Detecting the anomalous activity of a ship's engine

UNIVERSITY OF
CAMBRIDGE

Anastasia Alexandra Krivenkovskaya

Institute of Continuing Education

University of Cambridge

This academic report is submitted for the course

Applying statistics and core data science techniques in business

CAM_DS_C101

March 2025

Table of Contents

1. Introduction

In this mini-project, I explored various approaches to detect anomalies in a dataset capturing the performance metrics of a ship's engine. The core research question driving this analysis was: *How can unsupervised learning techniques be effectively used to identify anomalous engine behaviour that could indicate potential malfunction, inefficiency, or safety concerns?*

Detecting anomalies is critical in industrial contexts, as unexpected deviations from normal engine operation may lead to serious breakdowns, costly delays, or even safety hazards for crew and cargo. The goal was to identify such deviations early and build an anomaly detection system capable of flagging unusual behaviour for maintenance intervention.

2. Dataset Description

The dataset included 19,535 rows, each representing a snapshot of engine performance. It featured six key variables: Engine RPM, Lubrication Oil Pressure, Lubrication Oil Temperature, Coolant Pressure, Coolant Temperature, and Fuel Pressure. These features are commonly used to monitor the operational health of heavy machinery in real time. Before any modelling, I performed data exploration to understand the distributions and identify any skewness or irregularities. Notably, Lubrication Oil Temperature and Fuel Pressure exhibited heavy right-skew, suggesting the presence of extreme values or potential anomalies. No missing or duplicate values were found, which allowed me to proceed directly to analysis.

3. Methods

3.1 Statistical Method: Interquartile Range (IQR)

For the statistical method, I used the Interquartile Range (IQR) to identify outliers. This involved computing the 25th (Q1) and 75th (Q3) percentiles of each feature and defining the IQR as Q3 - Q1. Any data point lying below Q1 - 1.5×IQR or above Q3 + 1.5×IQR was flagged as an outlier. To reduce noise, I only considered a row to be anomalous if it contained outliers in two or more features. This approach led to the detection of 422 anomalous samples, roughly 2.16% of the dataset, fitting the expected anomaly range for real-world scenarios. Features like lubrication oil temperature and fuel pressure were the most common contributors to outliers. Figure 1 shows a bar chart illustrating which features were most often involved in anomalous readings.

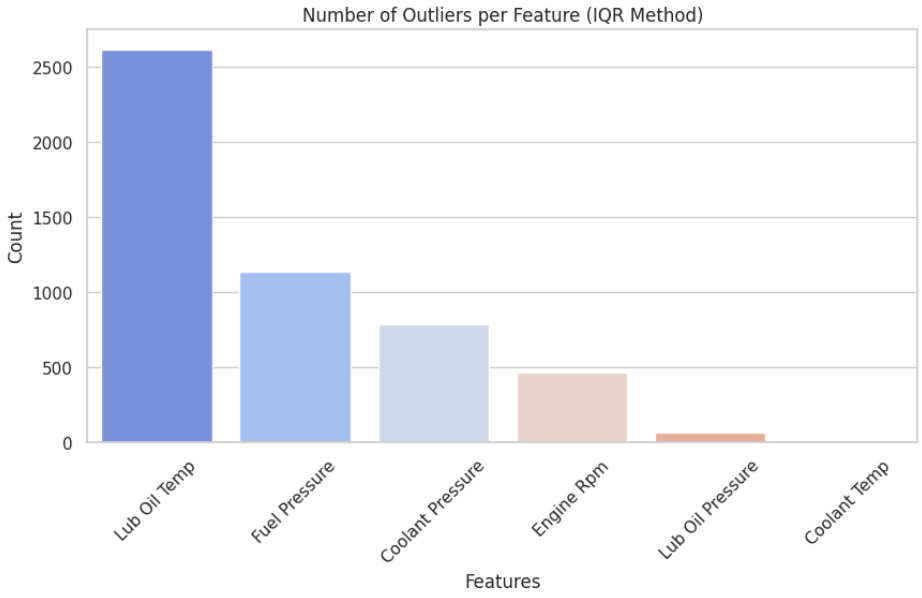

Figure 1. Number of outliers per feature detected using the IQR method.

Lubrication oil temperature is the most frequent contributor to anomalies, followed by fuel pressure and coolant pressure.

3.2 Machine Learning Models

Next, I applied machine learning methods: One-Class Support Vector Machine (SVM) and Isolation Forest. For SVM, I first scaled the features using StandardScaler, since SVM is sensitive to the magnitude of feature values. I tuned parameters like 'nu' and 'gamma', and found that setting nu=0.02 and gamma='auto' produced good results, with 400 anomalies detected (2.05%). One-Class SVM attempts to learn a boundary that encloses normal data points and labels anything outside as an anomaly. The model's performance was visualised using PCA (Principal Component Analysis) to project the high-dimensional data into 2D space. Figure 2 displays the SVM results, with outliers clearly situated on the periphery of the main data cluster.

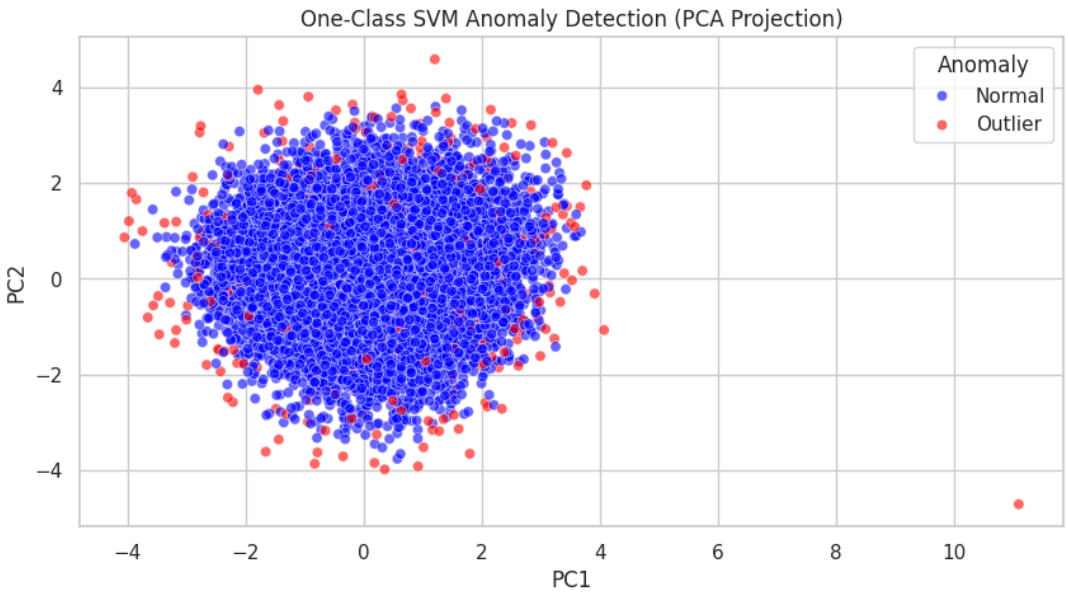

Figure 2. One-Class SVM Anomaly Detection (PCA Projection)

Anomalies are marked in red and appear clearly on the outer edges of the main data cluster. This indicates that the model effectively isolates unusual engine behavior from normal performance.

Isolation Forest, on the other hand, does not require feature scaling and is known for its efficiency and performance in high-dimensional settings. It works by randomly selecting features and thresholds to 'isolate' each data point. Anomalies are isolated more quickly and thus receive higher anomaly scores. Using a contamination rate of 0.02, the model identified 391 anomalies (2.00%). This method was notably fast and stable. Like with SVM, I used PCA to visualise the separation between normal and anomalous samples. Figure 3 presents this plot, where again the anomalies are clustered at the edges. The consistency between models increased my confidence in the validity of the detected outliers.

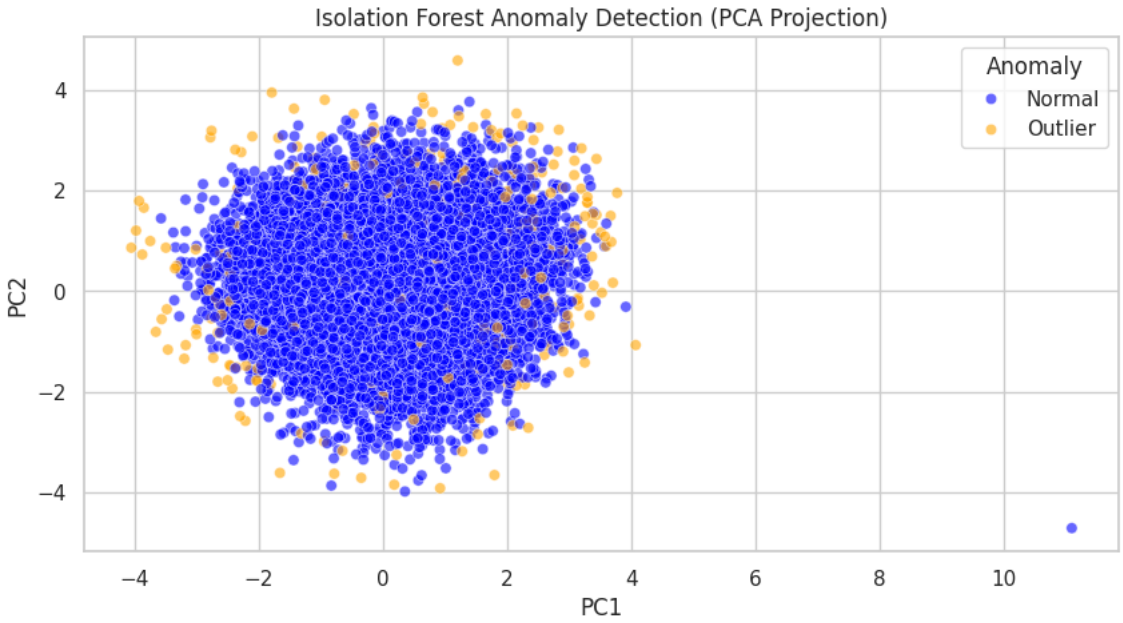

Figure 3. Isolation Forest Anomaly Detection (PCA Projection)

The orange points represent anomalies, mostly located at the margins of the dense central cluster. This suggests the model's ability to separate out rare and distinct observations from the rest of the data.

4. Results and Observations

Across all methods, several important patterns emerged. Most anomalies were not caused by a single unusual feature, but by combinations of abnormal values across multiple sensors. For instance, high lubrication oil temperature alone did not always indicate a problem — but when it coincided with high fuel pressure or elevated RPM, it became a strong warning sign. This highlights the value of multivariate anomaly detection. Among the six features, Lubrication Oil Temperature, Fuel Pressure, and Coolant Pressure were most frequently associated with anomalous readings. Engine RPM and Coolant Temperature were more stable but contributed in cases of combined deviations.

5. Insights and Recommendations

Based on these observations, I recommend using a hybrid approach. Isolation Forest should be adopted as the primary detection mechanism in real-time systems, due to its speed and reliability. IQR remains a valuable supporting method, particularly for transparent audits and explainable alerts. Additionally, the company should focus on multivariate monitoring — paying special attention to situations where multiple features exceed their 95th percentile simultaneously. Static alert thresholds should be replaced with dynamic, quantile-based boundaries to better adapt to evolving operational norms.

6. Conclusion

This project demonstrated the value of combining data exploration, statistical reasoning, and machine learning to address practical challenges in industry.

Isolation Forest emerged as the most robust and convenient model for this dataset, while PCA proved to be an effective tool for visual validation and communication.

The anomalies detected through these techniques — representing approximately 2% of the 19,535 engine performance records — can serve as early warnings, helping to prevent malfunctions before they cause major disruptions.

Appendix: Figures

Figure 1: Feature-wise IQR Outliers Bar Chart

Figure 2: PCA Projection of Anomalies (One-Class SVM)

Figure 3: PCA Projection of Anomalies (Isolation Forest)