ANOMALY DETECTION

Detecting the Anomalous Activity of a Ship's Engine

Presented by – Poonam Vyas

Date – 15th September 2025

Table of Contents

1. Problem Statement

Ship engines operate under complex conditions, and even minor anomalies in parameters such as engine RPM, oil pressure, fuel pressure, coolant pressure, and temperatures can signal potential mechanical issues or future failures. Detecting these anomalies early is crucial for preventive maintenance, operational safety, and minimizing downtime.

The challenge is to identify unusual patterns or outliers in engine data within the expected 1–5% anomaly range, using robust statistical and machine learning methods that balance accuracy with interpretability.

2. Methodology

Dataset

The ship engine data consists of 6 features (columns): Engine RPM, Lubrication oil pressure, Fuel pressure, Coolant pressure, Lubrication oil temperature and Coolant temperature. These features contain overall 19535 observations.

Data Preparation

The dataset was structured into a Data Frame. Initial exploration included:

- **Histograms** to understand the distribution of values and highlight potential skewness or heavy tails.
- **Descriptive statistics** (mean, median, quartiles, 95th percentile) to summarize central tendency and spread.
- **Boxplots** for each feature to visualize the range and detect extreme values.

Anomaly Detection Methods

Method	Type	Approach	Parameters
IQR	Statistical	Applied the Interquartile Range rule to identify outliers per feature. We also evaluated anomalies triggered when multiple features exceeded the outlier threshold simultaneously.	No. Of features that contain outliers
One-Class SVM	Machine Learning	An unsupervised method based on kernel functions. Dimensionality reduction via PCA enabled 2D scatterplot visualization of results.	nu, gamma
Isolation Forest	Machine Learning	A tree-based algorithm, controlling the proportion of anomalies detected	Number of estimators, contamination

Approach

We tested parameter variations for each method while targeting an expected anomaly range of 1-5%. Finally, we compared overlaps in anomalies detected across methods to assess consistency.

3. Analysis

IQR Method

- For each feature, I calculated Q1, Q3, and IQR = Q3 – Q1.
- Observations outside the bounds (Q1 – 1.5*IQR, Q3 + 1.5*IQR) were flagged as anomalies.
- To align with the business definition of anomalies (1–5%), I focused on the **percentage of observations where at least 2 features simultaneously satisfied the outlier condition**.
- This yielded **422 anomalies (≈ 2.16%)**, which lies neatly within the expected range.

One-Class SVM

- Applied after scaling the dataset with **Standard Scaler**.
- Explored different parameters:
- **nu-values:** 0.01, 0.03, 0.05
- **gamma-values:** 0.5, 0.3, 0.1
- The results indicated that **gamma changes had little impact**, while **nu values slightly shifted the anomaly count**.
- Two main configurations (nu=0.05, gamma=0.1 vs nu=0.03, gamma=0.5) showed almost identical scatter plots after PCA, suggesting stability in detection.
- Result: **978 anomalies (~5%)**.

Isolation Forest

- Tested using different parameters:
- Estimators = 100, 200
- Contamination = 0.01, 0.03, 0.05
- Increasing estimators had no meaningful effect, while contamination directly scaled the anomaly count.
- Results:
- Contamination=0.01 → 196 anomalies (~1%)
- Contamination=0.03 → 587 anomalies (~3%)
- Contamination=0.05 → 977 anomalies (~5%)
- Scatter plots confirmed that anomaly clusters grew visibly larger with higher contamination.

Overlap of Methods

- IQR (422) vs One-Class SVM (978): **233 overlaps**
- IQR (422) vs Isolation Forest (977): **315 overlaps**
- One-Class SVM (978) vs Isolation Forest (977): **617 overlaps**

This shows strong agreement between **One-Class SVM and Isolation Forest**, while IQR, being rule-based, identified fewer anomalies with only partial overlap.

