Creating a URL Phishing Detector

by Aggrey Timbwa, Richard Macharia, Pamela Jepkorir Chebii, Cynthia Njambi, Omara Waldea

What is Phishing?

Deceptive Tactics

- Fraudsters use email, social media, and other online channels to trick users into revealing sensitive information.

Real-World Examples

- Emails impersonating banks, government agencies, and trusted brands are common phishing techniques.

Understanding Phishing Attacks

1

Business Risks

- Financial losses
- reputational damage.

2

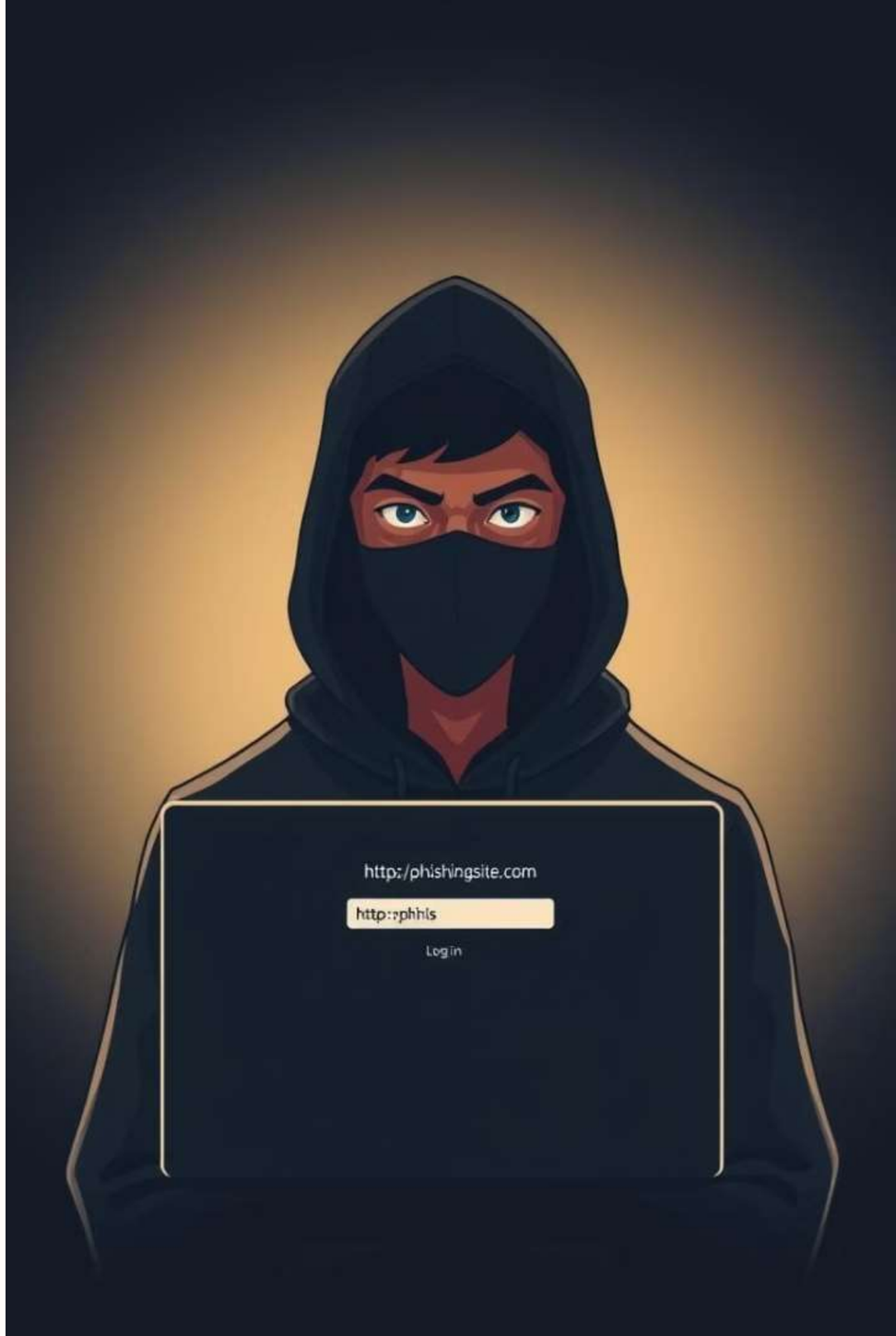
Cybercrime Sophistication

- Advanced tactics
- Social engineering.

