Fixing data quality issues in CMDB in IT and OT by using machine learning algorithms. Forecasting for IT procurement

Author: Szymon Niewiadomski

Date: 19 września 2025

Summary

Configuration Management Databases (CMDBs) serve as the reference repository for ITSM practices, including, among others, change management as well as capacity and performance management. Their data quality directly affects service continuity and an organization's cyber-resilience. This dissertation presents a transparent, modular, and on-premises machine-learning architecture aimed at systematically improving CMDB data quality while meeting the security, auditability, and compliance requirements characteristic of critical-infrastructure operators. The research was carried out as an industrial doctorate in cooperation with an energy-sector enterprise, under conditions in which processing remains local, models are not authorized to modify production data autonomously, and expert-verification throughput is constrained by an operational budget.

Methodologically, a monthly data-quality evaluation pipeline is developed that *nominates* records with elevated non-conformity risk for expert review. Textual, numerical, and categorical (dictionary) attributes are represented in extensive, automatically generated feature spaces (including n-gram counts and one-hot encoding), followed by threshold filtering (variance/frequency) and density-based discriminative scoring. Recursive kernel density estimators with bandwidth adaptation and a rollback mechanism are proposed to enable non-parametric learning in the presence of feedback and label uncertainty. Feature selection is guided by criteria of independence and error-detection capability. Entropy-based Rajski/Jaccard measures are employed. A four-layer neural classifier integrates the evidence provided by individual features to produce a ranked list of records from most uncertain to most trustworthy. Such a list can be used to plan inspections and audits. Complementary modules detect numerical and structural anomalies: distance-based methods (full all-pairs variant and k-NN) and structure-aware checks using a graph representation of the CMDB, including identifier-driven relationship validation (a case study of VIN→manufacturer). The proposed system implements a human-in-the-loop paradigm: every model recommendation is fully auditable; verified decisions feed adaptive updates; and inspection volume and cadence are aligned with available resources.

In empirical evaluation, identifier models achieve high accuracy at moderate computational cost, enabling reliable structural-consistency tests. Precision–throughput curves confirm effective prioritization of erroneous records under varying inspection cut-offs. Experiments show that the pipeline captures a broad spectrum of non-conformities (duplicates, missing unique identifiers, typographical errors, field misuse, unfinished operational procedures, and lifecycle mislabels) without relying on brittle rule sets. Architectural choices—on-premises deployment, low computational requirements, explainable scoring, and offline traceability—support security and compliance needs and facilitate knowledge transfer across related CMDBs within data governance.

A second contribution provides a forecasting-and-optimization approach to IT procurement planning. The dissertation formulates a cost-minimization problem with predictive

models for prices and demand and proposes a genetic algorithm (GA) to solve the resulting complex, non-linear problem. A bespoke method for crossover and mutation of IT resource acquisition strategies is introduced. Experimental results reveal emergent strategies (e.g., order consolidation versus last-minute decisions) under differing market conditions and budget/logistics constraints. Recommendations are also provided for monitoring market trends, predicting delivery times, tracking the evolution of licensing models, and assessing the implications of cloud-service adoption for procurement policy.

In summary, mathematically grounded and transparent ML methods measurably improve CMDB data quality and support more efficient procurement planning under public-procurement regimes.

Keywords: ITSM; ITIL; data governance; CMDB data quality management; data cleaning; graph representation; mutual information; Rajski distance; Jaccard entropy; kernel density estimation; outlier detection; interpretable neural networks; demand and price forecasting; IT procurement optimization; genetic algorithm.