Malicious or Benign Websites Research Report

Christopher Heimbuch – Flatiron School Data Science

As of August 22, 2024

Section I

Executive Summary

Executive Summary

As a reminder, the dataset was curated using various verified sources of benign and malicious URLs, captured in a low-interactive client honeypot to isolate network traffic. Additional tools were employed to gather further information, such as server country data through Whois.

Key Findings:

- **Countries hosting malicious websites:**
  - Spain, the United States, Czech Republic, Russia, and Great Britain are the top 5 countries hosting malicious websites.
- **Server Operating Type:**
  - Malicious websites are predominantly hosted on Apache servers, followed by Nginx. Nginx, known for handling high traffic, was originally designed to address the “c10k” problem faced by Apache.
- **Malicious to Benign Ratio:**
  - The dataset contains nearly six times more benign websites than malicious ones.
- **Hypothesis Testing Results:**
  - **IP Packets for Websites:** No significant difference was found in the mean number of IP packets generated during communication between the honeypot and the server for malicious versus benign websites.
  - **DNS Packets Generated:** A statistically significant difference was observed in the mean number of DNS packets generated between malicious and benign websites.
  - **Different Server Operating System’s TCP Packets Exchanged:** A statistically significant difference was found in the mean TCP packets exchanged for Apache, Nginx, and “Other” server types between the honeypot and the server.
- **Machine Learning Results:**
  - **Shotgun Approach:** Out of 9 tested classification models, the Random Forest, Stochastic Gradient Descent (SGD), and Gradient Boosting classifiers were selected for further testing.
  - **Hyperparameter Tuning:** Improved accuracy and performance were achieved for the Random Forest and Gradient Boosting classifiers through hyperparameter tuning, while the SGDclassifier showed poorer results. The final models chosen were Random Forest and Gradient Boosting.
  - **Class Rebalance:** SMOTE was applied to rebalance the dataset due to a severe class imbalance, initially leading to model overtraining on benign websites.
  - **Results:** The Gradient Boosting Classifier proved to be the most effective, achieving 98.4% accuracy with 98% precision, recall, and F1 score across all classes.

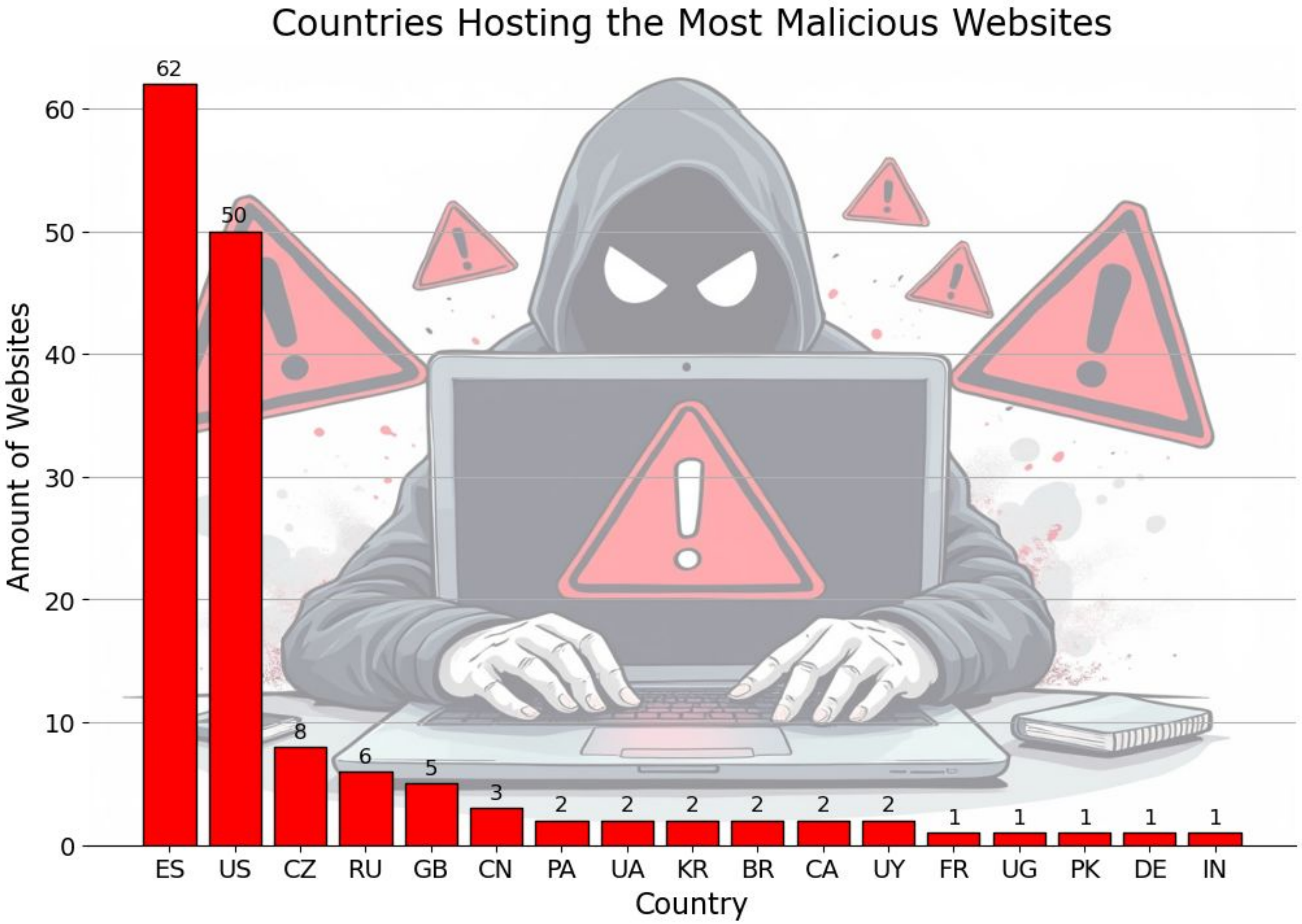
Recommendations for Companies and individuals:

This project demonstrates a practical, real-world model that can be implemented by individuals or organizations to protect against the evolving threat of malicious web activity. With the ever-present risk of contracting spyware, malware, ransomware, trojan horses, and more from visiting malicious websites, it is crucial to safeguard digital environments proactively.

Section II

Data Visualization, Inferential Analysis,
& Machine Learning Approach

Which countries host the most malicious websites?



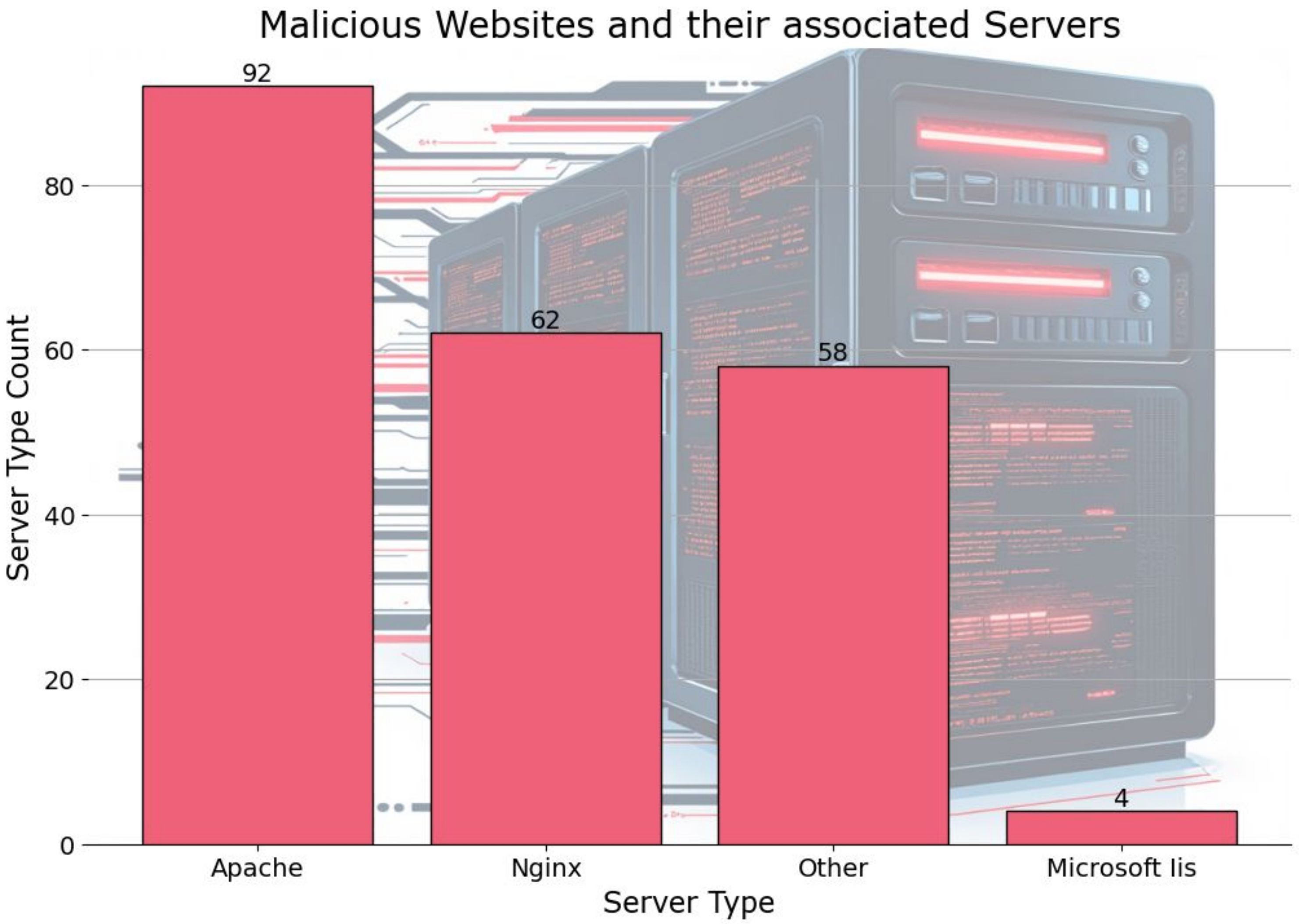
- **Top 5 Countries:** Spain, the United States, Czech Republic, Russia, and Great Britain are the leading countries hosting the highest number of malicious websites.
- **Countries with Fewer Malicious Websites:** France and India are among the countries that host the least number of malicious websites.

Of Malicious websites, what are the most common server operating system type?

- **Apache Servers:** Apache is the most prevalent server type hosting malicious websites, followed closely by Nginx. Apache dominates the web server market, holding over 33% of the market share across all sectors.

- **Technical Structure:** Apache operates using a thread-based structure, where each user request is handled by a separate thread. However, this architecture has a limitation—Apache struggles to manage more than 10,000 simultaneous connections, a challenge known as the "c10k problem."

- **Nginx Solution:** Nginx was developed in 2004 specifically to address the c10k problem. It was designed to efficiently manage high-traffic environments, making it a preferred choice for websites with heavy traffic demands.



CH CONSULTING

Inferential Analysis Summary

Data Distribution:

- **Normality Testing:** Data was tested for normality using a KDE plot and the Shapiro-Wilk test, revealing that the data was not normally distributed.

IP App Packets Generated:

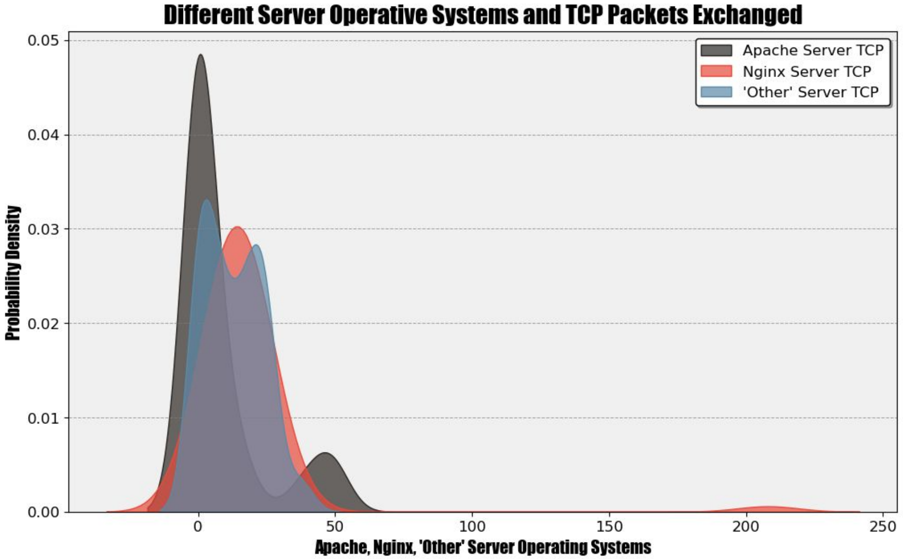
- **Research Question:** Is there a significant difference in the total number of IP app packets generated between benign and malicious websites during communication between the honeypot and the server?
- **Test:** Mann-Whitney U Test
- **Result:** No significant difference was found in the means of IP app packets generated between benign and malicious websites.

DNS Packets Generated:

- **Research Question:** Is there a significant difference in DNS packets generated between benign and malicious websites?
- **Test:** Mann-Whitney U Test
- **Result:** A significant difference was found in the means of DNS packets generated between benign and malicious websites.

Different Server Operating System TCP Packets Exchanged:

- **Research Question:** For malicious websites, is there a significant difference in the TCP packets exchanged among Apache, Nginx, and other servers?
- **Test:** Kruskal-Wallis H Test
- **Result:** The null hypothesis was rejected, indicating a significant difference in the means of TCP packets exchanged among the different server types.



***Note: All tests were ran with a 95% significance level.**

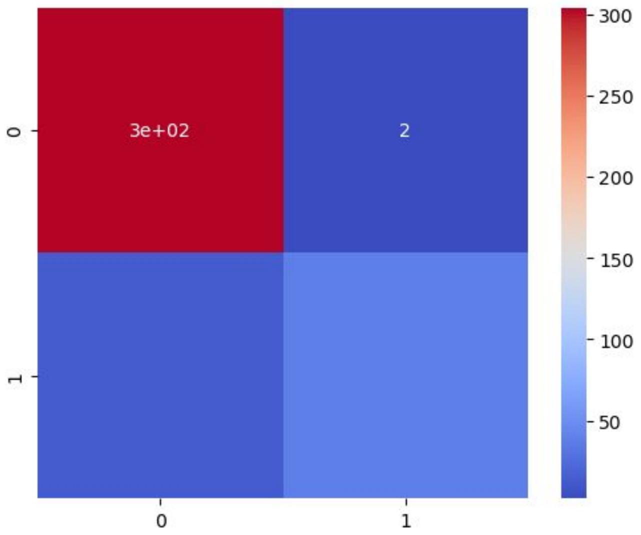
Machine Learning Approach (1 of 2)

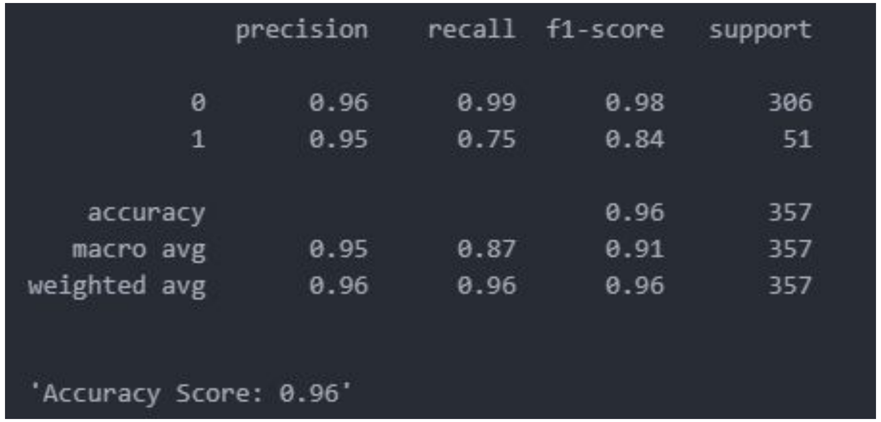
- **Data Preparation:** I undertook extensive data cleaning and preprocessing to ensure that the dataset was well-prepared for machine learning models.

- **Model Selection:** Adopting a “Shotgun” approach, I initially tested 9 different classifiers: Logistic Regression, KNeighborsClassifier, DecisionTreeClassifier, SVC, GradientBoostingClassifier, AdaBoostClassifier, RandomForestClassifier, SGDClassifier, and Multinomial Naive Bayes.

- **Top Performers:** The top three models were the GradientBoostingClassifier, RandomForestClassifier, and SGDClassifier. The Gradient Boosting Classifier achieved 96% accuracy, while the SGD and Random Forest classifiers both reached 95% accuracy.

- **Challenge of Class Imbalance:** Despite the high accuracy, the models were overfitting on benign websites and despite a decent f1 score for malicious websites, struggled to accurately predict malicious websites due to class imbalance. The recall for malicious websites was lower than desired, with the model incorrectly predicting the target class about 25% of the time.



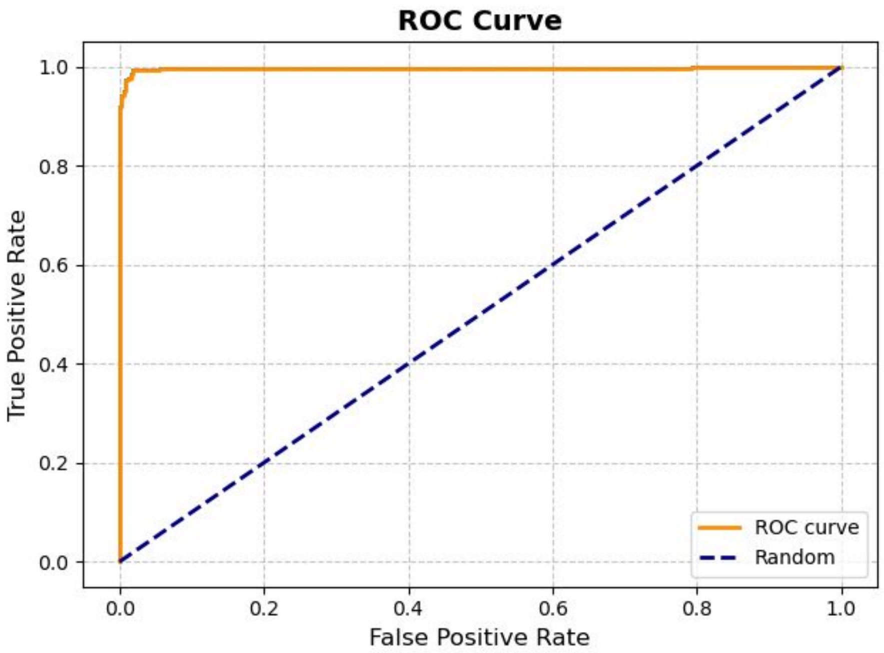


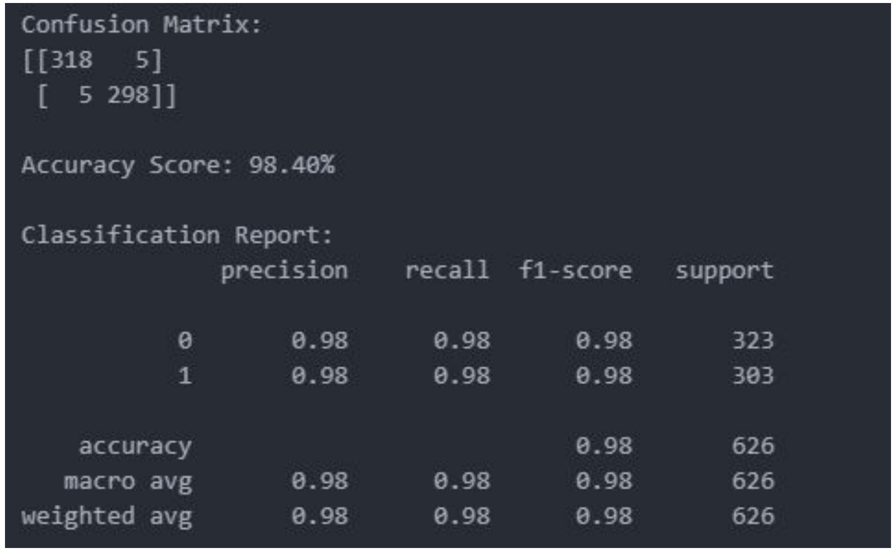
```
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       306
           1       0.95      0.75      0.84        51

    accuracy                           0.96       357
   macro avg       0.95      0.87      0.91       357
weighted avg       0.96      0.96      0.96       357


'Accuracy Score: 0.96'
```

Machine Learning Approach (2 of 2)

Addressing Class Imbalance

- **SMOTE Application:** To further balance the classes, I applied the SMOTE technique, which helped to even out the dataset, particularly enhancing the identification of malicious websites.
- **Impact on Accuracy:** The application of SMOTE resulted in a 2.4% increase in accuracy on the testing data with the Gradient Boost Classifier, achieving an accuracy score of 98.4%.
- **Model Performance:** The Random Forest model performed exceptionally well, achieving an accuracy of 98.08%, just slightly behind the Gradient Boost Classifier.
- **Prediction Strength:** The final model is highly effective at predicting whether a website is malicious—a critical outcome for this analysis—or if it is benign.



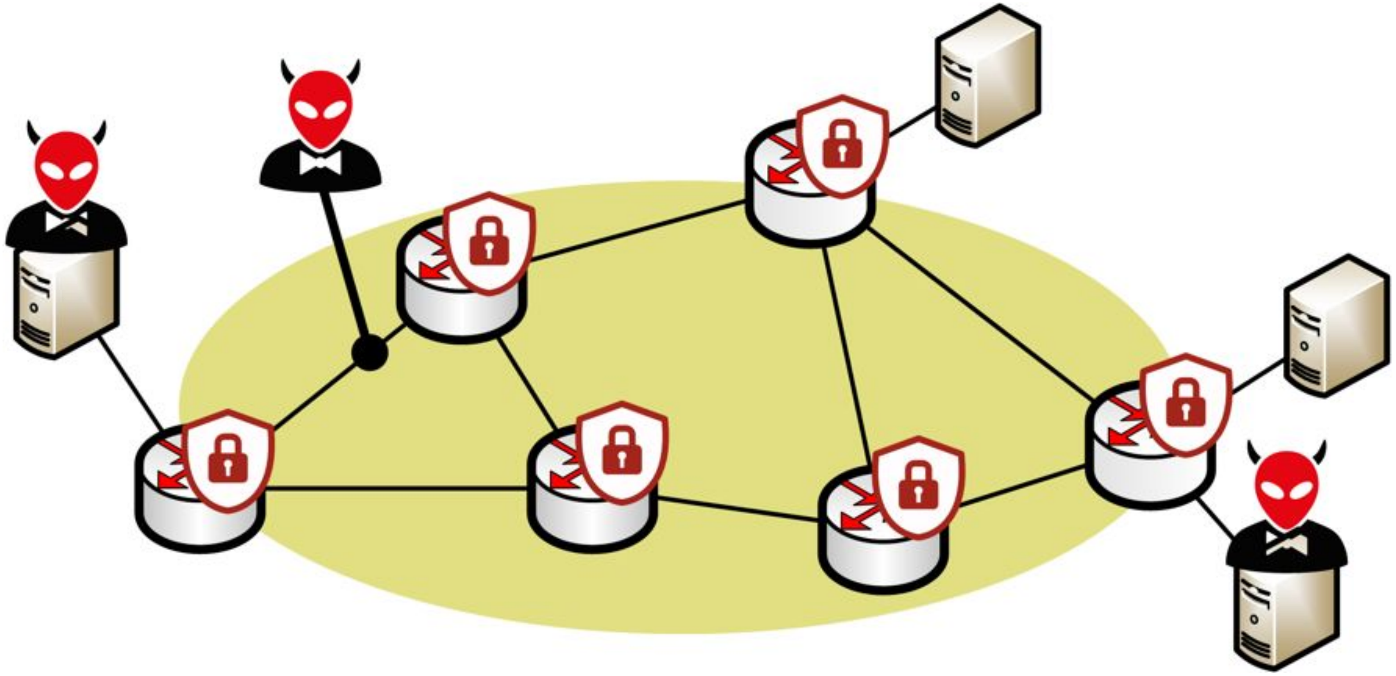


```
Confusion Matrix:
[[318   5]
 [  5 298]]

Accuracy Score: 98.40%

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       323
           1       0.98      0.98      0.98       303

    accuracy                           0.98       626
   macro avg       0.98      0.98      0.98       626
weighted avg       0.98      0.98      0.98       626
```

Recommendation

- This model provides a practical solution for individuals and organizations to safeguard against the pervasive threat of malicious online activity. The growing danger of spyware, malware, ransomware, trojan horses, and other malicious software poses a significant risk to both personal and business infrastructures. Malicious websites can cripple entire systems through ransomware attacks or secretly download malware designed to steal sensitive information, such as banking details and credentials, or even seize control of your computer.

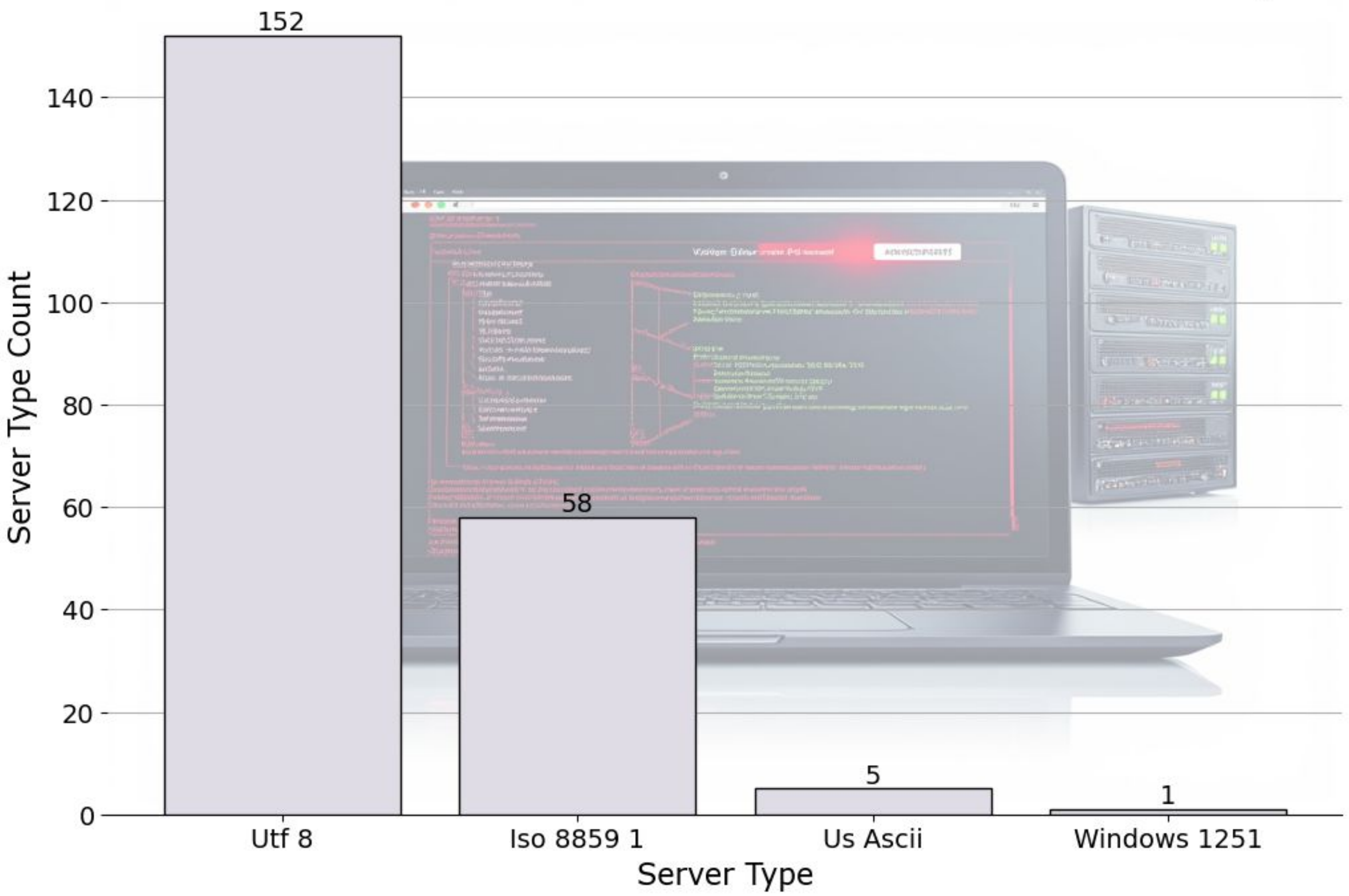
- Given the gravity of these threats, deploying this model in real-world scenarios is a promising prospect. It has the potential to offer robust protection against a wide array of cyber threats, ensuring the security and integrity of digital environments.

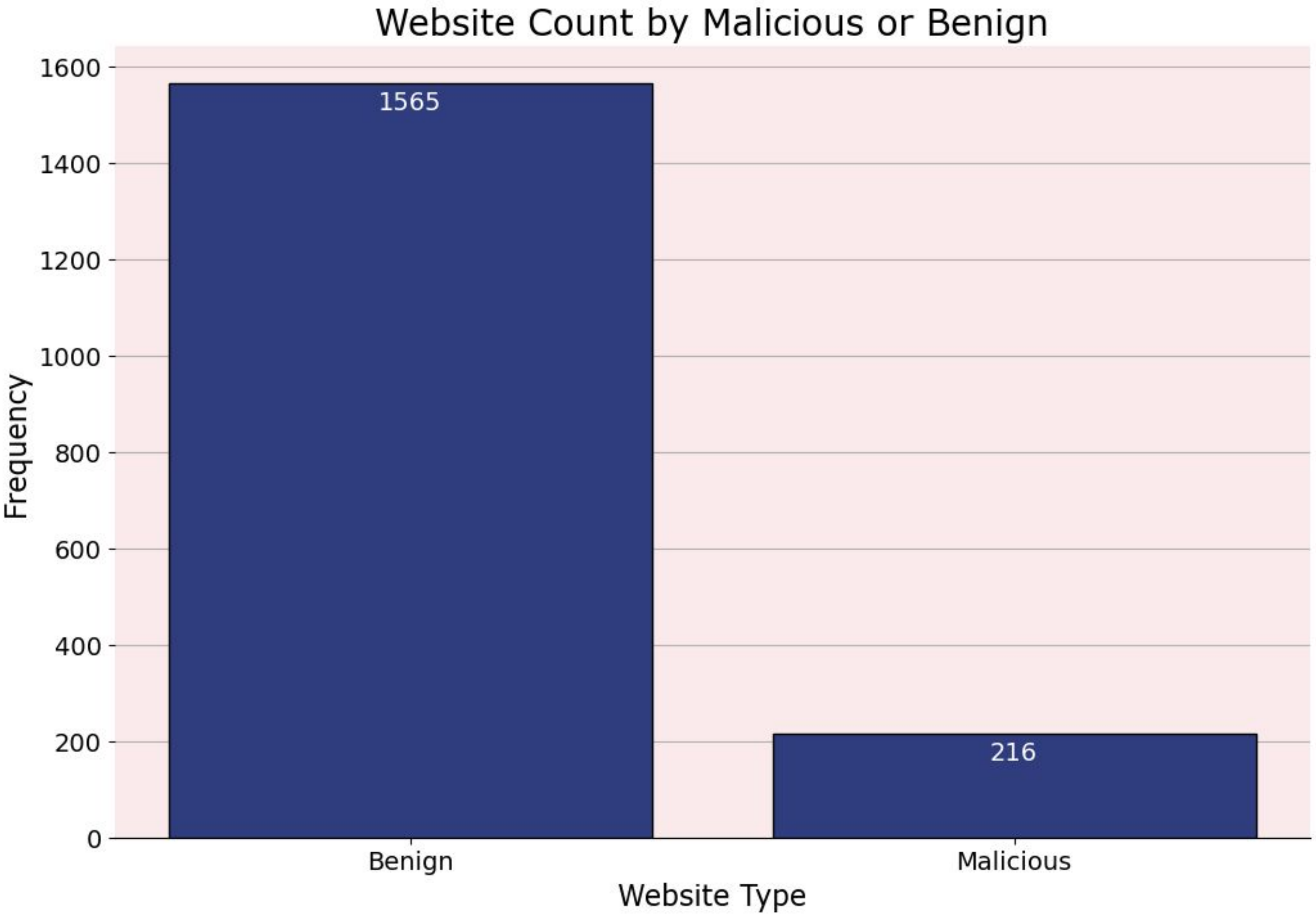


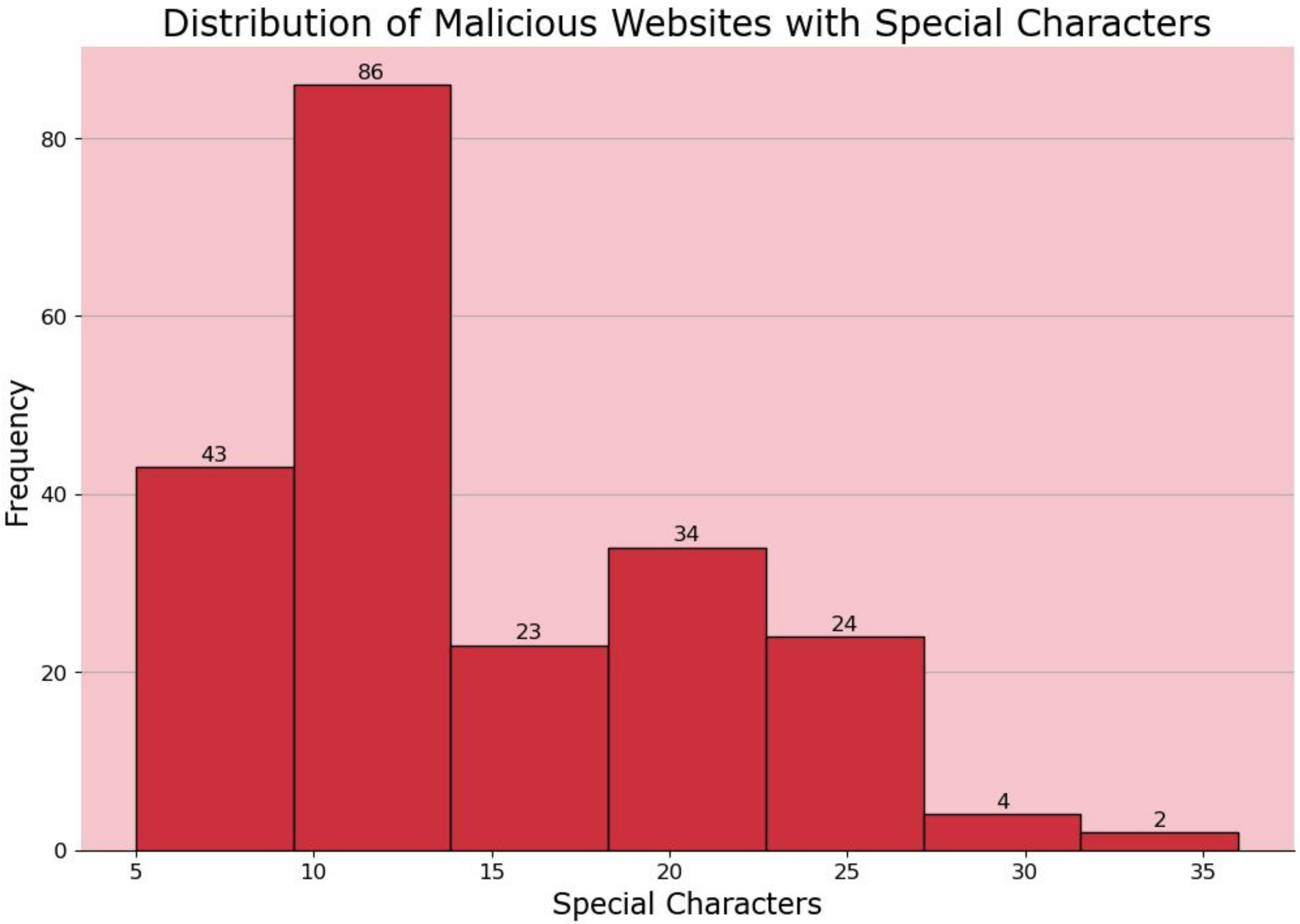
Section III

Appendix

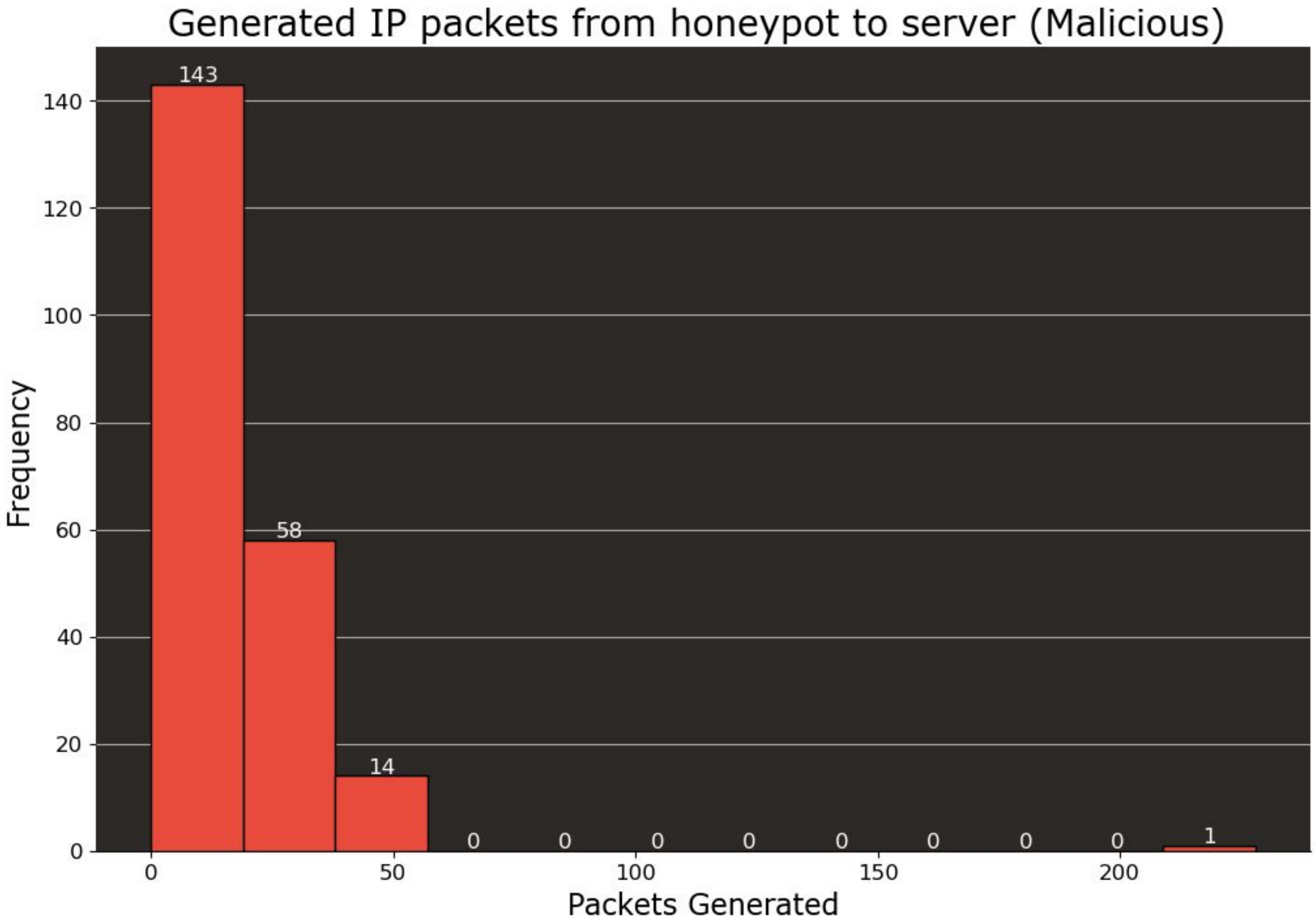
Malicious Websites and their associated Character Encoding



