DETECTING THE ANOMALOUS ACTIVITY OF A SHIP'S ENGINE

Stakeholder Report

Prepared by Matt Cocker

Contents

Introduction

The business requires a robust anomaly detection system to evaluate engine functionality. This system is capable of identifying engines with potential issues, allowing the business to schedule specific maintenance for the engines that need it and prevent malfunctions. Keeping the engines well maintained will increase operational efficiency, ensure timely delivery, and therefore improve customer satisfaction and profits.

Figure 1: *Project Flow Chart*

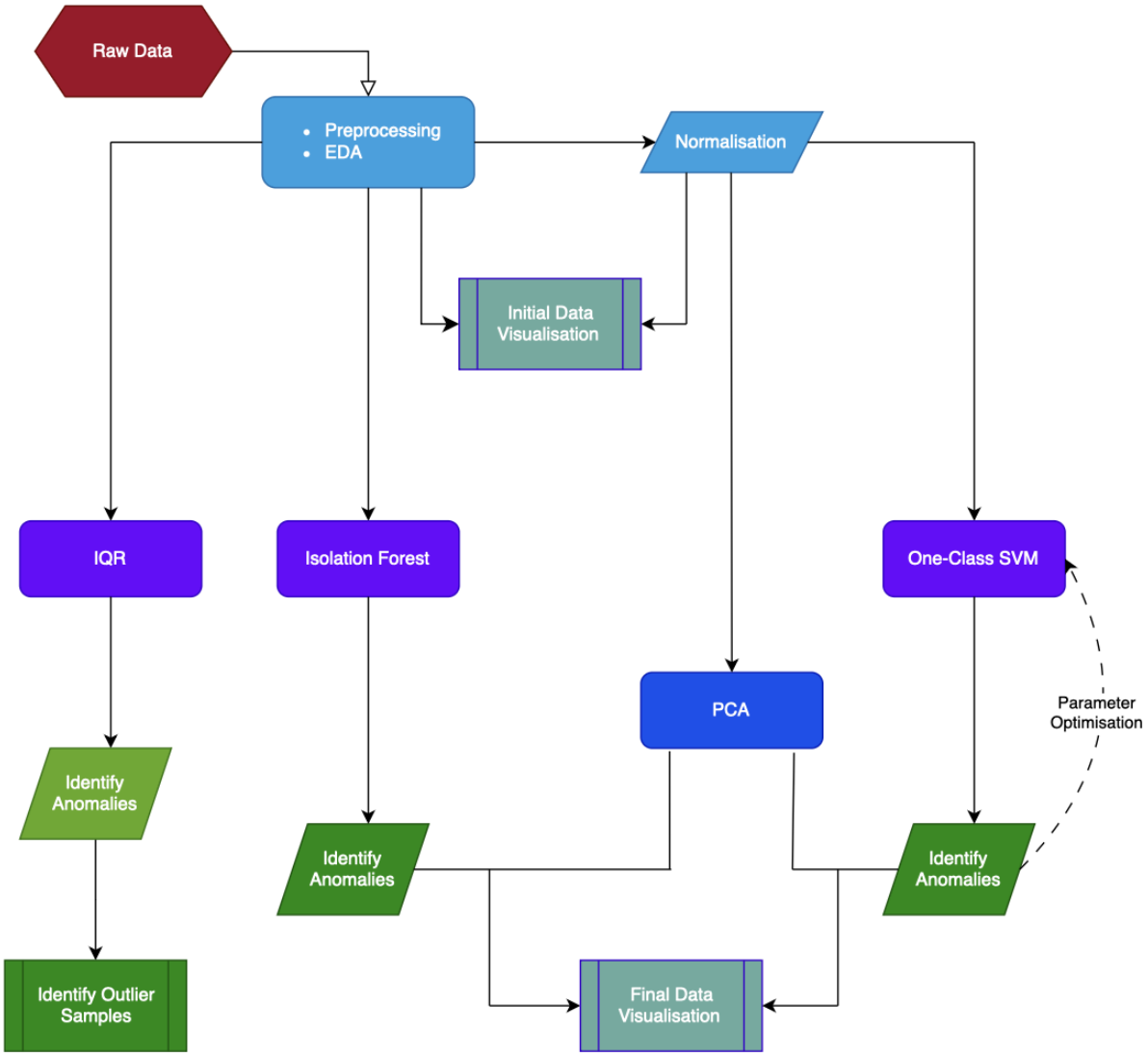


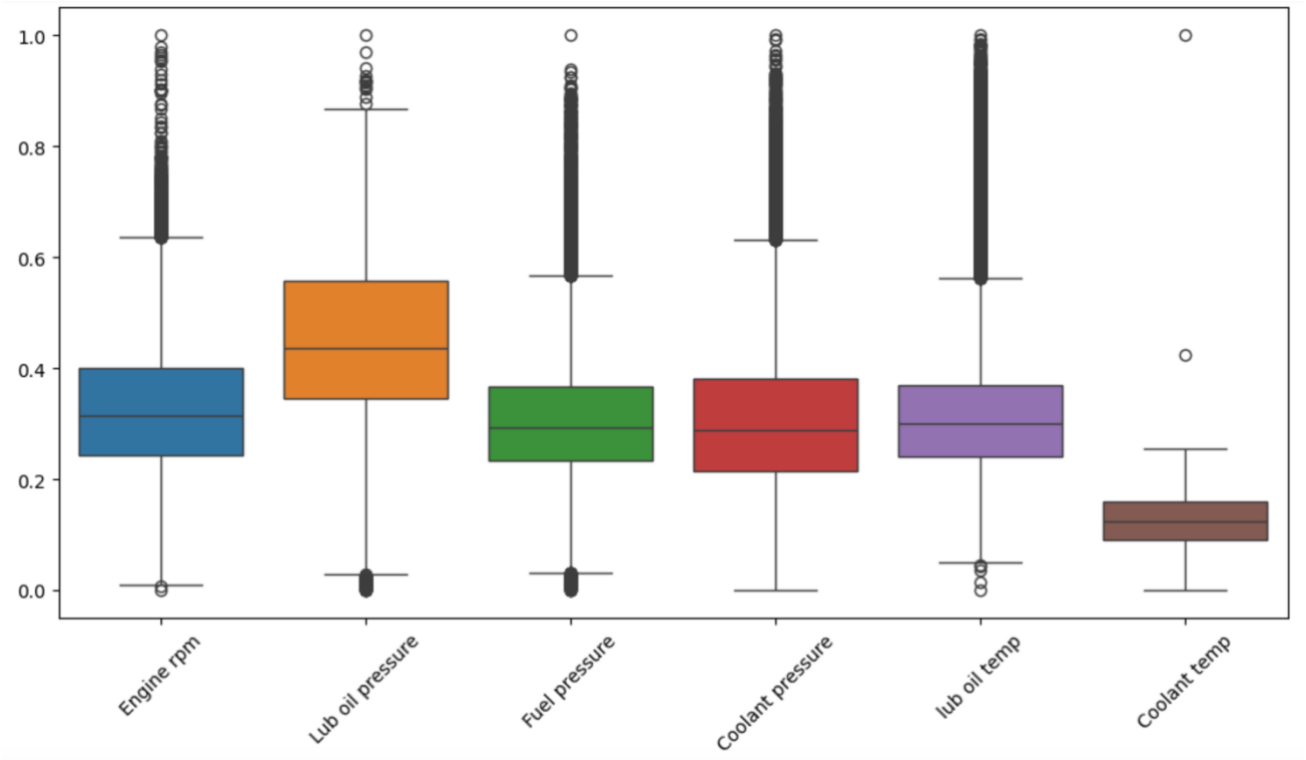
Figure 1 describes the process of creating the anomaly detection system. Data preprocessing ensured there were no missing or duplicate values. Initial data was plotted, and descriptive statistics were generated, to enable understanding of the dataset. Three anomaly detection methods were then tested. Principal Component Analysis (PCA) was performed to reduce the dimensions of the dataset, allowing it to be visualised in 2D. These charts were analysed and the results of the three methods were compared.

Data

The dataset contained 19,535 samples, each showing data for 6 features.

During the Exploratory Data Analysis (EDA) stage, the data was visualised in *Figure 2.* The box plot shows that the Fuel Pressure, Coolant Pressure and Lubricant Oil Temperature features have a higher proportion of outliers in the data. This may suggest that engine malfunctions are more likely to be caused by certain specific errors; it would be helpful to discuss this with the company's engineers.

Figure 2: *Feature box plots – normalised data*



Methods

