Los investigadores de ciberseguridad han identificado una vulnerabilidad crítica en un proveedor de inteligencia artificial (IA) como servicio, Replicate, que podría haber permitido a actores maliciosos acceder a modelos de IA propietarios y datos sensibles.

> *«Explotar esta vulnerabilidad habría permitido el acceso no autorizado a los comandos y resultados de IA de todos los clientes de la plataforma de Replicate»*, declaró la empresa de seguridad en la nube Wiz en un informe publicado esta semana.

El problema surge porque los modelos de IA generalmente están empaquetados en formatos que permiten la ejecución de código arbitrario, lo que un atacante podría utilizar para realizar ataques cruzados entre inquilinos mediante un modelo malicioso.

Replicate utiliza una herramienta de código abierto llamada Cog para contenerizar y empaquetar modelos de aprendizaje automático que luego pueden desplegarse en un entorno autohospedado o en Replicate.

Wiz informó que creó un contenedor Cog malicioso y lo subió a Replicate, empleándolo para lograr la ejecución remota de código en la infraestructura del servicio con privilegios elevados.

> *«Sospechamos que esta técnica de ejecución de código es un patrón común, donde las empresas y organizaciones ejecutan modelos de IA de fuentes no confiables, a pesar de que estos modelos pueden ser potencialmente maliciosos»,* dijeron los investigadores de seguridad Shir Tamari y Sagi Tzadik.

La técnica de ataque desarrollada por la empresa aprovechó una conexión TCP ya establecida, asociada con una instancia de servidor Redis dentro del clúster de Kubernetes alojado en la plataforma Google Cloud, para inyectar comandos arbitrarios.

Además, dado que el servidor Redis centralizado se utiliza como una cola para gestionar múltiples solicitudes de clientes y sus respuestas, los investigadores descubrieron que podría ser abusado para facilitar ataques cruzados entre inquilinos al manipular el proceso para insertar tareas maliciosas que podrían afectar los resultados de los modelos de otros clientes.

Estas manipulaciones maliciosas no solo amenazan la integridad de los modelos de IA, sino que también representan riesgos significativos para la precisión y confiabilidad de los resultados impulsados por la IA.

> *«Un atacante podría haber consultado los modelos de IA privados de los clientes, exponiendo potencialmente conocimientos propietarios o datos sensibles involucrados en el proceso de entrenamiento del modelo. Además, interceptar los comandos podría haber expuesto datos sensibles, incluida información de identificación personal (PII)»,* dijeron los investigadores.

La vulnerabilidad, que fue revelada de manera responsable en enero de 2024, ha sido corregida por Replicate. No hay evidencia de que la vulnerabilidad haya sido explotada en la práctica para comprometer los datos de los clientes.

La divulgación se produce poco más de un mes después de que Wiz detallara riesgos ya solucionados en plataformas como Hugging Face, que podrían permitir a actores maliciosos escalar privilegios, obtener acceso cruzado entre inquilinos a los modelos de otros clientes e incluso tomar el control de las tuberías de integración y despliegue continuo (CI/CD).

> *«Los modelos maliciosos representan un riesgo importante para los sistemas de IA, especialmente para los proveedores de IA como servicio, porque los atacantes pueden aprovechar estos modelos para realizar ataques cruzados entre inquilinos»,* concluyeron los investigadores.

> *«El impacto potencial es devastador, ya que los atacantes podrían acceder a millones de modelos de IA y aplicaciones privadas almacenadas dentro de los proveedores de IA como servicio».*