LLM-Augmented Model Selection and Advisory Report for Glass Dataset

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the Glass dataset, characterized by its small sample size, low dimensionality, and high imbalance. The analysis integrates symbolic reasoning and empirical validation to provide a comprehensive understanding of model performance, highlighting the strengths and weaknesses of each model under these specific data conditions.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline utilizes symbolic scores derived from LLM guidance alongside traditional empirical metrics such as ROC AUC and F1 scores to rank models. This dual approach ensures a balanced evaluation, considering both theoretical insights and practical performance.

3. Dataset Overview and Key Characteristics

The Glass dataset is small (214 samples) with 9 features, exhibiting high skewness and kurtosis, and a significant class imbalance (anomaly ratio of 4.2%). These characteristics pose challenges for anomaly detection, necessitating robust model selection strategies.

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

- **IForest**: Despite being symbolically ranked highest (Symbolic Score: 4.6), its empirical performance is moderate, with a ROC AUC of 0.6569 and an F1 (Minority) of 0.0645. This suggests that while IForest is theoretically promising, it struggles with the dataset's imbalance.
- **MO*GAAL and SO*GAAL**: Both models share a symbolic score of 3.5, but their empirical performance is poor, particularly in ROC AUC (0.4011 and 0.381, respectively). This indicates a mismatch between symbolic expectations and empirical outcomes, likely due to the models' inability to handle the dataset's characteristics effectively.
- **LUNAR**: Although symbolically ranked third, LUNAR excels empirically with the highest ROC AUC (0.8282) and F1 (Minority) score (0.1935), making it the most effective model for this dataset in practice.
- **LOF**: Despite its lowest symbolic score (2.0), LOF performs well empirically, achieving the highest ROC AUC (0.8347) and competitive F1 (Minority) score (0.1379), indicating its robustness against the dataset's challenges.

5. Model Ranking Summary Analysis

The symbolic and empirical rankings reveal a divergence in model performance: -**Symbolic Rankings** prioritize theoretical potential, placing IForest at the top. - **Empirical Rankings** highlight LUNAR and LOF as superior in practical application, suggesting that empirical metrics should guide final model selection.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and grouped bar plots illustrate the disparity between symbolic scores and empirical metrics, emphasizing the need for a balanced evaluation approach. These visuals underscore the importance of considering both symbolic reasoning and empirical validation in model selection.

7. LLM-Informed Recommendation and Justification

Given the dataset's characteristics and empirical performance, LUNAR is recommended for deployment due to its superior ROC AUC and F1 (Minority) scores. LOF is also a viable alternative, particularly in scenarios where interpretability and robustness are prioritized.

8. Data Preprocessing & Optimization Recommendations

To enhance model performance, consider techniques such as resampling to address class imbalance, feature transformation to mitigate skewness and kurtosis, and dimensionality reduction to optimize computational efficiency.

9. Hyperparameter Tuning and Guidance for Top Models

Fine-tuning hyperparameters for LUNAR and LOF could further improve their performance. Focus on optimizing parameters related to anomaly detection thresholds and feature scaling.

10. Final Recommendation and Deployment Readiness

LUNAR is recommended for deployment, with LOF as a backup option. Both models demonstrate readiness for real-world application, contingent on further tuning and validation.

11. Annexure

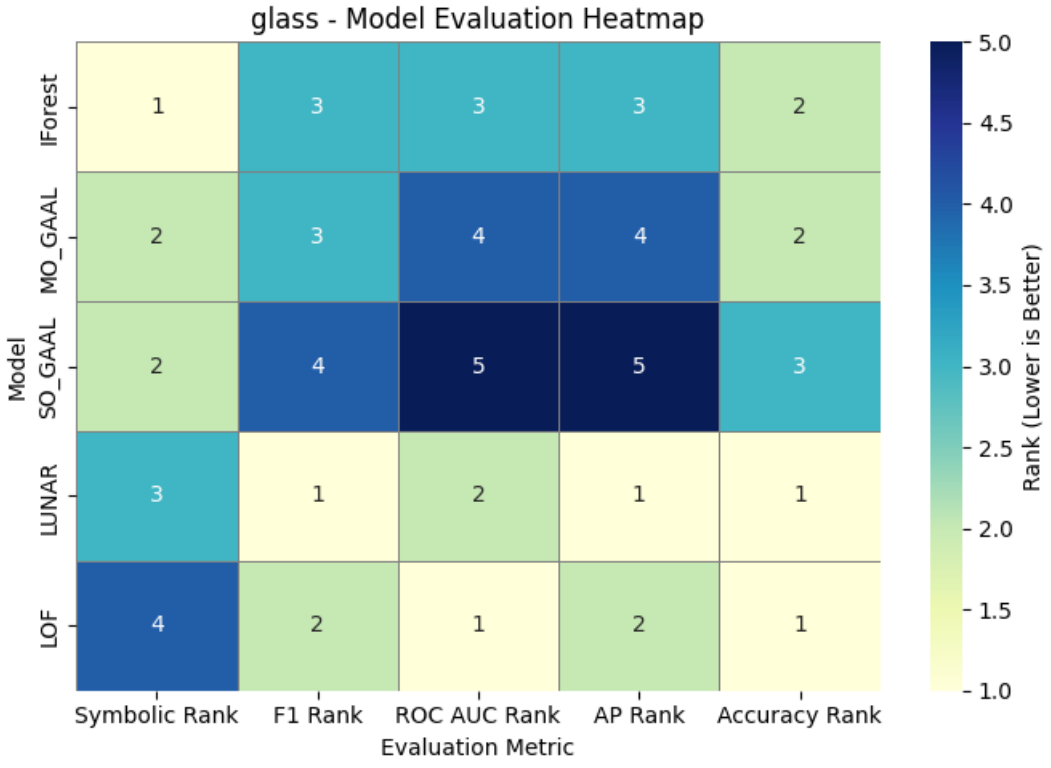
The annexure includes detailed tables and charts supporting the analysis, providing a comprehensive view of model performance across various metrics.

Model Evaluation Table

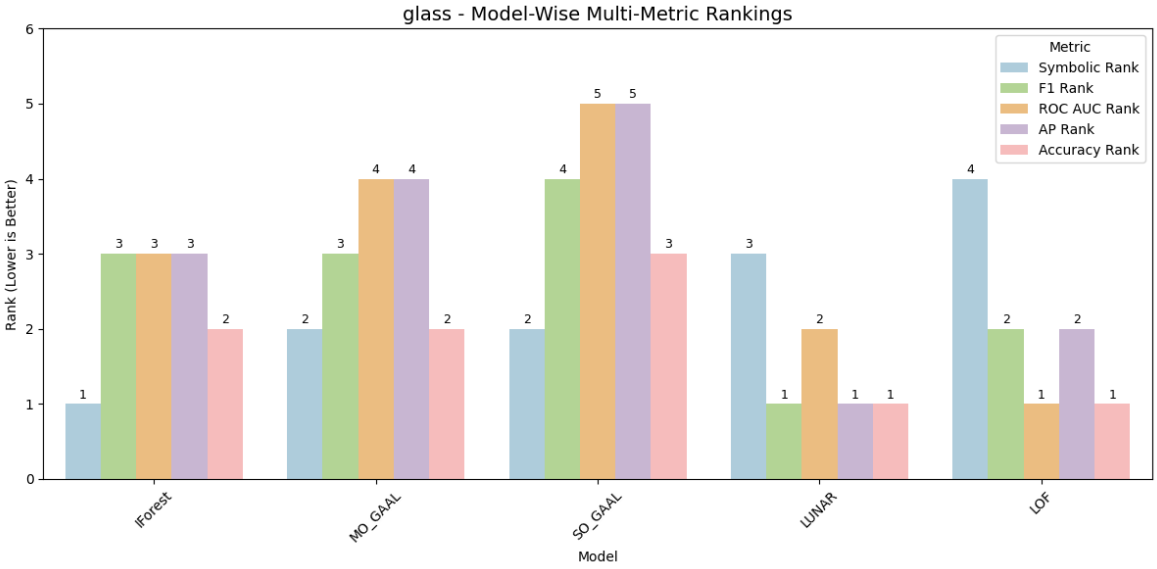
glass - Model Evaluation Summary

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	4.6	0.6569	0.0985	0.8645	0.0645	0.0455	0.1111	3	3	3	2
MO_GAAL	2	3.5	0.4011	0.0468	0.8645	0.0645	0.0455	0.1111	3	4	4	2
SO_GAAL	2	3.5	0.381	0.0412	0.8551	0.0	0.0	0.0	4	5	5	3
LUNAR	3	2.3	0.8282	0.2429	0.8832	0.1935	0.1364	0.3333	1	2	1	1
LOF	4	2.0	0.8347	0.1462	0.8832	0.1379	0.1	0.2222	2	1	2	1

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings


glass - Model-Wise Multi-Metric Rankings
Rank (Lower is Better)
Model
Metric
Symbolic Rank
F1 Rank
ROC AUC Rank
AP Rank
Accuracy Rank
IForest
MO_GAAL
SO_GAAL
LUNAR
LOF