# Model Evaluation Summary Table for cardio

Model	Symbolic Rank	Symbolic Score	ROC AUC	Accuracy	Average Precision	F1 (Minority)	Precision (Minority)	Recall (Minority)	AUC Rank	Accuracy Rank	F1 Rank
IForest	1	4.2	0.9205	0.9066	0.5816	0.5237	0.5137	0.5341	1	1	1
LUNAR	2	3.1	0.6368	0.8596	0.2302	0.2841	0.2787	0.2898	4	4	4
SO_GAAL	2	3.1	0.7794	0.8798	0.3572	0.3855	0.3791	0.392	3	3	3
MO_GAAL	2	3.1	0.7877	0.8957	0.4317	0.468	0.459	0.4773	2	2	2
LOF	3	2.0	0.5458	0.852	0.1543	0.1763	0.1895	0.1648	5	5	5

Model Evaluation Heatmap for cardio

cardio - Model Evaluation Heatmap
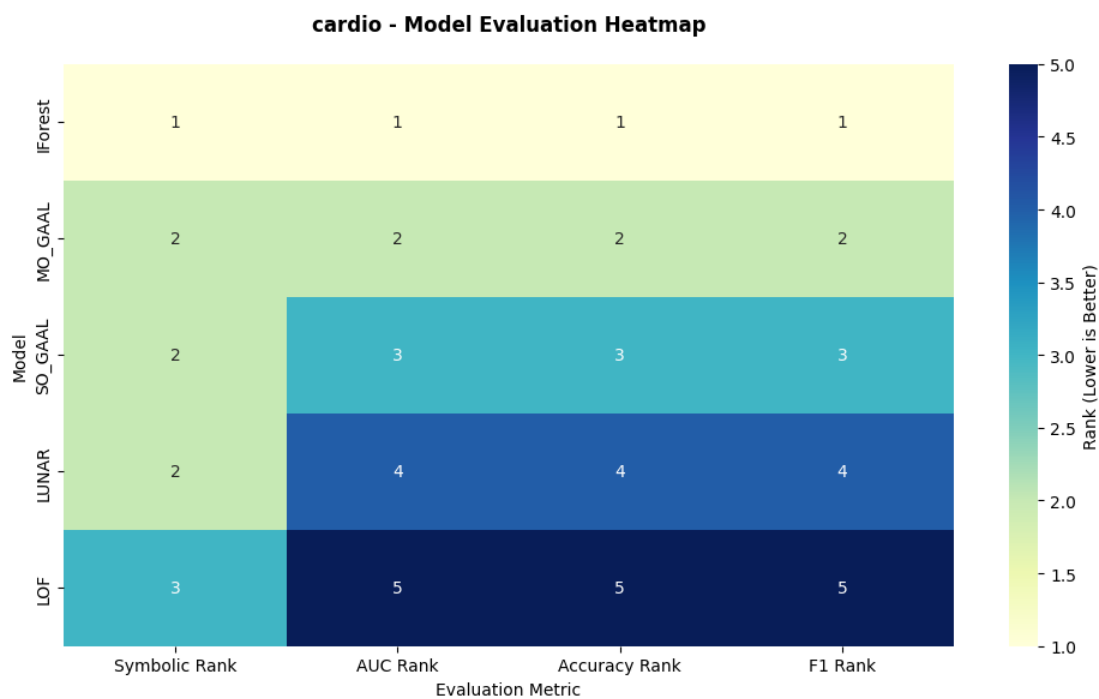
LLM Analysis and Recommendations for cardio

LLM Analysis & Model Recommendation for Cardio Dataset

Overview Summary for Cardio Dataset

The Cardio dataset is a medium-sized, structured dataset with 1831 samples and 21 features. It is clean with no missing values, and exhibits high skewness and kurtosis. The dataset is imbalanced, with an anomaly ratio of 0.096122.

Comparison of Symbolic vs. Actual Scores

The symbolic scoring system has successfully prioritized high-performing models. The IForest model, which was ranked first symbolically, also performed best in actual evaluations across all metrics (ROC AUC, Average Precision, Accuracy, F1 for Minority Class). Other models, such as LUNAR, SO$GAAL, and MO$GAAL, were symbolically ranked second and their actual performance also followed this order. The LOF model, ranked last symbolically, also performed worst in actual evaluations.

Model Suitability Summary Table for Cardio Dataset

The IForest model is the most suitable for this dataset, with the highest symbolic score of 4.2 and top actual performance across all metrics. It is well-suited for structured data, high dimensional data, high skewness, high kurtosis, and highly imbalanced data, which are all characteristics of the Cardio dataset.

Model Evaluation Heatmap for Cardio Dataset

The heatmap further confirms the superiority of the IForest model. It has the highest scores in all evaluation metrics, indicating its robust performance on the Cardio dataset.

Analysis of Summary Table and Heatmap

The summary table and heatmap clearly show that the IForest model outperforms the other models on the Cardio dataset. It has the highest symbolic score and actual performance metrics, making it the top recommended model.

LLM Recommendations and Justification for Model Selection

Given the characteristics of the Cardio dataset and the performance of the models, the IForest model is recommended for deployment. It aligns well with the dataset's properties and has demonstrated superior performance in both symbolic and actual evaluations.

Suggested Preprocessing Steps for Cardio Dataset

Given the high skewness and kurtosis of the dataset, it may benefit from transformations such as log or Box-Cox to reduce skewness and normalize the data. Feature scaling may also be beneficial, especially since the IForest model can be sensitive to the range of the data.

Hyperparameter Tuning Recommendations

For the IForest model, key hyperparameters to consider are the number of estimators and the contamination factor. The number of estimators can be increased to improve the model's performance, but this will also increase the computational cost. The contamination factor, which is an estimate of the proportion of outliers in the dataset, can be set close to the known anomaly ratio of 0.096122.

Final Professional Recommendation

In conclusion, the IForest model is recommended for deployment on the Cardio dataset. It has shown superior performance in both symbolic and actual evaluations, and aligns well with the dataset's properties. With appropriate preprocessing and hyperparameter tuning, it is expected to provide reliable anomaly detection.