

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

## 1. Executive Summary:

This report evaluates the performance of various anomaly detection models on the Musk dataset, characterized by medium sample size, high dimensionality, and a highly imbalanced class distribution. The analysis integrates symbolic reasoning with empirical validation to provide a comprehensive assessment of model performance.

## 2. Introduction: Methodology Behind the Recommendation

The AutoModelAdvisor pipeline employs symbolic scoring to rank models based on their theoretical suitability for the dataset's characteristics, followed by empirical validation through metrics like ROC AUC, Average Precision, and F1 Score for the minority class.

## 3. Dataset Overview and Key Characteristics

- **Sample Size:** 3062
- **Features:** 166
- **Anomaly Ratio:** 3.17%
- **Data Quality:** Clean with no missing values
- **Statistical Properties:** Low kurtosis and moderate skewness

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

The symbolic scoring system ranks models based on their theoretical fit to the dataset's characteristics, while empirical evaluation provides a practical performance measure.

## 5. Model Ranking Summary Analysis

- **IForest:** Symbolically ranked first with a score of 4.3, it also leads in empirical metrics with a near-perfect ROC AUC of 0.9998 and an F1 Minority score of 0.4802. Its perfect recall (1.0) indicates it captures all anomalies, but with a trade-off in precision (0.316), suggesting some false positives.
- **DeepSVDD:** Tied symbolically with AutoEncoder but performs better empirically. It ranks second in ROC AUC (0.9047) and has a balanced F1 Minority score (0.3416), indicating a good trade-off between precision (0.2248) and recall (0.7113).
- **AutoEncoder:** Despite a symbolic score of 4.0, its empirical performance lags with a ROC AUC of 0.8553 and a lower F1 Minority score of 0.2673. It struggles with precision and recall balance, suggesting it may miss some anomalies.
- **SO\_GAAL:** Symbolically ranked third, it shows moderate empirical performance with a ROC AUC of 0.8662. It has a balanced F1 Minority score (0.297), but lower precision (0.2103) and recall (0.5052) compared to DeepSVDD.
- **MO\_GAAL:** Despite sharing the third symbolic rank, it performs poorly empirically with the lowest ROC AUC (0.7683) and F1 Minority score (0.1244), indicating it is less suitable for this dataset.

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

The heatmap and bar plots illustrate the trade-offs between precision and recall across models, highlighting IForest's dominance in recall and DeepSVDD's balanced performance.

## 7. LLM-Informed Recommendation and Justification

For datasets like Musk with high dimensionality and imbalance, IForest is recommended for its superior recall and overall performance. DeepSVDD is a viable alternative when a balance between precision and recall is desired.

## 8. Data Preprocessing & Optimization Recommendations

Consider feature scaling and dimensionality reduction to enhance model performance, particularly for models like AutoEncoder that may benefit from reduced complexity.

## 9. Hyperparameter Tuning and Guidance for Top Models

Fine-tuning parameters such as the number of trees in IForest or the network architecture in DeepSVDD can further optimize performance.

10. Final Recommendation and Deployment Readiness

IForest is ready for deployment given its robust performance across metrics. DeepSVDD can be considered for scenarios requiring balanced precision and recall.

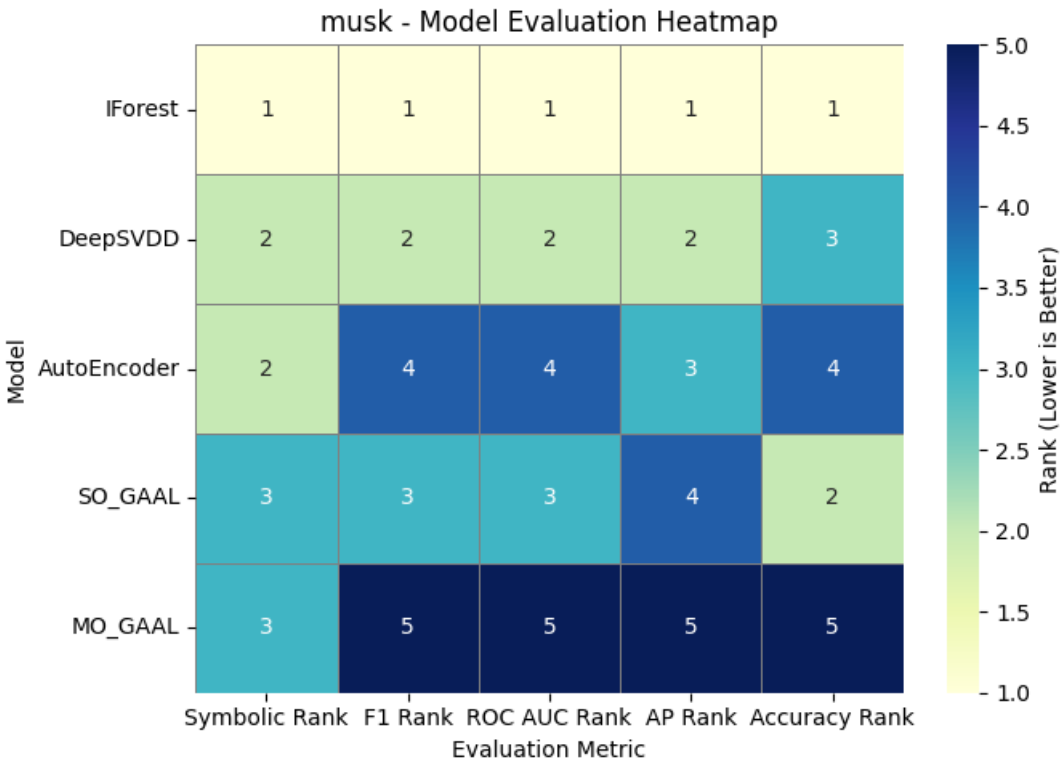
11. Annexure

Detailed metrics and additional insights are provided for further analysis and decision-making.

Model Evaluation Table

musk - 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	4.3	0.9998	0.9948	0.9314	0.4802	0.316	1.0	1	1	1	1
DeepSVDD	2	4.0	0.9047	0.5496	0.9131	0.3416	0.2248	0.7113	2	2	2	3
AutoEncoder	2	4.0	0.8553	0.4091	0.9033	0.2673	0.1759	0.5567	4	4	3	4
SO_GAAL	3	3.2	0.8662	0.172	0.9242	0.297	0.2103	0.5052	3	3	4	2
MO_GAAL	3	3.2	0.7683	0.0742	0.885	0.1244	0.082	0.2577	5	5	5	5

Model Evaluation Heatmap



Model-wise Multi-Metric Rankings

