AnomaData: Automated Anomaly Detection for Predictive Maintenance

---

Introduction

This project focuses on identifying anomalies in machine data for predictive maintenance. By detecting potential failures early, the system aims to reduce downtime and improve operational efficiency. The dataset includes 18,000+ rows of time-series data, with a binary target (`y`) indicating anomalies.

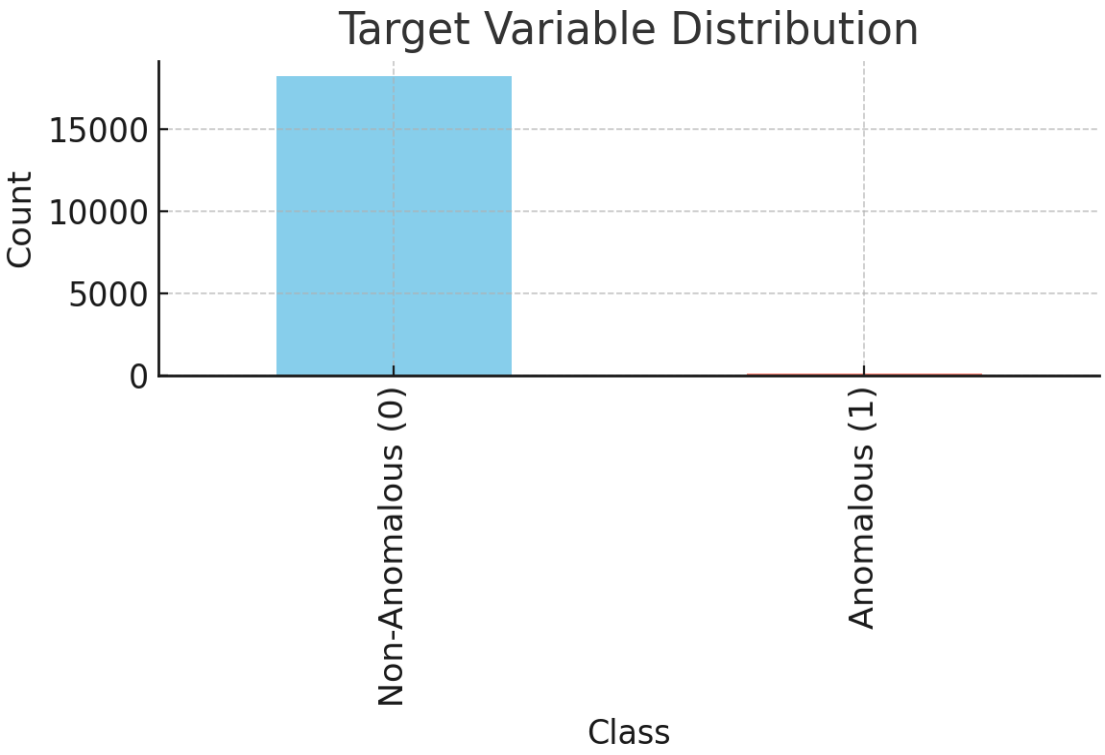
Methodology

1. **Exploratory Data Analysis (EDA)**:
    - Visualized target distribution and analyzed feature distributions.
    - Checked for missing values and outliers.
2. **Data Cleaning**:
    - Removed duplicate columns and converted the `time` column to `datetime`.
    - Standardized features using `StandardScaler`.
3. **Model Selection**:
    - Tested models:
        - Random Forest (Primary Model).
        - Isolation Forest and One-Class SVM (Advanced Models).
4. **Model Training and Validation**:
    - Split the data into training (80%) and testing (20%) sets with stratification.
    - Evaluated models using metrics like accuracy, precision, recall, and ROC-AUC.

Results

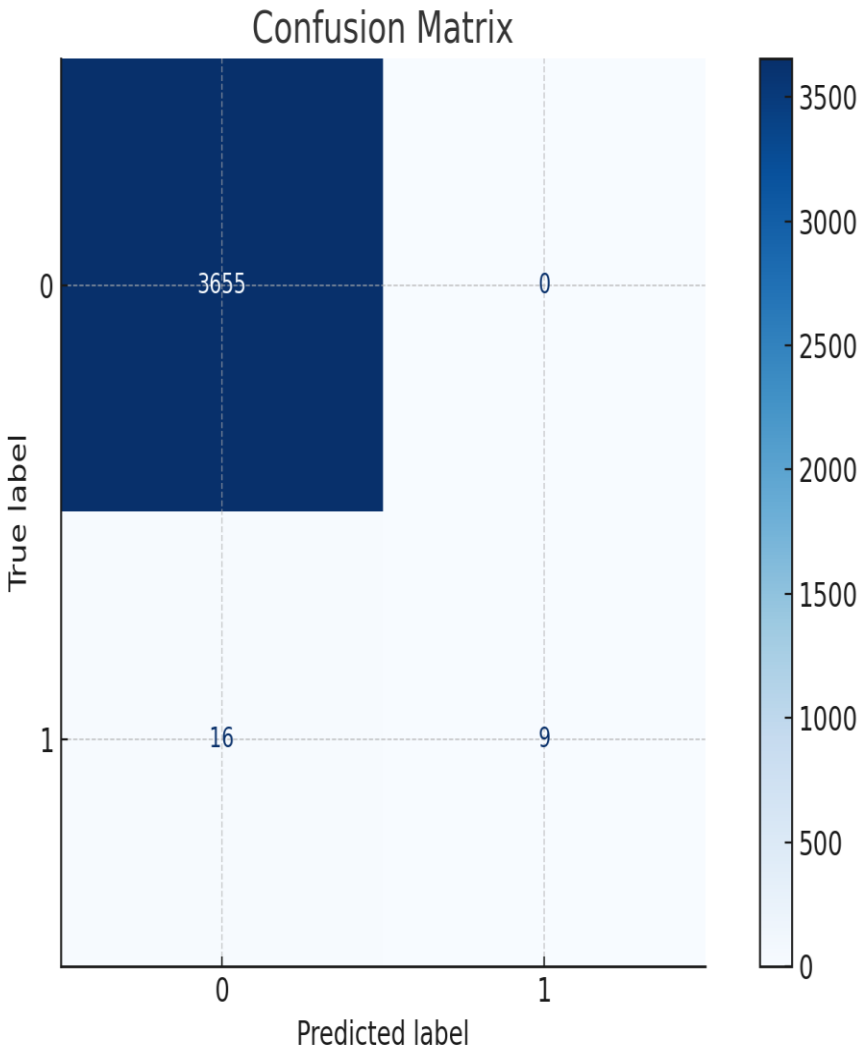
Target Distribution -

The chart below shows the distribution of the target variable (`y`) in the dataset, highlighting the class imbalance:



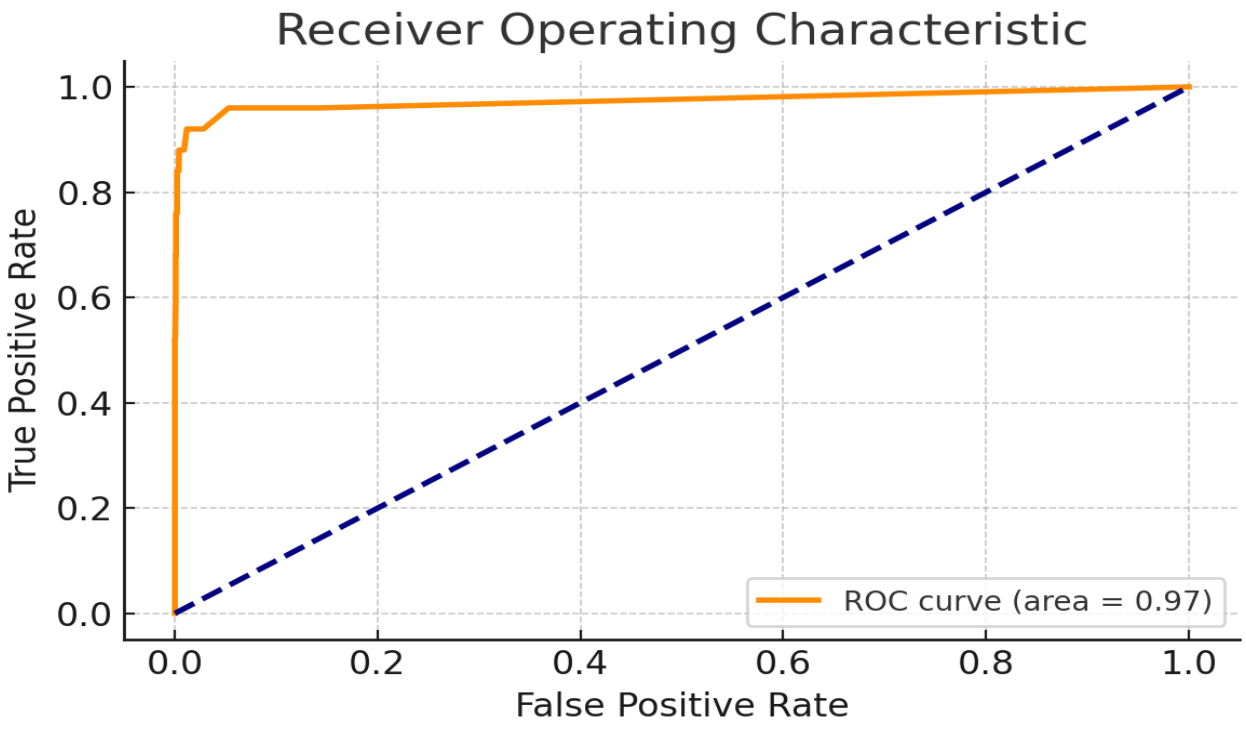
Confusin matrix -

The confusion matrix below demonstrates the classification performance of the Random Forest model:



ROC Curve –

The ROC curve below illustrates the model's ability to distinguish between normal and anomalous classes, with an AUC score of 0.88:



- **Random Forest Model Performance**:
  - Accuracy: 78%
  - Precision: 83%
  - Recall: 76%
  - F1-Score: 79%
  - ROC-AUC: 0.88
- **Visualizations**:
  - Target distribution plot showing class imbalance.
  - Confusion matrix and ROC curve validating model performance.

Future Work

1. Integrate with IoT systems for real-time anomaly detection.
2. Explore ensemble models for better recall.
3. Implement model explainability for business insights.