Model Training

Predictive Modeling & Risk Stratification

4.1 Objective

The objective of model training was to transform the validated **Data Quality Index (DQI)** into an **actionable, predictive risk stratification framework**, capable of identifying subjects requiring operational escalation. This phase focused on:

1. Defining a defensible risk stratification strategy
2. Engineering leakage-free predictive features
3. Training interpretable machine learning models
4. Validating that learned patterns align with clinical operations logic

---

4.2 Final Risk Stratification Strategy

4.2.1 Evaluation of Multi-Tier Risk Strategies

Three independent strategies were evaluated to derive risk tiers:

- **Distribution-based (percentile) stratification**
- **Outcome-aligned stratification using issue burden**
- **Clean-anchored stratification using internal quality flags**

All three strategies consistently demonstrated that a **stable and interpretable three-tier risk structure was not supported** by the data. Specifically:

- The final DQI exhibited **discrete ceiling effects** among clean subjects, collapsing percentile thresholds.
- Outcome-aligned metrics (issue counts) were **not universally defined** across all subjects.
- Clean-anchored thresholds cleanly separated high-risk subjects but did not support a meaningful intermediate tier.

4.2.2 Final Decision: Binary Risk Stratification

Based on convergent evidence, a **two-tier risk strategy** was adopted:

Risk Tier	Definition
High Risk (1)	`subject_data_quality_score ≤ 0.8`
Not High Risk (0)	`subject_data_quality_score > 0.8`

This binary formulation was selected to:

- Preserve semantic correctness
- Avoid artificial label construction
- Align with real clinical escalation workflows

This risk label was **locked prior to feature engineering** and used as the sole modeling target.

4.3 Feature Engineering

4.3.1 Design Principles

Feature engineering was guided by the following constraints:

- **No leakage** from DQI or downstream composite scores
- Use only **primitive, pre-aggregation operational signals**
- Favor interpretability over complexity

Composite variables such as `base_dqi`, `completeness_score`, `timeliness_score`, and `is_clean_subject_flag` were explicitly excluded from modeling.

4.3.2 Final Feature Set

The final feature matrix consisted of five categories:

Issue Burden

- `total_open_issue_count_per_subject`
- `edrr_open_issues_count`
- `open_sae_count`
- `pending_sae_dm_review_flag`
- `pending_sae_safety_review_flag`

Missingness Signals

- `missing_visits_count`

- `missing_pages_count`
- `missing_lab_issues_count`
- `lab_ranges_missing_flag`

Timeliness Extremes

- `max_days_visit_overdue`
- `max_days_page_missing`

Structural Integrity

- `inactivated_forms_count`
- `inactivated_records_flag`

Completeness Ratios

- `visit_completeness`
- `page_completeness`
- `lab_completeness`

Invariant features (e.g., `consistency_flag`) were excluded due to zero variance, as consistency had already been enforced during dataset construction.

4.4 Predictive Modeling

4.4.1 Primary Model — Logistic Regression

A **regularized Logistic Regression model** was selected as the primary predictive model due to its:

- Interpretability
- Stability under class imbalance
- Suitability for regulatory and operational environments

Class imbalance was handled using balanced class weights, and feature standardization was applied.

Performance Summary

- **ROC–AUC:** ~0.9999
- **Recall (High Risk):** 0.993
- **Precision (High Risk):** 0.987

- **False Negatives:** 1 subject
- **False Positives:** 2 subjects

This performance profile strongly prioritizes **escalation safety**, minimizing missed high-risk cases.

4.4.2 Coefficient Interpretation

The strongest positive predictors of High Risk status were:

- Open issue burden (`total_open_issue_count_per_subject` , `edrr_open_issues_count`)
- Severe timeliness delays (`max_days_visit_overdue`)
- Pending safety and data management SAE reviews

Protective signals included visit and page completeness, with smaller effect sizes.

The direction and magnitude of coefficients were fully aligned with:

- Phase-2 DQI construction logic
- Clinical operations intuition
- Regulatory escalation principles

4.5 Structural Validation — Decision Tree

To validate that the Logistic Regression model was not exploiting spurious correlations, a **shallow, regularized Decision Tree** was trained as a secondary model.

Purpose

- Not performance optimization
- Structural and logical validation

Key Findings

- **ROC–AUC:** ~0.997
- **False Negatives:** 1
- **False Positives:** 0

The tree relied almost exclusively on three features:

1. `pending_sae_safety_review_flag`
2. `total_open_issue_count_per_subject`
3. `max_days_visit_overdue`

The learned decision rules corresponded to intuitive escalation logic, such as:

- Immediate escalation for unresolved safety reviews
- Escalation driven by issue burden
- Severe visit delays triggering high risk classification

This confirmed that the Logistic Regression model's performance was driven by **genuine operational signals**, not model artifacts.

4.6 Calibration Considerations

Although the Logistic Regression model demonstrated excellent ranking performance, probability calibration showed **over-confident predictions** near 0 and 1. This behavior is expected given the strong separability of the feature space and the presence of near-deterministic escalation signals.

Accordingly, predicted probabilities are best interpreted as **risk scores**, not literal likelihood estimates, unless post-hoc calibration is explicitly applied.

Table: Comparison of Predictive Models Used

Aspect	**Logistic Regression (Primary Model)**	**Decision Tree (Validation Model)**
Model Role	**Primary production model**	**Structural validation & interpretability model**
Objective	Predict High Risk status with maximum recall and stable generalization	Verify that escalation logic can be rediscovered via simple rules
Target Variable	Binary High Risk label (`1 = High Risk`)	Same binary High Risk label
Feature Set	Identical, leakage-free primitive operational signals	Identical feature set

Aspect	Logistic Regression (Primary Model)	Decision Tree (Validation Model)
Handling of Class Imbalance	Balanced class weights	Balanced class weights
Model Complexity	Linear decision boundary	Shallow rule-based structure (`max_depth = 4`)
ROC–AUC	**0.9999**	**0.9967**
Accuracy	0.998	0.999
Recall (High Risk)	**0.993**	**0.993**
Precision (High Risk)	0.987	**1.000**
False Negatives (High Risk)	1	1
False Positives (High Risk)	2	0
Probability Output	Continuous probabilities (over-confident near extremes)	Hard class decisions only
Calibration Behavior	Excellent ranking, over-confident probabilities	Not probabilistic
Primary Risk Drivers Identified	Issue burden, timeliness delays, SAE review flags	Same three dominant drivers
Feature Importance Behavior	Distributed across multiple related features	Sparse; only 3 features used
Interpretability	Global, coefficient-based interpretation	Explicit decision rules
Robustness to Noise	High (regularization, linearity)	Lower (rule rigidity)
Primary Value Added	Accurate, stable, explainable predictions	Independent confirmation of escalation logic
Failure Mode if Used Alone	Over-confident probability interpretation	Over-simplified decision boundaries

---

Decision Tree as a Validation Model

The Decision Tree was intentionally trained under strict constraints (limited depth, minimum leaf size) to force it to learn **only the strongest and most obvious escalation rules**. Its purpose was **not** to maximize predictive performance, but to answer the following question:

> *Can a simple rule-based model independently rediscover the same escalation logic learned by the Logistic Regression model?*

The answer was **yes**.

Specifically, the tree:

- Split first on open issue burden
- Escalated immediately on pending safety reviews
- Triggered high risk classification for severe visit delays

These rules mirror:

- DQI penalty logic
- Logistic Regression coefficient rankings
- Clinical operations intuition

4.7 Final Modeling Conclusion

- The data supports a **binary escalation decision**, not a multi-tier risk gradient.
- Logistic Regression provides an **interpretable and stable.**
- Decision Tree analysis independently validates the learned risk structure.
- Additional complex models (e.g., ensembles) are unlikely to yield meaningful performance gains and are therefore unnecessary at this stage.

> **The modeling phase confirms that the DQI framework successfully encodes clinically meaningful escalation logic, and machine learning serves to formalize—not obscure—that logic.**

> **Logistic Regression provides the optimal balance of performance and stability for deployment, while the Decision Tree independently validates that the model's predictions are driven by clinically meaningful escalation rules rather than statistical artifacts.**
