Anomaly Detection in Sensor Data Using Logistic Regression

Project Overview

- **Title:** Anomaly Detection in Sensor Data Using Logistic Regression
- **Objective:** This project aims to detect anomalies in sensor data using logistic regression to identify irregularities that deviate from normal patterns.
- **Dataset:**
    - Total Records: 10,000
    - Features:
        - Temperature: Continuous feature representing sensor readings.
        - Anomaly: Target variable indicating anomalies (1 = anomaly, 0 = normal).
        - Timestamp: Represents the precise date and time when each sensor reading was recorded. This feature is crucial for understanding the temporal context of the data, identifying trends, and analysing anomalies over time.
        - Location: categorical features representing sensor locations.

Problem statement

The primary goal of this project is to develop an anomaly detection system for industrial equipment. The system's objective is to identify unusual behaviour in equipment data and thereby prevent equipment failure, reduce downtime, and improve operational efficiency.

However, the rarity of these anomalies in comparison to normal behaviour creates a significant class imbalance, which complicates the detection process. This project aims to address this challenge by developing a machine learning model that can accurately classify sensor data into normal and anomalous categories, even in the presence of imbalanced datasets. The solution will leverage logistic regression and advanced techniques like oversampling to improve detection performance and reliability.

Methodology

Overview:

This project develops a logistic regression model for anomaly detection using only the Temperature feature. The methodology focuses on simplicity, interpretability, and effectiveness in detecting anomalies.

EDA (Exploratory Data Analysis)

1. Class Distribution:

- **Observation:**
  - The dataset has a significant class imbalance, with normal records vastly outnumbering anomalous records.
- **Action:**
  - Decided to use SMOTE for oversampling the minority class to ensure the model learns effectively from anomalies.

2. Feature Correlation Analysis:

- **Observation:**
  - Temperature showed a strong correlation with the Anomaly target variable, making it the most critical feature for classification.
  - Other features like Hour, Day, Month, and Weekday had weak or no significant correlation with the target.
- **Action:**
  - Retained only the Temperature feature for the model and excluded less relevant features to simplify the model.

3. Feature Distributions:

- **Observation:**
  - The Temperature feature exhibited a normal-like distribution with a few extreme values (potential anomalies).

- **Action:**
    - Analyzed these extreme values further but did not apply outlier handling since anomalies are expected in this dataset.

4. **Timestamp:**

- **Observation:**
    - Timestamp feature did not show significant trends or periodicity in relation to anomalies.
- **Action:**
    - Removed Timestamp feature from the dataset.

5. **Location Features:**

- **Observation:**
    - Location did not show meaningful separation between anomalies and normal data.
- **Action:**
    - Removed the feature to streamline the model.

Summary of EDA Insights:

- The Temperature feature was identified as the most critical predictor for anomalies.
- Temporal and location-based features were removed due to weak relevance to the target variable.
- SMOTE was identified as a necessary technique to address the severe class imbalance in the dataset.

Model Building:

1. **Feature Selection:**

- Selected Temperature as the sole input feature for the logistic regression model.
- Excluded other features (Timestamp, Location, Hour, etc.) to maintain simplicity and focus on the most relevant sensor reading.

2. **Train-Test Split:**
   - Divided the dataset into training and testing sets with an 80-20 split.
   - Ensured the split was stratified to maintain the same class distribution in both sets.

3. **Handling Class Imbalance:**
   - Applied **Synthetic Minority Oversampling Technique (SMOTE)** and **Undersampling** (both for finer control) to address the imbalance in the training set.
   - first oversample the minority class using SMOTE to ensure it has a comparable number of samples to the majority class, and then perform under sampling on the majority class to create a balanced training dataset.
   - This ensured that the model could effectively learn to identify anomalies despite their rarity.

4. **Model Training:**
   - Trained a logistic regression model using the resampled training data.
   - Used class_weight ='balanced' to account for any residual class imbalance.

5. **Model Evaluation:**
   - Evaluated the model on the test set using metrics such as precision, recall, F1-score, and confusion matrix.
   - Analyzed the model's ability to detect anomalies effectively.

Results

Evaluation Metrics:

- **Precision:** 98%
- **Recall:** 100%
- **F1-Score:** 99%
- **Accuracy:** 100%

Confusion Matrix:

	Predicted Normal	Anomaly
Normal	1953	1
Anomaly	0	46

Conclusion

- **Summary:** The logistic regression model effectively detected anomalies in sensor data, achieving an F1-score of X%. The application of SMOTE helped address class imbalance, improving recall for anomalies.
- **Key Takeaways:** Logistic regression is a reliable and interpretable model for anomaly detection tasks when combined with oversampling techniques.