Machine Learning Security

OWASP Top Ten Machine Learning Risks

Made Aug 19, 2023

Machine Learning Risks

Top 10 Machine Learning Security Risks

- **ML01:2023 Adversarial Attack**
- **ML02:2023 Data Poisoning Attack**
- **ML03:2023 Model Inversion Attack**
- **ML04:2023 Membership Inference Attack**
- **ML05:2023 Model Stealing**
- **ML06:2023 Corrupted Packages**
- **ML07:2023 Transfer Learning Attack**
- **ML08:2023 Model Skewing**
- **ML09:2023 Output Integrity Attack**
- **ML10:2023 Neural Net Reprogramming**

- https://owasp.org/www-project-machine-learning-security-top-10/

- **ML01:2023 Adversarial Attack**
  - An attacker deliberately alters input data to mislead the model
  - This attack is also called **evasion**
  - Example: a model is trained to tell cat images from dog images. An attacker modifies a cat image so it is misclassified as a dog.
- **ML02:2023 Data Poisoning Attack**
  - An attacker manipulates the training data to cause the model to behave in an undesirable way
- **ML03:2023 Model Inversion Attack**
  - An attacker reverse-engineers the model to extract information from it
  - Example: a model is trained to recognize faces. An attacker inputs images of individuals into the model and and recovers the personal information of the individuals from the model's predictions, such as their name, address, or social security number.

- **ML04:2023 Membership Inference Attack**
  - An attacker manipulates the model's training data in order to cause it to behave in a way that exposes sensitive information
  - Example: A malicious attacker trains a machine learning model on a dataset of financial records and uses it to query whether or not a particular individual's record was included in the training data.
- **ML05:2023 Model Stealing**
  - An attacker gains access to the model's parameters
  - Example: Stealing a machine learning model from a competitor
- **ML06:2023 Corrupted Packages**
  - An attacker modifies or replaces a machine learning library or model that is used by a system

- **ML07:2023 Transfer Learning Attack**
  - An attacker trains a model on one task and then fine-tunes it on another task to cause it to behave in an undesirable way
  - Example: An attacker trains a machine learning model on a malicious dataset that contains manipulated images of faces. The attacker then transfers the model's knowledge to a target face recognition system. As a result, the face recognition system starts making incorrect predictions, allowing the attacker to bypass the security and gain access to sensitive information.
- **ML08:2023 Model Skewing**
  - An attacker manipulates the distribution of the training data to cause the model to behave in an undesirable way.
  - Example: The attacker provides fake feedback data to a loan-approving machine learning system. As a result, the model's predictions are skewed, and the attacker's chances of getting a loan approved are significantly increased.

- **ML09:2023 Output Integrity Attack**
  - An attacker aims to modify or manipulate the output of a machine learning model in order to change its behavior or cause harm to the system it is used in.
  - Example: An attacker has gained access to the output of a machine learning model that is being used to diagnose diseases in a hospital. The attacker modifies the output of the model, making it provide incorrect diagnoses for patients.
- **ML10:2023 Neural Net Reprogramming**
  - An attacker manipulates the model's parameters to cause it to behave in an undesirable way.
  - Example: A bank is using a machine learning model to identify handwritten characters on cheques. An attacker manipulates the parameters of the model by altering the images in the training dataset or directly modifying the parameters in the model. This can result in the model misidentifying characters, leading to incorrect amounts being processed.

ML-OWASP-ML