ANOMALY DETECTION IN SHIP ENGINE DATA: A COMPREHENSIVE ANALYSIS

Contents

1. Introduction

Ship engines play a critical role in ensuring the reliability and efficiency of maritime operations. Anomalies in engine performance can result in costly downtime, safety hazards, and delayed deliveries, impacting both revenue and customer satisfaction. This report explores anomaly detection methods to identify early indicators of engine malfunction based on six key features: Engine RPM, Lubrication Oil Pressure, Fuel Pressure, Coolant Pressure, Lubrication Oil Temperature, and Coolant Temperature.

The primary objective of this analysis is to develop a robust anomaly detection system to monitor engine functionality, reduce risks, and provide actionable insights to the business. The report documents the approach used, highlights key observations from the data, compares anomaly detection methods, and provides recommendations for optimizing maintenance strategies.

2. Data Exploration and Preprocessing

Data Overview

The dataset comprises 19,535 observations across six continuous features. Preliminary data checks revealed no missing or duplicate values, ensuring the integrity of the dataset for analysis. Feature scaling was performed to normalize the data, enabling the effective application of machine learning models.

Insights from Exploratory Data Analysis (EDA)

1. **Shape and Distribution**:

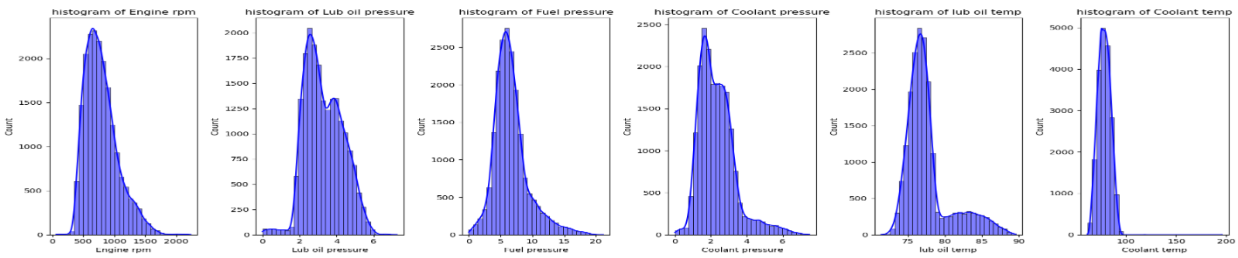


Figure 1 Engine Features Histogram

- All features exhibited slight right-skewness, except for Lubrication Oil Pressure, which appeared nearly normal.
- Boxplots revealed significant outliers for most features, with exceptions being Coolant Temperature and Coolant Pressure, which showed relatively stable behavior.

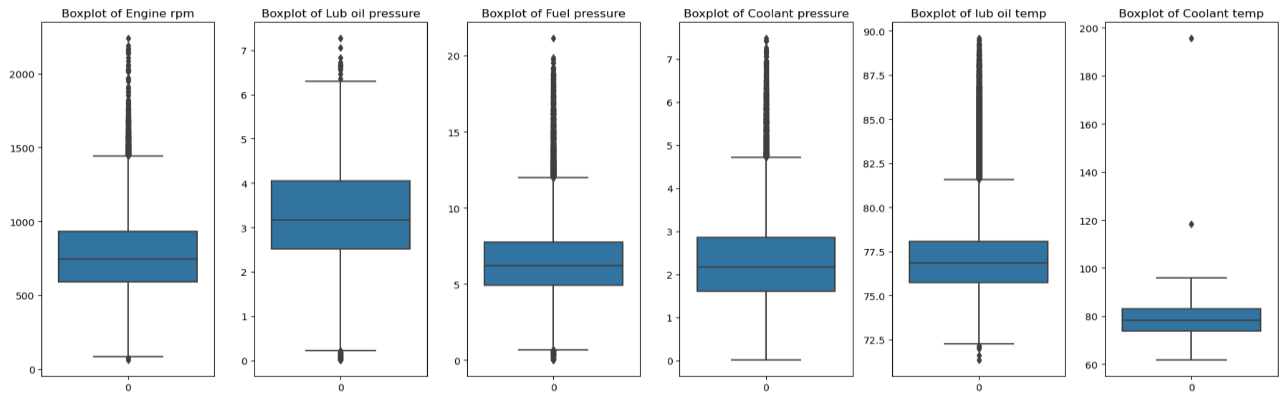


Figure 2 Engine Features Boxplot

2. **95th Percentile Values**:
   - Engine RPM values exceeding 1324 and Lubrication Oil Pressure values over 5.058 were identified as outliers.
   - Recommendations: Set monitoring thresholds at 1300 RPM and 4.5 Lubrication Oil Pressure for pre-emptive alerts, as confirmed through consultations with ship engineers.

