Nathan Weber

DDS 8500: Principles of Data Science

Assignment 1: CRISP-DM Paper

A Data Science Approach to Predictive Maintenance and Asset Risk Prioritization in Electric Utility Protection Systems Using the CRISP-DM Framework

National University

December 2025

Abstract

Electric utilities operate complex, asset-intensive systems in which failures of protection equipment can lead to service interruptions, equipment damage, and safety risks. As utility infrastructure ages and operational complexity increases, data-driven decision support systems offer opportunities to improve maintenance planning and asset risk prioritization. This paper proposes a high-level data science application for predictive maintenance and asset risk prioritization within electric utility protection systems, using the CRISP-DM methodology. The proposed approach integrates asset inventory data, maintenance records, inspection histories, and environmental factors to estimate relative failure risk and support informed maintenance and replacement decisions. Emphasis is placed on interpretability, robustness, and alignment with engineering judgment rather than automation of protection design. The analysis demonstrates how data science can complement protection engineering by enhancing situational awareness, supporting iterative decision-making, and improving long-term system reliability.

1. Introduction

Data science has emerged as a multidisciplinary field that integrates mathematics, statistics, programming, advanced analytics, machine learning, and domain expertise to extract actionable insights from data and support decision-making (IBM, 2014). Unlike traditional data analytics, which often focuses on descriptive or retrospective analysis, data science encompasses a broader life cycle that includes data acquisition, preparation, modeling, evaluation, and iterative refinement. This life-cycle-oriented approach is particularly relevant in operational environments where decisions must balance reliability, safety, cost, and uncertainty.

Electric utilities represent a compelling application domain for data science due to their reliance on long-lived physical assets and their responsibility to deliver reliable service under regulatory and operational constraints. Within these systems, protection equipment—such as relays, breakers, and associated control assets—plays a critical role in isolating faults and preventing cascading failures. While protection engineering relies on rigorous analysis and established engineering standards, maintenance and replacement decisions are often constrained by incomplete data, aging infrastructure, and competing priorities.

This paper explores how data science techniques can support predictive maintenance and asset risk prioritization for electric utility protection systems. Using the CRISP-DM methodology, the proposed application emphasizes decision support rather than automation of engineering judgment. By integrating historical asset data, maintenance records, and environmental factors, the approach seeks to enhance situational awareness, prioritize inspections and maintenance activities, and inform long-term asset management strategies.

To ground this application in an operational context, the following sections apply the CRISP-DM framework, beginning with a detailed examination of the business problem, stakeholders, and decision environment associated with protection system maintenance in electric utilities.

2. Business Understanding (CRISP-DM Step 1)

2.1 Problem Definition

Electric utilities operate large fleets of long-lived physical assets that are critical to system reliability and public safety. Within these systems, protection equipment—including protective relays, circuit breakers, and associated control components—plays a central role in detecting abnormal conditions and isolating faults. Failures or degradation of protection assets can contribute to extended outages, equipment damage, and increased operational risk.

Many electric utilities face challenges related to aging infrastructure, constrained maintenance resources, and incomplete historical data. Traditional maintenance strategies often rely on fixed inspection schedules, reactive repairs, or expert judgment informed by limited visibility across the full asset population. As a result, maintenance and replacement decisions may not consistently reflect relative risk or evolving asset conditions.

The problem addressed in this application is how to support more consistent, data-informed maintenance and asset prioritization decisions for utility protection systems. Rather than automating protection design or replacing engineering judgment, the goal is to enhance situational awareness and support planning decisions through structured analysis of historical and operational data.

2.2 Stakeholders and Requirements

Key stakeholders involved in this application include protection engineers, operations and maintenance teams, asset management and planning teams, utility leadership, regulators, and customers. Stakeholder requirements emphasize transparency, interpretability, and alignment with established engineering practices. Outputs must support decision-making without introducing undue complexity or obscuring underlying assumptions.

