INTELLIGENT DATA QUALITY MONITORING USING MACHINE LEARNING FOR REGULATORY REPORTING SYSTEMS

AIMLCZG628T – Dissertation (Mid-Semester Report)

Student Name: Sujatha Chittiri
BITS ID: 2023AC05729
Programme: M.Tech Artificial Intelligence & Machine Learning
Organisation: HSBC Global Technologies Pvt Ltd, Hyderabad
Research Area: Financial Risk Analytics

ABSTRACT

Regulatory reporting in large financial institutions relies on complex data pipelines that integrate information from multiple heterogeneous source systems. Ensuring high data quality in such environments is critical, as inaccurate or inconsistent data can lead to regulatory breaches, supervisory observations, and reputational risk. Traditional rule-based validation mechanisms, while necessary, are often insufficient to detect subtle, non-obvious anomalies in large-scale datasets.

This dissertation proposes an intelligent data quality monitoring framework that combines deterministic data quality rules with machine learning–based anomaly detection techniques. The solution integrates Isolation Forest and Autoencoder models to identify hidden data quality issues, while maintaining explainability and regulatory defensibility. A unified scoring mechanism consolidates rule-based violations and ML-based anomaly scores, and the results are visualised through an interactive dashboard for operational monitoring. The work demonstrates how artificial intelligence can enhance data quality assurance in regulatory reporting systems while aligning with supervisory expectations for transparency and auditability.

1. BROAD AREA OF WORK

The project lies at the intersection of Artificial Intelligence, Machine Learning, and Financial Risk Analytics. It focuses on applying intelligent techniques to improve data quality monitoring in regulatory reporting processes for supervisory authorities such as the PRA, ECB, and EBA.

The work covers multiple AIML dimensions including:
• Rule-based data validation and profiling

• Unsupervised machine learning for anomaly detection
• Explainable AI principles for regulatory compliance
• MLOps-oriented pipeline design and monitoring

2. BACKGROUND AND MOTIVATION

Regulatory reports are generated from data aggregated across core banking, risk, treasury, and finance systems. Due to differences in data definitions, transformation logic, and operational controls, data quality issues such as missing values, invalid codes, duplicate records, and extreme outliers frequently arise.

