Malicious URL Classification

A.Y. 2023/2024

Gabriele Pongelli

AI4Cybersecurity Course

Malicious URL Classification

Goals

1. Extract some features form the URLs
2. Train a decision tree and extract the final rules
3. Compare those rules with results in the literature

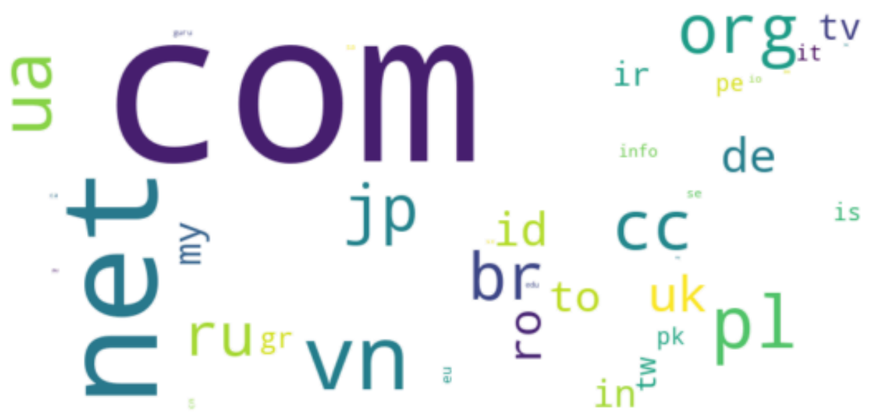
Malicious URL Classification

Extracted Features

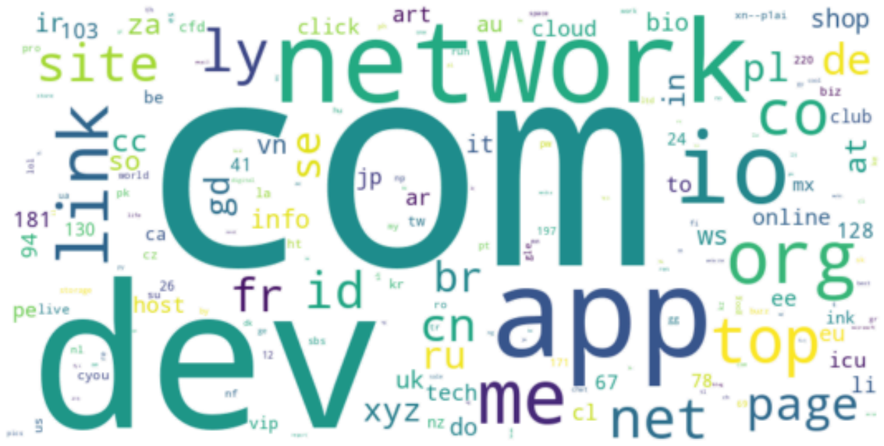
URL Length	Top-Level Domain	# of Subdomains
URL Entropy	Website Traffic	Suspicious Domain Age
# of Parameters	Use a URL Shortening Service	Has iframe
Has ‘-’ in the Domain	Has a DNS Record	Suspicious Forwardings

Data Exploration

Benign TLD

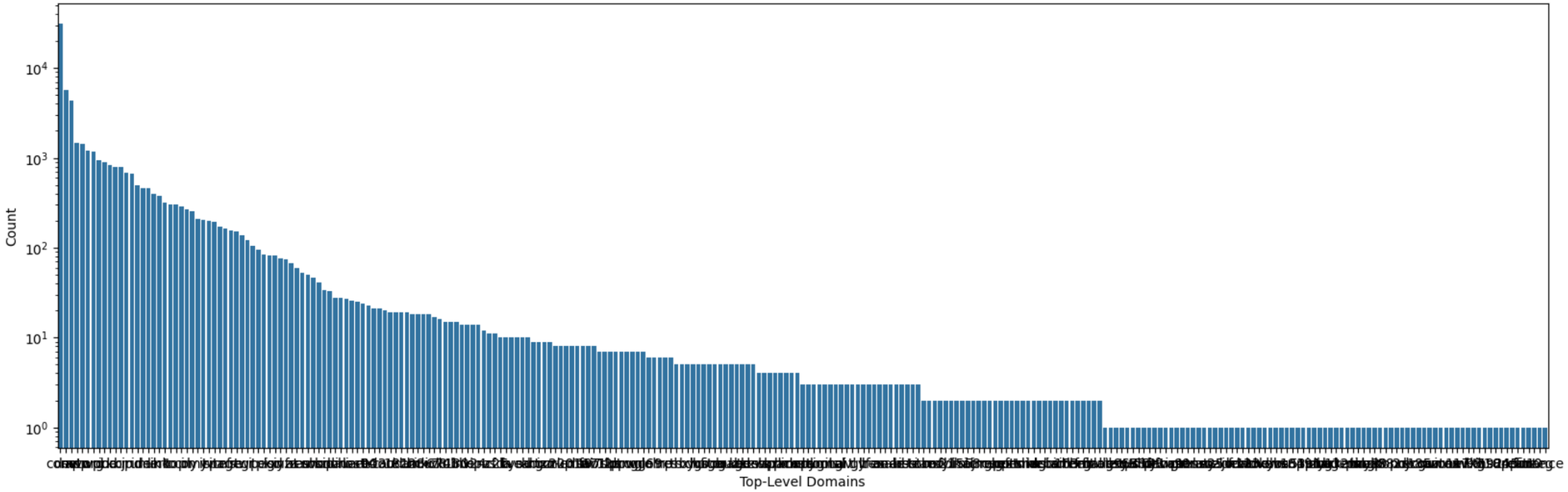


Malicious TLD



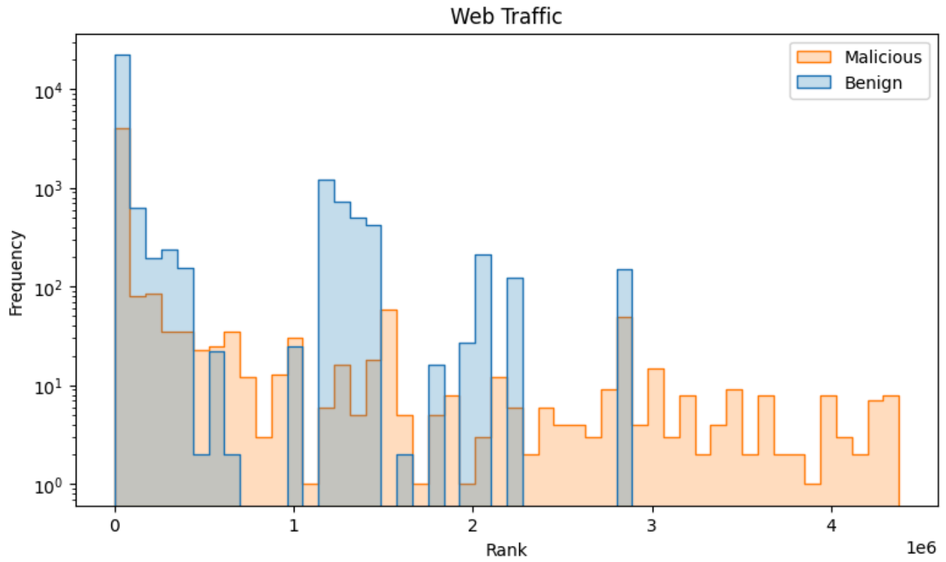
Malicious URL Classification

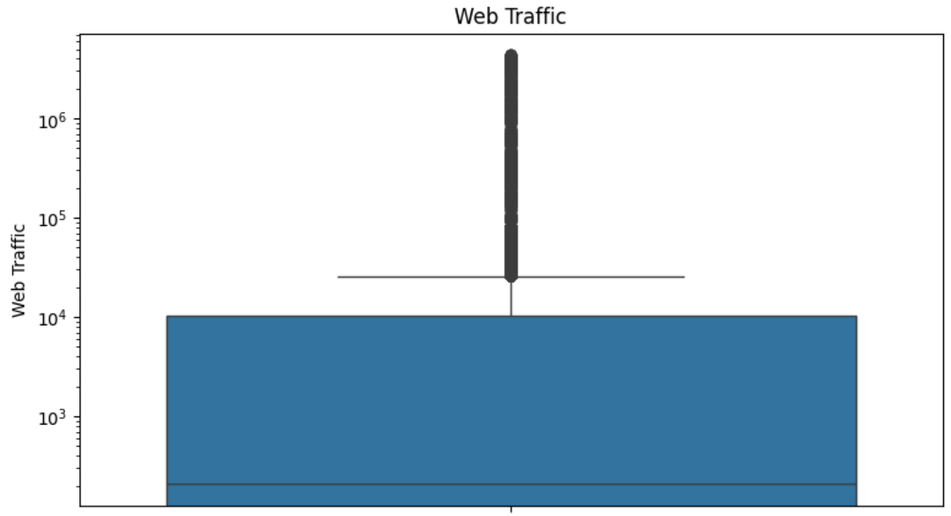
Data Exploration



Malicious URL Classification

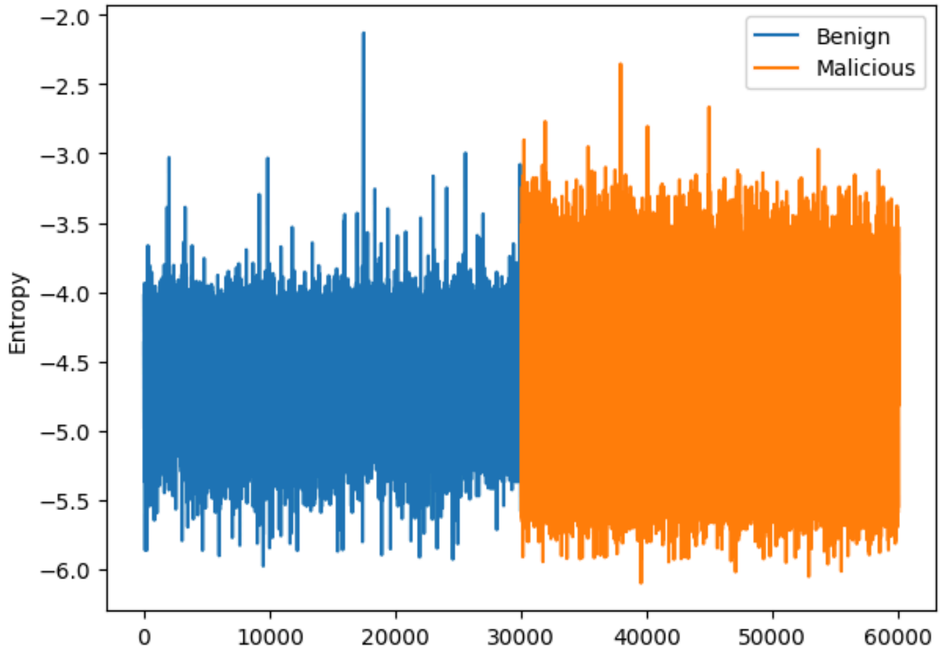
Data Exploration

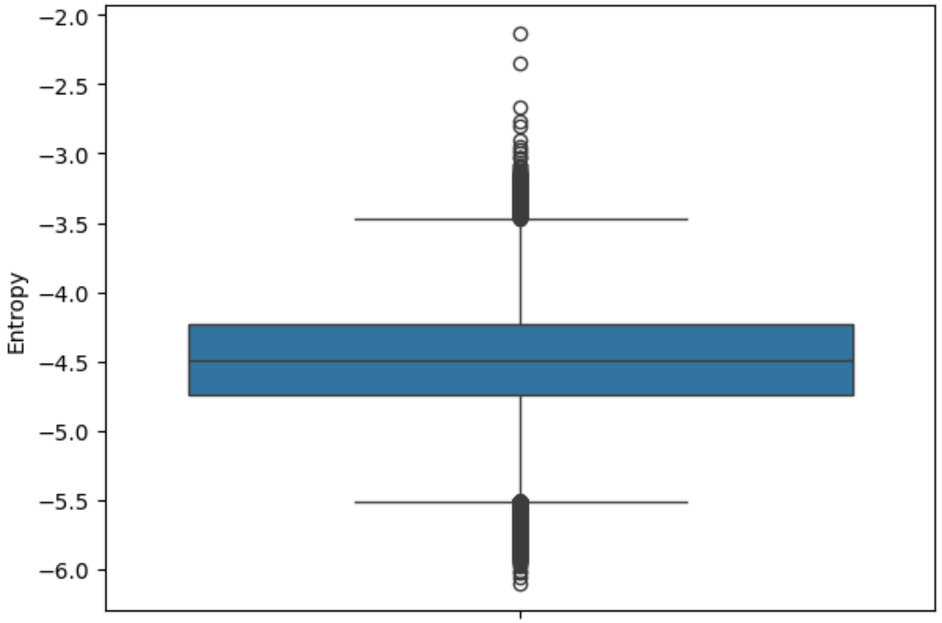




Malicious URL Classification

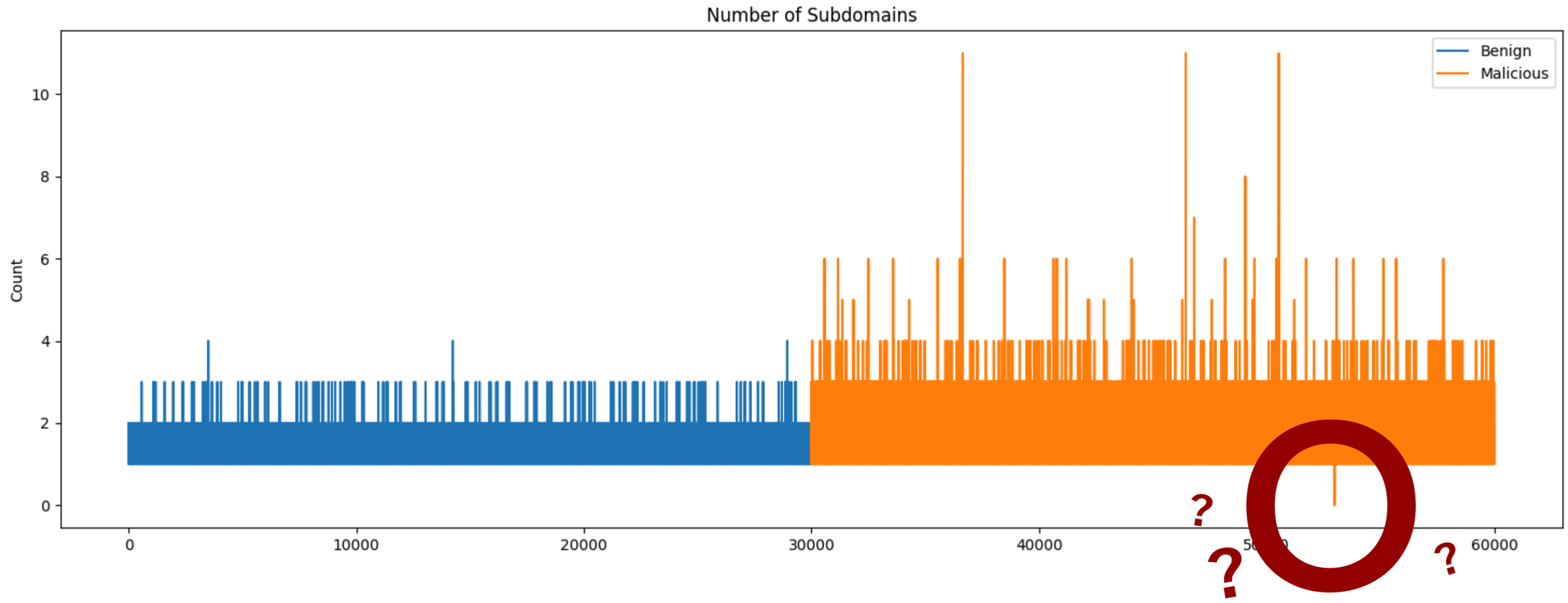
Data Exploration





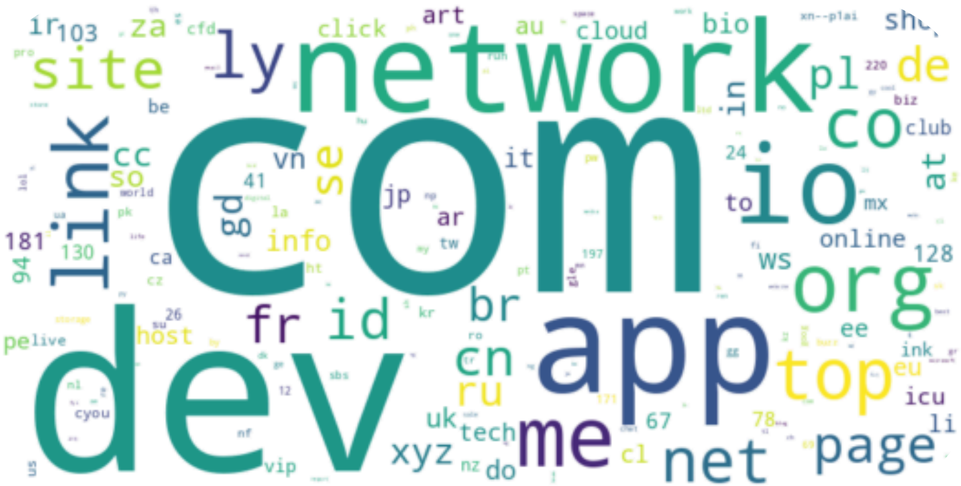
Malicious URL Classification

Data Exploration



Malicious URL Classification

Data Cleaning



Others

Malicious URL Classification

Data Cleaning

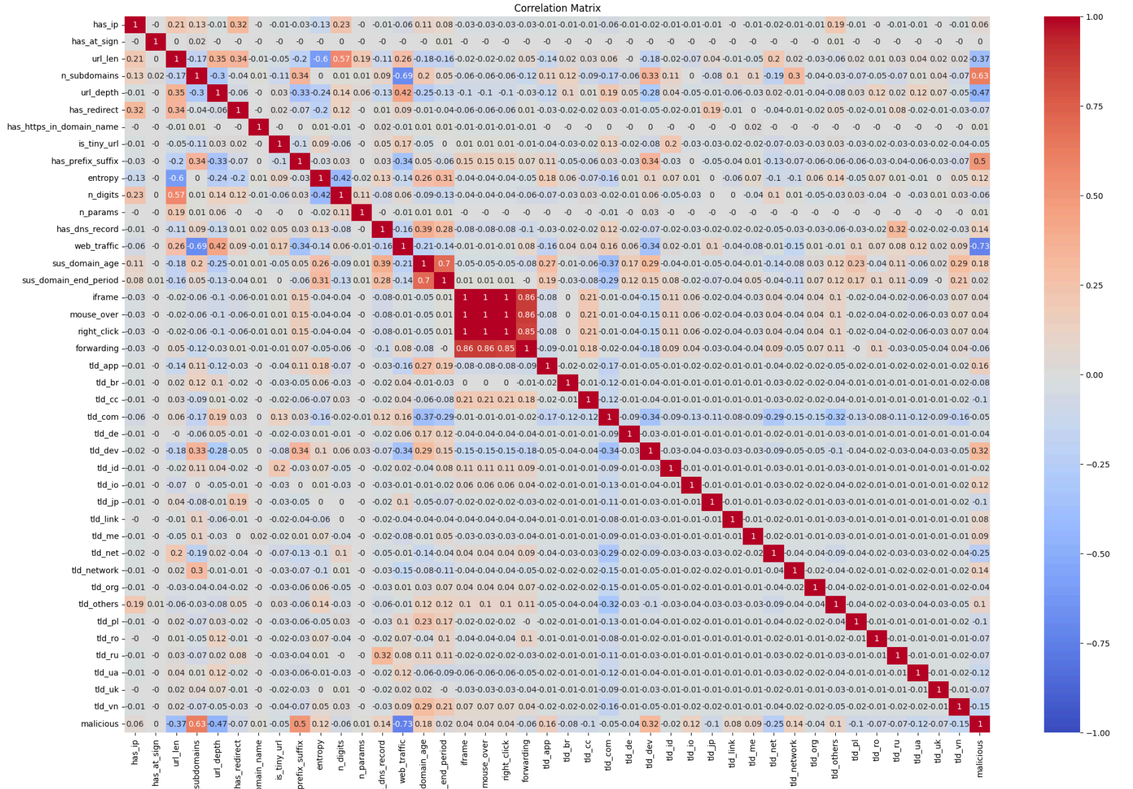
- Compress the Top-Level Domains
- One-Hot Encode Top-Level Domains
- Rearrage Web Traffic values
- Discard null values and duplicates
- Keep outliers
- Remove strage samples

Malicious URL Classification

Data Preprocessing

Before:

41 features

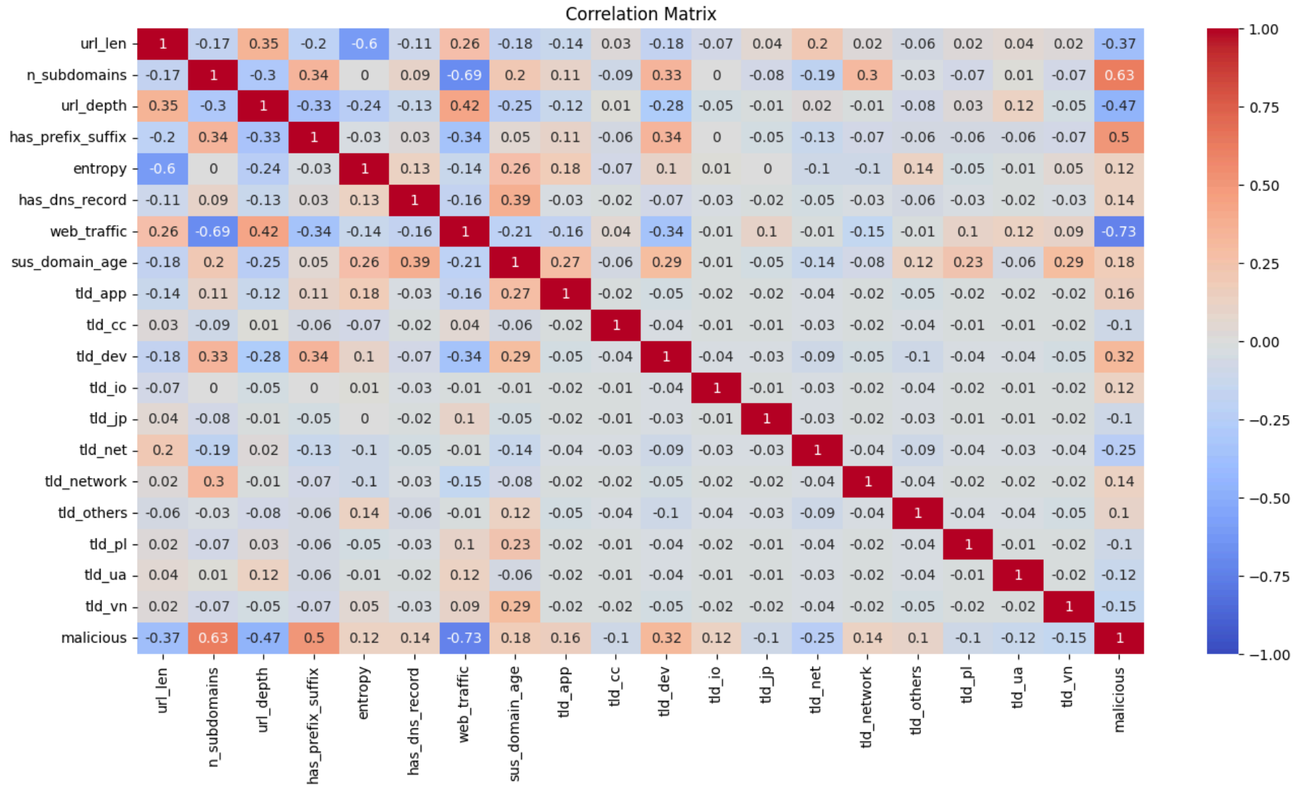


Malicious URL Classification

Data Preprocessing

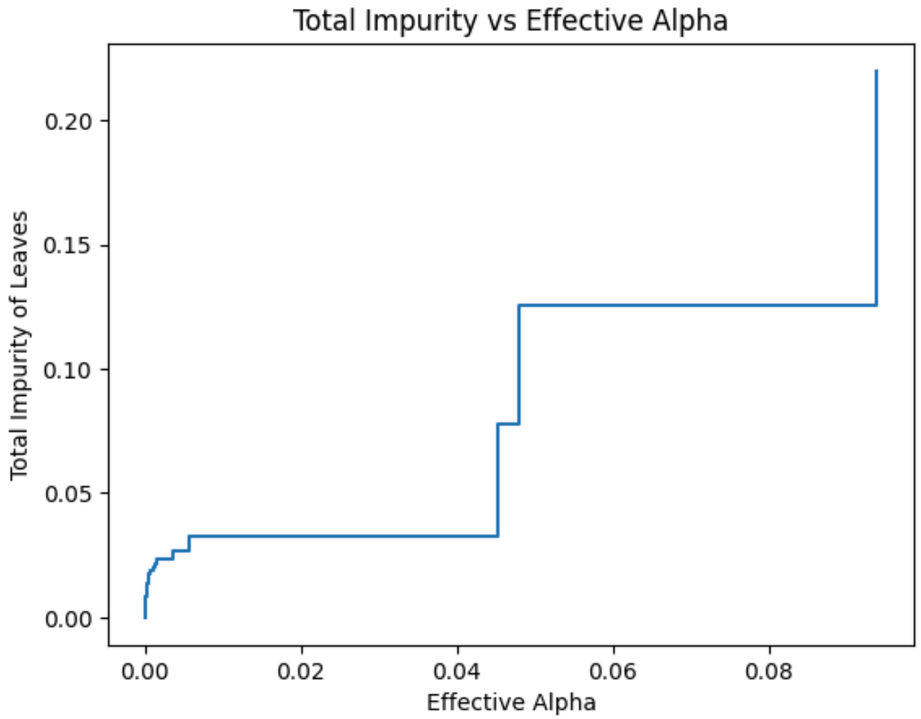
After:

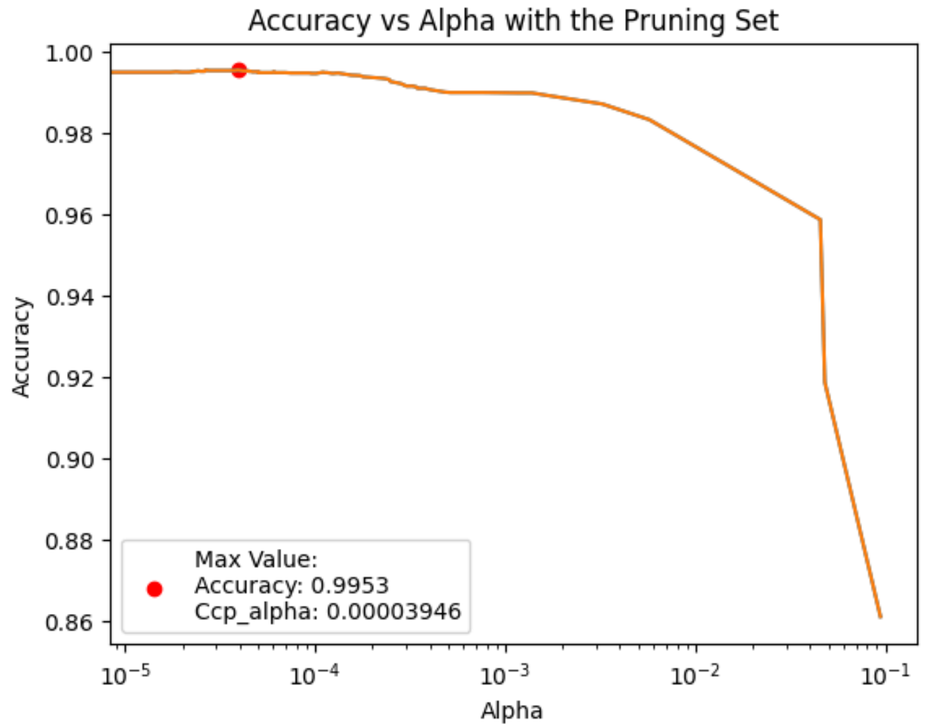
19 features



Malicious URL Classification

Data Processing





Malicious URL Classification

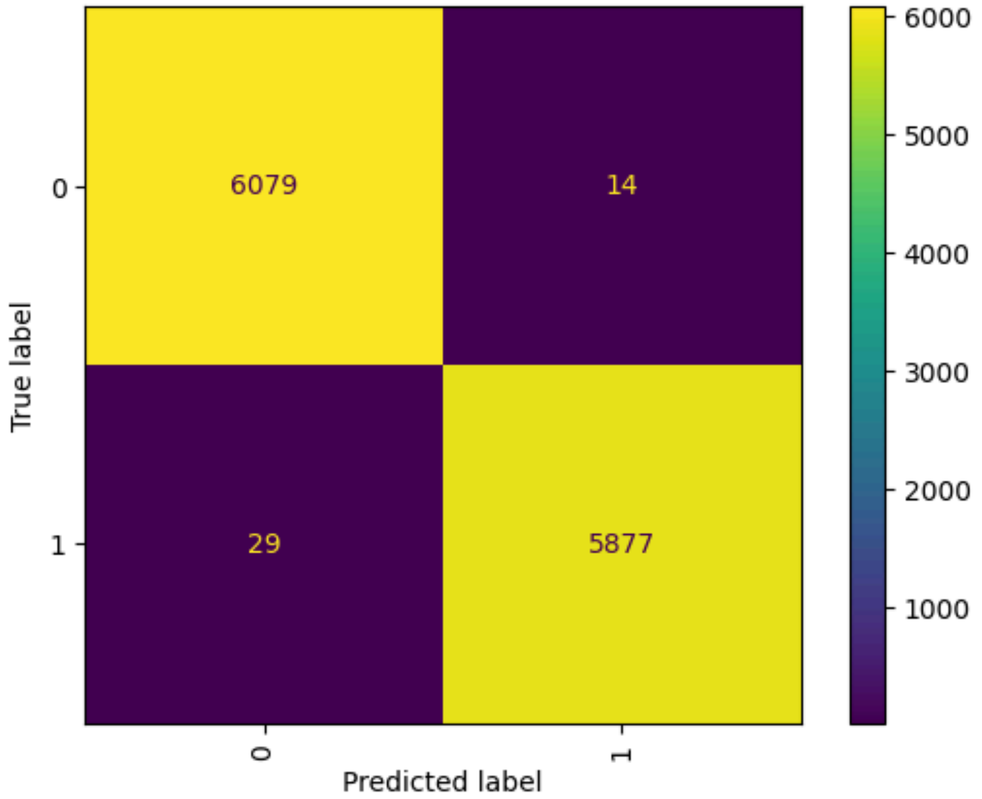
Validation

Accuracy: 0.9964

Precision: 0.9953

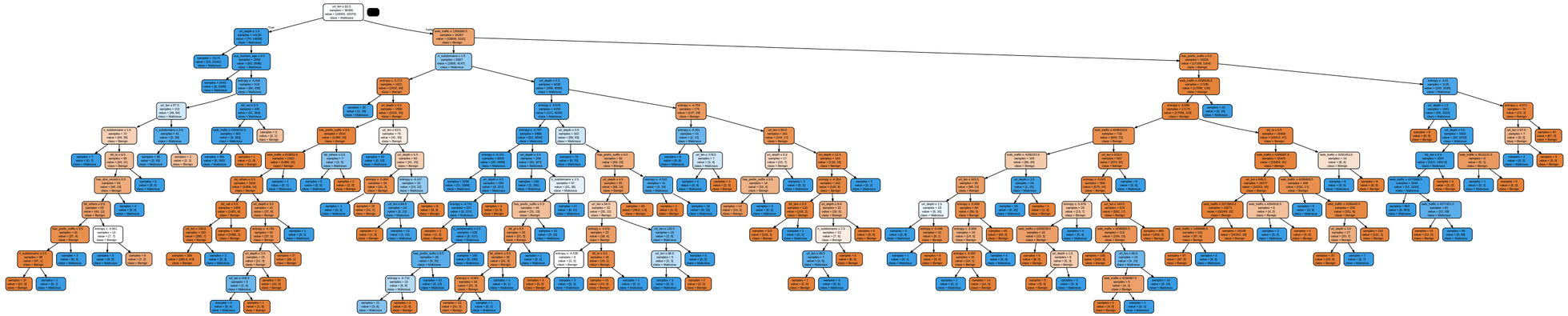
Recall: 0.9977

F1-Score: 0.9965



Malicious URL Classification

Conclusions



Malicious URL Classification

Conclusions

- A lot of the classic features proposed in the literature are not used
- Some of the non-classical ones instead revealed to be useful
- To classify URLs, an active scanning is needed

Further Improvements

- Analyze information related to the domain age, creation and end dates
- Analyze other classical features
- Analyze new features not present in the literature

Malicious URL Classification

Demo Application

URL Checker

Enter a URL to check

Submit

Malicious URL Classification

Thanks for the Attention