2.3 Business Goals

The primary business goals of this data science application are to improve system reliability by reducing the likelihood of protection equipment failures; support proactive maintenance and inspection planning; prioritize assets based on relative risk rather than fixed schedules alone; and inform long-term replacement and capital planning decisions. Success is measured not solely by predictive accuracy, but by the usefulness of outputs in operational planning and decision-making contexts.

2.4 Decision Context and Use of Results

The results of the proposed application are intended to support decisions such as which protection assets should be inspected or tested first; where maintenance resources should be allocated; and which assets warrant closer monitoring or accelerated replacement. Model outputs are designed to be advisory and contextual, providing ranked risk indicators or health scores that complement engineering expertise. Decisions remain the responsibility of engineers and planners, with data science serving as a decision-support tool rather than an automated decision-maker.

With the business objectives and decision context established, the next step is to examine the nature, sources, and limitations of the data available to support these decisions.

3. Data Understanding (CRISP-DM Step 2)

3.1 Nature of Available Data

Electric utilities typically maintain a variety of operational and asset-related datasets that can support predictive maintenance and risk assessment. These datasets may include asset inventories, maintenance records, inspection logs, outage and fault histories, and environmental data related to asset location. The data is predominantly structured, with a combination of numerical, categorical, and time-based attributes. Asset lifecycles often span decades, resulting in long time horizons and evolving data quality over time.

3.2 Data Sources and Collection Methods

Potential data sources for this application include asset management systems containing equipment specifications and installation dates; maintenance management systems documenting inspections, testing, and repairs; outage management systems recording fault events and service interruptions; and environmental datasets capturing temperature, weather exposure, or geographic factors. Data is typically collected as part of routine operations, compliance reporting, and maintenance activities rather than for analytical purposes, which influences its completeness and consistency.

3.3 Dataset Size and Sufficiency

Utility datasets often contain records for thousands of assets across extended time periods. While the volume of data may be substantial, sufficiency depends on the completeness and consistency of records, particularly for failure events, inspections, and condition assessments. In some cases, failure events may be relatively rare, resulting in imbalanced datasets that require careful handling during modeling and evaluation.

3.4 Data Quality Issues and Limitations

Common data quality challenges include missing or incomplete inspection records; inconsistent asset identifiers across systems; changes in data collection practices over time; and legacy systems with limited integration. These issues must be addressed during preprocessing and acknowledged as limitations in model interpretation.

3.5 Biases and Constraints

Potential biases may arise from historical prioritization practices influencing which assets received more attention; underreporting of minor faults or degradation; and environmental or geographic factors correlated with asset age. Recognizing these biases is essential to ensure that model outputs are interpreted appropriately and do not reinforce existing blind spots.

Understanding these data characteristics, quality issues, and potential biases informs how the data must be prepared, integrated, and managed prior to analysis. These considerations are addressed in the following section.

4. Data Preprocessing, Storage, and Integration (CRISP-DM Step 3)

4.1 Data Storage and Management

Electric utilities typically manage asset and operational data across multiple systems, including asset management platforms, maintenance management systems, and outage management systems. For a data science application supporting predictive maintenance, these data sources must be consolidated into a secure, centralized analytical environment. Data storage considerations include access control, auditability, and long-term retention, particularly given the regulatory and safety-critical nature of utility operations. Maintaining clear data lineage and versioning supports transparency and enables review of historical analyses as system conditions evolve.

4.2 Data Preprocessing Techniques

Preprocessing is a critical step in preparing utility data for analysis, as operational datasets are often incomplete or inconsistent. Common preprocessing techniques for this application include identification and treatment of missing values in inspection or maintenance records; normalization of numerical attributes such as asset age or time since last maintenance; and detection and handling of outliers that may reflect data entry errors rather than true conditions. These steps help ensure that downstream analyses reflect underlying asset behavior rather than artifacts of data collection.

4.3 Data Integration Challenges

Integrating data across systems presents several challenges, including inconsistent asset identifiers, differing data formats, and misaligned timestamps. Addressing these challenges requires careful reconciliation of records and explicit documentation of assumptions made during integration. Given the long operational lifecycles of utility assets, changes in data collection practices over time must also be considered. Acknowledging these challenges is essential for interpreting model outputs responsibly.

5. Data Preparation (CRISP-DM Step 4)

5.1 Feature Selection

Feature selection focuses on identifying variables that plausibly reflect asset condition and operational risk. For protection system assets, relevant features may include asset age and installation date; frequency and type of maintenance activities; historical fault or outage associations; and environmental exposure factors such as temperature extremes or location. The selection of features is guided by domain knowledge and literature rather than purely statistical criteria.

5.2 Dataset Splitting

Datasets are split into training, validation, and testing subsets to support reliable model development. This structure helps reduce overfitting and provides a basis for evaluating generalization to unseen data. Where event outcomes are rare, stratified sampling or time-based splits may be used to preserve outcome representation and operational realism.

5.3 Scaling and Transformation

Scaling and transformation may be applied where appropriate to improve model performance and interpretability. Numerical features may be normalized or standardized, and categorical variables may be encoded using established methods suitable for the selected models.

6. Feature Engineering (CRISP-DM Step 5)

6.1 Engineered Features

Feature engineering enhances raw data by constructing variables that better capture underlying patterns. Examples include aggregated maintenance indicators (e.g., number of inspections over a defined period), time-based features representing degradation trends, and composite asset health indices combining multiple attributes. These engineered features support more meaningful comparisons across assets while remaining interpretable to stakeholders.

6.2 Handling Categorical Variables

Categorical variables such as asset type, manufacturer, or location are encoded using standard approaches appropriate for the modeling technique. Care is taken to avoid introducing unnecessary complexity or obscuring interpretability, particularly in a decision-support context.

7. Modeling Technique and Validation (CRISP-DM Steps 6–7)

7.1 Modeling Approach

The modeling approach prioritizes interpretability, robustness, and alignment with operational decision-making. Rather than deploying complex or opaque models, this application considers established techniques suitable for risk ranking and classification. Candidate models include logistic regression as a transparent baseline and tree-based methods that can capture non-linear relationships while remaining interpretable. The purpose of modeling is to estimate relative risk and support prioritization, not to produce deterministic predictions of failure.

7.2 Model Training and Tuning

Models are trained using historical data and evaluated through standard validation techniques. Hyperparameter tuning is conducted conservatively to avoid overfitting and ensure that performance generalizes beyond the training data. Where failure events are rare, class imbalance strategies such as resampling, class weighting, or threshold tuning may be used to improve sensitivity to high-risk outcomes.

7.3 Validation Techniques

Validation focuses on assessing model robustness and reliability rather than maximizing performance metrics alone. Techniques may include cross-validation to evaluate consistency across data subsets; sensitivity analysis to assess the impact of key features; and review of model outputs against known historical outcomes. These steps support confidence in the model's usefulness as a decision-support tool.

7.4 Limitations and Risk Considerations

Model limitations are explicitly acknowledged, including dependence on historical data quality and the potential for unobserved factors to influence outcomes. These limitations reinforce the role of engineering judgment and underscore the importance of using model outputs as guidance rather than directives.

8. Evaluation of the Model (CRISP-DM Step 8)

8.1 Evaluation Metrics

Model evaluation must reflect the operational context in which results are used. In electric utility protection systems, the consequences of missed high-risk assets can be significant, while unnecessary maintenance also carries cost and resource implications. As a result, evaluation focuses on metrics that balance sensitivity to risk with practical decision-making needs. Common evaluation metrics include precision, recall, and area under the receiver operating characteristic curve (ROC-AUC). Recall is particularly important in identifying assets at elevated risk, as false negatives may correspond to missed opportunities for preventative maintenance. Precision provides insight into the proportion of flagged assets that truly warrant attention, supporting efficient allocation of limited maintenance resources. In addition to classification metrics, ranking-based evaluations assess the model's ability to prioritize assets relative to one another. Ranking stability across validation sets is considered an important indicator of robustness and operational reliability.

8.2 Interpretation of Results

Model outputs are interpreted as relative indicators of risk rather than deterministic predictions of failure. Risk scores or rankings are intended to support planning conversations and prioritization decisions rather than to prescribe specific engineering actions. Interpreting results in collaboration with protection engineers and operations personnel helps ensure that outputs align with domain knowledge and operational realities. Assets flagged as high risk may warrant further inspection, testing, or review rather than immediate replacement. Evaluation emphasizes whether outputs are useful for decision-making rather than whether performance metrics are maximized in isolation.

8.3 Operational Relevance

The practical value of the application is assessed by its ability to inform maintenance scheduling, inspection prioritization, and long-term asset planning. Evaluation therefore considers how outputs integrate into existing workflows and whether they provide actionable insight to stakeholders. This decision-centric framing aligns with the broader goal of data science to support informed decision-making rather than to generate standalone analytical results.

9. Triangulation of Results (CRISP-DM Step 9)

9.1 Triangulation with Historical Decisions

To assess the reliability of model outputs, results can be triangulated against historical maintenance and inspection decisions. Comparing model-generated risk rankings with past prioritization practices helps identify areas of agreement as well as potential blind spots. Discrepancies are not necessarily indicative of error; they may reflect evolving asset conditions, changes in operating environments, or limitations in historical data. These comparisons provide opportunities for learning and refinement rather than simple validation or rejection of model results.

9.2 Comparison with Alternative Approaches

Triangulation also involves comparing model outputs with alternative methods, such as rule-based heuristics or expert judgment. In utility contexts, these traditional approaches are often well-established and grounded in engineering experience. By examining where data-driven insights align with or diverge from expert assessments, stakeholders can better understand the strengths and limitations of each approach. This comparison supports a balanced view in which data science complements, rather than replaces, established engineering practices.

9.3 Generalizability and Limitations

The generalizability of the application depends on the availability and quality of data, as well as on organizational and environmental factors specific to each utility. Differences in asset types, maintenance practices, and regulatory environments may influence model performance and applicability. Acknowledging these limitations is essential to responsible deployment. The application is intended to be adapted and iteratively refined within specific operational contexts.

9.4 Iterative Improvement

Consistent with the dynamic data science life cycle, triangulation results inform subsequent iterations of the application. Feedback from engineers, maintenance teams, and asset managers can guide adjustments to feature selection, modeling approaches, and evaluation criteria. This iterative process supports continuous improvement in asset management and reliability outcomes.

10. Conclusion

This paper proposed a high-level data science application for predictive maintenance and asset risk prioritization in electric utility protection systems using the CRISP-DM framework. By framing the problem as a decision-support capability rather than an automation effort, the approach emphasizes interpretability, robustness, and alignment with engineering judgment. The proposed workflow integrates common utility data sources, addresses quality and bias considerations, and applies conservative modeling and validation strategies suitable for operational contexts. Evaluation and triangulation reinforce the practical goal of supporting maintenance prioritization and planning decisions. Finally, the application reflects an iterative life cycle in which insights and feedback inform continuous refinement, helping utilities improve reliability and manage risk over time.

References

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

Griffin, P., Khadake, J., LeMay, K., Lewis, S., Orchard, S., Pask, A., Pope, B., Roessner, U., Russell, K., Seemann, T., Treloar, A., Tyagi, S., Christiansen, J., Dayalan, S., Gladman, S., Hangartner, S., Hayden, H., Ho, W., Keeble-Gagnère, G., & Schneider, M. (2018). Best practice data life cycle approaches for the life sciences. F1000Research, 6, 1618. https://doi.org/10.12688/f1000research.12344.2

IBM. (2014). What is data science? https://www.ibm.com/topics/data-science

Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied Artificial Intelligence, 17(5–6), 375–381. https://doi.org/10.1080/713827180