Model Validation Report

# Model Performance Metrics

This section includes generic model performance metrics.

**Prediction Distribution**

Probability distribution from model prediction output.

|  |  |  |
| --- | --- | --- |
| Testing Dataset |  |  |
| Test | Model |  |

# Statistical Metrics

This section includes statistical metrics.

**PSI**

A metric to measure how much a variable has shifted in distribution between two samples over time. Please refer the link for more details: https://towardsdatascience.com/psi-and-csi-top-2-model-monitoring-metrics-924a2540bed8

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Testing Dataset |  | PSI |  | train\_count | train\_perc | test\_count | test\_perc | index\_value |
| Test | Model | 0.0998621 | (0.00, 0.10] | 246 | 0.034 | 14 | 0.017 | 0.012 |
|  |  |  | (0.10, 0.20] | 1994 | 0.273 | 161 | 0.193 | 0.028 |
|  |  |  | (0.20, 0.30] | 1702 | 0.233 | 213 | 0.255 | 0.002 |
|  |  |  | (0.30, 0.40] | 1021 | 0.14 | 138 | 0.165 | 0.004 |
|  |  |  | (0.40, 0.50] | 711 | 0.097 | 115 | 0.138 | 0.014 |
|  |  |  | (0.50, 0.60] | 465 | 0.064 | 70 | 0.084 | 0.005 |
|  |  |  | (0.60, 0.70] | 362 | 0.05 | 44 | 0.053 | 0.0 |
|  |  |  | (0.70, 0.80] | 199 | 0.027 | 41 | 0.049 | 0.013 |
|  |  |  | (0.80, 0.90] | 370 | 0.051 | 25 | 0.03 | 0.011 |
|  |  |  | (0.90, 1.00] | 223 | 0.031 | 13 | 0.016 | 0.01 |
|  | Benchmark Model | 0.01767149 | (0.00, 0.10] | 1587 | 0.221 | 171 | 0.208 | 0.001 |
|  |  |  | (0.10, 0.20] | 3439 | 0.478 | 398 | 0.485 | 0.0 |
|  |  |  | (0.20, 0.30] | 984 | 0.137 | 97 | 0.118 | 0.003 |
|  |  |  | (0.30, 0.40] | 147 | 0.02 | 19 | 0.023 | 0.0 |
|  |  |  | (0.40, 0.50] | 212 | 0.029 | 23 | 0.028 | 0.0 |
|  |  |  | (0.50, 0.60] | 194 | 0.027 | 40 | 0.049 | 0.013 |
|  |  |  | (0.60, 0.70] | 0 | 0.0 | 0 | 0.0 | 0.0 |
|  |  |  | (0.70, 0.80] | 309 | 0.043 | 36 | 0.044 | 0.0 |
|  |  |  | (0.80, 0.90] | 148 | 0.021 | 19 | 0.023 | 0.0 |
|  |  |  | (0.90, 1.00] | 175 | 0.024 | 18 | 0.022 | 0.0 |
|  |  |  |  |  |  |  |  |  |

# Transparency Metrics

This section includes LIME and SHAP interpretability under the framework of MAS's FEAT metrics.

Local interpretability refers to the ability to explain the behavior of a model on a specific instance or observation. It focuses on understanding why a model made a particular prediction or decision for a single data point. Local interpretability techniques provide explanations that are specific to a particular instance, and they typically focus on the features that the model used to make its decision.

Global interpretability, on the other hand, refers to the ability to explain the behavior of a model across its entire input space. It focuses on understanding how the model works in general, rather than on a specific instance. Global interpretability techniques aim to provide insights into the relationships between features and how they contribute to the overall behavior of the model.

**LIME Interpretability**

LIME is a method that helps to explain the predictions of any machine learning model by approximating the model locally with an interpretable model. LIME works by perturbing the input data and observing the changes in the output of the model. The perturbed data is then used to train a new, interpretable model that approximates the behavior of the original model in that local region. The resulting model can then be used to explain why the original model made a particular prediction.   
 The LIME chart shows the coefficients assigned to each feature by the interpretable model. The coefficients indicate the direction and magnitude of the impact of each feature on the model's output. Positive coefficients indicate that the feature has a positive impact on the Class1, while negative coefficients indicate a positive impact on Class 0.   
 Please refer to the link for more details: https://arxiv.org/abs/1602.04938)

**Local LIME Interpretability**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Testing Dataset | Model | | Benchmark Model | |
| Test |  | |  |  |