MINI-PROJECT 5.3 DETECTING THE ANOMALOUS ACTIVITY OF A SHIP'S ENGINE

Juan Pablo Salazar

2024/12/09

FOURTH REV

UNIVERSITY OF CAMBRIDGE
INSTITUTE OF CONTINUING EDUCATION

I. Introduction and problem statement:

A shipping supply chain company seeks to detect anomalous engine functionality across its fleet by analyzing data from six critical engine features. These features include *engine revolutions per minute (RPM), lubrication oil pressure, fuel pressure, coolant pressure, lubrication oil temperature, and coolant temperature*, which are indicators of engine health and performance.

Engine malfunctions can pose significant safety risks, increase fuel consumption, and lead to delayed deliveries, negatively impacting revenue and customer satisfaction. To address these issues, this project aims to develop an anomaly detection system capable of predicting engine problems and enabling timely maintenance. By employing exploratory data analysis combined with basic statistical methods and machine learning models, the company will be able to identify malfunctioning engines, minimize costs associated with downtime and repairs, and prevent accidents.

II. Data description and visualization

Table I: Data description of the six engine features.

Engine feature	Engine rpm	Lubricant oil pressure	Fuel pressure	Coolant pressure	Lubricant oil temperature	Coolant temp
Count	19535.00	19535.00	19535.00	19535.00	19535.00	19535.00
Mean	791.20	3.30	6.70	2.30	77.60	78.40
Standard Dev.	267.60	1.00	2.80	1.00	3.10	6.20
Min value	61.00	0.00	0.00	0.00	71.30	61.70
5%	443.00	1.90	3.10	1.10	74.30	68.40
25%	593.00	2.50	4.90	1.60	75.70	73.90
50%	746.00	3.20	6.20	2.20	76.80	78.30
75%	934.00	4.10	7.70	2.80	78.10	82.90
95%	1324.00	5.10	12.20	4.40	84.90	88.60
Max value	2239.00	7.30	21.10	7.50	89.60	195.50
% of outliers	2.38%	0.34%	5.81%	4.02%	0.01%	13.40%

When analyzing the data description in Table 1, we can observe key statistical metrics of the six engine features, including the mean, standard deviation, minimum and maximum values, and the 5th, 25th, 75th, and 95th percentiles. The dataset contains information collected from 19,535 ships.

The *Engine RPM* feature has a mean value of 791.2 rpm, while the median is 746 rpm. The 5th and 95th percentiles are 443 rpm and 1324 rpm, respectively, which indicates the range where most engine RPM values fall. For the *Lubrication Oil Pressure* feature, the mean value is 3.3 a.u., with a median of 3.20 a.u.. The 5th and 95th percentiles are 1.90 a.u. and 5.10 a.u., respectively, showing the range of extreme oil pressure readings.

From Figure I, we observe that the distribution of the *Engine RPM* feature is slightly right-skewed but could still be considered approximately normal. Its corresponding boxplot highlights numerous outliers, which will be identified using the IQR method. The remaining distribution plots and boxplots are available in the Colab notebook. The distribution of the other five features also seems right-skewed, with several data points falling outside the lower and higher range limits.

