Auteur: Vincent van Dijk, eigenaar van Security Scientist. Hij is bereikbaar via: vincent@securityscientist.net

Wat ik leerde van het bouwen van AI-systemen

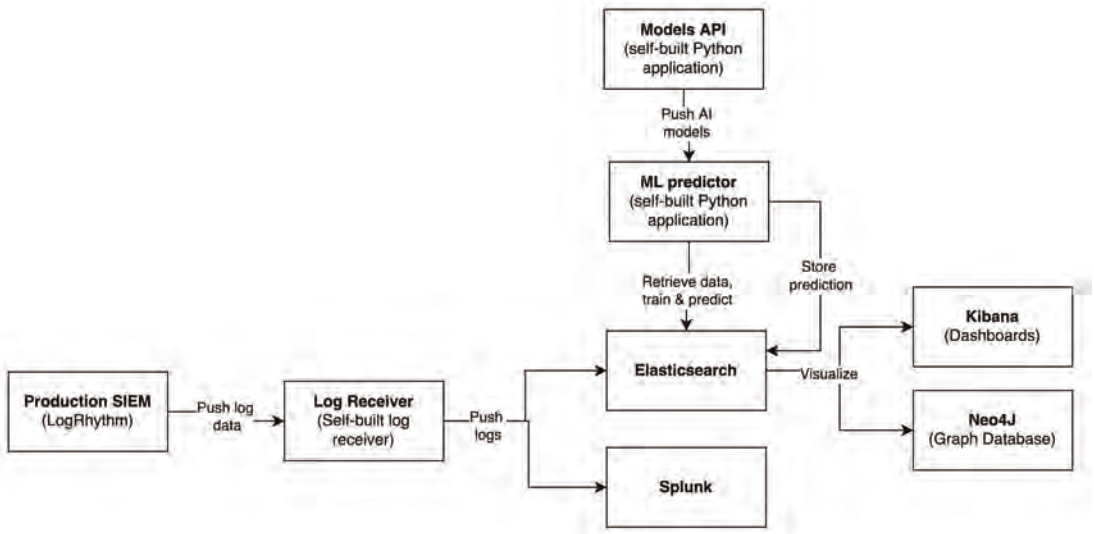
Jan, werkzaam in de verkoop, had problemen met zijn gezin. Hij kon werken, maar zijn gedachten waren elders. Zijn productiviteit daalde en AI classificeerde hem met 'snel productiviteitsverlies'. Kort daarna werd Jan ontslagen. Dit is niet echt gebeurd, met Jan gaat het prima. Het was een risico van één van de AI-modellen die ik in 2019 beschreef.

Voor mijn eerste werkgever, een cybersecurity bedrijf, maakte ik in 2019 een AI-systeem bestemd voor beveiligingsdoeleinden, maar - naar later bleek - met het risico van privacyschending als gevolg van classificering van de prestaties van werknemers. Ik hield het internetgedrag van werknemers bij door middel van een Excel-sheet met 40 rijen en één kolom per maand. Elke cel vertegenwoordigde het internetgedrag van een werknemer voor die maand, inclusief informatie over het meest dominante of afwijkende gedrag en hoe vaak dit zich voordeed.

Dit werd niet gerealiseerd met een nieuw Large Language Model. Het was pas 2019 en mijn oplossing werd al uitgedacht en ontwikkeld door Hugo Steinhaus in 1956. Met de juiste data en een paar creatieve ideeën, maak je zowel de slechtste en beste AI's met eenvoudige modellen. Zelfs kostenefficiënter en sneller (met een training in luttele seconden) dan elk groot taalmodel.

Onethische modellen

Ons idee over AI is verkeerd. Wij denken dat generatieve AI (GenAI) de veroorzaker van schade zal zijn, maar dat is niet



Figuur 1: De infrastructuur van het op maat gemaakte SIEM Machine Learningsysteem.

waar! Oudere modellen kunnen net zo slim zijn als de GenAI. Onder het mom van veiligheidsdenken creëren wij onethische modellen ogenschijnlijk met volkomen legitieme redenen.
Ook Amazon, met 1,5 miljoen werknemers wereldwijd (8,5% afgezet tegenover de Nederlandse bevolking), bouwde een dergelijk AI-systeem om beslissingen van werknemers te automatiseren. Jan kent een levensechte tegenhanger bij Amazon: Stephen. Hij werd door AI ontslagen.
Wij allen schatten in dat Algemene Kunstmatige Intelligentie een reëel gevaar betekent, misschien in de toekomst. Nu weerhoudt deze angst ons om te leren waar AI echt over gaat. En dat is dataverwerking.

