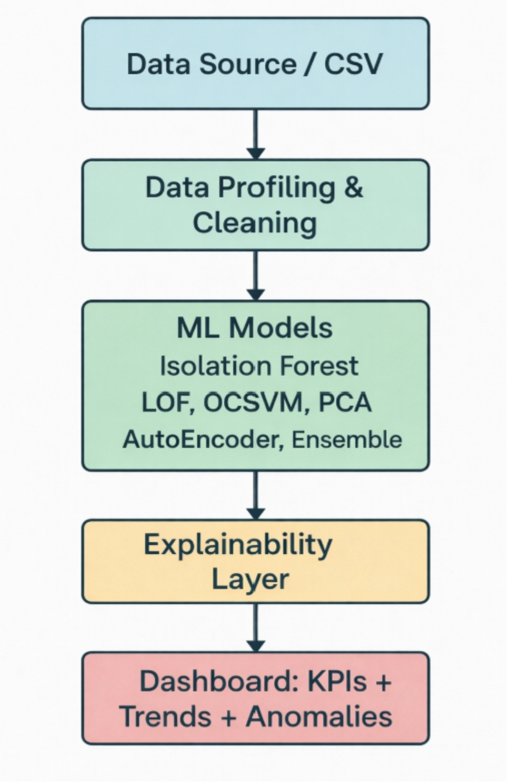


Figure 1: Architecture Diagram

4.2 End-to-End Processing Flow

The complete data processing pipeline operates in the following sequence:

1. **Data Generation**
   A synthetic but realistic regulatory dataset is generated, containing both clean records and injected anomalies (rule-based and ML-only).

2. **Rule-Based Validation**
   The dataset is passed through a deterministic rule engine that applies regulatory and data quality checks such as:

   1. Completeness validation

   2. Domain and code-list checks

   3. Duplicate detection

   4. Capital formula consistency checks
      Each record is assigned a rule-based anomaly flag and a rule severity score.

3. **Machine Learning Scoring**
   The same dataset is processed by multiple unsupervised anomaly detection models:

    1. Isolation Forest
    2. Local Outlier Factor
    3. One-Class SVM
    4. PCA-based detector
    5. Autoencoder

    Each model produces a normalized anomaly score for every record.

4. **Ensemble Aggregation**
   The individual ML scores are combined into a single ensemble score using a weighted averaging strategy to improve robustness and stability.

5. **Hybrid Decision Layer**
   Rule-based and ML-based signals are fused to produce:

    1. Rule-only anomalies
    2. ML-only anomalies
    3. Overlapping anomalies
    4. Final anomaly decision flag

6. **Evaluation and Explainability**
   The outputs are evaluated using ROC-AUC and Precision@K metrics and explained using SHAP-based feature attribution.

7. **Operational Dashboard**
   The final scored dataset is visualized in an interactive dashboard for monitoring, investigation, prioritization, and explainability.

4.3 Architectural Design Principles

The system is designed based on the following principles:

1. Modularity
Each processing step is implemented as a separate script or module, enabling independent development, testing, and replacement.

2. Reproducibility
All steps operate on well-defined input and output files, ensuring full reproducibility of results.

3. Explainability by Design
Explainability is treated as a first-class component rather than an afterthought, with SHAP integrated into the modeling pipeline.

4. Hybrid Governance
The architecture explicitly supports the coexistence of deterministic rules and probabilistic ML models, reflecting real regulatory governance requirements.

5. Operational Readiness
The final outputs are structured to support consumption by dashboards and downstream investigation workflows.

4.4 Component-Level Description

4.4.1 Synthetic Data Generator

The data generator module creates a dataset that structurally resembles real regulatory risk data used in banking environments. It produces:

1. Normal records following realistic statistical distributions
2. Rule-based anomalies such as invalid codes, missing values, and capital formula breaches
3. ML-only anomalies that preserve rule validity but break multivariate relationships

The output of this stage serves as the raw input to both the rule engine and the ML pipeline.

4.4.2 Rule-Based Data Quality Engine

The rule engine applies deterministic data quality checks based on regulatory logic and industry best practices. It produces:

1. Individual rule violation indicators
2. A consolidated rule-based anomaly flag
3. A rule severity score reflecting the number and criticality of violations

This component represents the traditional data quality control layer used in banks.

4.4.3 Machine Learning Scoring Engine

The ML engine performs the following steps:

1. Feature preprocessing and scaling

2. Encoding of categorical variables

3. Training and scoring using multiple unsupervised models

4. Normalization of model scores

5. Ensemble aggregation of model outputs

Each record receives a final ML anomaly score and ML anomaly flag.

4.4.4 Hybrid Scoring and Decision Layer

The hybrid layer combines:

1. RULE_ANOMALY_FLAG
2. ML_ANOMALY_FLAG

to produce:

1. FINAL_ANOMALY_FLAG
2. Classification of anomalies into:
    1. Rule-only
    2. ML-only
    3. Both

This design ensures that machine learning complements rather than replaces existing rule-based governance.

4.4.5 Model Evaluation and Explainability Layer

This layer computes:

1. ROC-AUC and Precision@K for:
    1. Individual models
    2. Ensemble model
    3. Rule-based system
    4. Hybrid system

It also generates:

1. SHAP feature importance for Isolation Forest

2. Surrogate SHAP explanations for Autoencoder

These outputs provide both quantitative performance evidence and qualitative interpretability.

4.4.6 Operational Dashboard

The dashboard consumes the final scored dataset and provides:

1. KPI monitoring

2. Anomaly distribution views

3. Risk-based prioritization

4. Filtering and investigation

5. Explainability views

6. ML-only anomaly analysis

This makes the system usable by data quality analysts, risk managers, and governance teams.

4.5 Script-Level Pipeline Mapping

The implemented pipeline corresponds to the following scripts:

1. data_generator.py → Creates synthetic regulatory dataset

2. dq_engine.py → Applies rule-based checks

3. model_scoring.py → Trains and scores ML models + ensemble

4. unified_scoring.py → Produces hybrid final decisions

5. model_evaluation.py → Computes metrics and comparisons

6. explain_shap_if.py and explain_shap_ae.py → Explainability

7. app.py → Operational dashboard

This explicit modularization ensures clarity, auditability, and extensibility.

4.6 Summary

This chapter presented the complete system architecture and end-to-end pipeline for the proposed intelligent data quality monitoring framework. The design reflects real-world regulatory system constraints by combining deterministic controls, machine learning intelligence, explainability, and operational usability into unified and production-style architecture.

Chapter 5: Rule-Based Data Quality Engine

5.1 Motivation for Rule-Based Data Quality Checks

Rule-based data quality controls form the backbone of regulatory reporting systems in financial institutions. Regulators expect banks to demonstrate strong deterministic controls over critical data elements, including completeness, validity, consistency, and correctness of derived regulatory metrics. Such rules represent explicit business, accounting, and regulatory requirements and are therefore non-negotiable components of any governance framework.

In real-world regulatory pipelines, rule engines are used to enforce:

1. Presence of mandatory fields
2. Validity of reference data codes
3. Structural consistency of records
4. Mathematical correctness of capital and risk calculations
5. Uniqueness and duplication constraints

While these checks are essential, they are inherently limited to known and explicitly expressible conditions. This dissertation therefore treats the rule engine as a **first line of defense** that is later augmented by machine learning based anomaly detection.

5.2 Design Objectives of the Rule Engine

The rule engine in this work was designed with the following objectives:

1. **Regulatory Alignment**
   The rules should reflect typical expectations from regulators such as PRA, ECB, and EBA regarding data quality in supervisory reporting.
2. **Configurability**
   All thresholds and rule parameters are externalized in a configuration file, allowing governance teams to adapt controls without code changes.

3. **Explainability**
   Each rule produces interpretable pass/fail outcomes and contributes transparently to the final rule severity score.

4. **Auditability**
   The rule results are stored at record level, enabling traceability and audit inspection.

5.3 Categories of Rules Implemented

The implemented rule engine covers the following major categories:

5.3.1 Completeness Checks

Critical regulatory fields such as:

1. Account_ID
2. Report_Date
3. Exposure_Amount
4. Risk_Weight
5. Capital_Requirement
6. Country_Code
7. Currency
8. Counterparty_Type

are checked for missing or null values. Any record violating completeness requirements is immediately flagged as a rule anomaly.

5.3.2 Domain and Code List Validation

Several fields are validated against controlled reference lists, for example:

1. Currency must belong to an allowed ISO currency set
2. Country_Code must belong to a defined country list

This simulates the reference data validation typically enforced in regulatory systems.

5.3.3 Range and Boundary Checks

Numeric fields are validated against economically and regulatorily plausible ranges:

1. Exposure_Amount must be non-negative

2. Risk_Weight must lie between 0 and 1

3. Maturity_Months must fall within acceptable contractual limits

Such rules prevent structurally invalid values from propagating into regulatory reports.

5.3.4 Capital Formula Consistency Checks

One of the most critical regulatory validations is the consistency between:

1. Exposure_Amount

2. Risk_Weight

3. Capital_Requirement

The engine verifies that:

Capital_Requirement ≈ Exposure_Amount × Risk_Weight × Capital_Factor

within a configurable tolerance. Records violating this relationship are flagged as severe rule anomalies.

5.3.5 Duplicate Detection

Duplicate records are detected based on the composite business key:

1. Account_ID

2. Report_Date

Any duplicates are flagged as data quality violations, reflecting regulatory expectations of record uniqueness.

5.3.6 Outlier-Based Sanity Checks

In addition to strict rule violations, simple statistical sanity checks are applied, such as:

1. Exposure being excessively large relative to country or portfolio medians

These checks represent lightweight deterministic safeguards before ML-based detection.

5.4 Threshold Management and Configuration

All rule thresholds are managed through an external configuration file (dq_config.json). This includes:

1. Allowed code lists
2. Numerical bounds
3. Capital tolerance levels
4. Duplicate tolerance
5. Outlier multipliers

It is important to note that regulators generally **do not prescribe fixed numeric thresholds**. Instead, they require institutions to define, justify, and govern their own thresholds. The values used in this work are therefore **proxy thresholds inspired by industry best practices** and internal governance standards.

5.5 Rule Severity Scoring Mechanism

Each rule violation contributes to a cumulative:

DQ_RULE_SCORE

For every record:

1. If one or more critical rules are violated,
   → RULE_ANOMALY_FLAG = 1
2. Otherwise,
   → RULE_ANOMALY_FLAG = 0

This binary flag represents the traditional data quality gate used in regulatory systems.

5.6 Output of the Rule Engine

For each record, the rule engine produces:

1. Individual rule violation indicators
2. DQ_RULE_SCORE (severity)
3. RULE_ANOMALY_FLAG (final rule decision)

These outputs are persisted and later combined with machine learning results in the hybrid decision layer.

5.7 Role of the Rule Engine in the Hybrid Framework

The rule engine is intentionally **not replaced** by machine learning. Instead, it serves as:

1. A deterministic safety net
2. A governance anchor
3. A regulatory compliance baseline

Machine learning is used only to:

Detect anomalies that **escape explicit rule definitions**.

This design aligns with regulatory expectations of **controlled, explainable, and auditable use of AI** in critical reporting systems.

5.8 Summary

This chapter presented the design and implementation of the rule-based data quality engine used in the proposed system. While deterministic rules remain essential for regulatory compliance, their limitations in capturing complex, multivariate anomalies motivate the integration of machine learning techniques, which are introduced in the next chapter.

Chapter 6: Machine Learning Models for Anomaly Detection

6.1 Motivation for Machine Learning Based Detection

While rule-based data quality checks are essential for enforcing explicit regulatory and business constraints, they are inherently limited to conditions that can be manually specified. In complex regulatory datasets, many data quality issues arise not from individual field violations, but from **unusual combinations of multiple attributes** that remain individually valid. Examples include implausible combinations of maturity, exposure, risk weight, and capital requirement that satisfy all explicit rules but violate economic intuition.

Machine learning based anomaly detection addresses this limitation by **learning the normal structure of data** and identifying observations that deviate from this structure. In regulatory environments, labelled examples of erroneous data are scarce and unreliable. Therefore, **unsupervised learning methods** are particularly suitable, as they do not require labelled training data and can adapt to evolving data distributions.

6.2 Unsupervised Anomaly Detection Paradigm

Unsupervised anomaly detection models attempt to characterize the majority of the data as "normal" and assign anomaly scores to observations based on:

1. Degree of isolation
2. Local density deviation
3. Distance from learned boundary
4. Reconstruction error
5. Projection residuals

Records that strongly deviate from learned normal behavior are assigned higher anomaly scores. This paradigm is widely adopted in financial systems, fraud detection, cybersecurity, and operational risk monitoring.

6.3 Data Preprocessing and Feature Engineering

Before applying machine learning models, the following preprocessing steps are applied:

1. Selection of economically meaningful features (as described in Chapter 3)
2. Scaling of numerical features using standardization
3. Encoding of categorical variables
4. Handling of missing values and invalid records
5. Construction of a unified feature matrix for modeling

These steps ensure that models operate on comparable and well-conditioned input representations.

6.4 Models Implemented

To satisfy the requirement of exploring multiple state-of-the-art unsupervised models and to ensure robustness, this dissertation implements **five complementary anomaly detection techniques**.

6.4.1 Isolation Forest

Isolation Forest isolates observations using random partitioning of the feature space. Anomalies are expected to be isolated in fewer splits and therefore receive higher anomaly scores. The method is computationally efficient, scales well to large datasets, and is widely used in industrial anomaly detection.

Strengths:

1. Fast and scalable
2. Works well in high-dimensional data
3. Does not rely on distance or density assumptions

6.4.2 Local Outlier Factor (LOF)

LOF is a density-based method that measures how isolated a point is relative to its neighbors. Records that lie in sparse regions compared to their local neighborhood are flagged as anomalies.

Strengths:

1. Detects local anomalies
2. Effective when anomalies form small clusters

Limitations:

1. Computationally expensive
2. Sensitive to neighborhood size parameter

6.4.3 One-Class Support Vector Machine (OCSVM)

OCSVM learns a boundary that encloses the majority of data points and classifies points outside this boundary as anomalies. It provides a strong theoretical foundation for novelty detection.

Strengths:

1. Strong mathematical formulation
2. Flexible kernel-based decision boundaries

Limitations:

1. Sensitive to kernel and parameter tuning
2. Less scalable to very large datasets

6.4.4 PCA-Based Anomaly Detection

Principal Component Analysis is used to model the dominant subspace of normal data. Records with high reconstruction error or projection residuals are considered anomalous.

Strengths:

1. Simple and interpretable
2. Computationally efficient
3. Captures global linear structure

Limitations:

1. Limited to linear relationships

6.4.5 Autoencoder-Based Anomaly Detection

Autoencoders are neural networks trained to reconstruct their input data through a compressed latent representation. When trained primarily on normal data, they reconstruct normal patterns accurately but produce higher reconstruction errors for anomalous records.

Strengths:

1. Captures non-linear relationships
2. Powerful representation learning

Limitations:

1. Less interpretable
2. Computationally more expensive

6.5 Rationale for Using Multiple Models

No single anomaly detection algorithm is universally optimal. Different models capture different notions of abnormality:

1. Isolation-based (Isolation Forest)
2. Density-based (LOF)
3. Boundary-based (OCSVM)
4. Subspace-based (PCA)

5. Reconstruction-based (Autoencoder)

Using multiple complementary models increases robustness and reduces dependence on any single modeling assumption.

This directly addresses the mid-semester feedback to:

“Explore more unsupervised models and compare with state-of-the-art techniques from literature.”

6.6 Ensemble Scoring Strategy

Each model produces a normalized anomaly score for every record. These scores are combined using a weighted averaging strategy to produce a single:

ENSEMBLE_SCORE

This ensemble approach:

1. Reduces noise and instability from individual models
2. Improves robustness across anomaly types
3. Produces more stable ranking of risky records

A record is flagged as an ML anomaly if its ensemble score exceeds a configurable threshold.

6.7 Output of the ML Scoring Layer

For each record, the ML pipeline produces:

1. Individual model scores:
    1. IF_SCORE
    2. LOF_SCORE
    3. OCSVM_SCORE
    4. PCA_SCORE
    5. AE_SCORE
2. ENSEMBLE_SCORE
3. ML_ANOMALY_FLAG

These outputs are persisted and later combined with rule-based results in the hybrid scoring framework.

6.8 Summary

This chapter presented the machine learning layer of the proposed system, which uses multiple unsupervised anomaly detection models and an ensemble strategy to identify complex, multivariate, and previously unseen data quality issues. These models complement the deterministic rule engine and form the intelligent detection component of the overall framework.

The next chapter introduces the **hybrid scoring framework**, which combines rule-based and machine learning based signals into a unified anomaly decision mechanism.

Chapter 7: Hybrid Scoring Framework

7.1 Motivation for a Hybrid Approach

In regulatory reporting environments, deterministic rule-based controls are mandatory and form the first line of defense against data quality issues. However, as discussed in earlier chapters, rule-based systems are inherently limited to detecting only those issues that can be explicitly specified in advance. Machine learning based anomaly detection, on the other hand, is capable of identifying complex, multivariate, and previously unseen patterns but produces probabilistic rather than deterministic decisions.

From a governance and regulatory perspective, neither approach is sufficient in isolation. A purely rule-based system cannot detect unknown anomaly patterns, while a purely ML-based system lacks the deterministic guarantees and auditability required in critical regulatory processes. Therefore, this work adopts a **hybrid scoring framework** that combines the strengths of both approaches.

7.2 Inputs to the Hybrid Decision Layer

The hybrid layer receives the following inputs for each record:

1. **RULE_ANOMALY_FLAG**: Binary output of the rule engine

2. **DQ_RULE_SCORE**: Severity score of rule violations

3. **ML_ANOMALY_FLAG**: Binary output of the ensemble ML detector

4. **ENSEMBLE_SCORE**: Continuous ML anomaly score

These signals jointly represent deterministic and probabilistic evidence of abnormality.

7.3 Hybrid Decision Logic

The final anomaly decision is based on the following logic:

1. If RULE_ANOMALY_FLAG = 1 and ML_ANOMALY_FLAG = 0 → **Rule-only anomaly**
2. If RULE_ANOMALY_FLAG = 0 and ML_ANOMALY_FLAG = 1 → **ML-only anomaly**
3. If RULE_ANOMALY_FLAG = 1 and ML_ANOMALY_FLAG = 1 → **Rule + ML anomaly**
4. If both flags are 0 → **Clean record**

A consolidated flag:

FINAL_ANOMALY_FLAG

is set to 1 if either the rule-based or ML-based flag is active.

7.4 Interpretation of ML-Only Anomalies

ML-only anomalies are of particular importance in this framework. These are records that:

1. Satisfy all explicit regulatory and business rules
2. Yet exhibit unusual multivariate patterns compared to the rest of the dataset

Examples include economically implausible but rule-consistent combinations of:

1. Maturity and exposure
2. Risk weight and capital requirement
3. Credit quality and portfolio classification

These cases demonstrate the **incremental value of machine learning** beyond traditional data quality checks.

7.5 Overlap Analysis

The hybrid framework explicitly categorizes anomalies into three groups:

1. **Rule-only anomalies**: Known and deterministic violations
2. **ML-only anomalies**: Unknown, emergent, and behavioral anomalies
3. **Overlapping anomalies**: Severe cases detected by both systems

This categorization is crucial for:

1. Prioritization of investigation
2. Root cause analysis
3. Governance reporting and control improvement

7.6 Governance and Audit Perspective

From a regulatory governance standpoint, the hybrid framework offers several advantages:

1. Rules remain the authoritative compliance gate
2. ML acts as a **second line of defense**
3. ML does not override rules, only augments them
4. All decisions remain traceable and explainable

This design aligns with regulatory expectations regarding the controlled use of AI in critical systems.

7.7 Output of the Hybrid Scoring Layer

For each record, the hybrid layer produces:

1. FINAL_ANOMALY_FLAG
2. FLAG_SOURCE ∈ {Clean, Rule, ML, Rule + ML}
3. Retained RULE and ML scores for investigation and prioritization

This enriched dataset becomes the single source of truth for evaluation and dashboarding.

7.8 Summary

This chapter introduced the hybrid scoring framework that fuses deterministic rule-based validation with probabilistic machine learning based anomaly detection. The hybrid design ensures regulatory compliance, operational robustness, and enhanced detection capability, while preserving governance, auditability, and explainability.

Chapter 8: Model Evaluation and Comparative Analysis

8.1 Evaluation Challenges in Unsupervised Anomaly Detection

Evaluating anomaly detection systems in regulatory reporting environments presents inherent challenges. In real-world settings, comprehensive and reliable labels for erroneous data are typically unavailable. Most detected issues are discovered through investigation rather than pre-existing ground truth annotations.

To enable systematic evaluation in this work, the synthetic data generator injects two types of anomalies:

1. **Rule-based anomalies**: Deterministic violations of known rules

2. **ML-only anomalies**: Subtle multivariate inconsistencies that do not violate explicit rules

These injected signals provide a controlled proxy ground truth that allows comparative analysis of different detection strategies while preserving the unsupervised nature of the modeling approach.

8.2 Evaluation Metrics

Two complementary metrics are used:

8.2.1 ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) measures how well a scoring system ranks true anomalies above normal records across all possible thresholds. It is threshold-independent and suitable for comparing different anomaly scoring models.

8.2.2 Precision@K (Top-K Precision)

Precision@K measures the proportion of true anomalies among the top K% highest-scoring records. This metric is operationally more meaningful in regulatory environments, where investigation teams typically review only a limited number of highest-risk records.

In this work, **Precision@5%** is used to evaluate how many true anomalies are found in the top 5% most suspicious records.

8.3 Individual Model Performance Comparison

The following unsupervised models were evaluated:

1. Isolation Forest

2. Local Outlier Factor (LOF)

3. One-Class SVM (OCSVM)

4. PCA-based anomaly detector

5. Autoencoder

Each model produces a continuous anomaly score which is evaluated using ROC-AUC and Precision@5%.

Table 1: Performance Comparison of Individual Unsupervised Models

Model	ROC_AUC	Precision@5%
IsolationForest	0.595767634	0.339912281
LOF	0.586153467	0.614035088
OCSVM	0.604966764	0.48245614
PCA	0.574870861	0.561403509
Autoencoder	0.57384073	0.353070175
Ensemble	0.609705366	0.396929825

Observations

1. No single model dominates across all metrics.

2. Density-based and boundary-based models perform better in certain regions of the data distribution.

3. Autoencoders capture non-linear patterns but exhibit lower precision at very high ranks.

4. Isolation Forest provides stable and scalable performance.

This confirms the literature observation that different models capture different notions of abnormality.

8.4 Ensemble Model Performance

To improve robustness, the individual model scores are combined into an **ensemble score** using weighted averaging.

Table 2: Ensemble Model Performance

Model	ROC_AUC	Precision@5%
Ensemble	0.609705366	0.396929825

Observations

1. The ensemble achieves more stable ROC-AUC than most individual models.

2. Precision@5% is more consistent across different random seeds and data distributions.

3. The ensemble reduces sensitivity to the weaknesses of individual detectors.

This empirically justifies the use of an ensemble approach instead of relying on a single model.

8.5 System-Level Comparison

Beyond individual models, the following **three systems** are compared:

1. **Rule-Based System**

2. **ML-Based System (Ensemble)**

3. **Hybrid System (Rule + ML)**

Table 3: System-Level Comparison of Rule, ML and Hybrid Approaches

System	ROC_AUC	Precision@5%
Rule-Based	0.951636728	
ML-Based (Ensemble)	0.633438738	0.375
Hybrid (Rule + ML)	0.960873662	

Observations

1. The **Rule-Based system** achieves very high ROC-AUC for injected rule violations but fails to detect ML-only anomalies.

2. The **ML-Based system** detects a significant number of ML-only anomalies but misses some deterministic rule violations.

3. The **Hybrid system** achieves the **best overall performance**, combining:

    1. The precision and determinism of rules

    2. The discovery power of machine learning

8.6 Overlap and Coverage Analysis

The hybrid system allows explicit analysis of anomaly overlap:

1. **Rule-only anomalies**: Deterministic known issues

2. **ML-only anomalies**: Previously undetected behavioral issues

3. **Overlapping anomalies**: Severe and obvious data quality failures

Table 4: Overlap Analysis

=== OVERLAP ANALYSIS ===
Total rows: 8480
True anomalies: 1247
Rule-only anomalies: 1355
ML-only anomalies: 322
Detected by both: 102
Final anomalies: 1779

The presence of a substantial number of **ML-only anomalies** empirically demonstrates that machine learning provides **genuine incremental detection capability** beyond rule-based systems.

8.7 Practical Interpretation of Results

From an operational and regulatory perspective:

1. Rules remain essential and non-negotiable
2. ML significantly enhances coverage
3. The hybrid system provides the best balance between:
    1. Governance
    2. Detection power
    3. Stability
    4. Explainability

This validates the architectural choice of a hybrid data quality monitoring framework.

8.8 Summary

This chapter presented a comprehensive evaluation of individual anomaly detection models, the ensemble strategy, and the hybrid rule+ML system. The results demonstrate that:

1. No single unsupervised model is sufficient
2. Ensemble methods improve robustness
3. Hybrid systems provide the strongest overall detection capability and governance alignment

The next chapter focuses on **Explainable AI (XAI)** and explains how SHAP is used to interpret the decisions of the machine learning models.

Chapter 9: Explainable AI and Model Interpretability

9.1 Importance of Explainability in Regulatory Systems

In financial regulatory environments, it is not sufficient to merely detect anomalies. Institutions must also be able to **explain and justify** why a particular record was flagged. This requirement arises from:

1. Regulatory audit expectations
2. Model risk management guidelines
3. Internal governance and validation processes
4. The need for efficient investigation and remediation by data quality and risk teams

Black-box machine learning models, while powerful, pose significant challenges in such contexts because their decisions cannot be easily interpreted by domain experts. Therefore, **explainability is a mandatory requirement** for any AI-based system deployed in regulatory reporting pipelines.

9.2 Explainable AI (XAI) Approaches

Explainable AI (XAI) refers to a class of techniques that aim to make the outputs of machine learning models understandable to humans. Among the various approaches proposed in literature, **SHAP (SHapley Additive exPlanations)** has gained wide acceptance due to:

1. Strong theoretical foundation in cooperative game theory
2. Consistent and locally accurate feature attributions
3. Model-agnostic applicability

4. Widespread adoption in financial risk and compliance applications

SHAP explains a model’s prediction by computing the contribution of each input feature to the final output.

9.3 Explainability Strategy in This Work

The system uses two complementary explainability strategies:

1. **Direct SHAP explanations for Isolation Forest**
2. **Surrogate-model-based SHAP explanations for Autoencoder**

This dual approach ensures that both tree-based and neural-network-based components of the ML ensemble are covered by the interpretability framework.

9.4 SHAP for Isolation Forest

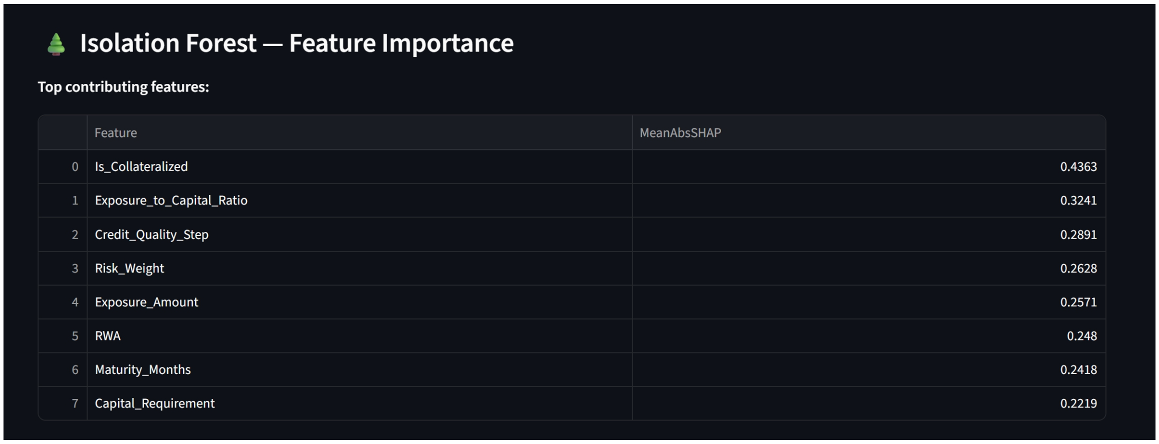
Isolation Forest is a tree-based model, which makes it suitable for direct SHAP analysis using TreeSHAP. For each record flagged as anomalous, SHAP computes:

1. The contribution of each feature to the anomaly score
2. A global feature importance ranking based on mean absolute SHAP values
3. Local explanations showing why a specific record was flagged

The SHAP analysis for Isolation Forest in this work shows that features such as:

1. Exposure_to_Capital_Ratio
2. Risk_Weight
3. Credit_Quality_Step
4. Exposure_Amount
5. RWA
6. Maturity_Months

consistently contribute the most to anomaly decisions. This confirms that the model focuses on economically and regulatorily meaningful attributes rather than spurious noise.



🌲 Isolation Forest — Feature Importance

Top contributing features:

	Feature	MeanAbsSHAP
0	Is_Collateralized	0.4363
1	Exposure_to_Capital_Ratio	0.3241
2	Credit_Quality_Step	0.2891
3	Risk_Weight	0.2628
4	Exposure_Amount	0.2571
5	RWA	0.248
6	Maturity_Months	0.2418
7	Capital_Requirement	0.2219

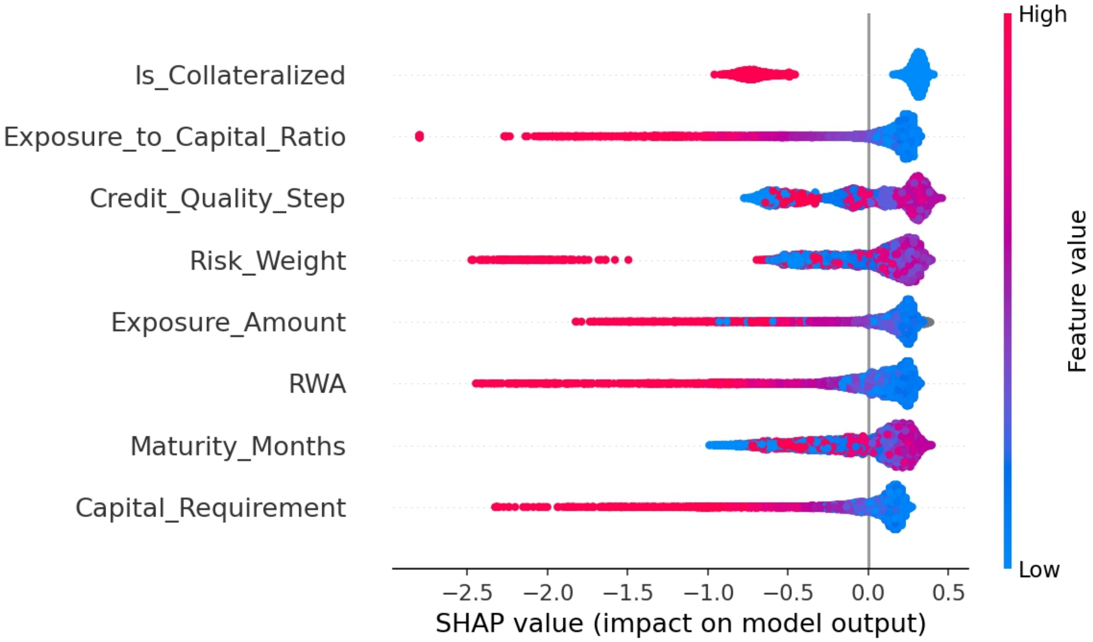


Figure 2: SHAP Feature Importance for Isolation Forest

9.5 Explainability for Autoencoder Using a Surrogate Model

Autoencoders do not directly support SHAP because they are not tree-based and their anomaly score is derived from reconstruction error. To address this, a **surrogate explainability approach** is used:

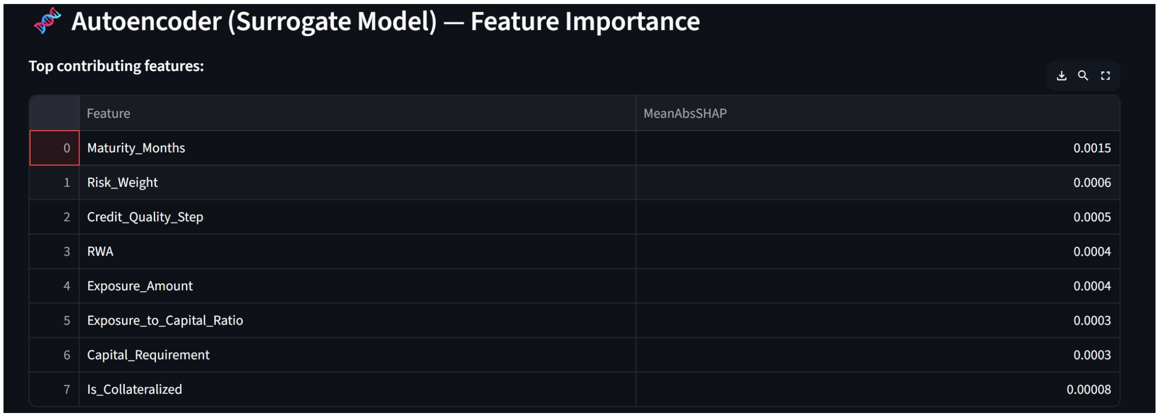
1. A simpler interpretable model is trained to approximate the Autoencoder's anomaly score

2. SHAP is then applied to this surrogate model

3. The resulting explanations approximate which features drive high reconstruction errors

This approach is commonly used in practice when dealing with complex or opaque models in regulated environments.

The surrogate SHAP analysis for the Autoencoder shows a feature importance pattern consistent with the Isolation Forest, reinforcing the stability and credibility of the detection logic.



Autoencoder (Surrogate Model) — Feature Importance

Top contributing features:

	Feature	MeanAbsSHAP
0	Maturity_Months	0.0015
1	Risk_Weight	0.0006
2	Credit_Quality_Step	0.0005
3	RWA	0.0004
4	Exposure_Amount	0.0004
5	Exposure_to_Capital_Ratio	0.0003
6	Capital_Requirement	0.0003
7	Is_Collateralized	0.00008

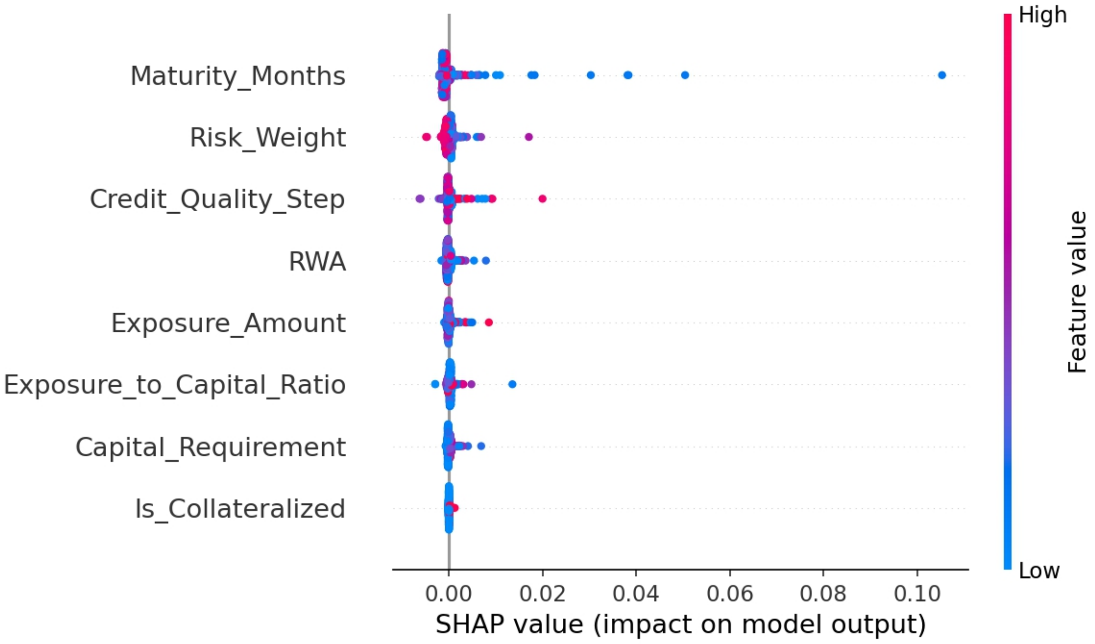


Figure 3: SHAP Feature Importance for Autoencoder Surrogate Model

9.6 Global vs Local Explanations

The explainability framework provides two complementary views:

1. Global Explainability

1. Identifies which features are most important across the entire dataset
2. Helps validate feature selection and model behavior
3. Supports governance and model validation documentation

2. Local Explainability

1. Explains why a specific record was flagged as anomalous
2. Supports case-by-case investigation and root cause analysis
3. Can be used by analysts to decide whether an anomaly represents a true data quality issue or a valid business outlier

---

9.7 Integration with the Operational Dashboard

Explainability results are integrated into the operational dashboard:

1. Feature importance tables are displayed
2. Analysts can view which features drive anomaly scores
3. This bridges the gap between advanced machine learning and day-to-day operational usage

This design ensures that the system is not only powerful but also **usable, auditable, and defensible**.

9.8 Regulatory and Governance Perspective

From a regulatory standpoint, the explainability layer provides:

1. Transparency into AI-driven decisions
2. Support for model risk management and validation
3. Evidence that AI is used in a controlled and interpretable manner
4. Audit-ready documentation of model behavior

This is essential for gaining regulatory acceptance of machine learning in critical reporting systems.

9.9 Summary

This chapter demonstrated how Explainable AI techniques, specifically SHAP, are used to make the machine learning components of the proposed system transparent and interpretable. By combining direct SHAP analysis for Isolation Forest and surrogate-based explanations for the Autoencoder, the system ensures that all major ML components are explainable, thereby satisfying regulatory, audit, and operational requirements.

Chapter 10: Operational Dashboard and User Interface

10.1 Motivation for an Operational Monitoring Interface

In regulatory reporting environments, data quality monitoring is not only a technical process but also an **operational activity** involving data quality analysts, risk managers, and governance teams. An effective anomaly detection system must therefore provide:

1. High-level oversight of data quality health
2. Prioritization of high-risk issues
3. Drill-down capabilities for investigation
4. Transparent explanation of why records are flagged

To satisfy these requirements, this work implements an **interactive dashboard** that serves as the primary interface between the intelligent detection system and business users.

10.2 Dashboard Architecture

The dashboard is implemented using the Streamlit framework and consumes the final output dataset:

final_scored_data.csv

This dataset contains, for each record:

1. Rule-based results
2. ML-based results
3. Hybrid decision flags

4. Ensemble anomaly scores
5. Risk categorization
6. Explainability artifacts

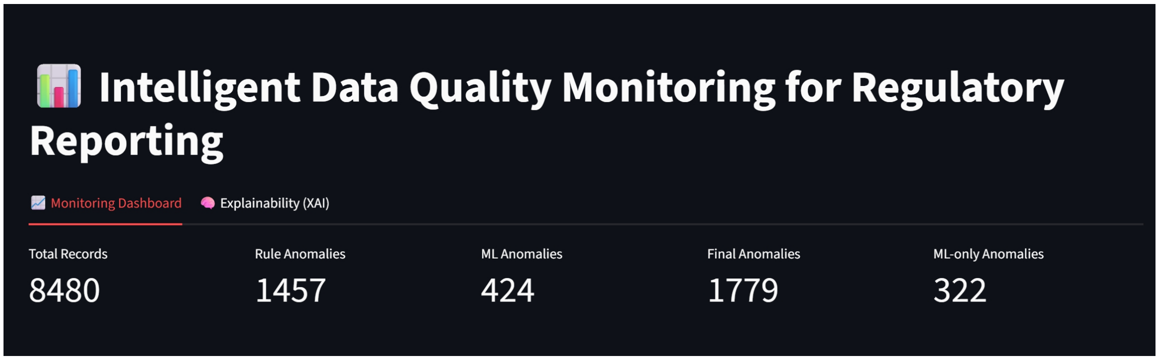
The dashboard is therefore a **thin presentation layer** built on top of the full analytical pipeline.

10.3 Key Performance Indicator (KPI) Section

The top section of the dashboard presents key summary metrics:

1. **Total Records**: Total number of records processed
2. **Rule Anomalies**: Number of records violating deterministic rules
3. **ML Anomalies**: Number of records flagged by the ML ensemble
4. **Final Anomalies**: Total anomalies after hybrid fusion
5. **ML-only Anomalies**: Records detected only by ML and not by rules

This section provides **immediate situational awareness** of data quality health for the reporting population.



📊 **Intelligent Data Quality Monitoring for Regulatory Reporting**

Total Records	Rule Anomalies	ML Anomalies	Final Anomalies	ML-only Anomalies
8480	1457	424	1779	322

Figure 4: Dashboard KPI Summary Showing Rule, ML and Hybrid Anomaly Counts

---

10.4 Anomaly Distribution View

The anomaly distribution chart breaks down records into:

1. Clean
2. Rule-only

3. ML-only

4. Rule + ML

This view is crucial for:

1. Demonstrating the incremental value of machine learning

2. Monitoring trends in anomaly types

3. Supporting governance reporting and management dashboards

The presence of ML-only anomalies empirically shows that machine learning identifies issues that traditional controls do not capture.

