LLM-Augmented Model Selection and Advisory Report for Satimage2

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the satimage2 dataset. The models are ranked based on symbolic scores and empirical metrics to provide a comprehensive understanding of their effectiveness. The dataset is characterized by a large sample size, medium dimensionality, and a highly imbalanced class distribution.

2. Introduction: Methodology Behind the Recommendation

The analysis integrates symbolic reasoning, empirical validation, and LLM guidance to assess model performance. Symbolic scores provide an initial ranking based on theoretical expectations, while empirical metrics such as ROC AUC, Average Precision, and F1 (Minority) offer practical performance insights.

3. Dataset Overview and Key Characteristics

- **Sample Size:** 5803
- **Features:** 36
- **Anomaly Ratio:** 1.22%
- **Data Quality:** Clean with no missing values, low skewness, and low kurtosis.

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

All models except IForest share a symbolic score of 3.8, indicating a theoretical expectation of similar performance. However, empirical metrics reveal significant differences:

- **VAE** and **IForest** are top performers in terms of ROC AUC and F1 (Minority), indicating strong detection capabilities.
- **IForest** outperforms others in ROC AUC (0.9941) and Average Precision (0.9154), suggesting superior precision and recall balance.
- **LOF** shows a stark contrast with a low ROC AUC (0.5364) and Average Precision (0.0313), highlighting its inadequacy for this dataset.

5. Model Ranking Summary Analysis

- **IForest**: Despite a lower symbolic rank (2), it excels empirically with the highest scores across ROC AUC, Average Precision, and F1 (Minority). This model is particularly effective in detecting anomalies with high recall (0.9718).
- **VAE**: Matches the symbolic score of 3.8 but ranks second empirically. It offers a balanced performance with high recall (0.9296) but lower precision (0.1136).
- **AutoEncoder** and **DeepSVDD**: Both models have similar symbolic scores but lag behind in empirical performance, particularly in precision and recall metrics.
- **LOF**: Despite a symbolic score of 3.8, its empirical performance is poor, making it unsuitable for this dataset.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and bar plots illustrate the disparity between symbolic scores and empirical metrics, highlighting the importance of empirical validation in model selection.

7. LLM-Informed Recommendation and Justification

Based on the analysis, **IForest** is recommended for deployment due to its superior empirical performance, particularly in highly imbalanced datasets like satimage2. Its ability to maintain high precision and recall makes it ideal for anomaly detection tasks.

8. Data Preprocessing & Optimization Recommendations

- **Feature Scaling**: Ensure features are normalized to enhance model performance.
- **Imbalance Handling**: Consider oversampling techniques to improve minority class detection.

9. Hyperparameter Tuning and Guidance for Top Models

- **IForest**: Optimize the number of estimators and contamination parameter to further enhance detection accuracy.
- **VAE**: Adjust latent space dimensions and learning rate for better anomaly representation.

10. Final Recommendation and Deployment Readiness

Deploy **IForest** for its robust performance in detecting anomalies with high precision and recall. Ensure continuous monitoring and retraining to adapt to any changes in data distribution.

11. Annexure

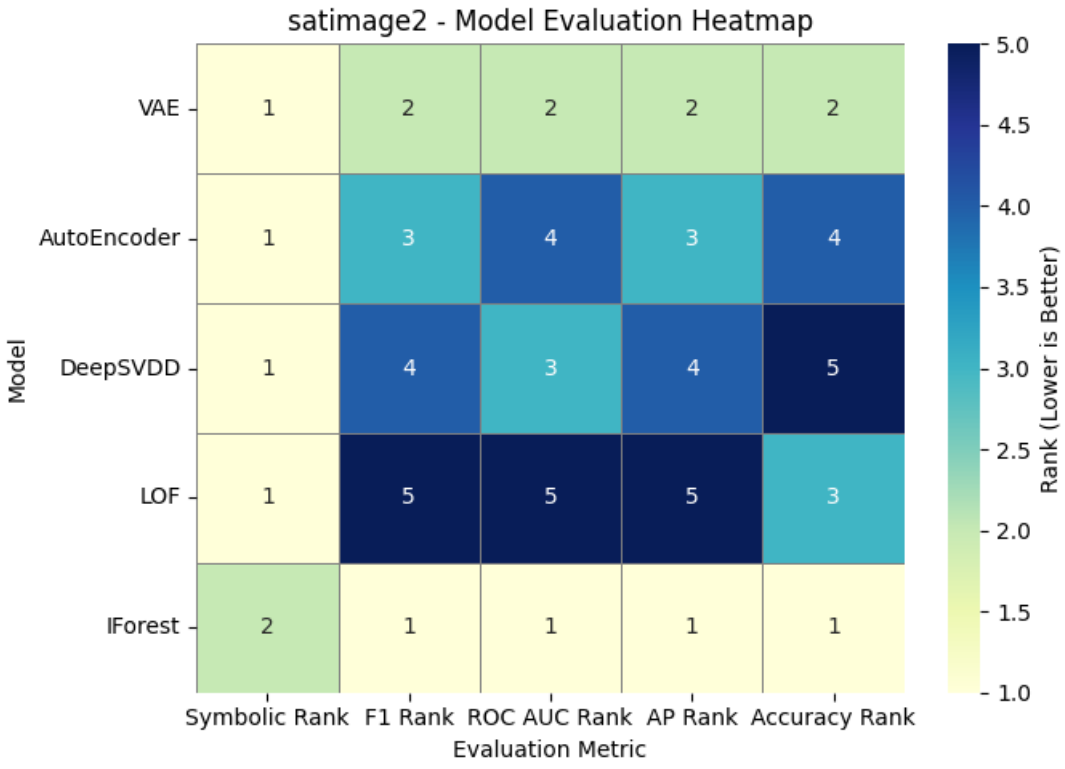
Detailed tables and heatmaps are provided for further reference, illustrating the comparative analysis of model performance metrics.

Model Evaluation Table

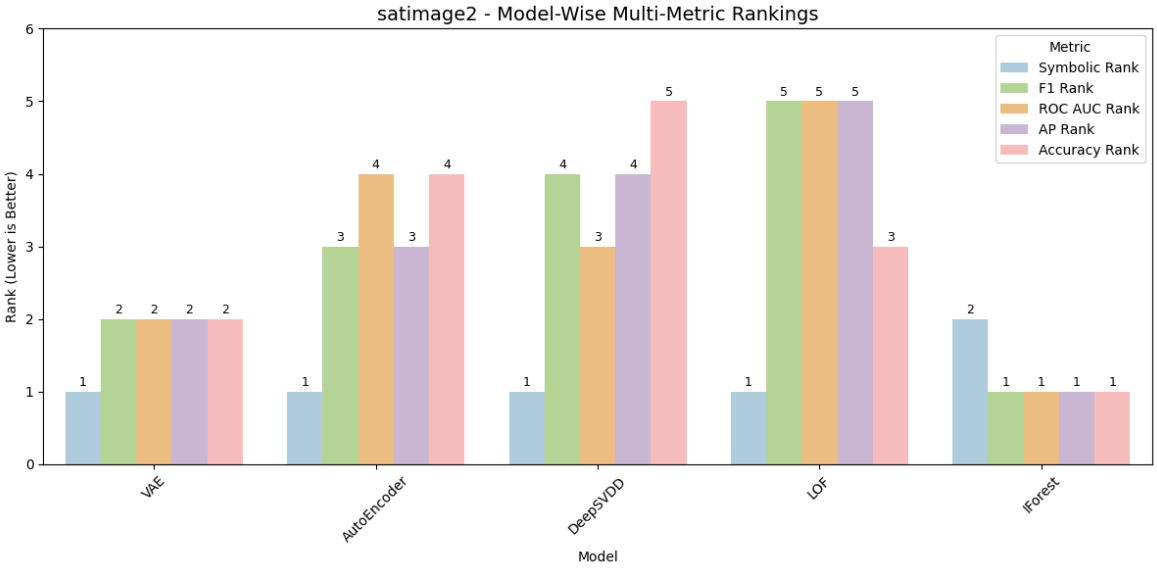
satimage2 - Model Evaluation Summary

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
VAE	1	3.8	0.9837	0.7958	0.9104	0.2025	0.1136	0.9296	2	2	2	2
AutoEncoder	1	3.8	0.9037	0.2987	0.9045	0.1503	0.0843	0.6901	3	4	3	4
DeepSVDD	1	3.8	0.9121	0.1992	0.9035	0.1411	0.0792	0.6479	4	3	4	5
LOF	1	3.8	0.5364	0.0313	0.9082	0.0698	0.0398	0.2817	5	5	5	3
IForest	2	3.2	0.9941	0.9154	0.9114	0.2117	0.1188	0.9718	1	1	1	1

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings


satimage2 - Model-Wise Multi-Metric Rankings
Rank (Lower is Better)
Model
Metric
Symbolic Rank
F1 Rank
ROC AUC Rank
AP Rank
Accuracy Rank
VAE
AutoEncoder
DeepSVDD
LOF
iForest