Mijn ontdekking in 2019

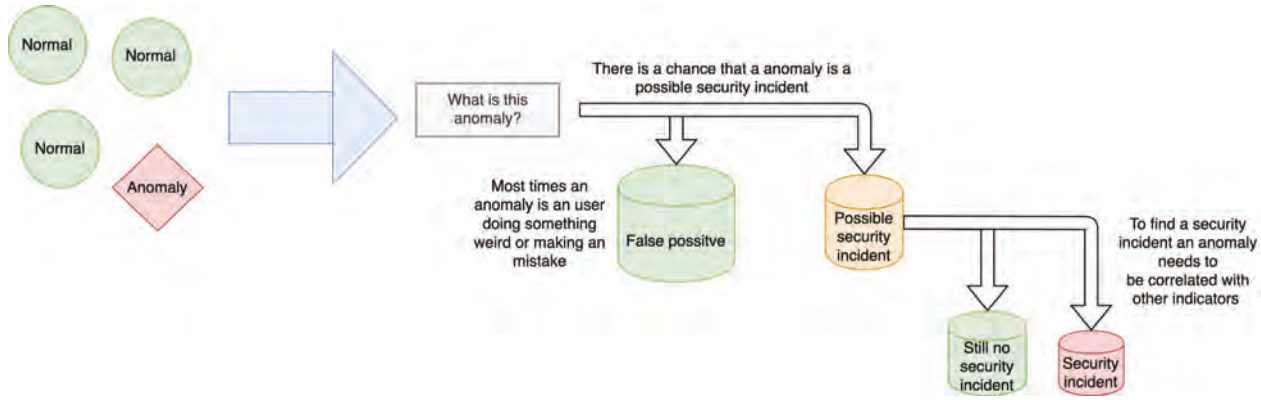
Ik bouwde elk Machine Learning-model dat ik wilde voor een Nederlands beveiligingsbedrijf. Ik had mijn eigen 'schaduw-SIEM' - van hun LogRhythm SIEM - in Elasticsearch. Met mijn schaduw was ik 'in control': ik verrijkte mijn logs, implementeerde geautomatiseerd modellen en realiseerde geautomatiseerde hertraining. Alles voor een complete AI-gerichte SIEM. De modellen konden het gedrag van gebruikers en systemen classificeren en vonden problemen die een compleet SOC-team vanuit Spanje niet kon detecteren: sneller, indringender en met betere resultaten dan Machine Learning-pakketten zoals: Splunk en LogRhythm. Dat project is helaas nooit afgerond.

Complexe modellen, slechte prestaties

Eind 2018 bezocht ik een conferentie over mijn bachelor school-project in Zwolle. Ik was vroeg en dat was mijn geluk want er was een sneeuwstorm die latere treinen blokkeerde. Op de conferentie was ik één van de vijf aanwezige scholieren; de anderen waren thuis gebleven. Ik ontmoette een CEO en als een van de weinige daar sprak hij mij aan. Ik had net voor mijn studie bij een technologisch gedreven verkeersinfrastructuurbedrijf het Machine Learning-project afgerond en zonder te weten wat hij wilde horen, bleef ik er maar over praten. Ik werd daarop aangenomen. Mijn volgende Machine Learning-project zou bij zijn beveiligingsbedrijf zijn. Mijn functie daar: datawetenschapper. AI was 'hot'. Voor hun Machine Learning-project kreeg ik mijn eigen servers. Ik bouwde mijn schaduw-SIEM, iedereen was geïnteresseerd. Nu begon de pret.

Het werd een op maat geschreven applicatie om logs te ontvangen (tot 20.000 per seconde!), een Machine Learning API, voor AI-modellen, opslag en publicatie. Een applicatie die deze modellen automatisch gebruikte en trainde op basis van de gegevens die waren opgeslagen. Ik had zelfs een Splunk-instance testlicentie om mijn tools te vergelijken. Alles bij elkaar zeven systemen, daarvan drie volledig op maat gemaakt en alle integraties daartussen eveneens.

Ik startte mijn onderzoek naar Machine Learning-modellen om beveiligingsproblemen te vinden. Ik deed veel modellenon-

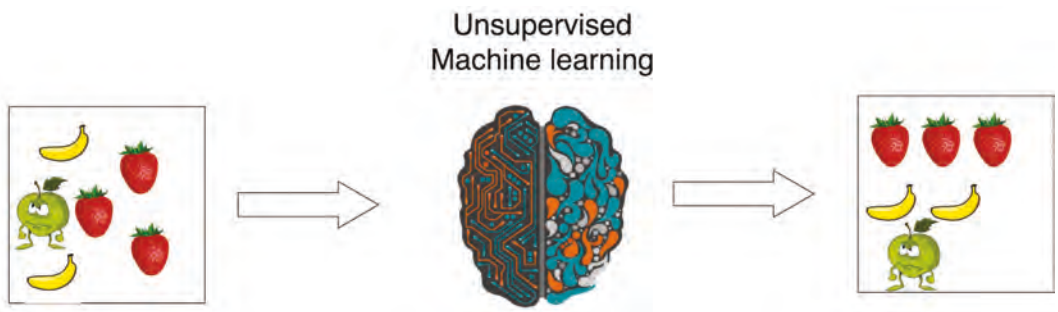


Figuur 2: Anomaliedetectie in cyberbeveiliging.

derzoek: van vroege Large Language Models en Deep Learning tot modellen zonder supervisie. Tot mijn verbazing ondervond ik dat hoe complexer het model, hoe slechter de prestatie. Deze modellen konden geen interessante conclusies trekken uit de gegevens. De reden: op welke locatie moesten deze modellen zoeken? Elk beveiligingsincident is uniek, beveiligingsincidenten komen niet zo vaak voor. Het Verizon Data Breach Investigation Report bevat slechts 1.000+ incidenten per jaar. Dat zijn maar een paar megabytes aan gegevens tegenover de dagelijkse petabytes-overdracht, dat is een speld in een hooiberg zoeken.

Ik wijzigde mijn aanpak liever dan beveiligingsincidenten te voorspellen. Ik zocht anomalieën met eenvoudig uit te leggen modellen. Anomaliedetectie kent echter ook problemen: is het een beveiligingsincident of een systeem dat, of een persoon die, zich afwijkend gedraagt? Elke dag handelt iemand ongewoon. Er logt iemand in op een applicatie wat hij niet eerder had gedaan, omdat hij als chef besloot dat dit zijn verantwoordelijkheid werd. Of een systeembeheerder zet terabytes aan data over wegens een nieuwe architectuur, zie figuur 2 voor de werkelijkheid anno 2019.

Ik wist dat niet elke anomalie een waarschuwing moest triggeren, omdat de meeste niet gerelateerd zijn aan beveiligingsvraagstukken. Tijd voor een compleet nieuwe aanpak. Ik verdiepte mij in clustermodellen. Clusteren, een techniek om te groeperen in datasets. Het algoritme voor machinaal leren probeert de dataset te begrijpen en vormt clusters van vergelijkbare datagroepen.



Figuur 3: Model voor toezichtloos machinaal leren en clusteren.

K-means clustering

Ik kwam uit bij K-Means clustering (1). Het verdeelt de gegevens in K-clusters (vectorkwantisatie - signaalanalyse), waarbij elk gegevenspunt behoort tot het cluster met het dichtst nabijgelegen gemiddelde. Een eenvoudig voorbeeld:

Voedsel	Gezondheidsscore
Appel	88
Banaan	90
Big Mac	10
Koekje	20

Appel en Banaan bezitten vergelijkbare scores, zij vormen een groep. Big Mac en Koekje hebben ook een vergelijkbare maar lagere score, zij zijn de tweede groep. K-Means maakt twee groepen. Dit is een voorbeeld waarbij een mens de groepen gemakkelijk overziet, maar in het echt zijn er honderden kolommen en soms miljoenen rijen. Dat is waar het K-Means-model uitblinkt.

K-Means is efficiënter en extreem snel geworden. De eenvoudigste versies van Deep Learning-modellen kunnen gemakkelijk uren kosten om te trainen. Dit K-Means-model traint in slechts enkele seconden. Dat maakt toegankelijke hertraining mogelijk, zodat het model zich gemakkelijk aanpast aan het dagelijkse, soms vreemde gedrag van mensen en technologie.

Met K-Means bij de hand vond ik een databron voor testen: de Palo Alto firewalls. Die informatie was geweldig omdat, met de juiste configuratie, je ziet welke applicaties gebruikers benutten, hoe vaak en hoeveel gegevens er werden overgedragen. Ik begon met interne tests, waarbij ik al het netwerkverkeer van de medewerkers had verzameld en gefilterd, zie voorbeeld onderaan pagina 19 Vereenvoudigde fictieve gegevens van de Palo Alto firewalls.

Stel een dataset voor van 10.000 records, 40 Azure AD-gebruikers en honderden applicaties. Alleen daaruit al trek je opmerkelijke conclusies. Je achterhaalt wie cloudapplicaties of schaduw-IT gebruikt, die ze niet mogen gebruiken. Toch zegt dit niets over de beveiliging. Ze gebruiken de cloudapplicatie of schaduw-IT wellicht om persoonlijke redenen. Misschien is het normaal gedrag? Dat riep de vraag op: 'Wat is echt abnormaal/ongewoon gedrag?'

Met K-Means zoeken wij beter uit wat normaal gedrag is en wat niet. Zoals eerder groepeer je data, maar in plaats van een appel en een Big Mac, groeperen wij specifieke gebruikers en hun applicatiegedrag.

Het resultaat:

Azure AD-gebruikersnaam	Toepassing
Jan	Groep 1
Peter	Groep 1
David	Groep 2
Kees	Groep 2

K-Means groepeerde op basis van toepassingengebruik. Jan en Peter delen een groep, zo ook David en Kees. Geweldig, maar je trekt hieruit niet echt een conclusie. Zijn David of Kees in groep 2 abnormaal? Of zijn beide groepen normaal? En waarom zijn ze bij elkaar ingedeeld? Je krijgt geen antwoord op

Voorbeeld:

Azure AD-gebruikersnaam	Toepassing	Gebruikte tijden	Datum
Jan	Google Drive	10	10-01-2019
Peter	OneDrive	20	10-01-2019
David	YouTube	12	10-01-2019
Kees	Amazon Web Services	19	10-01-2019

Vereenvoudigde, fictieve gegevens van de Palo Alto firewalls.

deze vragen, tenzij je verder onderzoekt. K-Means is een vrij eenvoudig instrument en interpreteert de resultaten niet voor je.

Naamgeving groepen

Deze vragen hielden mij wakker. Met Machine Learning krijg je spannende resultaten, maar de moeilijkheid is: wat doe je er mee? In cyberbeveiliging onderzoek je de onderliggende redenering om te zien of de gebeurtenis kwaadaardig is of dat het gaat om een willekeurige, legitieme afwijking.
Mijn oplossing: geef de groepen een naam als alternatief voor een nummering. Een naam op basis van de meest dominante applicatie. Groep 1 kan bestaan uit Google Drive-, OneDrive- en iCloud-gebruikers. Werd Google Drive het meest gebruikt, dan heet de groep 'Google Drive'. Als er meerdere groepen zijn met meerdere dominante applicaties, scheidde ik de groepen op basis van het totaal aantal activaties, bijvoorbeeld Google Drive-Hoog en Google Drive-Laag. Tot slot werd elke toepassing die niet binnen de gebruikte kaders viel, als 'Abnormale toepassing' gekenmerkt en kreeg hogere prioriteit.

Het model opnieuw uitgevoerd:

Azure AD-gebruikersnaam	Toepassing
Jan	Google Drive - Hoog
Peter	Google Drive - Hoog
David	YouTube - Hoog
Kees	Abnormale toepassing - Laag

Nu werden de resultaten resultaten interessant, zie tabel onderaan pagina 20. Wij zagen een trend in hoe mensen zich gedroegen en op basis daarvan werden geclassificeerd. Met verder onderzoek verbonden wij zelfs conclusies aan de verschillende indelingen:

- Het toepassen van Google Documenten was de primaire manier van werken. Wij ontdekten bij 'Google Drive – Hoog' meestal productieve werknemers. Bij 'Google Drive – Laag', betekende dit vaak dat een medewerker ziek was of op vakantie ging;
- Bij een groot IT-project werd de systeembeheerder vaak ingedeeld bij 'Abnormale toepassing';
- In geen geval mag iemand buiten de IT-afdeling in de groep 'Abnormale toepassing' vallen. Dit was waarschijnlijk een indicatie van niet-goedgekeurde schaduw-IT taken.

