Rapport de stage 3A: Machine Learning appliqué à la cybersécurité

Mathieu Thomassin

2024-08-31

Download PDF

Table des matières

- Titre du stage: Machine Learning appliqué à la cybersécurité

Introduction

L'Insee a renforcé récemment ses capacités opérationnelles en terme de cybersécurité via notamment la création de son SOC (Security Operation Center) en septembre 2023. Devant la quantité de données et leur diversité et face aux évolutions des techniques et tactiques des attaquants, les méthodes de détection de cyberattaques peuvent avoir des limites. L'application d'algorithmes de Machine Learning peut aider les analystes SOC à repérer des attaques. Se déroulant au sein de l'équipe SOC construite depuis peu, le stage va permettre d'appliquer des techniques de Machine Learning et de deep learning sur des jeux de données réelles, afin de participer à la détection d'incidents de sécurité.

- Présentation de l'organisation : L'Insee a renforcé récemment ses capacités opérationnelles en terme de cybersécurité via notamment la création de son SOC (Security Operation Center) en septembre 2023. Cette équipe est répartie entre les sites de la DR de Nantes et la DR de Metz et a pour objectif de renforcer la sécurité du système d'information (SI).
- Contexte général du stage: Devant la quantité de données et leur diversité et face aux évolutions des techniques et tactiques des attaquants, les méthodes de détection de cyberattaques peuvent avoir des limites. L'application d'algorithmes de Machine Learning peut aider les analystes SOC à repérer des attaques. Le stage avait donc pour but de permettre d'appliquer des techniques de Machine Learning et de deep learning sur des jeux de données réelles, afin de participer à la détection d'incidents de sécurité.
- Objectifs du Stage :
- Objectifs principaux: Il s'agissait d'appliquer des techniques de détection de requêtes malveillantes arrivant dans le SI de l'Insee. Les requêtes arrivant au sein du SI peuvent être centralisées par un SIEM (Security information and event management).
- Résultats attendus:
- Offrir un service s'ajoutant au SIEM permettant d'identifier des requêtes malveillantes.
- Pouvoir comparer différents meilleurs modèles (au sens d'une recherche dans les hyper paramètres) entre eux
- Favoriser les bonnes pratiques du MLOps: reproductibilité, contrôle de version, automatisation, surveillance, collaboration
- Étendre la méthodologie à des jeux de données publics différents
- Explorer les algorithmes de détection d'anomalie

- Importance de la mission pour l'Insee: à titre d'exemple, l'Insee détient l'application Elire. Une attaque réussie sur cette application perturberait le bon déroulement de la vie démocratique.

Contexte général du stage

Ce stage s'inscrit pleinement dans le parcours que j'ai progressivement construit tout au long de ma carrière à l'Insee. Débutant il y a plusieurs années sur un poste de contrôleur-programmeur, j'ai pu apprécier les cours d'informatique et de statistiques pendant mon parcours à l'Ensai. J'y ai suivi les options menant progressivement à se spécialiser jusqu'au master en Informatique et traitement des données.

Au cours de ma préparation et de l'obtention de la qualification d'analyste, ma curiosité pour la sécurité des systèmes d'information s'est particulièrement développée. L'analyse des requêtes au sein d'un réseau soulève de nombreuses questions : volume des données, méthode de traitement statistique avec le machine learning, rapidité et efficacité de ce traitement. C'est un domaine où les techniques de machine learning présentent un intérêt certain.

Les enjeux de la cybersécurité peuvent se révéler particulièrement lourds, comme on peut le découvrir dans les journaux pour de nombreuses organisations. Ce stage au SOC représente une opportunité unique de confronter ces intérêts théoriques à des problématiques concrètes de cybersécurité, renforçant ainsi mes compétences et ma compréhension dans un domaine en pleine expansion.

Importance de la cybersécurité et du machine learning dans ce domaine

La cybersécurité et le machine learning sont cruciaux pour les entreprises aujourd'hui pour plusieurs raisons. D'une part, la cybersécurité est essentielle pour protéger les données sensibles contre les cyberattaques qui sont de plus en plus sophistiquées et fréquentes. Les conséquences d'une faille de sécurité peuvent être dévastatrices pour l'Insee, incluant des pertes de données, des dommages à la réputation et la crédibilité de l'Institut, et des impacts sur les utilisateurs.

D'autre part, le machine learning offre des outils puissants pour détecter et prévenir les menaces en temps réel. Grâce à ses capacités de traitement et d'analyse de vastes quantités de données, le machine learning peut identifier des modèles et des anomalies que les méthodes traditionnelles pourraient manquer. Par exemple, la DSI a déployé sur le poste de chaque agent HarfangLab, un outil de sécurité informatique reposant sur des modèles de machine learning pour détecter la présence de logiciels malveillants, des malware, et permettant d'isoler le poste compromis.

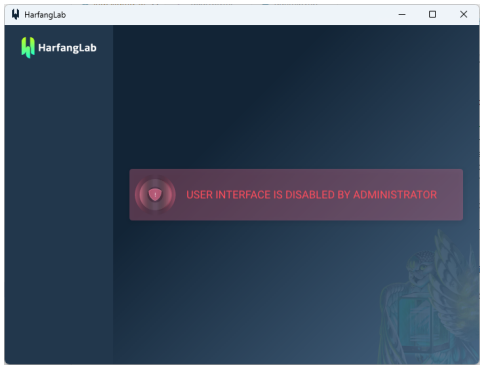

Figure 1: HarfangLab, déployé sur le poste de travail

Le machine learning permettrait également de réduire le travail des analystes du SOC en classifiant automatiquement les requêtes comme bégnines ou malveillante.

Objectifs du stage

En débutant ce stage, j'avais pour but de pouvoir réaliser des objectifs à la fois académiques, en lien avec mes intérêts développés lors de la scolarité, ainsi que des objectifs professionnels, me permettant de m'insérer au mieux dans la structure de l'Insee pour débuter ma carrière d'attaché.

1. **Alignement avec mes objectifs de carrière** : Ce stage présentait une opportunité unique de développer des compétences à la fois en statistique et en informatique. De plus, au lieu de se dérouler dans un service de développement, il a eu lieu dans un service de production, ce qui m'a permis de découvrir une autre perspective de l'informatique par rapport à l'informatique de type "data science" pratiquée à l'Ensai. En renforçant mes compétences en machine learning et en informatique, j'ai donc pu acquérir de l'expérience pour de futurs rôles au sein de l'Insee.
2. **Développement de compétences spécifiques** : Pour réaliser la tâche principale de ce stage, j'avais besoin de maîtriser des techniques de machine learning et de deep learning appliquées à la cybersécurité, ainsi que de les inscrire dans une approche de type DevOps ou MLOps. Il m'a donc fallu utiliser des outils comme Scikit-learn, MLflow, des techniques de deep learning, et utiliser le cluster Kubernetes pour entraîner et déployer un modèle avec une API.
3. **Application pratique des connaissances théoriques** : Si j'ai appris à faire du machine learning à l'école et y ai découvert des notions de DevOps, ce n'est qu'en arrivant en stage que j'ai pu découvrir l'étendue des problèmes pratiques que cela peut poser. Isolés, dans des environnements de travail bien conçus, voire seulement à travers une présentation théorique, les cours m'ont permis d'acquérir des connaissances. Cependant,

c'est seulement en travaillant sur des projets réels que j'ai pu répondre de façon pratique en piochant dans la boîte à outils de mes cours.

4. **Contribution à l'organisation d'accueil** : Débutant dans le domaine de la cybersécurité, je n'avais pas pour objectif d'apporter une solution exploitable en production. En revanche, il était essentiel pour moi de démontrer ma capacité future à prendre une position d'ingénieur capable d'envisager un problème nouveau et d'y apporter une solution pratique exploitable par une équipe. La création d'un socle d'entraînement pour un modèle de détection de requêtes malveillantes, destiné à être utilisé par des utilisateurs, sert cet objectif.
5. **Exploration des défis et opportunités** :

1. Découverte et Contexte

Premiers pas dans le stage : Enthousiasme et incertitude

Introduction à la cybersécurité : Définition et importance

Présentation de l'équipe SOC et du système d'information (SI)

Compréhension du SIEM et des logs

2. Recherche et Collecte de Données

Attente de Splunk et recherche de datasets pertinents

- Types de données : Réseau, logs, HTTP
- Sélection et préparation des datasets

Présentation des premières données obtenues

3. Choix des Problèmes et des Outils de Machine Learning

Définition des problèmes de machine learning en cybersécurité

Introduction à Scikit-learn et choix des modèles

Formation à l'interprétabilité des modèles

4. Exploration de Modèles et Outils Avancés

Essais avec XGBoost pour la classification des malwares

Découverte de MLflow et utilisation de l'API

Introduction au deep learning avec le livre "Deep Learning from Scratch" (DLFS)

Exploration des design patterns en deep learning

5. Anomaly Detection et Modèles de Classification

Détour par la détection d'anomalies : Difficultés et recentrage

Analyses simples avec KNN et clustering

Analyse des données HTTP et détection des attaques (ex. DDOS)

Focalisation sur l'analyse des URL et compréhension des attaques

6. Exploration d'Articles Scientifiques

Lecture et analyse de l'article "Machine Learning for Cybersecurity Applications" de la West Virginia University (WVU)

Étude de "A Comprehensive Review of Anomaly Detection in Web Logs" du Hasso Plattner Institute (HPI)

Analyse de l'article "CRISIS2020_EasyChair_PID_011"

Exploration de "Conf_SIN2022___SWAF" pour le développement d'un pare-feu applicatif basé sur du machine learning

7. Modélisation et Évaluation des Modèles

Modélisation des URL : KNN, SVM, regression logistique, CNN

Utilisation de ChatGPT et des GPU pour accélérer le développement

Organisation des priorités et gestion des expérimentations

8. Travail avec Splunk et Pipelines

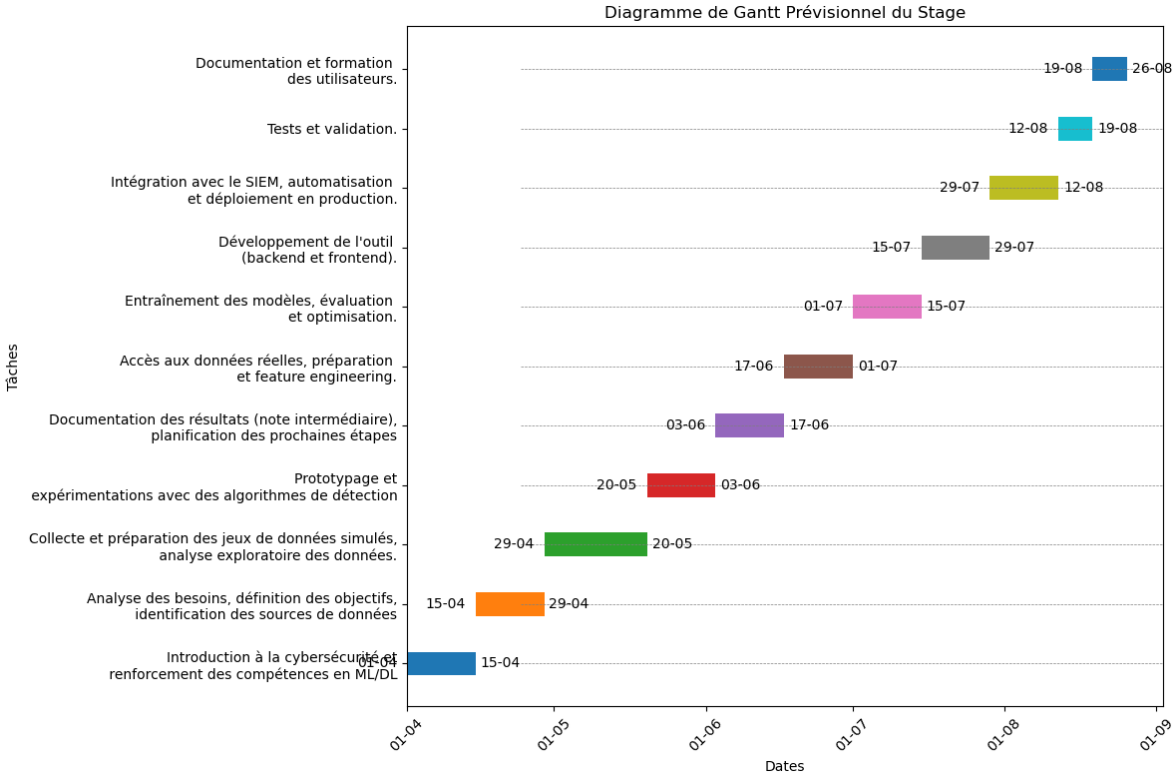

Figure 2: Gant

Détails techniques des implémentations et configurations

Bibliographie et références