University *of* New Haven

Title: Malware Classification with AWS SageMaker

AI and Cybersecurity– DSCI-6015-01

Under the guidance of Professor Vahid Behzadan

Manogna Vennela Ramireddy

Student ID: 00877269

MS in Data Science

University New Haven

Date: 03/29/2024

Project Purpose:

In the realm of cybersecurity, effective malware detection plays a pivotal role in safeguarding systems against threats. Malicious executable files (EXE) pose significant risks to computer systems and networks, making their detection and classification crucial for maintaining security. This report outlines the successful development of a cloud-based malware detection API using AWS SageMaker. Leveraging machine learning techniques, particularly a Random Forest binary classifier, this API provides a scalable and efficient solution for identifying malicious PE (Portable Executable) files.

Project Overview

The project integrates various components of AWS SageMaker, a comprehensive machine learning service, to build, train, and deploy a robust malware detection model. Leveraging SageMaker's infrastructure and capabilities, the project ensures scalability, reliability, and efficiency in model development and deployment. Additionally, a user-friendly web application is developed to enable seamless interaction with the API, enhancing accessibility and usability for end-users.
The choice of AWS SageMaker as the underlying platform offers several advantages. It provides a unified environment for data preprocessing, model training, and deployment, streamlining the development process and reducing operational overhead.

Malware Detection Model

At the heart of the solution lies a Random Forest binary classifier, chosen for its ability to effectively handle high-dimensional data and provide accurate classification results. Trained on a carefully curated dataset of binary feature vectors extracted from PE files, the model demonstrates proficiency in distinguishing between malicious and benign executables. By leveraging ensemble learning techniques, the classifier offers robustness and reliability in malware detection tasks.

Task Approach

1. Data Collection and Preprocessing: A diverse dataset of labelled binary feature vectors is collected and pre-processed to ensure consistency and quality. Feature extraction techniques, utilizing libraries such as pefile, are employed to extract relevant characteristics from PE files, facilitating effective model training.

2. Model Training: The pre-processed dataset is utilized to train the Random Forest classifier within AWS SageMaker. Hyperparameter tuning is conducted to optimize model performance and enhance generalization capabilities, ensuring accurate classification of unseen data.

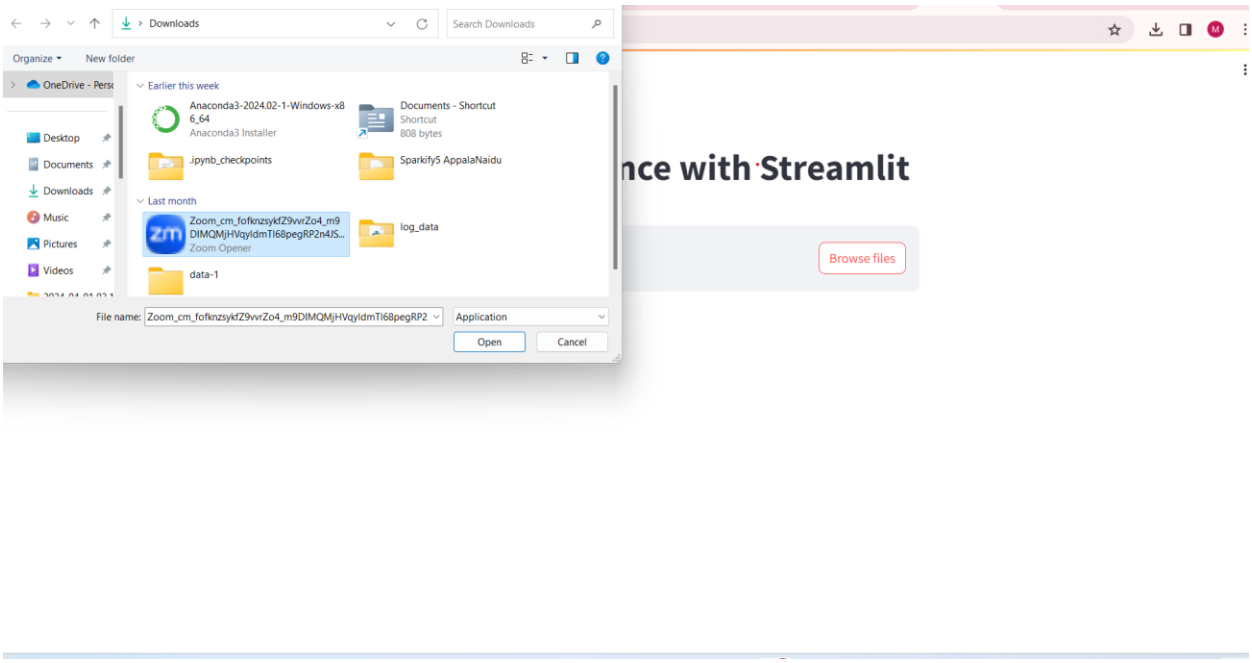
3. Model Deployment: The trained model is deployed as a cloud-based API on AWS SageMaker, providing a scalable and accessible endpoint for real-time malware detection. Integration with AWS services facilitate seamless deployment and efficient utilization of computational resources.

