AnomaData: Automated Anomaly Detection for Predictive Maintenance

1. Introduction

The goal of this project was to develop a predictive maintenance system for identifying machine breakdowns by detecting anomalies in sensor data. The system aims to predict machine failures by identifying deviations from normal behavior, allowing for timely intervention before equipment degradation occurs. The dataset consists of over 18,000 rows of time-series data, with a binary label 'y' indicating anomalies (1 for anomaly, 0 for normal).

2. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) was performed to gain insights into the data. Visualizations such as histograms, scatter plots, and box plots were used to understand the distribution of features and their correlation with the target variable (anomaly). Missing values and outliers were identified and handled appropriately. Descriptive statistics were computed to understand the central tendencies and variability of the data.

3. Data Preprocessing

The preprocessing steps included handling missing values through imputation, outlier detection and removal using z-scores, and scaling/normalization of features to ensure consistency. Features with datetime values were converted to appropriate formats for time-series analysis. After preprocessing, the data was split into training and test sets.

4. Model Selection and Training

Several machine learning models were considered for anomaly detection, including Logistic Regression, Random Forest, XGBoost, and Isolation Forest. Based on the nature of the problem, the Random Forest model was selected for its ability to handle both classification and anomaly detection tasks. The model was trained on the training set and evaluated using metrics such as accuracy, precision, recall, and F1-score. Hyperparameter tuning was conducted to improve the model's performance.

5. Model Evaluation and Validation

The model's performance was evaluated using the test set. Key evaluation metrics included accuracy, precision, recall, and F1-score. A confusion matrix was used to assess the false positives and false negatives. Cross-validation was performed to ensure the model's generalizability. The final model achieved a performance of over 75% accuracy, indicating its ability to reliably detect anomalies.

6. Future Work

Future work can involve further improvement of the model by incorporating additional features such as time-related variables (e.g., rolling averages) and exploring more advanced models like Autoencoders or Neural Networks. The model can also be deployed in a production environment where it can monitor live machine data and trigger alerts for potential failures.

7. Conclusion

This project successfully demonstrated the potential of machine learning models for predictive maintenance and anomaly detection. The trained model can predict machine breakdowns and help reduce downtime and maintenance costs. The solution can be extended to other machines or equipment in different industrial settings.