Course 101
5.3 Mini-project
Date: May 2024

Detecting the anomalous activity of a ship's engine

By Joshua Dixon

Executive Summary

In this analysis, statistical and machine learning methodologies were employed to detect anomalies in a ship's engine functionality, leveraging real-world data. The primary goal was to identify potential engine malfunctions that could lead to operational inefficiencies, increased fuel consumption, or safety hazards. Exploratory Data Analysis was carried out and anomaly detection implemented using the Interquartile Range (IQR), One-Class SVM, and Isolation Forest methods.

The rationale for using multiple anomaly detection methods lies in capturing complex patterns that might not be apparent through a single approach. The IQR method provides a straightforward baseline for outlier detection, while One-Class SVM and Isolation Forest offer advanced multivariate techniques to consider relationships between features. Utilising PCA for dimensionality reduction aids in visualising these complex relationships, thereby enhancing the interpretability and effectiveness of the models.

Key findings

Based on our analysis, several engine features exhibited more anomalies than others. This was apparent from the exploratory data analysis and generally consistent whether statistical or machine learning methodologies were used:

- Lubrication Oil Temperature
- Coolant Pressure
- Fuel Pressure
- Engine rpm

Recommendations

- Use the anomaly detection models developed to alert engineers of potential issues before they escalate, particularly for the key feature outline above, but not limited to.
- Schedule regular maintenance checks based on anomaly reports to ensure engine components are functioning within safe parameters.

Whilst this document identifies the features identified as outliers, a domain expert or the ship engineer should verify whether these outliers are normal/expected or investigate the context in which they occurred.

By implementing these recommendations, potential engine malfunctions and downtime will likely be significantly reduced. Proactive maintenance and timely identification of potential issues will lead to more efficient operations, reduced repair costs, and less frequent engine failures. This not only ensures the safety of the crew but also maintains timely deliveries, improving customer satisfaction and ultimately increasing the company's profitability.

Word Count: 979 (Excluding Page 2)

Table of Contents

1 Introduction

Efficient and safe operation of ship engines is crucial for timely deliveries, fuel efficiency, and the overall profitability of shipping companies. Detecting anomalies early can prevent costly breakdowns and ensure the safety of the crew.

This document outlines methodologies used to determine anomalies using statistical and machine learning methods. It does not determine the root cause of these anomalies.

2 Methodology

The Interquartile Range (IQR), One-Class SVM, and Isolation Forest, methods to detect anomalies in the engine’s performance data. PCA was used to reduce the feature set for visualisation, though the models were trained on the original feature set to ensure comprehensive anomaly detection.

The dataset included six key features: Engine RPM, Lubrication Oil Pressure, Fuel Pressure, Coolant Pressure, Lubrication Oil Temperature, and Coolant Temperature. These features were continuously monitored to assess the engine's health.

2.1 Key Steps in Analysis

- Exploratory Data Analysis
- IQR Method: Baseline outlier detection.
- Data Preprocessing: Scaling the data for machine learning algorithms.
- PCA: Reducing the feature set to two principal components for visualisation purposes.
- One-Class SVM analysis.
- Isolation Forest analysis.

Interpretations have been documented throughout the notebook. Several parameters were tested during the analysis, with the optimal plots summarised in this report.

3 Results

3.1 Interquartile Range (IQR) Method

- The interquartile method was used to identify outliers, this was done by flagging a feature with binary value to indicate whether it was an outlier or not for a given data point.

3.1.1 Inferences

has been confirmed that 1% to 5% of the dataset would be anomalies. It was imperative to analyse the interaction and combination of features to accurately discern anomalous activity. Data points had varying numbers of features with anomalies, as outlined below:

- 3 or More Outliers: Extremely rare, observed in only 11 data points (0.06% of the data).
- 2 or More Outliers: Uncommon but within the expected range, occurring in 422 data points (2.16% of the data).
- 1 or More Outliers: Relatively common, found in 19535 data points (23.73% of the data).
- All features are outliers: No occurrences were observed in the dataset.

The data points with 3 or more outliers should be flagged to the ship engineer to determine the cause of these outliers.

3.1.2 Effectiveness of the method

- The IQR method is straightforward, versatile, and robust to skewness and extreme values, making it a reliable tool for initial outlier detection.
- This method may flag dense clusters as outliers, relies on a fixed threshold, and is limited in multivariate contexts, potentially overlooking complex relationships between features.

3.2 One-class SVM

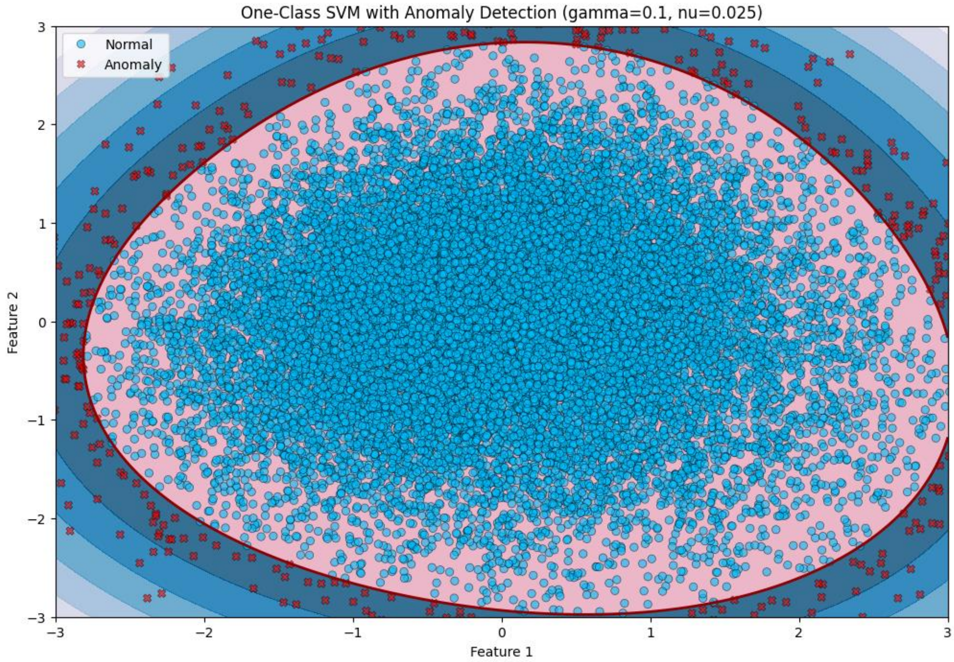

Figure 1: One-Class SVM with Anomaly Detection (gamma=0.1, nu=0.025)

- The figure above displays the One-Class SVM plot with gamma =0.1 and nu = 0.025
- The original features were reduced to two dimensions using PCA to enable visualisation with One-class SVM.
- 2.50% of the data points are classified as anomalies, which falls within the expected range of 1-5%. These are represented on the periphery of the data distribution.
- The data points that fall outside the boundary line should extracted and presented to the domain expert.

3.2.1 Inferences

MI	Values
Lub oil pressure	0.011499
Engine rpm	0.009873
Coolant pressure	0.008978
Lub oil temp	0.008338
Fuel pressure	0.005580
Coolant temp	0.004679

- The Mutual Information scores highlight which features are most influential in detecting anomalies. Specifically, lubrication oil pressure, engine rpm, coolant pressure, and lubrication oil temperature are the top features contributing to anomaly detection. These insights align with the results obtained from statistical methods and reinforce the importance of closely monitoring these features.

3.2.2 Effectiveness of the methodology

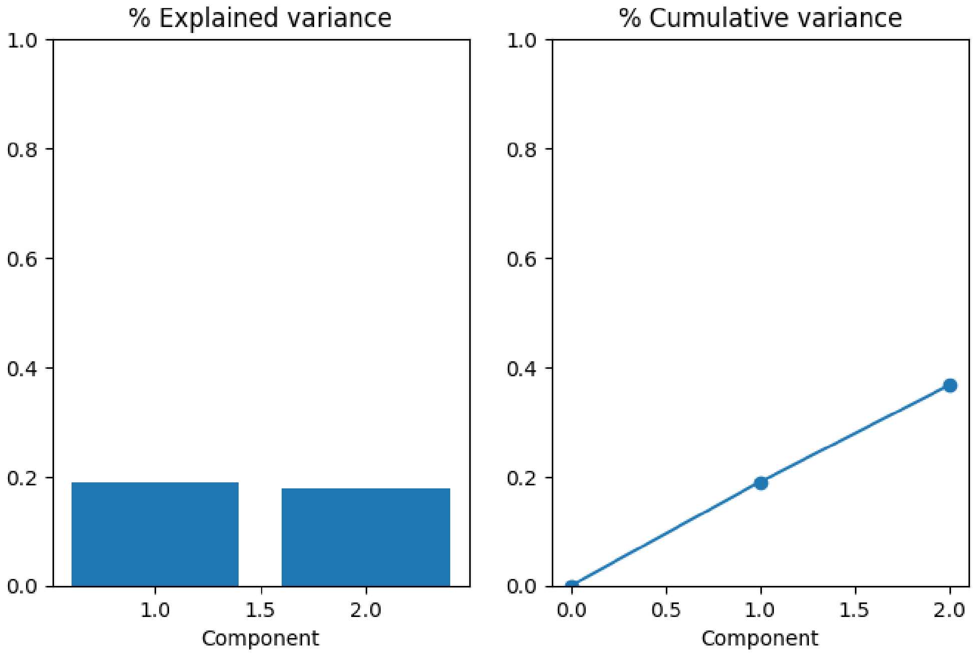

Figure 2: Principal Component Variance

- Variance and Visualisation: Approximately 40% of the variance is explained with two principal components, enabling visualization of high-dimensional data. However, this reduction might overlook important nuances, and One-Class SVM's limitation to two features for visualisation restricts a more comprehensive view.

- Feature Contribution and Dependency: Mutual Information scores provide insights into feature importance, highlighting parameters like lubrication oil pressure and engine rpm as critical for anomaly detection. Despite this, the relatively low scores indicate a weak dependency between individual features and anomalies, potentially limiting model robustness.

- It is difficult to interpret features when reduced to principal components.

3.3 Isolation Forest

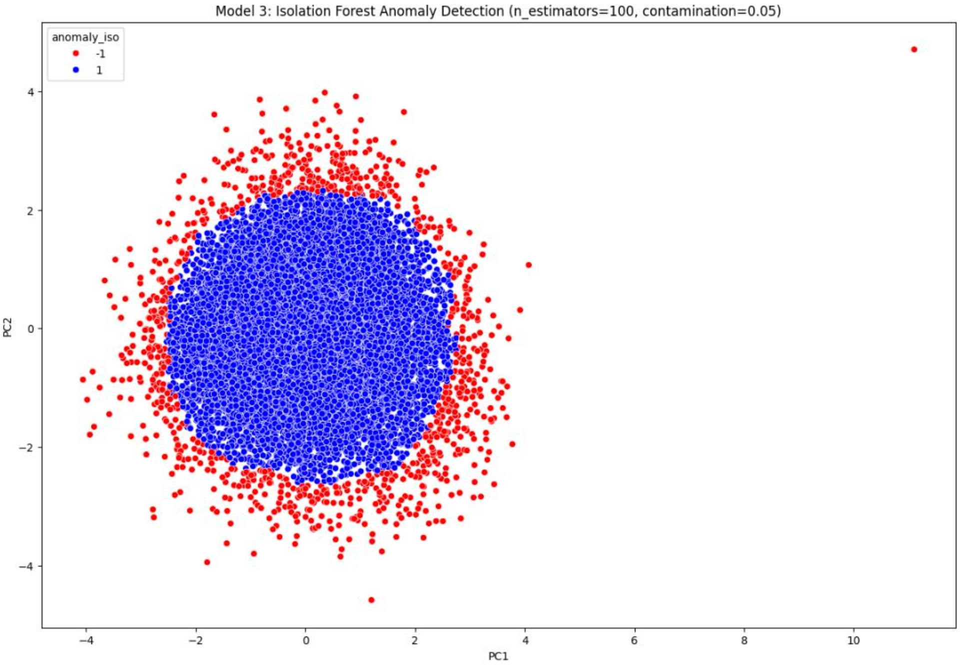

Figure 3: Model 3: Isolation Forest Anomaly Detection (n_estimators=100, contamination=0.05)

- Percentage of Anomalies: 5% of the data points are classified as anomalies, which falls within the expected range of 1-5%.
- Model Parameters: Using n_estimators=100 and contamination=0.05, the Isolation Forest identifies anomalies effectively.

3.3.1 Inferences

- Anomaly Distribution: Except for some extremities (i.e. top right of plot) anomalies (red points) are predominantly found at the periphery of the data distribution, indicating that these points deviate more from most of the data.
- This model configuration is suitable for detecting anomalies that are relatively rare but significant, providing valuable insights for identifying potential issues in the ship's engine data.
- The Mutual Information scores highlight which features are most influential in detecting anomalies. Specifically, engine rpm, coolant pressure, lubrication oil pressure, and lubrication oil temperature are the top features contributing to anomaly detection:

MI	Values
Engine rpm	0.013485
Coolant pressure	0.010940
Lub oil pressure	0.009262
Lub oil temp	0.007337
Fuel pressure	0.004173

3.3.2 Effectiveness of the methodology

- The strengths and limitations of the Isolation Forest method are similar to those of the One-Class SVM. Both methods use PCA for dimensionality reduction and display similar limitations in variance explanation and feature interpretation.

4 Conclusion

In this analysis, statistical and machine learning methodologies, including the Interquartile Range (IQR), One-Class SVM, and Isolation Forest, were employed to detect anomalies in a ship's engine functionality. The key features with higher anomalies were. Lubrication Oil Temperature, Coolant Pressure, Fuel Pressure, Engine rpm. This was generally consistent across all analysis methods.

Implementing the developed anomaly detection models and scheduling regular maintenance based on anomaly reports can significantly reduce potential engine malfunctions and downtime. These proactive measures will enhance operational efficiency, reduce repair costs, and ensure timely deliveries, ultimately improving profitability and safety.