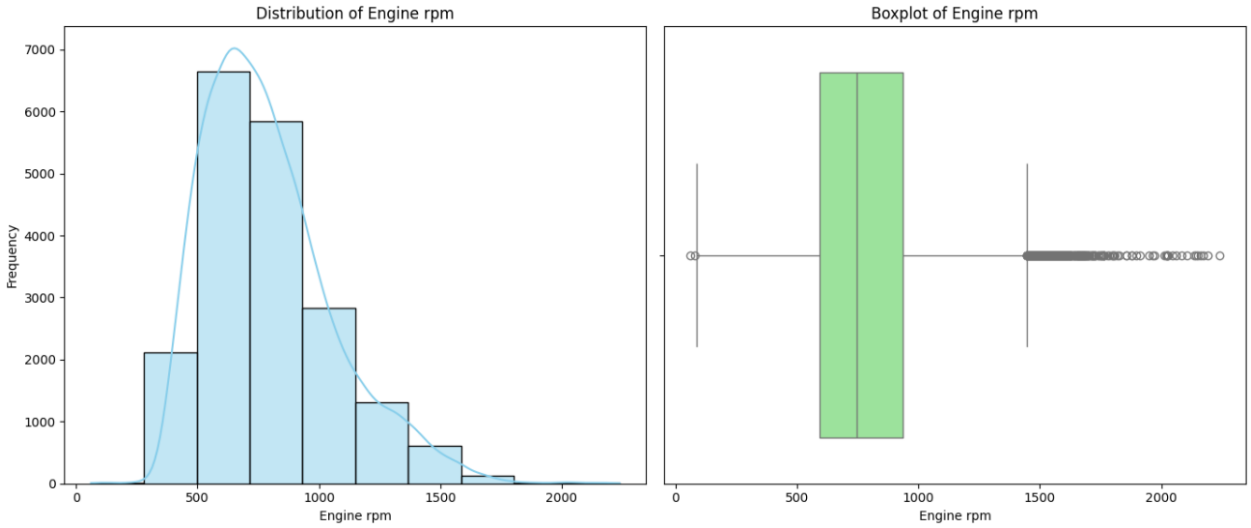

Figure I: Histogram of the distribution of Engine Rpm and its respective boxplot

III. Anomaly detection with the IQR method

For each feature, the percentage of outliers falling outside the lower and upper range limits was calculated and summarized in Table 1. Among the features, *Coolant Temperature* had the highest percentage of outliers at **13.40%**, while *Lubricant Oil Temperature* had the lowest at **0.01%**.

Nonetheless, having abnormal values in a single feature does not necessarily indicate a malfunctioning ship. To classify a ship as anomalous or malfunctioning, at least two features must exhibit abnormal values. Using the IQR method, we identified the percentage of ships with at least

two features outside the upper and lower range limits. This analysis revealed that **2.16%** of the fleet falls into this category, aligning with the client's expectation of anomalies affecting between **1% and 5%** of the ships.

IV. Anomaly detection with machine learning models

a) Anomaly detection with the one SVM machine learning model

The Support Vector Machine (SVM) method identifies anomalies by separating normal data points from potential outliers using a hyperplane in a high-dimensional space, effectively capturing multivariate relationships. Initially, we employed the one-class SVM method with an RBF kernel function, using the default settings where gamma was set to **0.5** and nu was set to **0.05**, identifying **5.34%** of the fleet as anomalies. To reduce this percentage and align with the client's expectations, the nu value was decreased to **0.025**, resulting in **3.42%** of ships being identified as anomalies, which is within the expected range. Finally, in a third approach, we adjusted gamma to **0.3** and nu to **0.025**, identifying **2.59%** of the fleet as anomalies.

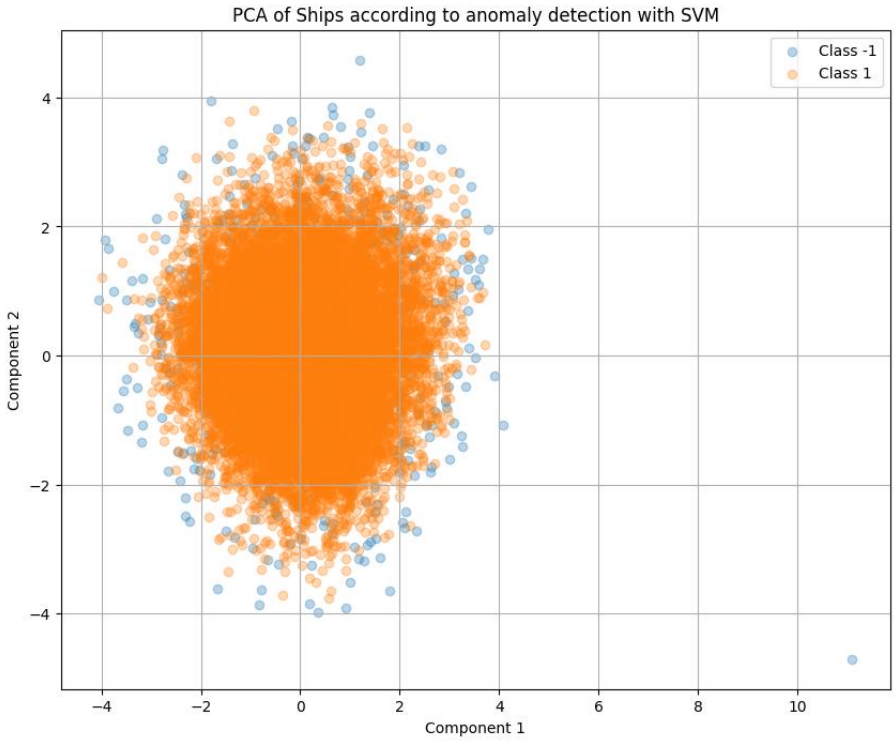

Figure II. PCA of ships according to anomaly detection with one SVM

Since the second and third approaches using the SVM method align with the client's expectation of 1% to 5%, either of these approaches is suitable for the client's expectations. For the PCA analysis shown in Figure II, the third approach has been chosen to visualize the anomalies in the data. From Figure II, we can observe that the data points agglomerate around a single cluster, with most anomalies represented by the blue dots labeled as class -1 located on the periphery of the cluster.

b) Anomaly detection with the isolation forest machine learning model

Isolation Forest is an unsupervised machine learning algorithm used for anomaly detection that isolates anomalies directly without the need to profile normal data points. We applied the Isolation Forest method with 100 estimators and a **0.05** contamination value, which identified **5%** of the ships as anomalies. In a second application of Isolation Forest, we adjusted the contamination value to **0.025**, identifying **2.5%** of the samples as anomalies. Both Isolation Forest approaches align with the client's expectations. The Isolation Forest with a 0.05 contamination value was used to plot Figure III. The plot is very similar to the PCA plot from the one-class SVM (fig. II), with most anomalies located on the peripheric of the cluster.

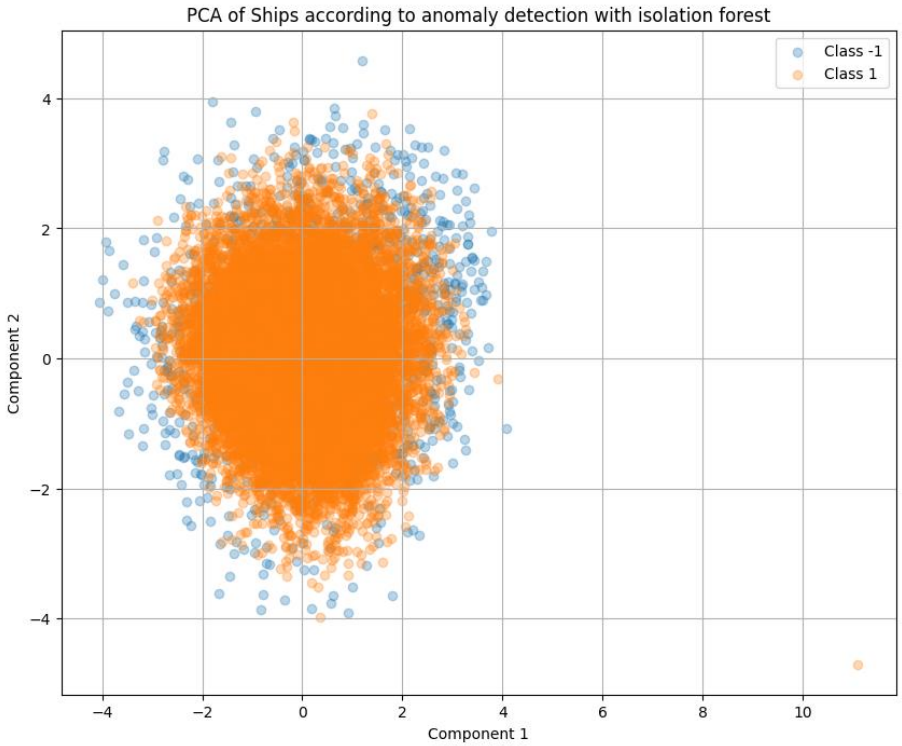

Figure III. PCA of ships according to anomaly detection with the isolation forest method

Conclusion:

The three methods employed for anomaly detection are all suitable for the client's expectations. The IQR method provides a simple and interpretable way of identifying outliers, although it does not consider the relationships between features. The One-Class SVM and Isolation Forest methods, on the other hand, account for multivariate relationships between the features. However, the One-Class SVM method is sensitive to outliers, computationally expensive, and heavily dependent on parameter tuning. In contrast, Isolation Forest is faster to train and scalable for larger datasets. Finally, identifying samples flagged as anomalous by all three methods could provide a more robust approach.