ANOMA-DATA

AUTOMATED ANOMALY DETECTION FOR PREDICTIVE MAINTENANCE

Project Documentation

Project Name: **AnomaData (Predictive Maintenance using Anomaly Detection in Machine Learning)**

Submitted by: **Komal Sahu**

INDEX :

1. Problem Statement :

Industries today face a significant challenge in maintaining their machinery efficiently. Unplanned downtime due to equipment failures leads to increased costs, reduced productivity, and operational inefficiencies. Traditional maintenance approaches such as reactive maintenance (fixing a machine after failure) and scheduled preventive maintenance are often inadequate, as they either lead to unexpected breakdowns or waste resources on unnecessary servicing.

This project aims to develop an Automated Anomaly Detection System using machine learning to predict potential failures in industrial machines. The system will analyse machine sensor data and detect anomalies, enabling predictive maintenance. This will allow industries to take proactive measures, reducing downtime, optimizing maintenance schedules, and enhancing the lifespan of machinery.

2. Data Sources :

The dataset used for this project consists of machine sensor readings collected over multiple days, stored in an Excel file named **AnomaData.xlsx**. The dataset contains **18,000+ Rows and 60+ Columns,** with each record representing a machine status at a particular time.

This dataset exhibits a significant imbalance, as the positive class (anomalous machine behaviour) accounts for only a small percentage of all records. This high level of imbalance presents challenges for detecting anomalies, as traditional machine learning models may struggle to accurately identify the minority class without appropriate handling of the imbalance. Addressing this imbalance is crucial to improving model accuracy and ensuring that critical failures are detected effectively.

Dataset Breakdown:

- **Target Variable : y (Binary Classification - 1 represents an anomaly, 0 represents normal operation)**
- **Features :** Multiple sensor readings capturing machine behaviour, environmental conditions, and system parameters.
- **Time Period :** Data collected over several days, allowing for time-series analysis.

This dataset enables a detailed study of machine behaviour over time , analyse historical patterns, detect hidden anomalies, and build machine learning models also capable of predictive maintenance, helping to identify conditions that lead to failures.

3. Data Preprocessing Steps :

1. **Loading and Inspecting the Dataset:**
    - The dataset was loaded into a Pandas Data Frame, and initial inspection was conducted to check for missing values, duplicate records, and data inconsistencies.
2. **Handling Missing Values:**
    - Any missing values found were treated using appropriate imputation techniques such as mean, median, or mode replacement, ensuring data completeness.
3. **Outlier Detection and Removal:**
    - Boxplots and the interquartile range (IQR) method were used to detect extreme outliers that could negatively impact the model's performance.
4. **Feature Scaling:**
    - Numerical features were normalized using StandardScaler to standardize their range, ensuring consistent model learning.
5. **Balancing the Dataset:**
    - The dataset exhibited class imbalance, with normal instances significantly outweighing anomalies. A hybrid approach (undersampling the majority class + oversampling the minority class) i.e. Resample approach were applied to balance the dataset.

4. Exploratory Data Analysis (EDA):

EDA helps in understanding data distributions, correlations, and key patterns:

- **Heatmaps** were used to analyse correlations between different sensor readings.
- **Boxplots** and **histograms** visualized the distribution of numerical features.
- **Time-series** Analysis examined how anomalies varied over time.
- Feature Importance Analysis identified the most significant variables influencing anomaly detection.

5. Feature Engineering :

Feature engineering plays a significant role in improving model performance. Several strategies were applied:

- **Time-Based Features:** Extracted meaningful time-related features such as time of day, day of the week, and seasonal patterns from the timestamp column.

- **Feature Transformation:** Applied logarithmic transformations to reduce skewness in numerical data.

- **Feature Selection:** Used correlation analysis and tree-based feature selection to retain only the most relevant features.

These feature engineering techniques help the model capture meaningful patterns in the data while reducing computational complexity and improve predictive performance.

6. Model Selection :

Machine Learning Models Evaluated:

The following machine learning algorithms were tested for anomaly detection:

- Naive Bayes
- KNN
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier (SVC)
- Logistic Regression

Hyperparameter Tuning and Optimization:

- Grid search with cross-validation was performed to optimize model parameters.
- Cross-validation helped assess model generalization across different subsets of data.

Best Model Selection: **Random Forest** emerged as the best model, achieving **99.9%** accuracy, making it the most suitable choice for detecting anomalies.

