Detecting Anomalous Activity in Ship Engine Data

Jack Goodbody

16 June 2025

Table of Contents

1 Introduction

This report aims to identify 1-5% of the data as outliers to support maintenance. Anomalous activity in a ship's engine can lead to significant challenges, including increased fuel consumption, safety hazards for the crew, and costly delays that impact delivery schedules. This report examines how data science techniques can identify these anomalies early, enabling the business to schedule timely maintenance, reduce operational risks, and enhance efficiency. By analysing a provided dataset, I aim to offer practical insights that support the company's goal of minimising downtime and improving customer satisfaction through reliable shipping operations.

2 Data Description

The dataset comprises 19,535 samples, each recorded with six critical engine features: Engine rpm, Lubrication oil pressure, Fuel pressure, Coolant pressure, Lubrication oil temperature, and Coolant temperature. These metrics are essential for assessing engine health, with anomalies potentially indicating issues like overheating or poor lubrication. Initial data checks using a range of different methods confirmed no missing or duplicate values, providing a robust foundation for analysis. The data's consistency allowed me to focus on extracting meaningful patterns without the need for extensive preprocessing.

3 Methods

A combination of statistical and machine learning techniques were employed to detect anomalies, as outlined below.

3.1 Exploratory Data Analysis (EDA)

EDA was used to understand the dataset's structure. This involved calculating descriptive statistics (means, medians, and 95th percentiles) to assess central tendencies and extremes. Visualisations, including histograms (Q-Q plots and box plots), were created to explore data distribution and identify potential outliers. Figure 1, a box plot, visually highlights the spread and outliers across all features, aiding initial insights.

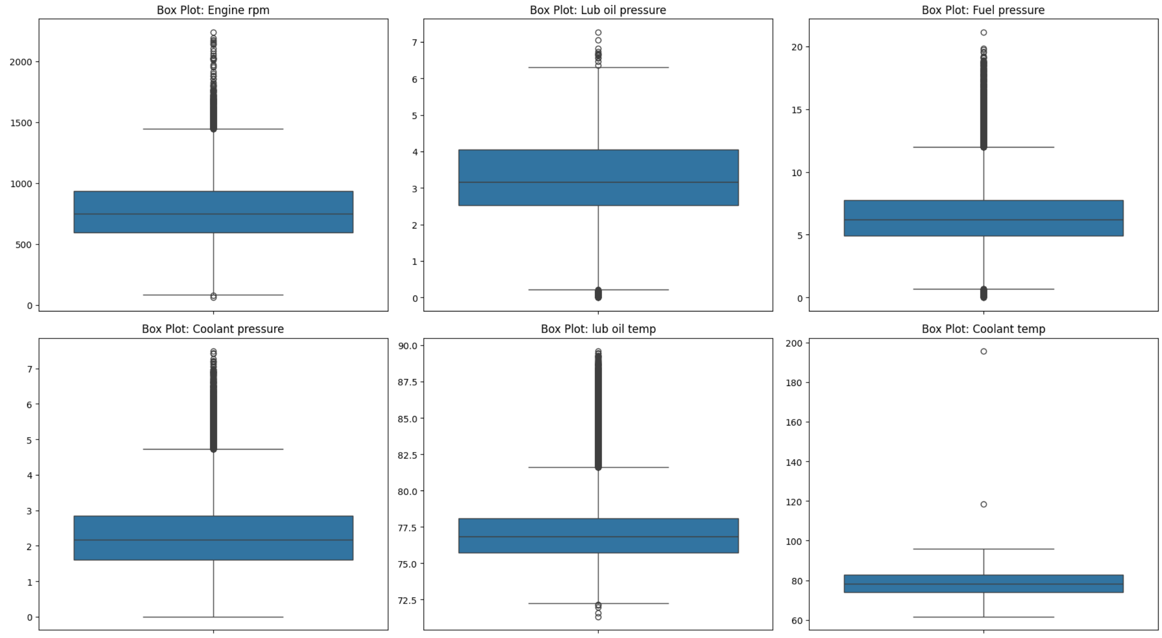

Figure 1: Box plot of orginal data showing outliers.

3.2 Statistical Method: IQR

The Interquartile Range (IQR) method was used to flag outliers systematically. For each feature, the first quartile (Q1), third quartile (Q3), and IQR was calculated, which determined lower and upper bounds. Values outside these bounds, such as Lubrication oil temperature below 50.5°C or above 95.5°C, were marked as outliers using binary `_flagged` columns. This information was used to identify that rows with two or more outliers could be classed as outliers. This equalled 422 rows falling within the expected 1-5% range.

3.3 Machine Learning Methods

Two machine learning approaches were tested: One-class Support Vector Machine (SVM) and Isolation Forest. Features were scaled using `StandardScaler` to ensure consistency. For One-class SVM, the nu parameter (e.g., 0.03) and gamma was adjusted to target 1-5% outliers, experimenting with values to refine results. Using this method, 585 outliers were identified, which also falls into the 1-5% range expected. See figure 2, a 2D PCA plot showing normal data points in blue and outliers in red, as detected by the OCSVM method.

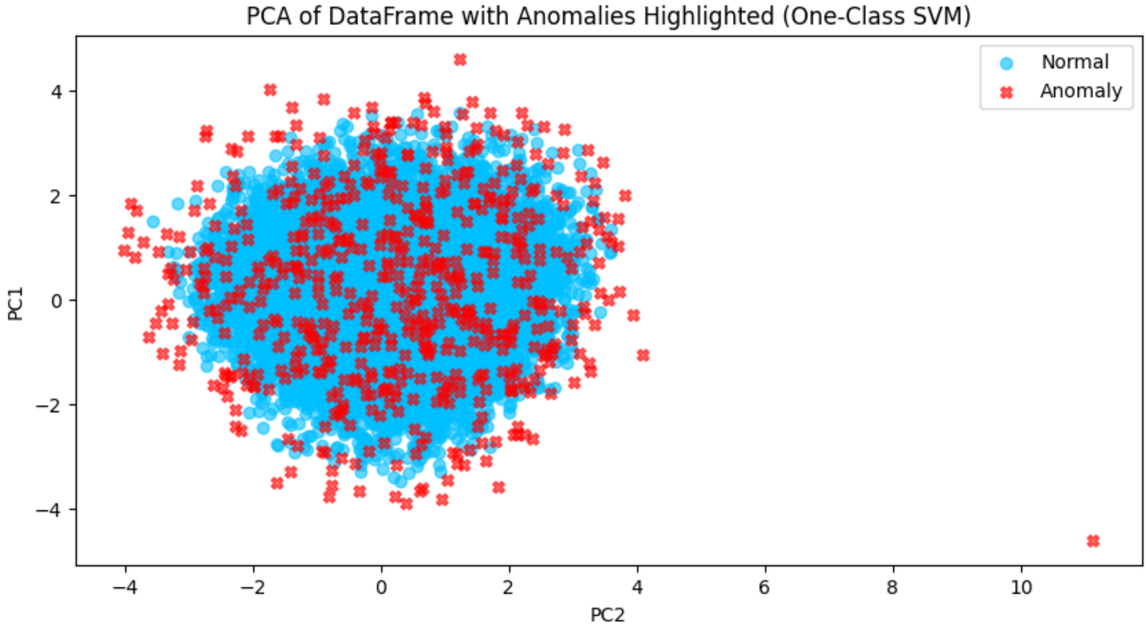

Figure 2: PCA visualisation of anomalies detected by One-Class SVM

For Isolation Forest, the contamination parameter (0.02) was tuned to achieve the same range, assessing its impact on outlier detection. This method was successful in finding 391 outliers, well within the 1-5% range expected. This method was the closest to the purely statistical IQR range, making it the more reliable machine learning method for this investigation. See figure 3.

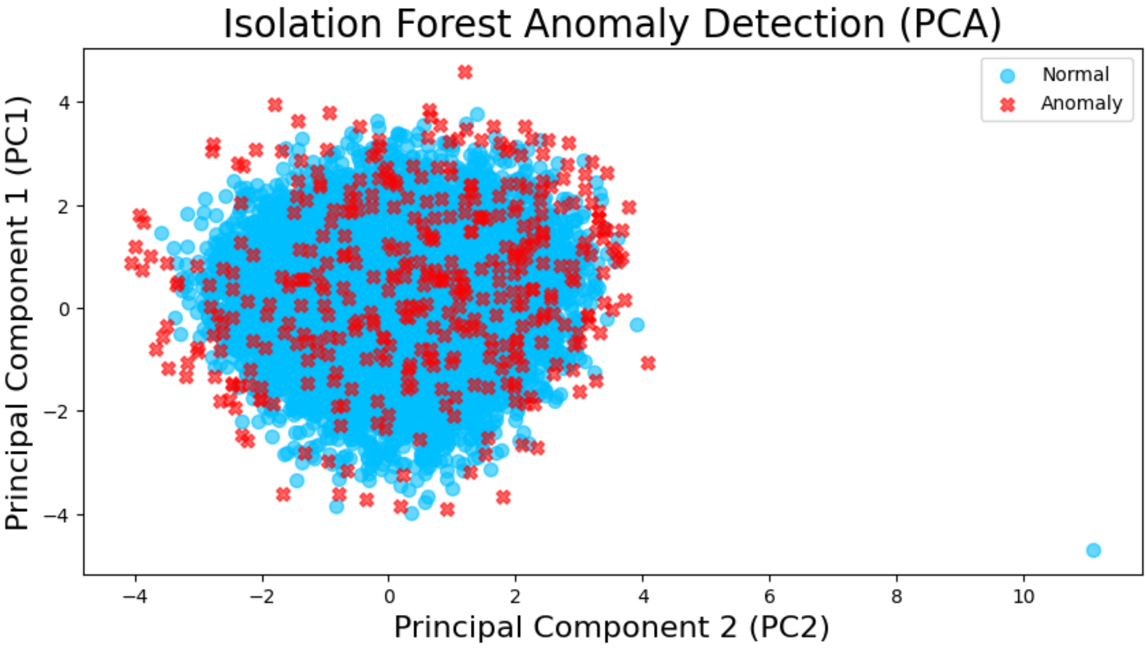

Figure 3: PCA visualisation of anomalies detected by Isolation Forest

Model	Parameter 1	Parameter 2	Outliers Detected	% Outliers
Isolation Forest	Contamination = 0.02	N/A	391	2.001
One-Class SVM	nu=0.03	gamma='scale'	585	2.990

Table 1: Summary of outliers detected by each model configuration.

3.4 Dimensionality Reduction

To visualise the results effectively, Principal Component Analysis (PCA) was applied to reduce the data to two dimensions. This enabled the plotting of normal points against outliers, using distinct colours for clarity, which is detailed in the results section.

4 Results

The IQR method identified Lubrication oil temperature and Coolant temperature as the features with the highest outlier frequencies, with bounds such as 50.5°C to 95.5°C for the former. This suggests potential overheating or inadequate lubrication, critical for engine health. Isolation Forest, with a contamination of 0.02, detected 391 outliers (2% of 19,535 samples), closely aligning with the IQR method's 422 outliers when focusing on samples with two or more flagged features. One-class SVM, with nu = 0.03 and gamma = 'scale', also identified around 2% outliers (585 samples), though it required more parameter adjustments to avoid over- or under-detection.

Method	Type	Outliers found	Percentage (%)
Interquartile Range (IQR)	Statistical	422	2.160
One Class SVM (OCSVM)	Unsupervised Machine Learning	585	2.990
Isolation Forest	Unsupervised Machine Learning	391	2.001

Table 2: Summary of outliers found by each method.

5 Conclusions

This analysis finds that temperature-related features, Lubrication oil temperature and Coolant temperature, are key indicators of engine anomalies, likely linked to overheating or poor lubrication, which could lead to malfunctions or safety risks. Isolation Forest, with a contamination rate of 0.02, emerged as the most reliable detection method, consistently identifying around 2% outliers, in line with IQR findings. It is recommended that a monitoring system be implemented to track these temperatures, using thresholds such as 50.5°C to 95.5°C for lubrication oil temperature, to trigger preventive maintenance. This could reduce downtime and costs, thereby enhancing fleet reliability.

However, the 2D PCA visualisations, while useful for learning, did not provide as much business value as the IQR method, as it was challenging to interpret the significance of outliers. Further testing and simpler reporting formats are needed to ensure the data is effectively communicated to stakeholders and delivers maximum value.