5.3 Min Project - Detecting Anomalous Activity in a Ship's Engine

Prepared by - Mohamed Nuri

Date - 09/12/2024

Client - International Shipping company

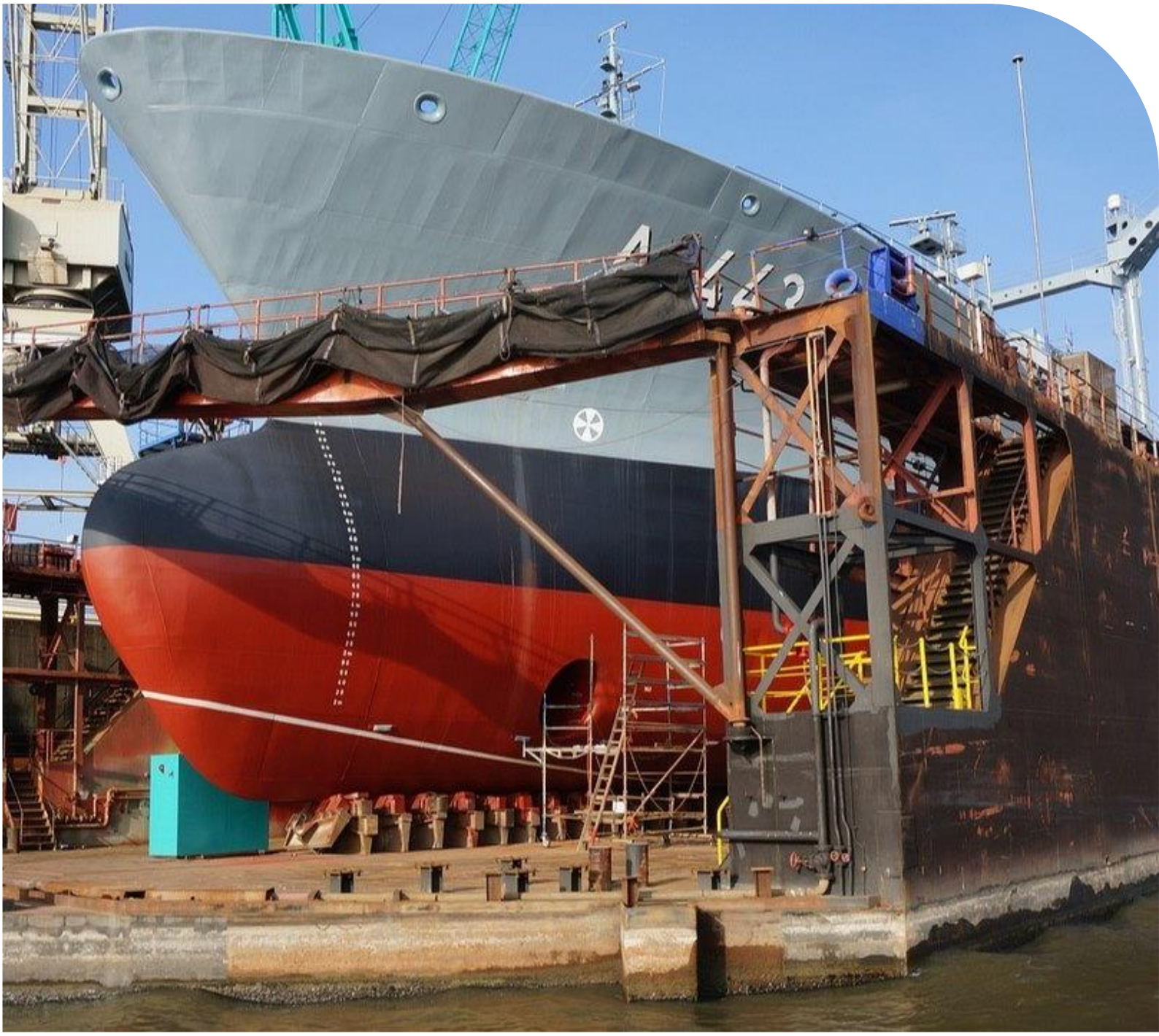

Table of Contents

Introduction

Problem Statement:

A shipping company aims to enhance revenue and customer satisfaction by ensuring the timely delivery of goods across its fleet. To achieve this, they seek to develop an anomaly detection system capable of identifying irregularities in a ship's engine functionality. This system will enable the company to proactively flag potential issues, allowing for timely investigation and maintenance. By addressing anomalies early, the system will help minimize downtime, improve crew safety, and ensure consistent performance, thereby reducing delays in delivery

Data Exploration

The client has provided us with a dataset that monitors a ships engine functionality (Devabrat, 2022). The data set provides us with data on six key features of a ships engine functionality for 19534 ships.

The six key features are:

- Engine rpm
- Lubrication oil pressure
- Fuel pressure
- Coolant pressure
- Lubrication oil temperature
- Coolant temperature

The dataset contains no missing values and no duplicate rows. After conducting exploratory data analysis, majority of the features showed a normal distribution, except coolant pressure and lubrication oil pressure which showed a bimodal shape in the histogram visualisation(fig 7). All six features showed that they are skewed from the boxplot distribution (fig 1-6).The correlation heatmap (fig 8) shows that the features have low correlation between each other.

Data Exploration

Boxplots for the different features:

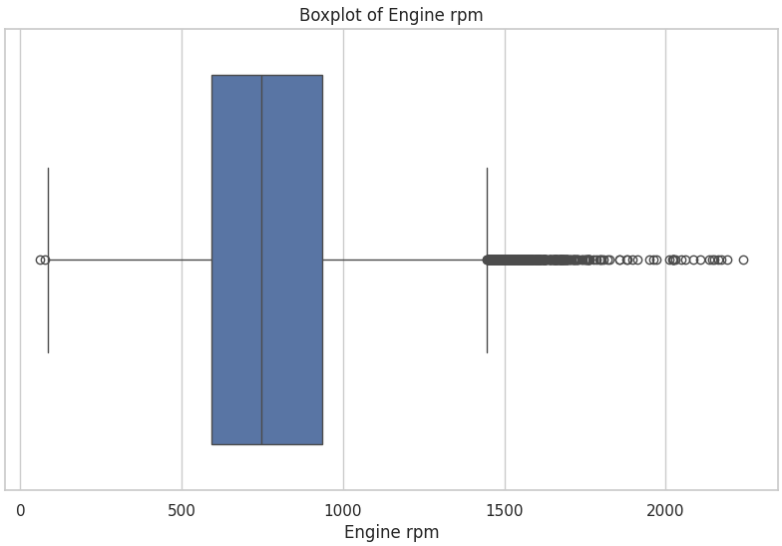

Figure 1

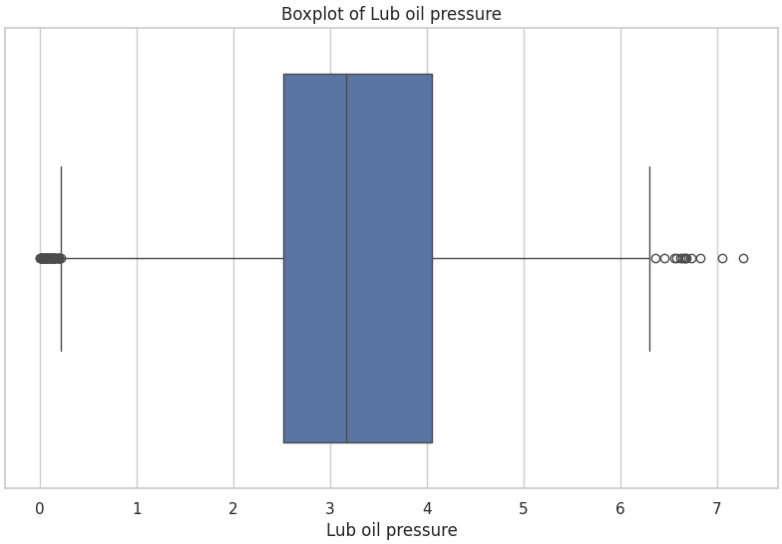

Figure 2

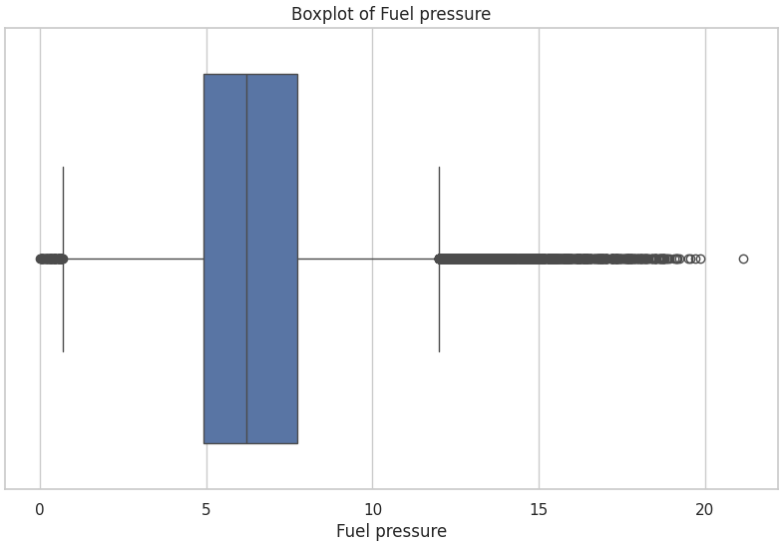

Figure 3

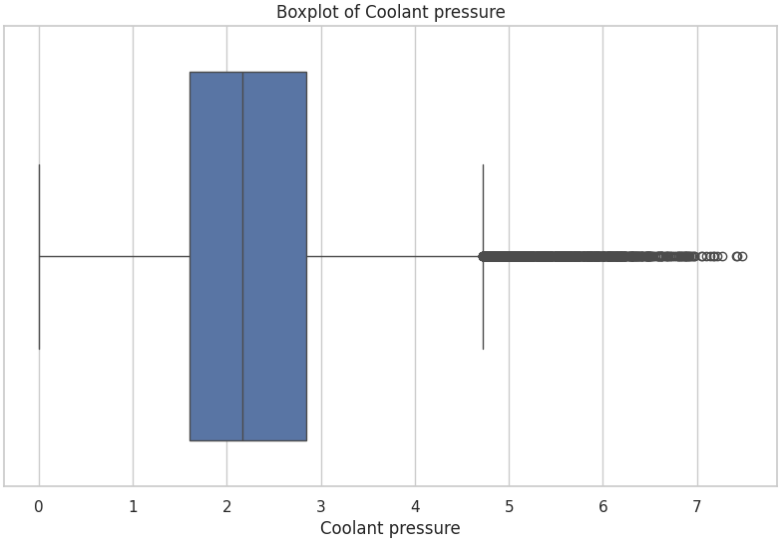

Figure 4

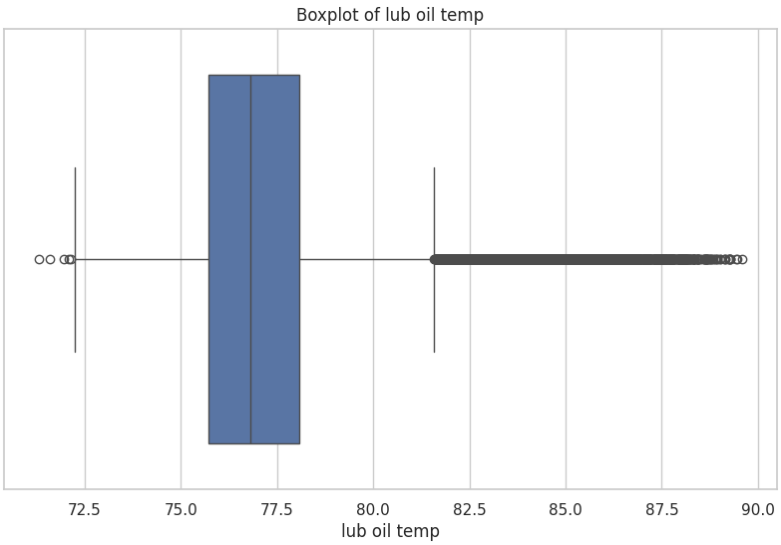

Figure 5

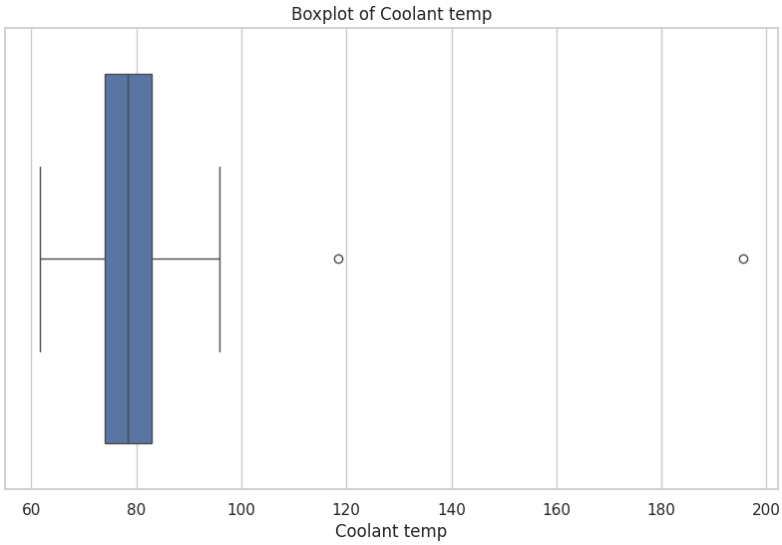

Figure 6

Histograms

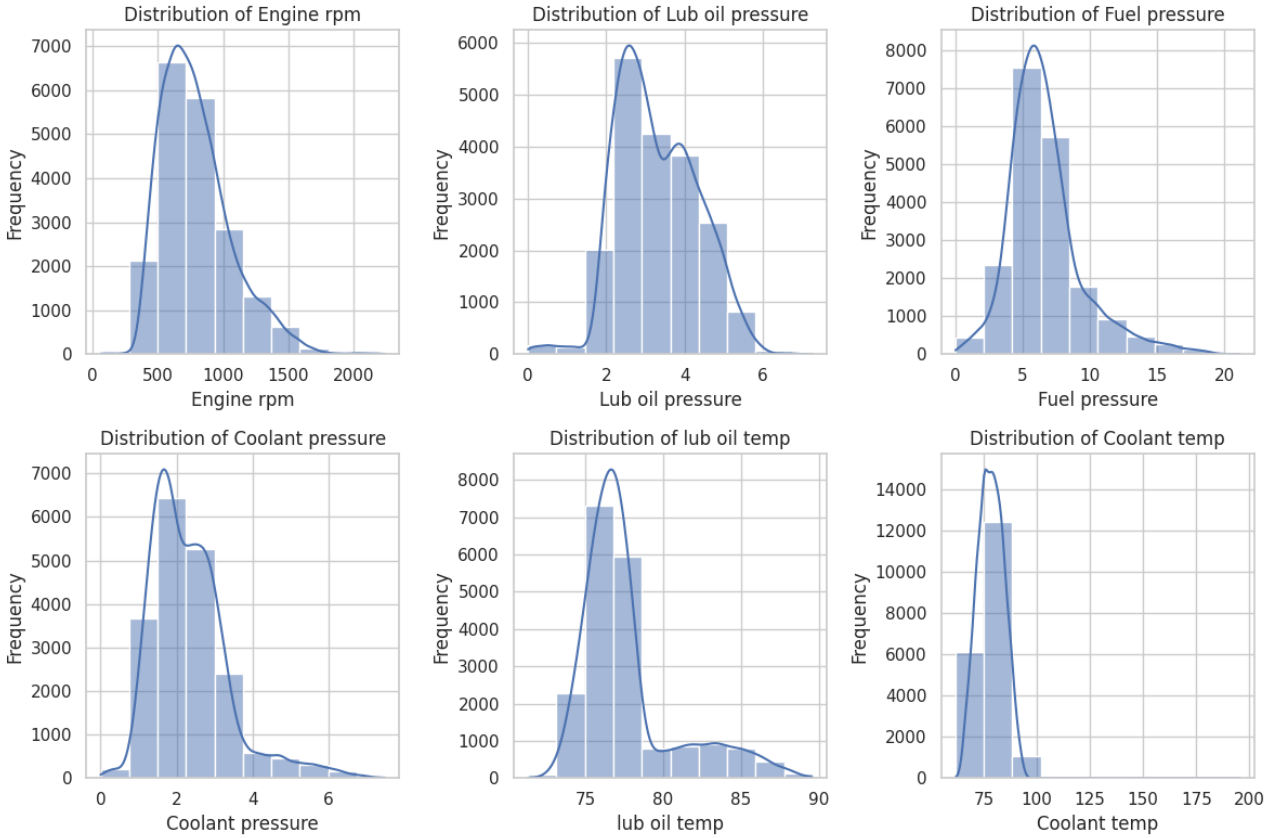

Figure 7

Correlation Heatmap of Features:

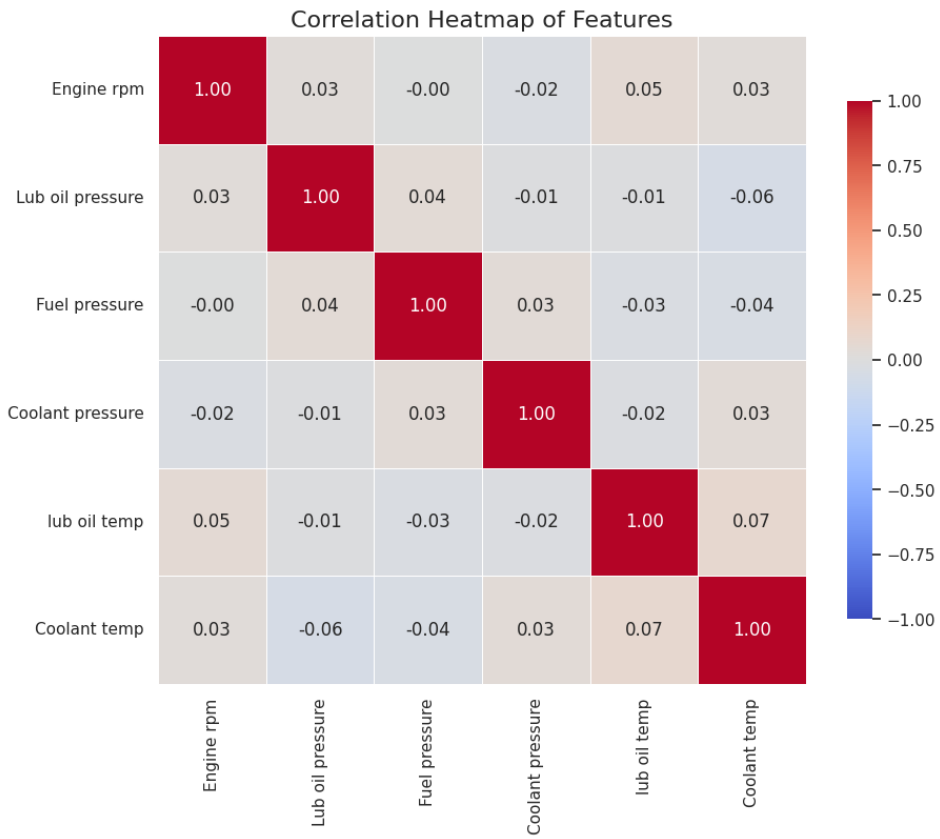

Figure 8

Data Analysis

To identify anomalous activity in the engine functionality we used three approaches:

- Approach 1 – Anomaly detection: Statistical Methods
- Approach 2 – Anomaly detection with OnceClass SV ML model
- Approach 3 – Anomaly detection with Isolation Forest ML model

We assumed that all ships in the dataset have the same ship and engine models for the analysis

Approach 1 – Anomaly detection: Statistical Methods

Method:

1) Calculated the IQR for each feature, (IQR= Q3-Q1)
2) Anomaly threshold established - any data value below Q1-1.5(IQR) or above Q3+1.5(IQR)
3) All data values in each feature is checked and flagged if its an anomaly.
4) The sum of anomalies for each ship is calculated.

Based on previous studies, approximately 1-5% of ships are expected to display anomalous engine functionality. To align with this expectation, we calculated the percentage of ships exhibiting anomalies across one, two, three, or four features

Results:

Number of features identified as anomalies	Total number of ships	Percentage(%)
1	4214	21.6
2	411	2.10
3	11	0.06
4	0	0

Figure 9

Conclusion:

- The majority of ships (21.6%) exhibited anomalies in a single feature, which is insufficient to justify maintenance intervention. It would be costly and not feasible to use this condition to identify ships requiring servicing and maintenance.
- Ships with anomalies in three or more features were exceedingly rare (<0.1%), and no ships had anomalies in all four features.Ships with anomalies in two features should be flagged for further investigation and potential maintenance.

Approach 2 – Anomaly detection with Once-Class SVM ML model

Method:

1) The data was first scaled using standardisation
2) A range of gamma and nu values were tested to identify the parameter combination that best aligns with the expected 1-5% anomaly rate. This is seen in Figure 10 & 11
3) We sum the number of ships with anomalies and calculate the percentage of ships overall that have anomalous engine functionality.

Gamma	Nu	Percentage(%)
0.167	0.2	20
0.167	0.05	5
0.167	0.02	2.05

Figure 10

Gamma	Nu	Percentage(%)
0.2	0.02	1.99
0.1	0.02	2.01
0.05	0.02	2.02

Figure 11

Conclusion

- On the other hand, with lower nu values (like 0.02), the model becomes stricter, potentially resulting in false negatives, where some anomalies go undetected
- The choice of nu = 0.05 when keeping gamma is 0.167 balances the need to detect engine issues while minimising unnecessary maintenance checks. It results in a 5% anomaly detection rate, aligning with the desired threshold.

PCA Visualisation:

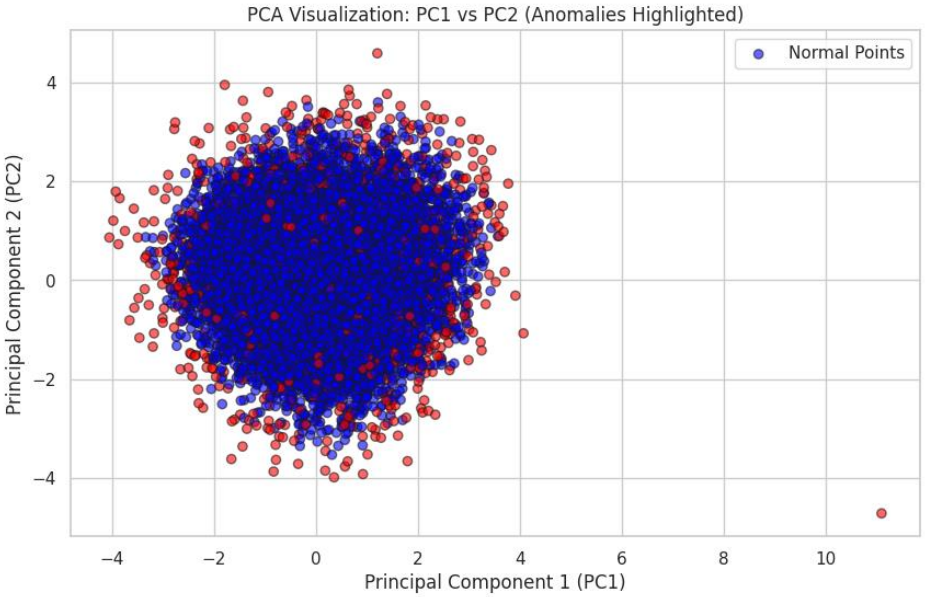

Figure 12- from One Class SVM anomaly detection method

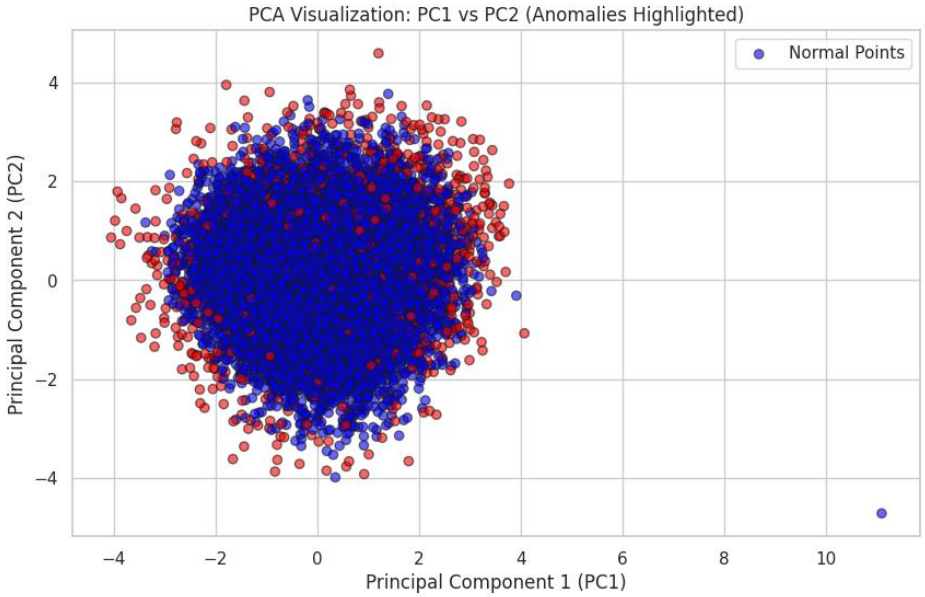

Figure 13- from Isolation Forest Anomoly detection method

- From the PCA visualisations, most of normal points are clustered tightly in the centre, which indicates that most ship engine data falls within normal range,
- There are red anomalies that are inside the blue clustered centre which suggest there are subtly issues that are not immediately obvious but need further investigation. It could be noise or an early indicator of engine performance issues
- Most of the anomalies form a ring around the blue normal data. This suggest that most anomalies are mild deviations, these data points could represent an early stage issue that warrants further monitoring.
- Few data points deviate significantly from the normal cluster and represent strong outliers, these ships may require immediate investigation

Approach 3 – Anomaly detection with Isolation Forest ML model

Method:

1) No Scaling was required for this algorithm
2) In Isolation Forest, the contamination parameter directly sets the percentage of anomalies, making it simple to control detection rates.
3) Contamination parameter was set to 5% as it ensures that more anomalies are detected within our 1-5% desired detection rate.
4) The n_ estimators parameter was set to 100 as it is a good balance between model stability and computational efficiency.

Contamination	Percentage (%) of anomalies detected
0.05	5
0.03	3

Figure 14

Conclusion:

- In Isolation Forest, the contamination parameter directly sets the percentage of anomalies, making it simple to control detection rates as seen in figure 13.
-

Evaluation:

Review of Approaches:

- The dataset contains 19,534 rows, which is relatively small for machine learning, and has low correlation between features, making anomaly detection challenging using the IQR method. Machine learning methods are more practical for this type of data as they can account for feature interactions, unlike IQR, which analyzes features independently and is less suitable for multivariate anomaly detection.
- One-Class SVM is better suited than the other two in capturing anomalies in higher-dimensional space.
- Given the scenario, machine learning methods are more practical, as IQR is designed to analyze each feature independently, making it less effective for detecting anomalies arising from interactions between multiple features. Machine learning models, on the other hand, can account for relationships between features.