Isolation Forest Analysis Report

Summary of Findings

1. Contamination Rate and Anomaly Detection

- **Contamination Rate 0.01**:
    - Total Anomalies Detected: **143**
    - This rate detected only the most extreme anomalies, resulting in a small anomaly set. Scores ranged narrowly from approximately -0.09 to 0.29.
    - Lower contamination resulted in higher average anomaly scores (above 0.20), indicating the model only flagged the most unusual outliers.
- **Contamination Rate 0.05**:
    - Total Anomalies Detected: **713**
    - As expected, increasing the contamination rate to 5% identified more anomalies, capturing additional outliers that were slightly less extreme but still significant.
    - The average anomaly score decreased to around 0.14, reflecting a broader anomaly set with some less severe outliers included.
- **Contamination Rate 0.1**:
    - Total Anomalies Detected: **1,425**
    - At 10%, the highest contamination rate included the widest variety of anomalies, with an even lower average score (~0.10).
    - This setting captures both significant and moderate outliers, trading off precision for a higher sensitivity to detect a larger set of outliers.

2. Impact of Number of Estimators on Anomaly Detection

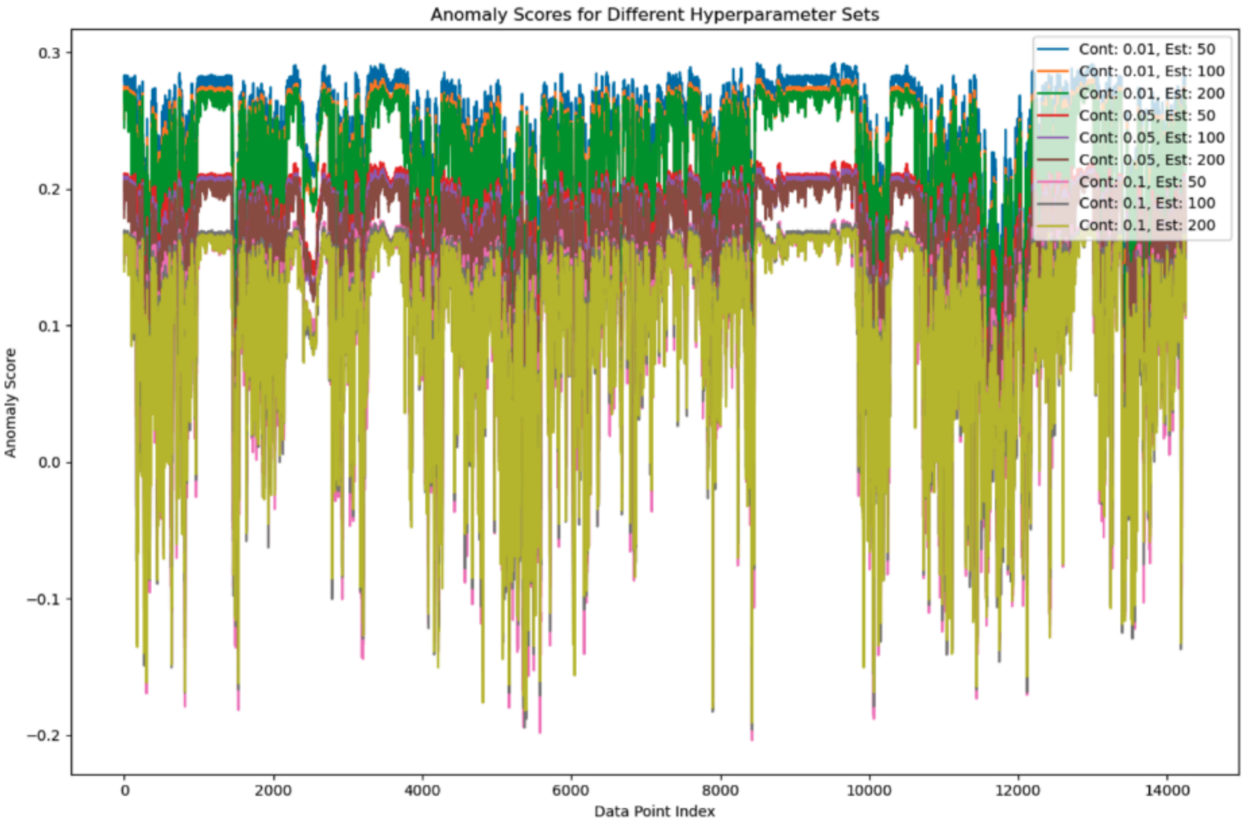
- With **50, 100, and 200 estimators**, the total number of anomalies detected remained **consistent** within each contamination rate (e.g., always 143 anomalies for 0.01 contamination).
- Increasing estimators led to:
    - **More stable anomaly scores**: As the number of estimators increased, there was a slight decrease in the range of scores (i.e., the minimum and maximum values became less extreme).
    - **Reduced score variability**: Higher estimator values reduced the score range slightly, likely due to increased robustness from additional ensemble averaging.
- **Conclusion**: Estimator count mainly affects scoring stability rather than the number of anomalies detected. Adding more estimators helps to fine-tune the model's precision but does not change the total anomalies flagged, as this is controlled by contamination.

3. Comparison of Anomaly Scores across Settings

- **Average Anomaly Scores**:
  - Higher contamination rates resulted in lower average anomaly scores, indicating that the model becomes less stringent about the level of deviation needed to classify a point as an anomaly as more points are flagged.
  - This suggests that a lower contamination rate provides a “sharper” anomaly score distinction, emphasizing only the most significant outliers, while higher contamination broadens the anomaly criteria.
- **Score Range**:
  - Lower contamination rates yielded a wider range in scores (minimum to maximum), while higher rates had more compressed score distributions.
  - As contamination increases, the range of scores narrows, showing a more relaxed standard for what constitutes an anomaly. This indicates that the highest contamination rate (0.1) flags even moderate deviations as anomalies.

Conclusions and Recommendations

1. **Optimal Parameter Setting**:
   - For **precision** in detecting only the most extreme anomalies, a contamination rate of **0.01** with at least **100 estimators** is recommended. This setting minimizes noise while ensuring only significant deviations are flagged as anomalies.
   - If the goal is to **capture a broader range** of outliers (including moderate anomalies), a **0.05 contamination rate** with 100 or 200 estimators provides a balance between specificity and sensitivity.
   - The **0.1 contamination rate** is best if there is a need to capture almost all possible anomalies, though it may include minor deviations and increase false positives.
2. **Estimator Recommendations**:
   - Using **100 to 200 estimators** is optimal for achieving stable and consistent anomaly scores. While 50 estimators provide similar results, using a higher number of estimators reduces variance, particularly important if the model needs to be applied on fluctuating datasets.
3. **Application-Specific Settings**:
   - **Critical Applications**: For applications where false positives are costly, such as quality control in manufacturing, the 0.01 contamination rate with 100+ estimators is ideal.
   - **Broad Outlier Detection**: In cases where capturing a wider set of anomalies is beneficial, such as in exploratory data analysis or preemptive equipment maintenance, the 0.05 rate is effective.



Interpretation and Analysis of Anomaly Score Plot

This plot represents the **anomaly scores** produced by the Isolation Forest model across the dataset for different combinations of **contamination rates** and **estimator counts**. Each colored line corresponds to a specific combination of contamination rate and estimator count, as labeled in the legend.

Key Observations

1. **Anomaly Score Variation Across Different Contamination Rates**:
   - Higher contamination rates (e.g., **0.1**) produce generally **lower anomaly scores** compared to lower contamination rates (e.g., **0.01**).
   - This is because higher contamination rates consider more data points as anomalies, thus **lowering the threshold** for what is flagged as an anomaly. The result is a more relaxed anomaly detection, with less extreme deviations needed to be classified as an anomaly.
2. **Impact of Estimator Count on Anomaly Scores**:
   - Within each contamination rate, changing the estimator count (e.g., from 50 to 200) has a **slight effect** on the stability of the anomaly scores.
   - Higher estimator counts appear to produce **more consistent and smooth scores** due to increased averaging from multiple trees. For instance, 200 estimators result in fewer fluctuations than 50 estimators, especially noticeable in sections where scores oscillate.

    - However, the **total count of anomalies remains unchanged** within a given contamination rate, regardless of the number of estimators. This reinforces that the estimator count mainly impacts the **reliability** of the score rather than the **quantity** of detected anomalies.

3. **Distinct Anomaly Clusters**:
    - The plot reveals **clusters of anomalies** in certain parts of the data, identifiable by the peaks or sharp score variations across contamination and estimator settings. These clusters suggest **periods of unusual activity**or **significant outliers** in the dataset, likely stemming from unusual sensor readings or irregular events.
    - Different contamination rates highlight these clusters with varying degrees of emphasis, indicating that lower contamination rates (e.g., 0.01) prioritize the most extreme anomalies within these clusters.
4. **Score Ranges and Consistency**:
    - Lower contamination rates (0.01) produce a **broader range of scores**, with scores reaching up to around **0.3**, while higher contamination rates (0.1) compress scores to around **0.1–0.2**.
    - This difference implies that a lower contamination rate maintains a **stricter anomaly threshold**, capturing only high-impact anomalies and ignoring minor fluctuations. In contrast, higher contamination rates are more **inclusive**, flagging both significant and moderate deviations.

Summary Report of Observations

1. **Anomaly Detection Sensitivity**:
    - Lower contamination rates (0.01) are more sensitive to extreme anomalies, yielding higher and more variable scores. This setting is suitable for applications needing a precise focus on the most significant outliers.
    - Higher contamination rates (0.1) are more inclusive, allowing for the detection of both major and minor anomalies but with reduced score sensitivity.
2. **Effect of Estimator Count**:
    - Increasing the number of estimators improves score stability but does not significantly alter the anomaly count. Using more estimators (100–200) is beneficial for producing smoother and more reliable anomaly scores.
3. **Identification of Anomaly Clusters**:
    - The dataset contains distinct clusters of high anomaly scores across all settings, suggesting recurring or concentrated outlier behavior in certain data segments. These clusters could represent specific periods where sensor behavior deviates from the norm, requiring further investigation.
4. **Parameter Tuning Recommendations**:
    - For **high-precision anomaly detection** (focusing on major outliers), a **contamination rate of 0.01** with at least **100 estimators** is recommended.