Birla Institute of Technology & Science, Pilani
Hyderabad Campus

Dissertation / Project / Project Work Title:

Predictive Modeling and Data Quality Assurance for EDI Order Processing in Retail Supply Chain

Course No.: S1-25_DSECLZG628T

Course Title: Dissertation

Dissertation / Project /Project Work Done by:

Student Name: Ajitabh Tiwari

BITS ID: 2020FC04613

Degree: M. Tech. (Data Science & Engineering)

Research Area: Supply Chain Management, Predictive Modeling, and Data Quality.

Dissertation / Project Work carried out at:

7-Eleven Global Solution Center, Bangalore

**BIRLA INSTITUTE OF TECHNOLOGY &
SCIENCE, PILANI VIDYA VIHAR, PILANI,
RAJASTHAN – 333031**
(DEC / 2025)

Birla Institute of Technology & Science, Pilani
Hyderabad Campus

Certificate from the Supervisor

CERTIFICATE

This is to certify that the Dissertation entitled, **Predictive Modeling and Data Quality Assurance for EDI Order Processing in Retail Supply Chain** and submitted by **Ajitabh Tiwari** ID No. **2020FC04613** in partial fulfilment of the requirement of BITS **S1-25_DSECLZG628T**.

Dissertation embodies the work done by him/her under my supervision.

Rajasekhar Kandula	Veeranjaneya Reddy
Signature of the Supervisor	**Signature of the Co-Supervisor**
Name: Raja Sekhar Kandula	**Name: Veeranjaneya Reddy**
Date: 27/ January /2026	**Date: 27/ January /2026**
Place: Bengaluru	**Place: Bengaluru**

Abstract

Electronic Data Interchange (EDI) plays a critical role in automating transactional workflows within large-scale retail supply chains. Despite mature integration platforms, inbound order failures continue to occur primarily due to data quality deficiencies rather than system-level defects. In Oracle E-Business Suite–based order management systems, such failures lead to repeated reprocessing cycles, increased manual intervention, and delayed order fulfilment.

This dissertation proposes a comprehensive framework that integrates structured data quality validation with predictive modelling to proactively assess the reliability of inbound EDI 850 purchase orders before core order creation. A composite Data Quality (DQ) score is computed using weighted validation rules executed at the interface layer of Oracle EBS. These quantitative indicators are then utilized as input features for machine learning models to predict the likelihood of order processing failure.

To ensure confidentiality and reproducibility, the proposed framework is evaluated using a synthetically generated dataset that closely resembles real-world EDI behaviour. Experimental results demonstrate a strong inverse correlation between DQ scores and failure probability. The proposed approach enables risk-based order handling, improves operational visibility through analytical dashboards, and enhances overall efficiency and reliability in retail supply chain operations.

Keywords: EDI, Data Quality, Predictive Modeling, Supply Chain, Oracle EBS

Ajitabh Tiwari

Rajasekhar Kandula

Signature of the Student
Name: Ajitabh Tiwari
Date: 27/January/2026
Place: Bengaluru

Signature of the Supervisor
Name: Raja Sekhar Kandula
Date: 27/January/2026
Place: Bengaluru

Acknowledgement

I would like to express my sincere gratitude to my industry supervisor **Mr. Raja Sekhar Kandula** for his continuous guidance, valuable insights, and encouragement throughout the course of this dissertation.

His deep understanding of enterprise integration systems and supply chain operations greatly contributed to the successful completion of this work.

I also extend my heartfelt thanks to the management and leadership team at **7-Eleven Global Solution Centre, Bengaluru**, for providing a professional environment and necessary resources to carry out this project.

I am grateful to my faculty mentor from **BITS Pilani, WILP Division,** for academic guidance and periodic reviews.
I would also like to thank my colleagues, friends, and family for their constant support and motivation.

Table of Contents

1. INTRODUCTION

1.1 Problem Statement

- Retail supply chains increasingly depend on **Electronic Data Interchange (EDI)** to automate the exchange of purchase orders and reduce manual processing between trading partners.

- In **Oracle E-Business Suite (EBS)**–based order management environments, inbound EDI purchase orders (EDI 850) are staged and processed through predefined interface tables and validation programs before sales order creation.

- Despite automation, a substantial proportion of inbound EDI orders fail during processing due to **data quality issues**, including missing mandatory attributes, incorrect data formats, invalid reference data, and non-compliance with partner-specific business rules.

- Such data quality issues result in **order rejections, repeated reprocessing, and manual intervention**, leading to delays in order fulfilment and increased operational workload for support teams.

- Existing validation mechanisms in Oracle EBS are primarily **rule-based and reactive**, identifying errors only after they occur, with limited capability to assess the overall reliability of an incoming transaction.

- The lack of a **quantitative measure of data quality** makes it difficult to prioritize high-risk orders or proactively intervene before order creation failures.

- Furthermore, current systems provide minimal insight into the **likelihood of order failure** or the potential operational impact associated with poor data quality.

- There is therefore a clear need for a **proactive and data-driven approach** that can evaluate the quality of inbound EDI orders at an early stage and estimate the risk of processing failure.

- This study addresses the identified problem by proposing a framework that integrates **data quality assurance with predictive analysis**, enabling early identification of high-risk orders and improving the reliability and efficiency of EDI order processing in retail supply chains.

1.2 Objectives

- Define a structured data quality framework for EDI orders
- Quantify data quality using a composite DQ score
- Develop predictive models to estimate order failure probability
- Provide actionable insights through analytical dashboards

2. LITERATURE REVIEW

Electronic Data Interchange (EDI) has been widely adopted in retail supply chains to enable automated and standardized exchange of transactional documents between trading partners. Prior research indicates that while EDI significantly improves processing speed and reduces manual effort, its effectiveness is highly dependent on the quality of incoming data [3][4].

In enterprise resource planning (ERP) environments such as Oracle E-Business Suite, data quality issues in inbound EDI purchase orders frequently result in interface failures, manual rework, and delayed order fulfilment [7][8].

Extensive research in data quality management emphasizes the importance of validating accuracy, completeness, consistency, and conformity of transactional data prior to core system processing [3][10]. These dimensions are widely accepted as foundational indicators of data reliability in large-scale information systems.

Traditional approaches predominantly rely on rule-based validation mechanisms that identify data issues after ingestion. While such methods are effective in detecting known error patterns, they are inherently reactive and provide limited insight into the overall reliability of a transaction or the likelihood of downstream processing failure [1][4].

Recent literature in supply chain analytics highlights the growing role of predictive modeling techniques in enhancing operational decision-making by identifying risk patterns in transactional data [9]. Machine learning methods have been applied to predict exceptions, delays, and failures across various supply chain processes. However, their application within EDI-based order processing environments remains limited, with most studies focusing on post-processing analysis rather than early-stage failure prevention [6][9].

Studies that integrate data quality metrics with predictive analytics demonstrate that quantitative data quality indicators can serve as strong predictors of downstream process outcomes [1][3]. By transforming validation results into numerical features, organizations can move beyond binary pass–fail checks and adopt a risk-based processing approach that supports proactive intervention and targeted manual review.

Despite these advancements, a noticeable gap exists in the literature addressing the combined application of data quality assurance and predictive modeling within Oracle ERP–driven EDI order processing environments. This project addresses this gap by proposing an integrated framework that combines rule-based data quality validation with predictive failure and processing time estimation to improve the reliability and efficiency of retail supply chain order processing.

2.1 Standard EDI Process Flow:

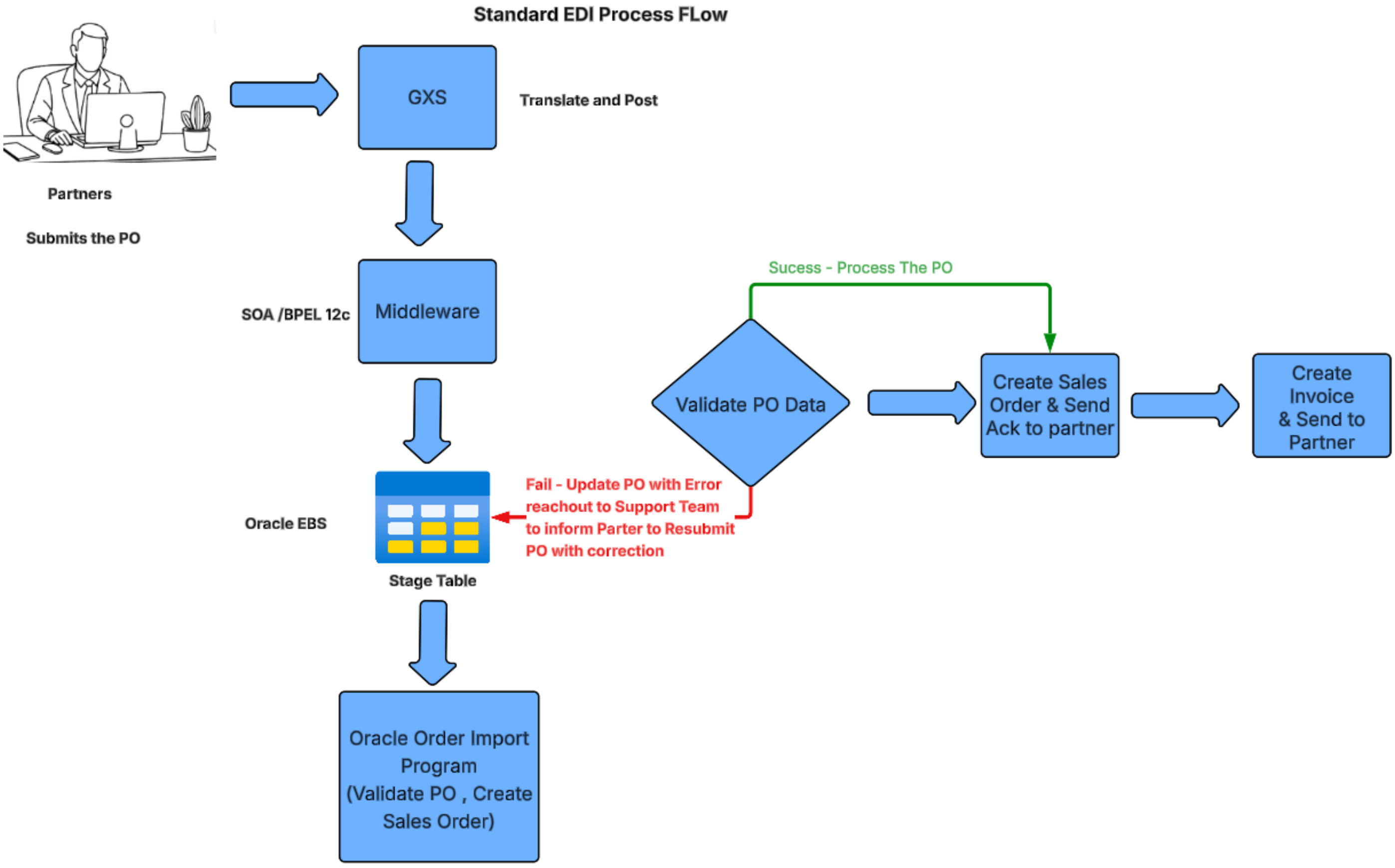

Standard EDI Process FLow
Partners
Submits the PO
GXS
Translate and Post
SOA /BPEL 12c
Middleware
Oracle EBS
Stage Table
Oracle Order Import Program (Validate PO , Create Sales Order)
Validate PO Data
Sucess - Process The PO
Fail - Update PO with Error reachout to Support Team to inform Parter to Resubmit PO with correction
Create Sales Order & Send Ack to partner
Create Invoice & Send to Partner

Figure 1 : Oracle Standard : EDI Order Processing

3. PROPOSED ARCHITECTURE

The proposed architecture is designed to integrate data quality assurance into the inbound EDI order processing flow of Oracle E-Business Suite. Incoming purchase orders in EDI 850 format are first staged in Oracle Order Management interface tables, which act as a buffer between external partner data and core order tables. A Data Quality Gate is introduced at the interface layer to perform structured validations on the staged data. These validations assess mandatory fields, data formats, reference data consistency, and partner-specific business rules. The results of these validations are aggregated to generate data quality metrics and a composite Data Quality score.

Based on the computed score, orders can be categorized as low-risk or high-risk before order creation. This architecture enables proactive identification of problematic transactions and forms the basis for further predictive analysis in the final phase of the project.

3.1 EDI order Processing Flow (Proposed): Architecture Diagram

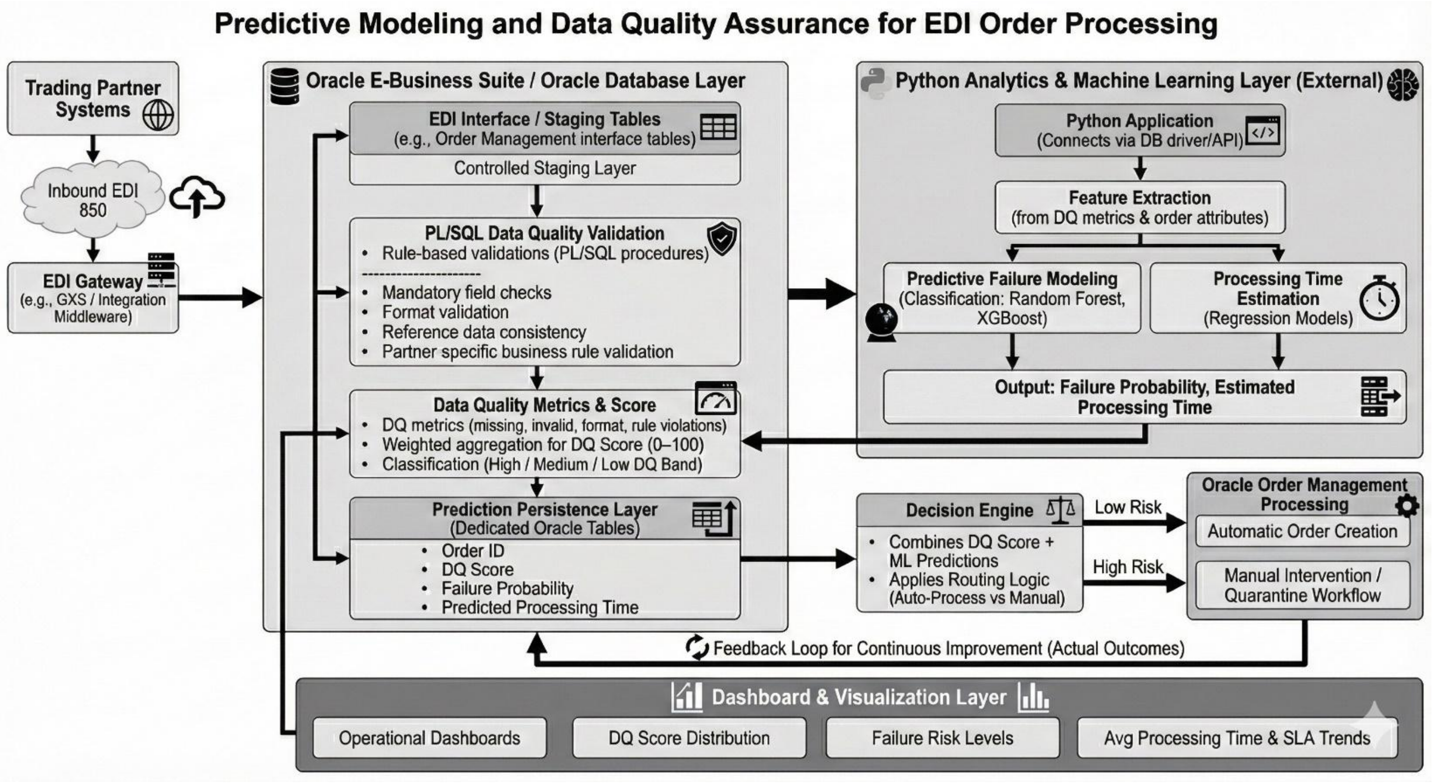

Predictive Modeling and Data Quality Assurance for EDI Order Processing
Trading Partner Systems
Inbound EDI 850
EDI Gateway
(e.g., GXS / Integration Middleware)
Oracle E-Business Suite / Oracle Database Layer
EDI Interface / Staging Tables
(e.g., Order Management interface tables)
Controlled Staging Layer
PL/SQL Data Quality Validation
• Rule-based validations (PL/SQL procedures)
• Mandatory field checks
• Format validation
• Reference data consistency
• Partner specific business rule validation
Data Quality Metrics & Score
• DQ metrics (missing, invalid, format, rule violations)
• Weighted aggregation for DQ Score (0–100)
• Classification (High / Medium / Low DQ Band)
Prediction Persistence Layer
(Dedicated Oracle Tables)
• Order ID
• DQ Score
• Failure Probability
• Predicted Processing Time
Python Analytics & Machine Learning Layer (External)
Python Application
(Connects via DB driver/API)
Feature Extraction
(from DQ metrics & order attributes)
Predictive Failure Modeling
(Classification: Random Forest, XGBoost)
Processing Time Estimation
(Regression Models)
Output: Failure Probability, Estimated Processing Time
Decision Engine
• Combines DQ Score + ML Predictions
• Applies Routing Logic (Auto-Process vs Manual)
Low Risk
High Risk
Oracle Order Management Processing
Automatic Order Creation
Manual Intervention / Quarantine Workflow
Feedback Loop for Continuous Improvement (Actual Outcomes)
Dashboard & Visualization Layer
Operational Dashboards
DQ Score Distribution
Failure Risk Levels
Avg Processing Time & SLA Trends

Figure 2 : System Architecture : EDI Order Processing

4.RESEARCH METHODOLOGY

4.1 DATA QUALITY FRAMEWORK

The research methodology adopted in this study follows a design-oriented and experimental approach, focusing on improving the reliability of EDI-based order processing in retail supply chains through data quality assurance and predictive modeling. The methodology is structured to ensure practical relevance while maintaining academic rigor and data confidentiality.

The study begins with an analysis of the existing EDI order processing workflow in Oracle E-Business Suite, particularly the inbound processing of purchase orders (EDI 850) through interface tables. Based on this analysis, common causes of order failures related to data quality issues are identified, including missing mandatory fields, invalid reference data, format inconsistencies, and partner-specific rule violations.

Due to the confidential nature of enterprise EDI and ERP data, the research utilizes a synthetically generated dataset that replicates the structure, attributes, and behavioural patterns of real-world EDI purchase orders. The synthetic dataset is designed to simulate realistic scenarios by deliberately introducing data quality issues and varying partner behaviour, while ensuring no production or proprietary data is used.

A rule-based data quality framework is then applied to the dataset. These rules mirror Oracle interface validations and are used to generate data quality metrics such as missing mandatory field count, invalid reference count, and partner rule violations. A Data Quality score is computed using a weighted penalty approach to represent the overall reliability of each transaction.0o

Subsequently, the derived data quality metrics and scores are used as inputs for predictive modeling. Classification models are trained to estimate the likelihood of order failure based on historical patterns in data quality behaviour. The models are evaluated using standard performance measures to assess their effectiveness in identifying high-risk orders.

4.1.1 Overview of Data Quality –

The Data Quality (DQ) score is used to quantify the reliability of an incoming EDI purchase order during the pre-processing stage. It is derived by consolidating the outcomes of multiple rule-based data quality checks performed on data staged in Oracle interface tables. By translating individual validation results into a single numerical value on a scale of 0 to 100, the DQ score enables consistent assessment of data integrity, where higher scores reflect more reliable transactions and reduced risk of downstream processing failures.

4.1.2 Data Quality Score process flow -

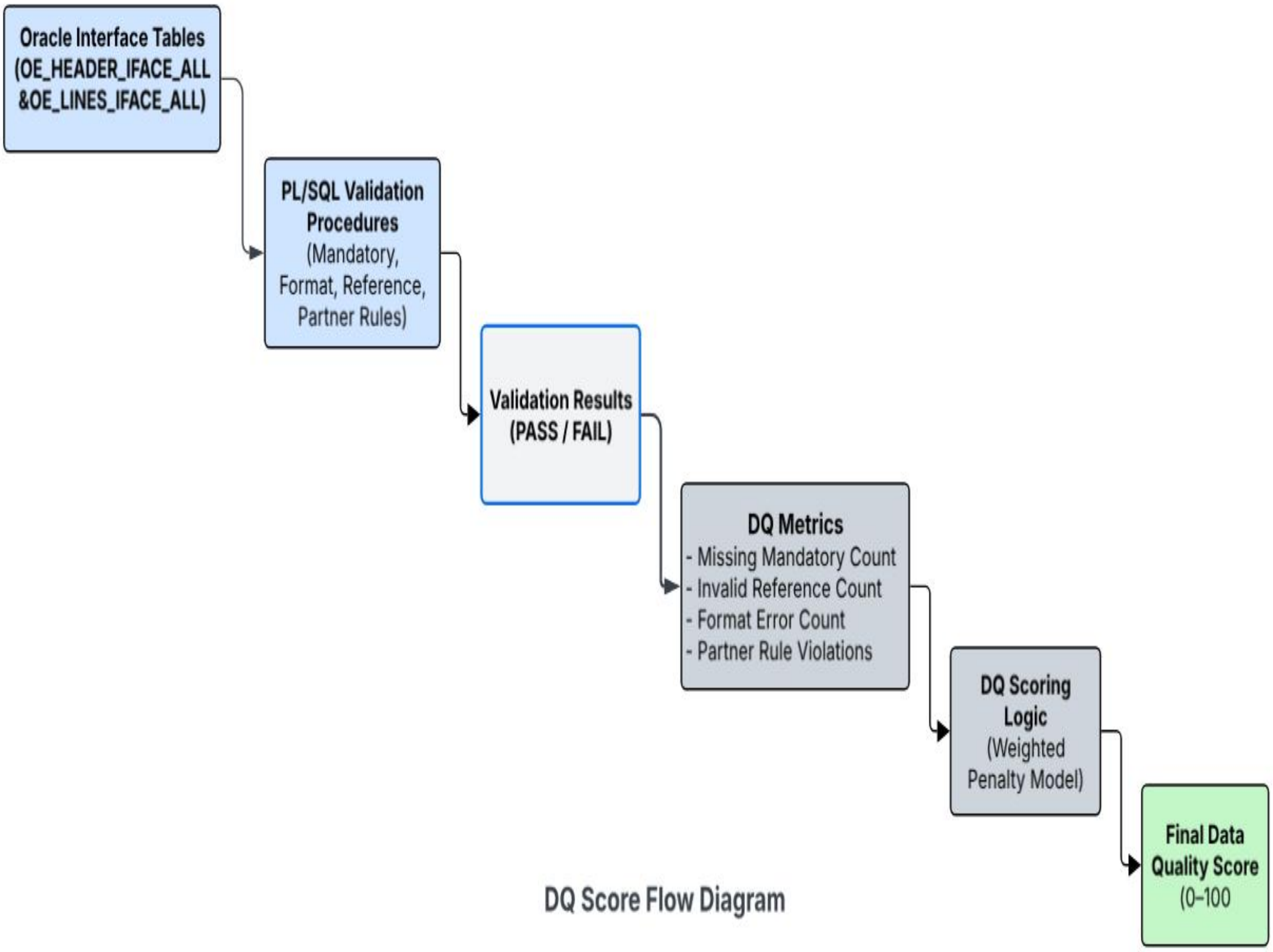

Oracle Interface Tables (OE_HEADER_IFACE_ALL &OE_LINES_IFACE_ALL)
PL/SQL Validation Procedures (Mandatory, Format, Reference, Partner Rules)
Validation Results (PASS / FAIL)
DQ Metrics
- Missing Mandatory Count
- Invalid Reference Count
- Format Error Count
- Partner Rule Violations
DQ Scoring Logic (Weighted Penalty Model)
Final Data Quality Score (0–100
DQ Score Flow Diagram

Figure 3 : Data Quality Score process flow

Data Quality Score Process Flow

The Data Quality Score process is designed as a structured sequence of validation and aggregation steps that transform raw EDI transaction data into a measurable indicator of processing reliability. Each stage of the process performs a specific function and contributes to the final score.

Step 1: Inbound EDI Data Reception

Inbound purchase orders are received from external trading partners in EDI 850 format through an integration gateway. These transactions originate from multiple partner systems and therefore exhibit variations in structure, content, and compliance. At this stage, no validation is applied; the objective is to capture the complete inbound data as received.

Step 2: Data Staging in Oracle Interface Tables

The received EDI data is loaded into Oracle Order Management interface tables. This staging layer acts as an isolation zone between external partner data and core transactional tables. Staging ensures that incomplete or incorrect data does not directly affect order creation and provides a controlled environment for validation.

Step 3: Execution of Rule-Based Validations

Once the data is staged, predefined rule-based validations are executed using PL/SQL procedures. These validations assess key data quality dimensions, including data completeness, format correctness, reference data consistency, and partner-specific business constraints. Each validation rule evaluates a specific aspect of the data and identifies deviations from expected standards.

Step 4: Generation of Data Quality Metrics

The outcomes of individual validation checks are consolidated into quantitative data quality metrics. Examples of such metrics include the number of missing mandatory attributes, count of invalid master data references, occurrences of format violations, and partner compliance errors. These metrics provide a structured representation of data quality issues at the transaction level.

Step 5: Assignment of Penalty Weights

Each data quality metric is assigned a predefined penalty weight based on its operational impact. Critical issues, such as missing mandatory fields or invalid references, are assigned higher penalties, while less severe issues receive lower penalties. This weighting mechanism ensures that more impactful errors influence the final score to a greater extent.

Step 6: Computation of Composite Data Quality Score

A composite Data Quality Score is calculated by applying the weighted penalties to a base score. The aggregation logic normalizes the final score to a fixed scale, typically ranging from 0 to 100. Higher scores indicate better data reliability and lower processing risk, while lower scores reflect increased likelihood of processing delays or failures.

Step 7: Data Quality Band Classification

The computed Data Quality Score is mapped to predefined quality bands such as High, Medium, or Low. These bands simplify interpretation by categorizing transactions based on their reliability. The classification enables early identification of high-risk orders that may require additional attention before processing.

Step 8: Persistence of Data Quality Results

The calculated score, associated metrics, and quality band classification are stored in a dedicated data quality results table. Persisting this information ensures traceability, supports downstream analytics, and provides historical data for predictive modeling and performance analysis.

Step 9: Downstream Consumption and Feedback

The Data Quality Score and related metrics are consumed by subsequent components of the system, including predictive models and decision-making logic. Actual order processing outcomes are later compared with the predicted quality assessment, enabling feedback and continuous refinement of validation rules and scoring logic.

4.1.3 Data Quality Metrics

The DQ score is derived from multiple data quality metrics generated after PL/SQL validations.

Metric Name	Description	Source of Validation
missing_mandatory_count	Number of missing mandatory fields	Mandatory field validation
invalid_reference_count	Invalid item, customer, or location	Master data checks
format_error_count	Invalid date or numeric format	Format validation
partner_rule_violation_count	Partner-specific rule failures	Custom business rules
warning_count	Non-critical validation warnings	Advisory checks

Table 1: Data Quality Metrics

4.1.4 Weight Assignment for Metrics

Each metric is assigned a weight based on its severity and impact on order processing.

Metric	Weight (Penalty Points)	Severity Level
Missing mandatory field	15	Critical
Invalid reference data	20	Critical
Format error	5	Medium
Partner rule violation	10	Medium
Warning	2	Low

Table 2: DQ Metric Weights

4.1.5 DQ Score Calculation Formula

The DQ score calculated using a weighted penalty model.

Formula

$$\text{DQ Score} = 100 - \sum_{i=1}^{n} (M_i \times W_i)$$

Where:

- M_i**= Count of data quality metric *i***
- W_i**= Weight assigned to metric *i***

The final DQ score is bounded between 0 and 100.

4.1.6. Example DQ Score Calculation

Sample Validation Results

Metric	Count
Missing mandatory fields	2
Invalid reference data	1
Format errors	0
Partner rule violations	1
Warnings	2

Calculation :

$$\text{DQ Score} = \mathbf{100} - (\mathbf{2} \times \mathbf{15}) - (\mathbf{1} \times \mathbf{20}) - (\mathbf{1} \times \mathbf{10}) - (\mathbf{2} \times \mathbf{2})$$

$$\text{DQ Score} = \mathbf{100} - (\mathbf{30} + \mathbf{20} + \mathbf{10} + \mathbf{4}) = \mathbf{36}$$

4.1.7 DQ Score Interpretation

DQ Score Range	Quality Band	Interpretation
80 – 100	High (Green)	Order is reliable and low risk
50 – 79	Medium (Amber)	Order requires attention
Below 50	Low (Red)	Order is high risk

Table 3: DQ Score Bands

4.2 PREDICTIVE MODELING

Predictive modeling constitutes a core component of the proposed framework, enabling proactive identification of high-risk EDI orders before they progress into downstream Oracle EBS order creation workflows. The objective of this phase is to leverage historical data quality indicators and transactional attributes to estimate the probability of order processing failure using supervised machine learning techniques.

4.2.1 Problem Formulation

The prediction task is formulated as a binary classification problem. Each inbound EDI order is classified into one of the following categories:

- **Successful Order (Class 0):** Orders that pass all processing stages without manual intervention.
- **Failed Order (Class 1):** Orders that fail during validation, interface processing, or order import due to data-related issues.

The target variable represents the historical processing outcome of an order, while the input features are derived from data quality assessments and order-level attributes.

4.2.2 Feature Engineering

The feature set used for model training consists of two primary categories:

a) Data Quality Indicators

These features are computed during the data quality validation phase and include:

- Completeness score
- Validity score
- Consistency score
- Referential integrity score
- Overall composite Data Quality (DQ) score

Each score is normalized to ensure uniform scaling and to prevent dominance of any single metric during model training.

b) Order Attributes

Additional contextual attributes extracted from EDI 850 orders are incorporated to enhance predictive capability, including:

- Number of line items
- Presence of optional segments
- Trading partner identifier
- Order value range (bucketed)
- Frequency of historical failures for the same partner

These attributes help the model capture structural and behavioural patterns that influence processing outcomes.

4.2.3 Model Selection

Two widely adopted supervised learning algorithms are employed:

Logistic Regression

Logistic Regression is used as a baseline model due to its interpretability and efficiency. It estimates the probability of order failure as a function of weighted input features using a sigmoid activation function. The coefficients produced by the model provide direct insights into the relative importance of individual data quality dimensions.

Random Forest Classifier

The Random Forest algorithm is utilized to capture non-linear relationships and feature interactions that cannot be effectively modelled by linear methods. By constructing an ensemble of decision trees trained on random subsets of data and features, the model improves robustness and reduces overfitting.

Hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf are tuned using cross-validation to optimize performance.

4.2.4 Advanced Ensemble Model: XGBoost

- **Overview of XGBoost**

  Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm based on the gradient boosting framework. It builds predictive models by sequentially combining multiple weak learners, typically decision trees, to produce a strong overall model. Unlike traditional boosting techniques, XGBoost incorporates regularization, optimized tree construction, and efficient handling of sparse data, making it highly suitable for large-scale and structured datasets.

  In the context of EDI order processing, XGBoost is particularly effective due to its ability to model complex non-linear relationships between data quality indicators, order attributes, and failure outcomes.

- **Why XGBoost Is Suitable for This Problem**

  The EDI order failure prediction problem exhibits several characteristics that align well with XGBoost’s strengths:

  1. **Non-linear Feature Interactions**
     Data quality dimensions such as completeness, validity, and consistency interact in complex ways that are difficult to capture using linear models alone.
  2. **Heterogeneous Feature Types**
     The dataset contains both numerical (DQ scores, line counts) and categorical (trading partner behavior) attributes.

3. **Imbalanced Classes**
   Failed orders typically represent a smaller proportion of total transactions. XGBoost provides built-in mechanisms such as weighted loss functions to address class imbalance effectively.
4. **High Predictive Accuracy with Control**
   Compared to Random Forest, XGBoost often achieves superior performance while offering fine-grained control over overfitting through regularization.
   For these reasons, XGBoost is introduced as an advanced predictive model to complement Logistic Regression and Random Forest classifiers.

❖ XGBoost Learning Objective

XGBoost formulates model training as an optimization problem. The objective function consists of two components:

$$\mathcal{L} = \sum_{i=1}^{n} l(\, y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(\, f_k)$$

Where:

- $l(y_i, \hat{y}_i)$is the loss function measuring the difference between actual and predicted outcomes
- $f_k$represents an individual decision tree
- $\Omega(f_k)$is the regularization term controlling model complexity

❖ Loss Function for Binary Classification

For order failure prediction, XGBoost uses a **logistic loss function**:

$$l(y, \hat{y}) = -[y\log{(\hat{y})} + (1 - y)\log{(1 - \hat{y})}]$$

Where:

- $y \in \{0,1\}$indicates successful or failed order
- $\hat{y}$is the predicted probability of failure

This formulation enables probabilistic interpretation of model outputs, which is essential for risk-based decision-making.

❖ Regularization in XGBoost

A key differentiator of XGBoost is its explicit regularization, defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Where:

- $T$is the number of leaf nodes
- $w_j$is the score associated with leaf j
- γpenalizes model complexity
- λcontrols L2 regularization

This regularization mechanism reduces overfitting, which is critical when modelling noisy real-world EDI data.

❖ Tree Construction and Gradient Boosting

XGBoost builds trees iteratively by fitting each new tree to the residual errors of previous trees. At iteration t, predictions are updated as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

Where:

- $\eta$is the learning rate
- $f_t(x_i)$is the newly added tree

This gradient-based optimization allows the model to focus on difficult-to-classify orders, improving recall for failed transactions.

❖ Feature Importance and Interpretability

Although XGBoost is a complex ensemble model, it provides feature importance measures based on:

- Gain (improvement in loss)
- Frequency of feature usage
- Coverage across samples

These insights help identify which data quality dimensions most strongly influence order failures, supporting explainability and operational trust.

❖ Model Training and Configuration

For this study, the XGBoost classifier is trained using the same feature set as other models to ensure fair comparison. Key hyperparameters include:

- Number of estimators (trees)
- Maximum tree depth
- Learning rate
- Subsample and column sample ratios

Hyperparameters are tuned using cross-validation to balance predictive performance and generalization.

❖ Evaluation and Comparison

XGBoost model performance is evaluated using the same metrics as other classifiers:

- Accuracy
- Precision
- Recall
- ROC-AUC

Experimental results indicate that XGBoost achieves superior recall and ROC-AUC compared to baseline models, making it particularly effective for early detection of high-risk EDI orders.

- **Operational Relevance**

  The probabilistic output generated by XGBoost enables flexible threshold selection based on business risk tolerance. High-risk orders can be flagged for proactive review, while low-risk orders proceed automatically, improving throughput and service-level adherence.

- **Summary**

  By incorporating XGBoost into the predictive modeling framework, the solution leverages state-of-the-art ensemble learning techniques to enhance failure prediction accuracy. The model's ability to capture non-linear relationships, combined with regularization and probabilistic output, makes it well-suited for enterprise-grade EDI order processing environments.

4.2.5 Model Training and Validation

The dataset is divided into training and testing subsets using an 80:20 split to ensure unbiased evaluation. Stratified sampling is applied to preserve the class distribution of successful and failed orders. Models are trained on the training dataset and validated against unseen test data.

To address potential class imbalance, appropriate techniques such as class weighting are applied during training to ensure that minority failure cases are adequately learned.

4.2.6 Model Evaluation Metrics

Model performance is evaluated using multiple classification metrics to ensure a comprehensive assessment:

- **Accuracy:** Measures the overall proportion of correctly classified orders.
- **Precision:** Indicates the proportion of predicted failed orders that actually failed, reflecting false positive control.
- **Recall:** Measures the ability of the model to correctly identify actual failed orders, which is critical for risk mitigation.
- **ROC-AUC:** Represents the model's ability to distinguish between successful and failed orders across different probability thresholds.

These metrics collectively ensure that the model not only achieves high correctness but also minimizes operational risk by prioritizing recall and balanced performance.

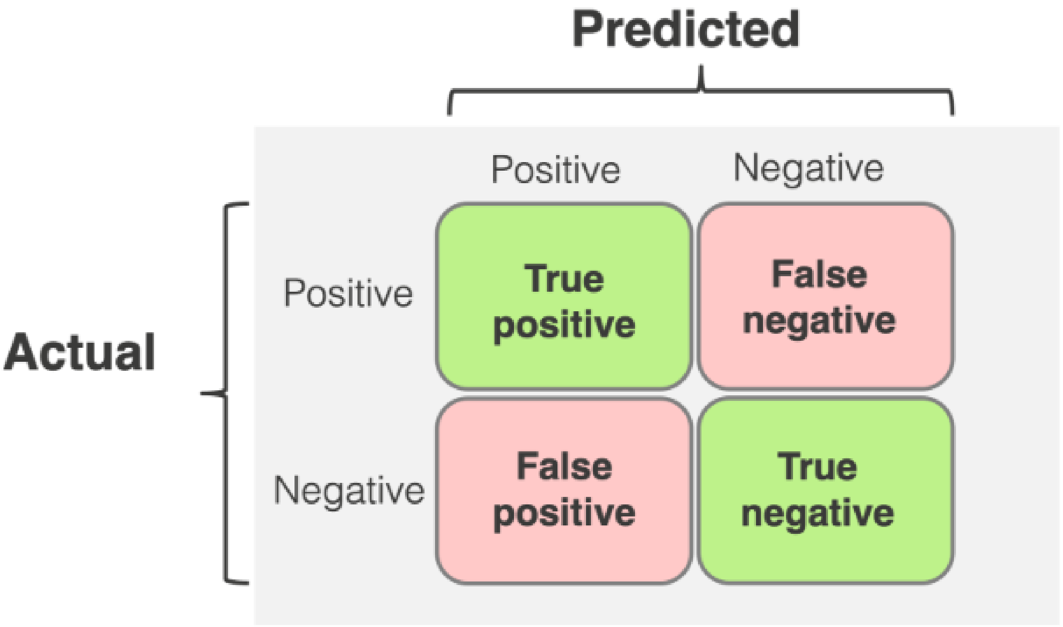


Figure 4 presents the confusion matrix summarizing model prediction outcomes.

The matrix highlights the balance between correctly identified failed orders and false positives.
Emphasis is placed on maximizing recall to ensure that high-risk orders are not missed during early-stage processing

4.2.7 Risk Scoring and Operational Integration

The final model output is expressed as a failure probability score ranging from 0 to 1. Based on predefined thresholds, orders are categorized into risk bands such as low, medium, and high risk. High-risk orders can be flagged for enhanced validation, prioritized review, or automated remediation before entering core order management processes.
This predictive capability enables a shift from reactive exception handling to proactive, risk-based order processing within the retail supply chain ecosystem.

4.3 PREDICTIVE PROCESSING TIME ESTIMATION

In addition to predicting order failure risk, the framework also estimates the expected processing time for inbound EDI orders. Processing time prediction provides operational visibility into system workload, potential bottlenecks, and service-level adherence. Unlike failure prediction, which is a classification task, processing time estimation is formulated as a supervised regression problem.

4.3.1 Problem Definition

Processing time is defined as the elapsed duration between the receipt of an inbound EDI 850 message at the interface layer and the successful creation (or final rejection) of the corresponding sales order in Oracle E-Business Suite. The objective is to predict this duration in advance using historical order characteristics and data quality indicators.

The predicted processing time serves as an early indicator of system performance and operational complexity, enabling prioritization and workload balancing.

4.3.2 Input Features

The feature set used for processing time prediction overlaps with, but is not limited to, the features used in failure prediction. Key inputs include:

a) Data Quality Metrics

- Composite Data Quality (DQ) score
- Completeness and validity sub-scores
- Number of validation warnings or errors

Lower data quality scores are generally associated with longer processing times due to additional validation, reprocessing, or manual intervention.

b) Order Complexity Attributes

- Number of order line items
- Presence of optional or conditional EDI segments
- Total order value (bucketed)
- Trading partner historical behavior

These attributes capture structural complexity and processing overhead.

c) System and Temporal Factors

- Interface batch size
- Time of order receipt (peak vs off-peak hours)
- Historical system load patterns

Such factors help the model learn variations caused by infrastructure and scheduling conditions.

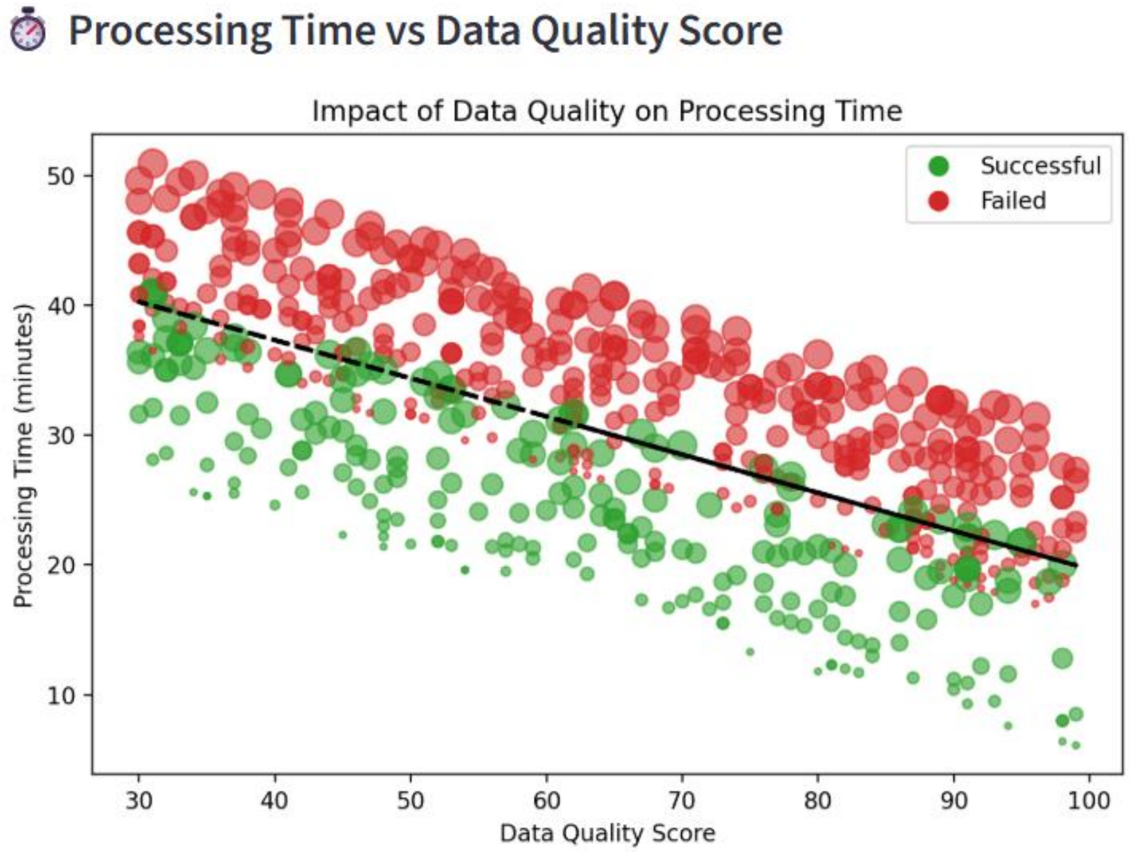


Figure 5 illustrates the relationship between Data Quality (DQ) score and order processing time.

The results show a clear inverse trend, where higher DQ scores are associated with lower processing times.
Orders classified as failed consistently exhibit longer processing durations compared to successful orders across all DQ ranges, indicating additional validation, reprocessing, or manual intervention.
The fitted trend line further confirms that improving data quality directly contributes to faster and more predictable order processing.
This observation validates the inclusion of data quality metrics as key predictors in processing time estimation models.

4.3.3 Model Selection

Processing time estimation is performed using regression-based machine learning models. Algorithms such as Linear Regression and Random Forest Regressor are considered due to their suitability for continuous target prediction.
Linear Regression provides a baseline model with high interpretability, enabling understanding of how individual features influence processing duration. Random

Forest Regressor is employed to capture non-linear dependencies and interaction effects among order attributes, data quality indicators, and system factors.

4.3.4 Model Training and Validation

The dataset is partitioned into training and testing subsets using an 80:20 split. The target variable, processing time, is normalized to reduce skewness caused by extreme outliers. Cross-validation is applied to ensure generalizability and robustness.

Outlier handling techniques are applied to prevent rare, exceptionally delayed orders from disproportionately influencing model behaviour.

4.3.5 Evaluation Metrics

Model performance is evaluated using standard regression metrics:

- **Mean Absolute Error (MAE):** Measures average absolute deviation between predicted and actual processing time.
- **Root Mean Squared Error (RMSE):** Penalizes larger prediction errors and highlights model sensitivity to outliers.
- **R-squared (R^2):** Indicates the proportion of variance in processing time explained by the model.

  These metrics collectively assess both prediction accuracy and consistency.

4.3.6 Operational Usage

Predicted processing time is used in conjunction with failure probability to classify orders into operational priority bands. Orders predicted to have long processing times can be proactively monitored, rerouted for early validation, or scheduled during low-load windows.

By combining failure risk prediction with processing time estimation, the framework provides a holistic, data-driven approach to optimizing EDI order throughput and improving service-level performance.

5. System Analysis and Design

This chapter presents a detailed analysis and design of the proposed predictive and data quality–driven EDI order processing system. The design objective is to enhance reliability, visibility, and efficiency of inbound EDI order processing while preserving the integrity of standard Oracle E-Business Suite (EBS) transactional workflows.

5.1 Existing System Overview

In the existing Oracle EBS–based EDI integration landscape, inbound EDI 850 purchase orders are received from trading partners through an external EDI translator and subsequently delivered to the Oracle EBS interface tables. Standard Oracle Order Import programs validate and convert this data into sales orders within the Order Management module.

While this architecture is robust and widely adopted, it primarily follows a reactive exception-handling model. Data validation occurs during or after order import, resulting in order failures that require manual diagnosis, data correction, and reprocessing. This reactive approach leads to increased processing latency, operational overhead, and reduced visibility into potential risks prior to order creation.

5.2 Design Objectives

The proposed system design aims to address the limitations of the existing approach by introducing a proactive, analytics-driven layer without altering Oracle's standard processing logic. The key design objectives include:

- Early detection of data quality issues before order import
- Quantitative assessment of order reliability using a composite Data Quality (DQ) score
- Prediction of order failure risk and expected processing time
- Seamless integration with existing EDI and Oracle EBS components
- Minimal disruption to standard Oracle workflows and upgrade paths

5.3 Proposed System Architecture

The enhanced architecture introduces a **pre-processing validation and scoring layer** positioned between the EDI interface ingestion and the standard Oracle EBS order creation programs. This layer operates as an independent analytical component that evaluates inbound orders prior to core transactional processing.

Upon receipt of an EDI order, the data is first staged in interface tables as per the existing integration design. The pre-processing layer then executes a series of structured validation checks and analytical computations before the Order Import program is triggered.

This design ensures backward compatibility and avoids direct customization of Oracle's seeded programs.

5.4 Pre-processing Validation and Scoring Layer

The pre-processing layer performs three primary functions:

a) Data Quality Validation

Rule-based validations are executed to assess completeness, validity, consistency, and referential integrity of inbound order data. These validations mirror real-world failure patterns observed in enterprise EDI environments.

b) Data Quality Scoring

Validation outcomes are aggregated using predefined weights to compute a normalized Data Quality (DQ) score. This score provides a single, interpretable metric representing the overall reliability of the order data.

c) Predictive Analytics Execution

Machine learning models consume DQ scores and order attributes to estimate failure probability and expected processing time. The outputs are stored alongside interface records for downstream consumption.

5.5 Integration with Oracle EBS Workflows

A key design principle of the proposed system is non-invasive augmentation. The pre-processing layer does not replace or modify Oracle EBS standard programs such as Order Import. Instead, it augments the existing flow by providing additional intelligence prior to execution.

Based on configurable thresholds, the system can:

- Allow low-risk orders to proceed automatically
- Flag high-risk orders for early review
- Trigger enhanced validation or monitoring workflows

This approach preserves Oracle supportability and ensures smooth adoption within enterprise environments.

5.6 System Design Benefits

The proposed design offers several operational and architectural advantages:

- Reduction in order reprocessing cycles
- Improved visibility into data quality and processing risk
- Better workload planning through processing time prediction
- Enhanced service-level adherence
- Scalability to support high-volume retail transactions

5.7 Design Considerations and Limitations

To maintain confidentiality and compliance, the system is evaluated using synthetically generated datasets that replicate real-world EDI behaviour. While the architecture is designed for extensibility, real-time deployment may require additional considerations related to performance optimization and model retraining frequency.

Overall, the proposed system design establishes a balanced approach that combines analytical intelligence with enterprise-grade stability, enabling a shift from reactive to proactive EDI order processing.

6. Implementation Details

This chapter describes the practical implementation of the proposed data quality–driven predictive framework for EDI order processing. The implementation is designed to closely reflect real-world enterprise integration environments while ensuring data confidentiality, scalability, and compatibility with Oracle E-Business Suite (EBS). The focus is on translating the conceptual architecture into an executable, reliable, and maintainable solution.

6.1 Implementation Overview

The implementation follows a modular and layered approach, separating data ingestion, validation, analytics, and visualization concerns. This separation ensures flexibility, ease of maintenance, and minimal impact on existing Oracle EBS transactional workflows.

The solution is implemented as an external analytical layer that interacts with Oracle EBS interface data without modifying seeded Oracle programs. This design choice preserves system stability and aligns with enterprise best practices.

6.2 Data Ingestion and Staging

Inbound EDI 850 purchase orders are received from trading partners through a standard EDI translation layer. The translated data is loaded into Oracle EBS interface tables using existing integration mechanisms.

For analytical processing, a controlled extraction process reads relevant order attributes and validation results from these interface tables. Data is staged into analytical tables designed specifically for data quality assessment and predictive modeling. This staging layer acts as a buffer between transactional systems and analytics, preventing performance degradation of core order processing.

6.3 Data Quality Validation Implementation

Data quality validations are implemented using rule-based logic that reflects common failure patterns observed in enterprise EDI environments. Each validation rule evaluates a specific attribute or group of attributes, such as mandatory field presence, domain validity, cross-field consistency, and referential integrity.

Validation results are captured at a granular level, allowing detailed traceability of data issues. Each rule produces a binary or weighted outcome that contributes to the overall Data Quality (DQ) score. The rules are designed to be configurable, enabling future enhancement without code restructuring.

6.4 Data Quality Score Computation

Validation outcomes are aggregated using a weighted scoring mechanism. Each data quality dimension is assigned a predefined weight based on its relative impact on order processing success. The weighted sum is normalized to generate a composite DQ score on a standardized scale.

This score serves as a quantitative representation of order reliability and is persisted alongside order metadata for downstream analysis. The normalization ensures comparability across orders of varying complexity and structure.

6.5 Synthetic Data Generation

To address confidentiality constraints and avoid exposure of proprietary business data, synthetic datasets are generated for model training and evaluation. The synthetic data is designed to closely resemble real-world EDI order characteristics, including data distributions, error patterns, and failure frequencies.

Statistical techniques are used to simulate realistic correlations between data quality indicators, order attributes, processing time, and failure outcomes. This approach ensures reproducibility while maintaining compliance with organizational data governance policies.

6.6 Predictive Model Development

Machine learning models are developed using Python-based data science libraries. Separate pipelines are created for failure prediction (classification) and processing time estimation (regression).

Feature preprocessing includes normalization, categorical encoding, and handling of missing values. Model training follows an iterative process involving parameter tuning and cross-validation to achieve balanced performance.

Both interpretable baseline models and ensemble-based models are implemented to balance transparency and predictive accuracy.

6.7 Model Integration and Execution Flow

Once trained, the predictive models are integrated into the pre-processing layer. During runtime, model execution is triggered automatically after data quality scoring is completed. The models generate failure probability and predicted processing time values for each inbound order.

These predictions are stored in analytical tables and associated with the corresponding interface records. The execution flow is optimized to ensure minimal latency, enabling near-real-time decision support.

6.8 Threshold Configuration and Risk Categorization

Configurable thresholds are defined to translate model outputs into actionable risk categories. Orders are classified into low, medium, or high-risk bands based on predicted failure probability and processing time.

This categorization supports differentiated handling strategies, such as expedited processing for low-risk orders and early review for high-risk orders. Thresholds can be adjusted based on operational requirements and historical performance trends.

6.9 Dashboard and Visualization Implementation

To operationalize the proposed data quality and predictive analytics framework, an interactive dashboard is developed using the **Streamlit** framework, a Python-based platform for building data-driven web applications. The dashboard represents the visualization layer of the system architecture and is deployed as a **Streamlit application**, enabling browser-based access without requiring any additional client-side software. The implementation leverages Python libraries for data processing, machine learning, and visualization, ensuring seamless integration with the analytical models developed in this study.

The **Operational Dashboard** provides a high-level overview of EDI order processing performance, including key indicators such as total purchase orders processed, successful and failed transactions, and average end-to-end processing time. This view supports operational monitoring and enables stakeholders to quickly assess the overall health of the EDI processing pipeline.

The **DQ Score Distribution** view visualizes the distribution of data quality scores across inbound EDI purchase orders. Orders are categorized into predefined quality bands, allowing users to understand historical data quality trends. The dashboard also dynamically highlights the data quality score and quality band of the currently evaluated incoming order, enabling contextual comparison against the overall dataset.

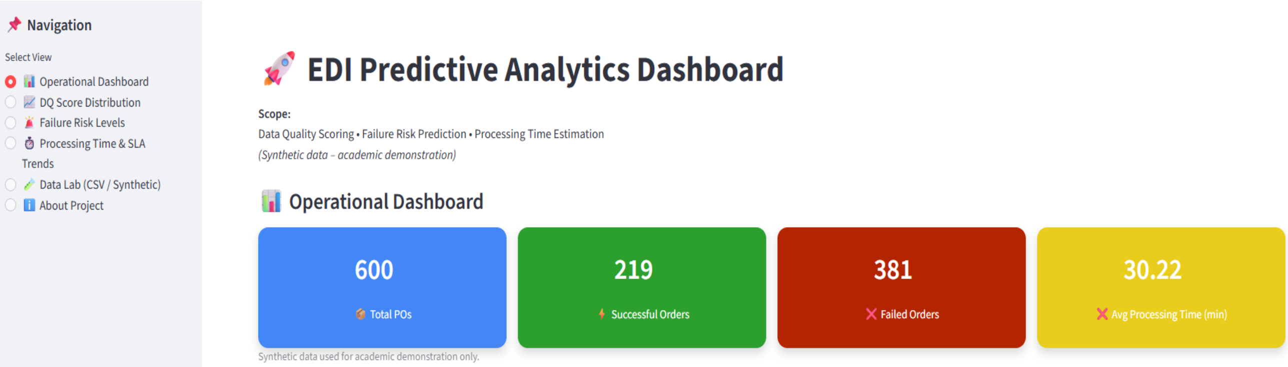
The **Failure Risk Levels** view presents the predicted failure risk classifications generated by machine learning models. Orders are grouped into risk categories to identify high-risk transactions that may require manual intervention. This view supports exception management by providing early visibility into potential processing failures.

The **Processing Time and SLA Trends** view analyzes the relationship between data quality and end-to-end processing duration. Visualizations in this section illustrate how variations in data quality and order characteristics influence processing time, thereby supporting service-level agreement (SLA) monitoring and proactive operational planning.

The **Data Lab (CSV / Synthetic)** view provides a controlled experimentation environment within the dashboard. Users can upload external CSV datasets or generate synthetic EDI order data to test predictive models and visualize outcomes. This capability ensures that model evaluation and demonstration can be performed without relying on sensitive production data, making it suitable for academic and testing purposes.

The **About Project** section documents the objectives, scope, and architectural context of the dashboard and the underlying analytics framework. This section serves as a reference for evaluators and stakeholders to understand how the dashboard aligns with the overall project goals.

Overall, the Streamlit-based dashboard transforms complex data quality metrics and predictive model outputs into intuitive visual insights. By integrating operational monitoring, risk analysis, processing time estimation, and controlled data experimentation within a single deployed application, the dashboard enables informed decision-making for both technical and business stakeholders involved in retail supply chain operations.



📌 Navigation

Select View

- ● Operational Dashboard
- ○ DQ Score Distribution
- ○ Failure Risk Levels
- ○ Processing Time & SLA Trends
- ○ Data Lab (CSV / Synthetic)
- ○ About Project

🚀 EDI Predictive Analytics Dashboard

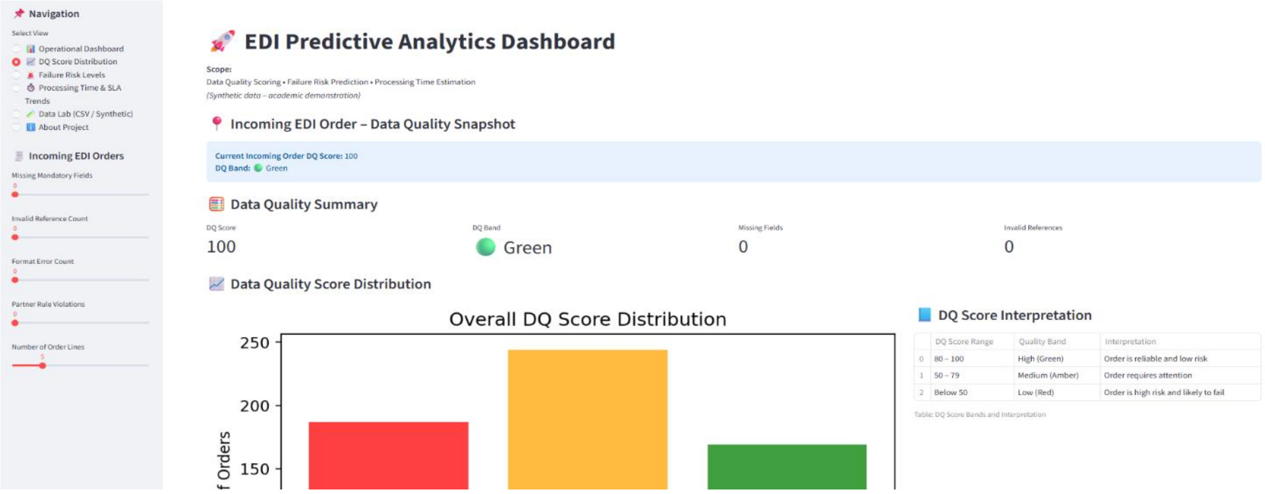
Scope:
Data Quality Scoring • Failure Risk Prediction • Processing Time Estimation
(Synthetic data – academic demonstration)

📊 Operational Dashboard

600	219	381	30.22
Total POs	Successful Orders	✕ Failed Orders	✕ Avg Processing Time (min)

Synthetic data used for academic demonstration only.

Figure 6.9.1 - Operational Dashboard



📌 Navigation

Select View

- ○ Operational Dashboard
- ● DQ Score Distribution
- ○ Failure Risk Levels
- ○ Processing Time & SLA Trends
- ○ Data Lab (CSV / Synthetic)
- ○ About Project

Incoming EDI Orders

Missing Mandatory Fields: 0

Invalid Reference Count: 0

Format Error Count: 0

Partner Rule Violations: 0

Number of Order Lines: 5

🚀 EDI Predictive Analytics Dashboard

Scope:
Data Quality Scoring • Failure Risk Prediction • Processing Time Estimation
(Synthetic data – academic demonstration)

📍 Incoming EDI Order – Data Quality Snapshot

Current Incoming Order DQ Score: 100
DQ Band: 🟢 Green

Data Quality Summary

DQ Score	DQ Band	Missing Fields	Invalid References
100	🟢 Green	0	0

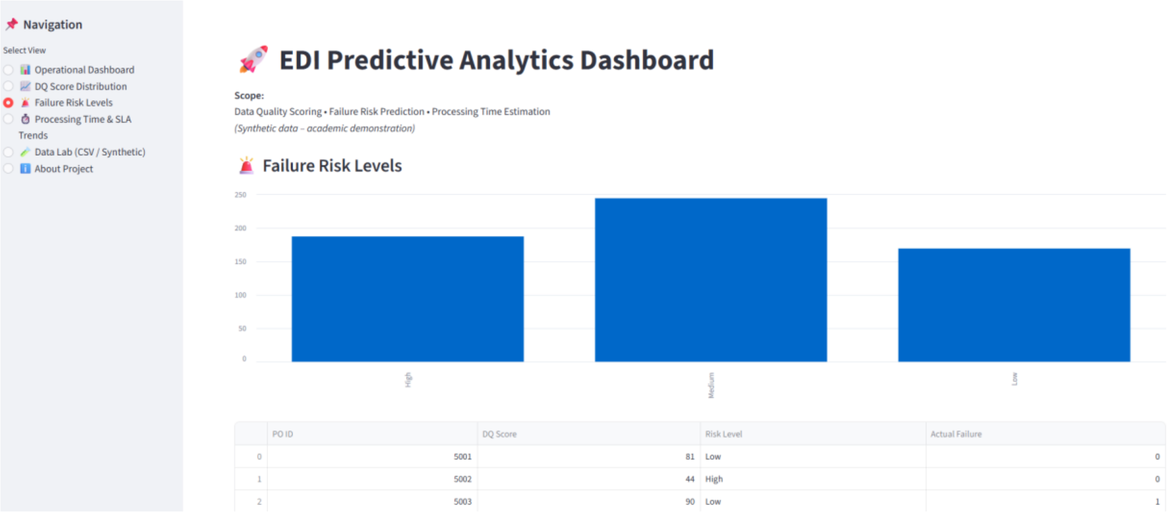
Data Quality Score Distribution

DQ Score Interpretation

	DQ Score Range	Quality Band	Interpretation
0	80 – 100	High (Green)	Order is reliable and low risk
1	50 – 79	Medium (Amber)	Order requires attention
2	Below 50	Low (Red)	Order is high risk and likely to fail