3

Protection Necessity

- Tools to identify phishing URLs for real-time detection.



Business Objectives

High-Accuracy Detection

User-Friendly Deployment

Real-Time Classification

Expected Impact Globally

1 Financial Security

- Reduce financial losses due to phishing scams.

2 User Trust

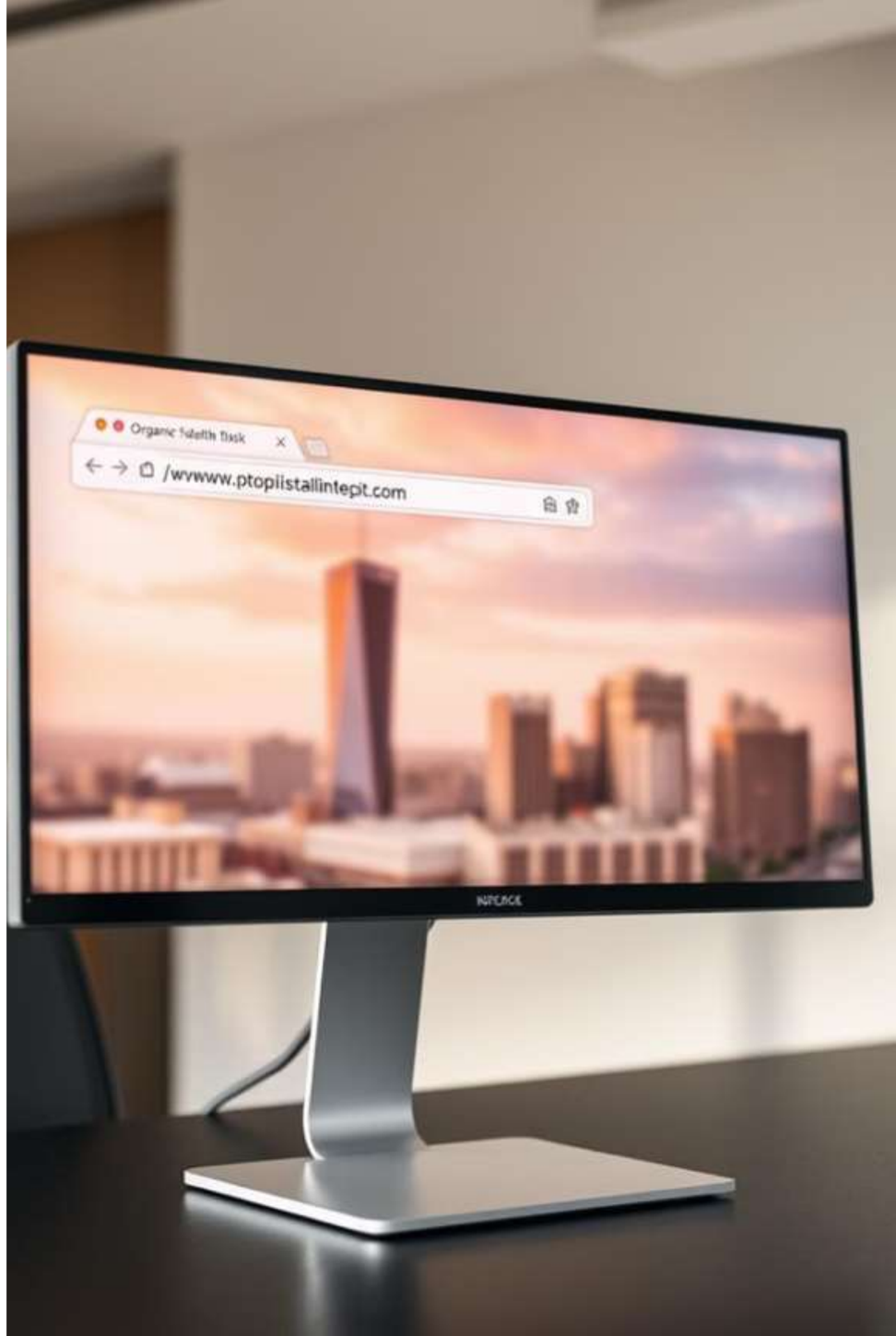
- Build confidence in online interactions.

