University of New Haven

AI and CyberSecurity
DSCI6015
Midterm Project

Jagadeesh Chandra Bose Mende

00869339

Dr. Vahid Behzadan

May 04 2024.

SUMMARY

This report outlines the successful development of a cloud-based Portable Executable (PE) static malware detection API using AWS Sagemaker. The API employs a LogisticRegression
classifier trained on a labeled dataset of binary feature vectors to classify PE files as either malicious or benign.

The project utilized AWS Sagemaker for both model construction and training, as well as for deploying the model. Additionally, a user-friendly web application was developed to allow remote users to upload their URL and assess potential threats. Python was the primary language used for the project, with machine learning libraries such as sklearn, pefile, and nltk employed for model creation and implementation.

INTRODUCTION

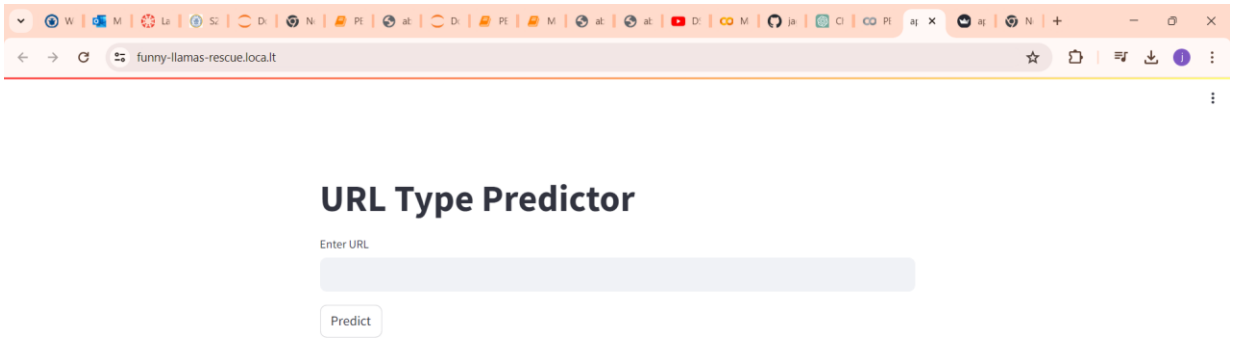
Executable files, known as PE files, are crucial components of Windows operating systems, serving as a standardized format for storing executable code and associated data. These files contain essential information required for program execution, encompassing machine instructions, resources, imported libraries, and metadata. Widely utilized for applications, drivers, and dynamic link libraries (DLLs), PE files adhere to a structured layout characterized by headers that provide insights into the file's attributes, including its architecture, entry point, and section arrangement. Proficiency in understanding the PE file format is invaluable for various tasks such as software analysis, reverse engineering, and malware detection, facilitating the examination and modification of executable content.

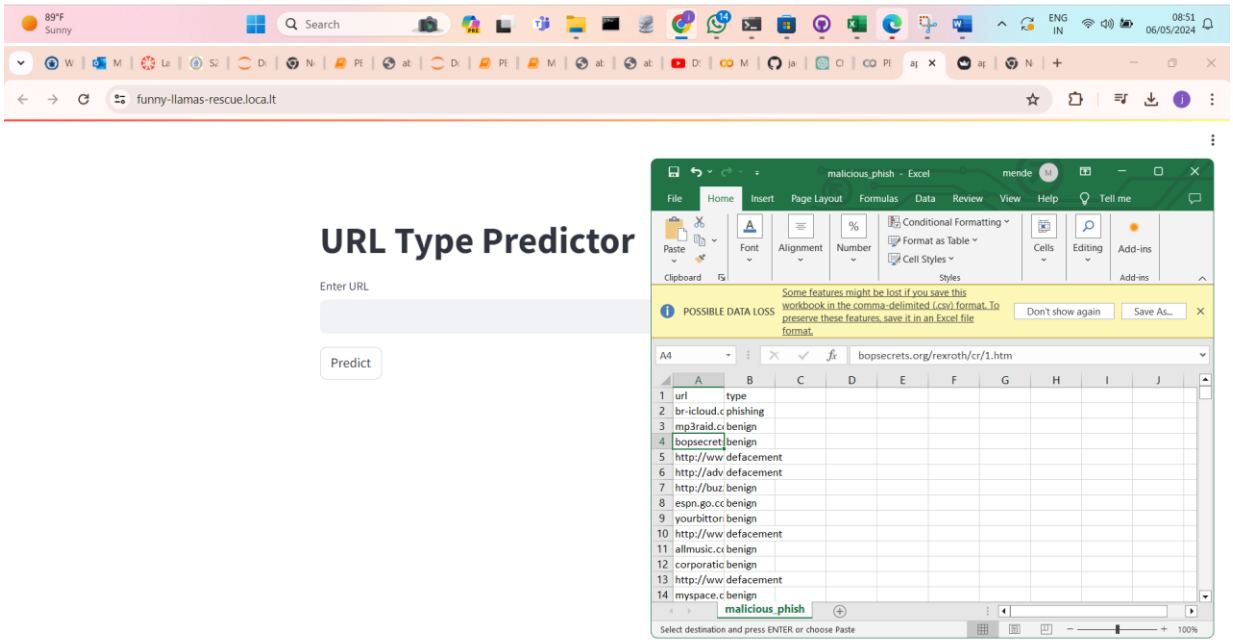
APPROACH SUMMARY:

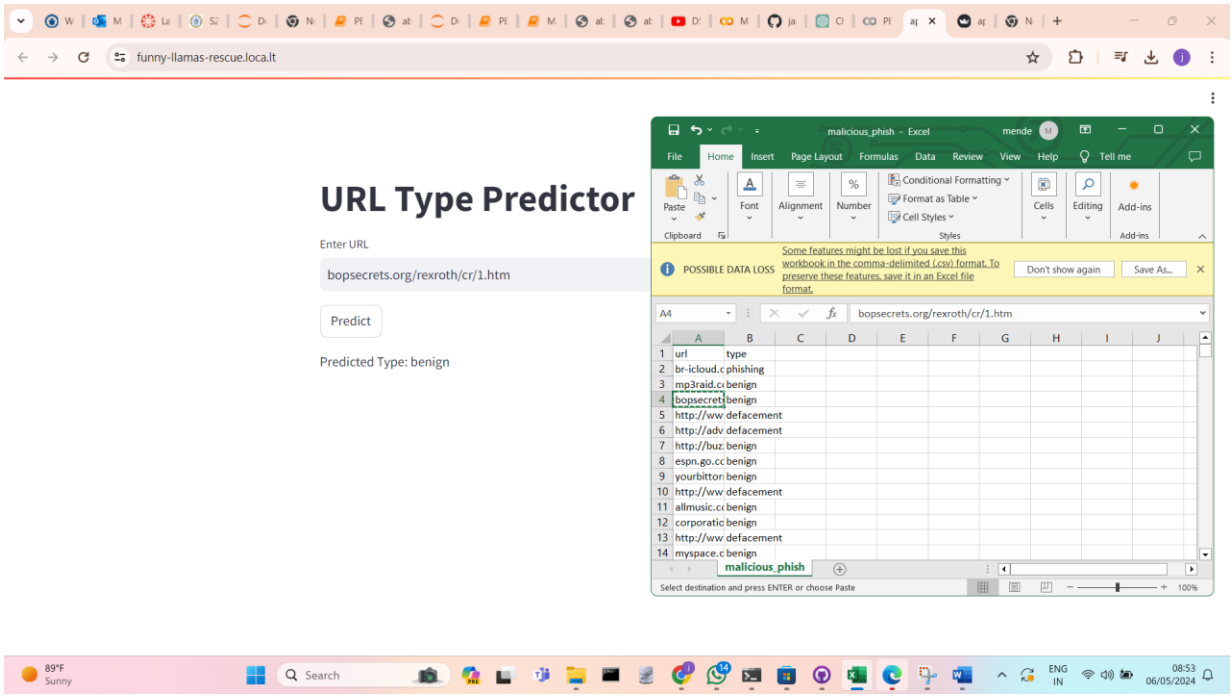
Model Building and Training: Utilizing an instance of scikit-learn version 1.2.1 on AWS Sagemaker, we trained a LogisticRegression
using a meticulously labeled dataset of binary feature vectors. Subsequently, the trained model was saved into a joblib file.

