Data Analysis on Cyclone Preheater Data

For Alog8.ai

Data Processing

Removing Duplicates

- Ensures no redundant readings in the dataset.
- Prevents misleading patterns due to repeated entries.

Handling Non-Numeric Data

- Converts all values to numeric, coercing invalid entries to NaN.
- Standardizes the dataset for anomaly detection models.

Feature Scaling (Normalization)

- Interpolation: Fills missing timestamps while preserving trends in cyclone preheater data.
- Why? Time series continuity is crucial for accurate anomaly detection.

Exporting Cleaned Data

- Used MinMaxScaler to scale all sensor readings between 0 and 1.
- Why? Ensures uniform weightage for different parameters (temperature, pressure)..

Exploratory data Analysis Insights

SEASONAL DECOMPOSE

- No clear seasonal pattern → Use trend-based/statistical models.
- High residual variation → External factors influencing data.
- Non-stationary data → Apply differencing or smoothing.

CORRELATION MATRIX

- Temprature reading are strongly Correlated
- Strong positive correlations between

ADFULLER TEST

- Strongly stationary
- Null Hypothesis Rejected
- No differencing needed

LOF MODEL, WHY?

- Works on Multivariate Data – Ideal for datasets with multiple interdependent features like temperature, gas flow, and pressure.
- Handles Non-Linear Patterns – Unlike traditional statistical methods, LOF can capture non-linear anomalies.
- Adaptive to Data Distribution – It adjusts based on local neighborhoods, reducing false positives in varying conditions.
- No Strict Assumptions – Unlike parametric models, LOF doesn’t assume normal distribution, making it versatile.

LOF MODEL, insights?

- After Finetuning the model, the anomaly points can be seen in the lof.ipynb
- The anomaly points can be seen and data fits well on the model
- There is scope of further fine-tuning on some wrongly classified data points

IsolationForest MODEL, WHY?

- Fast and Scalable – Uses tree-based partitioning, making it faster than density-based methods like LOF
- Non-Parametric Model – Does not assume any specific data distribution, making it versatile.
- Handles Multivariate Data – Suitable for detecting anomalies in complex interdependent variables like gas temperature, material flow, and pressure.
- Robust to Noise – Performs well even when some noise is present in the data.
- Unsupervised Learning – Does not require labeled anomaly data, making it ideal for real-world applications.

IsolationForest MODEL, insight?

- The model failed to fit well on the provided data even after fine tuning it in multiple ways, the best results that I could get are kept in the assignment
- Data distribution might not be suitable for the model
- Probably data is too clustered
- All in all this model is not a good fit for the particular task

Thank you

Yojit Ahuja