Rapport de stage 3A

Mathieu Thomassin

2024-08-31

Download PDF

Table des matières

- Introduction
- Découverte et Contexte
- Recherche et Collecte de Données
- Choix des Problèmes et des Outils de Machine Learning
- Exploration de Modèles et Outils Avancés
- Anomaly Detection et Modèles de Classification
- Exploration d'Articles Scientifiques
- Modélisation et Évaluation des Modèles
- Travail avec Splunk et Pipelines
- Expérimentations et Déploiement
- Reproductibilité et MLOps
- Conclusion
- Annexe

Introduction

Contexte général du stage

- Pourquoi ce stage est-il pertinent dans votre parcours académique et professionnel ?

Importance de la cybersécurité et du machine learning dans ce domaine

- Pourquoi la cybersécurité et le machine learning sont-ils cruciaux pour les entreprises aujourd'hui ?

Objectifs du stage

- Pourquoi avez-vous défini ces objectifs spécifiques ?

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 ?

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 ?