NOVARTIS SERVER HACK DATASET

By Nawfal Ahmed, Ahmed Tariq, Zeehan Rashid

Network Security Project Phase 01 Report November 2023
Supervisor: Faisal Iradat

DEPARTMENT OF COMPUTER SCIENCE - IBA
FOURTH YEAR PROJECT
CSE455-1/7675

NOVEMBER 2023

DECLARATION

This report entitled

NOVARTIS SERVER HACK DATA-SET PROJECT PHASE 01 REPORT

Was composed by our team and is based on our own work. Where the work of the others has been used, it is fully acknowledged in the text and in captions to table illustrations. This report has not been submitted for any other qualification.

Name: *Nawfal Ahmed khan [Team - Lead]*

Signed: . . . .

Date: 17th November 2023

Nawfal, Ahmed, Zeehan
Date: 17th Novembor 2023
Computer Science Students

Abstract

The frequency and sophistication of cyberattacks necessitate accurate prediction of server hacking incidents for proactive defense and risk mitigation. We propose a machine learning model using the Novartis data-set to predict server hacking incidents based on anonymized logging parameters. Our model involves data preprocessing, feature engineering, model selection and training, and model evaluation. We expect the model to achieve high accuracy and generalizability, demonstrating the effectiveness of data-driven approaches in cybersecurity.

Contents

Chapter 1

Introduction

1.1 About

The Novartis data-set, a collection of real-world network traffic data, offers a unique opportunity to train machine learning models for identifying and predicting server hacking incidents. The data-set encompasses a range of network events.[1] By analyzing these data features, machine learning algorithms can uncover hidden patterns and correlations that can be used to distinguish between normal and malicious network behavior.

1.2 Survey

The Novartis data-set provided ample data to train a classification model for predicting server hacking incidents, eliminating the need for a survey. The task was objective and the data-set provided clear-cut indicators of server activity, further supporting the decision to forgo a survey. Additionally, conducting a survey would have introduced additional time and resource constraints to the project. Therefore, the project team is focusing solely on the provided data-set to develop and deploy the classification model efficiently.

1.3 Novartis Competition Poster

Below attached is the official poster by Hacker Earth for the Novartis Data Science Competition.[1]

QMUL_SEMS_Projec_Template/Pictures/Novartis Data Science Competition.JPG

Figure 1.1. POSTER FOR COMPETITION

Chapter 2

Concept

2.1 Proposed Methodology

The development of the classification model is proposed to involve several key steps.

2.1.1 Data Pre-Processing

The data will be pre-processed to remove inconsistencies and irregularities that could hinder the learning process. This will include filtering out irrelevant data, handling missing values, and normalizing numerical features.

2.1.2 Feature Engineering

Additional features will be derived from the existing data to enhance the model's predictive power. This may involve extracting statistical summaries, transforming categorical variables, and creating new features based on domain knowledge.

2.1.3 Model Selection and Training

Various machine learning algorithms will be evaluated for their suitability in predicting server hacking incidents. The selected algorithm will be trained on the preprocessed and engineered data.

2.1.4 Model Evaluation

The trained model will be evaluated on a separate set of data to assess its ability to generalize to unseen instances. This will involve calculating performance metrics such as accuracy, precision, and recall.

2.2 Expected Results and Discussion

It is anticipated that the classification model will achieve promising results, demonstrating the feasibility of predicting server hacking incidents using machine learning techniques. The model is expected to exhibit high accuracy and generalizability, indicating its potential for real-world applications.

Chapter 3

Conclusion

3.1 Conclusion

The development of a machine learning model for predicting server hacking incidents using the Novartis data-set is expected to demonstrate the effectiveness of data-driven approaches in cybersecurity. The model's ability to identify patterns and anomalies in network traffic is expected to provide a valuable tool for proactive defense and risk mitigation.

3.2 Future work

Future directions for research include exploring the use of more advanced machine learning algorithms, integrating additional data sources, and developing real-time prediction systems for real-world deployment.

Bibliography

[1] Novartis, “Novartis data science competition,” 2023. [Online]. Available: https://www.hackerearth.com/challenges/machine-learning/novartis-data-science-competition/