Detecting Anomalies in Ship Engine Data using Statistical and Machine Learning Approaches

1 Introduction

The dataset used for this analysis consists of 19,535 entries with six numerical features describing ship engine performance metrics. The primary goal of this project is to identify anomalies that could indicate potential maintenance requirements. A combination of statistical and machine learning methods was employed to detect anomalies, evaluate their effectiveness, and provide actionable insights for ship operators.

The methodology for this experiment starts with an exploratory data analysis (EDA) to understand the data distribution and identify where potential outliers are likely to occur. Feature scaling and dimensionality reduction was conducted to prepare the data for the machine learning approaches. Finally, an evaluation of the ffectiveness of these approaches is presented along with visual representations of the results.

2 Exploratory Data Analysis

Initial inspection revealed that all features were continuous numerical data with no missing or duplicate entries, which simplified data preprocessing. However, the features varied in range, necessitating feature scaling for machine learning models (to avoid unwanted biases for certain features.

Histograms were used to visualise data distributions, highlighting the means, medians, and interquartile ranges (IQRs) (figure 1). Upon inspection of the distributions, it became clear that the data set was not normally distributed. It can be speculated that

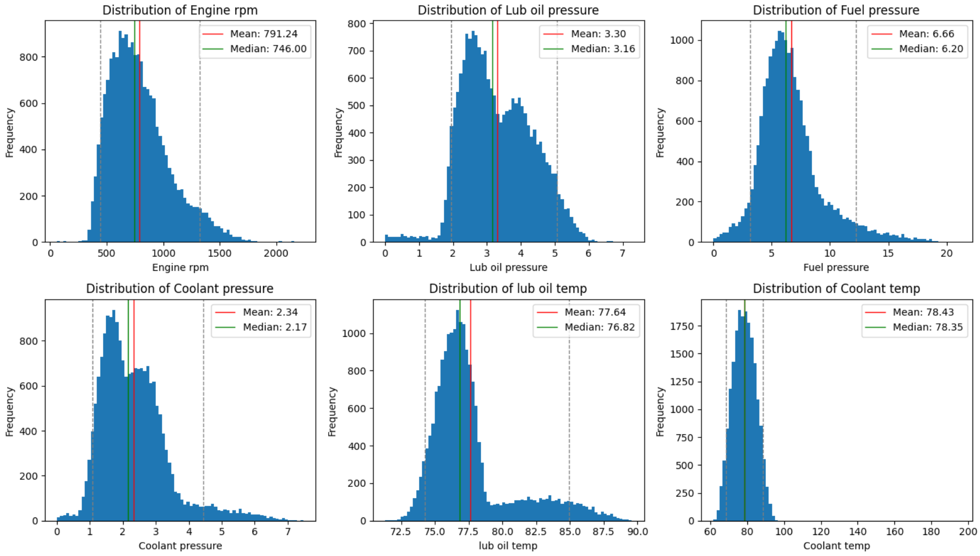

Figure 1: Data Distribution by Feature

the distribution of coolant temperature may be normal, but since the outliers must be included, the distribution becomes skewed. Box plots for each feature gave a closer look at potential outlier, revealing that features such as fuel pressure and lubrication oil temperature contained frequent outliers, suggesting that these values should be closely monitored.

