Analýza ML metód a algoritmov pre detekciu phishing a spoofing

Projekt INNOvative SOC/INNOSOC

Obsah

Úvod 3
Experimenty 4
Cieľ testovania 4
Dataset a predspracovanie 5
Výsledky: 5
K-Means 5
HDBSCAN 7
Zhrnutie 10
Záver 10

Úvod

Phishing je klamlivá kybernetická taktika, ktorú útočníci používajú na získanie citlivých informácií od nič netušiacich obetí. Táto metóda zvyčajne zahŕňa posielanie podvodných e-mailov alebo správ, ktoré napodobňujú legitímne zdroje, ako sú banky, vládne agentúry alebo známe spoločnosti. Cieľom je oklamať príjemcov, aby prezradili osobné údaje, ako sú prihlasovacie údaje, čísla kreditných kariet alebo iné citlivé údaje ako napr. číslo sociálneho zabezpečenia.

Phishing predstavuje typický príklad útoku sociálneho inžinierstva (angl. social engineering). Útoky sociálneho inžinierstva sú rôzne techniky, ktoré zneužívajú ľudskú psychológiu na škodlivé účely. Phishing e-mailov je asi najznámejší príklad, zahŕňa útočníkov posielajúcich e-maily, ktoré sa zdajú byť z dôveryhodných zdrojov, aby oklamali príjemcov a prinútili ich poskytnúť citlivé informácie alebo kliknúť na škodlivé odkazy. Tieto e-maily často napodobňujú vzhľad a jazyk legitímnej korešpondencie od bánk, poskytovateľov služieb alebo dokonca priateľov a kolegov, čo ich robí ťažko rozlíšiteľnými od skutočných komunikácií.

Phishing sa rozširuje aj na textové správy, známe ako SMS phishing alebo *smishing*. Pri týchto útokoch páchateľ posiela textové správy, ktoré lákajú príjemcu, aby klikol na škodlivý odkaz alebo poskytol osobné údaje. Tieto správy môžu pôsobiť ako správy od dôveryhodnej organizácie a často vytvárajú pocit naliehavosti na okamžitú reakciu na prijatú správu. *Vishing*, alebo hlasový phishing (angl. voice phishing), je ďalšia varianta, kde útočníci využívajú telefonáty na získanie citlivých informácií. V týchto situáciách sa útočník často vydáva za legitímneho zástupcu organizácie, aby získal informácie alebo priamu akciu od obete útoku.

Vydávanie sa za niekoho a pretextovanie sú sofistikované taktiky používané pri útokoch sociálneho inžinierstva, pri ktorých útočník predstiera určitú autoritu alebo vytvára presvedčivý scenár na zneužitie dôvery obete útoku. Prijatím identity autoritatívnej osoby ako je policajt, IT technik alebo firemný manažér, môžu útočníci manipulovať cieľmi útokov, aby prezradili dôverné informácie alebo vykonali kroky ohrozujúce bezpečnosť. Vytváranie presvedčivých scenárov si vyžaduje podrobné poznanie obete alebo jej organizácie, čo robí tieto útoky obzvlášť zákernými a účinnými.

Baiting (angl. baiting = vábiť, lákať) a *tailgating* (angl. tailgate = zadné vrátka) sú ďalšie rozšírené metódy sociálneho inžinierstva. *Baiting* útoky zahŕňajú lákanie obetí útoku sľubom tovaru alebo služieb, aby ukradli ich osobné údaje alebo infikovali ich systémy škodlivým softvérom. Môže to byť vo forme stiahnutia bezplatného softvéru alebo prístupu k exkluzívnemu obsahu. *Tailgating* je naopak taktika fyzického narušenia bezpečnosti, pri ktorej neoprávnená osoba získa prístup do zakázanej oblasti tým, že sleduje niekoho oprávneného. Tento útok využíva spoločenskú normu zdvorilosti a

prirodzenú tendenciu pomáhať druhým, čo útočníkom umožňuje preniknúť do zabezpečených prostredí.

Vyvolávanie a manipulácia sú jemné, no silné techniky sociálneho inžinierstva. Tieto metódy zahŕňajú získavanie informácií prostredníctvom zdanlivo nevinného rozhovoru, kde útočník zapája cieľ obeť do diskusie a nenápadne ho usmerňuje k získaniu požadovaných informácií. Útočníci často emocionálne manipulujú cieľmi, využívajúc ľudské sklony, ako je túžba byť obľúbený alebo strach z konfrontácie. Získanie dôvery prostredníctvom vzťahu a empatie je v týchto interakciách kľúčové, pretože znižuje obranu cieľa útoku a robí ho náchylnejším k prezradeniu citlivých informácií.

Každý útok sociálneho inžinierstva ukazuje zložitosť a rozmanitosť taktík, ktoré útočníci používajú na zneužitie ľudskej psychológie. Zdôrazňujú potrebu neustálej ostražitosti v osobnom aj pracovnom prostredí, aby sa zabránilo týmto manipulatívnym technikám. Pochopenie povahy týchto útokov je prvým krokom k rozvoju účinných stratégií na ich zvládnutie, pričom sa zdôrazňuje dôležitosť povedomia a vzdelávania v prebiehajúcom boji proti kybernetickým hrozbám.

Na boj proti phishingu musia byť jednotlivci a organizácie ostražití a skeptickí voči nevyžiadaným komunikáciám. Kľúčové preventívne opatrenia zahŕňajú neklikanie na odkazy ani sťahovanie príloh z neznámych zdrojov, overovanie pravosti žiadostí o osobné údaje a používanie anti-phishingových nástrojov dostupných vo väčšine moderných webových prehliadačov a e-mailových služieb. Vzdelávanie a školenia v identifikácii phishingových pokusov sú kľúčové pre vybudovanie prvej línie obrany proti týmto kybernetickým útokom.

Experimenty

Cieľ testovania

Cieľom pilotného testovania bolo **overiť a porovnať vhodnosť metód K-Means a HDBSCAN na identifikáciu phishingu** v dátach vo forme **sieťových logov**. Testovanie malo zistiť, či tieto metódy dokážu:

- identifikovať opakujúce sa vzorce a husté oblasti logov
- odhaliť anomálie a outlier-y, ktoré môžu indikovať phishingové alebo podozrivé udalosti
- porovnať rozdiely v prístupe oboch algoritmov, najmä schopnosť detekcie outlier-ov a štruktúru klastrov
- poskytnúť zrozumiteľnú interpretáciu klastrov.

Dataset a predspracovanie

Pre pilotné testovanie sme pracovali so súborom 57 000 logov, ktoré boli získané z prevádzky na FRI UNIZA, pôvodne obsahujúcim približne 160 premenných, z ktorých bolo identifikovaných 41 relevantných pre identifikáciu phishingu. Viaceré z týchto premenných boli kategorické, a preto bolo nevyhnutné ich predspracovať kombináciou one-hot encodingu a hashing tricku, čím sa dataset rozšíril na 2 009 premenných vhodných pre klastrovanie.

Pred klastrovaním boli všetky premenné štandardizované pomocou StandardScaleru, aby sa odstránili rozdiely v rozsahu hodnôt. Následne bola aplikovaná redukcia dimenzie PCA na 50 komponentov, čo umožnilo efektívnu prácu s algoritmami klastrovania pri zachovaní variability dát.

Tento krok bol nevyhnutný a časovo náročný, pričom má zásadný význam pre prípravu modelu strojového učenia. Logy pochádzajúce z rôznych systémov totiž obsahujú odlišné premenné, a preto bude pri tvorbe finálneho modelu na identifikáciu phishingu veľkou výzvou vyvinúť robustný skript, schopný spracovať aj neočakávané alebo nové premenné.

Výsledky:

K-Means

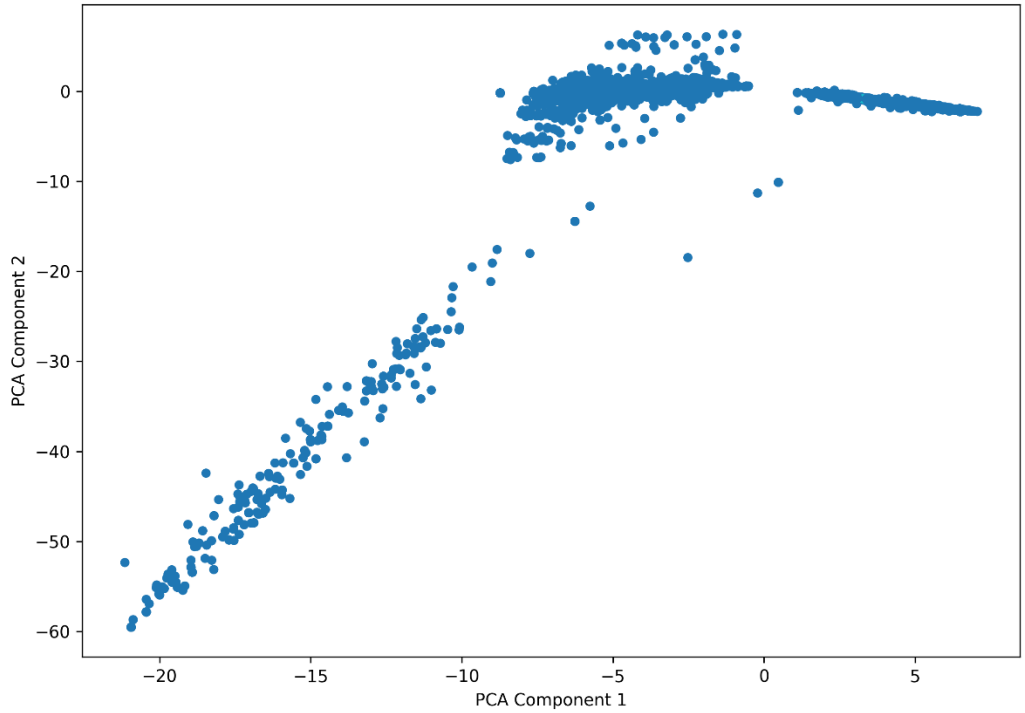
Pri použití algoritmu K-Means na náš dataset logov sme najprv predpokladali 3 klastre. Výsledné rozdelenie bolo veľmi nerovnomerné:

Cluster 0: 57 695 bodov

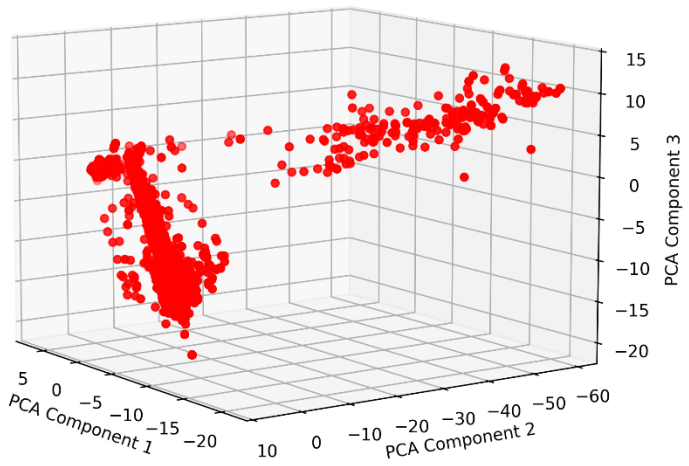
Cluster 1: 2 body

Cluster 2: 24 bodov

Výsledky K-Means sú vizualizované na obrázkoch 1 a 2.



Obr. 1: 2D vizualizácia K-Means



Obr. 2: 3D vizualizácia K-Means

Zmena predpokladaného počtu klastrov (napr. na 4 alebo 5) nepriniesla lepšie rozdelenie dát. Niektoré nové klastre obsahovali len 1–2 body, pričom hlavný klaster stále zachytával väčšinu dát. Toto nerovnomerné rozdelenie odráža prirodzenú štruktúru datasetu, kde je väčšina logov bežného typu a iba veľmi malé množstvo logov predstavuje podozrivé alebo phishingové udalosti.

