# LLM-Augmented Model Selection and Advisory Report for Letter Dataset

### 1. Executive Summary:

This report provides an in-depth analysis of the anomaly detection models evaluated on the Letter dataset. The models were ranked using both symbolic reasoning and empirical validation metrics. The primary objective is to identify the most suitable model for deployment based on the dataset's characteristics and the models' performance.

### 2. Introduction: Methodology Behind the Recommendation

The evaluation integrates symbolic scoring, which considers theoretical model capabilities, with empirical validation metrics such as ROC AUC, F1 Score, and others. This dual approach ensures a comprehensive assessment of each model's performance.

### 3. Dataset Overview and Key Characteristics

- **Sample Size:** 1600
- **Features:** 32
- **Anomaly Ratio:** 6.25%
- **Characteristics:** Medium sample size, medium dimensionality, imbalanced, clean data with low skewness and kurtosis.

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

All models except DevNet received a symbolic score of 3.8, indicating a theoretical potential for high performance. However, empirical results show significant variation:

- **LOF (Local Outlier Factor):** Achieved the highest empirical scores across all metrics, aligning well with its symbolic score.
- **AutoEncoder:** Despite a high symbolic score, it showed moderate empirical performance, particularly in precision and recall.
- **VAE (Variational Autoencoder) and DeepSVDD:** Both models had high symbolic scores but underperformed empirically, especially in ROC AUC and F1 Minority.
- **DevNet:** Although it had a lower symbolic score, its empirical performance was better than VAE and DeepSVDD, particularly in ROC AUC.

### 5. Model Ranking Summary Analysis

- **LOF** emerged as the top performer, excelling in both symbolic and empirical evaluations. It is particularly effective in scenarios requiring high recall.
- **AutoEncoder** is a viable option when a balance between precision and recall is needed, albeit with lower overall performance compared to LOF.
- **VAE and DeepSVDD** showed a mismatch between symbolic and empirical results, indicating potential issues in handling the dataset's characteristics.
- **DevNet**, while not top-ranked symbolically, provided a reasonable trade-off between symbolic and empirical results, especially in ROC AUC.

### 6. Visual Insights: Heatmap and Grouped Bar Plots Analysis

The heatmap and grouped bar plots illustrate the disparity between symbolic scores and empirical performance. LOF consistently shows darker shades in the heatmap, indicating superior performance across metrics.

### 7. LLM-Informed Recommendation and Justification

Based on the analysis, **LOF** is recommended for deployment due to its robust empirical performance and alignment with symbolic predictions. It is particularly suited for the dataset's imbalanced nature and clean structure.

### 8. Data Preprocessing & Optimization Recommendations

- Ensure data remains clean and structured to leverage LOF's capabilities.
- Consider balancing techniques to further enhance model performance on minority classes.

### 9. Hyperparameter Tuning and Guidance for Top Models

- **LOF:** Focus on tuning the number of neighbors and contamination parameters.

- **AutoEncoder:** Experiment with different architectures and learning rates to improve precision.

**10. Final Recommendation and Deployment Readiness**

Deploy **LOF** for its superior performance and alignment with dataset characteristics. Ensure continuous monitoring and periodic retraining to maintain performance.

**11. Annexure**

Includes detailed metric tables, heatmaps, and additional visual insights supporting the analysis.

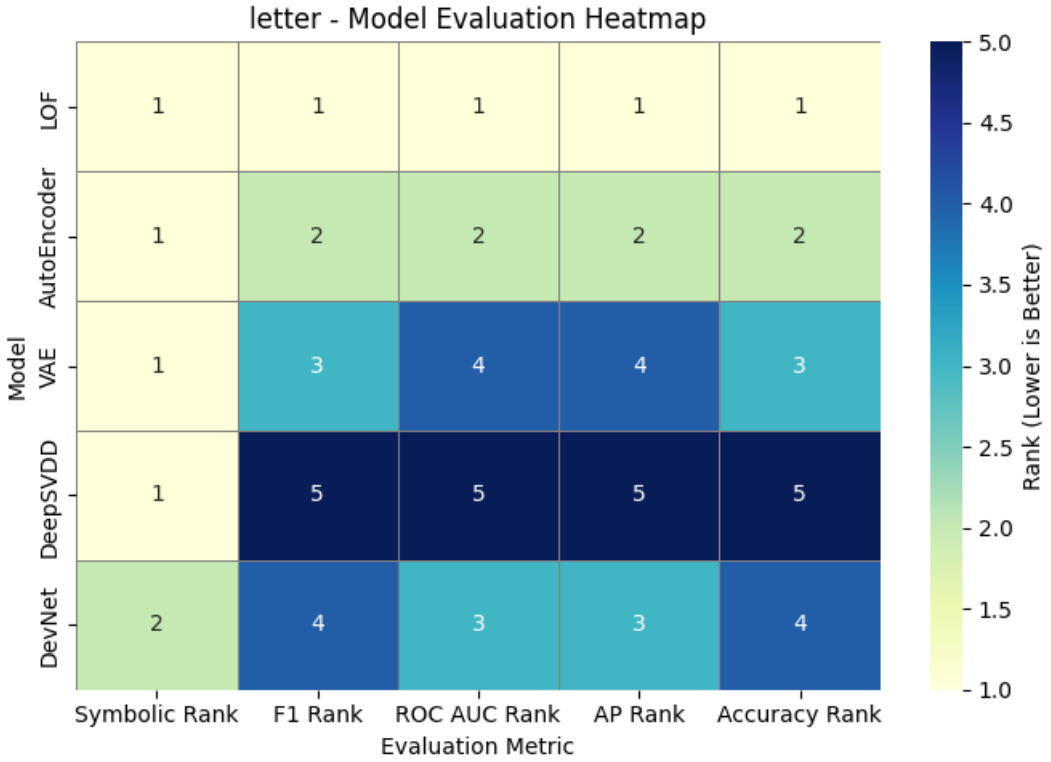
This report concludes that LOF is the most suitable model for the Letter dataset, offering a balance of theoretical potential and empirical validation.

# Model Evaluation Table

**letter - 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 | 0.8988 | 0.4759 | 0.9163 | 0.4508 | 0.3819 | 0.55 | 1 | 1 | 1 | 1 |
| AutoEncoder | 1 | 3.8 | 0.8365 | 0.2795 | 0.895 | 0.3538 | 0.2875 | 0.46 | 2 | 2 | 2 | 2 |
| VAE | 1 | 3.8 | 0.6039 | 0.1139 | 0.8669 | 0.1839 | 0.1491 | 0.24 | 3 | 4 | 4 | 3 |
| DeepSVDD | 1 | 3.8 | 0.5375 | 0.0744 | 0.8562 | 0.1154 | 0.0938 | 0.15 | 5 | 5 | 5 | 5 |
| DevNet | 2 | 2.9 | 0.6778 | 0.116 | 0.865 | 0.1692 | 0.1375 | 0.22 | 4 | 3 | 3 | 4 |

# Model Evaluation Heatmap




# Model-wise Multi-Metric Rankings