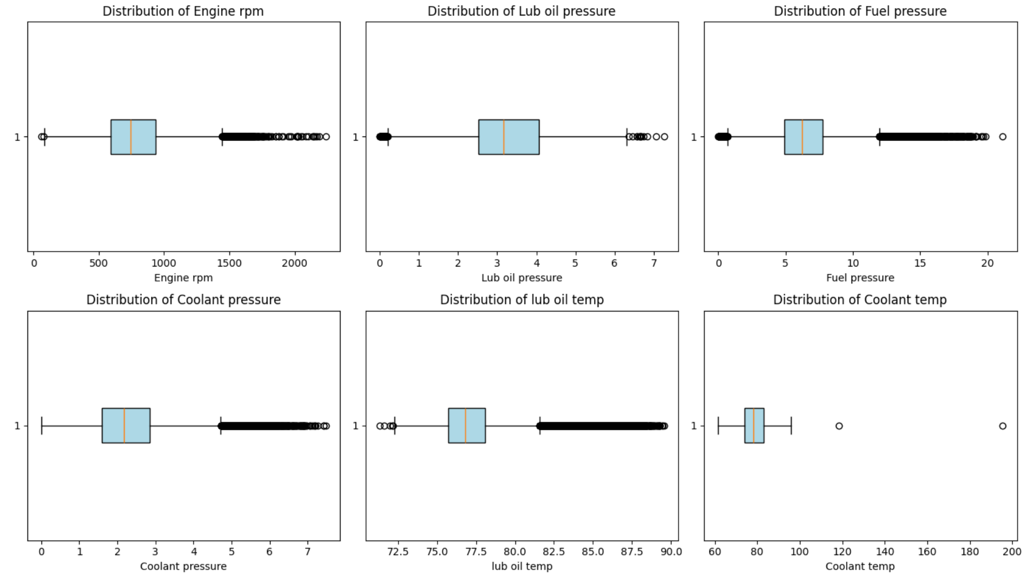

Figure 2: Box Plot by Feature

3 Anomaly detection

3.1 IQR Method

The first approach utilised the Interquartile Range (IQR) method, a statistical technique that identifies anomalies based on their deviation from the central 50% of the data. This method is advantageous for datasets with skewed distributions, as it is less affected by extreme values compared to other statistical approaches such as the Z-score approach.

To classify a data point as anomalous, it needed to be an outlier for at least two features simultaneously. This approach identified 2.16% of the data as anomalous, aligning with the expectations for this data set. The IQR method provided a simple and effective approach for anomaly detection, particularly for univariate analysis. As each feature had all their values independently identified.

3.2 Machine Learning Approaches

Machine learning models were implemented to capture more complex relationships in the data, serving as a multivariate approach to anomaly detection. Because the performance of machine learning methods can be sensitive to the shape of the input data, it was essential to conduct some pre-processing of the data.

The first step involved feature scaling to avoid biases due to variance in feature ranges. Z-score standardisation was the approach used for centring the data around 0 with unit variance. This method was suitable for handling outliers and ensured compatibility with machine learning algorithms.

The second step was dimensionality reduction in terms of principal component analysis (PCA). Reducing the dimensions from six features to two principal components helped remove redundant correlated features and provided the possibility for two-dimensional visualisation of the model performance.

3.2.1 One-class Support Vector Machine (SVM)

The one-class SVM creates a decision boundary around the normal data point. After dimensionality reduction, the SVM used the radial basis function (RBF) kernel to handle non-linear relationships between the principal components. Two key hyperparameters are what determine the tendencies of the decision boundary: the ν- and γ-values (nu and gamma). The model's tolerance for outliers is determined by ν while γ determines the size of the decision boundary. These values were fine-tuned to $\nu = 0.012$ and $\gamma = 0.1$.

With this hyperparameter configuration, the model classified 1.21% of the data as anomalous, thus consistent with the expectation of the data. As shown by the diagram in figure 3, the decision boundary was well-defined, ensuring no false-positive anomalous zones near the mean.

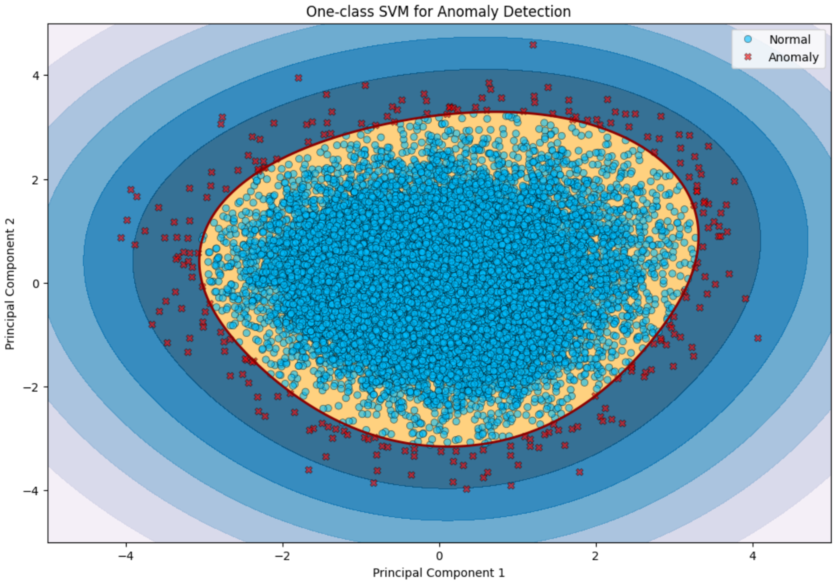

Figure 3: One-class SVM Decision Boundary

3.2.2 Isolation Forest

While the One-class SVM approach defines a decision boundary around the normal data, the Isolation Forest reverse-engineers the process and identifies anomalies directly, without the need for modelling an explicit boundary around the normal data. The two key hyperparameters fine-tuned for this approach are known as the contamination factor and the number of estimators. The contamination determines the proportion of the data that is expected to be anomalous, making it very easy to align this proportion with the expected outcome. The number of estimators decides how many times the data is segmented between outliers and normal data, where a higher value will yield more accurate results, it can become computationally intensive. The model used described in this report had 100 estimators and a contamination factor of 0.03. These values yielded a visually defined boundary and aligned with the expectations for this data set.

This approach identified anomalies with similar effectiveness to the SVM but provided a simpler mechanism for explaining outliers. Visualisations of the isolation shown in figure 4, presents the clearly defined boundary.

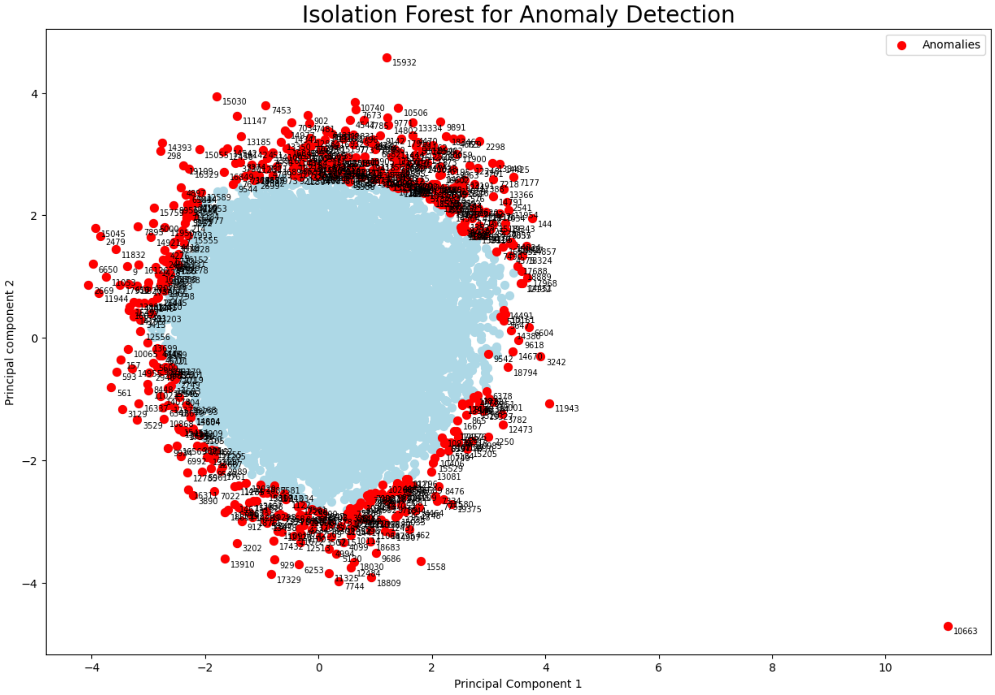

Figure 4: Isolation Forest Boundary

4 Conclusions

The analysis of this report yielded several key insights. Fuel pressure and lubrication oil temperature consistently exhibited outliers, suggesting that these features require close monitoring. Th IQR method provided straightforward thresholds for univariate analysis, while machine learning methods highlighted, multivariate interactions leading to anomalies. Comparing the methods, the IQR method was effective for quick, univariate anomaly detection. The One-class SVM and Isolation Forest provided robust multivariate detection with clear decision boundaries. For anomaly detection in a business context, the Isolation Forest is recommended for its scalability and simplicity. The One-Class SVM can complement this approach for more complex datasets requiring non-linear boundaries.