7. Performance Evaluation :

- **Accuracy**: Overall correctness of predictions.
- **Precision**: Percentage of correctly identified anomalies.
- **Recall (Sensitivity)**: Ability to detect actual anomalies.
- **F1-Score**: Balance between precision and recall.
- **ROC-AUC Score**: Measures the ability to distinguish between normal and anomalous instances.
- **Confusion Matrix & ROC Curve**: Visual representations of classification performance.

Metrics Used for Evaluation:

1. **Confusion Matrix:** Helps visualize model predictions by categorizing instances as True Positives, False Positives, True Negatives, and False Negatives.
2. **ROC-AUC Score:** Measures how well the model distinguishes between normal and anomalous machine states.
3. **Precision, Recall, and F1-Score:** Provides insights into the balance between false positives and false negatives.

Key Observations:

- Certain sensor readings were found to be highly correlated with machine failures.
- Addressing data imbalance improved model recall, reducing the chances of missing actual anomalies.
- Removing outliers resulted in increased model stability and robustness.

8. Benefits of the Model for Technical Users :

1. **Automated Detection & Early Warnings**
   - Helps engineers and IT teams identify potential system failures before they happen.
   - Reduces manual effort in monitoring time-series sensor data.
2. **Improved Data-Driven Decision Making**
   - Enables data scientists to analyse machine conditions using real-time data.
   - Supports decision-making through statistical analysis & ML insights.
3. **Scalability & Adaptability**
   - Can handle large-scale IoT sensor data and adapt to various machine types.
   - Can be integrated into cloud platforms for real-time monitoring.
4. **Advanced Feature Engineering & Model Optimization**
   - Uses state-of-the-art ML/DL techniques (e.g., Isolation Forest, Autoencoders).
   - Can be fine-tuned with hyperparameter tuning & anomaly scoring methods.
5. **Better Model Interpretability & Explainability**
   - Helps technical users understand why an anomaly is detected.
   - Can incorporate SHAP/LIME methods for transparency.

---

9. Benefits of the Model for Potential Users :

1. **Reduced Downtime & Maintenance Costs**
   - Prevents unexpected machine failures, reducing unplanned downtimes.
   - Saves costs by shifting from reactive to predictive maintenance.
2. **Increased Equipment Lifespan**
   - Regular monitoring helps extend the life of machinery.
   - Avoids unnecessary repairs by detecting early warning signs.
3. **Operational Efficiency & Productivity**
   - Enables real-time monitoring of industrial equipment.

  - Helps maintenance teams focus only on machines likely to fail, saving time.

4. **Compliance & Risk Mitigation**
   - Helps industries comply with safety regulations by preventing failures.
   - Reduces risks in critical infrastructure (manufacturing, healthcare, finance, etc.).

---

10. Future Work :

The future development of the fraud detection model could focus on enhancing accuracy, interpretability, and flexibility. Below are several areas for potential improvement and further work:

Future Enhancements:

- **Real-time Deployment:** Develop an API-based implementation for seamless real-time anomaly detection.
- **Integration with IoT Sensors:** Enable continuous data collection and monitoring.
- **Adaptive Learning:** Implement self-improving models that evolve with new data.
- **User Dashboard Development:** Build an interactive web interface for anomaly visualization.
- **Improved Data Privacy and Security**

---

11. Benefits of Anomaly Detection Model :

- **Predictive Maintenance:** Prevents failures, reduces downtime, and lowers maintenance costs.

- **Operational Efficiency:** Optimizes resource allocation, increases productivity, and enhances supply chain planning.

- **Cost Savings:** Reduces warranty claims, minimizes energy consumption, and improves profitability.

- **Safety & Compliance**: Prevents workplace accidents and ensures regulatory compliance.

- **Data-Driven Decisions:** Enables real-time monitoring, AI-driven insights, and continuous improvement.

---

12. Conclusion :

- The AnomaData project successfully demonstrated the application of machine learning techniques for predictive maintenance through anomaly detection. Among the evaluated models, the **Random Forest initial model** emerged as the best performer, achieving an outstanding accuracy of **99.9%**.
- This high accuracy, combined with superior precision, recall, F1-score, and AUC-ROC metrics, makes it the most reliable choice for detecting anomalies in the dataset.
- The model effectively differentiates between normal and anomalous conditions, ensuring timely identification of potential failures. By leveraging this high-performing model, industries can proactively optimize maintenance strategies, reduce operational downtime, and enhance overall efficiency.