Detecting Anomalous Activities of Ship's Engine

By Shishir. M. Paltanwale

Date: 28th March 2025

TABLE OF CONTENTS

PROBLEM STATEMENT

The company has faced significant losses due to unforeseen engine breakdowns, resulting in shipment delays and jeopardizing crew safety. To address this, the business aims to implement predictive maintenance for engines. The objective of this initiative is to identify and prioritize engines for routine maintenance, minimizing breakdowns and ensuring the seamless operation of the fleet.

DATA DESCRIPTION

The dataset consisted of 19535 records with 6 columns(features), which are as below:

1. Engine rpm
2. Lub oil pressure
3. Fuel pressure
4. Coolant pressure
5. Lub oil temperature
6. Coolant temp

OVERVIEW APPROACH

The first step involved **importing the necessary libraries** and loading the dataset. Next, **exploratory data analysis (EDA)** was performed, which included calculating the **mean, median, 95th percentile**, and identifying **missing or duplicate values** for each feature.

Following EDA, the data was visualized using **histograms and boxplots** to understand its distribution. **Outliers were identified using the interquartile range (IQR) method**, followed by **standardization** to ensure uniform scaling. **PCA scaling** was applied both for **data visualization** and for improving anomaly detection using **One-Class SVM** and **Isolation Forest**.

Various parameter combinations were explored for both methods, and the detected anomalies were analysed. The **One-Class SVM** detected **1,043 anomalies**, while **Isolation Forest** detected **977 anomalies**.

METHODS

Explanation and Insights to methods used:

1. **Dimensionality Reduction Principal Component Analysis (PCA)**

   - PCA is a dimensionality reduction method that transforms high-dimensional data into fewer dimensions while preserving maximum variance. It uses **linear algebra** to compute principal components as linear combinations of the original features.
   - For anomaly detection:
     - PCA simplifies visualization by reducing features to **2D or 3D** space.
     - Anomalies tend to deviate from clusters of normal points in this reduced space.

2. **One-Class SVM**

   - This is a kernel-based algorithm that identifies anomalies or novel patterns.
   - The algorithm:
     - Maps data into a higher-dimensional feature space.
     - Constructs a **decision boundary** around the normal data points.
     - Flags points outside this boundary as **anomalies**.
   - Hyperparameters like **gamma** (kernel influence) and **nu** (proportion of anomalies) directly impact the performance.

3. **Isolation Forest**

   - Isolation Forest uses random binary trees to isolate data points. Anomalies, due to their rarity or uniqueness, are isolated with fewer splits compared to normal points.
   - Unlike PCA, Isolation Forest can work directly on high-dimensional data without requiring scaling.

4. Standard Scaling

- is a feature scaling technique that transforms data to have a mean of 0 and a standard deviation of 1. This ensures that features with varying ranges and units are standardized, allowing them to contribute equally to model training.
- Features measured in different units (e.g., age in years vs. income in dollars) can disproportionately impact models like SVM or K-Means. Standard scaling mitigates this issue.

Xscaled=X−μσX_{\text{scaled}} = \frac{X - \mu}{\sigma}

After scaling:

- Mean (μ\mu) becomes 0.
- Standard deviation (σ\sigma) becomes 1.

5. Interquartile Range (IQR)

- measures the spread of the middle 50% of a dataset. It is the difference between the third quartile (Q3) and the first quartile (Q1)
- Since the IQR uses quartiles instead of the mean or standard deviation, it is unaffected by skewed data distributions.

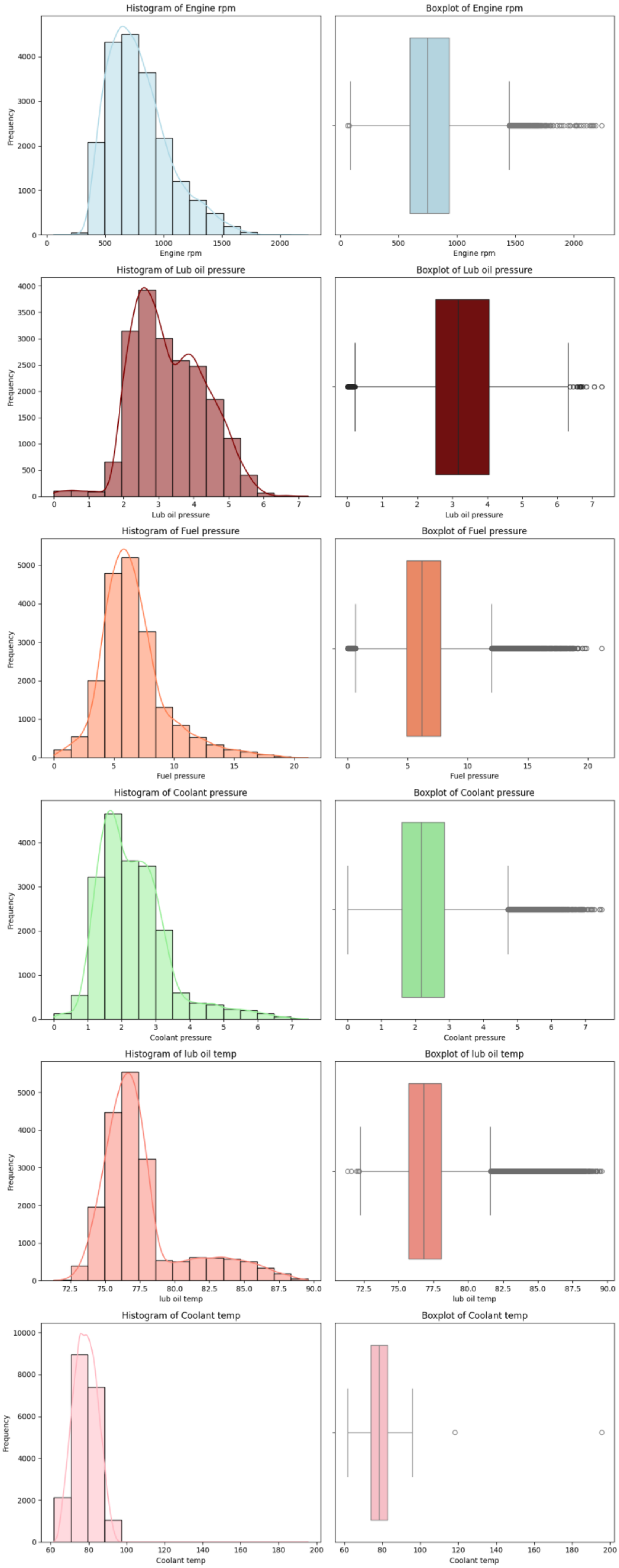
i. Lower limit is calculated as (Q1 – 1.5 * IQR)

ii. Upper limit is calculated as (Q3 – 1.5 * IQR)

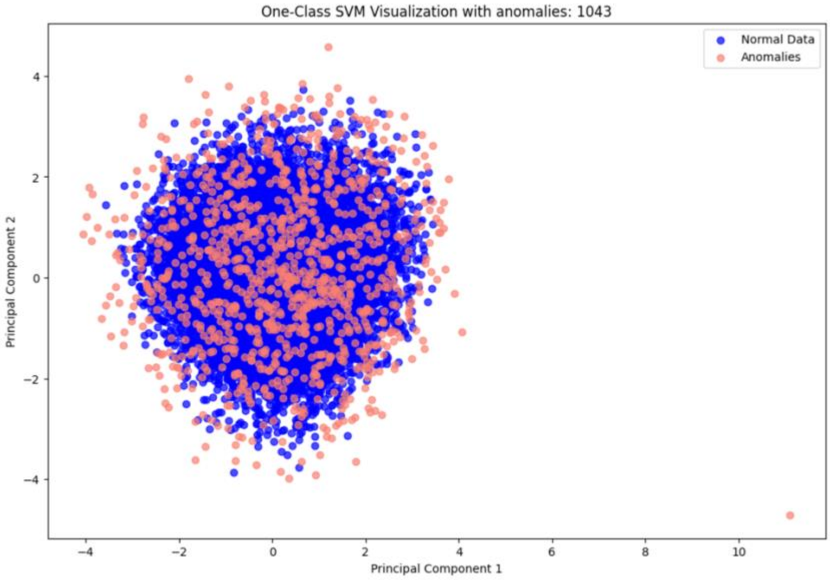
Anything above OR below these limits are anomalies

DATA VISUALISATION

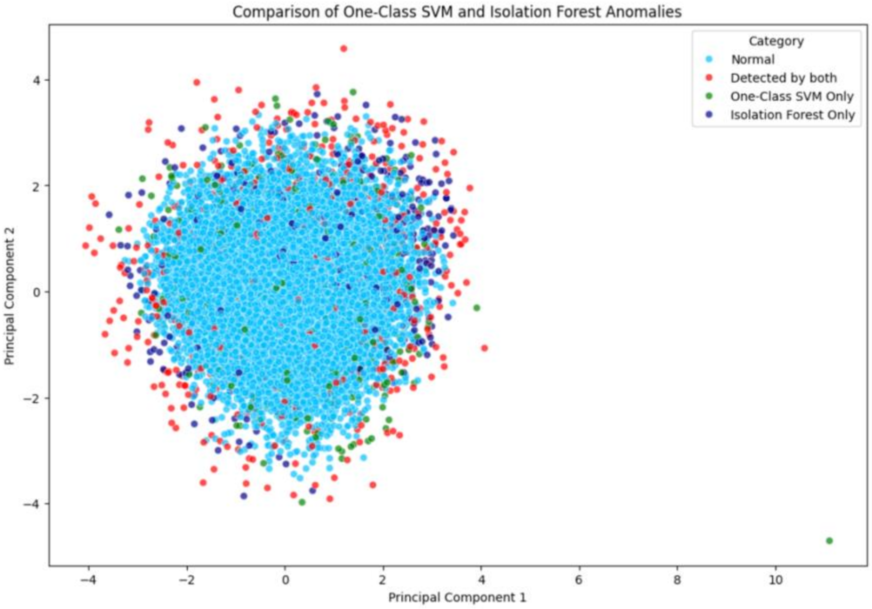
1. Histograms and Boxplots (all features)


Histogram of Engine rpm
Frequency
0
1000
2000
3000
4000
0
500
1000
1500
2000
Engine rpm
Boxplot of Engine rpm
0
500
1000
1500
2000
Engine rpm
Histogram of Lub oil pressure
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
0
1
2
3
4
5
6
7
Lub oil pressure
Boxplot of Lub oil pressure
0
1
2
3
4
5
6
7
Lub oil pressure
Histogram of Fuel pressure
Frequency
0
1000
2000
3000
4000
5000
0
5
10
15
20
Fuel pressure
Boxplot of Fuel pressure
0
5
10
15
20
Fuel pressure
Histogram of Coolant pressure
Frequency
0
1000
2000
3000
4000
0
1
2
3
4
5
6
7
Coolant pressure
Boxplot of Coolant pressure
0
1
2
3
4
5
6
7
Coolant pressure
Histogram of lub oil temp
Frequency
0
1000
2000
3000
4000
5000
72.5
75.0
77.5
80.0
82.5
85.0
87.5
90.0
lub oil temp
Boxplot of lub oil temp
72.5
75.0
77.5
80.0
82.5
85.0
87.5
90.0
lub oil temp
Histogram of Coolant temp
Frequency
0
2000
4000
6000
8000
10000
60
80
100
120
140
160
180
200
Coolant temp
Boxplot of Coolant temp
60
80
100
120
140
160
180
200
Coolant temp

2. PCA data visualisation.



3. Comparison between One-Class SVM and Isolation Forest methods



RESULT

The anomalies detected are as below:

1. Interquartile Range (IQR)

A total of **4,636 anomalies** were detected, accounting for **23.7%** of the dataset.

Among these:

- **Univariate anomalies**: **4,214 (21.57%)** – anomalies detected in a single feature.
- **Multivariate anomalies**: **422 (2.2%)** – anomalies present in more than one feature.
    - **411 records** had anomalies in **more than one features.**
    - **11 records** had anomalies in **more than two features**.
    - No observations exhibited anomalies in **more than three features**.

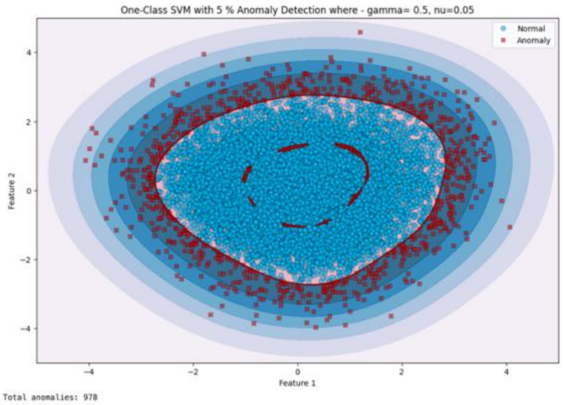
2. One-Class SVM

The anomaly detection results from **One-Class SVM** aligned well with the assumed **1–5% range**. The table below presents the closest anomaly count obtained after fine-tuning the **Gamma** and **Nu** parameters.

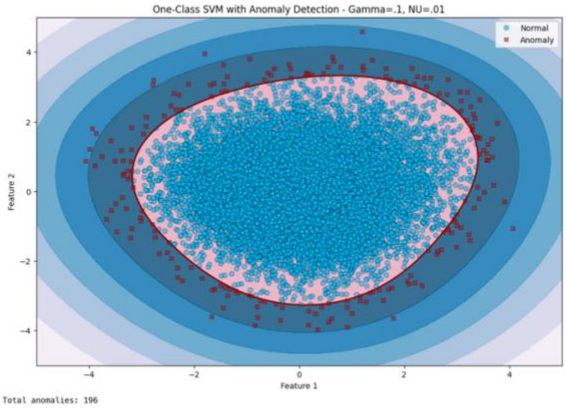
Table for One-Class SVM readings

Gamma	Nu	Actual Anomaly Observations	% of Dataset (19535)
0.5	0.05	978	976.5 (5%)
0.4	0.04	785	781.4 (4%)
0.3	0.03	592	586.05 (3%)
0.2	0.02	391	390.7 (2%)
0.1	0.01	196	195.35 (1%)

<u>One-Class SVM Anomaly detection</u>



Gamma = .5, Nu = 0.05



Gamma = 0.1, Nu = 0.01

3. Isolation Forest

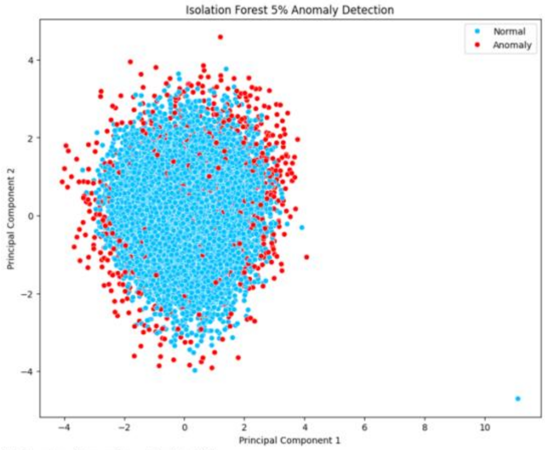
The **contamination factor** was adjusted to obtain an **anomaly count closest to the expected 1–5% range**. The visualizations show that **outliers (marked in red) tend to appear at the edges of the clusters**.

To achieve this, the data using PCA was **reduced from 6 to 2 features** for **2D visualization**. However, this dimensionality reduction makes it **impossible to pinpoint which specific features contribute to an anomaly** in the original dataset.

Table with Isolation Forest readings

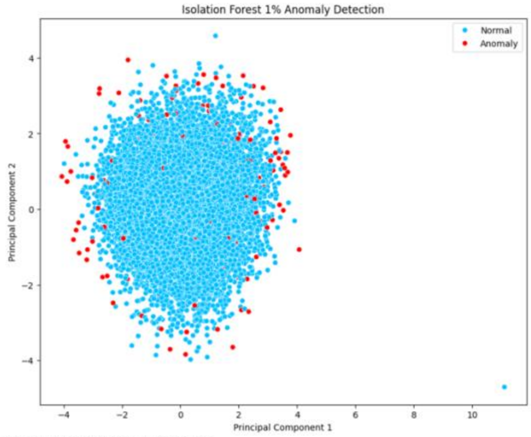
Contamination	Actual Anomaly Observations	% of Dataset (19535)
0.5	977	976.5 (5%)
0.4	782	781.4 (4%)
0.3	587	586.05 (3%)
0.2	391	390.7 (2%)
0.1	196	195.35 (1%)

Isolation Forest Anomaly detection



```
Total number of anomalies detected: 977
```

Contamination = 0.5



```
Total number of anomalies detected: 196
```

Contamination = .1

SUMMARY

To analyse the data and draw meaningful conclusions, **various anomaly detection methods** were applied. The dataset consisted of **19,535 records**, with the assumption that anomalies—representing **engines potentially requiring routine maintenance**—would account for approximately **1–5%** of the total dataset.

The **Interquartile Range (IQR) method** was not used, as it is designed to detect anomalies in **individual features** rather than identifying patterns across **multiple features**.

- **One-Class SVM** initially detected **1,043 anomalies**, which was later refined to **978 anomalies (5% of the dataset)**.
- **Isolation Forest** detected **977 anomalies**.

A key observation was the **significant number of uncommon anomalies** between the two methods, as illustrated in **point 3 of the Data Visualization section**, highlighting their **differences in anomaly identification**.

While **One-Class SVM** proved to be highly effective, **Isolation Forest** produced results that were **closer to the expected anomaly range** based on the initial assumption.

REFERENCES

Devabrat, M., 2022. Predictive Maintenance on Ship's Main Engine using AI. Available at: https://dx.doi.org/10.21227/g3za-v415. [Accessed 5 March 2024]