Project Methodology: Aircraft Hub Inspection Predictive Modeling

Executive Summary

This document outlines the comprehensive methodology employed in developing a predictive model for aircraft wheel hub crack detection. The project demonstrates a systematic approach to handling real-world, imbalanced datasets in a safety-critical aviation environment, with emphasis on strategic decision-making and analytical rigor. The complete analytical framework is illustrated in **Figure 1**, which will guide through each critical decision point in the methodology.

1. Problem Definition & Strategic Objectives

1.1 Business Context

Industry: Aviation Maintenance, Repair & Overhaul (MRO)
Critical Need: Early detection of structural defects in aircraft wheel hubs
Safety Imperative: Undetected cracks can lead to catastrophic wheel failure during critical flight phases

1.2 Project Goals

Primary: Develop ML model to predict crack occurrence with high recall (minimize false negatives)
Secondary: Provide interpretable insights for maintenance decision-making
Portfolio: Demonstrate data science capabilities for career transition from NDT Engineering

1.3 Success Criteria

- **Recall ≥ 70%:** Detect majority of actual cracks to ensure flight safety
- **Model Interpretability:** Clear feature importance for operational insights
- **Business Value:** Actionable recommendations for maintenance optimization

The problem definition phase is shown at the top of **Figure 1**, establishing the foundation for all subsequent analytical decisions.

2. Data Collection & Quality Assurance Strategy

2.1 Data Sources & Validation Framework

Primary Source: Handwritten inspection logbooks (Books 1-4)
Inspection Method: Eddy Current Testing (ET)
Operational Period: 5.6 years of continuous operations
Initial Volume: 8,870 inspection records

Digital Transformation Strategy: The transition from handwritten logs to analysis-ready datasets required robust validation protocols including dropdown lists for categorical consistency, conditional formatting for data type validation, cross-referencing with serial numbers, and manual verification of safety-critical fields.

2.2 Data Quality Assessment & Recovery

Challenge Identified: 26% missing data rate in original records
Recovery Strategy: Systematic imputation using cross-referencing techniques
Success Rate: 74% of missing values successfully recovered
Final Data Loss: 0.25% (1 record out of 386 for focused dataset)

This high recovery rate was critical for maintaining statistical power while preserving data integrity. As depicted in **Figure 1**, the quality assessment decision point (Missing Data >26%?) led to the recovery strategy implementation rather than record deletion, preserving valuable safety data.

2.3 Privacy & Confidentiality Protocol

Anonymization Strategy: Generic coding systems implemented for part numbers, inspector identifications, and temporal references while preserving analytical relationships essential for modeling.

3. Strategic Data Scoping Decision

3.1 Initial Challenge Assessment

Original Crack Rate: 0.43% (38 cracks in 8,870 records)
Class Imbalance Severity: 99.57% vs 0.43%
Modeling Feasibility: Extremely challenging with available sample size

3.2 Data-Driven Focusing Strategy

Analytical Approach:

1. Systematic calculation of crack rates across all hub types
2. Risk-based prioritization identifying Type 05 MW as highest-risk category
3. Strategic decision to focus modeling efforts on actionable subset

Strategic Outcome:

- **Type 05 MW Crack Rate:** 18.24% within category
- **Final Dataset:** 385 records with 14.03% overall crack rate
- **Rationale:** Optimal balance between sufficient positive cases and real-world applicability

This decision exemplifies data-driven problem scoping, prioritizing analytical feasibility while maintaining business relevance. **Figure 1** illustrates this critical decision point where the severe class imbalance, 0.43% crack rate triggered the focus strategy, resulting in the Type 05 MW subset with improved 14.03% crack rate

4. Exploratory Data Analysis Framework

4.1 Multi-Dimensional Analysis Strategy

Power BI Dashboard Development: Comprehensive visualization strategy encompassing hub type comparison, temporal trend analysis, inspector performance evaluation, part number reliability assessment, and hub cycle deterioration patterns.

4.2 Statistical Hypothesis Testing Framework

Methodological Approach:

- Point-biserial correlation for continuous-binary variable relationships
- Cramer's V for categorical association testing
- Chi-square contingency analysis for independence verification

Key Statistical Findings:

- Hub Cycle vs Crack: $r = 0.067$, $p = 0.19$ (weak, non-significant linear relationship)
- Inspector vs Crack: Cramer's $V \approx 0$ (no significant systematic association)
- Total Inspections vs Crack: $r = 0.0895$, $p = 0.079$ (marginally non-significant)

These results indicated that traditional linear relationships were insufficient, supporting the need for more sophisticated modeling approaches.

5. Feature Engineering & Preprocessing Strategy

5.1 Feature Engineering Philosophy

Categorical Treatment: One-hot encoding with drop-first strategy to prevent multicollinearity while preserving categorical information integrity.

Numerical Scaling: StandardScaler implementation for continuous variables to ensure algorithmic convergence and feature equality.

Target Variable: Binary encoding preserving safety-critical distinction between crack and no-crack outcomes.

5.2 Train-Test Split Strategy

Methodology: 80-20 split with stratification to maintain class distribution across training and testing sets, ensuring representative evaluation while maximizing training data availability.

Reproducibility: Fixed random state implementation for consistent model comparison and validation.

6. Machine Learning Model Development Strategy

6.1 Progressive Complexity Approach

Phase 1 - Baseline Establishment: Logistic Regression implementation to establish performance floor and validate data preprocessing pipeline. Complete failure (0% recall) confirmed the severity of class imbalance challenge.

Phase 2 - Tree-Based Enhancement: Random Forest with balanced class weighting demonstrated initial capability to predict positive class, achieving 45% recall and establishing model learning potential.

Phase 3 - Advanced Technique Integration: Systematic exploration of SMOTE (Synthetic Minority Oversampling Technique) and XGBoost implementation with weighted learning optimization.

6.2 Hyperparameter Optimization Strategy

XGBoost Configuration Rationale:

- Moderate tree depth (3) to prevent overfitting with limited dataset
- Conservative learning rate (0.1) for stable convergence
- AUCPR evaluation metric optimized for imbalanced classification
- Scale_pos_weight calculation based on class distribution for penalty optimization

Figure 1 shows the iterative model development process with the key success criterion (Recall ≥70%) as a decision point that determines whether to proceed to interpretation or return to model refinement.

7. Model Evaluation Framework

7.1 Metrics Selection Rationale

Primary Focus: Recall prioritization reflecting safety-critical nature where missed cracks (false negatives) carry significantly higher risk than false alarms.

Balanced Assessment: Precision, F1-Score, and ROC AUC included for comprehensive performance understanding while acknowledging recall primacy.

Business Alignment: Confusion matrix analysis providing detailed error type breakdown for operational decision-making.

7.2 Safety-First Evaluation Philosophy

Cost-Benefit Framework: Explicit acknowledgment that maintenance costs from false positives are acceptable trade-offs against catastrophic failure risks from missed cracks.

8. Model Interpretability & Business Intelligence

8.1 SHAP Implementation Strategy

Analytical Framework: Global feature importance ranking, individual prediction explanations, feature value impact distribution analysis, and specific case waterfall plots for comprehensive model understanding.

8.2 Operational Intelligence Generation

Business Value Creation:

- Cycle-based maintenance scheduling recommendations
- Part number reliability profiling for procurement decisions
- Inspector workload optimization for resource allocation
- Temporal pattern recognition for operational planning

9. Results Analysis & Validation

9.1 Optimal Model Performance

Weighted XGBoost Achievement:

- **82% Recall:** 9 out of 11 cracks successfully detected
- **21% Precision:** Acceptable trade-off for safety-focused application
- **Business Impact:** Significant reduction in undetected crack risk

9.2 Feature Importance Validation

Engineering Alignment: Top predictive features (Hub Cycle, Total Inspections, Part Number effects) align with domain expertise, validating model logical consistency and business applicability.

10. Future Enhancement Strategy

10.1 Technical Development Roadmap

Advanced feature engineering opportunities including temporal lag features and interaction terms, ensemble method development for improved robustness, and real-time prediction pipeline architecture for operational deployment.

10.2 Business Integration Pathway

Maintenance scheduling system integration, cost-benefit optimization modeling, extended data collection strategies, and cross-platform deployment considerations for enterprise implementation.

Conclusion: This methodology demonstrates systematic application of data science principles to safety-critical engineering challenges, emphasizing strategic decision-making, statistical rigor, and business value creation while maintaining transparency and reproducibility throughout the analytical process.

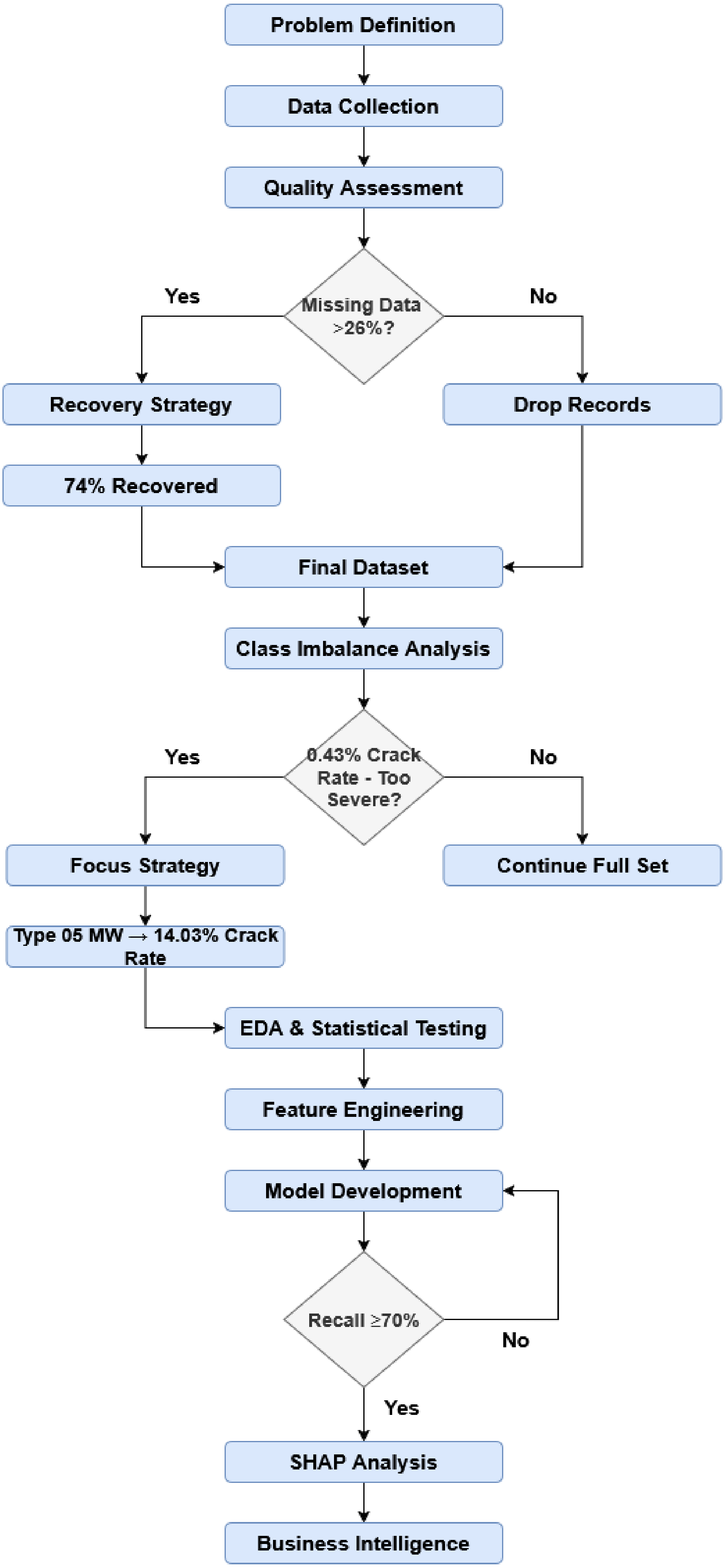

Figure 1: Analytical Framework