Enkele conclusies benadrukt, zie onderstaande tabel, cursief geschreven:

- Jan was niet op vakantie in februari, hij was onproductief. Waarschijnlijk door zijn verhuizing;
- Jan mag nooit worden ingedeeld bij 'Abnormale toepassingen', hij is geen IT-beheerder. Bij nader onderzoek bleek dat zijn computer spyware bevatte;
- Wij verwachtten geen hoge abnormale classificatie voor Kees in maart. Er waren geen grote projecten gepland. Het bleek dat Kees vanaf bedrijfssystemen toegang had tot zijn thuisserver via Amazon Web Services;
- David was niet productief. Zonder uitzondering stelden we vast dat alle gebruikers die met 'YouTube-Hoog' werden geclassificeerd, ondermaats presteerden.

Dit model was zeer sterk in het monitoren van medewerkers. Het was niet perfect omdat details ontbraken, maar gedrag kon op hoog niveau waargenomen en afgezet worden tegenover

Indien verrijkt met de code over de afgelopen 3 maanden:

Azure AD-gebruikersnaam	Januari	Februari	Maart
Jan	Google Drive - Hoog	*Google Drive - Laag*	*Abnormale toepassing - Laag*
Peter	Google Drive - Hoog	YouTube - Hoog	Google Drive - Laag
David	*YouTube - Hoog*	*YouTube - Hoog*	*YouTube - Hoog*
Kees	Abnormale toepassing - Laag	Google Drive - Hoog	*Abnormale toepassing - hoog*

Zelfs AI, zoals ChatGPT, heeft moeite met het logisch vinden van bedreigende actoren

verwachtingen. Door onderzoek konden we valideren of alle classificaties van gebruikers al dan niet overeenkwamen met verwachtingen en bepalen of daar een goede reden voor afwijking bestond.

Dit eenvoudige op K-Means gebaseerde model was... **griezelig**.

Conclusies

Zorgen rondom het model:

1. Hoe zit het met de privacy? Beveiliging verbeterde, maar ten koste van privacy;
2. Zonder begrip voor het model onderzochten onze junior analisten weinig succesvol de waarschuwingen. Dat vereiste een andere manier van denken;
3. Zonder organisatiebegrip waren deze modellen te complex. Doorgronden en inzicht creëren in de nuances van zo'n Machine Learning-project is nodig.

Gevolgen: de klassieke manier van beveiligingsmonitoring is voor klanten al complex, laat staan deze nieuwe modellen. Zonder training van veiligheidsanalisten zijn 'use cases' betekenisloos. Professionals met ervaring in Machine Learning en cyberbeveiliging zijn vereist, de organisatie moet zich aanpassen. De overstap naar nauwkeurige detecties op basis van AI was te complex. Er waren betere opties om de kwaliteit van onze beveiligingsdienstverlening te verhogen.

Essentiële vraag: 'Welk probleem proberen we op te lossen met AI? ' Proberen wij cyberbedreigingen efficiënter aan te pakken? Dan lijkt AI niet de juiste oplossing. Wij maken grotere sprongen door betere en gerichtere strategieën te definiëren.
Of proberen wij ons vermogen om cyberbedreigingen te detecteren te verbeteren? AI is dan ook niet de juiste keuze, omdat wij niet over grootschalige waarnemingen beschikken om gevaarlijke actoren te detecteren. De grootste dataset van incidenten bevat slechts een paar duizend per jaar, is te weinig gedetailleerd en onvoldoende voor een AI-model training. Wij moeten vertrouwen op anomaliedetectiemethoden zoals het K-Means, dat wel training en inspanning vereist.

Zelfs AI, zoals ChatGPT, heeft moeite met het logisch vinden van bedreigende actoren, omdat het vooral is getraind op internetgegevens en geen echte data heeft over bedreigende actoren.

Het ontwikkelen van AI-modellen leert ons veel. AI is niet het antwoord als je niet weet welke vraag je stelt. De meeste aangeboden AI-oplossingen schieten tekort; leveranciers maken slimme oplossingen, maar niet AI zorgt ervoor dat het werkt. AI in combinatie met bovengenoemde modellen is een leermodel, mits je tijd neemt voor het vinden en stellen van de juiste vragen, net als bij de griezelige K-Means AI, die ik maakte en onbedoeld mensen 'bespioneerde'.

Referentie

(1) https://en.wikipedia.org/wiki/K-means_clustering