Lehre & Studium

SCHUTZ VOR DNS-BASIERTER DATENEXFILTRATION MITHILFE VON MACHINE LEARNING

NIKLAS WILHELM (MASTERSTUDIUM INFORMATIK)

Betreuer: Prof. Dr. Reiner Hüttl, Prof. Dr. Markus Breunig

Motivation und Problemstellung

Das Domain Name System (DNS) spielt eine unverzichtbare Rolle im Internet: Es übersetzt menschenlesbare Domainnamen wie microsoft.com in IP-Adressen, damit Datenpakete ihr Ziel finden. Da das DNS für den laufenden Betrieb nahezu aller Netzwerke essenziell ist, bleibt der zugehörige Port 53 auch dann offen, wenn andere Verbindungen längst durch eine Firewall blockiert werden. Genau das machen sich Angreifer bei der sogenannten DNS-basierten Datenexfiltration zunutze. Sie verstecken gestohlene Daten in den Subdomains von DNS-Anfragen und schleusen sie so unbemerkt aus einem Unternehmensnetzwerk.

Laut einem Bericht von Palo Alto Networks nutzen 85 % aller Malware-Varianten das DNS für ihre Angriffe. Ein bekanntes Beispiel ist der Supply-Chain-Angriff auf SolarWinds im Jahr 2020, bei dem der Trojaner Sunburst über DNS-Anfragen Daten aus rund 18.000 betroffenen Unternehmen exfiltrierte. Der Versicherungsschaden wurde auf geschätzte 90 Millionen US-Dollar beziffert.

Wie eine Studie von IBM zeigt, lassen sich die Kosten eines Datenlecks durch den Einsatz automatisierter Erkennungssoftware im Schnitt um 33 % senken. Regelbasierte Erkennungssoftware wie Snort stößt dabei jedoch an ihre Grenzen: Sobald Angreifer ihre Kommunikation nur leicht verschleiern, versagen signatur- und schwellenwertbasierte Regeln.

Ziel dieser Arbeit war es daher, ein Machine-Learning-Modell zu entwickeln, das DNS-basierte Datenexfiltrationen zuverlässig und möglichst in Echtzeit erkennt.

Methodisches Vorgehen

Die Entwicklung des Modells erfolgte auf Basis von CRISP-DM (Cross-Industry Standard Process for Data Mining) mit seinen sechs iterativen Phasen von Business Understanding bis Deployment.

Als Klassifikationsmodell kam XGBoost (Extreme Gradient Boosting) zum Einsatz, ein Verfahren, bei dem viele Entscheidungsbäume nacheinander trainiert werden und jeder neue Baum die Fehler seiner Vorgänger korrigiert. Statt einzelner DNS-Pakete klassifiziert das Modell Query-Response-Paare, also jeweils eine DNS-Anfrage zusammen mit der zugehörigen Antwort. Dadurch stehen mehr Informationen für die Klassifikation zur Verfügung, was die Fehlalarmrate senkt und gleichzeitig eine nahezu echtzeitfähige Erkennung ermöglicht, da zwischen Anfrage und Antwort in der Regel nur wenige Millisekunden liegen.

Datensätze

Bestehende DNS-Datensätze weisen erhebliche Qualitätsprobleme auf: leicht zu klassifizierender Netzwerkverkehr, fehlerhaftes Labeling, eine zu geringe Vielfalt an Angriffsmustern oder fehlende Attribute. Aus diesem Grund wurden im Rahmen der Arbeit ein Basisdatensatz und ein Evaluationsdatensatz erstellt und veröffentlicht.

Der Basisdatensatz diente dem Training des Modells und umfasst nach der Bereinigung knapp 800.000 gutartige sowie rund 786.000 bösartige Query-Response-Paare. Der gutartige Anteil stammt aus zwei bestehenden, bereinigten und zusammengeführten Quellen. Für den bösartigen Anteil wurden neun verschiedene Exfiltrationsprogramme eingesetzt, darunter Cobalt Strike, Iodine, Dnscat2, DnsExfiltrator sowie eine modifizierte Version von DnsExfiltrator. Die neun Programme unterscheiden sich deutlich in Protokoll, Kodierung und verwendeten Record-Typen. Ein besonderes Angriffsmuster weist der modifizierte DnsExfiltrator auf, der eine Base32map-Kodierung verwendet. Dabei wird jedes kodierte Zeichen durch ein dreibuchstabiges englisches Wort ersetzt, um den erzeugten DNS-Verkehr stärker an legitime, natürlich wirkende DNS-Anfragen anzugleichen.

Der Evaluationsdatensatz wurde unabhängig vom Basisdatensatz erstellt und diente der abschließenden Bewertung der Modellrobustheit. Als Quelle für gutartigen Netzwerkverkehr wurde das Netz der Technischen Hochschule Rosenheim genutzt. Den bösartigen Anteil bilden vier reale Malware-Familien, auf die das Modell zuvor nicht trainiert wurde: Anchor, Saitama, FrameworkPOS und Trojan.Win32.lsmdoor.gen.

Features

Um gut- und bösartige Query-Response-Paare zu unterscheiden, nutzt das Modell insgesamt elf Features aus zwei Kategorien. Die eine Kategorie betrachtet die Subdomain der DNS-Query – etwa ihre Länge, die Entropie der Zeichenverteilung, die Wechselrate von Vokalen und Konsonanten oder den Anteil an Großbuchstaben. Die andere Kategorie untersucht die Answer-Records der zugehörigen DNS-Response – etwa deren Homogenität, Größe oder die maximale Time to Live.

Um zu verstehen, welche dieser Features für die Entscheidung des Modells besonders wichtig sind, wurde eine SHAP-Analyse (Shapley Additive Explanations) durchgeführt (siehe Abbildung 1). Sie zeigt für jedes Feature, wie stark und in welche Richtung es die Einschätzung des Modells beeinflusst. Bei einer sehr kurzen Subdomain zum Beispiel tendiert das Modell eher zur Einschätzung „gutartig“, bei einer sehr langen eher zu „bösartig“. Die Analyse zeigt, dass vor allem drei Features das Zünglein an der Waage sind: die Länge der Subdomain, die maximale Time to Live und die Entropie. Sie liefern dem Modell also die deutlichsten Anhaltspunkte dafür, ob ein Query-Response-Paar bösartig ist oder nicht. Der Anteil an Großbuchstaben und die Homogenität der Answer-Records spielen im Vergleich dazu eine untergeordnete Rolle.

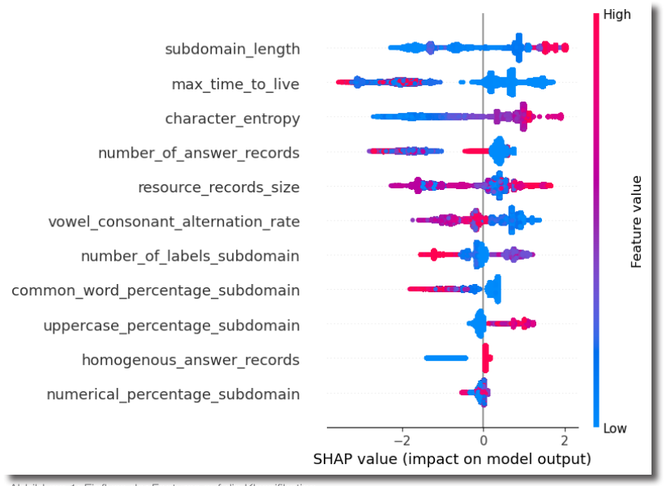

Abbildung 1: Einfluss der Features auf die Klassifikation

Evaluation

Die Evaluation zeigt, dass das entwickelte Modell grundsätzlich für den Praxiseinsatz geeignet ist. Es weist selbst in unbekannten Netzwerkumgebungen eine vergleichsweise niedrige Fehlalarmrate von 0,45 % auf. Die erzeugten Fehlalarme sind zudem von ihrer Struktur kaum von einer tatsächlichen Datenexfiltration zu unterscheiden, weshalb ihre fehlerhafte Klassifizierung nachvollziehbar ist. Da in großen Netzwerken dennoch eine beträchtliche Zahl an Fehlalarmen auftreten kann, lässt sich diese durch Whitelisting bekannter gutartiger Domains zusätzlich deutlich reduzieren – im Evaluationsdatensatz bereits um zwei Drittel durch den Ausschluss von nur drei Domains. Das Modell kann außerdem Angriffe durch Malware erkennen, auf die es nicht explizit trainiert wurde. Die Angriffe durch Anchor, Saitama, FrameworkPOS und Trojan.Win32.lsmdoor.gen hat das Modell noch nie zuvor gesehen. Dennoch ist es in der Lage, bei jeder Malware mindestens 25 % der Query-Response-Paare korrekt als bösartig zu identifizieren.

Eine ernüchternde Erkenntnis ist, dass das Modell aufgrund der niedrigen Erkennungsrate bei unbekannter Malware nur bedingt Angriffe in Echtzeit aufdecken kann. Sollten die nicht identifizierten bösartigen Query-Response-Paare zu Beginn eines Angriffs auftreten, könnte ein Unternehmen entsprechende Gegenmaßnahmen nur verzögert einleiten. Der verwendete Datensatz sollte aus diesem Grund regelmäßig um Angriffe neuartiger Malware erweitert und das Modell neu trainiert werden. Bei bekannter Malware werden 99,95 % der bösartigen Query-Response-Paare korrekt als bösartig identifiziert, wodurch eine Echtzeiterkennung von Angriffen in diesen Fällen nahezu garantiert ist.

Literatur

Kristijan Ziza, Pavle Vuletić und Predrag Tadić. „DNS exfiltration detection in the presence of adversarial attacks and modified exfiltrator behaviour”. In: International Journal of Information Security 22.6 (2023), S. 1865–1880.
Palo Alto Networks. Stop Attackers from Using DNS Against You. 2023.
Deb Radcliff. Taxonomy of SolarWinds SUNBURST DNS Abuse Tactics. 2021.
IBM. Cost of a Data Breach Report 2024.
Jacob Steadman. „Enhancing DNS data exfiltration protection through SDN data plane programming“. Diss. 2021.