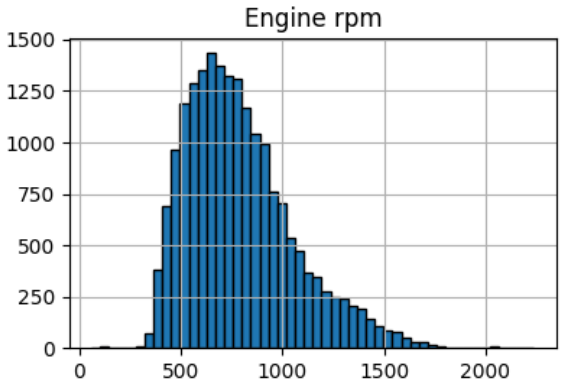
4. Insights

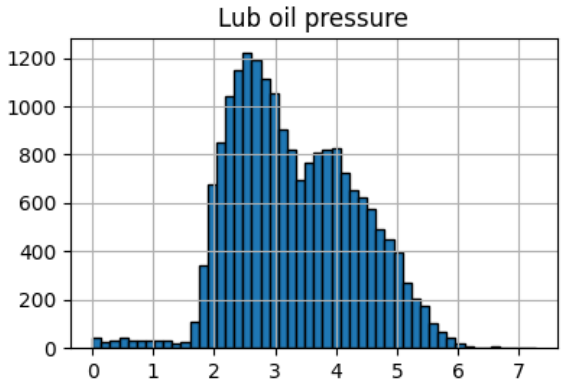
- **IQR** is a simple and interpretable method but limited in capturing complex anomaly patterns across multiple features.
- **One-Class SVM** provided consistent anomaly detection results regardless of minor parameter changes, but its visual clusters lacked strong separation.
- **Isolation Forest** emerged as the most flexible method, allowing precise tuning of anomaly proportions and producing visually clear anomaly groups.
- **Overlap analysis** indicates that anomalies consistently detected across methods (e.g., SVM and Isolation Forest) are likely true anomalies worth further monitoring.
- **2D visualizations (PCA scatterplots and boxplots)** were effective in conveying outlier behavior to non-technical stakeholders, though they reduce dimensional detail.

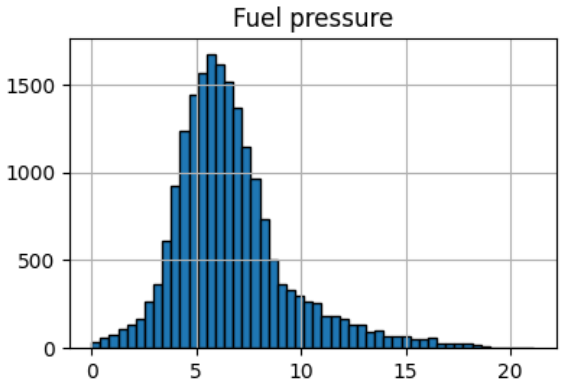
5. Visualizations

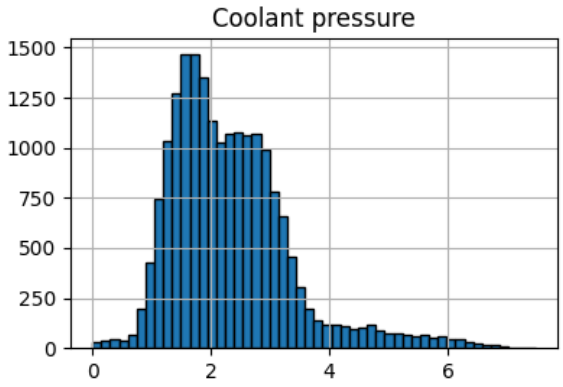
Histograms & Boxplots

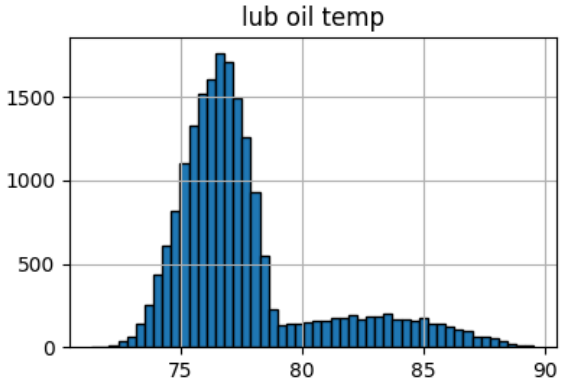
Histograms gave a first glance at data behavior for each feature independently, helping us identify which variables were stable, which were prone to fluctuations, and where anomalies might occur.

