LLM-Augmented Model Selection and Advisory Report for Vowels Dataset

1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the "vowels" dataset, characterized by medium sample size, medium dimensionality, and a highly imbalanced class distribution. The analysis integrates symbolic scoring and empirical metrics to provide a comprehensive model ranking and recommendation.

2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline combines symbolic reasoning, empirical validation, and LLM guidance to assess model performance. Symbolic scores are derived from a rule-based system considering dataset characteristics, while empirical metrics are obtained from model evaluations on the dataset.

3. Dataset Overview and Key Characteristics

- **Sample Size**: 1456
- **Features**: 12
- **Anomaly Ratio**: 3.43%
- **Data Quality**: Clean, with no missing values, low skewness (0.2532), and low kurtosis (0.4338).

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

All models except IForest received a symbolic score of 3.8, indicating a high potential fit for the dataset based on symbolic reasoning. However, empirical evaluations reveal significant variations in performance metrics like ROC AUC, Average Precision, and F1 scores.

5. Model Ranking Summary Analysis

- **LOF (Local Outlier Factor)**:
  - **Symbolic Score**: 3.8
  - **Empirical Performance**: Highest ROC AUC (0.943) and accuracy (0.921), indicating strong overall detection capability. However, its F1 score for the minority class is lower than AutoEncoder, suggesting potential trade-offs in precision and recall balance.
- **AutoEncoder**:
  - **Symbolic Score**: 3.8
  - **Empirical Performance**: Best F1 score for minority class (0.398) and highest Average Precision (0.5504), indicating superior performance in identifying anomalies despite slightly lower ROC AUC (0.9385) compared to LOF.
- **DeepSVDD**:
  - **Symbolic Score**: 3.8
  - **Empirical Performance**: Lower ROC AUC (0.7634) and F1 score (0.2041) suggest it is less effective for this dataset, despite its symbolic ranking.
- **VAE (Variational Autoencoder)**:
  - **Symbolic Score**: 3.8
  - **Empirical Performance**: Lowest across all metrics, indicating a mismatch between symbolic expectations and empirical results.
- **IForest (Isolation Forest)**:
  - **Symbolic Score**: 3.2
  - **Empirical Performance**: Consistently lower across all metrics compared to LOF and AutoEncoder, aligning with its lower symbolic score.

6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and bar plots illustrate the comparative performance of models across different metrics, highlighting the strengths of LOF and AutoEncoder in ROC AUC and F1 scores, respectively.

7. LLM-Informed Recommendation and Justification

Considering both symbolic and empirical analyses, AutoEncoder is recommended for scenarios prioritizing anomaly detection accuracy (F1 and Average Precision), while LOF is suitable for maximizing overall detection capability (ROC AUC and accuracy).

8. Data Preprocessing & Optimization Recommendations

- Ensure data normalization to enhance model performance.
- Consider feature engineering to improve model interpretability and detection accuracy.

9. Hyperparameter Tuning and Guidance for Top Models

- **AutoEncoder**: Experiment with different architectures and learning rates to optimize anomaly detection.
- **LOF**: Adjust the number of neighbors to balance sensitivity and specificity.

10. Final Recommendation and Deployment Readiness

AutoEncoder is recommended for deployment due to its superior performance in detecting anomalies, particularly in highly imbalanced datasets. LOF can be considered as an alternative where overall detection capability is prioritized.

11. Annexure

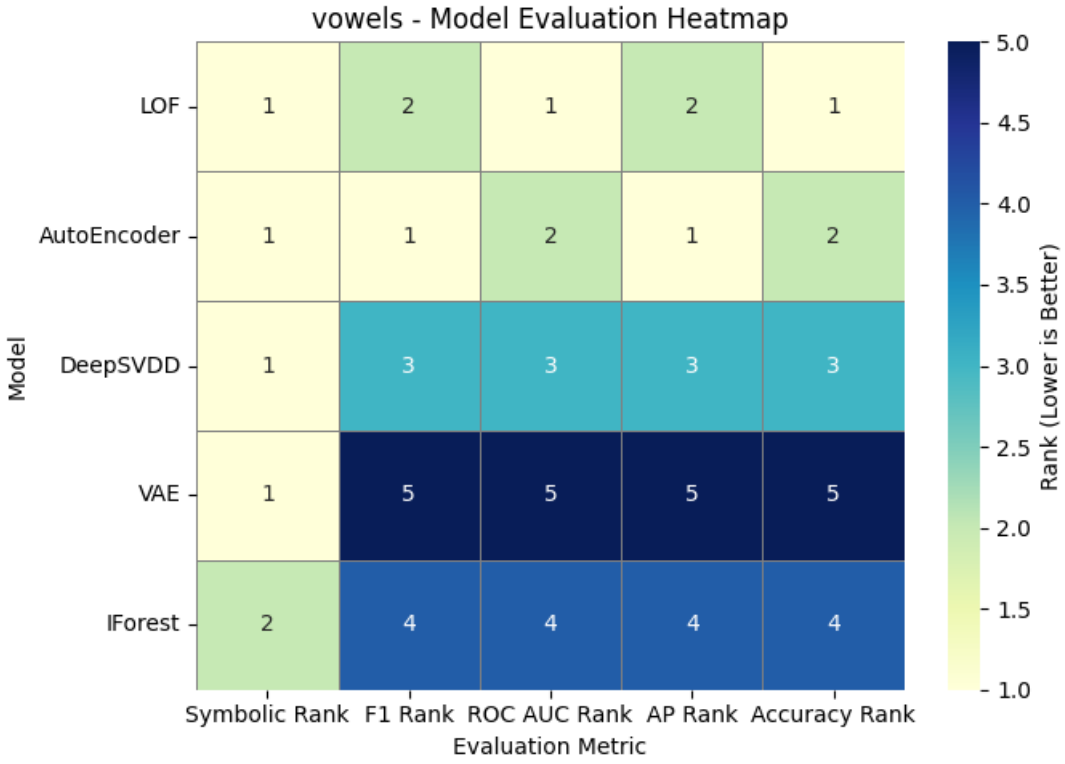
- Detailed empirical results and symbolic scoring criteria.
- Additional visualizations and model-specific insights.

Model Evaluation Table

vowels - 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
LOF	1	3.8	0.943	0.3162	0.921	0.3575	0.2481	0.64	2	1	2	1
AutoEncoder	1	3.8	0.9385	0.5504	0.919	0.398	0.2671	0.78	1	2	1	2
DeepSVDD	1	3.8	0.7634	0.2107	0.8929	0.2041	0.137	0.4	3	3	3	3
VAE	1	3.8	0.6077	0.0536	0.8771	0.0914	0.0612	0.18	5	5	5	5
IForest	2	3.2	0.7446	0.1171	0.8915	0.1939	0.1301	0.38	4	4	4	4

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings