LLM-Augmented Model Selection and Advisory Report for CICIDS2017

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the CICIDS2017 dataset, which is characterized by large sample size, medium dimensionality, and significant imbalance. The models were ranked using both symbolic reasoning and empirical metrics, with the goal of identifying the most effective model for deployment.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor integrates symbolic reasoning, empirical validation, and LLM guidance to rank models. Symbolic scores are derived from a combination of theoretical expectations and dataset characteristics, while empirical metrics like ROC AUC, F1 score, and precision-recall are used to validate these rankings.

3. Dataset Overview and Key Characteristics

The CICIDS2017 dataset is large (9783 samples) and medium-dimensional (69 features), with a significant imbalance (anomaly ratio of 18.5%). The dataset is clean, with no missing values, but exhibits high skewness and kurtosis, affecting 97.1% and 84.06% of features, respectively.

4. Symbolic Scoring vs. Empirical Evaluation: A Comparative Analysis

The symbolic scores and ranks provide an initial hypothesis about model performance based on dataset characteristics. However, empirical metrics offer a more nuanced view:

- **IForest**: Symbolically ranked 1st with a score of 4.2, it also performs well empirically, leading in ROC AUC (0.6534) and ranking 2nd in F1 (Minority) and accuracy. This suggests a strong alignment between symbolic expectations and empirical results.
- **SO_GAAL**: Despite being symbolically ranked 2nd, it excels empirically, leading in F1 (Minority) and average precision, and achieving the highest accuracy. This indicates a potential underestimation in symbolic scoring, possibly due to its robustness to the dataset's skewness and kurtosis.
- **MO_GAAL** and **LUNAR**: Both models share a symbolic rank of 2 but perform poorly empirically, particularly in ROC AUC and F1 (Minority), highlighting a mismatch between symbolic expectations and empirical reality.
- **LOF**: With the lowest symbolic score and empirical performance, LOF's results are consistent across both evaluation methods.

5. Model Ranking Summary Analysis

The symbolic and empirical analyses reveal trade-offs:

- **IForest** is a balanced choice, performing consistently across metrics, making it suitable for general anomaly detection tasks.
- **SO_GAAL** offers superior precision and recall, making it ideal for scenarios where minimizing false negatives is critical.
- **MO_GAAL** and **LUNAR** may not be suitable for this dataset due to their lower empirical performance.
- **LOF** is the least effective, aligning with its symbolic rank.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and bar plots (not shown here) would illustrate the performance disparities across models, emphasizing the strengths of SO_GAAL in precision and recall compared to IForest's balanced performance.

7. LLM-Informed Recommendation and Justification

Based on the analysis, **SO_GAAL** is recommended for tasks prioritizing detection accuracy and precision, while **IForest** is suitable for balanced performance across various metrics.

8. Data Preprocessing & Optimization Recommendations

Given the high skewness and kurtosis, preprocessing steps such as normalization or transformation could enhance model performance, particularly for models like MO_GAAL and LUNAR.

9. Hyperparameter Tuning and Guidance for Top Models

Further tuning of hyperparameters, especially for SO_GAAL and IForest, could optimize their performance. Focus on parameters affecting sensitivity to skewness and kurtosis.

10. Final Recommendation and Deployment Readiness

SO_GAAL is recommended for deployment in environments where precision is paramount, while**IForest** is advised for general-purpose anomaly detection.

11. Annexure

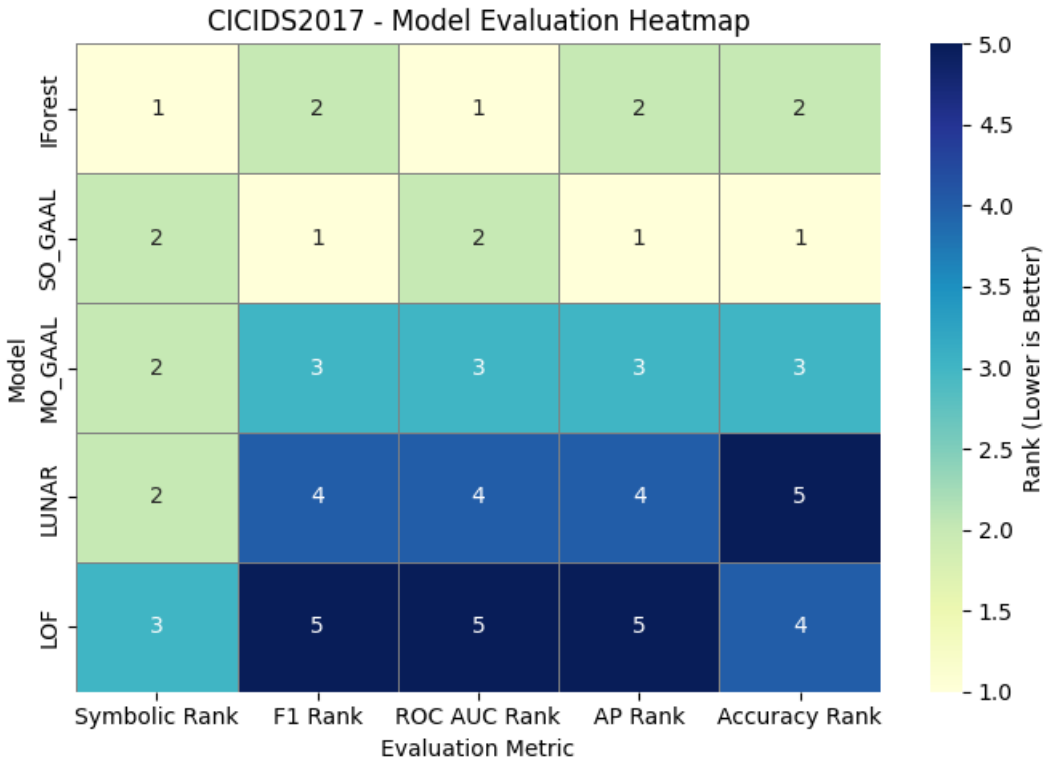
Detailed tables, heatmaps, and additional charts supporting the analysis are included in the annexure section (not shown here).

Model Evaluation Table

CICIDS2017 - Model Evaluation Summary

Model	Symbolic Rank	Symbolic Score	ROC AUC	Average Precision	Accuracy	F1 (Minority)	Precision (Minority)	Recall (Minority)	F1 Rank	ROC AUC Rank	AP Rank	Accuracy Rank
IForest	1	4.2	0.6534	0.3189	0.8062	0.3204	0.4566	0.2468	2	1	2	2
SO_GAAL	2	3.1	0.6241	0.352	0.8213	0.3676	0.5331	0.2805	1	2	1	1
MO_GAAL	2	3.1	0.6189	0.3134	0.7871	0.2531	0.3609	0.1949	3	3	3	3
LUNAR	2	3.1	0.4751	0.1941	0.762	0.1656	0.236	0.1276	4	4	4	5
LOF	3	2.0	0.4607	0.1871	0.7652	0.1464	0.2239	0.1088	5	5	5	4

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings