Advanced Data Quality Profiling and Anomaly Detection

1. Introduction

The objective of this project was to evaluate the quality of a sensor dataset, identify anomalies through statistical techniques, address data quality issues such as outliers, and provide actionable recommendations to enhance data reliability.

Sensor data is frequently impacted by noise, spikes, and drift resulting from hardware limitations or environmental factors. Ensuring high data quality is essential prior to utilizing such data in monitoring or decision-making systems.

This project emphasizes data profiling and explainable statistical anomaly detection, rather than predictive modeling.

2. Dataset Overview

The dataset comprises time-series readings from multiple sensors. Key attributes include:

- Time: Timestamps of sensor readings.
- Temperature: Ambient temperature values.
- Humidity: Environmental humidity levels.
- Light: Light intensity readings.
- Loudness: Sound intensity measurements.

Initial profiling revealed a complete dataset with no missing values. One column (Air Quality) contained constant values and was excluded from analysis, as it offered no analytical value.

3. Data Profiling Summary

Comprehensive data profiling was conducted to assess the dataset's structure and quality. The following checks were performed:

- Data types and structure validation.
- Missing value analysis.
- Uniqueness analysis.

- Statistical summary of numerical features.

Key observations include:

- No missing values, indicating strong data completeness.
- Sensor readings exhibited reasonable variability.
- One non-informative column (constant values) was identified and removed.

This step ensured that only relevant and reliable features were utilized for anomaly detection.

4. Anomaly Detection Approach

Two statistical methods were employed to identify anomalous sensor readings:

4.1 Z-Score Method

The Z-Score method was applied to detect extreme anomalies, flagging records where values deviated significantly from the mean (Z-Score > 3).

- Detected anomalies: Approximately 3–4% of records.
- Interpretation: Indicates sudden sensor spikes or abnormal operating conditions.

4.2 IQR (Interquartile Range) Method

The IQR method was used to identify distribution-based anomalies, capturing values outside the normal operating range.

- Detected anomalies: Higher count than Z-Score.
- Interpretation: Reflects gradual sensor drift or localized abnormal patterns.

Employing both methods provided a comprehensive view of global and local anomalies.

5. Outlier Handling Strategy

Rather than removing anomalous records, capping (winsorization) was implemented based on IQR limits.

Rationale:

- Deletion risks eliminating meaningful real-world events.

- Capping mitigates the impact of extreme values while preserving trends.

This approach supports safer downstream analysis and monitoring systems. Post-treatment, statistical distributions were more stable and realistic, maintaining data integrity.

6. Cleaned Dataset Output

Following profiling, anomaly detection, and outlier handling, the cleaned dataset was exported as a CSV file:

- File name: cleaned_sensor_readings.csv

This dataset is now suitable for further analysis, visualization, or integration into monitoring pipelines.

7. Corrective Actions & Recommendations

Based on identified data quality issues and anomaly patterns, the following recommendations are proposed:

- **Regular Sensor Calibration**: Conduct periodic calibration to mitigate drift and ensure consistent readings.
- **Real-Time Anomaly Alerts**: Implement threshold-based alerts for immediate detection of extreme values.
- **Data Validation at Ingestion**: Enforce range checks and validation rules during data collection to prevent invalid entries.
- **Anomaly Logging Instead of Deletion**: Log anomalies for review, as they may signify important operational events.
- **Continuous Data Quality Monitoring**: Perform ongoing profiling to sustain long-term data reliability.

8. Conclusion

This project illustrates a structured methodology for data quality assessment and anomaly detection using explainable statistical techniques. Through thorough profiling, dual anomaly detection methods, and prudent outlier handling, the dataset was refined into a reliable, analysis-ready format.

The approach aligns with practical data engineering and analytics practices and can be readily adapted to real-world sensor monitoring systems.