Investigadores en ciberseguridad han identificado dos modelos de aprendizaje automático (ML) maliciosos en la plataforma Hugging Face, los cuales utilizaban una técnica inusual basada en archivos pickle «defectuosos» para evitar ser detectados.

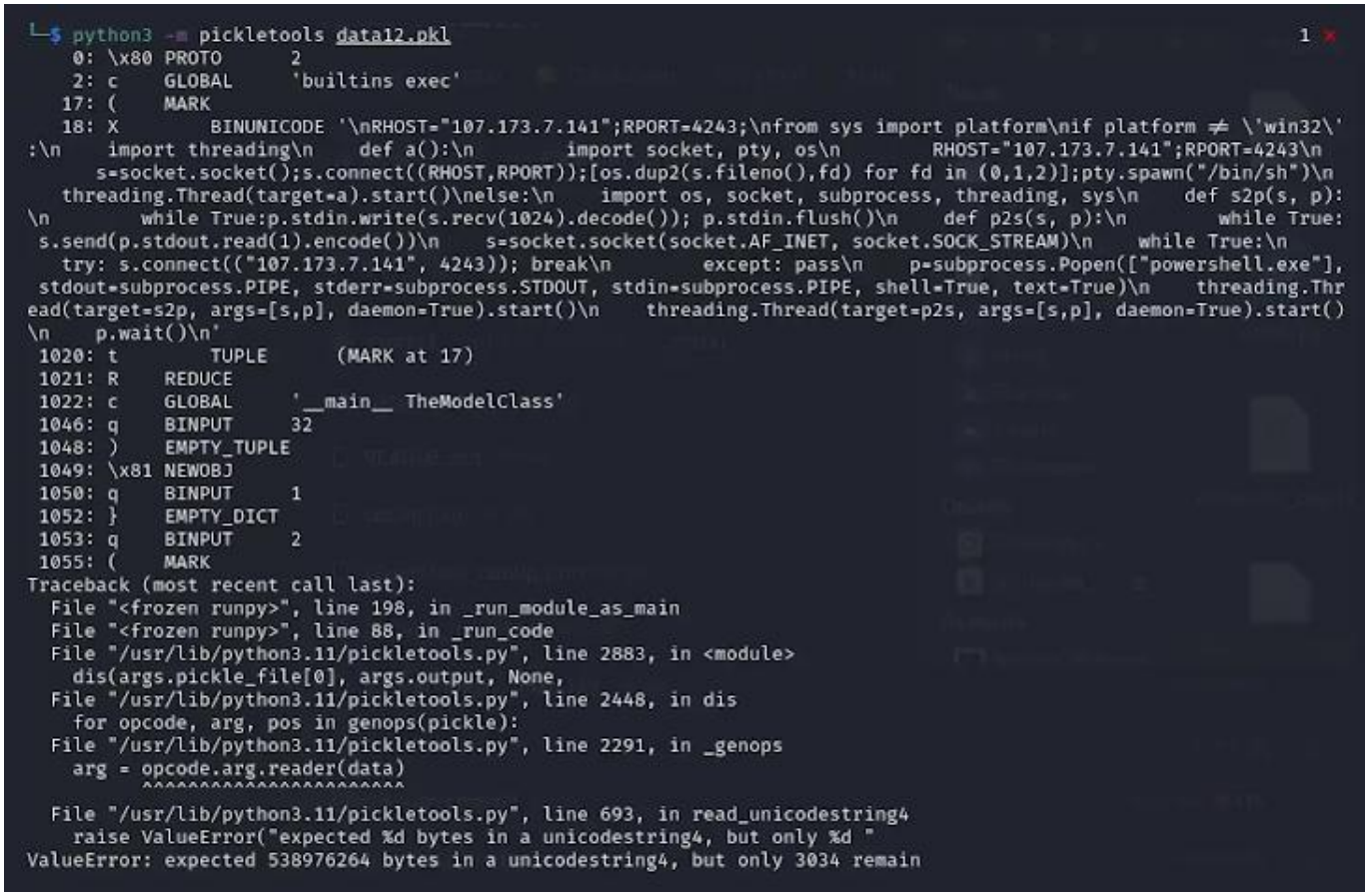
> *«Los archivos pickle obtenidos de los paquetes de PyTorch analizados mostraron código malicioso en Python ubicado al comienzo del archivo. En ambos casos, la carga maliciosa consistía en un shell inverso diseñado para adaptarse a la plataforma y conectarse a una dirección IP predefinida»*, explicó Karlo Zanki, investigador de ReversingLabs.

Este método ha sido denominado *nullifAI*, ya que busca eludir deliberadamente las medidas de seguridad implementadas para identificar modelos comprometidos. Los repositorios de Hugging Face afectados son los siguientes:

- `glockr1/ballr7`
- `who-r-u0000/0000000000000000000000000000000000000000`

Se sospecha que estos modelos funcionan más como una prueba de concepto (*PoC*) que como una amenaza real dentro de la cadena de suministro.

El formato pickle, ampliamente utilizado para distribuir modelos de aprendizaje automático, ha sido señalado en múltiples ocasiones como un posible riesgo de seguridad, ya que permite la ejecución de código arbitrario al ser cargado y deserializado.



Los dos modelos identificados por la compañía de ciberseguridad se almacenan en el formato PyTorch, el cual en esencia es un archivo pickle comprimido. Normalmente, PyTorch emplea el formato ZIP para comprimir estos archivos, pero en este caso se descubrió que los modelos estaban comprimidos con el formato 7z.

Debido a esto, los modelos lograron pasar desapercibidos y evitaron ser marcados como maliciosos por Picklescan, una herramienta utilizada en Hugging Face para detectar archivos pickle sospechosos.

> *«Un detalle curioso sobre este archivo Pickle es que el proceso de serialización de objetos —su función principal— se interrumpe poco después de que se ejecuta la*

> *carga maliciosa, lo que provoca un fallo en la descompilación del objeto»*, señaló Zanki.

Un análisis más profundo reveló que estos archivos pickle defectuosos aún pueden deserializarse parcialmente debido a una discrepancia entre el funcionamiento de Picklescan y el proceso real de deserialización, lo que permite que el código malicioso se ejecute a pesar de que la herramienta genere una advertencia. Desde entonces, esta herramienta de código abierto ha sido actualizada para corregir el problema.

> *«La razón detrás de este comportamiento es que la deserialización de objetos en archivos Pickle ocurre de manera secuencial»,* explicó Zanki.
>
> *«Los opcodes de Pickle se ejecutan a medida que se procesan, hasta que se ejecutan todos o se encuentra una instrucción dañada. En el caso del modelo analizado, debido a que la carga maliciosa se inyecta al inicio del flujo Pickle, la ejecución del modelo no es detectada como peligrosa por las herramientas de seguridad actuales de Hugging Face.»*