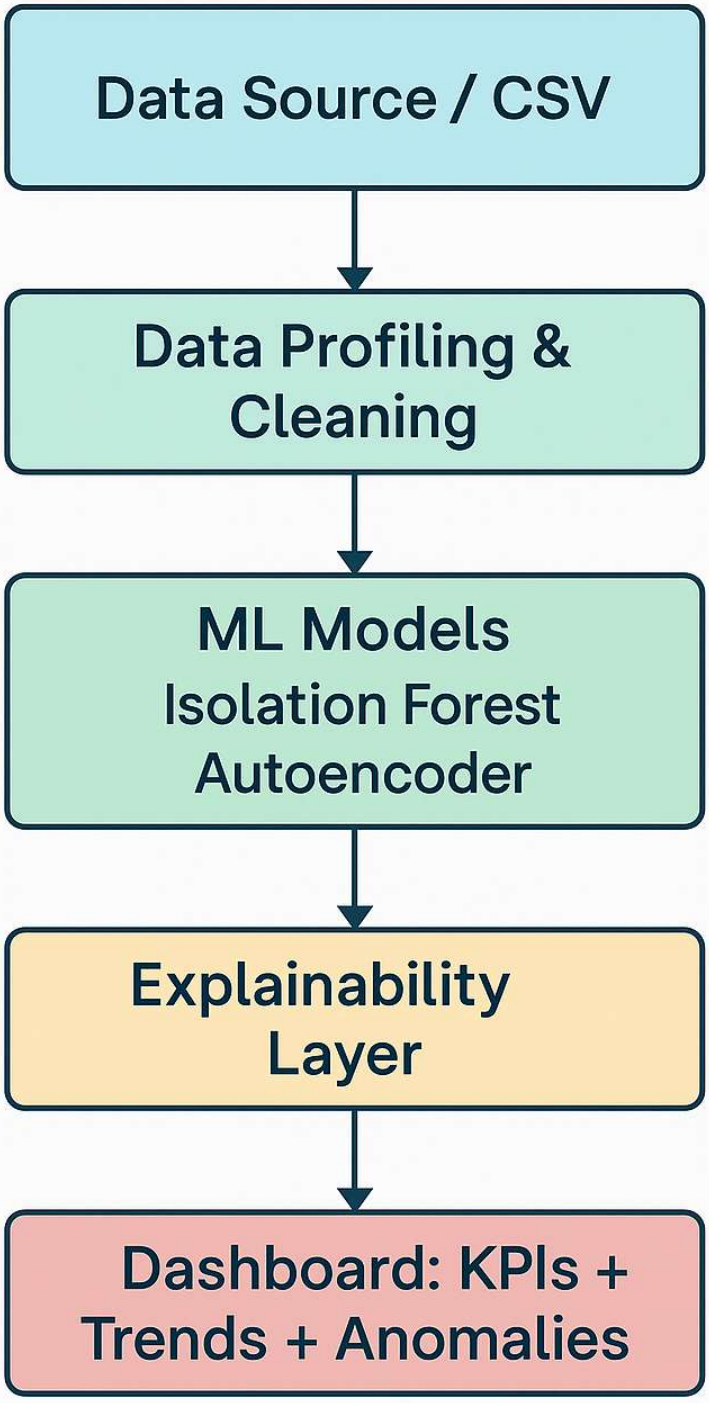
Supervisory bodies increasingly expect banks to demonstrate proactive data quality monitoring rather than reactive issue resolution. While rule-based checks remain essential, they are limited to known scenarios and predefined thresholds. This motivates the adoption of machine learning approaches that can learn data patterns and identify abnormal behaviour without explicit rules.

3. SYSTEM ARCHITECTURE AND PIPELINE

The proposed solution follows a modular pipeline architecture:

1. Data Ingestion: Synthetic regulatory-style datasets are generated to simulate exposure-level reporting data.
2. Rule-based Data Quality Engine: Deterministic checks identify explicit regulatory violations.
3. Machine Learning Models: Unsupervised models detect latent and non-linear anomalies.
4. Unified Anomaly Scoring: Rule and ML outputs are combined into a single decision framework.
5. Dashboard Visualisation: Results are presented for monitoring and analysis.

This layered design ensures robustness, explainability, and ease of extension.


Data Source / CSV
Data Profiling & Cleaning
ML Models
Isolation Forest
Autoencoder
Explainability Layer
Dashboard: KPIs + Trends + Anomalies

4. TECHNICAL SPECIFICATIONS OF THE SYSTEM

Technical Parameter	Specification
Dataset Size	~7,000 synthetic PRA-like records
Rule Engine	Completeness, validity, boundaries, duplicates
ML Models Compared	IF, AE, LOF, OCSVM
Selected Models	Isolation Forest + Autoencoder
Explainability	SHAP + FI
Tools	Python, sklearn, Streamlit
Deployment	Colab / Local
Outputs	Anomaly report + Dashboard

5. DATA QUALITY RULE ENGINE

The rule-based engine implements regulatory-style validations including:
- Completeness checks on mandatory reporting fields
- Validity checks for country and currency codes
- Non-negativity checks for exposure values
- Risk weight boundary checks
- Duplicate record detection
- Statistical outlier detection based on country-level exposure distributions

Each rule produces an individual flag, enabling transparent root-cause analysis.

6. MACHINE LEARNING MODEL COMPARISON

Multiple unsupervised anomaly detection algorithms were evaluated:

- Isolation Forest
- Local Outlier Factor (LOF)
- One-Class SVM
- Autoencoder

Weak labels derived from rule violations were used for comparative evaluation. Performance was assessed using ROC-AUC and relative anomaly separation metrics.

Experimental results indicated that Autoencoders achieved the highest anomaly separation capability, reflecting their ability to capture complex non-linear relationships. LOF showed moderate performance but was not selected due to scalability limitations. One-Class SVM demonstrated sensitivity to hyperparameters and higher computational cost.

Isolation Forest, while exhibiting a comparatively lower ROC-AUC, was selected due to its scalability, robustness on large tabular datasets, and compatibility with SHAP-based explainability. A hybrid approach combining Isolation Forest and Autoencoder was therefore adopted to balance interpretability and detection performance.

7. REGULATORY THRESHOLDS AND ACCEPTABLE DATA QUALITY LEVELS

Regulatory authorities do not prescribe fixed numerical thresholds for acceptable data quality defect rates. Instead, they expect institutions to define internal thresholds aligned with risk appetite and reporting materiality.

In this project, proxy thresholds were defined based on industry best practices and regulatory guidance:
- Completeness issues: < 1–2% of records
- Invalid reference data codes: < 0.5%
- Duplicate records: 0% tolerance
- Extreme outliers: Investigate top 1–5%
- Aggregate anomaly rate: < 5%

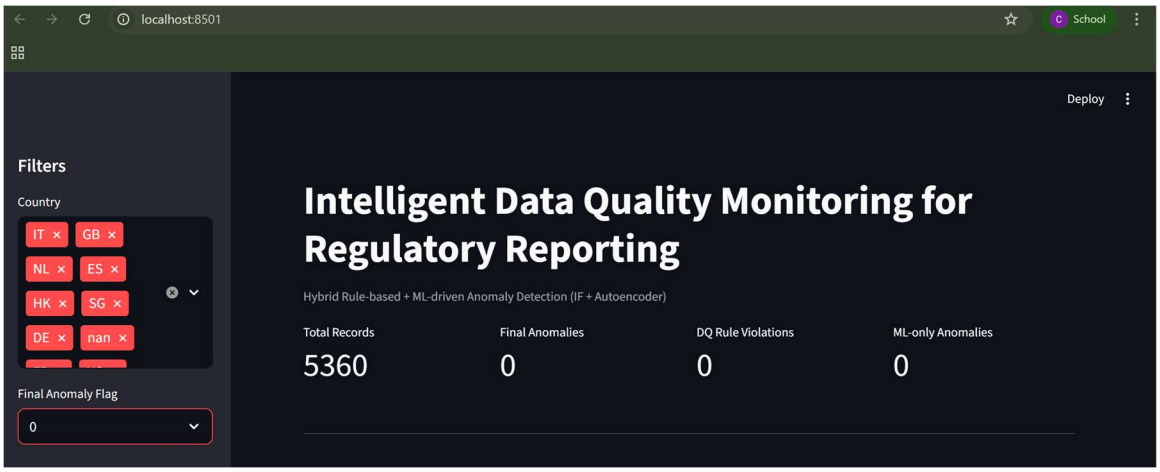
These thresholds are externalised in a configuration file, allowing transparent governance and easy updates. The approach aligns with PRA and ECB expectations for institution-owned controls rather than prescriptive limits.

8. DASHBOARD VISUALISATION

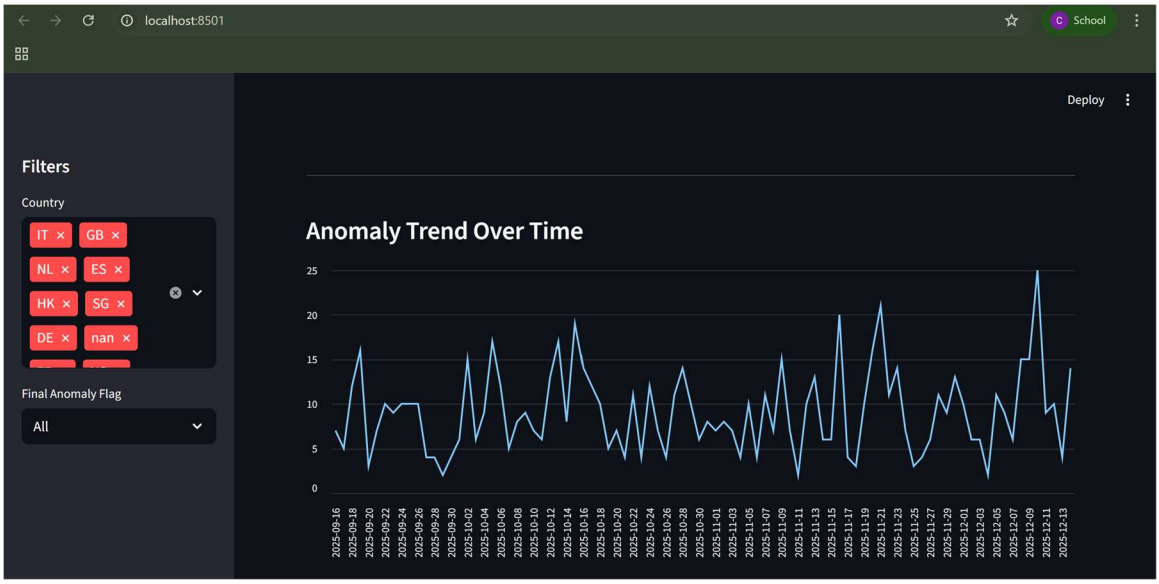
An interactive dashboard was developed using Streamlit to present:
- Key performance indicators for data quality
- Temporal trends of detected anomalies
- Severity-based prioritisation
- Drill-down views for anomalous records

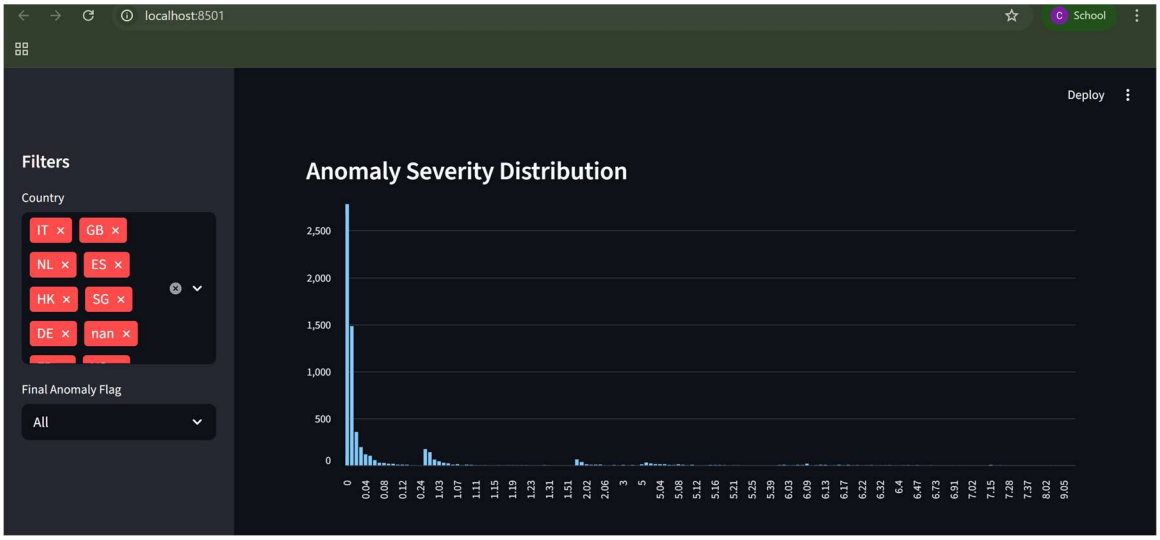
The dashboard supports operational monitoring and provides evidence of proactive data quality management. (Dashboard screenshots to be attached in the final submission.)

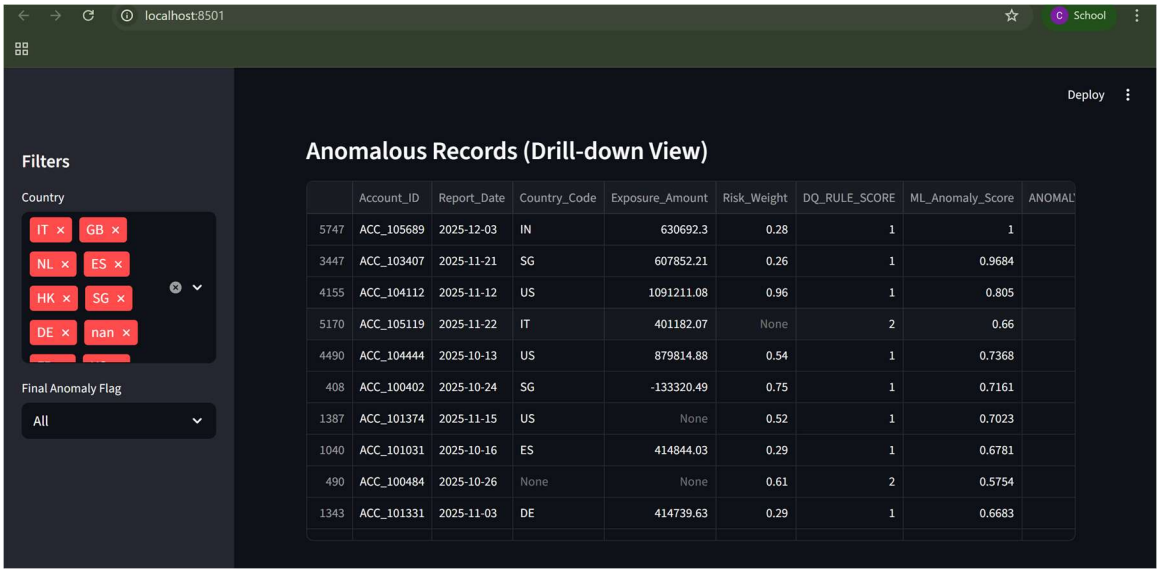

localhost:8501
School
Deploy
Filters
Country
IT × GB ×
NL × ES ×
HK × SG ×
DE × nan ×
Final Anomaly Flag
All
Intelligent Data Quality Monitoring for Regulatory Reporting
Hybrid Rule-based + ML-driven Anomaly Detection (IF + Autoencoder)
Total Records
6180
Final Anomalies
820
DQ Rule Violations
820
ML-only Anomalies
3


localhost:8501
School
Deploy
Filters
Country
IT × GB ×
NL × ES ×
HK × SG ×
DE × nan ×
Final Anomaly Flag
0
Intelligent Data Quality Monitoring for Regulatory Reporting
Hybrid Rule-based + ML-driven Anomaly Detection (IF + Autoencoder)
Total Records
5360
Final Anomalies
0
DQ Rule Violations
0
ML-only Anomalies
0

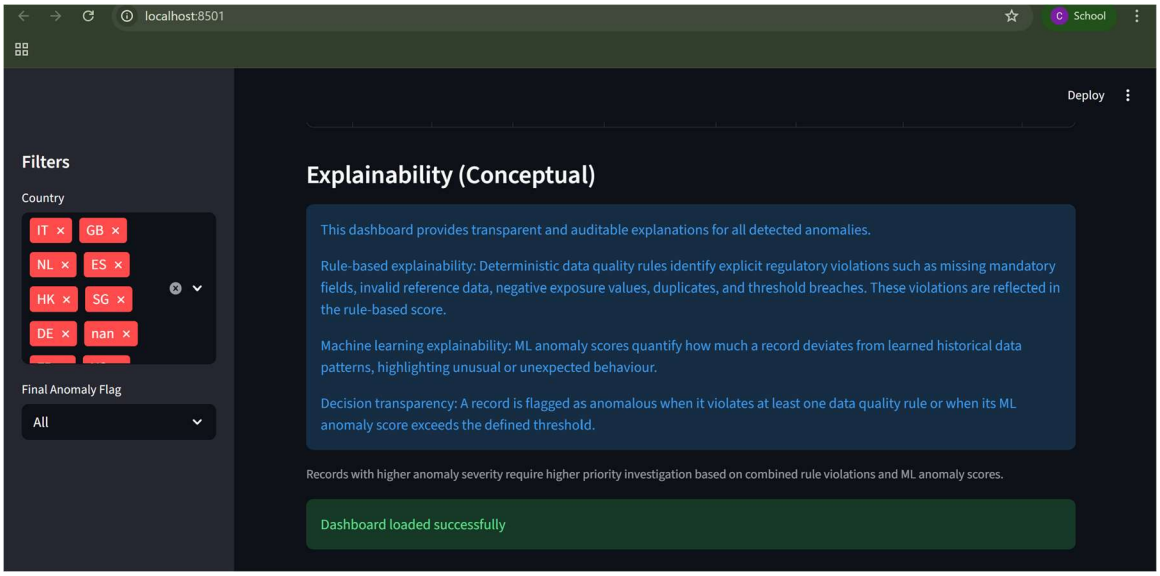

localhost:8501
School
Deploy
Filters
Country
IT × GB ×
NL × ES ×
HK × SG ×
DE × nan ×
Final Anomaly Flag
1
Intelligent Data Quality Monitoring for Regulatory Reporting
Hybrid Rule-based + ML-driven Anomaly Detection (IF + Autoencoder)
Total Records
820
Final Anomalies
820
DQ Rule Violations
820
ML-only Anomalies
3


localhost:8501
School
Deploy
Filters
Country
IT
GB
NL
ES
HK
SG
DE
nan
Final Anomaly Flag
All
Anomaly Trend Over Time
25
20
15
10
5
0
2025-09-16
2025-09-18
2025-09-20
2025-09-22
2025-09-24
2025-09-26
2025-09-28
2025-09-30
2025-10-02
2025-10-04
2025-10-06
2025-10-08
2025-10-10
2025-10-12
2025-10-14
2025-10-16
2025-10-18
2025-10-20
2025-10-22
2025-10-24
2025-10-26
2025-10-28
2025-10-30
2025-11-01
2025-11-03
2025-11-05
2025-11-07
2025-11-09
2025-11-11
2025-11-13
2025-11-15
2025-11-17
2025-11-19
2025-11-21
2025-11-23
2025-11-25
2025-11-27
2025-11-29
2025-12-01
2025-12-03
2025-12-05
2025-12-07
2025-12-09
2025-12-11
2025-12-13


localhost:8501
School
Deploy
Filters
Country
IT
GB
NL
ES
HK
SG
DE
nan
Final Anomaly Flag
All
Anomaly Severity Distribution
2,500
2,000
1,500
1,000
500
0
0
0.04
0.08
0.12
0.24
1.03
1.07
1.11
1.15
1.19
1.23
1.31
1.51
2.02
2.06
3
5
5.04
5.08
5.12
5.16
5.21
5.25
5.39
6.03
6.09
6.13
6.17
6.22
6.32
6.4
6.47
6.73
6.91
7.02
7.15
7.28
7.37
8.02
9.05



	Account_ID	Report_Date	Country_Code	Exposure_Amount	Risk_Weight	DQ_RULE_SCORE	ML_Anomaly_Score	ANOMAL
5747	ACC_105689	2025-12-03	IN	630692.3	0.28	1	1	
3447	ACC_103407	2025-11-21	SG	607852.21	0.26	1	0.9684	
4155	ACC_104112	2025-11-12	US	1091211.08	0.96	1	0.805	
5170	ACC_105119	2025-11-22	IT	401182.07	None	2	0.66	
4490	ACC_104444	2025-10-13	US	879814.88	0.54	1	0.7368	
408	ACC_100402	2025-10-24	SG	-133320.49	0.75	1	0.7161	
1387	ACC_101374	2025-11-15	US	None	0.52	1	0.7023	
1040	ACC_101031	2025-10-16	ES	414844.03	0.29	1	0.6781	
490	ACC_100484	2025-10-26	None	None	0.61	2	0.5754	
1343	ACC_101331	2025-11-03	DE	414739.63	0.29	1	0.6683	



9. ABBREVIATIONS

DQ – Data Quality

ML – Machine Learning

IF – Isolation Forest

AE – Autoencoder

LOF – Local Outlier Factor

OCSVM – One-Class SVM

PRA – Prudential Regulation Authority

ECB – European Central Bank

EBA – European Banking Authority

SHAP – Shapley Additive Explanations

10. WORK COMPLETED TILL MID-SEMESTER

The following milestones have been completed:
• Synthetic regulatory dataset generation
• Rule-based data quality engine implementation
• Machine learning model comparison and selection
• Unified anomaly scoring framework
• Prototype dashboard development
• Git-based project structure and version control

11. CHALLENGES FACED

• 1. Real Data Availability – Regulatory datasets cannot be used due to confidentiality; synthetic datasets were required.

• 2. Regulatory Threshold Definition – PRA/ECB guidelines do not specify numeric thresholds; proxy thresholds had to be designed based on best practices.

• 3. Model Explainability – Unsupervised models pose challenges for SHAP; workaround approaches were required.

• 4. Hybrid Scoring Balance – Needed careful tuning of rule-based vs ML-based anomaly contributions.

• 5. Dashboard Integration – Combining rule flags, ML scores, SHAP values, and trends required complex dataset merging.

• 6. Computational Constraints – Notebook environments such as Colab reset, requiring pipeline re-runs.

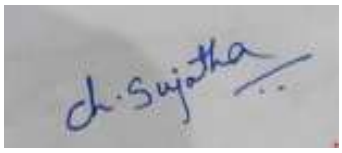
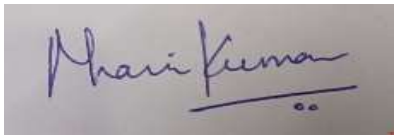
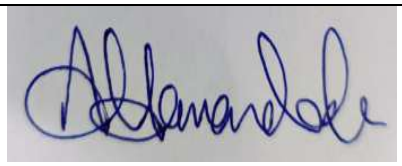
• 7. Regulatory Alignment – Ensuring explainability, auditability, and validation aligned with PRA/ECB expectations required extra design effort.

12. ROADMAP FOR FINAL SUBMISSION

The remaining work planned for the final submission includes:
• Integration of SHAP-based explainability for Isolation Forest
• Enhanced dashboard visualisations and filters
• Model tuning and sensitivity analysis
• Validation using additional datasets
• Final dissertation documentation and evaluation

13. CONCLUSION

This mid-semester work demonstrates a robust and explainable approach to intelligent data quality monitoring for regulatory reporting. By combining deterministic rules with machine learning–based anomaly detection, the proposed framework enhances detection capability while maintaining regulatory transparency. The progress achieved provides a strong foundation for the final dissertation submission.

		
Signature of Student	**Signature of Supervisor**	**Signature of Additional Examiner**
Name: Sujatha Chittiri	**Name: Phani Kumar PVSKP**	**Name: Hemendra LK Akuthota**