Detecting the anomalous activity of a ship’s engine

Fasone Alessandro

07/06/2025

Table of content

Problem statement

The company requires a robust anomaly analysis and detection model for its ship engines due to the critical risks of unmanaged malfunctions, including crew safety hazards, operational disruptions, and significant financial losses. This initiative aims to promptly identify potential failures, allowing the company to prioritize high-risk cases and intervene proactively. Effective maintenance will minimize unplanned interventions, optimize costs, and enhance operational efficiency. The project seeks to develop a system that not only detects anomalies but also provides actionable insights to improve overall fleet functionality.

Methods

Our approach involved breaking down the problem into sequential phases, culminating in the evaluation of various anomaly detection models.
Initially, we conducted an exploratory data analysis (EDA), both statistical and visual, to understand the distribution of the 19,535 dataset values, identify missing or duplicate entries, and detect preliminary anomalies.
Following the EDA, three distinct anomaly detection approaches were applied, utilizing various statistical and machine learning models to identify outliers. These approaches are summarized in Table 1:

Table 1: Anomaly detection methods

Model	Description	Parameters
IQR Method	Identify anomalies as observation that fall out the lower and upper bound, defined as Q1 - 1.5IQR and Q3 + 1.5IQR	
One-class SVM	Detects anomalies by constructing a separation hyperplane that encloses normal data points in the feature space. Observations outside this decision boundary are classified as outliers	nu and gamma are the parameters tested with these combinations: nu = 0.02 & gamma = 0.1 nu = 0.02 & gamma = 0.2 nu = 0.04 & gamma = 0.2
Isolation Forest	Identifies anomalies by isolating observations using random trees; anomalies show significantly shorter paths	contamination is the only parameter tested in this case Contamination = 0.02, 0.03, 0.05

For the IQR Method, anomalies were initially identified per single column. Subsequently, observations with multiple anomalies were considered, leading to a consolidated evaluation of total anomalies and their percentage.

For machine learning models (as summarized in Table 1), various parameter combinations were tested to ensure robust and reliable results. Each outcome was then graphically represented in a two-dimensional space using PCA on the original dataset for visualization.

Results

From the initial Exploratory Data Analysis (EDA) (Figure 1), after verifying the absence of null values and duplicate rows, we observe that, even if similar, the mean consistently exceeds the median across all variables, suggesting the presence of anomalies in the right tail of their distributions. This hypothesis is further supported by the variable distributions, which largely appear normal, with the exception of Lub oil pressure and Lub oil temp displaying bimodal patterns. Anomalies are present in every variable, being particularly pronounced for Engine rpm, Fuel pressure, and Coolant temp. These three variables were also identified as critical based on their 'extremity index' (Table 2), calculated from the proportion of observations exceeding the 95th percentile relative to their standard deviation.

Figure 1: Distribution and outliers for each feature

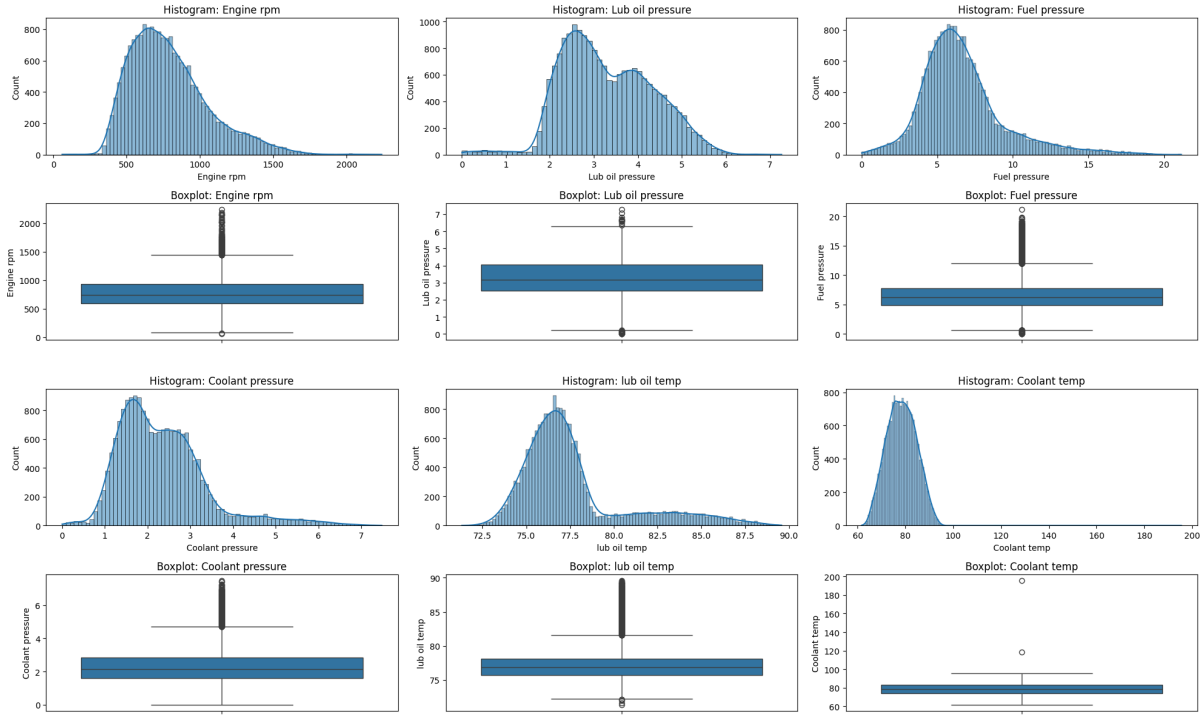


Table 2: Statistical informations for each feature with extremity_index

	mean	median	std	percentile_95	max	extremity_index
Engine rpm	791.24	746.00	267.61	1324.00	2239.00	3.42
Lub oil pressure	3.30	3.16	1.02	5.06	7.27	2.17
Fuel pressure	6.66	6.20	2.76	12.21	21.14	3.24
Coolant pressure	2.34	2.17	1.04	4.44	7.48	2.92
lub oil temp	77.64	76.82	3.11	84.94	89.58	1.49
Coolant temp	78.43	78.35	6.21	88.61	195.53	17.22

IQR Method: The IQR method identified 23.73% (4636 observations) of anomalies when assessed univariately, exceeding expectations. Aligning with business requirements for simultaneous anomalies, a multivariate analysis showed a maximum of 3 concurrent anomalies, occurring in only 0.1% of cases. However, considering at least 2 simultaneous anomalies, the outlier percentage (422 observations, 2.16%) falls perfectly within the suggested 1-5% range.

One-class SVM:For the One-Class SVM method, 25 combinations of nu and gamma parameters were initially tested. The number and percentage of identified outliers for each combination were recorded, leading to the selection of three specific combinations, which are graphically represented through PCA dimensionality reduction (Table 3).

Table 3: Parameter combination for one-class SVM

Model	Nu	Gamma	Anomalies
ocsvm	0.02	0.1	392 (2.01%)
ocsvm 1	0.02	0.2	389 (1.99%)
ocsvm 2	0.04	0.2	790 (4.04%)

Isolation forest: In the case of the Isolation Forest ML model, which allows pre-setting the expected anomaly percentage, it was tested for three different percentages within the 1-5% range, as detailed in Table 4. The number of estimators remained constant for these tests and equal to 100.

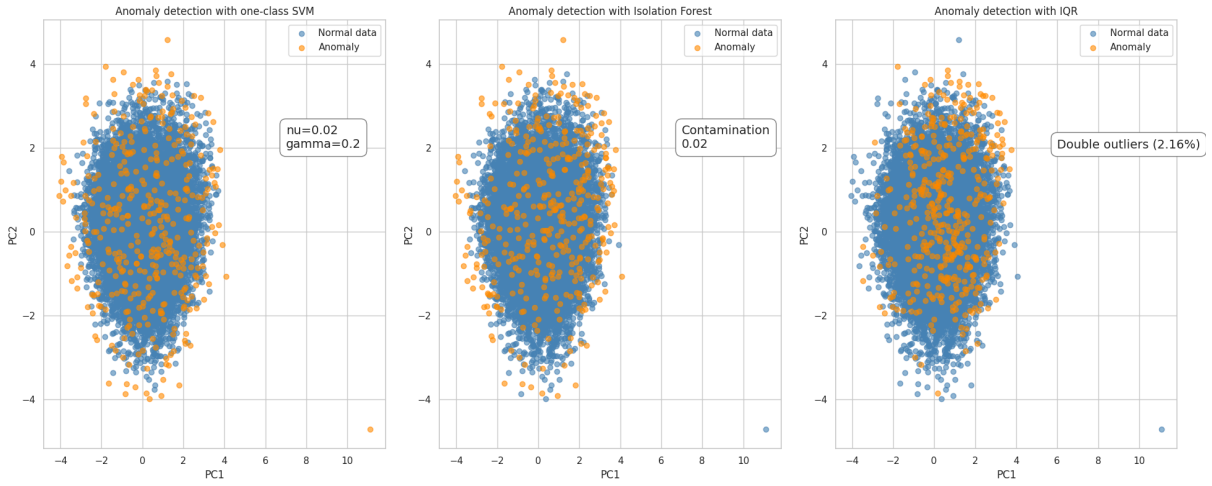
Table 4: Contamination parameter for Isolation Forest

Model	Contamination	Anomalies
iso_forest	0.02	391 (2%)
iso_forest 1	0.03	587 (3%)
iso_forest 2	0.05	977 (5%)

After applying PCA for 2D visualization, models with similar anomaly percentages (around 2%) were selected and compared (Figure 2). Across all models, a clear separation between anomalies and normal data was not observed, which might stem from visualization distortion due to PCA. Despite similar total anomaly counts, the three models shared only 106

common anomalies. Machine Learning models (One-Class SVM and Isolation Forest) shared 228 anomalies, while the IQR method and One-Class SVM showed greater differences, with only 123 shared anomalies. These results suggest a higher consistency between the ML models but highlight the critical need for a more accurate anomaly identification approach.

Figure 2: Anomaly detection through different models



Conclusion

In conclusion, both Machine Learning models proved effective for anomaly identification within this company's practical context. The IQR method is less suitable for multivariate assessments, primarily excelling in univariate anomaly detection.
Given that not all variables exhibit a normal distribution (e.g., Lub oil pressure, Lub oil temp), Isolation Forest emerges as a particularly advantageous model. Its non-parametric nature and faster implementation make it a robust solution, easily adaptable for the company in similar contexts.

Implementing this model in production will enable the company to optimize maintenance interventions, reducing time and operational costs. This will translate into fewer breakdowns, lower fuel and oil consumption, and consequently, increased customer satisfaction due to more punctual deliveries.