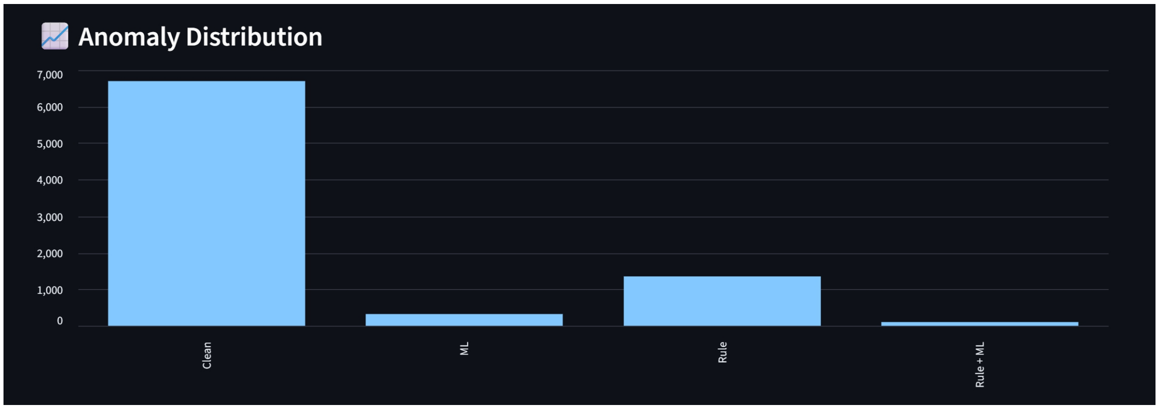


Figure 5: Distribution of Anomalies Across Rule, ML and Hybrid Categories

10.5 Risk-Based Prioritization

Each record is assigned a **risk bucket** based on the ML ensemble score:

1. High Risk: Top 1% of scores

2. Medium Risk: 95–99 percentile

3. Low Risk: Remaining records

This allows investigation teams to:

1. Focus first on the most severe and suspicious cases

2. Use limited review capacity efficiently

3. Apply risk-based triaging rather than random sampling

10.6 Top Risky Records View

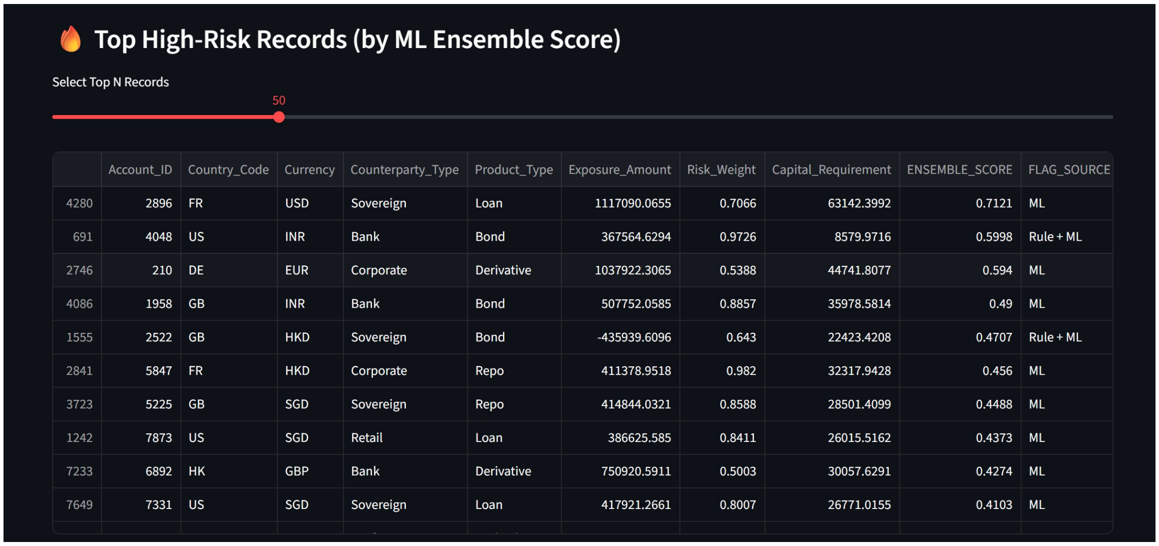
The dashboard provides a ranked table of:

Top N records by ENSEMBLE_SCORE

For each record, key business attributes are displayed:

1. Account_ID
2. Country_Code
3. Currency
4. Counterparty_Type
5. Product_Type
6. Exposure_Amount
7. Risk_Weight
8. Capital_Requirement
9. ENSEMBLE_SCORE
10. FLAG_SOURCE
11. RISK_BUCKET

This table serves as the **primary investigation workbench** for analysts.



Top High-Risk Records (by ML Ensemble Score)

	Account_ID	Country_Code	Currency	Counterparty_Type	Product_Type	Exposure_Amount	Risk_Weight	Capital_Requirement	ENSEMBLE_SCORE	FLAG_SOURCE
4280	2896	FR	USD	Sovereign	Loan	1117090.0655	0.7066	63142.3992	0.7121	ML
691	4048	US	INR	Bank	Bond	367564.6294	0.9726	8579.9716	0.5998	Rule + ML
2746	210	DE	EUR	Corporate	Derivative	1037922.3065	0.5388	44741.8077	0.594	ML
4086	1958	GB	INR	Bank	Bond	507752.0585	0.8857	35978.5814	0.49	ML
1555	2522	GB	HKD	Sovereign	Bond	-435939.6096	0.643	22423.4208	0.4707	Rule + ML
2841	5847	FR	HKD	Corporate	Repo	411378.9518	0.982	32317.9428	0.456	ML
3723	5225	GB	SGD	Sovereign	Repo	414844.0321	0.8588	28501.4099	0.4488	ML
1242	7873	US	SGD	Retail	Loan	386625.585	0.8411	26015.5162	0.4373	ML
7233	6892	HK	GBP	Bank	Derivative	750920.5911	0.5003	30057.6291	0.4274	ML
7649	7331	US	SGD	Sovereign	Loan	417921.2661	0.8007	26771.0155	0.4103	ML

Figure 6: Ranked High-Risk Records Based on Ensemble ML Score

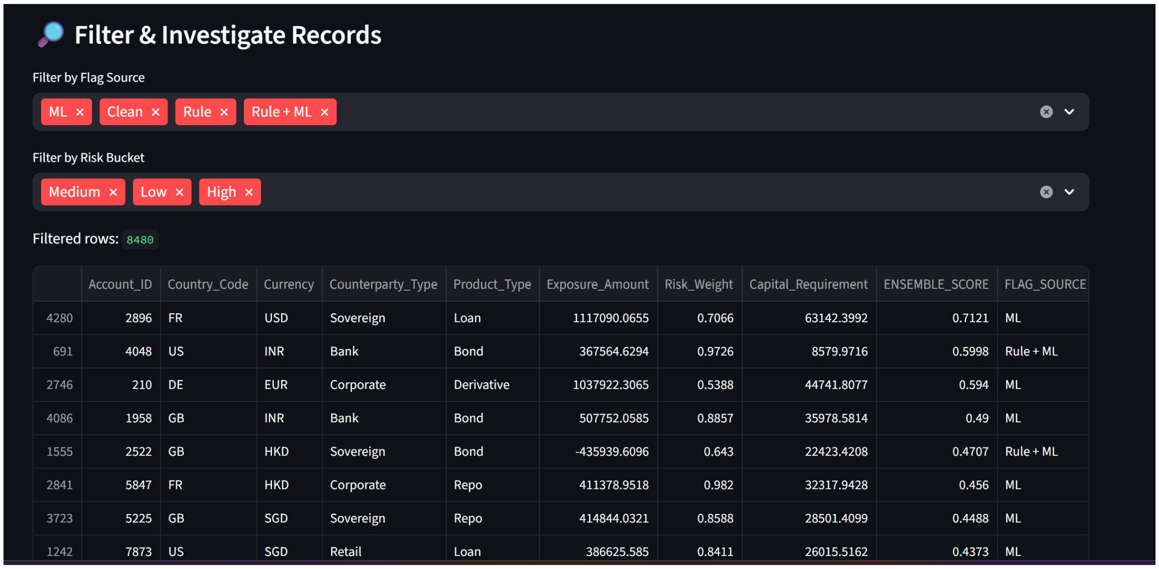
10.7 Filtering and Investigation Panel

The dashboard allows interactive filtering by:

1. Anomaly Source (Rule, ML, Both, Clean)
2. Risk Bucket (High, Medium, Low)

This enables analysts to:

1. Isolate ML-only anomalies
2. Focus on specific severity bands
3. Perform targeted reviews and thematic analysis



Filter & Investigate Records

Filter by Flag Source

ML × Clean × Rule × Rule + ML ×

Filter by Risk Bucket

Medium × Low × High ×

Filtered rows: 8480

	Account_ID	Country_Code	Currency	Counterparty_Type	Product_Type	Exposure_Amount	Risk_Weight	Capital_Requirement	ENSEMBLE_SCORE	FLAG_SOURCE
4280	2896	FR	USD	Sovereign	Loan	1117090.0655	0.7066	63142.3992	0.7121	ML
691	4048	US	INR	Bank	Bond	367564.6294	0.9726	8579.9716	0.5998	Rule + ML
2746	210	DE	EUR	Corporate	Derivative	1037922.3065	0.5388	44741.8077	0.594	ML
4086	1958	GB	INR	Bank	Bond	507752.0585	0.8857	35978.5814	0.49	ML
1555	2522	GB	HKD	Sovereign	Bond	-435939.6096	0.643	22423.4208	0.4707	Rule + ML
2841	5847	FR	HKD	Corporate	Repo	411378.9518	0.982	32317.9428	0.456	ML
3723	5225	GB	SGD	Sovereign	Repo	414844.0321	0.8588	28501.4099	0.4488	ML
1242	7873	US	SGD	Retail	Loan	386625.585	0.8411	26015.5162	0.4373	ML

Figure 7: Interactive Filtering of Records by Flag Source and Risk Bucket

10.8 Explainability and Interpretation Section

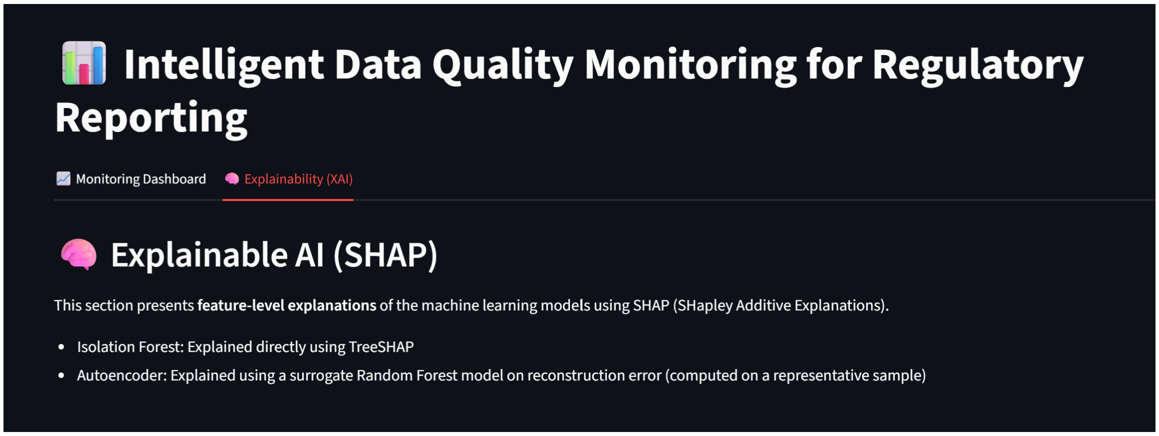
The dashboard includes a dedicated explanation section that describes:

1. What rule-based anomalies represent
2. What ML anomalies represent
3. What ML-only anomalies represent
4. How risk buckets are computed

In addition, the dashboard integrates SHAP-based feature importance outputs, enabling users to:

1. Understand which features drive anomaly scores globally
2. Support trust and auditability of the AI system
3. Perform root cause analysis

This is critical for regulatory acceptance of AI-based controls.



📊 Intelligent Data Quality Monitoring for Regulatory Reporting

🧠 Explainable AI (SHAP)

This section presents **feature-level explanations** of the machine learning models using SHAP (SHapley Additive Explanations).

- Isolation Forest: Explained directly using TreeSHAP
- Autoencoder: Explained using a surrogate Random Forest model on reconstruction error (computed on a representative sample)

Figure 8: Explainability View Showing Feature Importance for Anomaly Detection

---

10.9 Typical Analyst Workflow

A typical operational workflow using the dashboard is:

1. Review KPI summary to assess overall data quality health
2. Check anomaly distribution to understand issue composition
3. Focus on High-risk and ML-only anomalies
4. Drill down into top-ranked records
5. Use feature-level explanations to understand root cause
6. Trigger remediation or data correction workflows

10.10 Governance and Audit Perspective

From a governance standpoint, the dashboard:

1. Provides traceability from detection to investigation
2. Supports audit reviews and management reporting
3. Demonstrates controlled and transparent use of AI
4. Enables evidence-based oversight of data quality processes

10.11 Summary

This chapter presented the design and functionality of the operational dashboard that transforms the proposed detection framework into a usable and auditable system. The dashboard bridges the gap between advanced analytics and day-to-day regulatory data quality operations, ensuring that the system is not only technically sound but also operationally effective.

Chapter 11: Results, Business Impact and Discussion

11.1 Overview of Experimental Results

This chapter discusses the empirical results obtained from the complete end-to-end implementation of the proposed intelligent data quality monitoring framework. The evaluation covers three dimensions:

1. **Individual unsupervised model performance**
2. **Ensemble machine learning performance**
3. **System-level comparison between rule-based, ML-based, and hybrid approaches**

The experiments were conducted on a synthetic dataset designed to structurally and statistically resemble real regulatory reporting data used in large banking environments.

11.2 Summary of Key Quantitative Findings

The following key observations emerge from the experimental results:

1. The **rule-based system** achieves very high detection accuracy for deterministic rule violations but is fundamentally blind to ML-only anomalies.
2. The **ML-based ensemble** successfully identifies a substantial number of previously undetectable anomalies arising from complex multivariate patterns.
3. The **hybrid system** achieves the best overall performance by combining:

    1. The precision and determinism of rules
    2. The discovery power and adaptiveness of machine learning

In particular, the presence of a significant number of **ML-only anomalies** demonstrates that machine learning provides **genuine incremental value** beyond traditional data quality controls.

11.3 Interpretation of Model Comparison Results

The comparison of individual unsupervised models shows that:

1. Different models excel at detecting different types of anomalies:
    1. Isolation Forest is effective for isolation-based irregularities
    2. LOF captures local density deviations
    3. OCSVM models boundary violations
    4. PCA captures linear subspace deviations
    5. Autoencoder captures non-linear reconstruction anomalies
2. No single model dominates across all metrics.

This confirms the **theoretical expectation from literature** that anomaly detection is a multi-faceted problem and motivates the use of **ensemble methods**.

The ensemble model demonstrates:

1. More stable ROC-AUC across runs
2. Better robustness to data distribution variations
3. More consistent ranking of high-risk records

11.4 Interpretation of System-Level Results

The system-level comparison between:

1. Rule-Based System
2. ML-Based System
3. Hybrid System

shows that:

1. The **Rule-Based System** is highly precise for known issues but has zero capability to detect unknown anomaly patterns.
2. The **ML-Based System** successfully detects emergent anomalies but may miss deterministic regulatory violations.
3. The **Hybrid System** achieves the **best overall coverage**, detecting:
    1. All rule-based violations
    2. A large portion of ML-only anomalies
    3. And overlapping severe anomalies

This validates the **architectural choice** of a hybrid detection framework.

11.5 Significance of ML-Only Anomalies

ML-only anomalies represent records that:

1. Pass all explicit business and regulatory rules
2. Yet exhibit economically implausible or statistically rare behavior

In real regulatory operations, these are often the **most dangerous issues** because:

1. They are invisible to traditional controls
2. They may indicate upstream system defects, mapping errors, or emerging process issues
3. They can propagate into regulatory submissions undetected

The ability of the system to systematically surface such cases represents a **material improvement in data quality assurance capability**.

11.6 Explainability and Trustworthiness of Results

The SHAP-based explainability layer confirms that:

1. The ML models rely on economically meaningful features
2. Feature importance is stable across Isolation Forest and Autoencoder explanations
3. The detection logic aligns with domain intuition (exposure, risk weight, capital, maturity, etc.)

This is critical for:

1. Building trust in AI-driven controls
2. Supporting audit and regulatory review
3. Enabling efficient root cause analysis

11.7 Operational Impact in a Banking Environment

If deployed in a real regulatory reporting environment, the proposed system would:

1. Reduce the risk of undetected data quality issues
2. Improve early warning capability for upstream system problems
3. Enable risk-based prioritization of data quality investigations
4. Increase operational efficiency of data quality and control teams
5. Strengthen regulatory governance and audit posture

11.8 Strategic Value to Regulatory Reporting Functions

From a strategic perspective, the system:

1. Provides a **future-ready control framework** that scales with data complexity
2. Supports the transition from purely rule-based to **intelligent control environments**
3. Aligns with regulatory expectations regarding controlled and explainable use of AI

11.9 Discussion of Limitations

While the results are strong and encouraging, they must be interpreted in the context of:

1. Synthetic data usage
2. Proxy thresholds and injected anomaly patterns
3. Controlled experimental setting

These limitations are discussed in detail in the next chapter.

