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.

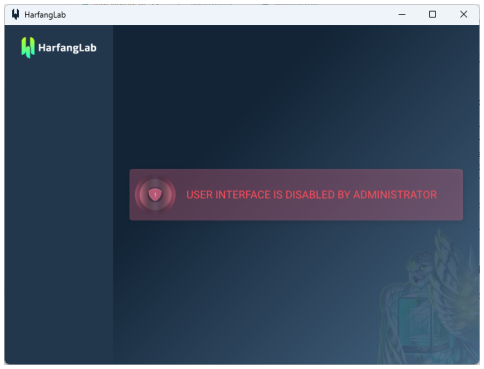

Figure 1: HarfangLab, déployé sur le poste de travail

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

- Pourquoi avez-vous défini ces objectifs spécifiques ?

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** :
4. **Contribution à l'organisation d'accueil** :
5. **Exploration des défis et opportunités** :

1. Découverte et Contexte

Premiers pas dans le stage : Enthousiasme et incertitude

- Pourquoi avez-vous ressenti ces sentiments au début ?

Introduction à la cybersécurité : Définition et importance

- Pourquoi est-il important de comprendre les bases de la cybersécurité dès le début ?

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

- Pourquoi la compréhension de l'équipe et du SI est-elle cruciale pour votre stage ?

Compréhension du SIEM et des logs

- Pourquoi le SIEM et les logs sont-ils essentiels dans le contexte de la cybersécurité ?

2. Recherche et Collecte de Données

Attente de Splunk et recherche de datasets pertinents

- Pourquoi avez-vous choisi d'utiliser Splunk et quels critères ont guidé la recherche des datasets ?
  - Types de données : Réseau, logs, HTTP
    - Pourquoi ces types de données spécifiques sont-ils importants ?
  - Sélection et préparation des datasets
    - Pourquoi cette étape est-elle critique pour la suite du projet ?

Présentation des premières données obtenues

- Pourquoi ces données sont-elles pertinentes pour votre projet ?

3. Choix des Problèmes et des Outils de Machine Learning

Définition des problèmes de machine learning en cybersécurité

- Pourquoi est-il crucial de bien définir les problèmes avant de choisir les outils ?

Introduction à Scikit-learn et choix des modèles

- Pourquoi avez-vous choisi Scikit-learn et ces modèles spécifiques ?

Formation à l'interprétabilité des modèles

- Pourquoi l'interprétabilité est-elle importante dans le contexte de la cybersécurité ?

4. Exploration de Modèles et Outils Avancés

Essais avec XGBoost pour la classification des malwares

- Pourquoi avez-vous choisi XGBoost pour ce problème ?

Découverte de MLflow et utilisation de l'API

- Pourquoi MLflow est-il utile pour votre projet ?

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

- Pourquoi ce livre et le deep learning sont-ils pertinents pour vos objectifs ?

Exploration des design patterns en deep learning

- Pourquoi est-il important de comprendre ces design patterns ?

5. Anomaly Detection et Modèles de Classification

Détour par la détection d'anomalies : Difficultés et recentrage

- Pourquoi avez-vous rencontré ces difficultés et comment avez-vous réajusté votre approche ?

Analyses simples avec KNN et clustering

- Pourquoi avez-vous choisi ces méthodes pour l'analyse initiale ?

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

- Pourquoi ces types d'attaques sont-ils une priorité dans votre analyse ?

Focalisation sur l'analyse des URL et compréhension des attaques

- Pourquoi l'analyse des URL est-elle cruciale dans la détection des attaques ?

6. Exploration d'Articles Scientifiques

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

- Pourquoi cet article est-il pertinent pour votre travail ?

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

- Pourquoi cet article est-il pertinent pour votre travail ?

Analyse de l'article "CRISIS2020_EasyChair_PID_011"

- Pourquoi cet article est-il pertinent pour votre travail ?

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

- Pourquoi cet article est-il pertinent pour votre travail ?

7. Modélisation et Évaluation des Modèles

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

- Pourquoi avez-vous choisi ces modèles spécifiques pour la modélisation des URL ?

Utilisation de ChatGPT et des GPU pour accélérer le développement

- Pourquoi ces outils ont-ils été utilisés pour accélérer le développement ?

Organisation des priorités et gestion des expérimentations

- Pourquoi cette organisation est-elle importante pour la réussite de votre projet ?

8. Travail avec Splunk et Pipelines

Traitement des données Splunk en préproduction

- Pourquoi le traitement des données en préproduction est-il nécessaire ?

Tokenization des URL et limitations

- Pourquoi avez-vous utilisé cette méthode de tokenization et quelles en sont les limites ?

Développement de pipelines et utilisation de GridsearchCV

- Pourquoi ces techniques sont-elles cruciales pour optimiser les modèles ?

Évaluation des performances des modèles

- Pourquoi une évaluation rigoureuse des performances est-elle essentielle ?

9. Expérimentations et Déploiement

Utilisation de MLFlow pour la gestion des expérimentations

- Pourquoi MLFlow est-il un choix stratégique pour gérer les expérimentations ?

Apprentissage à requêter et déployer un modèle via une API

- Pourquoi cette compétence est-elle importante dans le cadre de votre stage ?

Déploiement sur un cluster Kubernetes et gestion du preprocessing

- Pourquoi Kubernetes et le preprocessing sont-ils importants pour le déploiement ?

Introduction à Metaflow (Netflix) et tests préliminaires

- Pourquoi avez-vous exploré Metaflow et quels avantages cela apporte-t-il ?

10. Reproductibilité et MLOps

Importance de la reproductibilité et utilisation des cours de l'ENSAE

- Pourquoi la reproductibilité est-elle critique dans le domaine du machine learning ?

Réalisation d'un projet MLOps :

- Pourquoi ces aspects spécifiques du MLOps sont-ils importants ?

Exploration de Spark pour le calcul et le streaming

- Pourquoi Spark est-il un outil pertinent pour vos objectifs ?

Conclusion

Bilan du stage et accomplissements

- Pourquoi ces réalisations sont-elles significatives ?

Perspectives futures : Améliorations et extensions possibles

- Pourquoi ces améliorations sont-elles envisagées ?

Remerciements et réflexions personnelles

- Pourquoi ces personnes/expériences sont-elles importantes pour vous ?

Annexe

Ressources supplémentaires : Notebooks, articles, tutoriels, dépôts Github

- Pourquoi ces ressources sont-elles utiles ?

Gantt

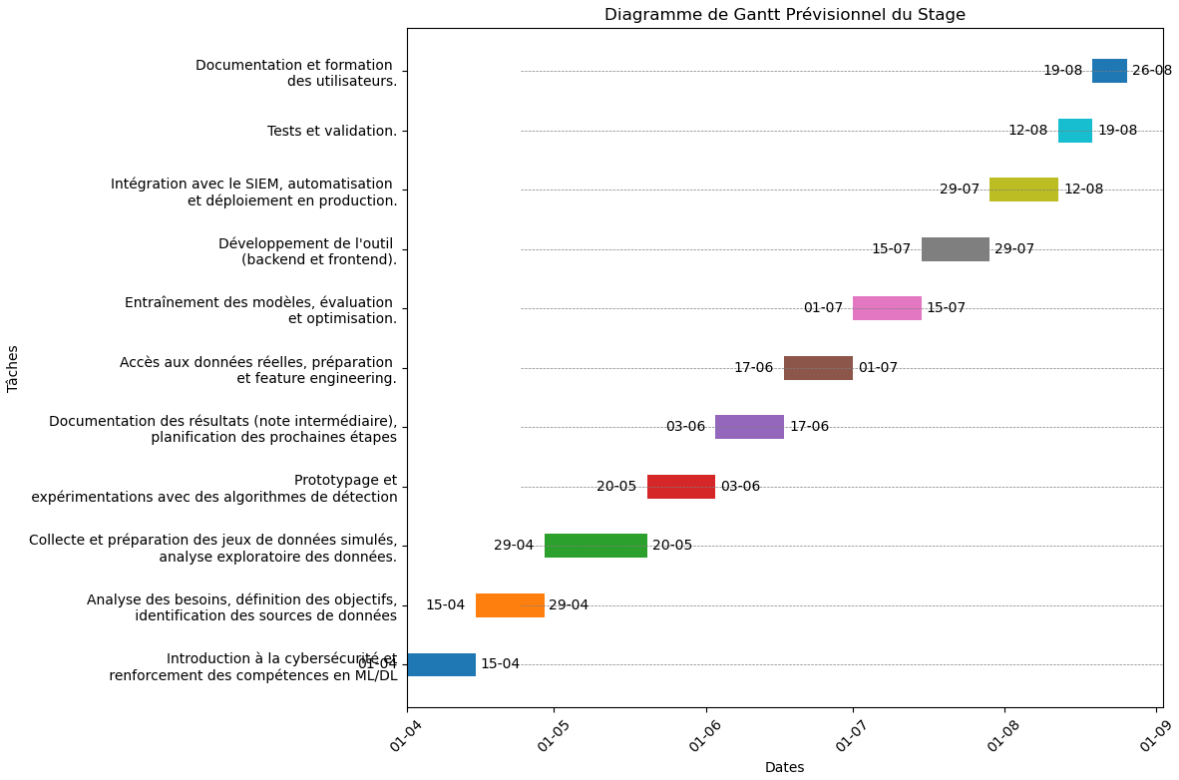

Figure 2: Gant

Détails techniques des implémentations et configurations

- Pourquoi ces détails techniques sont-ils importants pour la compréhension de votre travail ?

Bibliographie et références

- Pourquoi ces références ont-elles été utilisées ?