LLM-Augmented Model Selection and Advisory Report for Shuttle Dataset

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the Shuttle dataset, characterized by large sample size, low dimensionality, and significant class imbalance. The analysis integrates symbolic reasoning and empirical validation to provide a comprehensive understanding of model performance.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline combines symbolic scoring, which assesses models based on theoretical suitability, with empirical metrics derived from model evaluation on the dataset. This dual approach ensures robust model selection by balancing theoretical expectations with actual performance.

3. Dataset Overview and Key Characteristics

The Shuttle dataset consists of 49,097 samples with 9 features. It is highly imbalanced, with an anomaly ratio of approximately 7.15%. The dataset is clean, with no missing values, but exhibits high skewness and kurtosis, indicating potential challenges for models sensitive to distributional assumptions.

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

- **IForest**: Symbolically ranked highest with a score of 4.2, it also leads in empirical metrics with a ROC AUC of 0.9975 and an F1 Minority score of 0.822. This alignment suggests that IForest is both theoretically and practically well-suited for this dataset.
- **MO*GAAL, LUNAR, SO*GAAL**: All share a symbolic rank of 2 with a score of 3.1. However, their empirical performance varies significantly. MO_GAAL ranks second in empirical metrics but shows a considerable drop in ROC AUC (0.8265) and F1 Minority (0.4448) compared to IForest.
- **LOF**: Despite being symbolically ranked third, its empirical performance is the weakest, with a ROC AUC of 0.521 and an F1 Minority score of 0.1267, indicating a mismatch between symbolic expectations and actual results.

5. Model Ranking Summary Analysis

- **IForest** consistently ranks first across all empirical metrics, confirming its robustness and adaptability to the dataset's characteristics.
- **MO_GAAL** performs moderately well but struggles with precision and recall, highlighting potential issues in detecting minority classes effectively.
- **LUNAR and SO_GAAL** show poor empirical performance despite their symbolic ranking, suggesting that their theoretical strengths do not translate well to this dataset.
- **LOF**'s low symbolic and empirical scores indicate it is unsuitable for this dataset.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and grouped bar plots (not shown here) would likely illustrate the stark contrast between IForest's performance and other models, emphasizing its superior ROC AUC and F1 scores.

7. LLM-Informed Recommendation and Justification

Based on both symbolic and empirical analyses, IForest is recommended for deployment. Its high performance across all metrics and alignment with symbolic expectations make it the most reliable choice for anomaly detection in the Shuttle dataset.

8. Data Preprocessing & Optimization Recommendations

Given the dataset's high skewness and kurtosis, consider normalization or transformation techniques to further enhance model performance, particularly for models sensitive to distributional assumptions.

9. Hyperparameter Tuning and Guidance for Top Models

For IForest, explore tuning the number of trees and sample size to optimize detection capabilities. For MO_GAAL, adjustments in the learning rate and number of generators could improve precision and recall.

10. Final Recommendation and Deployment Readiness

IForest is ready for deployment, given its strong empirical performance and alignment with symbolic reasoning.

Ensure continuous monitoring and periodic retraining to maintain its effectiveness.

11. Annexure

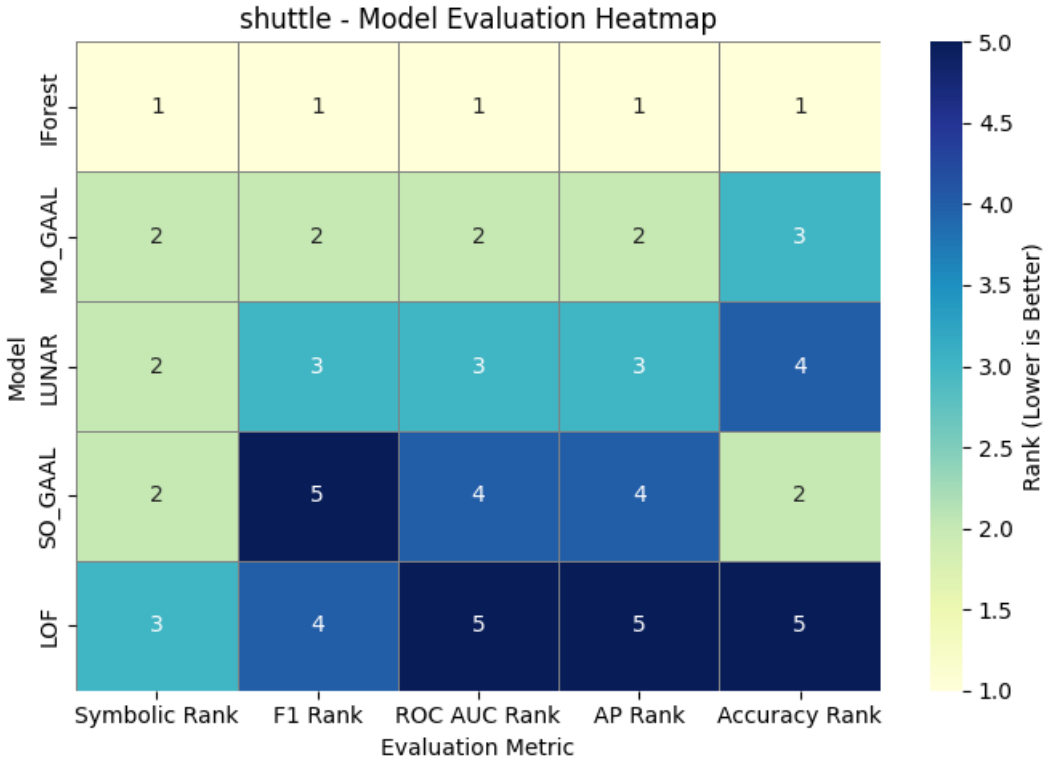
Detailed metrics, model configurations, and additional analyses are available upon request to support further exploration and validation of the findings.

Model Evaluation Table

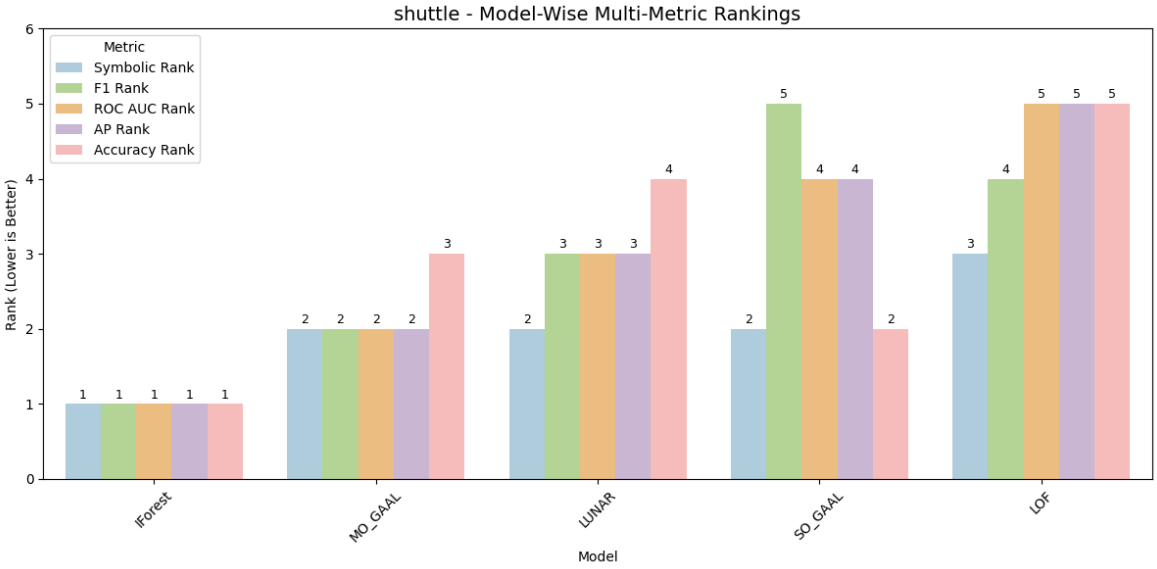
shuttle - Model Evaluation Summary

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	0.9975	0.981	0.9695	0.822	0.7049	0.9858	1	1	1	1
MO_GAAL	2	3.1	0.8265	0.3516	0.9079	0.4448	0.391	0.5158	2	2	2	3
LUNAR	2	3.1	0.6214	0.1707	0.8625	0.1981	0.1699	0.2375	3	3	3	4
SO_GAAL	2	3.1	0.5425	0.1147	0.9285	0.0	0.0	0.0	5	4	4	2
LOF	3	2.0	0.521	0.1085	0.8624	0.1267	0.116	0.1396	4	5	5	5

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings


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