Benefits to Stakeholders

- **Business Users**
  Helps secure business operations by preventing phishing scams, protecting financial data, and ensuring safe online transactions. Example is banking services.

- **Individual Users (General Public)**
  Provides an easy way to avoid phishing attacks and safeguard personal information while browsing the internet.

- **Cybersecurity Teams**
  Supports proactive threat prevention by identifying phishing URLs and integrating into existing security measures.



Preparing a Training Dataset

1

Data Collection

- Gather diverse labelled URLs from Mendely Dataset .

2

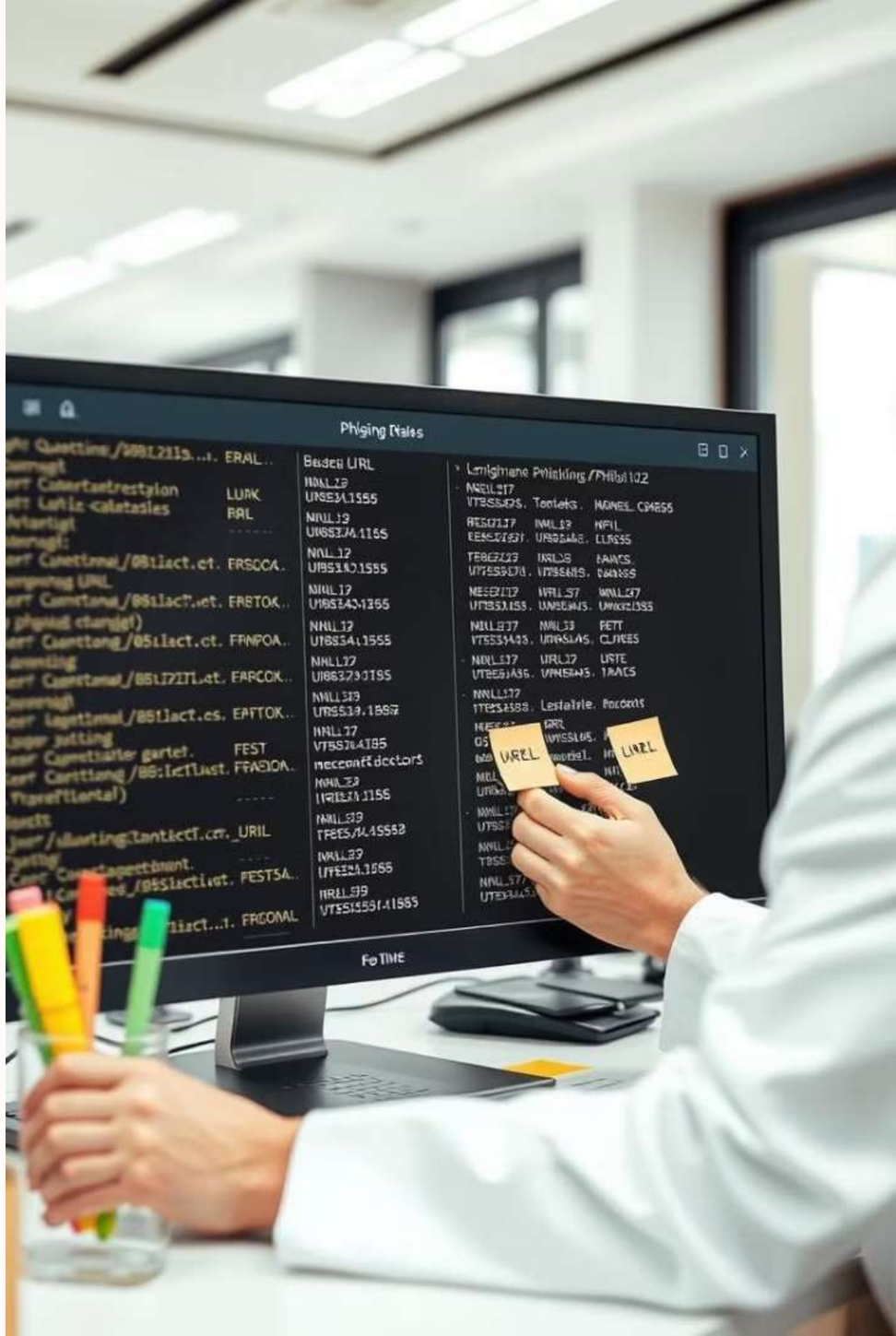
Data Preprocessing

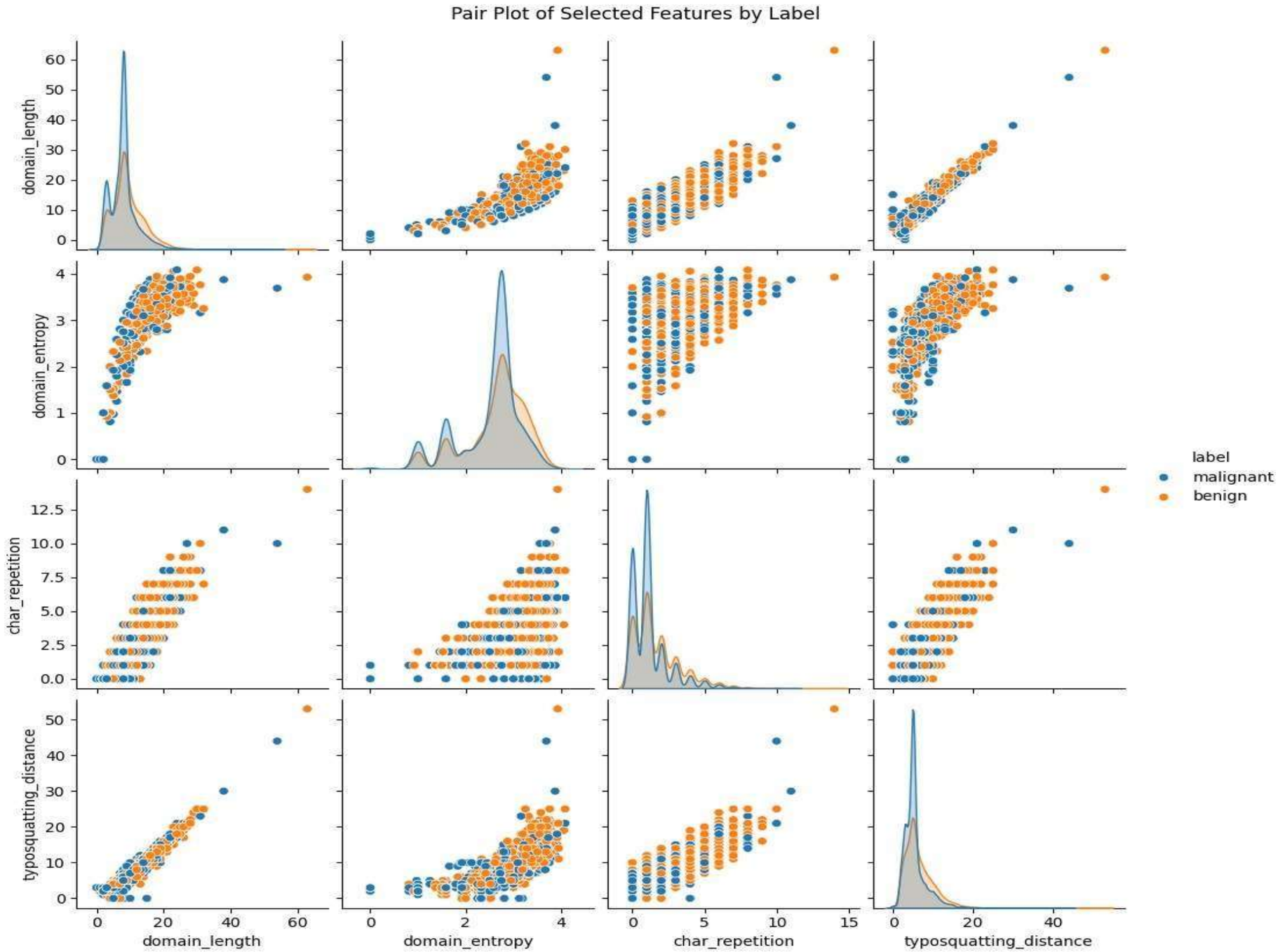
- Clean dataset by removing duplicates and invalid entries.

3

Feature Extraction and Engineering

- Extract relevant URL attributes like domain age, URL length.





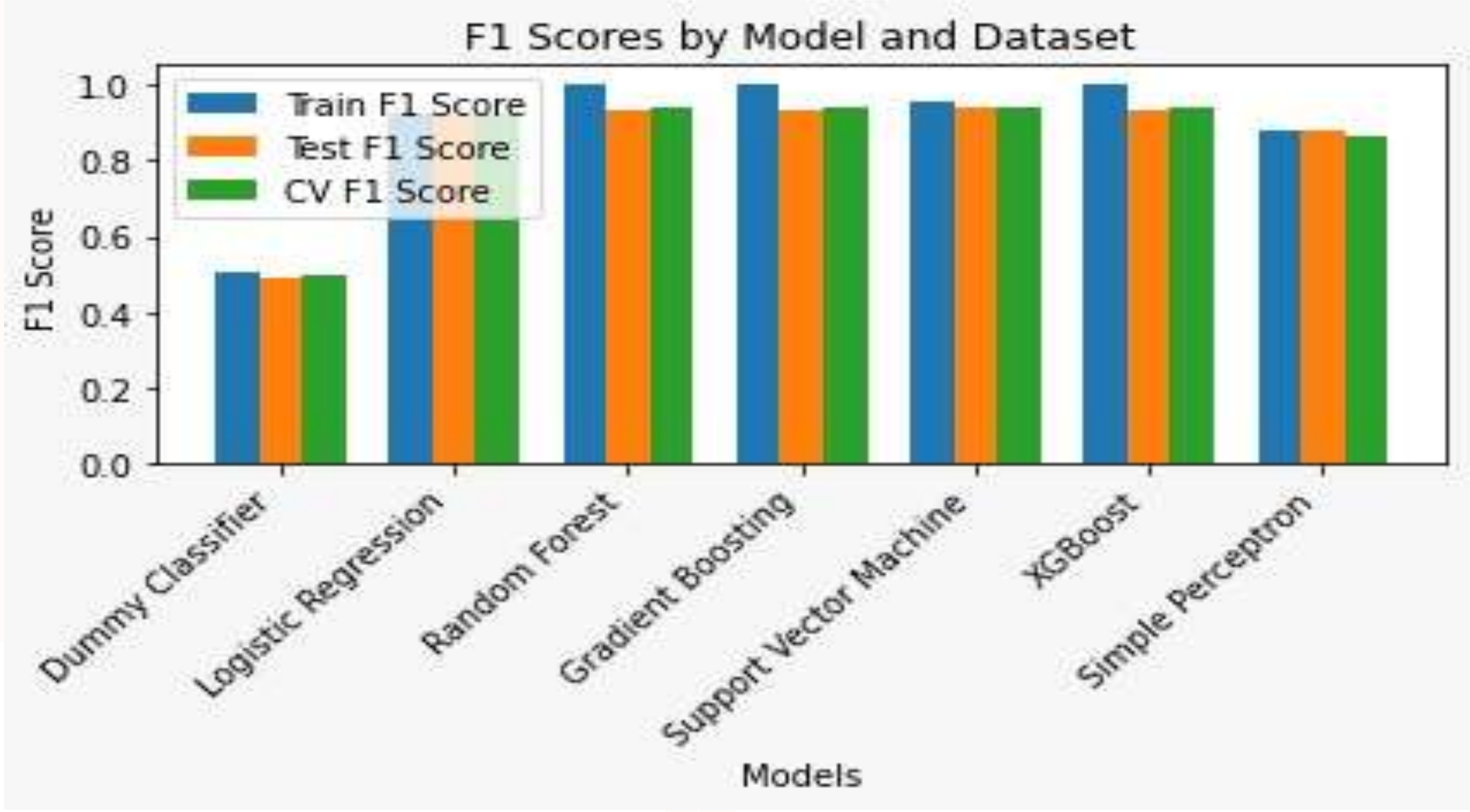
VISUALIZATION

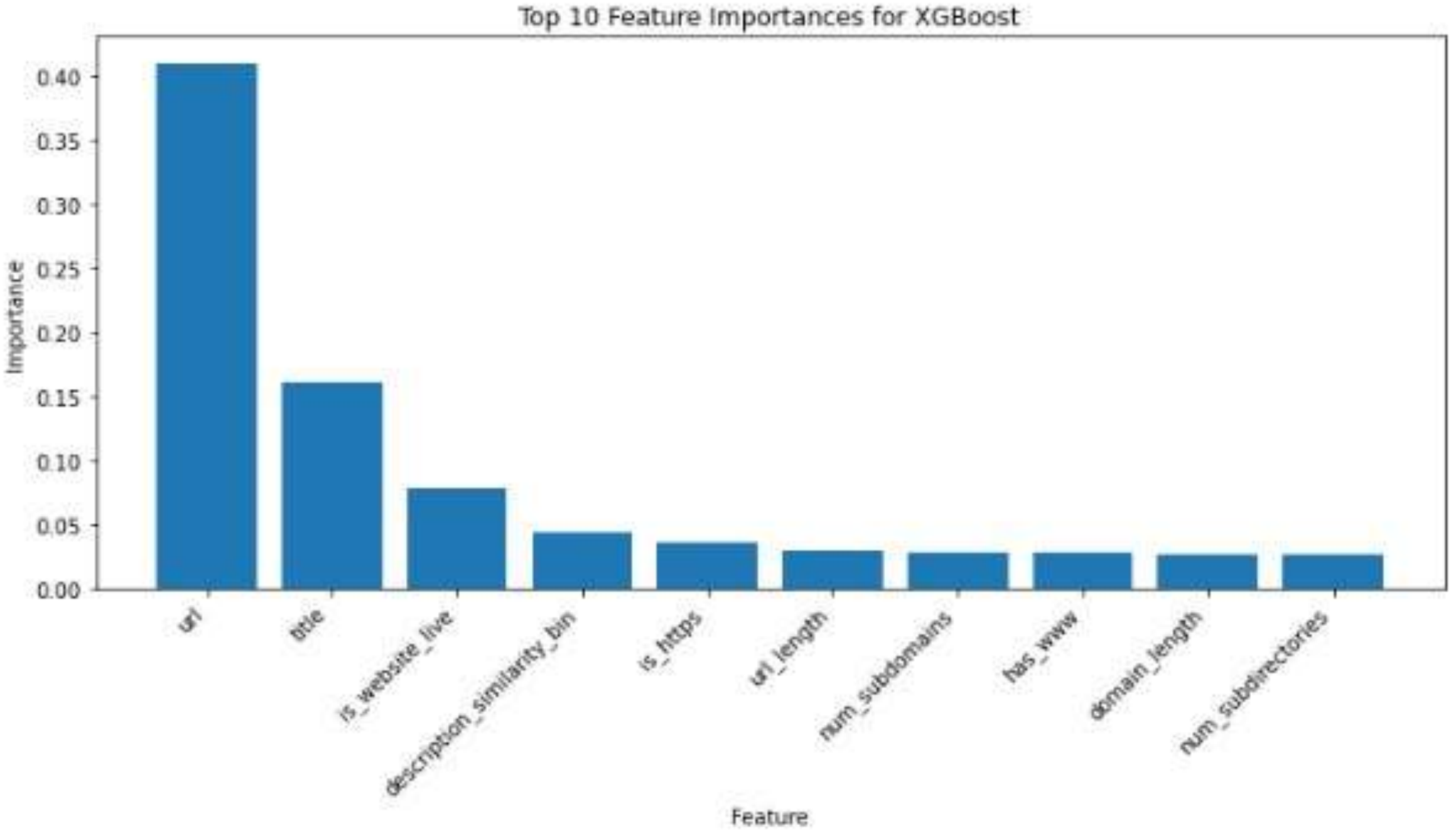
- **Distinct Distributions:**
  Some features show clear differences between benign and malignant labels.

- **Class Separation:**
  Features like domain length and typosquatting distance aid in distinguishing classes.

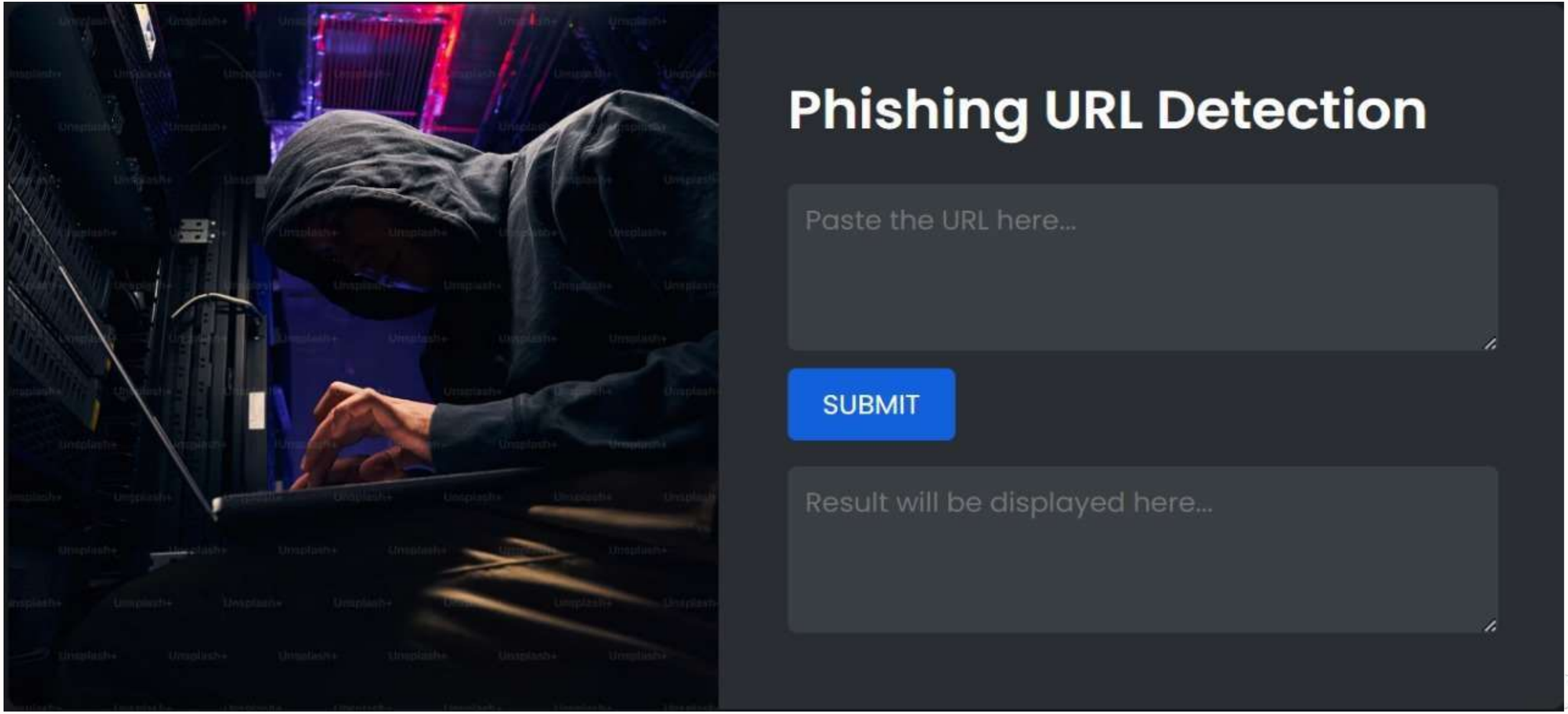
- **Significant Features:**
  Domain length, entropy, and character repetition stand out for classification.

Model Training and Evaluation



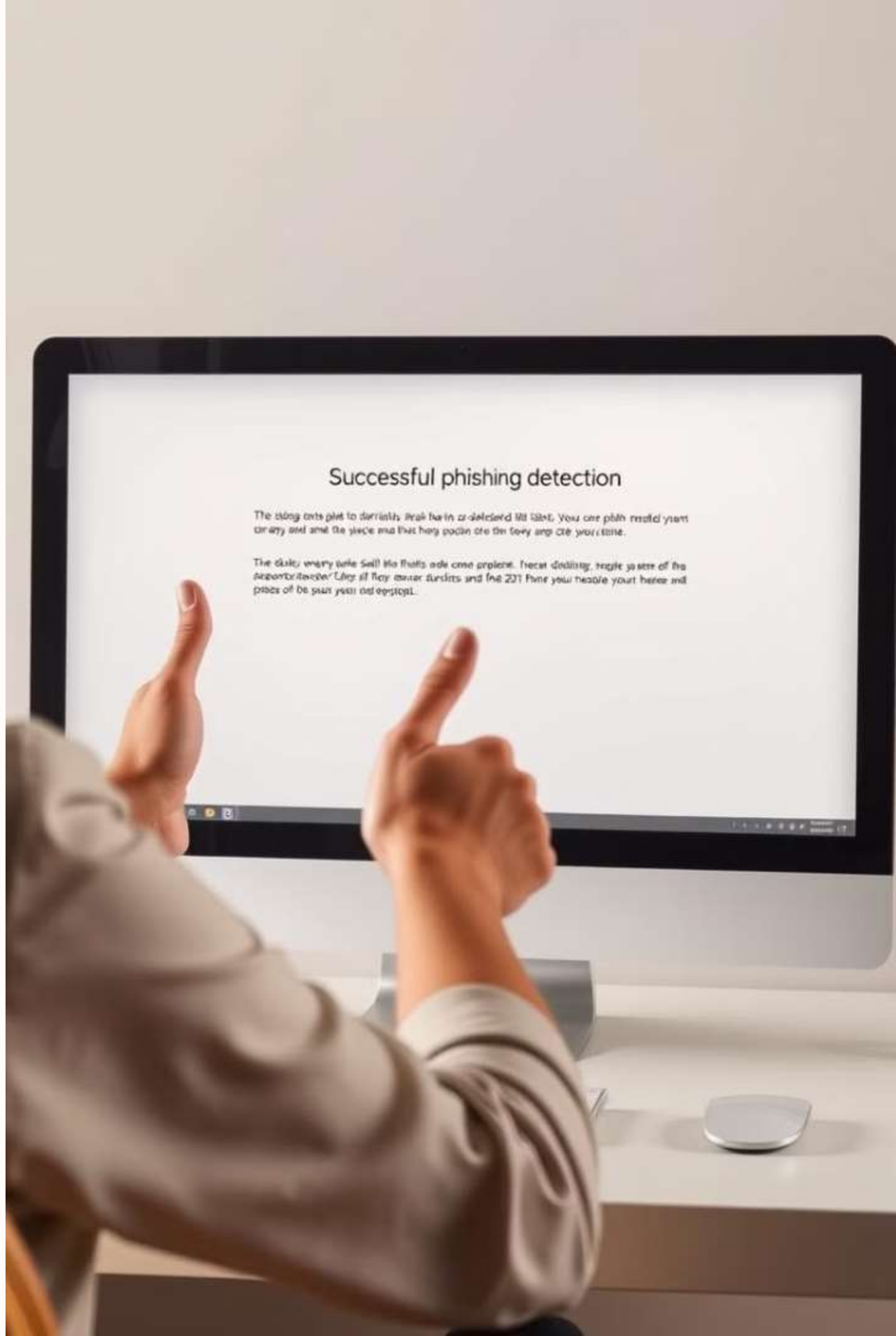

Top 10 Feature Importances for XGBoost
Importance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
url
title
is_website_live
description_similarity_bin
is_https
url_length
num_subdomains
has_www
domain_length
num_subdirectories
Feature

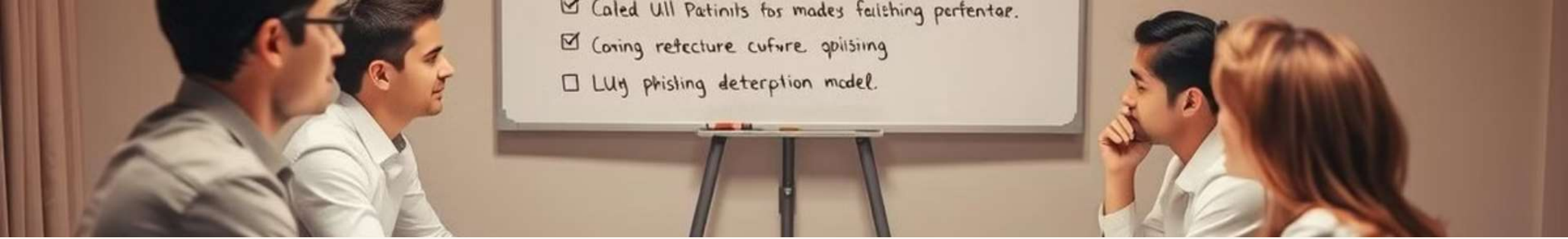
Deploying the Detector



Conclusion

- High-Accuracy Phishing Detection.
- Balanced F1-Score
- User-Friendly Web Deployment
- Real-Time Classification
- Identify Important Features





Recommendations

- Implement Continuous Model Retraining
- Train the Model Using More URLs
- Consider Deployment as a Browser Extension
- User Feedback

Q&A

Thank You