Swap Attack Impact Distribution: Malicious Logit Injection Analysis

Min: 0.031
Median: 0.039
Max: 0.060
Std: 0.007

0.053
0.045
0.039
0.036
0.032
0.031
0.031

Threshold Coverage (%)

Model Separation Value