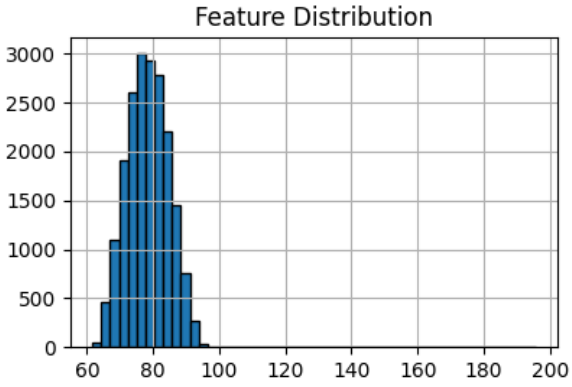




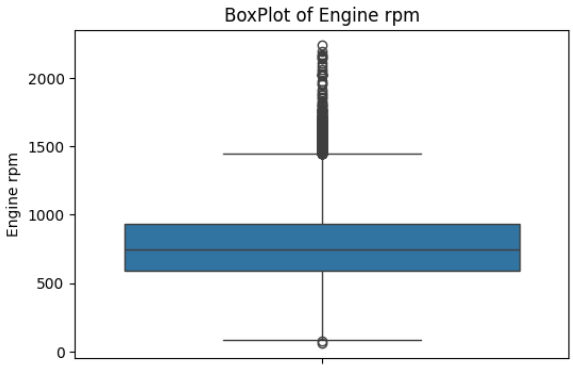


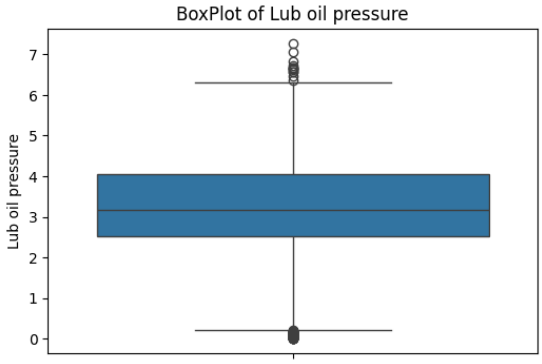


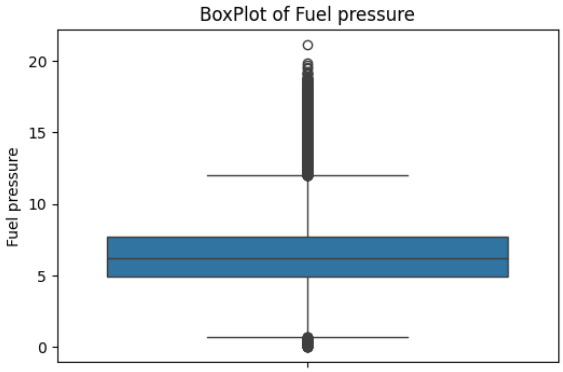


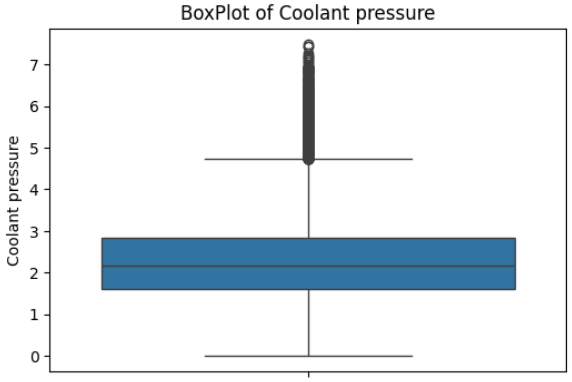


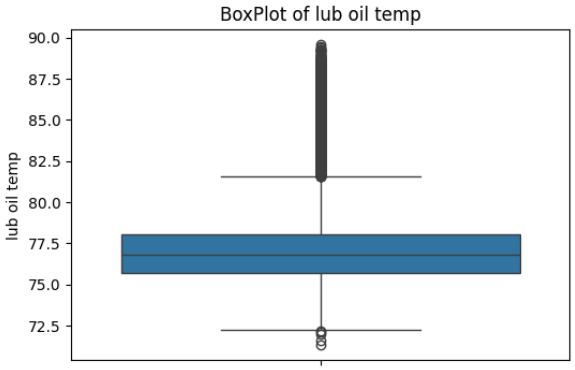
Boxplots provided feature-level outlier identification, making it easy to see extreme values.

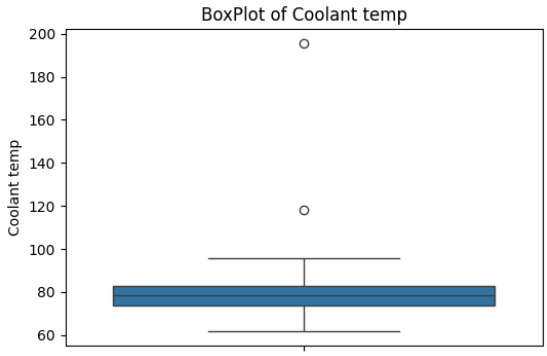












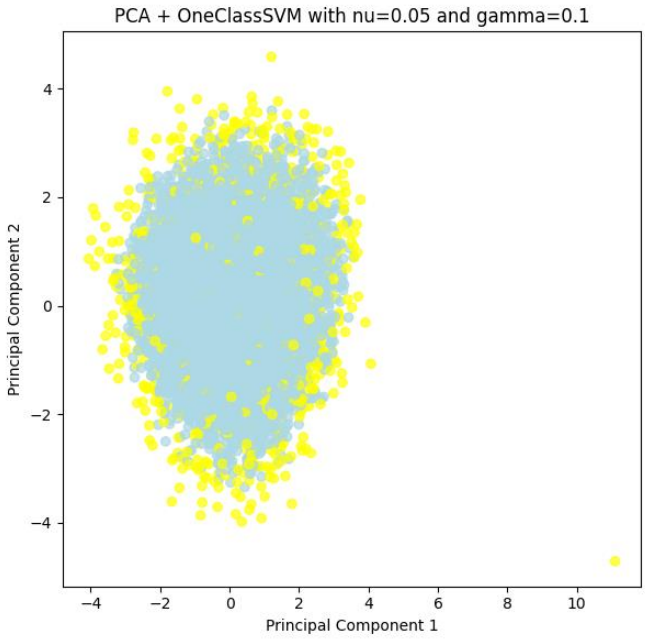
PCA Scatterplots

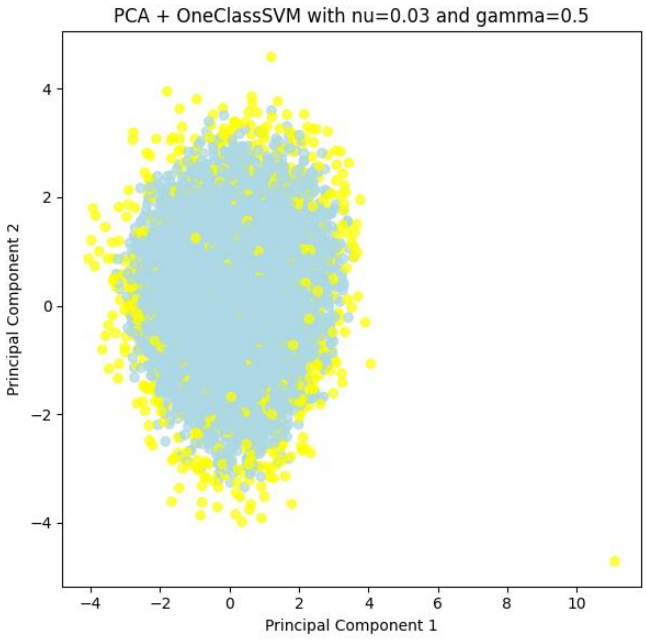
2D visualizations, especially PCA-based scatter plots were highly effective in:

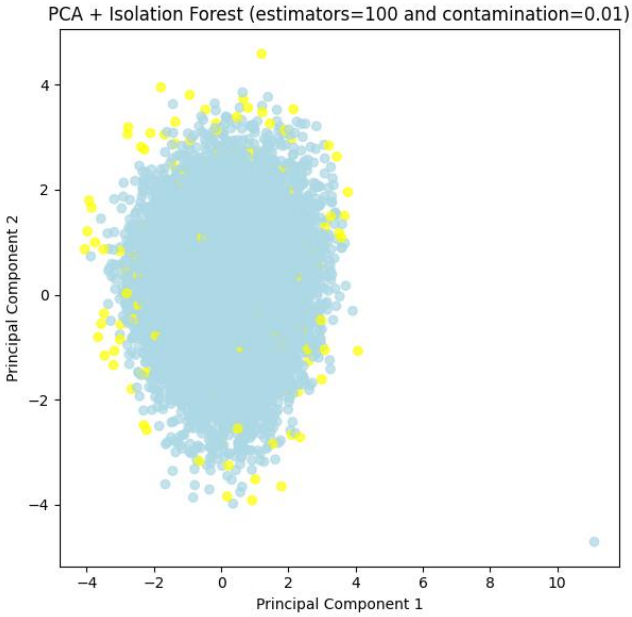
- Highlighting separation between normal and anomalous data points.
- Comparing parameter impacts across different models
- Communicating results to non-technical stakeholders, who could easily see “clusters” of anomalies.

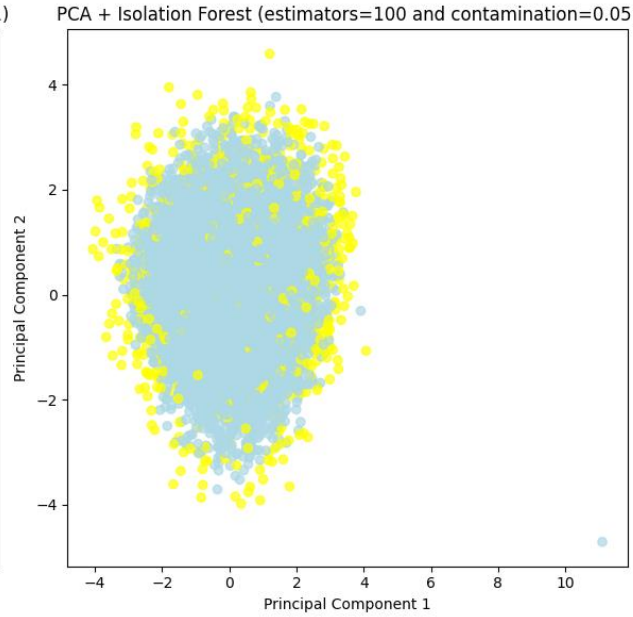
However, limitations exist:

- Reducing six dimensions to two inevitably loses some information
- Some anomalies may not appear distinct in 2D even if they are statistically significant in higher dimensions
- Visualization is therefore best used as supporting tool, not the sole method of validation.






PCA + Isolation Forest (estimators=100 and contamination=0.01)
Principal Component 2
Principal Component 1
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10


PCA + Isolation Forest (estimators=100 and contamination=0.05)
Principal Component 2
Principal Component 1
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10

6. Conclusion

When comparing statistical and machine learning methods, the results suggest:

- **IQR** provides simple and interpretable thresholds for individual features (univariate anomalies), it struggles to capture complex relationships across features.
- **One-Class SVM** offered moderate stability but limited separation in results.
- **Isolation Forest** emerged as an effective technique, offering flexibility in tuning anomaly proportions and providing clear separation between normal and anomalous observations in multivariate settings.
- From practical perspective, the company should use a **multivariant approach** for predictive maintenance, since anomalies often arise from the combined behaviour of multiple features rather than single thresholds.