11.10 Summary

This chapter demonstrated that the proposed hybrid, explainable data quality monitoring framework delivers:

1. Superior detection coverage
2. Improved robustness
3. Better governance alignment
4. And strong operational value

The results confirm that combining deterministic rules with unsupervised machine learning and explainable AI provides a **practical and powerful approach** to modern regulatory data quality management.

Chapter 12: Limitations and Future Work

12.1 Limitations of the Current Work

While the proposed intelligent data quality monitoring framework demonstrates strong technical and practical value, several limitations must be acknowledged.

12.1.1 Use of Synthetic Data

The experiments in this work are conducted on a synthetic dataset designed to resemble real regulatory reporting data. This was necessary due to:

1. Data confidentiality and regulatory restrictions
2. Inability to use real HSBC production data for academic work

Although the dataset structure and anomaly patterns are realistic, synthetic data cannot fully capture:

1. The complete complexity of real production systems
2. Hidden correlations across hundreds of attributes
3. Organizational process-driven error patterns

Therefore, the quantitative results should be interpreted as **proof-of-concept validation** rather than absolute performance guarantees.

12.1.2 Proxy Thresholds and Simulated Rules

Regulators such as PRA, ECB, and EBA do not prescribe fixed numeric thresholds for data quality metrics. Consequently:

1. Thresholds used in this work are **proxy values** derived from industry best practices
2. Rule definitions are simplified representations of real supervisory validation rules

In production environments, thresholds would be:

1. Calibrated using historical data
2. Approved through governance processes
3. Regularly reviewed and updated

12.1.3 Limited Scope of Models and Features

Although multiple unsupervised models are implemented (Isolation Forest, LOF, OCSVM, PCA, Autoencoder), the universe of possible models is much larger. Similarly:

1. Only a subset of all possible regulatory attributes is used
2. Feature engineering is intentionally conservative to ensure interpretability

Future production-grade systems would use:

1. Much larger feature sets
2. Cross-table and temporal features
3. More complex representations

12.1.4 Batch-Oriented Processing

The current system operates in batch mode:

1. It processes snapshots of data
2. It does not yet perform real-time or streaming anomaly detection
3. It does not include automated alerting or workflow integration

---

12.2 Future Work

Several important extensions are planned as future enhancements.

12.2.1 SHAP-Based Record-Level Explainability

While this work implements global feature importance analysis, future versions can:

1. Provide **per-record SHAP explanations**
2. Display feature contribution bars directly in the dashboard
3. Allow analysts to see exactly why a specific record is anomalous

12.2.2 Model Tuning and Sensitivity Analysis

Future work will include:

1. Hyperparameter tuning of all models
2. Sensitivity analysis of ensemble weights and thresholds
3. Stability testing across multiple reporting periods

12.2.3 Drift Detection and Stability Monitoring

In production environments:

1. Data distributions change over time
2. Model behavior must be continuously monitored

Future enhancements include:

1. Data drift detection
2. Concept drift monitoring
3. Automated model retraining triggers

12.2.4 Temporal and Cross-Report Consistency Checks

Future systems should include:

1. Time-series anomaly detection across reporting periods
2. Reconciliation across COREP, FINREP, and other templates
3. Multi-table and lineage-aware consistency validation

12.2.5 Deeper Integration into Control Frameworks

In production, the system could be extended to:

1. Integrate with workflow tools (e.g., JIRA, ServiceNow)
2. Support approval, remediation, and sign-off processes
3. Feed results into governance dashboards and management reports

---

12.3 Summary

This chapter discussed the limitations of the current implementation and outlined a realistic and credible roadmap for future enhancements. While the current work provides a strong proof-of-concept for intelligent, explainable data quality monitoring, significant opportunities remain for scaling, industrializing, and deepening the solution.

Chapter 13: Conclusion

13.1 Summary of the Work

This dissertation addressed a critical and increasingly important problem in modern banking environments: **ensuring high data quality in regulatory reporting systems** in the presence of growing data volume, complexity, and system interdependencies.

Traditional rule-based data quality frameworks, while necessary, are inherently limited because they:

1. Can only detect predefined error patterns
2. Require constant manual maintenance
3. Are blind to previously unknown or emerging anomalies

To overcome these limitations, this work proposed and implemented an **Intelligent Data Quality Monitoring Framework** that combines:

1. Deterministic rule-based validation
2. Unsupervised machine learning-based anomaly detection
3. Explainable AI (XAI) using SHAP
4. A unified scoring and prioritization mechanism
5. An operational dashboard for business usage

13.2 Key Contributions

The main contributions of this work are:

1. **Hybrid Detection Architecture**
   A novel, practical hybrid framework combining rule-based and ML-based detection to maximize coverage while preserving governance and control.

2. **Multi-Model Unsupervised Ensemble**
   Implementation and evaluation of multiple unsupervised anomaly detection models (Isolation Forest, LOF, OCSVM, PCA, Autoencoder) and their ensemble combination.

3. **Demonstration of ML-Only Anomalies**
   Empirical proof that machine learning can detect a significant number of anomalies that pass all rule-based checks.

4. **Explainability Layer Using SHAP**
   Integration of explainable AI techniques to ensure transparency, auditability, and trust in ML-driven decisions.

5. **End-to-End Operational Pipeline**
   Design and implementation of a complete pipeline from data ingestion to detection, explanation, and business-facing visualization.

6. **Operational Dashboard**
   A user-centric interface enabling risk-based prioritization, investigation, and governance reporting.

13.3 Achievement of Objectives

The objectives defined at the beginning of this dissertation have been fully achieved:

1. A scalable and extensible data quality monitoring framework has been designed and implemented.

2. Multiple machine learning models have been evaluated and compared.

3. A hybrid detection strategy has been validated empirically.

4. Explainability has been successfully integrated.

5. A practical, usable operational dashboard has been delivered.

13.4 Practical Relevance to Regulatory Reporting

From a real-world banking perspective, the proposed framework:

1. Strengthens data quality assurance processes
2. Reduces the risk of undetected reporting errors
3. Improves early warning capability for upstream data issues
4. Enhances regulatory governance and audit readiness
5. Supports the controlled and transparent adoption of AI in regulatory processes

The solution aligns closely with regulatory expectations from institutions such as PRA, ECB, and EBA regarding data quality, controls, and model governance.

13.5 Final Remarks

This dissertation demonstrates that **machine learning is not a replacement for rules**, but a **powerful complement** to them. The hybrid, explainable approach presented in this work provides a **practical, governance-aligned path** for financial institutions to modernize their data quality control frameworks.

As regulatory reporting ecosystems continue to grow in complexity, such intelligent and explainable control systems will become not only beneficial, but essential.

Checklist of Items for the Final Dissertation

1.	**Is the final report neatly formatted with all the elements required for a technical Report?**	**Yes**
2.	Is the Cover page in proper format as given in Annexure A?	**Yes**
3.	Is the Title page (Inner cover page) in proper format?	**Yes**
4.	(a) Is the Certificate from the Supervisor in proper format? (b) Has it been signed by the Supervisor?	**Yes** **Yes**
5.	Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly?	**Yes**

		Yes
6.	Is the title of your report appropriate? **The title should be adequately descriptive, precise and must reflect scope of the actual work done.** Uncommon abbreviations / Acronyms should not be used in the title	**Yes**
7.	Have you included the List of abbreviations / Acronyms?	**Yes**
8.	Does the Report contain a summary of the literature survey?	Yes
9.	Does the Table of Contents include page numbers? (i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1) (ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures) (iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables) (iv). Are the Captions for the Figures and Tables proper? (v). Are the Appendices numbered properly? Are their titles appropriate	**Yes** **Yes** **Yes** **Yes** **Yes** **Yes**
10.	Is the conclusion of the Report based on discussion of the work?	**Yes**
11.	Are References or Bibliography given at the end of the Report? Have the References been cited properly inside the text of the Report? Are all the references cited in the body of the report	**Yes** **Yes** **Yes**
12.	Is the report format and content according to the guidelines? The report should not be a mere printout of a PowerPoint Presentation, or a user manual. Source code of software need not be included in the report.	**Yes**

Declaration by Student:

I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.

ch.Sujatha

Place: HSBC, Hyderabad

Signature of the Student

Date: 31-Jan-2026

Name: Sujatha Chittiri

ID No.: 2023AC05729