3. Anomaly Detection: Interquartile Range (IQR) Method

Approach

The IQR method was used to identify outliers for each feature by computing the lower and upper bounds:

- Lower Bound = Q1 - 1.5 × IQR

- Upper Bound = Q3 + 1.5 × IQR

A binary column was created for each feature, indicating whether a value was an outlier. A sample was classified as an anomaly only if at least two features marked it as an outlier.

Findings

1. **Overall Anomaly Rate**:
   - The IQR method flagged 2.16% of data points as anomalies, consistent with the expected 1-5% range.
2. **Combination of Features**:
   - The combination of **Lubrication Oil Temperature** and **Fuel Pressure** produced the highest number of anomalies (152 instances).
   - The next most frequent combination was **Coolant Pressure** and **Lubrication Oil Temperature**.
   - The recurring appearance of **Lubrication Oil Temperature** highlights that it may play a critical role in engine performance and failures.

Insights and Recommendations:

- Given the consistent involvement of Lubrication Oil Temperature in anomalies, further investigation is necessary. Collaborate with ship engineers to understand its influence on engine reliability.

4. Anomaly Detection Using Machine Learning

4.1 One-Class SVM

1. **Model Performance**:
   - Initial settings (gamma=0.5, nu=0.05) resulted in an outlier percentage above the desired range.

- After parameter tuning (gamma=0.03, nu=0.0216), the model identified anomalies within the expected 1-5% range, ensuring better precision.

2. **Visual Insights**:

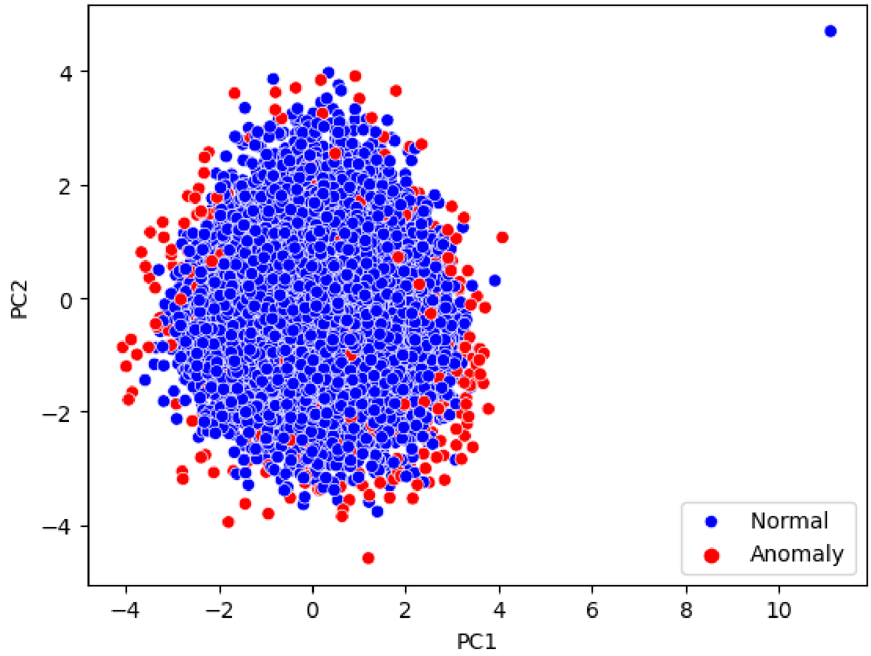


Figure 3 One-class SVM PCA Plot

- PCA-based 2D visualization confirmed that outliers were clustered at the periphery of the feature space.
- The ability to visually separate anomalies from normal observations enhances interpretability for stakeholders.

4.2 Isolation Forest

1. **Model Performance**:
   - Initial settings (n_estimators=100, contamination=0.05) produced a higher outlier rate than desired.
   - After tuning (n_estimators=100, contamination=0.025), the model achieved an anomaly rate of 2.5%, aligning closely with business requirements.

2. **Visual Insights**:

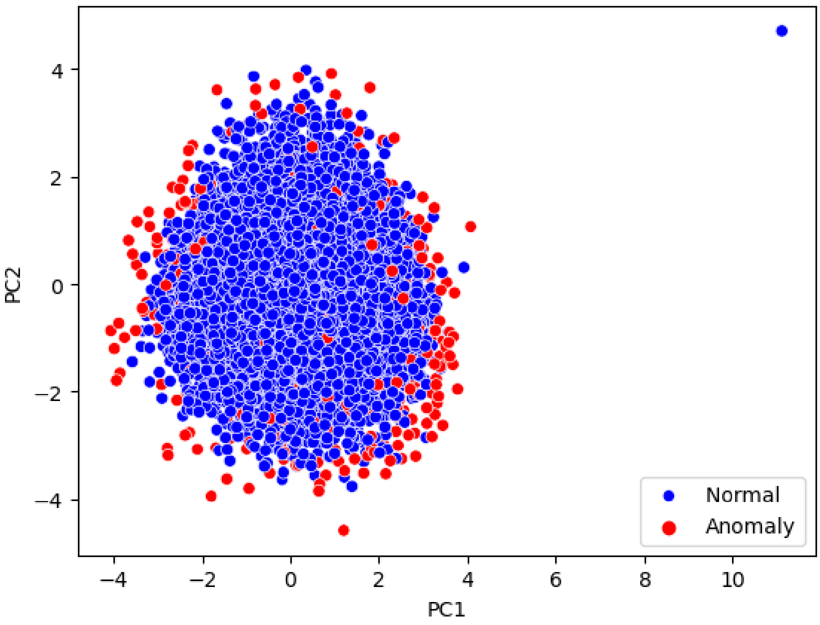


Figure 4 2 Isolation Forest PCA Plot

- PCA-based 2D plots showed that Isolation Forest effectively isolated anomalies, with fewer false positives compared to One-Class SVM.
- Unlike One-Class SVM, fewer outliers were detected at the top of the graph, suggesting a more refined anomaly detection.

5. Observations and Recommendations

5.1 Key Findings

- **Lubrication Oil Temperature** emerged as a critical factor influencing engine anomalies, frequently appearing in outlier combinations.
    - Recommendation: Collaborate with ship engineers to investigate its role further, enabling targeted maintenance.
- **Engine RPM** and **Lubrication Oil Pressure** were identified as priority features for monitoring, given their association with overheating and mechanical wear.

5.2 Model Effectiveness

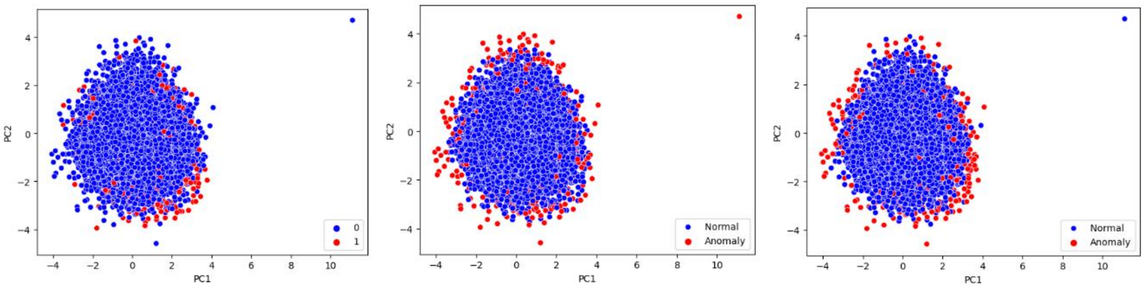
- **One-Class SVM** was highly sensitive but required fine-tuning to achieve the desired anomaly detection range.

- **Isolation Forest**, with its intuitive parameter tuning, demonstrated slightly superior performance, effectively identifying anomalies with fewer false positives.

5.3 Visualization Effectiveness

- PCA-based 2D visualizations were invaluable for communicating results, providing a clear depiction of outliers separated from normal data clusters.
- These visualizations can assist stakeholders in understanding the findings, supporting data-driven decision-making.

6. IQR Vs One-class SVM Vs Isolation Forest and Conclusion



	IQR	One-class SVM	Isolation Forest
Anomaly percentage	2.160225%	2.513437% at run time	2.503199 at run time
Observations	All observed outliners appear to be closer to the periphery of the PCA-based 2D visualization.		
	IQR and Isolation Forest appear to have similar clustering of outliners, although IQR detected fewer outliners.		

This analysis highlights the critical role of robust anomaly detection systems in enhancing ship engine reliability. Among the methods explored, Isolation Forest proved the most effective, offering an optimal balance of sensitivity and precision while aligning closely with business objectives. A key reason for its selection was its ability to exclude a distant data point that the One-Class SVM model incorrectly classified as an outlier. Notably, the IQR method did not flag this data point as an outlier, supporting the assumption that it represents normal behaviour rather than an anomaly. This reinforces the reliability of Isolation Forest in accurately identifying genuine anomalies while minimizing false positives. Key features like **Lubrication Oil Temperature** and **Engine RPM** were identified as critical for monitoring, providing actionable insights for maintenance planning.

By integrating these findings into operational workflows, the business can reduce engine failure risks, optimize maintenance schedules, and enhance fleet reliability, ultimately driving cost savings and customer satisfaction.