Anomaly Detection in Multivariate Time Series Data for Printer Diagnostics

1. Project Overview

Collaborators:

- **Institution:** IIT Roorkee
- **Corporate Partner:** HP

Project Goals:

The primary goal of the project was to develop a Machine Learning (ML) model capable of detecting or predicting mechanical failures in printers using the Print Engine Failure Sensor (PEFS) data. A significant aspect of the project involved ensuring that the ML model incorporated explainability features, enabling stakeholders to understand the decision-making process behind the predictions.

Data Provision:

The PEFS data, which formed the backbone of this project, was provided by the HP team. This dataset included acoustic signals and relevant metadata aimed at facilitating anomaly detection and predictive maintenance.

2. Objectives and Deliverables

Objectives:

1. Develop machine learning models to detect and predict mechanical failures in printer components using PEFS data.
2. Explore and implement multiple algorithms for anomaly detection.
3. Integrate explainability mechanisms to ensure transparency in the decision-making process of the ML models.
4. Create visualizations and reports to present insights derived from the models.
5. Validate and tune models to achieve optimal performance for the given dataset.
6. Maintain a centralized GitHub repository for storing all project-related files, including documentation, visualizations, and notebooks.

Planned Deliverables:

1. Comprehensive documentation of algorithms, hyperparameter tuning, and model results.
2. Interactive visualizations and diagnostics tools for anomaly detection.
3. Final project report detailing findings, methodologies, and future recommendations.
4. Well-maintained GitHub repository containing all project artifacts.

Metrics for Success:

1. Model performance evaluated through metrics such as AUC, PR curve analysis, and confusion matrix insights.
2. Demonstration of explainability features within the models to aid in stakeholder understanding.
3. Clear, actionable insights presented through visualizations and reports.
4. Organized and accessible GitHub repository with all relevant project files.

Responsibilities:

- **HP Team:**

    - Provide PEFS data and detailed background material.
    - Facilitate data transfer and infrastructure setup.
- **IIT Roorkee Team:**
    - Develop, validate, and deliver ML models.
    - Create all necessary documentation and visualizations.
    - Procure any additional equipment required for the project.
    - Maintain and update the GitHub repository with all project-related files

3. Project Phases:

- Background and Data Exploration:
    - Studied research papers provided by HP to gain insights into anomaly detection techniques.
        - **Digital Signal Processing for Laser Printer Noise Source Detection and Identification (2019)**
          This paper explores the application of digital signal processing techniques to detect and identify noise sources in laser printers, focusing on HP LaserJet M603. Key methods include discrete Fourier transform (DFT) for power spectrum density analysis, Butterworth filters, and Hilbert transforms to isolate and analyze squeaking signals. The study highlights the integration of acoustical and mechanical characteristics for accurate source identification, proposing a next step toward machine learning-based detection.
        - **Using Acoustic Information to Diagnose the Health of a Printer (2020)**
          This research presents a sound-based anomaly detection system for printer diagnostics, emphasizing the use of acoustic data augmentation to improve model performance. Techniques include

principal component analysis (PCA) for feature extraction and the use of One-Class SVM and Random Forest models. The study demonstrates the effectiveness of augmented datasets in enhancing detection accuracy, offering a pipeline for identifying printer health anomalies based on sound.

- **Acoustic Print Engine Failure Sensor (2018)**
  This paper discusses the grand vision of predictive maintenance using acoustic sensors in HP printers. The study focuses on analyzing sound signatures to detect trends and identify both known and unknown failures across printer fleets. The integration of machine learning algorithms for real-time failure predictions showcases the potential for proactive maintenance and cost reduction, highlighting the utility of frequency and modulation analysis.
- **A Noise Source Detector for Squeaking or Rattling Issues of Printers (2018)**
  This publication describes the development of a noise source detector aimed at addressing squeaking or rattling issues in printers. The solution employs advanced signal processing and noise characterization techniques to pinpoint fault sources, providing a foundation for predictive diagnostics in printer systems. The study emphasizes the potential to integrate such systems into operational workflows for enhanced reliability.

- Explored the sample dataset provided by HP to understand its structure and characteristics.
- Set up a base framework to prepare for the larger dataset.

- Data Analysis:
    - Addressed the complication of multiple rows with identical timestamps, which introduced redundancy and ambiguity in analysis. This issue was resolved through coordination with HP professionals, where a methodology was finalized to group rows with the same timestamps effectively.
    - Finalized a data labeling approach with HP professionals due to the absence of predefined labels.
    - The dataset provided by HP for this project included multiple features that were critical for analyzing and identifying anomalies in printer diagnostics. Below is an explanation of the key features included in the dataset:

        Features in the Dataset:

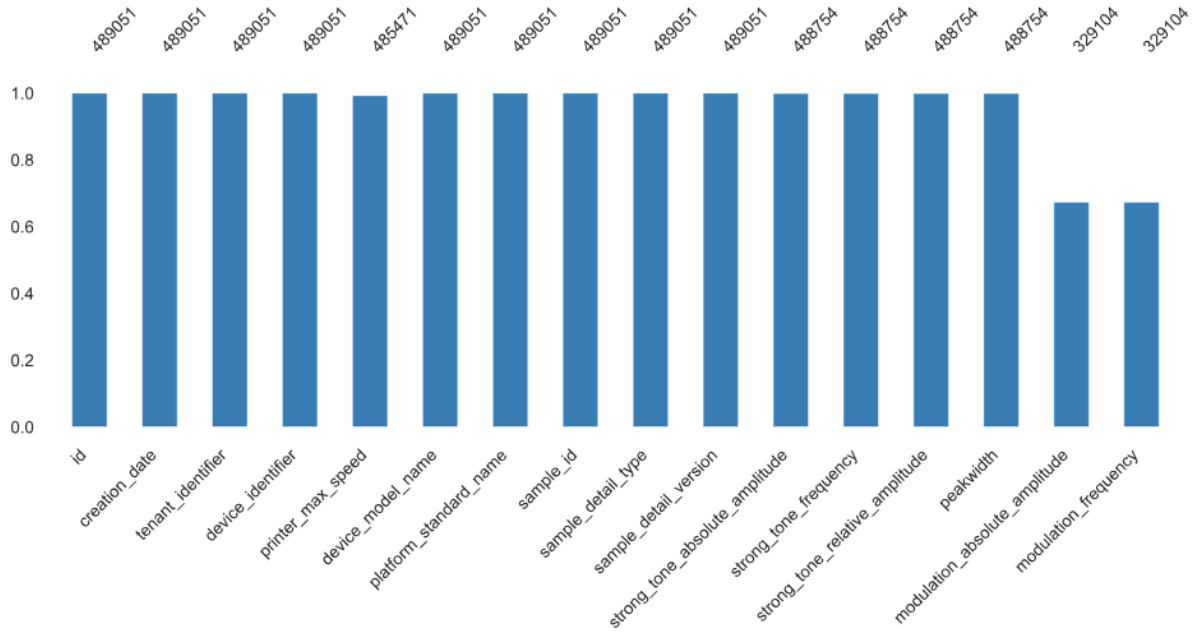
        1. **creation_date:**
            - The timestamp indicating when the data sample was created. This feature is essential for tracking temporal patterns and analyzing time-based anomalies.
        2. **tenant_identifier:**
            - A unique identifier for the tenant or organizational unit associated with the data. Useful for multi-tenant systems and segregating data for analysis.
        3. **device_identifier:**
            - A unique identifier for the printer or device generating the data. This allows for device-specific diagnostics and tracking.
        4. **printer_max_speed:**
            - The maximum operational speed of the printer, measured in pages per minute. Variations in this feature may correlate with performance issues or mechanical wear.
        5. **device_model_name:**

    - The model name of the printer device. This helps contextualize the data and align it with device-specific parameters or configurations.
6. **platform_standard_name:**
    - The platform or standard associated with the printer device. It provides additional context about the operational environment.
7. **sample_id:**
    - A unique identifier for each data sample. This ensures traceability and aids in dataset management.
8. **sample_detail_type:**
    - The type of detail or data being captured in the sample. Helps classify and categorize different data subsets.
9. **sample_detail_version:**
    - The version of the sample detail, indicating updates or revisions to the data collection protocol.
10. **strong_tone_absolute_amplitude:**
    - The absolute amplitude of the strongest tone in the signal. High amplitudes may indicate significant mechanical activity or potential faults.
11. **strong_tone_frequency:**
    - The frequency of the strongest tone in the signal, which often corresponds to the dominant mechanical process within the printer.
12. **strong_tone_relative_amplitude:**
    - The relative amplitude of the strongest tone compared to the overall signal. Deviations may signal abnormal behavior.
13. **peakwidth:**

        - The width of the peak in the frequency domain, indicating the spread of dominant frequencies. Narrow peaks often signify stable operations, while wide peaks may indicate anomalies.
    14. **modulation_absolute_amplitude:**
        - The absolute amplitude of the modulation in the signal, representing the intensity of variations in mechanical activity.
    15. **modulation_frequency:**
        - The frequency at which the signal's amplitude varies, often tied to cyclic or oscillatory components in the printer.
- Conducted Exploratory Data Analysis (EDA) on the complete dataset received.
- Collaborated with the HP team to discuss and implement data cleaning mechanisms to handle issues like large amounts of NaN values.
- Statistics collected from a subset of data:
    - **NULL values per column:**

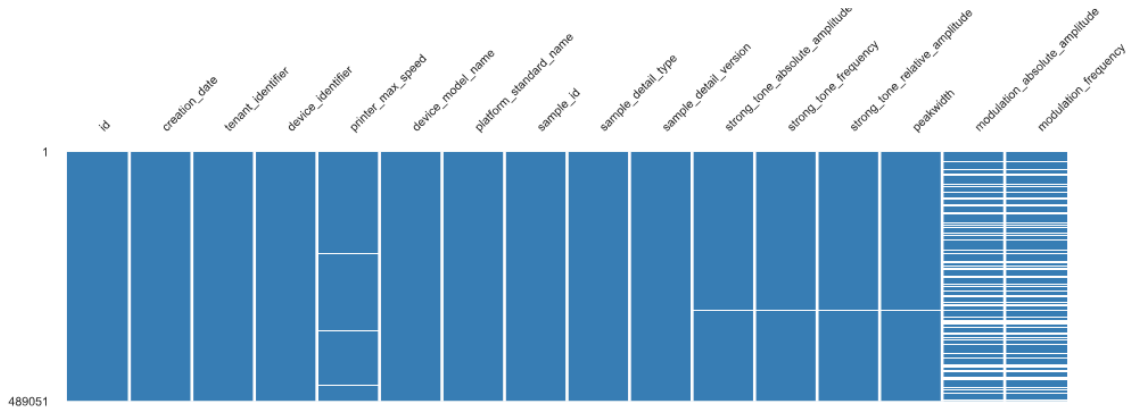
      The bar chart visualizes the count of null values for each column in a subset of the dataset containing 489,051 rows.

      modulation_absolute_amplitude and **modulation_frequency**: These columns have significantly more null values, with around 329,104 non-null rows, implying approximately 159,947 rows with missing data.



- **Nullity Matrix:**

  Nullity matrix is a data-dense display which lets us quickly visually pick out patterns in data completion.



Key Data Issues Identified

While analyzing the dataset, several critical data issues were identified. These issues impacted the quality and usability of the dataset and required extensive preprocessing efforts to mitigate. Below are the three most significant data issues:

1. Presence of Multiple Rows with Identical Timestamps:

- **Description:**
  - The dataset contained numerous instances where multiple rows shared the same timestamp. This redundancy created ambiguity in understanding whether these rows represented duplicate entries or separate events occurring simultaneously.
- **Impact:**
  - This issue complicated the interpretation of the data and introduced noise, making it challenging to analyze temporal trends and patterns accurately.
- **Resolution Approach:**

    - Collaborated with HP professionals to finalize a grouping methodology, where rows with identical timestamps were aggregated based on their mean or sum for numerical features. While this approach provided a temporary solution, it lacked robustness and could lead to information loss.

2. **High Volume of Missing Values in Key Columns:**

- **Description:**
    - Certain columns, such as `modulation_absolute_amplitude` and `modulation_frequency`, had a significant proportion of missing values.
    - `strong_tone_absolute_amplitude`, `strong_tone_frequency`, `strong_tone_relative_amplitude`, and `peakwidth`: These columns have fewer null entries.
- **Impact:**
    - These columns were critical for identifying anomalies related to modulation patterns. The high proportion of missing data reduced the effectiveness of models that relied on these features.
- **Resolution Approach:**
    - Multiple strategies were considered, including imputation using statistical measures (e.g., mean or median) and the use of advanced techniques such as k-Nearest Neighbors (k-NN) imputation. However, the effectiveness of imputation remained limited, leading to a cautious approach in using these features for model training.

3. Lack of Labels for Supervised Learning:

- **Description:**
    - The dataset lacked predefined labels indicating whether a sample was anomalous or normal. This absence of ground truth significantly hindered the ability to directly evaluate model performance.

- **Impact:**
    - The lack of labels necessitated the use of unsupervised learning techniques, which are generally less interpretable and harder to validate.
    - Hypotheses, such as the contamination window approach (assuming anomalies occurred 1 to 7 days before a reported fault), introduced assumptions that might not align perfectly with real-world scenarios.
- **Resolution Approach:**
    - Developed a hypothesis-driven labeling strategy in collaboration with HP professionals. This included using fault detection dates to create a contamination window for anomaly labeling.
    - Despite this approach, the lack of direct labels limited the ability to evaluate true positives and false positives rigorously.

Model Building:

Explored and implemented various models for unsupervised anomaly detection on multivariate time-series data. Below is a summary of the key models used:

1. **Isolation Forest (ISF):**
    - A tree-based model designed for detecting anomalies in high-dimensional data by isolating data points that appear different from the majority.
    - Strengths: Effective for handling high-dimensional datasets and detecting outliers.
    - Challenges: Requires careful hyperparameter tuning to avoid overfitting or underfitting.
2. **Autoencoders:**
    - Neural network-based models trained to reconstruct input data. Anomalies are identified by high reconstruction errors.

    - Strengths: Suitable for capturing non-linear relationships in the data.
    - Challenges: Sensitive to the choice of network architecture and training parameters.
3. **k-Nearest Neighbors (KNN):**
    - A distance-based model that detects anomalies by measuring the distance of a data point from its nearest neighbors.
    - Strengths: Simple to implement and interpret.
    - Challenges: Computationally expensive for large datasets.
4. **Local Outlier Factor (LOF):**
    - A density-based model that identifies anomalies by comparing the local density of a point to its neighbors.
    - Strengths: Effective for detecting clusters of anomalies.
    - Challenges: Sensitive to the choice of neighborhood size (k).
5. **Gaussian Mixture Model (GMM):**
    - A probabilistic model that assumes data is generated from a mixture of Gaussian distributions.
    - Strengths: Provides probabilistic scores for anomaly detection.
    - Challenges: Assumes the data follows Gaussian distributions, which may not always be true.
6. **Kernel Density Estimation (KDE):**
    - A non-parametric model used to estimate the probability density function of the data.
    - Strengths: Effective for detecting anomalies in low-dimensional data.
    - Challenges: Performance deteriorates in high-dimensional spaces.
7. **One-Class Support Vector Machine (OCSVM):**
    - A kernel-based model that learns a decision boundary to separate normal data from potential anomalies.
    - Strengths: Effective for datasets with well-defined normal behavior.
    - Challenges: Sensitive to kernel choice and hyperparameter settings.

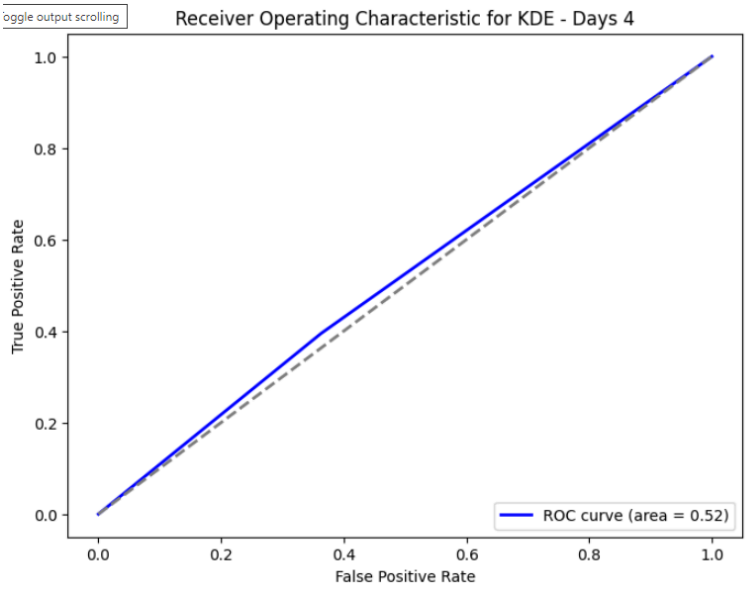
Inferencing:

- Applied the cleaned and prepared data to all explored models.
- Each model's performance was evaluated by varying the contamination window (1 to 7 days before fault detection) to determine the optimal range for anomaly detection. This approach was critical given the lack of explicit labels and was finalized in collaboration with HP professionals.
- Documented, visualized, and compared the results across different algorithms.
- Maintained and continuously updated the GitHub repository to include all project artifacts, including notebooks, visualizations, and documentation.

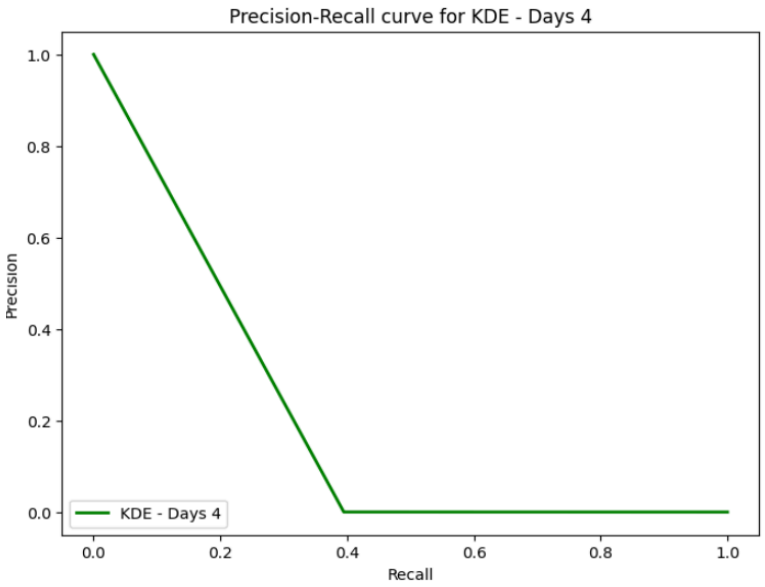
4. Results and Visualizations

The following models were implemented and evaluated for anomaly detection on multivariate data:

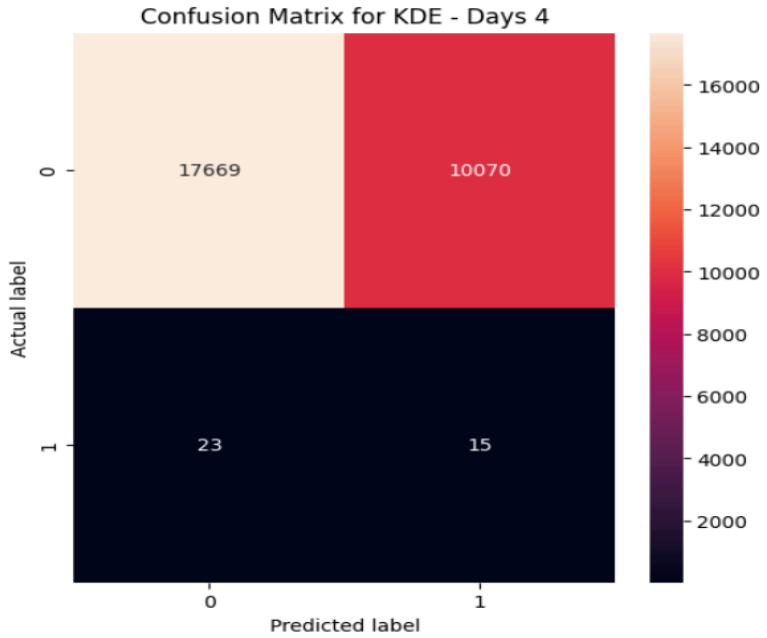
1. **Kernel Density Estimation (KDE):**
   - **AUC Score:** 0.52 (close to random guessing).
   - **Key Observations:**
     - The ROC curve followed a diagonal path, indicating low discriminative power.
     - The PR curve showed high precision at low recall but dropped sharply, suggesting a struggle to balance sensitivity and specificity.
   - **Confusion Matrix Analysis:**
     - High false positive rate.
     - Low true positive rate, consistent with low recall.
   - ROC Curve:


oggle output scrolling
Receiver Operating Characteristic for KDE - Days 4
True Positive Rate
False Positive Rate
ROC curve (area = 0.52)
0.0
0.2
0.4
0.6
0.8
1.0

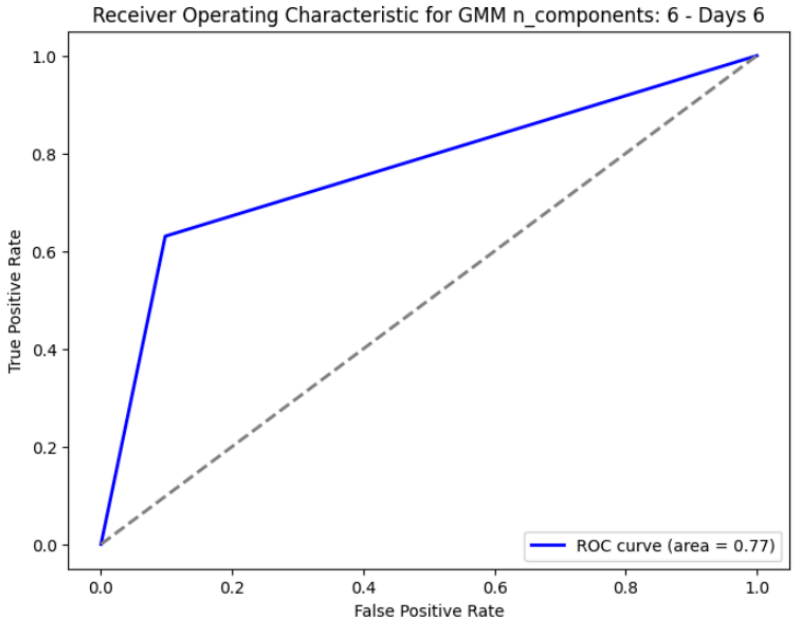
- PR Curve:



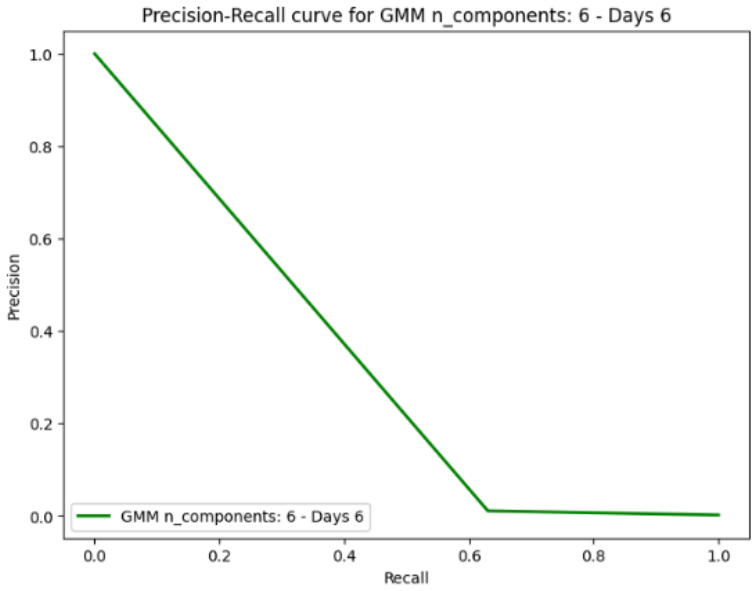
- Confusion Matrix



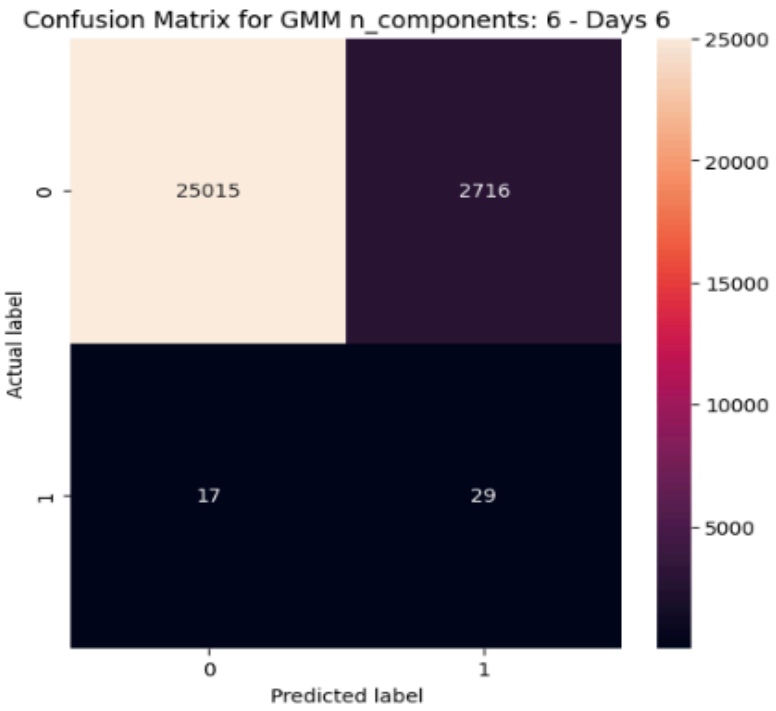
2. **Gaussian Mixture Model (GMM):**
    - **AUC Score:** 0.77 (indicating reasonably good performance).
    - **Key Observations:**
        - ROC curve showed sharp initial rise, achieving high TPR with low FPR initially.
        - PR curve exhibited high precision at low recall but declined sharply with increasing recall.
    - **Confusion Matrix Analysis:**
        - High true negative count but struggled with true positives due to low recall.
    - ROC Curve:


Receiver Operating Characteristic for GMM n_components: 6 - Days 6
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.77)

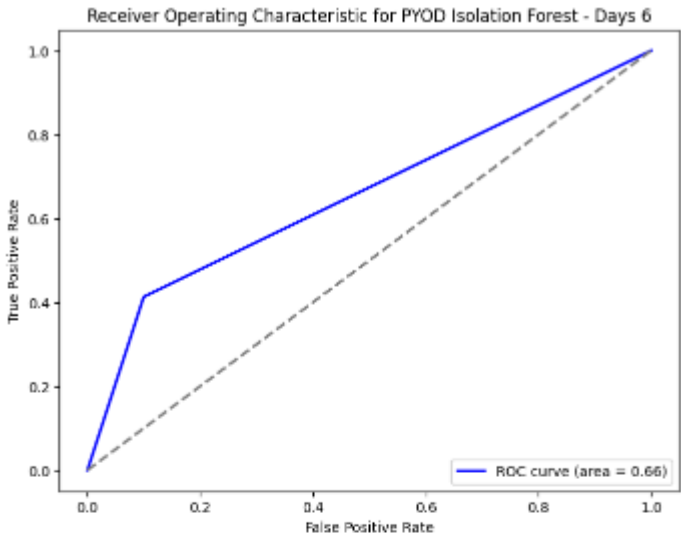
- PR Curve:



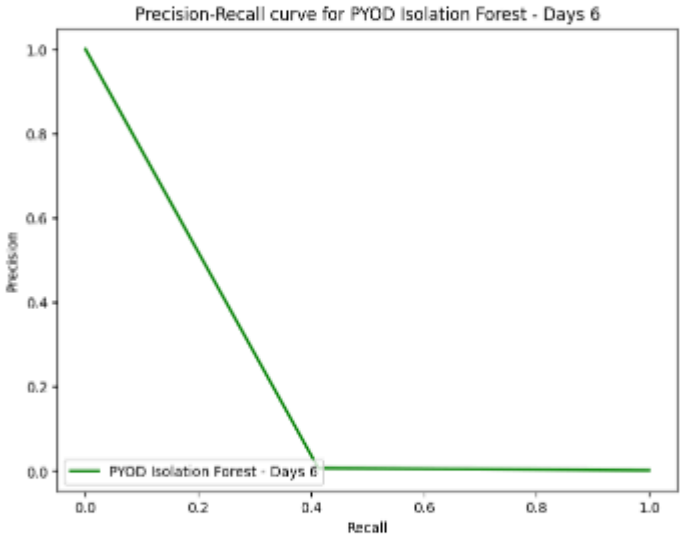
- Confusion Matrix



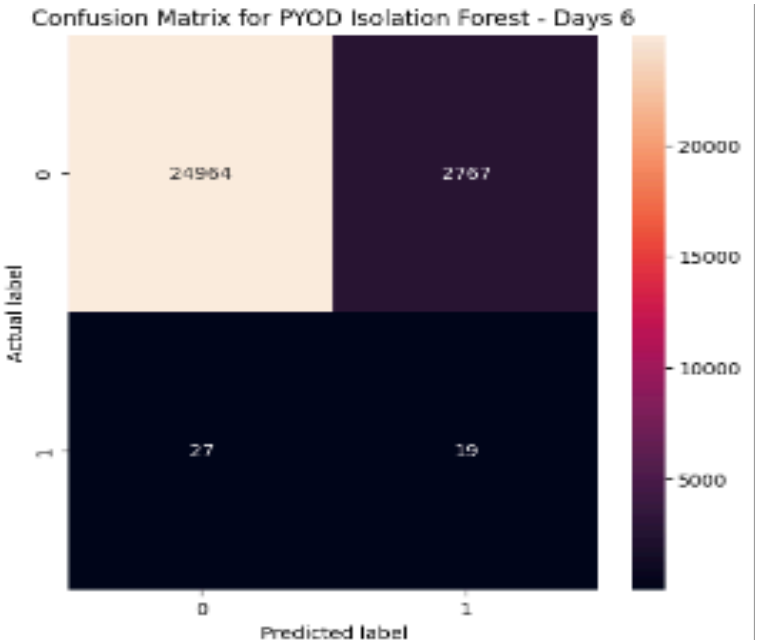
3. **Isolation Forest (ISF):**
    - **AUC Score:** 0.66 (moderate performance).
    - **Key Observations:**
        - ROC curve showed steep rise initially but neared the diagonal with increasing thresholds.
        - PR curve highlighted high precision at low recall but a significant drop with increased recall.
    - **Confusion Matrix Analysis:**
        - High true negative rate but a limited ability to identify true positives.
    - ROC Curve:


Receiver Operating Characteristic for PYOD Isolation Forest - Days 6
True Positive Rate
False Positive Rate
ROC curve (area = 0.66)
0.0
0.2
0.4
0.6
0.8
1.0

- PR Curve:

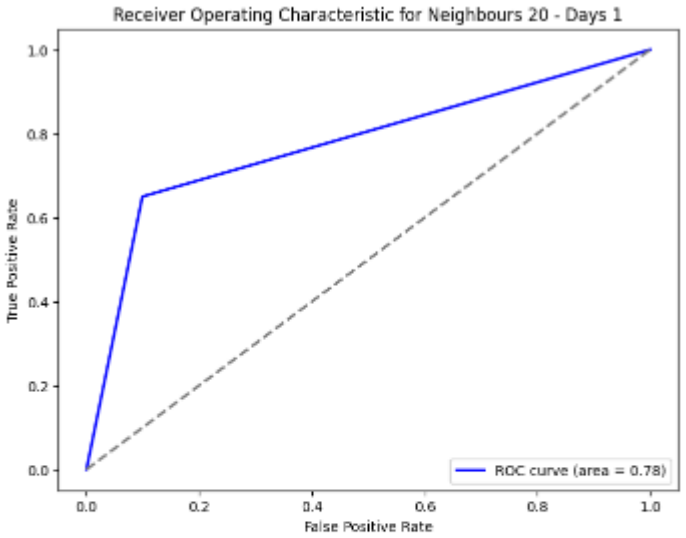

Precision-Recall curve for PYOD Isolation Forest - Days 6
Precision
Recall
PYOD Isolation Forest - Days 6
0.0
0.2
0.4
0.6
0.8
1.0

- Confusion Matrix

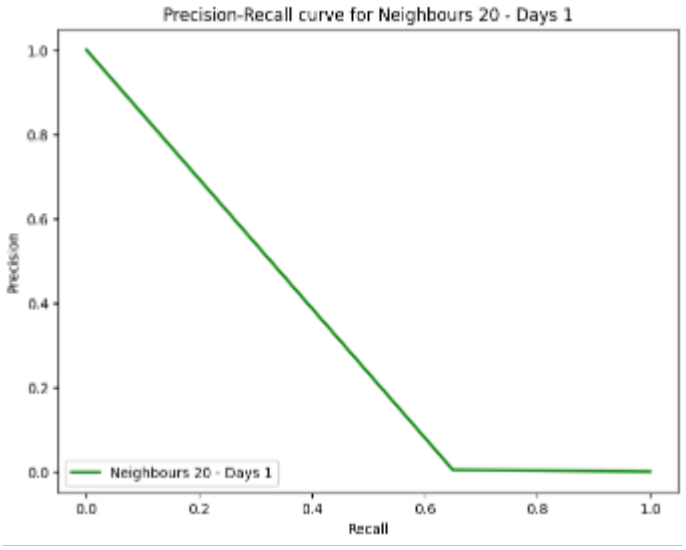


4. **K-Nearest Neighbors (KNN):**
   - **AUC Score:** 0.78 (fairly good performance).
   - **Key Observations:**
     - ROC curve showed a sharp rise initially, indicating effective TPR to FPR trade-off.
     - PR curve demonstrated high initial precision, but precision dropped with higher recall.
   - **Confusion Matrix Analysis:**
     - High number of true negatives with relatively better recall compared to other models.

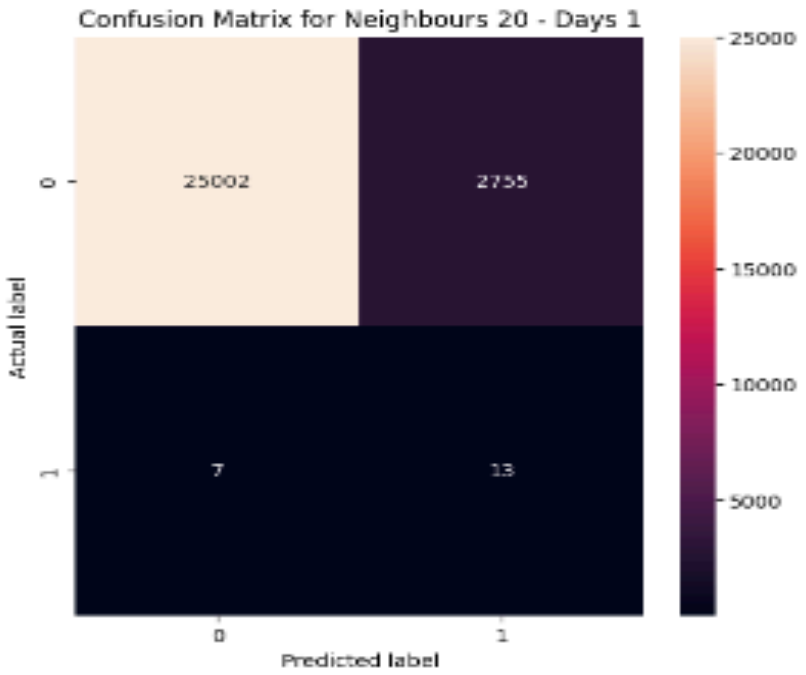
- ROC Curve:



- PR Curve:

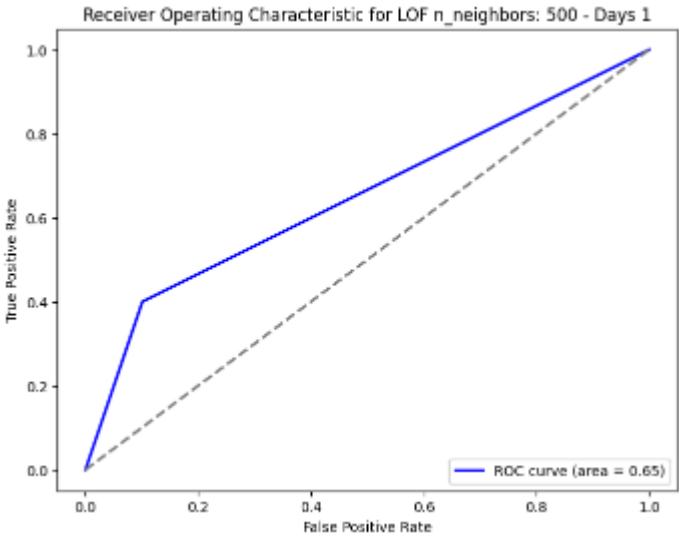


- Confusion Matrix

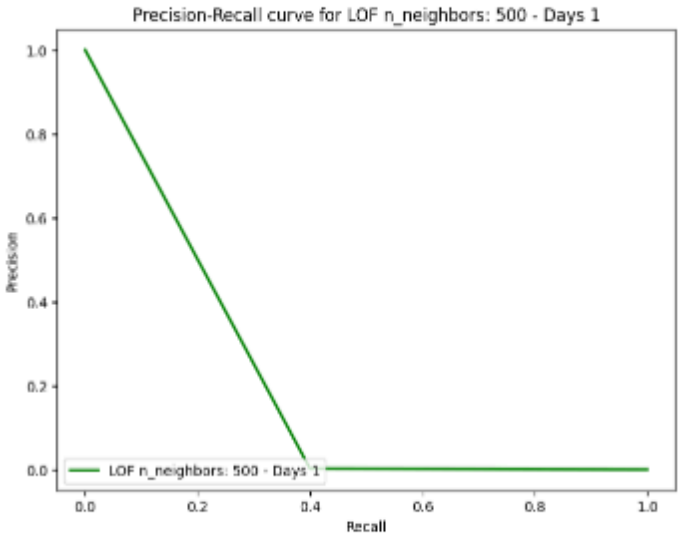


5. **Local Outlier Factor (LOF):**
    - **AUC Score:** 0.65 (limited discriminative ability).
    - **Key Observations:**
        - ROC curve displayed limited performance improvements over random guessing.
        - PR curve showed a steep drop in precision with increasing recall, indicating poor balance.
    - **Confusion Matrix Analysis:**
        - Struggled with identifying true positives effectively.

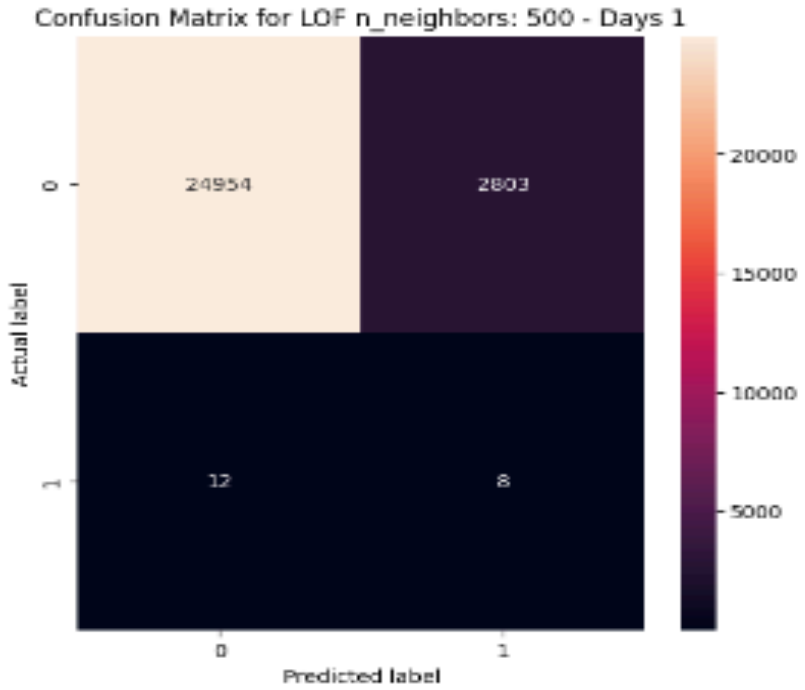
- ROC Curve:



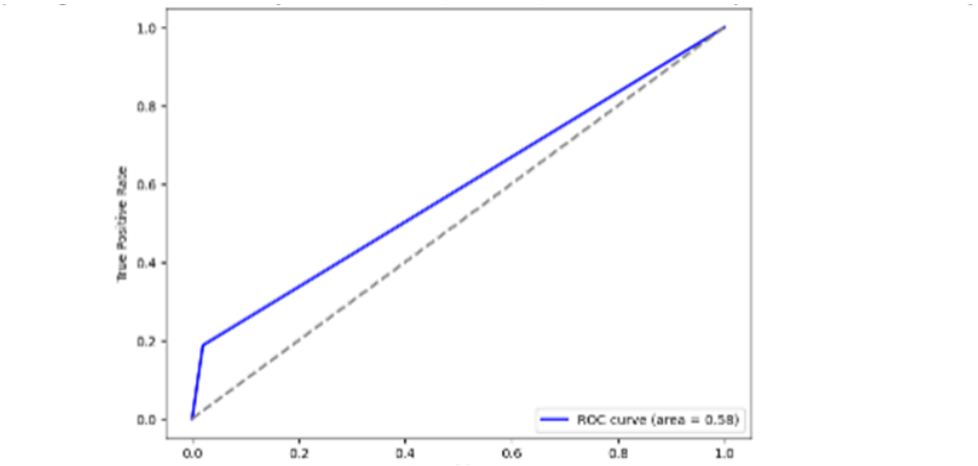
- PR Curve:



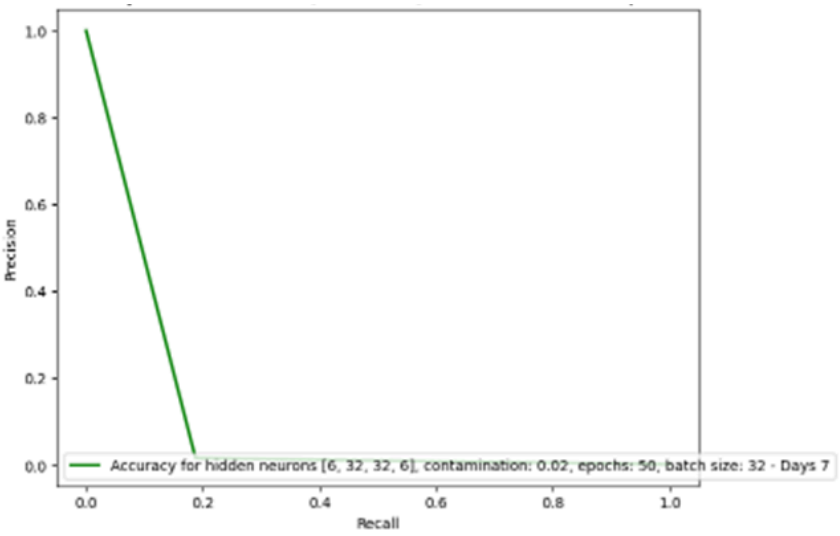
- Confusion Matrix



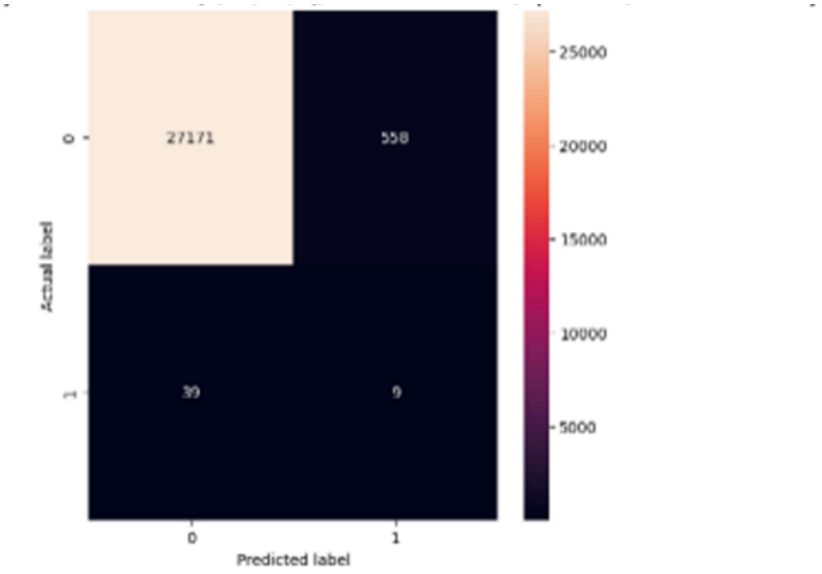
6. **Autoencoder:**
    - **AUC Score:** 0.58 (marginally better than random guessing).
    - **Key Observations:**
        - ROC curve was marginally above the diagonal, suggesting weak performance.
        - PR curve showed significant trade-offs, with precision dropping drastically at higher recall.
    - **Confusion Matrix Analysis:**
        - Heavily biased towards true negatives; poor true positive identification.
    - ROC Curve:



- PR Curve:

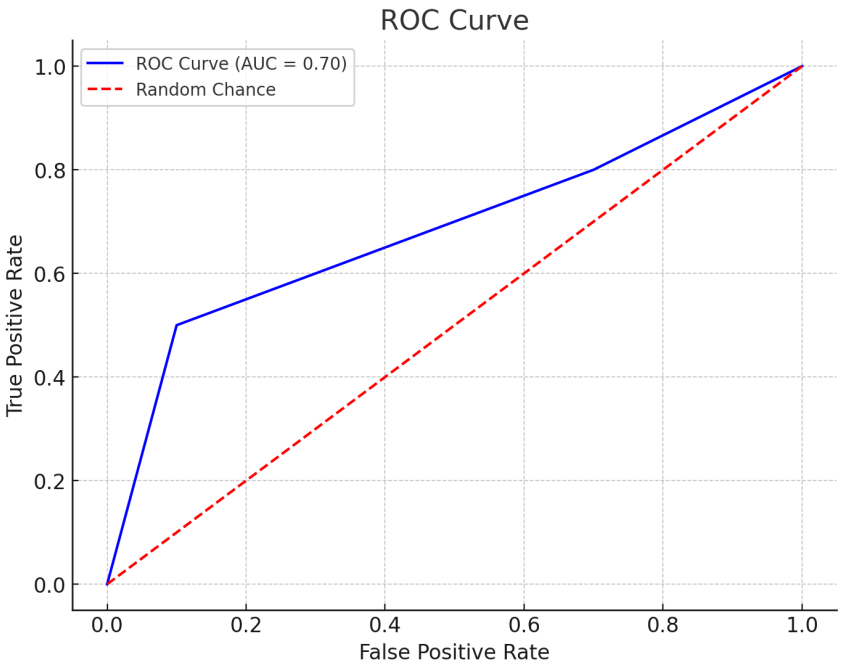


- Confusion Matrix

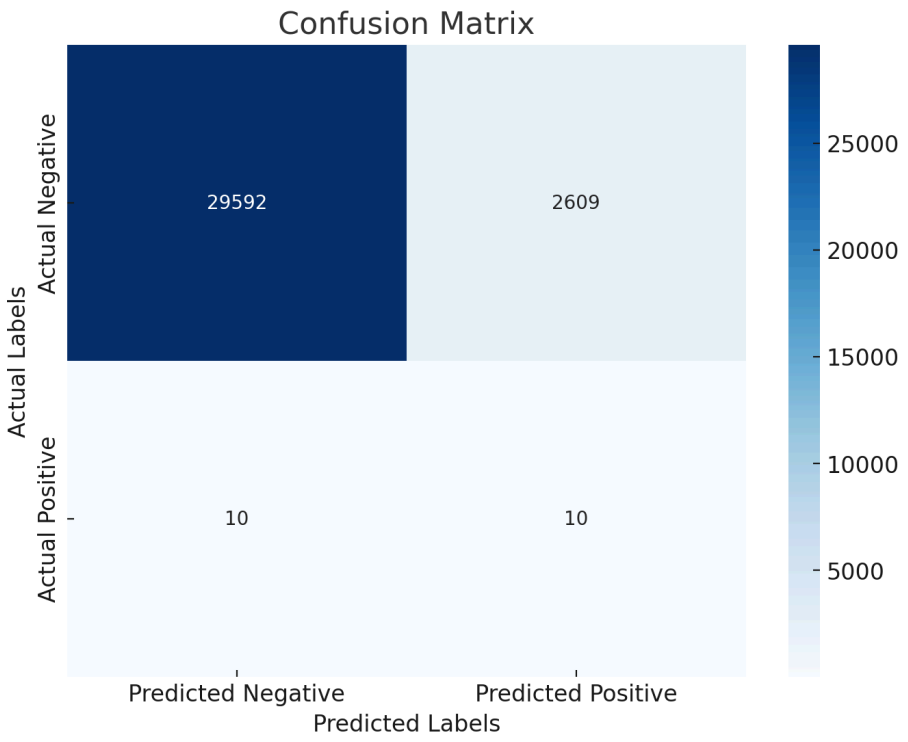


7. OCSVM

- **AUC Score:** 0.70 (fair performance)**.**
- **Key Observations:**
    - ROC curve showed a steep ascent initially, suggesting good TPR at low FPR but plateaued as thresholds increased.
    - PR curve indicated high precision at low recall but a significant drop as recall increased.
- **Confusion Matrix Analysis:**Balanced false positives and false negatives, with moderate identification of true positives.
- ROC Curve:



- **Confusion Matrix**



5. Lessons Learned

1. **Challenges in Data Quality:**

- Managing data discrepancies, such as multiple rows with identical timestamps, proved to be more complex than initially anticipated. Although a methodology for grouping these rows was developed in collaboration with HP professionals, the approach lacked robustness and led to potential ambiguities in the processed dataset.
- The contamination window hypothesis provided a framework for data labeling in the absence of predefined labels. However, it introduced assumptions that may not fully align with the actual data patterns, potentially impacting model performance.
- Imbalanced Data: The dataset was heavily imbalanced, with a significantly smaller proportion of anomaly instances compared to normal instances.

2. **Model Evaluation Limitations:**
   - Differentiating model results was particularly challenging due to overlapping performance metrics. The absence of clearly defined benchmarks or a gold standard for anomaly detection made it difficult to ascertain which models were truly effective.
   - Imbalanced data further complicated the evaluation, as models often struggled to identify true anomalies while maintaining acceptable false positive rates.
3. **Explainability Gaps:**
   - Providing interpretability for unsupervised models remains a significant challenge. Despite efforts to incorporate explainability features, stakeholders found it difficult to understand and trust the outputs of the models.

6. Future Scope

1. **Enhanced Data Processing Frameworks:**

    - Develop and integrate more sophisticated data preprocessing pipelines to handle discrepancies like redundant timestamps or missing values more effectively.
2. **Robust Model Evaluation Techniques:**
    - Establish a standardized framework for comparing model performance, including the use of domain-specific metrics and validation strategies to mitigate the impact of data imbalances.
    - Introduce ensemble methods to leverage strengths across multiple models and improve overall anomaly detection accuracy.
3. **Advanced Labeling Strategies:**
    - Investigate semi-supervised or active learning methods to reduce reliance on assumptions like contamination windows, enabling more accurate data labeling.
4. **Real-Time Integration:**
    - Design and deploy real-time anomaly detection systems capable of integrating seamlessly with printer diagnostic processes, providing actionable insights in operational settings.
5. **Improved Explainability Features:**
    - Focus on developing tools for visualizing and interpreting model decisions, such as feature importance heatmaps or scenario-based explanations, to enhance stakeholder confidence in the outputs.

7. Milestones

Key Milestones Achieved:

1. **Background and Data Exploration:**
    - Explored relevant research to identify techniques suitable for unsupervised anomaly detection.

    - Set up the initial framework for data analysis and processing using sample datasets provided by HP.
2. **Data Cleaning and Analysis:**
    - Conducted comprehensive Exploratory Data Analysis (EDA) to uncover patterns and inconsistencies in the dataset.
    - Collaborated with HP professionals to address data quality issues, such as missing values and redundant timestamps, leading to the development of a grouping methodology.
    - Finalized a contamination window hypothesis to address the lack of explicit labels, enabling a structured approach to model evaluation.
3. **Model Building and Evaluation:**
    - Implemented and tested multiple unsupervised anomaly detection algorithms, including Isolation Forest, Autoencoders, KNN, LOF, GMM, KDE and OCSVM.
    - Conducted hyperparameter tuning using tools like wandb to improve model performance and documented the findings.
4. **Visualization and Documentation:**
    - Created comprehensive visualizations, including ROC and PR curves, confusion matrices, and time-series anomaly plots.
    - Maintained a centralized GitHub repository to store and share all project artifacts, ensuring accessibility and transparency.

---

8. Handover and Transition

Deliverables Provided:

1. **Final Models:**

    - All trained and validated machine learning models, including their configurations and tuning details.
2. **Data Analysis Reports:**
    - Reports detailing the EDA process, data quality challenges, and final cleaning methodologies.
3. **Documentation:**
    - Technical documentation for all implemented algorithms, including code explanations and hyperparameter tuning strategies.
4. **GitHub Repository:**
    - A well-maintained repository containing:
        - Source code for all models.
        - Notebooks for EDA and visualization.
        - Final visualizations and project reports.

Transition Notes:

- The contamination window hypothesis and grouping methodology require further validation in real-world scenarios.

---

9. Conclusion

The project aimed to develop machine learning models capable of detecting anomalies in printer diagnostics using PEFS data. Despite significant challenges, such as data quality issues and the lack of explicit labels, the collaboration between IIT Roorkee and HP resulted in the creation of a structured approach to anomaly detection.

Key Takeaways:

1. **Strengths:**
    - The project delivered foundational insights into handling multivariate data for anomaly detection.
    - Collaboration fostered innovative solutions, such as the contamination window hypothesis and grouping methodology.
2. **Challenges:**
    - Data inconsistencies, such as redundant timestamps, and assumptions in data labeling hindered model performance.
    - The dataset contained a large number of features, which increased computational complexity and made model optimization more challenging. This required careful feature selection and dimensionality reduction techniques.
    - The dataset was heavily imbalanced, with a significantly smaller proportion of anomaly instances compared to normal instances. This imbalance made it difficult for models to effectively identify anomalies without overfitting to normal patterns.
    - The lack of clear criteria for what constituted an anomaly introduced ambiguity into the labeling process. The assumptions made for the contamination window might not have captured all meaningful anomalies.
3. **Future Potential:**
    - Improved data preprocessing and advanced labeling techniques can significantly enhance model accuracy and reliability.
    - Real-time anomaly detection systems hold promise for integration into printer diagnostic workflows, enabling proactive maintenance.

In conclusion, while the project highlighted critical limitations in current methodologies, it also laid the groundwork for future research and development in anomaly detection for printer diagnostics. The deliverables and insights provided offer a robust starting point for subsequent iterations and improvements.