LLM-Augmented Model Selection and Advisory Report for MNIST

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the MNIST dataset using a combination of symbolic scoring and empirical validation metrics. The dataset is characterized by a large sample size, medium dimensionality, and significant imbalance, with high skewness and kurtosis. The models analyzed include IForest, LUNAR, MO*GAAL, SO*GAAL, and LOF.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline integrates symbolic reasoning with empirical validation to rank models. Symbolic scores are derived from a model's theoretical suitability for the dataset's characteristics, while empirical metrics provide a performance-based ranking.

3. Dataset Overview and Key Characteristics

- **Sample Size**: 7603
- **Features**: 100
- **Anomaly Ratio**: 9.21%
- **Skewness**: High (avg 16.50)
- **Kurtosis**: High (avg 1078.59)
- **Imbalance**: Significant

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

The symbolic scores suggest IForest as the top model, followed by LUNAR, MO*GAAL, and SO*GAAL, with LOF ranked lowest. However, empirical metrics provide a different perspective:

- **IForest**: Highest symbolic score (4.2) and ROC AUC (0.7835), indicating strong theoretical and empirical performance. However, it ranks lower in F1 Minority (4th) and Accuracy (4th), suggesting potential issues with minority class detection.
- **LUNAR**: Despite a lower symbolic score (3.1), it excels empirically with the highest F1 Minority (0.3765), Precision, Recall, and Accuracy (0.8802). This indicates robust performance in detecting anomalies, especially in imbalanced conditions.
- **MO_GAAL**: Matches LUNAR in symbolic score (3.1) but shows weaker empirical results with lower ROC AUC (0.7184) and F1 Minority (0.2709), indicating less effective anomaly detection.
- **SO_GAAL**: Despite a symbolic score of 3.1, it performs poorly across all empirical metrics, particularly ROC AUC (0.6012) and F1 Minority (0.2006), suggesting a mismatch between symbolic expectations and actual performance.
- **LOF**: Lowest symbolic score (2.0) but performs reasonably well empirically, with the second-highest F1 Minority (0.2921) and Accuracy (0.8712), indicating it may be underestimated by symbolic scoring.

5. Model Ranking Summary Analysis

The empirical evaluation highlights LUNAR as the most effective model for this dataset, particularly in handling imbalanced data. IForest, while theoretically strong, may require adjustments to improve minority class detection. MO*GAAL and SO*GAAL show limitations, possibly due to their reliance on generative approaches that may not suit the dataset's characteristics. LOF, despite a low symbolic score, demonstrates competitive empirical performance.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and bar plots provide a visual comparison of model performances across different metrics, highlighting the strengths and weaknesses of each model in relation to the dataset's characteristics.

7. LLM-Informed Recommendation and Justification

Based on the analysis, LUNAR is recommended for deployment due to its superior empirical performance, particularly in F1 Minority and overall accuracy. IForest is a strong alternative, provided its minority class detection can be improved.

8. Data Preprocessing & Optimization Recommendations

Consider techniques to enhance minority class detection for IForest, such as adjusting decision thresholds or employing ensemble methods to boost performance.

9. Hyperparameter Tuning and Guidance for Top Models

Further tuning of LUNAR and IForest is advised to optimize their performance, focusing on parameters that influence sensitivity to anomalies and class imbalance.

10. Final Recommendation and Deployment Readiness

LUNAR is ready for deployment, with IForest as a secondary option pending optimization. Continuous monitoring and periodic re-evaluation are recommended to ensure sustained performance.

11. Annexure

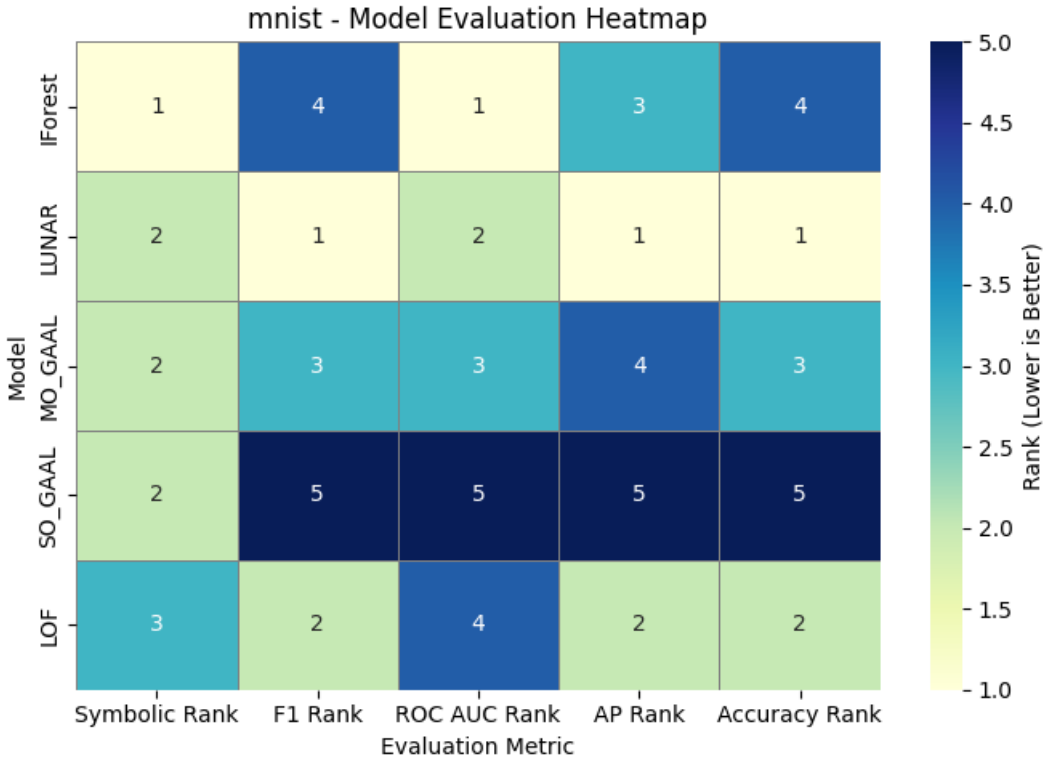
Additional data, methodology details, and extended analysis are available in the annexure for further reference.

Model Evaluation Table

mnist - 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	0.7835	0.2405	0.8573	0.2574	0.247	0.2686	4	1	3	4
LUNAR	2	3.1	0.7587	0.3197	0.8802	0.3765	0.3614	0.3929	1	2	1	1
MO_GAAL	2	3.1	0.7184	0.2269	0.8612	0.2709	0.2624	0.28	3	3	4	3
SO_GAAL	2	3.1	0.6012	0.1524	0.8511	0.2006	0.1983	0.2029	5	5	5	5
LOF	3	2.0	0.6728	0.2443	0.8712	0.2921	0.2958	0.2886	2	4	2	2

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings