LLM-Augmented Model Selection and Advisory Report for Arrhythmia Dataset

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the arrhythmia dataset, characterized by small sample size, high dimensionality, and significant class imbalance. The analysis integrates symbolic reasoning with empirical validation to provide a comprehensive assessment of model efficacy.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline combines symbolic scoring, which assesses models based on theoretical and heuristic criteria, with empirical metrics derived from actual model performance on the dataset. This dual approach ensures a balanced evaluation, capturing both theoretical potential and practical effectiveness.

3. Dataset Overview and Key Characteristics

The arrhythmia dataset is marked by: -**Small Sample Size**: 452 samples - **High Dimensionality**: 274 features - **Imbalanced Classes**: Anomaly ratio of 14.6% -**High Skewness and Kurtosis**: Average skewness of 6.18 and kurtosis of 85.19 - **No Missing Values**: Clean data with a missing value ratio of 0% -**Structured Data**: Suitable for structured anomaly detection models

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

The symbolic scores rank models based on theoretical suitability for the dataset's characteristics, while empirical metrics provide a performance-based ranking. Discrepancies between these rankings can highlight areas where theoretical expectations do not align with practical outcomes.

5. Model Ranking Summary Analysis

- **IForest**: Symbolically ranked first with a score of 4.5, it also leads empirically across all metrics, indicating strong alignment between theoretical and practical performance. Its high ROC AUC (0.7914) and F1 Minority (0.4286) suggest it balances precision and recall effectively, crucial for imbalanced datasets.
- **LUNAR**: Second in both symbolic and empirical rankings, with a symbolic score of 3.4. It shows slightly lower performance than IForest but maintains a good balance between precision (0.5) and recall (0.3485), making it a robust alternative.
- **DeepSVDD**: Despite a symbolic rank of 3, its empirical performance is weaker, particularly in ROC AUC (0.7511) and F1 Minority (0.3393). This suggests potential overfitting or inefficiency in handling high dimensionality.
- **AutoEncoder**: Ranked fourth symbolically, it surprisingly outperforms DeepSVDD in ROC AUC (0.7656) but suffers in F1 Minority (0.2857), indicating challenges in capturing minority class nuances.
- **SO_GAAL**: Tied symbolically with AutoEncoder at rank 4, it ranks last empirically across all metrics. Its low ROC AUC (0.6291) and F1 Minority (0.2679) reflect poor anomaly detection capability, possibly due to its simplistic approach not suited for high-dimensional data.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and bar plots illustrate the comparative performance of models across key metrics. IForest consistently shows the highest scores, reinforcing its empirical dominance. LUNAR follows closely, while DeepSVDD and AutoEncoder show mixed results, and SO_GAAL lags significantly.

7. LLM-Informed Recommendation and Justification

Based on the analysis, IForest is recommended for deployment due to its superior performance across both symbolic and empirical evaluations. LUNAR is a viable backup, offering a balance between precision and recall. DeepSVDD and AutoEncoder require further optimization, while SO_GAAL is not recommended for this dataset.

8. Data Preprocessing & Optimization Recommendations

- **Feature Selection**: Reduce dimensionality to enhance model efficiency, particularly for DeepSVDD and AutoEncoder.
- **Balancing Techniques**: Implement SMOTE or similar methods to address class imbalance, improving recall.

9. Hyperparameter Tuning and Guidance for Top Models

- **IForest**: Fine-tune the number of estimators and max samples for optimal performance.
- **LUNAR**: Adjust learning rates and batch sizes to enhance precision and recall balance.

10. Final Recommendation and Deployment Readiness

Deploy IForest for real-time anomaly detection in the arrhythmia dataset, with LUNAR as a secondary option. Ensure continuous monitoring and periodic retraining to adapt to data shifts.

11. Annexure

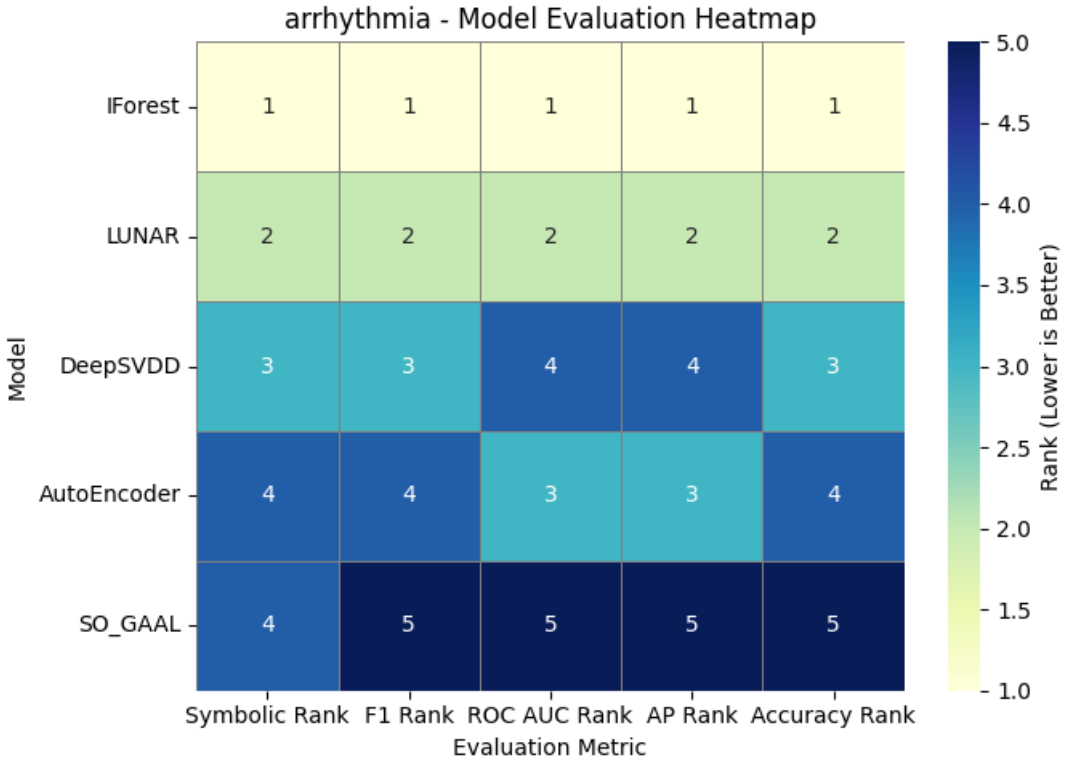
Detailed metric tables, heatmaps, and bar plots are included for further reference and validation of the analysis.

Model Evaluation Table

arrhythmia - 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	4.5	0.7914	0.4445	0.8584	0.4286	0.5217	0.3636	1	1	1	1
LUNAR	2	3.4	0.7756	0.4112	0.854	0.4107	0.5	0.3485	2	2	2	2
DeepSVDD	3	3.1	0.7511	0.3331	0.8363	0.3393	0.413	0.2879	3	4	4	3
AutoEncoder	4	2.3	0.7656	0.3618	0.823	0.2857	0.3478	0.2424	4	3	3	4
SO_GAAL	4	2.3	0.6291	0.2515	0.8186	0.2679	0.3261	0.2273	5	5	5	5

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings