AIShield
Powered by Bosch

Vulnerability Analysis

Tabular Classification

Attack Type
Evasion

Date
2024-08-05

Author
AIShield

Job Id
gAAAAABm...JdESi-3A==

Executive Summary:

Our analysis unveils substantial vulnerabilities, highlighting a heightened risk of attackers exploiting the model through Evasion with 82.0% efficacy. These vulnerabilities could lead to significant financial, legal, and reputational repercussions. Employing AIShield's sample attack vectors for adversarial hardening or implementing our Threat Informed Defense Engine for real-time protection are essential. Proactive measures will act as enablers of regulatory preparedness and minimize potential risks.

This assesment aligns with the following OWASP Machine Learning Security Top-10 Vulnerabilities(v0.3 Draft):

- ML01:2023 Input Manipulation Attack

The proposed measures align with these CVE/CWE Entries:

- ML01:2023 Input Manipulation Attack
- ML07:2023 Transfer Learning Attack
- ML09:2023 Output Integrity Attack

For more details about OWASP Vulnerabilities, please refer to Section 4.4 of the report.
For more details about CVE/CWE Entries, please refer to Section 4.5 of the report

1. Security:

1.1 Strength Vs Efficacy on Extracted Model:

Below table shows the change in epsilon vs attack efficacy

Epsilon	Attack Efficacy %
0.05	8.0
0.1	18.0
0.2	32.0
0.3	76.0
0.4	78.0
0.5	82.0

For more details on the metrics, refer to the Appendix Section of the report

1.2 Adversarial samples:

Refer to the explainability section to visualize the adversarial samples.

2. Performance:

2.1 Inference Time of models:

Inference time of a model refers to the amount of time it takes for the model to generate predictions on new input data. It is the time taken by the model to process input data and produce output. inference time is influenced by factors such as the model's complexity, the input data's size, and the computational resources available for inference. It is an important metric to consider when deploying machine learning models in production environments where fast and efficient processing is necessary.

In the context of this scenario, Original Model Inference Time in ms is 35.04.

More details on the hardware infrastructure, and inference time distributions will be populated in the upcoming reports.

2.2 Accuracy and F1 score of Models (Model):

Accuracy and F1 score on original data is necessary to understand the model performance even under imbalanced data distribution. The below table represents the performance comparison of two models - the Original model and the Extracted model - based on the number of samples, their relative accuracy, and the relative F1 score performed on original data as ground truth.

Model	Number of Samples	Model Accuracy	F1 Score
Model	100	98.0	0.98

2.3 Adversarial Accuracy on Original Model:

The below table shows the concentration of adversarial samples vs accuracy (adversarial accuracy)

Original samples %	Adversarial samples %	Model Accuracy %
100	0	98.0
80	20	57.25
60	40	33.33
40	60	18.52
20	80	7.84
0	100	0.0

3. Explainability Report:

Explainability report provides a clear and concise explanation of the factors that contribute to a model's behavior or decisions, helping stakeholders to better understand and trust the model's outputs. The below report typically includes information on the data used to train the model, the model's architecture and performance, and the techniques or tools used to extract insights from the model.

3.1 Contrastive Analysis:

Example 1, Attack Strength 20%, Original Class 0, Adversarial Class 1, Max Index-duration

Features	Original(0)	Adversarial(1)	Difference
age	0.3209876543209876	0.32879256021463943	0.007804905893651848
job	0.8181818181818182	0.8463562271279014	0.02817440894608314
marital	0.3333333333333333	0.3092540681229825	0.02407926521035081
education	0.6000000000000001	0.6681747968042719	0.06817479680427185
default	0.0	0.0	0.0
housing	1.0	0.9731832061695924	0.02681679383040758
loan	0.0	0.038691776485834675	0.038691776485834675
contact	1.0	0.9833476590371395	0.0166523409628605
duration	0.0180967873119154	0.1615830867527004	0.143486299440785
campaign	0.0	0.0	0.0
pdays	1.0	0.9902406401844728	0.009759359815527224
previous	0.0	0.0	0.0
poutcome	0.5	0.4393635917100754	0.060636408289924615
emp.var.rate	0.6875	0.6868599715551794	0.000640028444820584
cons.price.idx	0.3893219017926768	0.44980475668848424	0.06048285489580746
cons.conf.idx	0.3682008368200835	0.3525559369413311	0.015644899878752394
euribor3m	0.7977782815688053	0.8470640791030958	0.04928579753429052
nr.employed	0.8778827977315693	0.8508504170732628	0.027032380658306487

Example 2, Attack Strength 30%, Original Class 0, Adversarial Class 1, Max Index-duration

Features	Original(0)	Adversarial(1)	Difference
age	0.3209876543209876	0.3635886292062518	0.04260097488526421
job	0.8181818181818182	0.8376857211678947	0.019503902986076516
marital	0.3333333333333333	0.2914776185620941	0.04185571477123923
education	0.6000000000000001	0.6992901516911842	0.09929015169118416
default	0.0	0.005034406850456829	0.005034406850456829
housing	1.0	0.9629158053419311	0.03708419465806889
loan	0.0	0.07239916546341245	0.07239916546341245

contact	1.0	0.9630070545674653	0.03699294543253473
duration	0.0180967873119154	0.21637706671254422	0.1982802794006288
campaign	0.0	0.0	0.0
pdays	1.0	0.9778231370706505	0.022176862929349506
previous	0.0	0.0	0.0
poutcome	0.5	0.40564264659330934	0.09435735340669066
emp.var.rate	0.6875	0.6743881865803134	0.013111813419686591
cons.price.idx	0.3893219017926768	0.4730631484159863	0.0837412466233095
cons.conf.idx	0.3682008368200835	0.32497869774758165	0.04322213907250183
euribor3m	0.7977782815688053	0.8889207846445416	0.09114250307573635
nr.employed	0.8778827977315693	0.8319164320759709	0.04596636565559842

Example 3, Attack Strength 10%, Original Class 0, Adversarial Class 1, Max Index-nr

Features	Original(0)	Adversarial(1)	Difference
age	0.654320987654321	0.6294170156981266	0.024903971956194426
job	0.4545454545454546	0.42741373782557446	0.027131716719880128
marital	0.3333333333333333	0.33723941188032636	0.0039060785469930415
education	0.0	0.002251034686907026	0.002251034686907026
default	0.5	0.4747868382506451	0.025213161749354895
housing	1.0	1.0	0.0
loan	0.0	0.0045947021695546715	0.0045947021695546715
contact	0.0	0.014854885841477716	0.014854885841477716
duration	0.0650671004473363	0.11443395824298967	0.04936685779565338
campaign	0.0	0.0	0.0
pdays	1.0	0.993661677159225	0.006338322840774957
previous	0.0	0.01543412009670977	0.01543412009670977
poutcome	0.5	0.4879626628088382	0.01203733719116179
emp.var.rate	0.1041666666666667	0.10547258926179336	0.0013059225951266573
cons.price.idx	0.0	0.035294015128206825	0.035294015128206825
cons.conf.idx	0.8117154811715479	0.7974086063380278	0.014306874833520111
euribor3m	0.0564497846293357	0.02686741658651232	0.029582368042823378
nr.employed	0.4257088846880883	0.37513426828474433	0.050574616403343986

Example 4, Attack Strength 50%, Original Class 0, Adversarial Class 1, Max Index-duration

Features	Original(0)	Adversarial(1)	Difference
age	0.3209876543209876	0.3930351265553434	0.07204747223435581

job	0.0909090909090909	0.23294544714751528	0.14203635623842437
marital	0.3333333333333333	0.25233213694626805	0.08100119638706527
education	0.0	0.05994566762828802	0.05994566762828802
default	0.0	0.04223535468360085	0.04223535468360085
housing	1.0	0.915761472901434	0.08423852709856605
loan	0.0	0.07547906653142057	0.07547906653142057
contact	1.0	0.9296996395409536	0.07030036045904642
duration	0.2189914599430663	0.41485983711415086	0.19586837717108457
campaign	0.0	0.0	0.0
pdays	1.0	0.9565431504584976	0.04345684954150242
previous	0.0	0.0	0.0
poutcome	0.5	0.5709810828371795	0.07098108283717952
emp.var.rate	0.9375	0.9966183386681592	0.05911833866815919
cons.price.idx	0.6987529228371017	0.7057313290248778	0.0069784061877761205
cons.conf.idx	0.6025104602510458	0.5844226427408082	0.01808781751023758
euribor3m	0.9573792790750396	0.9995737927907504	0.04219451371571081
nr.employed	0.8597353497164448	0.7615086249237968	0.09822672479264805

4. Appendix:

This section offers supplementary data on our methodology, the architectures of both the Original and Extracted Models, and other pertinent data like the Confusion matrix and Classification reports.

4.1 Model Architecture:

Below image represent the original model architecture.

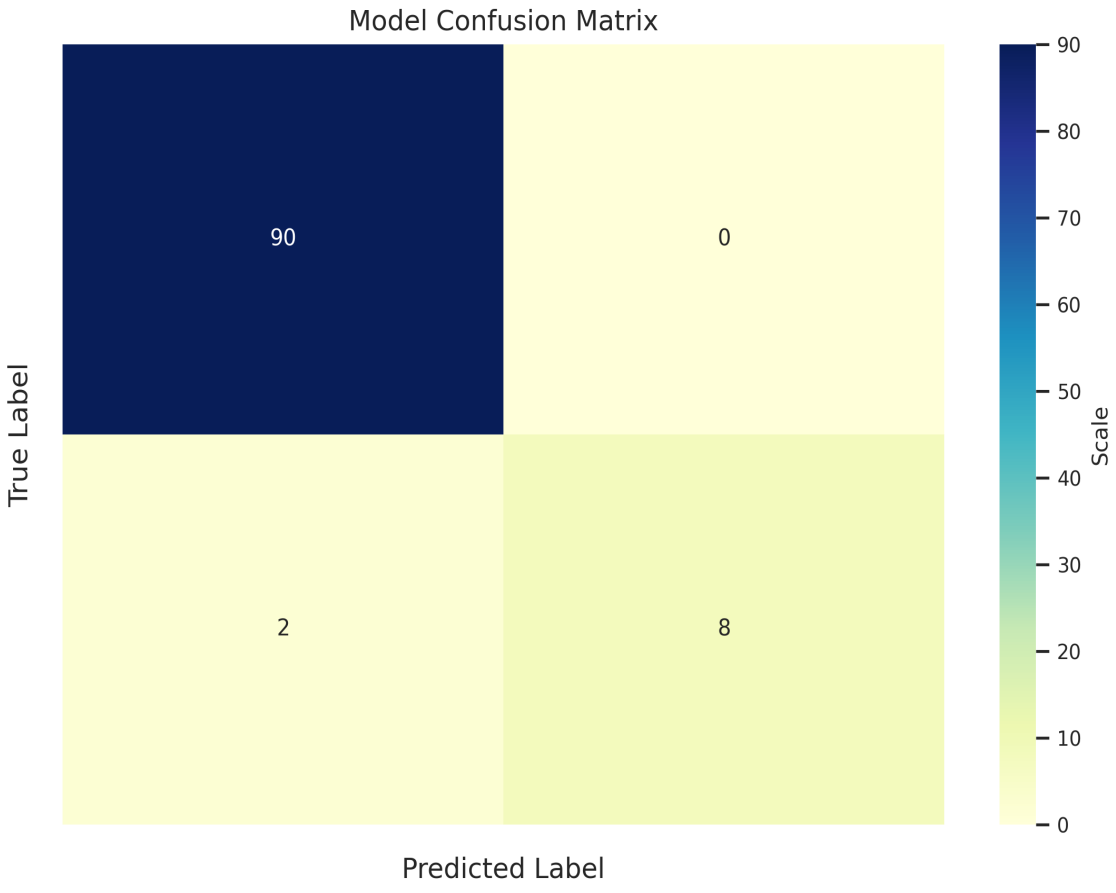
Model Architecture

Model	Parameters
Model Name	XGBClassifier
n_estimators	100
missing	nan
n_classes_	2

4.2 Confusion Matrix:

Confusion matrix provides a comprehensive overview of the classification model's performance. The rows of the matrix represent the actual class labels, while the columns represent the predicted class labels. Each cell of the matrix shows the of samples that belong to a particular combination of actual and predicted class labels.

Below images represent the Confusion Matrix.



4.3 Classification Report:

Classification report helps to evaluate the performance of your model for each class and make decisions about which classes need further improvement. It can also be used to compare the performance of different models or different versions of the same model.

Below images represent the Classification Report.

Model Classification Report

	precision	recall	f1-score	support
0	0.98	1.00	0.99	90
1	1.00	0.80	0.89	10
accuracy			0.98	100
macro avg	0.99	0.90	0.94	100
weighted avg	0.98	0.98	0.98	100

4.4 OWASP Vulnerabilities(v0.3 Draft):

ML01:2023 Input Manipulation Attack
This weakness directly relates to adversarial attacks where an automated mechanism, such as machine learning, doesn't properly detect or handle inputs that have been modified to mislead the model.

ML07:2023 Transfer Learning Attack
If the machine learning model does not ensure that structured messages or data are well-formed, it can be susceptible to attacks like the Transfer Learning Attack.

ML09:2023 Output Integrity Attack
The description of the adversarial attack mentions the importance of input validation to detect and prevent adversarial attacks.

For an expanded understanding of the OWASP Top - 10 ML Vulnerabilities, please visit https://owasp.org/www-project-machine-learning-security-top-10/

4.5 CVE(Common Vulnerabilities and Exposures)/CWE(Common Weakness Enumeration):

CWE-20: Improper Input Validation
A weakness where the product receives input or data but does not validate or incorrectly validates that the input has the properties required to process the data safely and correctly. This can lead to various security issues, such as altered control flow, arbitrary control of a resource, or arbitrary code execution.

CWE-707: Improper Neutralization
The product does not ensure or incorrectly ensures that structured messages or data are well-formed and that certain security properties are met before being read from an upstream component or sent to a downstream component.

CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
This weakness occurs when a product uses an automated mechanism, to recognize complex data inputs, but it does not properly detect or handle inputs that have been modified or constructed such that it causes the mechanism to detect an incorrect concept.

CWE-1288: Improper Validation of Consistency within Input
The product receives a complex input with multiple elements or fields that must be consistent with each other, but it does not validate or incorrectly validates that the input is actually consistent.This can allow attackers to trigger unexpected errors, cause incorrect actions to take place, or exploit latent vulnerabilities.

For comprehensive data related to CVE/CWE list,please visit https://cwe.mitre.org/

4.6 Utilized Attack Methods in Tabular Analysis:

In the comprehensive vulnerability analysis of the Tabular Classification model, a variety of sophisticated evasion attack methods are employed.

AIShield integrates its proprietary evasion attack methodologies with an array of additional evasion attack techniques, ensuring a thorough and multi-dimensional assessment. The following section outlines the specific evasion attack techniques utilized in the vulnerability assessment by AIShield attack engine

List of Blackbox Attacks:

AIShield Propreitary methods; Auto Projected Gradient Descent (Auto-PGD); Basic Iterative Method (BIM); Fast Gradient Method; Feature Adversaries; Optimisation attack (L-BFGS); Feature importance informed attack; transferability attacks; Projected Gradient Descent (PGD); Universal Perturbation; and Hybrid(combination of above)

List of Greybox Attacks:

AIShield Propreitary methods; Auto Projected Gradient Descent (Auto-PGD); Basic Iterative Method (BIM); Fast Gradient Method; Feature Adversaries; Optimisation attack (L-BFGS); Projected Gradient Descent (PGD); transferability attacks; Universal Perturbation; and Hybrid(combination of above)

Report can be verified for its integrity with SHA-256 checksum:
786b19a18f2a286cc3f6b5e7dc0aea56b68bbbe9a7ab12ab76f679347b2655ea