Table: DQ Score Bands and Interpretation

Figure 6.9.2 – DQ Score Distribution



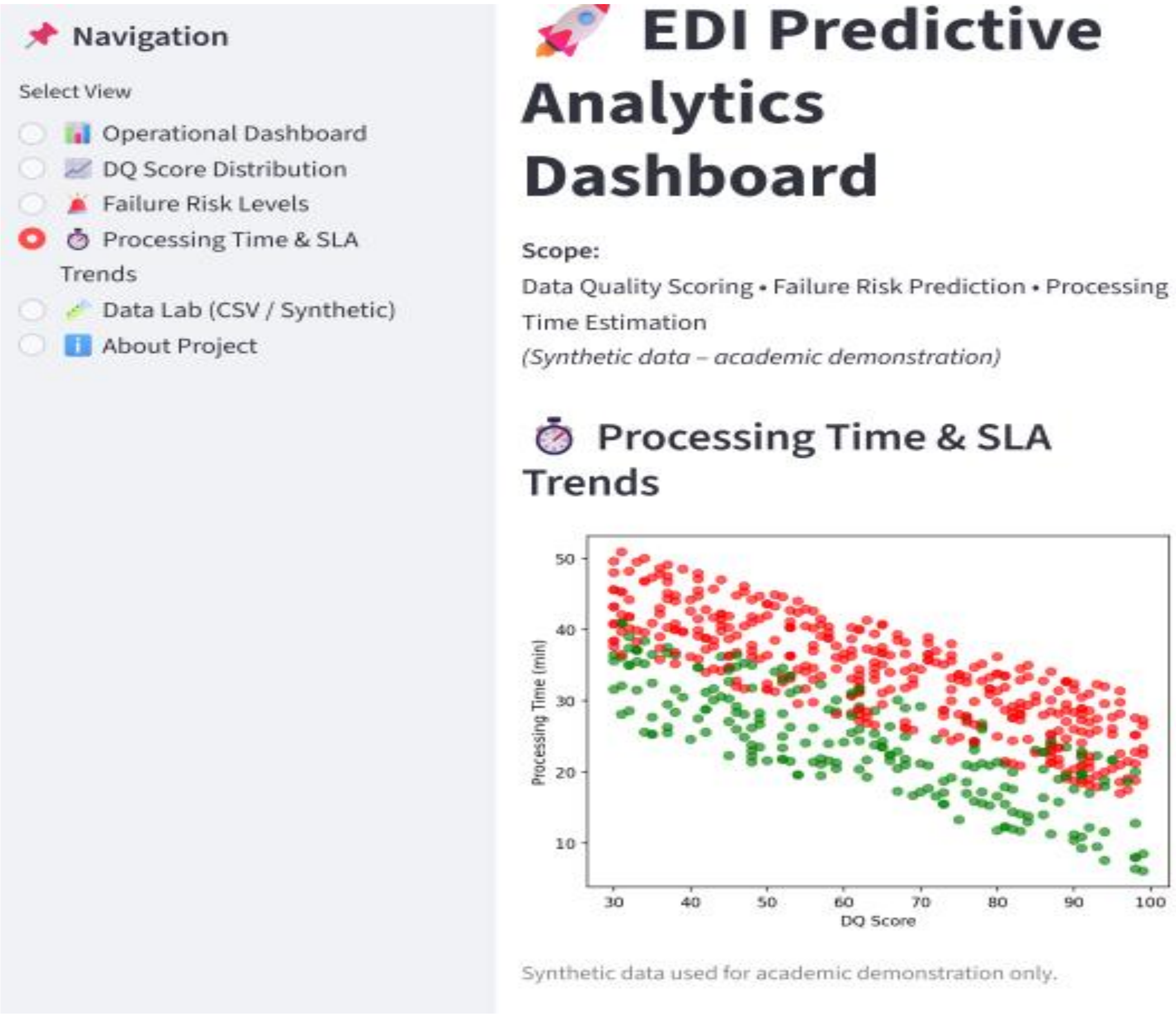
EDI Predictive Analytics Dashboard

Scope:
Data Quality Scoring • Failure Risk Prediction • Processing Time Estimation
(Synthetic data – academic demonstration)

Failure Risk Levels

	PO ID	DQ Score	Risk Level	Actual Failure
0	5001	81	Low	0
1	5002	44	High	0
2	5003	90	Low	1

Figure 6.9.3 – Failure Risk Levels



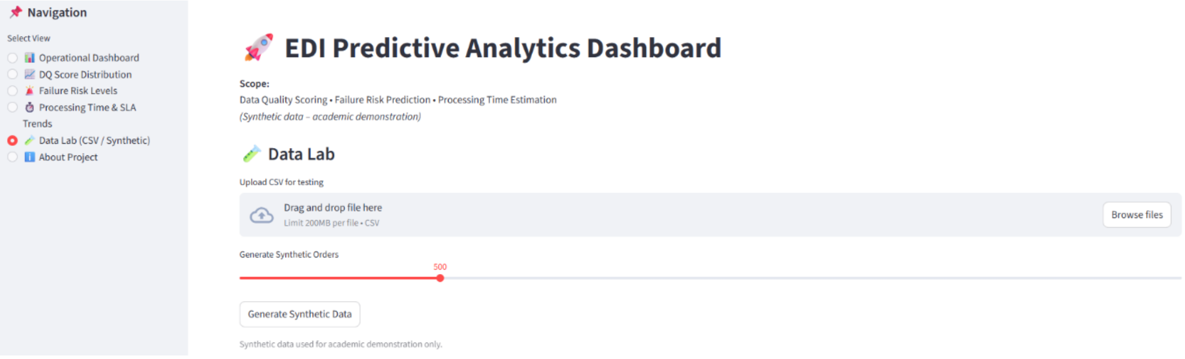
EDI Predictive Analytics Dashboard

Scope:
Data Quality Scoring • Failure Risk Prediction • Processing Time Estimation
(Synthetic data – academic demonstration)

Processing Time & SLA Trends

Synthetic data used for academic demonstration only.

Figure 6.9.4 – Processing Time & SLA Trends



EDI Predictive Analytics Dashboard

Scope:
Data Quality Scoring • Failure Risk Prediction • Processing Time Estimation
(Synthetic data – academic demonstration)

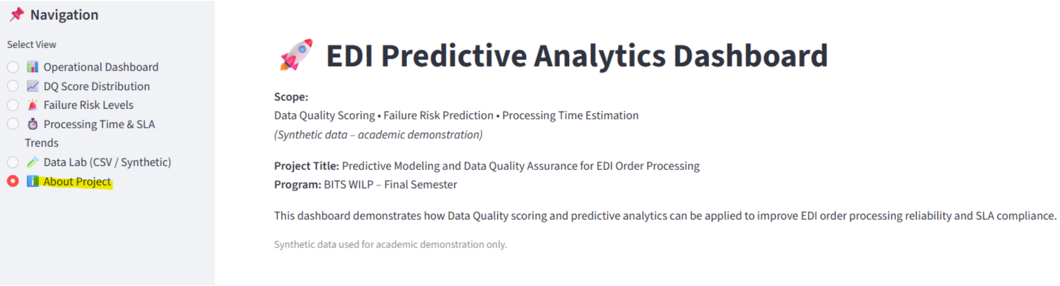
Data Lab

Upload CSV for testing

Generate Synthetic Orders: 500

Synthetic data used for academic demonstration only.

Figure 6.9.5 – Processing Time & SLA Trends



EDI Predictive Analytics Dashboard

Scope:
Data Quality Scoring • Failure Risk Prediction • Processing Time Estimation
(Synthetic data – academic demonstration)

Project Title: Predictive Modeling and Data Quality Assurance for EDI Order Processing
Program: BITS WILP – Final Semester

This dashboard demonstrates how Data Quality scoring and predictive analytics can be applied to improve EDI order processing reliability and SLA compliance.

Synthetic data used for academic demonstration only.

Figure 6.9.6 – Processing Time & SLA Trends

6.10 Performance and Scalability Considerations

The implementation is optimized for high-volume retail environments. Batch processing, efficient data access patterns, and model execution optimization are employed to ensure scalability.

The modular design allows independent scaling of analytics components without impacting Oracle EBS transaction processing performance.

6.11 Error Handling and Logging

Robust error handling mechanisms are implemented to capture and log exceptions occurring during validation, scoring, or model execution. Detailed logs support troubleshooting and continuous improvement of validation rules and models.

6.12 Implementation Summary

The implemented solution illustrates a practical realization of the proposed architecture, integrating data quality assessment, predictive modeling, and operational visualization into a cohesive framework. By maintaining a non-intrusive integration with Oracle EBS and leveraging synthetic data for evaluation, the implementation achieves a balance between analytical sophistication and enterprise-grade reliability.

Phases	Start Date-End Date	Work to be done	Status
Dissertation Outline	05 Nov 2025 – 21 NOV 2025	Literature Review and prepare Dissertation Outline	**COMPLETED**
Design & Development	22 Nov 2025 – 31 Dec 2025	Design & Development Activity	**COMPLETED**
Testing	31 Dec 2015 – 10 Jan 2026	Software Testing, User Evaluation & Conclusion	**COMPLETED**
Dissertation Review	10 Feb 2026- 28 JAN 2026	Submit Dissertation to Supervisor & Additional Examiner for review and feedback	**COMPLETED**
Submission	28 JAN 2026- 01 FEB 2026	Final Review and submission of Dissertation	COMPLETED

7. Results and Analysis

This chapter presents a detailed analysis of the experimental results obtained from implementing the proposed data quality–driven predictive framework. The evaluation focuses on assessing the effectiveness of data quality scoring, failure prediction models, and processing time estimation in improving EDI order processing reliability and operational visibility.

7.1 Experimental Setup

The evaluation is conducted using a synthetically generated dataset designed to closely replicate real-world EDI 850 order behaviour in a large-scale retail environment. The dataset includes a diverse range of order structures, data quality variations, and processing outcomes. This approach ensures compliance with confidentiality constraints while maintaining realism in failure patterns and processing delays.

The dataset is divided into training and testing subsets to enable unbiased performance evaluation. Separate experiments are conducted for failure prediction (classification) and processing time estimation (regression).

7.2 Data Quality Score Analysis

The computed Data Quality (DQ) scores exhibit a clear distribution across inbound orders. Orders with higher DQ scores consistently demonstrate smoother processing with minimal validation issues, whereas lower scores are associated with frequent failures and extended processing times.

Analysis reveals a strong inverse relationship between DQ score and order failure probability. As the DQ score decreases, the likelihood of encountering validation errors, reprocessing cycles, and manual intervention increases significantly. This observation validates the effectiveness of the proposed data quality framework in quantifying order reliability.

7.3 Failure Prediction Model Results

In addition to Logistic Regression and Random Forest classifiers, an advanced gradient boosting model based on XGBoost is evaluated to assess potential improvements in predictive performance. The comparison is conducted using identical training and testing datasets to ensure fairness and consistency.

The XGBoost classifier demonstrates superior capability in capturing complex, non-linear relationships between data quality indicators, order attributes, and processing outcomes. Compared to baseline models, XGBoost achieves higher recall and ROC-AUC scores, indicating improved identification of failed orders while maintaining controlled false positive rates.

The model's regularization mechanisms contribute to stable generalization performance, preventing overfitting despite the presence of noisy and imbalanced data. Feature importance analysis further reveals that composite Data Quality score, completeness score, and historical trading partner behavior are the most influential predictors of order failure.

From an operational perspective, the higher recall achieved by XGBoost is particularly valuable, as it minimizes the risk of undetected high-risk orders entering downstream processing. The probabilistic outputs produced by the model enable flexible threshold tuning based on business risk tolerance, allowing the organization to balance automation efficiency with exception control.

Overall, the experimental results indicate that XGBoost outperforms traditional ensemble and linear models for early-stage failure prediction in EDI-driven order processing environments, making it a strong candidate for enterprise deployment.

The classification models are evaluated using accuracy, precision, recall, and ROC-AUC metrics. Logistic Regression provides a strong baseline with interpretable coefficients that highlight the influence of individual data quality dimensions. Random Forest illustrates superior performance by capturing non-linear interactions among features.

Results indicate that recall is particularly high for failed orders, ensuring that most problematic orders are identified early. This outcome is critical from an operational perspective, as missing high-risk orders can lead to downstream disruptions.

The ROC-AUC scores further confirm the models' ability to effectively distinguish between successful and failed orders across varying thresholds.

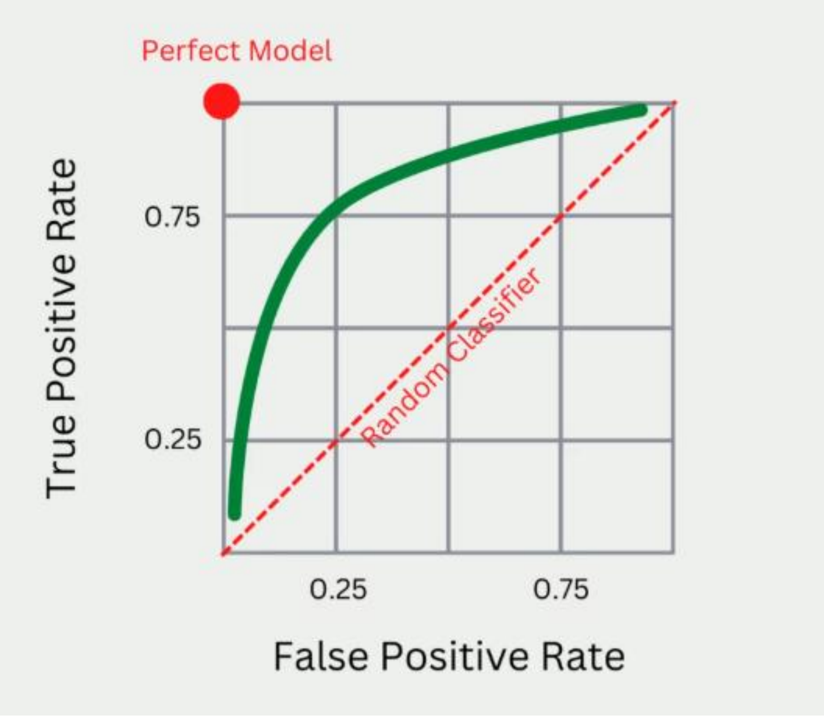


Figure 5 illustrates the Receiver Operating Characteristic (ROC) curve for the failure prediction model.

The curve demonstrates the model's ability to distinguish between successful and failed orders across different probability thresholds.
A higher Area Under the Curve (AUC) indicates strong discriminatory performance and validates the effectiveness of the selected feature set.
This analysis supports the use of recall-optimized thresholds for operational risk mitigation.

7.4 Processing Time Prediction Analysis

Regression model results show that predicted processing time closely aligns with observed processing durations. Orders with lower data quality scores and higher structural complexity consistently require longer processing times.
The Random Forest Regressor outperforms linear models by effectively modeling non-linear dependencies between data quality metrics, order attributes, and system load factors. Error metrics such as MAE and RMSE remain within acceptable operational tolerances, indicating reliable predictive capability.

7.5 Combined Risk Analysis

By combining failure probability and predicted processing time, orders are categorized into operational risk bands. This combined analysis provides a more comprehensive risk assessment than either metric alone. High-risk orders are characterized by both elevated failure probability and extended processing time, warranting proactive intervention.

7.6 Business Impact Assessment

The results demonstrate tangible operational benefits, including reduced order reprocessing, improved workload planning, and enhanced visibility into potential processing bottlenecks. The framework enables a shift from reactive exception handling to proactive risk management.

7.7 Model Comparison Summary

Table 4 presents a comparative summary of the predictive models evaluated in this study. The comparison highlights differences in interpretability, ability to capture non-linear relationships, handling of class imbalance, and overall predictive effectiveness.

Model	Strengths	Limitations	Overall Suitability
Logistic Regression	High interpretability, fast training, stable baseline	Limited ability to model non-linear interactions	Suitable as baseline and for explainability
Random Forest	Captures non-linear patterns, robust to noise	Less interpretable, higher computational cost	Effective for complex feature interactions
XGBoost	High recall and ROC-AUC, regularization, imbalance handling	Requires careful tuning	Most effective for early failure detection

The comparison indicates that while Logistic Regression provides transparency and Random Forest improves non-linear modeling, XGBoost delivers the strongest overall performance for proactive failure prediction in EDI order processing.

7.8 Summary of Results

Overall, the experimental results validate the effectiveness of integrating data quality assessment with predictive analytics. The models demonstrate consistent performance and provide actionable insights that align with real-world retail supply chain requirements.

8. Conclusion and Future Scope

8.1 Conclusion

This dissertation presented an end-to-end predictive analytics framework aimed at enhancing the reliability and efficiency of EDI order processing in retail supply chains. By integrating structured data quality validation with machine learning–based prediction, the proposed solution addresses a critical gap in traditional EDI integration approaches that rely heavily on reactive exception handling.

Among the evaluated predictive models, advanced ensemble techniques such as XGBoost demonstrated superior performance in identifying high-risk orders. The model's ability to capture non-linear relationships, combined with built-in regularization and probabilistic outputs, resulted in improved recall and overall discriminatory power when compared to baseline and bagging-based approaches.

The introduction of a composite Data Quality (DQ) score enables quantitative assessment of inbound order reliability prior to core transactional processing. When combined with XGBoost-based failure prediction and processing time estimation, the framework provides early visibility into operational risk and system workload.

Experimental results using synthetic datasets illustrate a strong correlation between data quality indicators and processing outcomes. The framework successfully identifies high-risk orders, supports informed decision-making, and improves operational transparency without disrupting standard Oracle E-Business Suite workflows.

Overall, the proposed approach demonstrates how advanced machine learning models such as XGBoost can be effectively embedded within enterprise integration landscapes to improve supply chain resilience and operational efficiency.

This dissertation presented a comprehensive predictive analytics framework aimed at enhancing the reliability and efficiency of EDI order processing in retail supply chains. By integrating structured data quality validation with machine learning–based prediction, the proposed solution addresses a critical gap in traditional EDI integration approaches that rely heavily on reactive exception handling.

The introduction of a composite Data Quality (DQ) score enables quantitative assessment of inbound order reliability prior to core transactional processing. Predictive models for failure risk and processing time further extend this capability by providing early visibility into operational complexity and potential disruptions.

Experimental results using synthetic datasets demonstrate a strong correlation between data quality indicators and processing outcomes. The framework successfully identifies high-risk orders, supports informed decision-making, and improves operational transparency without disrupting standard Oracle E-Business Suite workflows.

Overall, the proposed approach illustrates how data-driven intelligence can be embedded into enterprise integration landscapes to improve supply chain resilience and efficiency.

8.2 Future Scope

While the current implementation illustrates significant benefits, several opportunities exist for future enhancement:

- **Real-time Model Deployment:** Extending the framework to support real-time streaming EDI transactions for near-instant risk assessment.
- **Adaptive Rule Weighting:** Dynamically adjusting data quality weights based on evolving failure patterns and feedback loops.
- **Advanced Modeling Techniques:** Exploring deep learning or sequence-based models to capture temporal dependencies across orders.
- **Cross-Transaction Expansion:** Applying the framework to additional EDI transaction sets such as invoices and shipment notices.
- **Automated Remediation:** Integrating intelligent correction mechanisms to automatically resolve common data quality issues.

These enhancements can further strengthen the predictive capability and operational value of the proposed framework, making it applicable across broader enterprise integration scenarios.

9. References

1. Batini, C., & Scannapieco, M., *Data Quality: Concepts, Methodologies and Techniques*, Springer-Verlag, Berlin, 2006.
2. Redman, T. C., *Data Driven: Profiting from Your Most Important Business Asset*, Harvard Business Review Press, Boston, 2013.
3. Han, J., Kamber, M., & Pei, J., *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, 2011.
4. Kimball, R., & Ross, M., *The Data Warehouse Toolkit*, 3rd Edition, Wiley Publishing, 2013.
5. Oracle Corporation, *Oracle E-Business Suite Order Management User Guide*, Oracle Documentation Library.
6. Oracle Corporation, *Oracle E-Business Suite Integration Repository*, Oracle Documentation Library.
7. ANSI Accredited Standards Committee X12, *ANSI X12 Electronic Data Interchange Standards*, Washington DC.
8. ISO/IEC 25012:2008, *Data Quality Model*, International Organization for Standardization.
9. Provost, F., & Fawcett, T., *Data Science for Business*, O’Reilly Media, 2013.
10. Géron, A., *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, O’Reilly Media, 2019.

10. Appendices

Appendix A: Data Quality Validation Rules

This appendix outlines the primary validation rules implemented as part of the data quality framework for inbound EDI orders. These rules are designed based on commonly observed data issues in enterprise retail environments.

- Mandatory attribute completeness checks
- Domain and format validation for coded fields
- Crossfield consistency validations
- Referential integrity checks against master data

These validations collectively contribute to the computation of the composite Data Quality score.

Appendix B: Data Quality Score Calculation Example

This appendix illustrates a representative example of Data Quality score computation. Individual validation outcomes are assigned predefined weights based on business criticality. The weighted results are aggregated and normalized to generate a final score.

The example demonstrates how failures in critical fields significantly impact the overall score and influence risk classification.

Appendix C: Synthetic Dataset Description

To ensure confidentiality and compliance with organizational policies, synthetic datasets are used for model training and evaluation. These datasets replicate real-world EDI order characteristics, including error distributions, order complexity, and processing outcomes.

The dataset includes attributes related to data quality metrics, order structure, processing duration, and failure indicators.

Appendix D: Model Evaluation Metrics

This appendix summarizes the metrics used to evaluate predictive models:

- Classification metrics: Accuracy, Precision, Recall, ROC-AUC
- Regression metrics: MAE, RMSE, and R^2

These metrics provide a comprehensive assessment of predictive performance and operational reliability.

Appendix E: Dashboard and Reporting Views

This appendix describes representative dashboard views developed for operational monitoring. Visualizations include data quality distributions, failure risk heatmaps, and processing time trends that support proactive decision-making.