Three anomaly detection methods were tested, 1 statistical and 2 machine learning (ML). Typically, anomalies would make up between 1% and 5% of the dataset, therefore the models were calibrated to ensure the number of anomalies detected fell within this range.

IQR

Anomalies were detected using the Interquartile Range (IQR) method, where the IQR is defined as the range (Q3 – Q1) between the 25th percentile (Q1) and the 75th percentile (Q3), and using multivariate analysis. Anomalous datapoints (outliers) are calculated as data points that fall below the lower bound (Q1 - 1.5 * IQR) or above the upper bound (Q3 + 1.5 * IQR) in a feature.

4,636 samples (23.7%) had an outlier in at least one feature, 422 samples (2.16%) had outliers in at least two, and 11 samples (0.06%) had outliers in three or more.

Consequently, as 2.16% is within the 1% to 5% range, samples which contained **2 or more outliers** were classified as anomalies.

One Class SVM

Anomalies are calculated using the One-Class SVM method by training a model to learn the boundary of the normal data and identifying data points that fall outside this boundary as anomalies. To ensure the number of anomalies detected fell in the expected range, parameter optimisation was performed. Two parameters were investigated, with the results in *Figure 3*:

- Gamma – Impacts the fit of the boundary around the data
- Nu – Impacts the size of the boundary around the data

Figure 3: *One-Class SVM Parameter Optimisation*

Gamma	Nu	No. of Anomalies (Percentage of Dataset)
0.5	0.05	975 (5%)
0.5	0.03	587 (3%)
0.75	0.03	587 (3%)
0.95	0.03	584 (3%

Figure 3 suggests that nu has a greater impact on the number of anomalies generated by the model than gamma. A gamme of 0.75 and a nu of 0.03 ensured the model accurately reflected the shape of the distribution of anomalies, but reduced the risk of overfitting.

Isolation Forest

The Isolation Forest method by generating splits in the dataset and identifying points that require fewer splits to isolate, which are considered anomalies. Our Isolation Forest model only uses one parameter, contamination, which was set at 0.03 to ensure that the model identified 3% of the dataset as outliers. This allowed the results to be compared directly to those from One-Class SVM.

Results

Principal Component Analysis (PCA) was performed to reduce the dimensions of the data, allowing it to be plotted in 2D. The results of the One-Class SVM and Isolation Forest models are shown on *Figure 4* and *Figure 5* respectively.

Figure 4: *One-Class SVM Scatter*

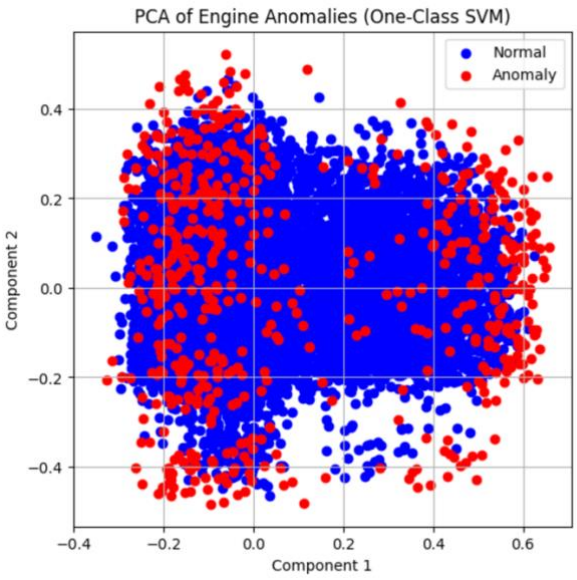
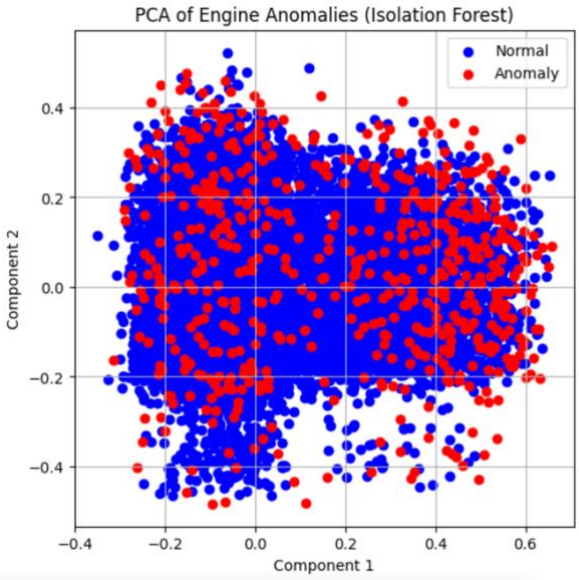


Figure 5: *Isolation Forest Scatter*



Both scatter plots are somewhat uninterpretable. The anomalies in *Figure 4* seem to be mostly concentrated in two vague clusters, one where Component 1 is -0.15 and one where Component 1 is 0.55, but those clusters are not tight. *Figure 5* shows a somewhat similar distribution of anomalies, but with less distinct clustering. As Component 1 and Component 2 are complexly defined, there is a limit to the insight that can be drawn from these charts.

293 samples are classified as anomalies by both ML models. This suggests that, whilst there is significant overlap, care must be taken when choosing which model to use for the anomaly detection system, as the models classify anomalies differently.

Conclusions & Recommendations

Whilst One-Class SVM is a suitable model, this report would recommend Isolation Forest for the anomaly detection system, as it handles large datasets more efficiently, can be easily calibrated to the expected proportion of outliers, and is less sensitive to different scales and outliers. Studies [1, 2, 3] have also shown it to be better at anomaly detection in similar use cases. The IQR method is suited to individual features, making it unsuitable when looking at features which interact with each other.

This system is capable of identifying engines with potential issues and, once in regular use, can be continuously improved to more accurately detect engines that require maintenance. Enabling engineers to schedule timely maintenance on the fleet of engines will increase both the safety of the fleet and operational efficiency. This increased efficiency should ensure timely delivery, which will boost customer satisfaction and improve profits.

References

1. Lukito, K. et. Al. (2023). Comparison of Isolation Forest and One Class SVM in Anomaly Detection of Gas Pipeline Operation. doi:10.1109/ICICyTA60173.2023.10428838.

2. Darrab, S. et al. (2024). Anomaly Detection Algorithms: Comparative Analysis and Explainability Perspectives. doi: 10.1007/978-981-99-8696-5_7

3. Lu Haowen (2025). Evaluating the Performance of SVM, Isolation Forest, and DBSCAN for Anomaly Detection. doi: 10.1051/itmconf/20257004012