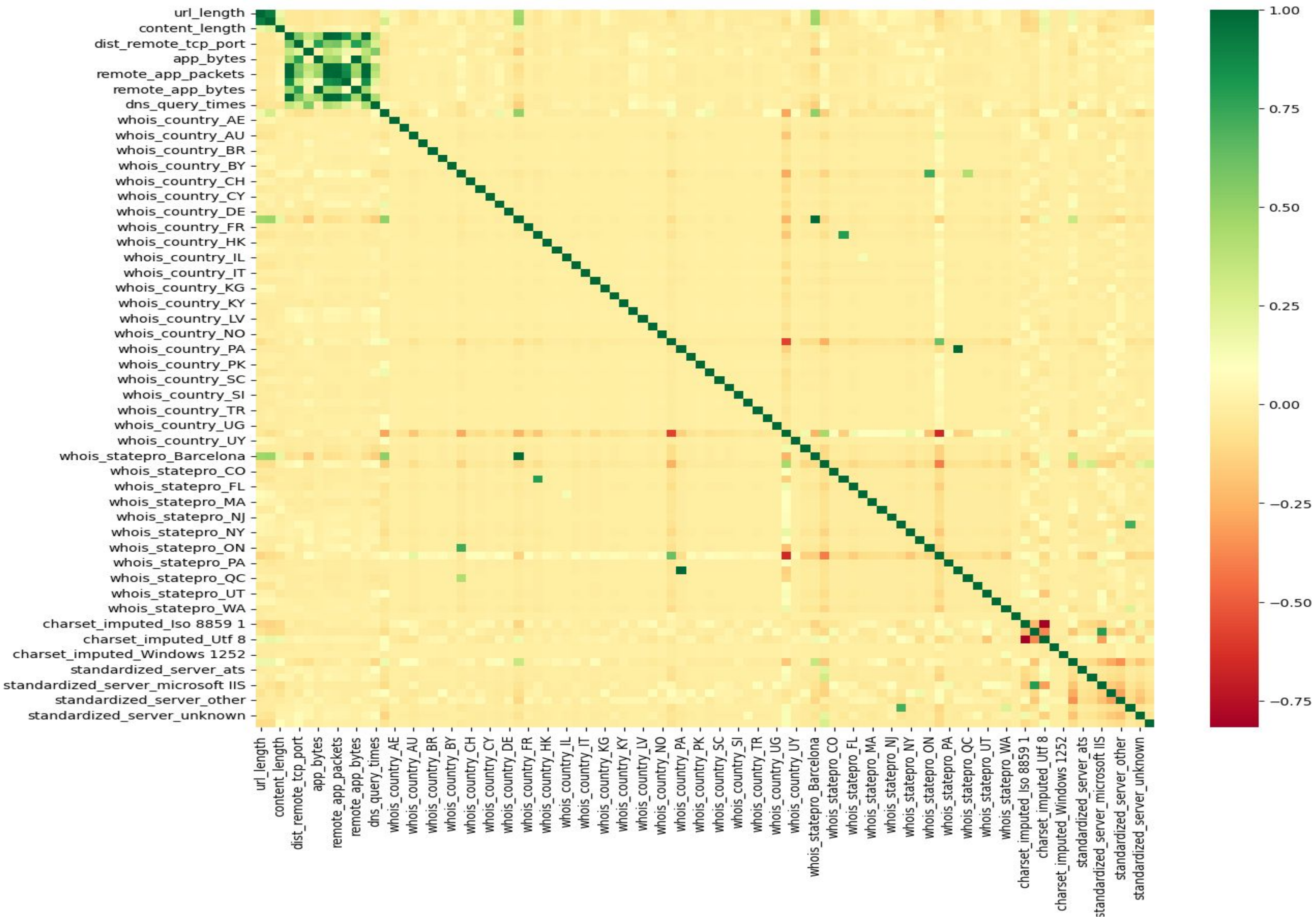

Website Count by Malicious or Benign
Frequency
1600
1400
1200
1000
800
600
400
200
0
1565
216
Benign
Malicious
Website Type



Distribution of Malicious Websites with Special Characters


Generated IP packets from honeypot to server (Malicious)
Frequency
Packets Generated
0
20
40
60
80
100
120
140
0
50
100
150
200
143
58
14
0
0
0
0
0
0
0
0
1

Correlation of Features


url_length
content_length
dist_remote_tcp_port
app_bytes
remote_app_packets
remote_app_bytes
dns_query_times
whois_country_AE
whois_country_AU
whois_country_BR
whois_country_BY
whois_country_CH
whois_country_CY
whois_country_DE
whois_country_FR
whois_country_HK
whois_country_IL
whois_country_IT
whois_country_KG
whois_country_KY
whois_country_LV
whois_country_NO
whois_country_PA
whois_country_PK
whois_country_SC
whois_country_SI
whois_country_TR
whois_country_UG
whois_country_UY
whois_statepro_Barcelona
whois_statepro_CO
whois_statepro_FL
whois_statepro_MA
whois_statepro_NJ
whois_statepro_NY
whois_statepro_ON
whois_statepro_PA
whois_statepro_QC
whois_statepro_UT
whois_statepro_WA
charset_imputed_Iso 8859 1
charset_imputed_Utf 8
charset_imputed_Windows 1252
standardized_server_ats
standardized_server_microsoft IIS
standardized_server_other
standardized_server_unknown