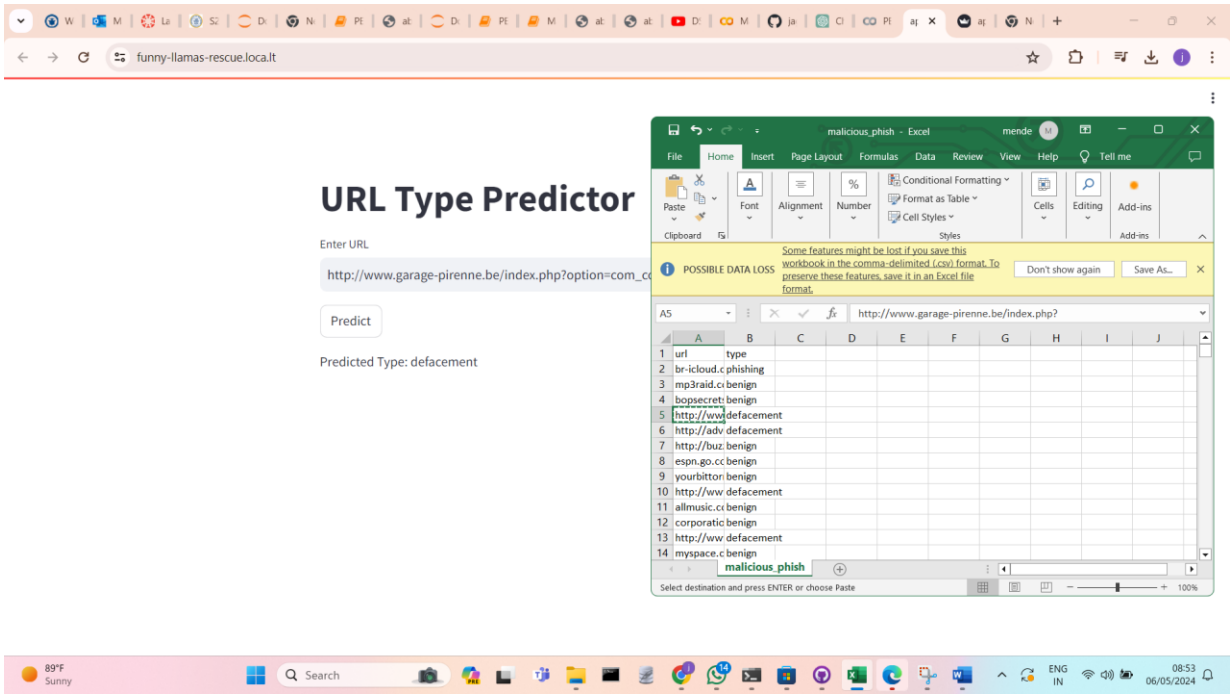
Model Deployment as a Cloud API: Leveraging Amazon Sagemaker, we deployed the trained model to create an endpoint, thereby establishing a cloudbased API for real-time predictions. The model was loaded from the saved joblib file and deployed using the sagemaker.SKLearn module. This involved configuring and creating an endpoint to host the model.

Client Application Development: A Streamlit web application was developed to furnish users with an intuitive interface. Through this application, users can upload executable URL. The application, utilizing pefile library and other pre-trained data, extracts requisite features from the uploaded .URL, converts them into JSON format, and dispatches them to the deployed API. The application then showcases the classification results, discerning between Malware (Danger) and Benign (Safe) files. For client deployment, Google Colab was utilized to initiate the Streamlit application.


funny-llamas-rescue.loca.lt
URL Type Predictor
Enter URL
Predict


89°F
Sunny
Search
ENG
IN
08:51
06/05/2024
funny-llamas-rescue.loca.lt
URL Type Predictor
Enter URL
Predict
malicious_phish - Excel
mende
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Tell me
Paste
Font
Alignment
Number
Conditional Formatting
Format as Table
Cell Styles
Cells
Editing
Add-ins
Clipboard
Styles
POSSIBLE DATA LOSS
Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.
Don't show again
Save As...
A4
bopsecrets.org/rexroth/cr/1.htm
url type
br-icloud.c phishing
mp3raid.c benign
bopsecret benign
http://ww defacement
http://adv defacement
http://buz benign
espn.go.cc benign
yourbittor benign
http://ww defacement
allmusic.c benign
corporatic benign
http://ww defacement
myspace.c benign
malicious_phish
Select destination and press ENTER or choose Paste
100%
89°F
Sunny
Search
ENG
IN
08:52
06/05/2024


funny-llamas-rescue.loca.lt
URL Type Predictor
Enter URL
bopsecrets.org/rexroth/cr/1.htm
Predict
Predicted Type: benign
malicious_phish - Excel
POSSIBLE DATA LOSS
Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.
Don't show again
Save As...
A4
bopsecrets.org/rexroth/cr/1.htm
malicious_phish
Select destination and press ENTER or choose Paste
89°F Sunny
Search
ENG IN
08:53 06/05/2024


funny-llamas-rescue.loca.lt
URL Type Predictor
Enter URL
http://www.garage-pirenne.be/index.php?option=com_c
Predict
Predicted Type: defacement
malicious_phish - Excel
POSSIBLE DATA LOSS
Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.
Don't show again
Save As...
A5
http://www.garage-pirenne.be/index.php?
malicious_phish
Select destination and press ENTER or choose Paste
89°F Sunny
Search
ENG IN
08:53 06/05/2024

RESULTS

The project has effectively delivered its desired results:

Developed a proficient Malware detection model that can classify PE files as either malicious or benign after thorough training.

Deployed the trained model on Amazon Sagemaker, establishing a real-time prediction API accessible via the internet.

Designed a web-based interface for end-users to upload files and verify their malicious status.

CONCLUSION

The obtained accuracy of 99.58% (0.99582770439428) suggests that the model accurately predicts the type of exe file most of the time, achieving a high level of correctness in its predictions.