Los expertos en ciberseguridad han identificado varias vulnerabilidades que afectan herramientas y plataformas de aprendizaje automático (ML) de código abierto, como MLflow, H2O, PyTorch y MLeap, que podrían ser explotadas para ejecutar código malicioso.

Estas fallas, descubiertas por la empresa JFrog, forman parte de un grupo más amplio de 22 problemas de seguridad que la compañía había revelado previamente el mes pasado.

A diferencia de las vulnerabilidades iniciales, que estaban relacionadas con problemas del lado del servidor, las más recientes afectan a los clientes de ML y se encuentran en bibliotecas diseñadas para manejar formatos de modelos considerados seguros, como Safetensors.

> *«El compromiso de un cliente de ML dentro de una organización puede permitir a los atacantes realizar un movimiento lateral significativo en su infraestructura. Un cliente de ML probablemente tendrá acceso a servicios clave, como registros de modelos de ML o pipelines de MLOps»,* indicó la compañía.

Esto podría dar lugar a la exposición de datos sensibles, como credenciales de registro de modelos, lo que facilitaría a los atacantes insertar puertas traseras en los modelos almacenados o ejecutar código malintencionado.

A continuación, se presenta un resumen de las vulnerabilidades detectadas:

- CVE-2024-27132 (puntuación CVSS: 7.2): Un fallo de sanitización insuficiente en MLflow que permite un ataque de cross-site scripting (XSS) al ejecutar una receta no confiable en Jupyter Notebook, resultando en la ejecución remota de código (RCE) en el cliente.
- CVE-2024-6960 (puntuación CVSS: 7.5): Un problema de deserialización insegura en H2O que, al importar un modelo no confiable, podría permitir la ejecución remota de código.
- Una vulnerabilidad de recorrido de directorios en la función TorchScript de PyTorch

que puede provocar la sobrescritura arbitraria de archivos, lo que podría usarse para realizar un ataque de denegación de servicio (DoS) o ejecutar código (sin identificador CVE asignado).
- CVE-2023-5245 (puntuación CVSS: 7.5): Un problema de recorrido de directorios en MLeap al cargar un modelo comprimido, que permite una vulnerabilidad conocida como Zip Slip, con capacidad de sobrescribir archivos arbitrariamente y ejecutar código.

JFrog advirtió que los modelos de aprendizaje automático no deben cargarse sin precaución, incluso si utilizan formatos considerados seguros, como Safetensors, ya que podrían ser utilizados para ejecutar código arbitrario.

> *«Las herramientas de inteligencia artificial y aprendizaje automático tienen un gran potencial para la innovación, pero también representan un riesgo si son aprovechadas por atacantes para causar daños importantes en una organización»,* afirmó Shachar Menashe, vicepresidente de investigación en seguridad de JFrog.
>
> *«Para mitigar estas amenazas, es fundamental conocer los modelos que se están utilizando y evitar cargar aquellos de procedencia no confiable, incluso si provienen de un repositorio considerado 'seguro'. Hacerlo podría permitir la ejecución remota de código, con consecuencias graves para la organización».*