K-Means priradí každý bod do nejakého klastru, bez možnosti označiť body ako body ležiace mimo zhluku alebo klastru (angl. outlier). V dôsledku toho sú extrémne alebo neštandardné hodnoty automaticky zahrnuté do niektorého z klastrov, často do malých

alebo marginálnych klastrov. Tento fakt môže znížiť schopnosť K-Means samostatne detegovať netypické alebo podozrivé logy.

Celkovo sa ukazuje, že K-Means je vhodný pre rýchle a deterministické rozdelenie dát do predom stanoveného počtu klastrov, avšak jeho schopnosť zachytiť jemnejšie rozdiely alebo identifikovať odchýlky v logoch je obmedzená. Nerovnomerné rozdelenie v tomto prípade korešponduje s prirodzenou nerovnováhou dát, kde je počet phishingových logov veľmi nízky.

HDBSCAN

Algoritmus HDBSCAN bol aplikovaný na dataset logov s parametrami min_cluster_size=3000 a min_samples=20. Tieto nastavenia umožnili identifikovať 6 hlavných klastrov a zároveň vyznačiť odľahlé hodnoty (outliery), ktoré nepripadali do žiadneho klastru.

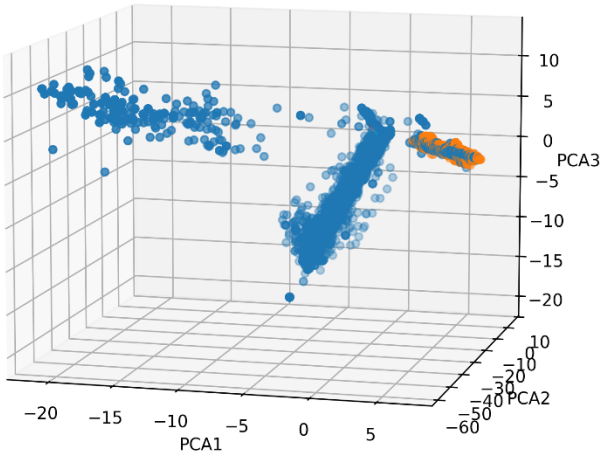
Výsledné rozdelenie počtu logov v jednotlivých klastroch bolo nasledovné:

Outliery (label = -1): 14 913 bodov
Cluster 0: 13 493 bodov
Cluster 1: 6 702 bodov
Cluster 2: 7 704 bodov
Cluster 3: 7 891 bodov
Cluster 4: 3 363 bodov
Cluster 5: 3 655 bodov

