AIShield
Powered by Bosch

Vulnerability Analysis

Timeseries Forecasting

Attack Type
Extraction

Date
2024-02-21

Author
AIShield

Job Id
gAAAAABl...76E-uIoA==

Executive Summary:

Attack Type: blackbox	Attack Queries: 50000	Stolen Model Relative RMSE: 0.05	Alert:Critical	Defense Recommended: Yes

Our analysis unveils substantial vulnerabilities, highlighting a heightened risk of attackers exploiting the model through blackbox Extraction with 0.0439036716 RMSE. 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):

- ML03:2023 Model Inversion Attack
- ML05:2023 Model Stealing

The proposed measures align with these CVE/CWE Entries:

- CWE-20: Improper Input Validation
- CWE-212: Improper Removal of Sensitive Information Before Storage or Transfer
- CWE-226: Sensitive Information in Resource Not Removed Before Reuse
- CWE-501: Trust Boundary Violation
- CWE-707: Improper Neutralization
- CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
- CWE-1357: Reliance on Insufficiently Trustworthy Component

For more details about OWASP Vulnerabilities, please refer to Section 3.2 of the report.
For more details about CVE/CWE Entries, please refer to Section 3.3 of the report

1. Security:

1.1 Relative RMSE:

Relative model RMSE compares the root mean squared error between the predictions of two models, measuring the similarity in their prediction errors. It provides a quantitative measure of agreement between the models' outputs, with a lower relative RMSE indicating better agreement and similarity.

In the context of this scenario, Relative RMSE is 0.0439036716

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 568.97.

In the context of this scenario, Extracted Model Inference Time in ms is 1.25.

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

2.2 RMSE Score of Models (Original Model, Extracted Model):

RMSE 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, and their relative RMSE score performed on original data as ground truth.

Model	Number of Samples	Model RMSE
Original Model	31439	0.01
Extracted Model	31439	0.05

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

3. 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.

3.1 Model Architecture:

Below image represent the model architecture.

Original Model architecture can not be displayed due to the access restriction.

Extracted Model Architecture

Model	Parameters
Model Name	LGBMRegressor
num_leaves	31
max_depth	-1
learning_rate	0.1
n_estimators	100
subsample_for_bin	200000
min_split_gain	0.0
min_child_weight	0.001
min_child_samples	20
subsample	1.0
subsample_freq	0
colsample_bytree	1.0
reg_alpha	0.0
reg_lambda	0.0
_best_iteration	0
_n_features	100
_n_features_in	100
_n_classes	-1

3.2 OWASP Vulnerabilities(v0.3 Draft):

ML03:2023 - Model Inversion Attack
This weakness occurs when an automated mechanism doesn't properly detect or handle inputs that have been modified or constructed such that it causes the mechanism to detect an incorrect concept. This is related to the concept of model inversion where the attacker can reverse-engineer the model's predictions.

ML05:2023 - Model Stealing
Model stealing can occur as a result of inadequate detection mechanisms, especially in machine learning systems.

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

3.3 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-212: Improper Removal of Sensitive Information Before Storage or Transfer
This weakness occurs when a product stores, transfers, or shares a resource that contains sensitive information, but it does not properly remove that information before the product makes the resource available to unauthorized actors.

CWE-226: Sensitive Information in Resource Not Removed Before Reuse
This weakness occurs when a product releases a resource such as memory or a file so that it can be made available for reuse, but it does not clear the information contained in the resource before the product performs a critical state transition or makes the resource available for reuse by other entities.

CWE-501: Trust Boundary Violation
A base level weakness that occurs when a product mixes trusted and untrusted data in the same data structure or structured message. A trust boundary violation occurs when a program blurs the line between what is trusted and what is untrusted. By combining trusted and untrusted data in the same data structure, it becomes easier for programmers to mistakenly trust unvalidated data.

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-1357: Reliance on Insufficiently Trustworthy Component
This weakness occurs when a product is built from multiple separate components, but it uses a component that is not sufficiently trusted to meet expectations for security, reliability, updateability, and maintainability.

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

3.4 Utilized Attack Methods in Regression Analysis:

In the comprehensive vulnerability analysis of the Regression model, a variety of sophisticated extraction attack methods are employed.

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

List of Blackbox Attacks:

AIShield Propreitary methods; Active Learning-based Extraction; Distillation-Based Attacks; Knock-off Nets; Copycat CNN; functionally-equivalent model; Active Learning-based Extraction; Extraction through Synthetic Attack vectors; Neural Architecture Search (NAS) for Model Approximation; Data-Free Model Extraction; Model Stealing via Prediction APIs; and Hybrid (combination of above)

List of Greybox Attacks:

AIShield Propreitary methods; Augmentation-based attacks; Active Learning-based Extraction; Distillation-Based Attacks; functionally-equivalent model; Active Learning-based Extraction; Extraction through Synthetic Attack vectors; Neural Architecture Search (NAS) for Model Approximation; Model Stealing via Prediction APIs; and Hybrid (combination of above)

Report can be verified for its integrity with SHA-256 checksum:
9d7f4967c0e23cb0be1c083cd2ce1fc505d73e017d828ae5a8f8497ffa7a00e6