LLM-Augmented Model Selection and Advisory Report for PIMA

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the PIMA dataset, focusing on both symbolic and empirical metrics. The symbolic scores, derived from symbolic reasoning, are juxtaposed with empirical validation metrics such as ROC AUC, Average Precision, and F1 scores to provide a comprehensive analysis of model performance.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline integrates symbolic reasoning with empirical validation and LLM guidance to rank models. Symbolic scores are computed based on theoretical model capabilities, while empirical metrics are derived from model performance on the dataset.

3. Dataset Overview and Key Characteristics

The PIMA dataset is characterized by: - Medium sample size (768 samples) - Low dimensionality (8 features) - Balanced class distribution with an anomaly ratio of 34.9% - Clean data with no missing values - Moderate skewness and kurtosis, with some features exhibiting higher skewness and kurtosis.

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

All models received a symbolic score of 2.8, indicating a theoretical equivalence in potential performance. However, empirical metrics reveal significant differences:

- **LUNAR**: Achieves the highest empirical performance with a ROC AUC of 0.6722 and an F1 (Minority) of 0.2551. This suggests strong discriminatory power and reasonable balance between precision and recall.
- **AutoEncoder**: While it shares the top symbolic rank, its empirical performance is slightly lower, with a ROC AUC of 0.6278 and an F1 (Minority) of 0.2145. This indicates a trade-off between symbolic potential and actual performance.
- **VAE**: Despite a lower ROC AUC of 0.5396, it ranks second in F1 (Minority), suggesting it may better capture minority class instances compared to AutoEncoder.
- **LOF**: Exhibits moderate empirical performance with a ROC AUC of 0.6014 but struggles with minority class detection, as indicated by its F1 (Minority) of 0.1502.
- **DevNet**: Shows the weakest empirical performance across all metrics, with a ROC AUC of 0.4061 and F1 (Minority) of 0.1333, indicating poor anomaly detection capability.

5. Model Ranking Summary Analysis

The symbolic scores suggest theoretical equivalence, but empirical results highlight LUNAR as the superior model in practical terms. The discrepancy between symbolic and empirical rankings underscores the importance of empirical validation in model selection.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and grouped bar plots (not provided here) would typically illustrate the comparative performance across models, highlighting the strengths of LUNAR in ROC AUC and F1 metrics.

7. LLM-Informed Recommendation and Justification

Based on empirical evidence, LUNAR is recommended for deployment due to its superior ROC AUC and F1 scores, indicating robust anomaly detection performance. The symbolic score alignment further supports its selection.

8. Data Preprocessing & Optimization Recommendations

Given the dataset's characteristics, minimal preprocessing is required. However, addressing feature skewness and kurtosis through transformations could enhance model performance.

9. Hyperparameter Tuning and Guidance for Top Models

Fine-tuning LUNAR's hyperparameters, such as learning rate and regularization, could further optimize its performance. Exploring ensemble methods may also enhance robustness.

10. Final Recommendation and Deployment Readiness

LUNAR is deemed ready for deployment, contingent on further hyperparameter tuning and validation on a holdout set to ensure generalizability.

11. Annexure

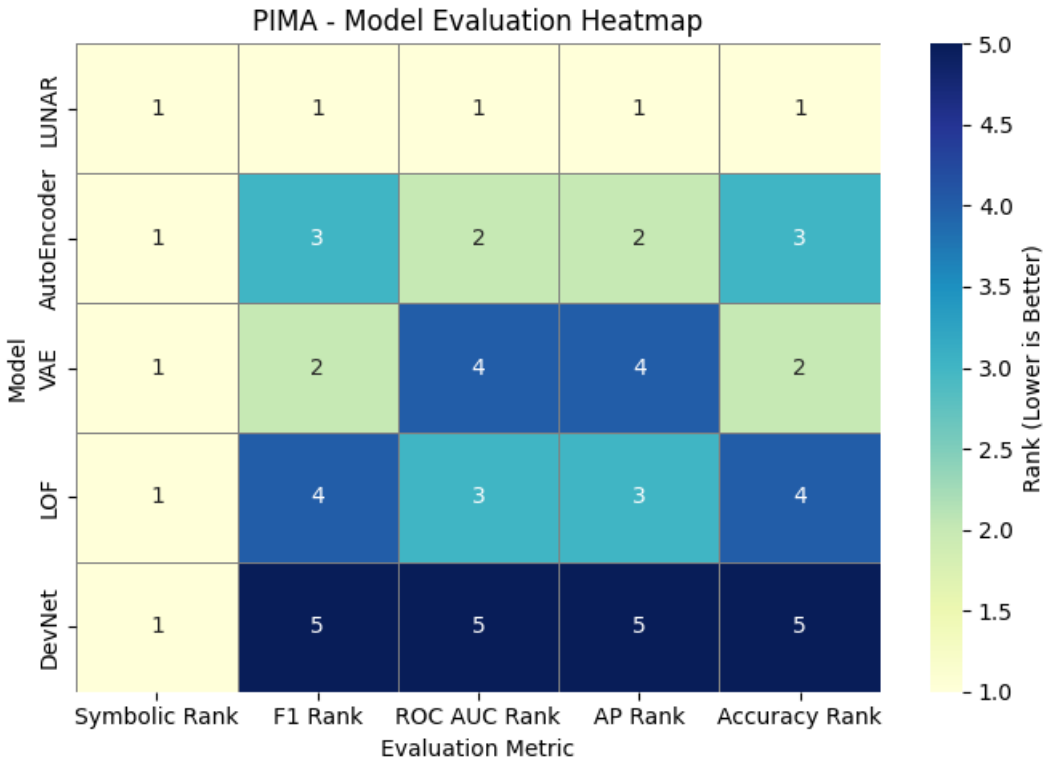
Detailed metrics, model configurations, and additional insights are available in the annexure for further reference.

Model Evaluation Table

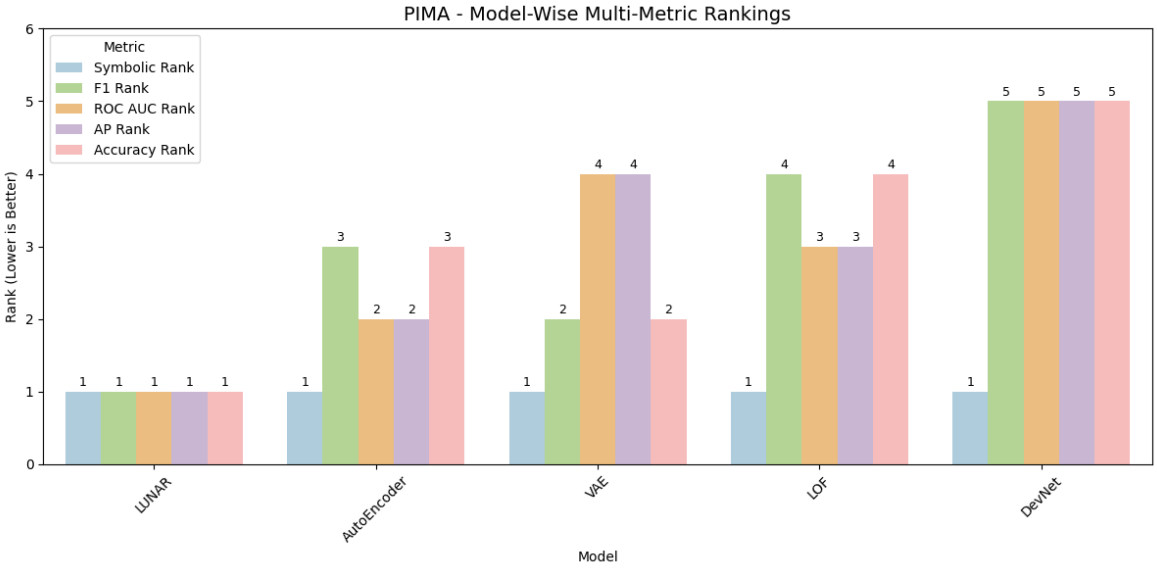
PIMA - Model Evaluation Summary

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
LUNAR	1	2.8	0.6722	0.5031	0.6654	0.2551	0.5714	0.1642	1	1	1	1
AutoEncoder	1	2.8	0.6278	0.4458	0.6471	0.2145	0.4805	0.1381	3	2	2	3
VAE	1	2.8	0.5396	0.3989	0.6484	0.2197	0.4872	0.1418	2	4	4	2
LOF	1	2.8	0.6014	0.418	0.6315	0.1502	0.3846	0.0933	4	3	3	4
DevNet	1	2.8	0.4061	0.312	0.6107	0.1333	0.2987	0.0858	5	5	5	5

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings


PIMA - Model-Wise Multi-Metric Rankings
Metric
Symbolic Rank
F1 Rank
ROC AUC Rank
AP Rank
Accuracy Rank
Rank (Lower is Better)
Model
LUNAR
AutoEncoder
VAE
LOF
DevNet