LLM-Augmented Model Selection and Advisory Report for WBC Dataset

1. Executive Summary:

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

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline utilizes symbolic scores derived from model characteristics and empirical metrics such as ROC AUC, Average Precision, and F1 Score for minority classes. This dual approach ensures a balanced evaluation of models.

3. Dataset Overview and Key Characteristics

The WBC dataset is small (378 samples) with 30 features, exhibiting high skewness and kurtosis, and a low anomaly ratio (5.56%). The absence of missing values and low sparsity ratio indicate a clean dataset, suitable for anomaly detection tasks.

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

Symbolic scores prioritize models based on theoretical suitability for the dataset's characteristics, while empirical metrics provide performance validation. Discrepancies between these evaluations highlight the importance of context-specific model selection.

5. Model Ranking Summary Analysis

- **IForest**: Achieves the highest symbolic score (3.4) and ranks first in ROC AUC and F1 for minority classes. Its high recall (0.7143) suggests strong anomaly detection capability, though precision is moderate (0.3947), indicating potential false positives.

- **LUNAR**: Shares the symbolic rank with SO*GAAL and MO*GAAL but outperforms them empirically, with a respectable ROC AUC (0.8683) and identical F1 score to IForest. Its symbolic score (2.3) reflects its robustness in handling imbalanced data.

- **SO*GAAL and MO*GAAL**: Despite sharing a symbolic rank with LUNAR, these models perform poorly empirically, with negligible ROC AUC and F1 scores. Their symbolic scores may overestimate their capability in this context.

- **LOF**: Although symbolically ranked lower (2.0), LOF shows strong empirical performance, with the second-highest ROC AUC (0.9384) and the highest accuracy (0.9259). Its balanced precision and recall make it a reliable choice for datasets with high skewness.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and bar plots illustrate the disparity between symbolic and empirical evaluations, emphasizing the need for a nuanced approach in model selection. IForest and LOF consistently demonstrate superior performance across key metrics.

7. LLM-Informed Recommendation and Justification

Based on the analysis, IForest is recommended for its overall balance of recall and precision, making it suitable for detecting anomalies in highly skewed datasets. LOF is a strong alternative, particularly where accuracy is prioritized.

8. Data Preprocessing & Optimization Recommendations

Given the dataset's skewness and kurtosis, consider normalization or transformation techniques to enhance model performance. Addressing class imbalance through resampling or synthetic data generation may further improve detection rates.

9. Hyperparameter Tuning and Guidance for Top Models

Fine-tuning parameters such as the number of estimators for IForest and the contamination factor for LOF can optimize detection capabilities. Cross-validation should be employed to ensure robust performance across different

subsets.

10. Final Recommendation and Deployment Readiness

IForest is the preferred model for deployment, offering a balanced trade-off between detection sensitivity and precision. LOF serves as a viable backup, particularly in scenarios prioritizing accuracy.
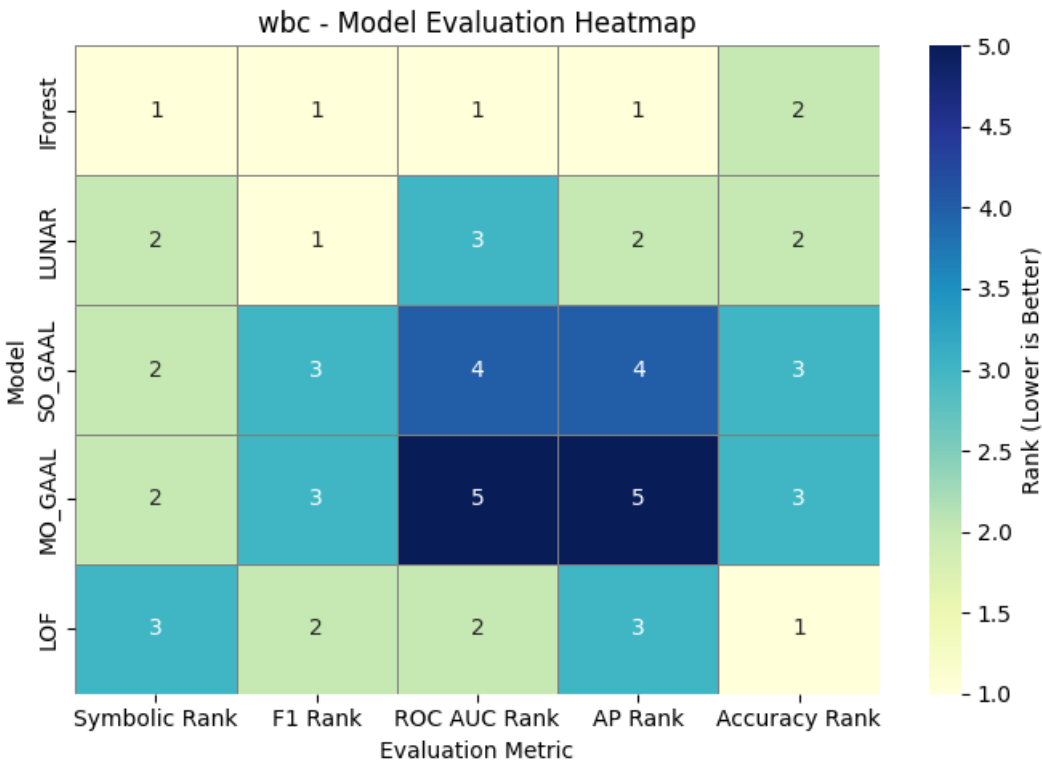
11. Annexure

Detailed metrics, model configurations, and additional visualizations are provided to support the analysis and recommendations outlined in this report.

Model Evaluation Table

wbc - 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	3.4	0.9492	0.6155	0.9233	0.5085	0.3947	0.7143	1	1	1	2
LUNAR	2	2.3	0.8683	0.5103	0.9233	0.5085	0.3947	0.7143	1	3	2	2
SO_GAAL	2	2.3	0.2346	0.0363	0.8439	0.0	0.0	0.0	3	4	4	3
MO_GAAL	2	2.3	0.1459	0.0324	0.8439	0.0	0.0	0.0	3	5	5	3
LOF	3	2.0	0.9384	0.5069	0.9259	0.5	0.4	0.6667	2	2	3	1

Model Evaluation Heatmap

Model-wise Multi-Metric Rankings

wbc - Model-Wise Multi-Metric Rankings