Detecting the Anomalous activity of a Ship's engine

Vanessa Ezeoke

21/Sep/2025

Table of Contents

Problem Statement

Ship engine failures pose significant risks to maritime operations, including costly downtime, fuel inefficiency, delivery delays, and safety concerns. To mitigate these challenges, this project seeks to design an anomaly detection system that leverages six critical engine parameters—engine rpm, lubrication oil pressure, fuel pressure, coolant pressure, lubrication oil temperature, and coolant temperature. The system will evaluate engine performance and flag abnormal patterns that warrant further inspection. By enabling timely and targeted maintenance, this solution will help reduce operational risks, improve efficiency, enhance customer satisfaction, and ultimately drive revenue growth.

Data Overview

The ship engine dataset consists of six features and 19535 entries. The features provided are: engine rpm, lubrication oil pressure, fuel pressure, coolant pressure, lubrication oil temperature, and coolant temperature.

Method

Exploratory data analysis was performed on the data, then three anomaly detection methods were applied on the data. The methods applied are outlined in the table below. Based on information provided by the domain expert, it is expected that the anomalies will make up about 1-5% of the data. Hence, for all methods, the goal was to ensure the engine anomalies detected fell within the aforementioned range.

Method	Type	How the Method works	Key parameters tested
IQR Method	Statistical	Anomalies are entries that exceed the inter-quartile range: entries > Q3 + (IQR*1.5) or entries < Q1 - (IQR*1.5)	Number of Features that contained anomalous points
One Class SVM	Machine Learning	It learns the boundary of “normal” data in feature space and classifies new points as either inside (normal) or outside (anomalous) the boundary.	Gamma, Nu
Isolation Forest	Machine Learning	An ensemble method that detects anomalies by randomly splitting data points into trees. Anomalies are detected by highlighting data points that are isolated in fewer splits (shorter tree paths).	contamination

Table 1: summary of each method used

The Exploratory data analysis performed on the data include assessing the data for missing and duplicate data, generating descriptive statistics and visualising the data via histogram and box plots.

IQR Method:
In this statistical method, anomalies were identified for each feature independently. Subsequently, entries were flagged as anomalies if multiple features on that entry were flagged as an anomaly.

One Class SVM and Isolation Forest:

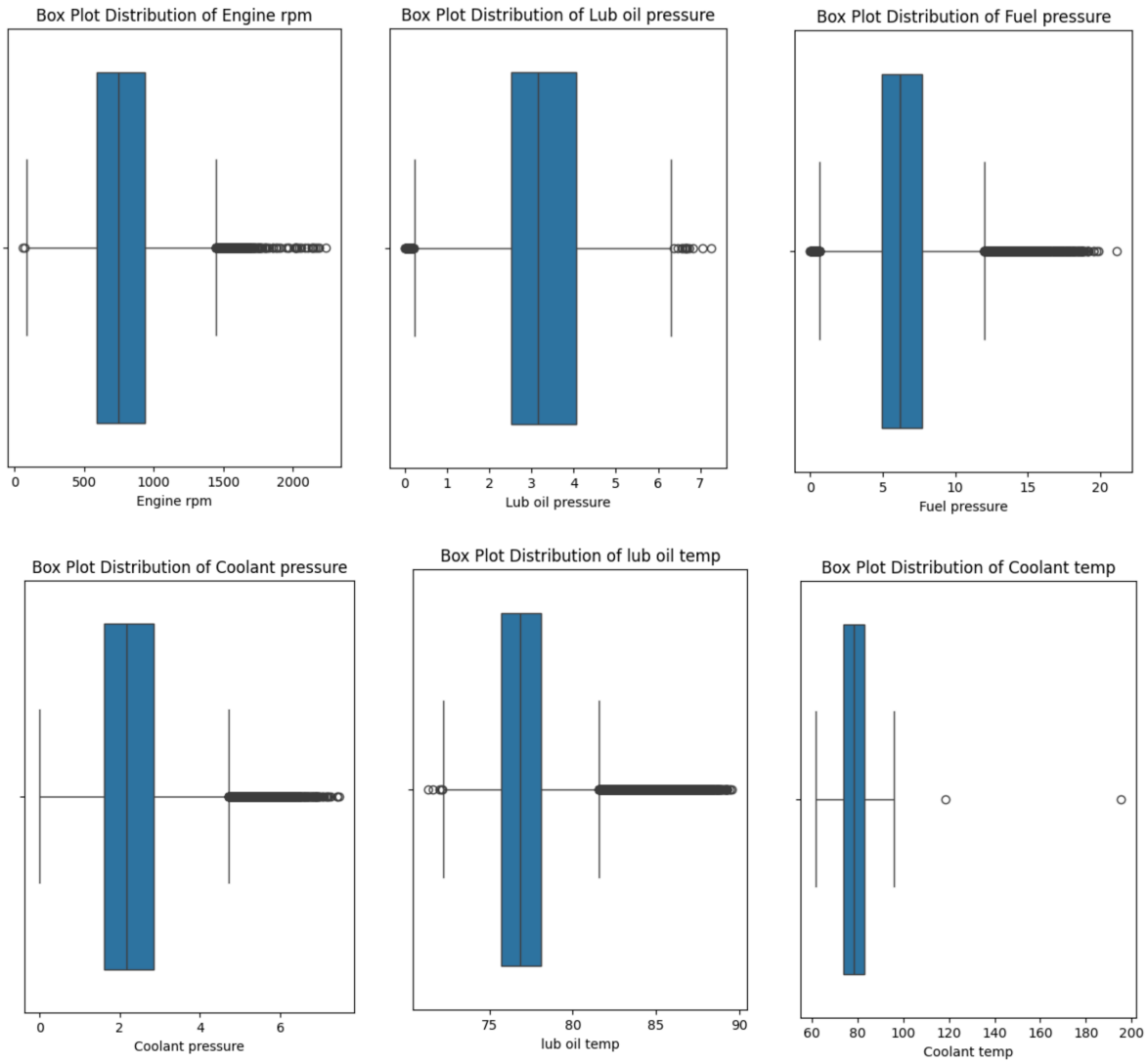
For the Machine learning methods, different parameters were tested for optimisation. In one-class SVM, the features were normalised using the standardization method before testing. However, in the Isolation Forest method, the data was imputed directly without normalisation. For both methods, after the algorithms were applied, the was visualised to view the model outputs. Since the

data had 6 features, the data was reduced using Principal Components Analysis (PCA) in order to visualise the predictions in 2D. PCA requires standardized data, hence the data was normalised before PCA. The ideal parameters were chosen based on the visualisation.

Results

Box plots were used to show the distribution of data for each data point. The Engine rpm, Fuel pressure, coolant pressure and lubrication oil temperature had the most outliers. The coolant temperature and pressure did not have outliers in the lower limit.

Fig 1: Box plot visualisation of the features



IQR Method:

Using the IQR method, the following results were attained and used to decide on the ideal classification of anomalies.

Test	Entries flagged	Ratio to total data
≥ 1 feature	4636	23.73%
≥ 2 feature	422	2.16%
≥ 3 features	11	0.06%
≥ 4 features	0	0%

Table 2: IQR Test results

Due to the high ratio of entries flagged as anomalies from univariate classification, we decided on a multivariate method, with at least 2 features flagged as an outlier for the entry to be considered an anomaly.

One Class SVM Method:

Using the one-class SVM, the combination of gamma and nu that result in a suitable ratio of anomalies are given in the table below:

Gamma	Nu	Anomaly Count	Ratio to total data
0.1	0.025	490	2.51%
0.25	0.1	230	1.18%
0.25	0.025	501	2.56%
0.5	0.01	611	3.13 %
0.5	0.025	668	3.42%

Table 3: Combination of Gamma and Nu that result in a reasonable number of anomalies

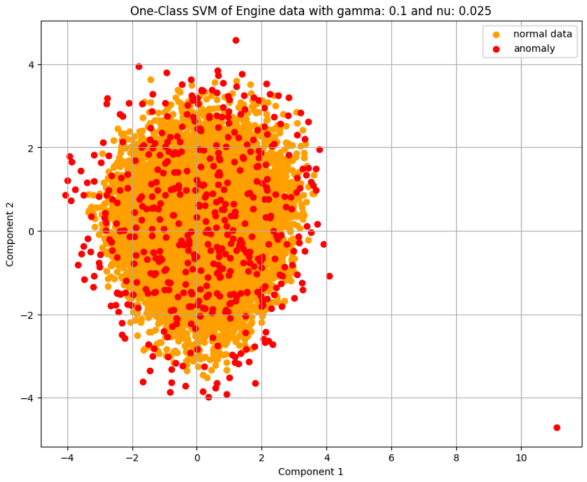
Isolation Forest:

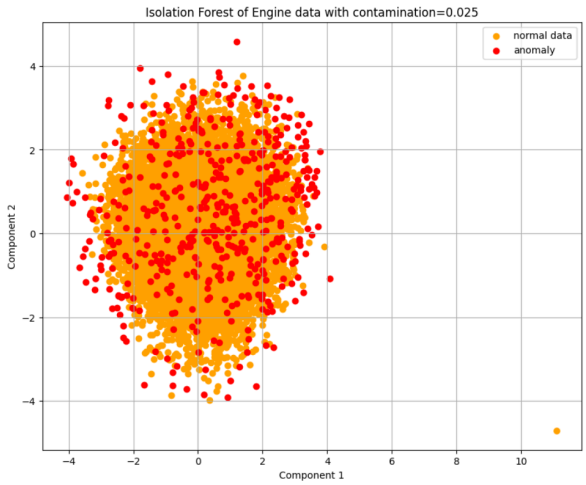
For Isolation, the contamination parameter allows the user to specify the ratio of anomalies to be identified in the data. Hence, any value between 0.01 and 0.05 was appropriate.

Visualising between the One-class SVM and Isolation Forest models that identifies 2.5% of the data as anomalies. The graphs were created by plotting the output of a 2-component PCA and the model outputs.

Figure 2: Visualisation of One Class SVM (A) and Isolation Forest (B) in 2D using a 2-component PCA of the engine data

A B





Conclusion

In conclusion, either of the Machine learning model methods are ideal. One-class SVM was able to identify the data point at the extreme end as an outlier. However, Isolation Forest was faster, handled large data well and is more robust in handling different data scales. The IQR Method, while not requiring any hyperparameter tuning, is dependent on univariate analysis to flag each feature as an anomaly before it can be applied to the entire dataset, which is not ideal for creating an anomaly detection system. Lastly, further engagement would be required with the data owners to validate the entries flagged as anomalies in order to decide the best performing method.