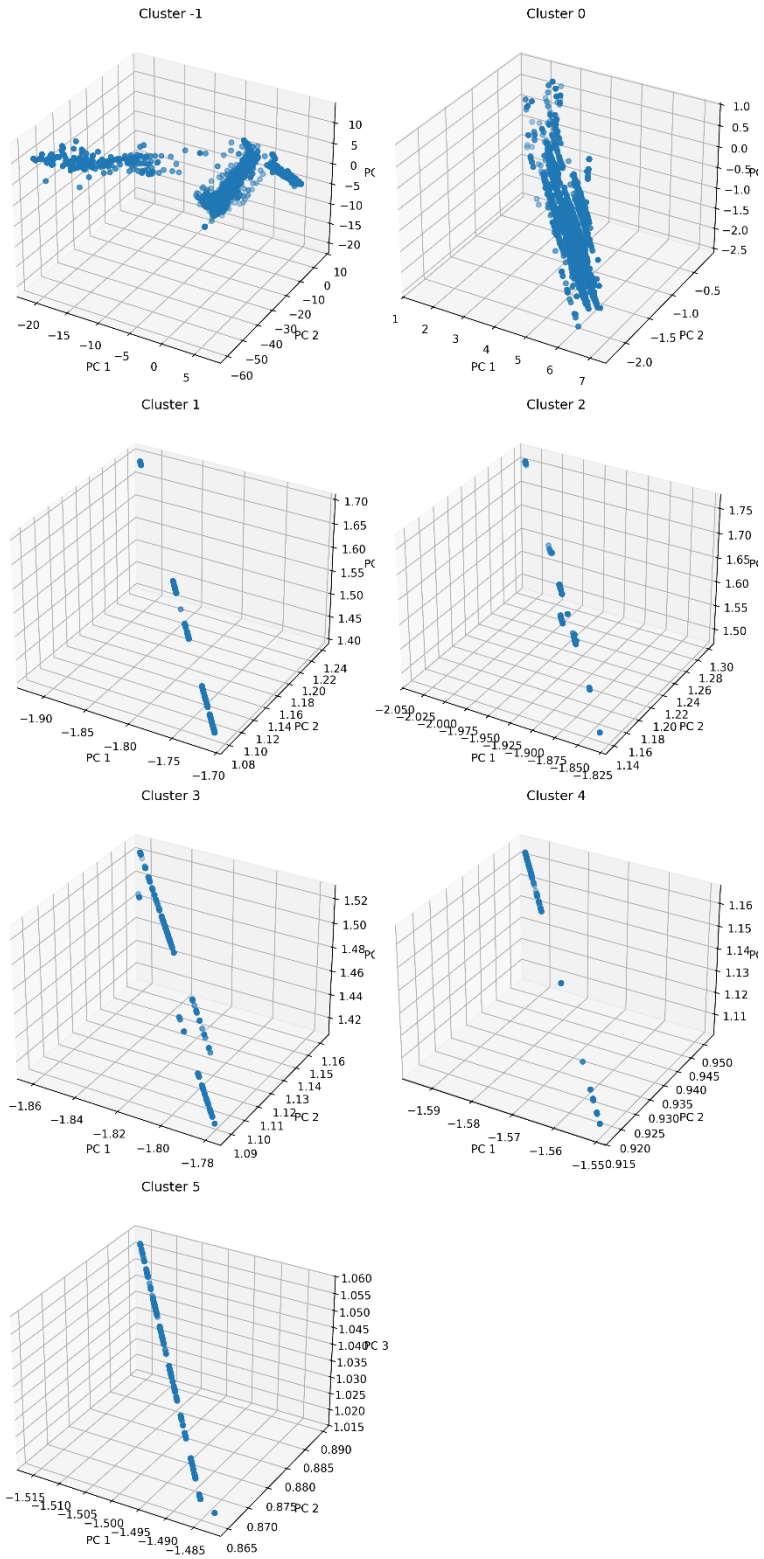
Interpretácia výsledkov

1) Outliery (-1):
   - algoritmus označil 14 913 logov ako odľahlé, ktoré nepatrili do žiadneho hustého klastru
   - tieto logy pravdepodobne obsahujú netypické alebo potenciálne podozrivé udalosti, čo je výhodou pre detekciu phishingu alebo anomálií
2) Hlavné klastre (0–5):
   - klastre sú pomerne rovnomerne zastúpené, pričom najväčšie klastre obsahujú 13 493 a 7 891 bodov a menšie klastre 3 363 a 3 655 bodov
   - toto rovnomernejšie rozdelenie v porovnaní s K-Means umožňuje lepšie zachytiť jemnejšie rozdiely medzi logmi a identifikovať štruktúru hustoty v dátach
3) Význam parametrov:
   - min_cluster_size=3000 zabezpečilo, že vznikajú len významné klastre, ktoré obsahujú dostatočný počet bodov
   - min_samples=20 určil citlivosť na hustotu a prispel k detekcii outlierov

Výsledky HDBSCAN sú vizualizované na obrázkoch 3 a 4.



Obr. 3: 3D vizualizácia HDBSCAN



Obr. 4: 3D vizualizácia jednotlivých klastrov pri HDBSCAN

Na obrázku 4 vidíme, že logy priradené do jednotlivých klastrov sú rozložené v rovnobežných rovinách, zatiaľ čo logy označené ako odľahlé hodnoty (outliery) sa nachádzajú mimo týchto hlavných rovín. Tento vizuálny efekt naznačuje, že HDBSCAN

dokáže efektívne rozlíšiť typické správanie od netypických alebo podozrivých udalostí, na rozdiel od K-Means, ktorý priraďuje každý bod do nejakého klastra.

V tomto kontexte predstavuje využitie vzdialenosti logu od roviny hlavného klastra ako pravdepodobnostného skóre phishingu nový a inovatívny spôsob interpretácie dát. Tento prístup umožňuje kvantifikovať mieru odchýlky jednotlivých logov a potenciálne identifikovať podozrivé alebo škodlivé udalosti s jemnejšou granularitou než bežné binárne označenie outlierov.

Zároveň je potrebné tento koncept v rámci ďalšieho výskumu overiť, aby sa potvrdila jeho účinnosť pri automatizovanej detekcii anomálií a phishingových útokov, a aby bolo možné vyvinúť robustný systém, ktorý kombinuje vzdialenostné skóre s ďalšími charakteristikami logov (napr. čas, IP adresy, typ udalosti) pre presnejšiu identifikáciu podozrivých aktivít.

Zhrnutie

HDBSCAN sa preukázal ako vhodný nástroj pre analýzu logov s nerovnomernou distribúciou, keďže dokáže:

- identifikovať husté oblasti logov
- označiť odľahlé hodnoty (outliery), ktoré môžu predstavovať phishing alebo iné podozrivé udalosti
- zachovať rovnomernejšie rozdelenie klastrov v porovnaní s K-Means, čo uľahčuje interpretáciu výsledkov a ďalšiu analýzu
- nový a inovatívny prístup: ak je nový log priradený do niektorého klastru, zvyčajne nejde o anomáliu ani potenciálny útok. Ak však log nie je zaradený do žiadneho klastra, jeho miera závažnosti sa môže kvantifikovať pomocou vzdialenosti od najbližšieho klastru, čím vzniká pravdepodobnostné skóre phishingu. Tento prístup umožňuje jemnejšie kvantifikovať riziko podozrivých udalostí a podporuje automatizovanú detekciu anomálií.

Celkovo HDBSCAN poskytuje robustný a flexibilný prístup ku klastrovaniu sieťových logov, najmä pri veľkých a heterogénnych datasetoch, pričom nový koncept pravdepodobnostného skórovania outlierov predstavuje sľubnú oblasť pre ďalší výskum a overenie jeho účinnosti.

Viac informácií k popisu spomínaných algoritmov pre klastrovanie možno nájsť v podkladovom dokumente Clustering.pdf

Záver

K-Means je vhodný, ak je potrebný rýchly a deterministický rozklad, avšak nedeteguje outliery a nerovnomerne rozdeľuje body v nerovnomerne hustých dátach.

HDBSCAN je vhodný pre robustnú analýzu logov, pretože dokáže detegovať husté oblasti aj odľahlé hodnoty, čo je významnou výhodou pri identifikácii phishingu a iných anomálií v sieťových logoch. Na rozdiel od algoritmov, ktoré priraďujú každý bod do klastru, HDBSCAN umožňuje označiť netypické alebo odľahlé logy (outliery), ktoré nespĺňajú kritériá hustoty potrebné pre vznik klastra. Týmto spôsobom algoritmus efektívne filtruje šum a sústredí sa na logy, ktoré reprezentujú opakovateľné a významné vzory správania.

V rámci pilotného testovania sa HDBSCAN preukázal ako vhodnejší na zachytenie jemnejších rozdielov medzi logmi, ktoré by mohli byť indikátormi podozrivých alebo neštandardných udalostí. Algoritmus dokáže rozdeliť dáta do klastrov podľa hustoty bodov, pričom zachováva štruktúru hierarchie a umožňuje identifikovať menšie, ale významné skupiny logov, ktoré by pri použití klasického K-Means mohli byť „stratené“ v masívnych klastroch.

Hlavnou prednosťou HDBSCAN voči K-Means je teda jeho schopnosť:

- automaticky určiť počet klastrov podľa hustoty dát
- robustne identifikovať outliery, ktoré môžu indikovať anomálie alebo phishingové aktivity
- zachovať jemné rozdiely medzi logmi, čo zlepšuje interpretáciu a analytickú hodnotu výsledkov.

Celkovo HDBSCAN poskytuje flexibilný a robustný prístup k analýze logov, ktorý zohľadňuje prirodzenú variabilitu dát a umožňuje efektívnu detekciu podozrivých udalostí bez nutnosti predom stanoveného počtu klastrov.