Rapport: Phishing Detection med Maskininlärning

Introduktion

Detta projekt syftar till att utveckla en maskininlärningsmodell som kan identifiera om webbsidor innehåller phishing eller inte. Det är viktigt att upptäcka phishing eftersom dessa attacker kan leda till att känslig information och data kompromitteras, stjäls eller krypteras. Med hjälp av maskininlärning kan vi automatisera identifieringen och snabbare upptäcka nuvarande och kommande phishingförsök genom att känna igen mönster baserat på historisk och ny data.

Förberedelse och Data

Datasetet kommer från UCI Machine Learning Repository med ID 327. Det innehåller 11,055 webbsidor med olika features som beskriver varje webbsida. Alla webbsidor i datasetet är klassificerade med -1 (phishing) och 1 (legitimate). Datasetet är relativt balanserat med 44% phishing-sidor och 56% legitima sidor.

Datan delades upp i 80% träningsdata och 20% testdata. Träningsdatan används för att modellen ska lära sig mönster, medan testdatan nyttjas för att utvärdera modellens prestanda på nya och tidigare osedda sidor.

Metod

Random Forest är den maskininlärningsmodell som valdes för detta projekt eftersom den är mer robust än andra modeller och fungerar bra för klassificeringsproblem. Modellen använder sig av 100 beslutsträd som tillsammans röstar om den slutgiltiga klassificeringen.

Träningsdatan består av 8,844 webbsidor och testdatan består av 2,211 webbsidor. Random Forest generaliserar väl, vilket innebär att modellen fungerar bra även på ny data som den inte har sett tidigare.

Resultat

Modellen uppnådde en accuracy på 96.61% på testdatan, vilket innebär att den klassificerade 2,136 av 2,211 sidor korrekt. 75 sidor var felklassificerade.

För att analysera vilka typer av fel modellen gör användes en confusion matrix, vilket gjorde resultaten mer förståeliga och tydliga:

- **False Negatives (48 st):** Phishing-sidor som modellen felaktigt klassificerade som legitima. Detta är det farligaste felet eftersom användare kan besöka dessa sidor och bli hackade.
- **False Positives (27 st):** Legitima sidor som felaktigt klassificerades som phishing. Detta är mindre allvarligt men kan innebära att användare inte kan komma åt vissa säkra sidor.

Diskussion och Slutsats

En accuracy på 96.61% är ett bra resultat, men vad innebär detta i praktiken? Även med hög accuracy kan ett enda lyckat phishing-angrepp kompromettera en hel IT-miljö och orsaka stora skador.

En annan stor risk är att phishing-tekniker utvecklas ständigt och i rasande takt. Min analys är att modellen behöver tränas kontinuerligt och eventuellt autonomt. Jag tror att man bör använda sig av flera modeller med olika inlärningssätt för att få en högre accuracy och jämnare träffsäkerhet. Dessutom krävs mer träningsdata och justeringar för att minska false negatives, även om det innebär fler false positives.

Sammanfattningsvis är maskininlärning ett användbart verktyg för att skydda mot phishing, men det kräver löpande underhåll och uppdateringar.