GLOBAL FINANCIAL SERVICES COMPANY USES HIDDENLAYER TO MINIMIZE CUSTOMER EXPERIENCE ISSUES WHILE COMBATING FRAUD

CASE STUDY

HIDDENLAYER
PROTECT YOUR ADVANTAGE

COMPANY OVERVIEW

A financial services company engaged the HiddenLayer Professional Services team to conduct a red team evaluation of machine learning models used to detect and intercept fraud during financial transactions. The primary purpose of the assessment was to identify weaknesses in their AI model classifications that could be exploited by adversaries to conduct fraudulent activities on the platform without triggering detection, resulting in millions of dollars of potential losses annually.

CHALLENGES

Hitting the Target: Ensuring an Optimal Customer Experience and Lowering Losses Amidst Rising Fraud Risks

With over 50 million users and facilitating more than 5 billion transactions annually, our customer grappled with the ongoing challenge of minimizing customer experience issues while simultaneously combating fraud. The delicate balance required cutting-edge AI and ML models to detect and intercept fraudulent transactions effectively. With these models at the core of their transaction operations, their commitment to stellar customer experiences and the need for advanced security for AI led them to engage with HiddenLayer for a comprehensive red teaming initiative.

DISCOVERY AND SELECTION OF HIDDENLAYER

The customer was referred to HiddenLayer by an existing customer who recognized our deep domain expertise in cyber and data science, along with our experience in automated adversarial attack tools. Additionally, HiddenLayer's flexible pricing model aligned with the customer needs, making HiddenLayer the clear choice for their red teaming endeavor.

Key Selection Criteria

- **Deep Expertise**: HiddenLayer's proficiency across cyber and data science modalities stood out.
- **Adversarial Attack Experience**: Their experience in dealing with automated adversarial attack tools was a crucial factor.
- **Flexible Pricing**: HiddenLayer's pricing model offered the flexibility we sought for our unique requirements.

Objectives of Red Teaming

- **Identify Vulnerable Features**: Pinpoint features within the models susceptible to an attacker's influence, with a potentially substantial impact on classification outcomes trending towards legitimacy.
- **Create Adversarial Examples**: Develop adversarial examples by modifying the least amount of features in inputs classified as fraudulent, transitioning the classification from fraudulent to legitimate.
- **Improve Model Classification**: Identify areas for improvement within the target models to enhance the accuracy of classifying fraudulent activities.

HIDDENLAYER'S MITIGATION

Alongside existing security controls, the introduction of inference-time monitoring for model inputs and outputs to detect targeted attacks against the models may help to flag and block suspected adversarial abuse.

HiddenLayer's AI Security (AISec) Platform which includes AI Detection and Response for Gen AI (AIDR) provides real-time, scalable, and unobtrusive inference-time monitoring for all model types. AIDR can be used to audit all existing models for adversarial abuse and ongoing prevention of abuse. AIDR does not require access to the customer's data or models as all detections are performed using vectorized inputs and outputs.

AIDR provides protection against common adversarial techniques including model extraction/theft, tampering, data poisoning/model injection, and inference.

IMPACT OF HIDDENLAYER

The core purpose of HiddenLayer's red teaming assessment was to uncover weaknesses in model classification that adversaries could exploit for fraudulent activities without triggering detection. Now armed with a prioritized list of identified exploits, our client can channel their invaluable resources, involving data science and cyber teams, towards mitigation and remediation efforts with maximum impact. The result is an enhanced security posture for the entire platform without introducing additional friction for internal or external customers.

HiddenLayer's product has proven instrumental in fortifying our defenses, allowing us to address vulnerabilities effectively and elevate our overall security stance while maintaining a seamless experience for our users.