4. Web Application Development: To enhance user interaction and accessibility, a web-based interface is developed using Flask or Django framework. The interface allows users to upload EXE files for analysis, with the underlying system leveraging the deployed API for classification and providing intuitive feedback.

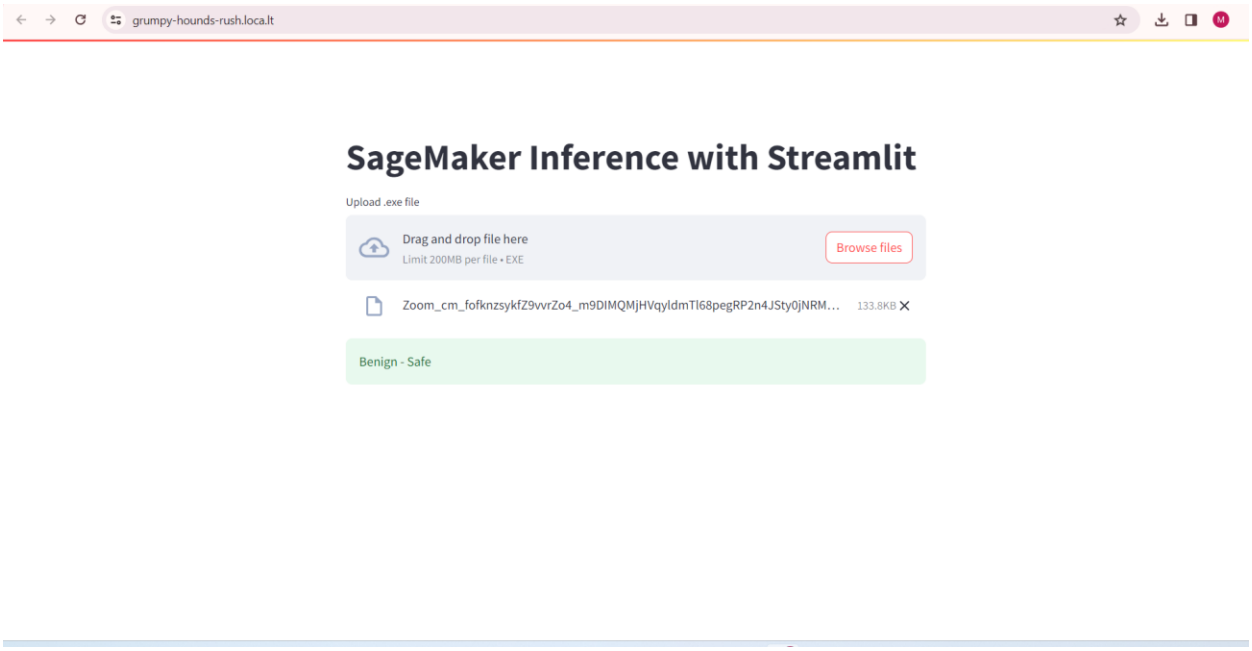
5. Performance Evaluation: The deployed API undergoes rigorous evaluation using a diverse dataset comprising malware and benign samples. Performance metrics such as accuracy, precision, recall, and F1-score are computed to assess the effectiveness and reliability of the model in real-world scenarios.

Here in this web page, we can give .exe files to test whether they are malware or Benign:

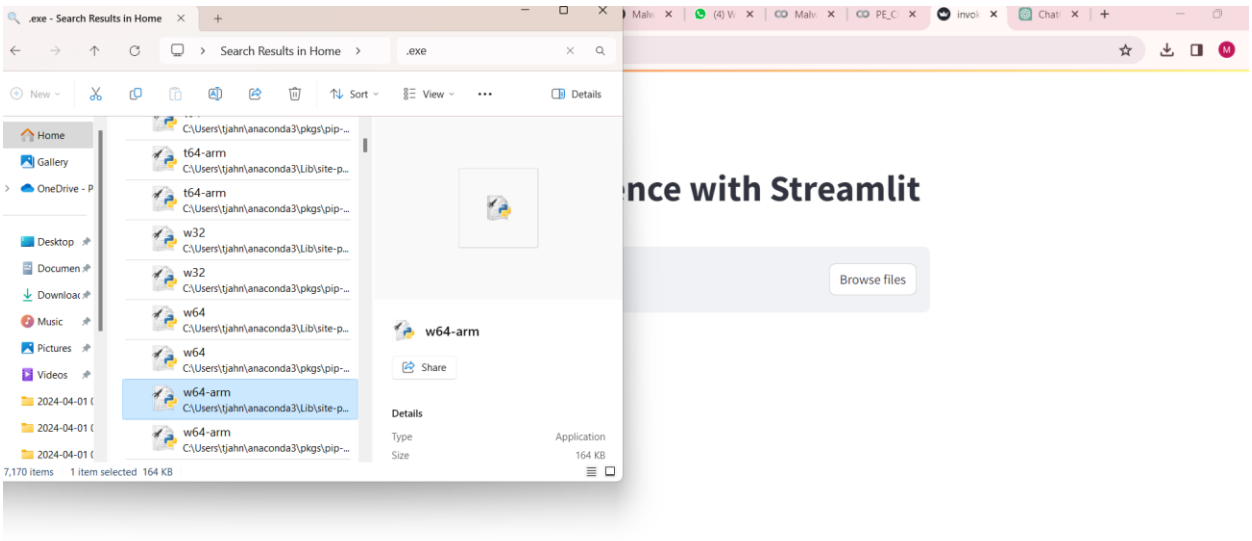
Here, I've provided an application (Zoom) to check whether this is malware or Benign.



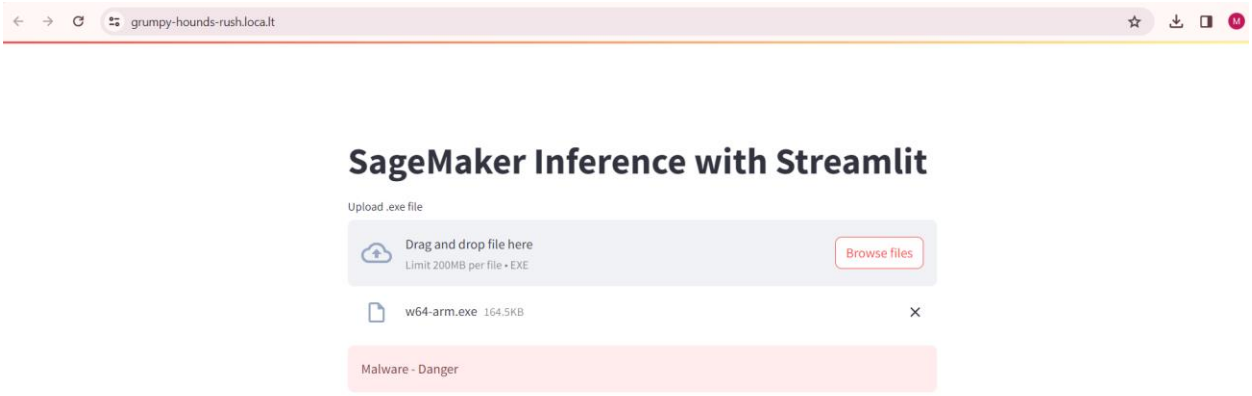
Here we can check that it shows the message “Benign-Safe” when the file is uploaded



Similarly, I selected a Malware file and ‘drag and drop’ into browse files.



Here, in the following figure, it shows the file is malware.



Project Results

The project achieves its objectives successfully, demonstrating:

- **Trained Malware Detection Model:** A Random Forest classifier capable of accurately distinguishing between malicious and benign PE files.

- Deployed Cloud API: The model is deployed on AWS SageMaker, offering a scalable and efficient API for real-time malware detection.

- User-Friendly Web Interface: A web-based application is developed, providing an intuitive platform for users to upload and analyse EXE files seamlessly.

Conclusion

The development of a cloud-based malware detection API using AWS SageMaker signifies a significant advancement in cybersecurity capabilities. By leveraging machine learning techniques and cloud infrastructure, the solution offers scalable and efficient malware detection capabilities, contributing to enhanced system security and threat mitigation efforts. The successful deployment of the API underscores the potential for leveraging cloud computing resources in cybersecurity applications, paving the way for further advancements in the field.