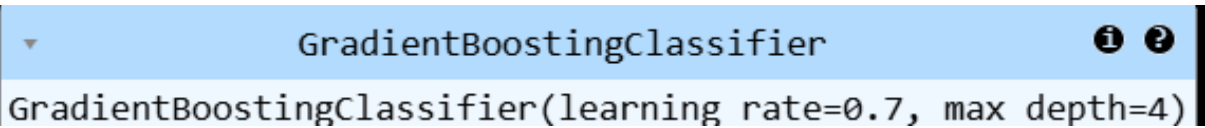## Abstract

Phishing remains a critical cybersecurity threat, causing significant financial and data loss. This project addresses this challenge by developing a high-performance machine learning model to automatically classify website URLs as either benign (legitimate) or malicious (phishing). An XGBoost (Extreme Gradient Boosting) classifier was trained on 30 features engineered from a dataset of over 11,000 URLs. The resulting model is both highly accurate, achieving 97% on unseen data, and practical, with a functional real-time prediction system.
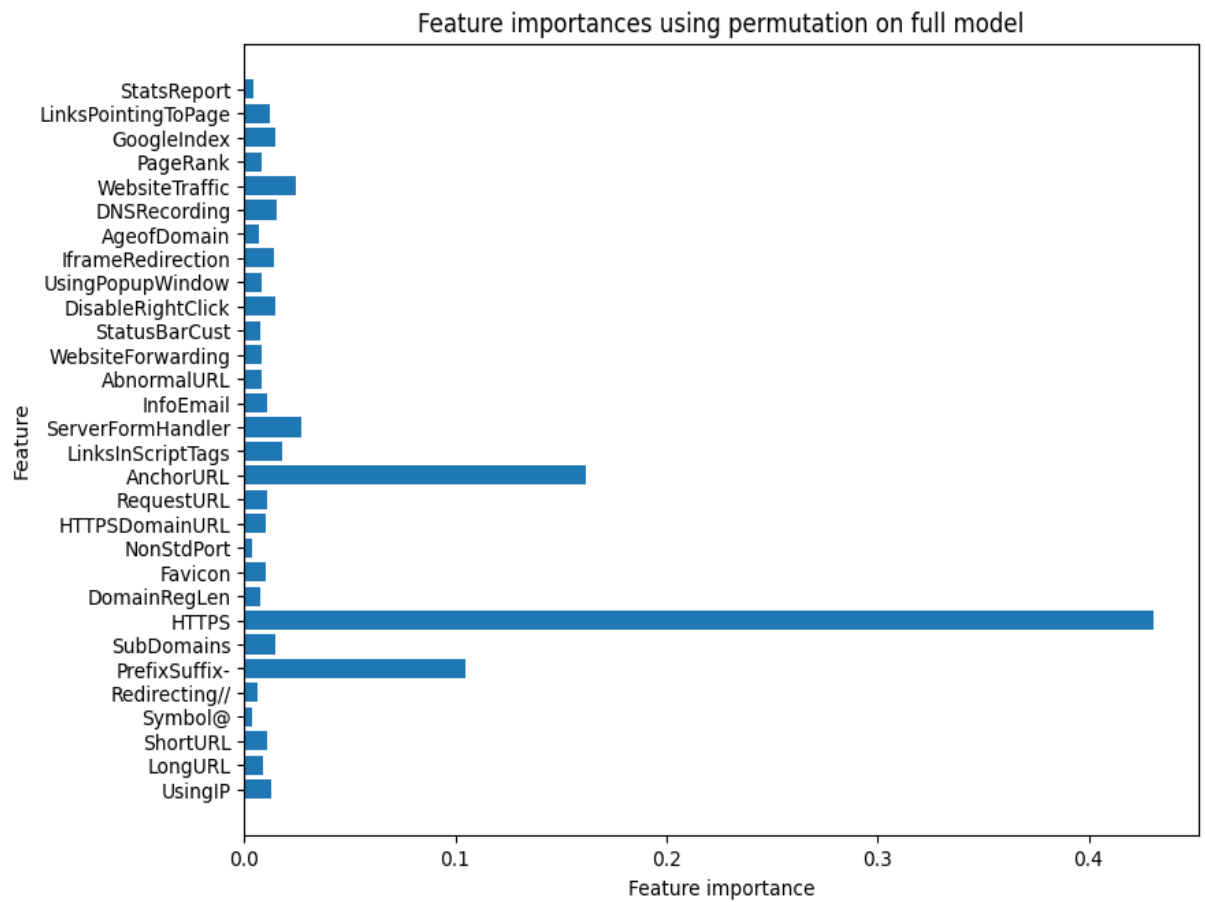
Data and Processing

- Loaded the dataset of over **11,000 URLs** from the phishing.csv file.
- Cleaned the data and standardized the labels (1 for Benign, -1 for Malicious).
- Confirmed the dataset was **well-balanced**, with a nearly even split between benign and malicious URLs.
- The model was trained using **30 pre-calculated features** from the file.
- A separate function was built to extract features from live URLs for the real-time prediction system.
- Split the data into **80% for training** and **20% for testing** to ensure a fair and reillaliable evaluation of the model.

Model:

GradientBoostingClassifier(learning rate=0.7, max depth=4)

Feature Importance:

Training and Testing Accuracy:

Metrics:

	precision	recall	f1-score	support
0	0.98	0.95	0.97	976
1	0.97	0.98	0.97	1235
accuracy			0.97	2211
macro avg	0.97	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

Outcomes:
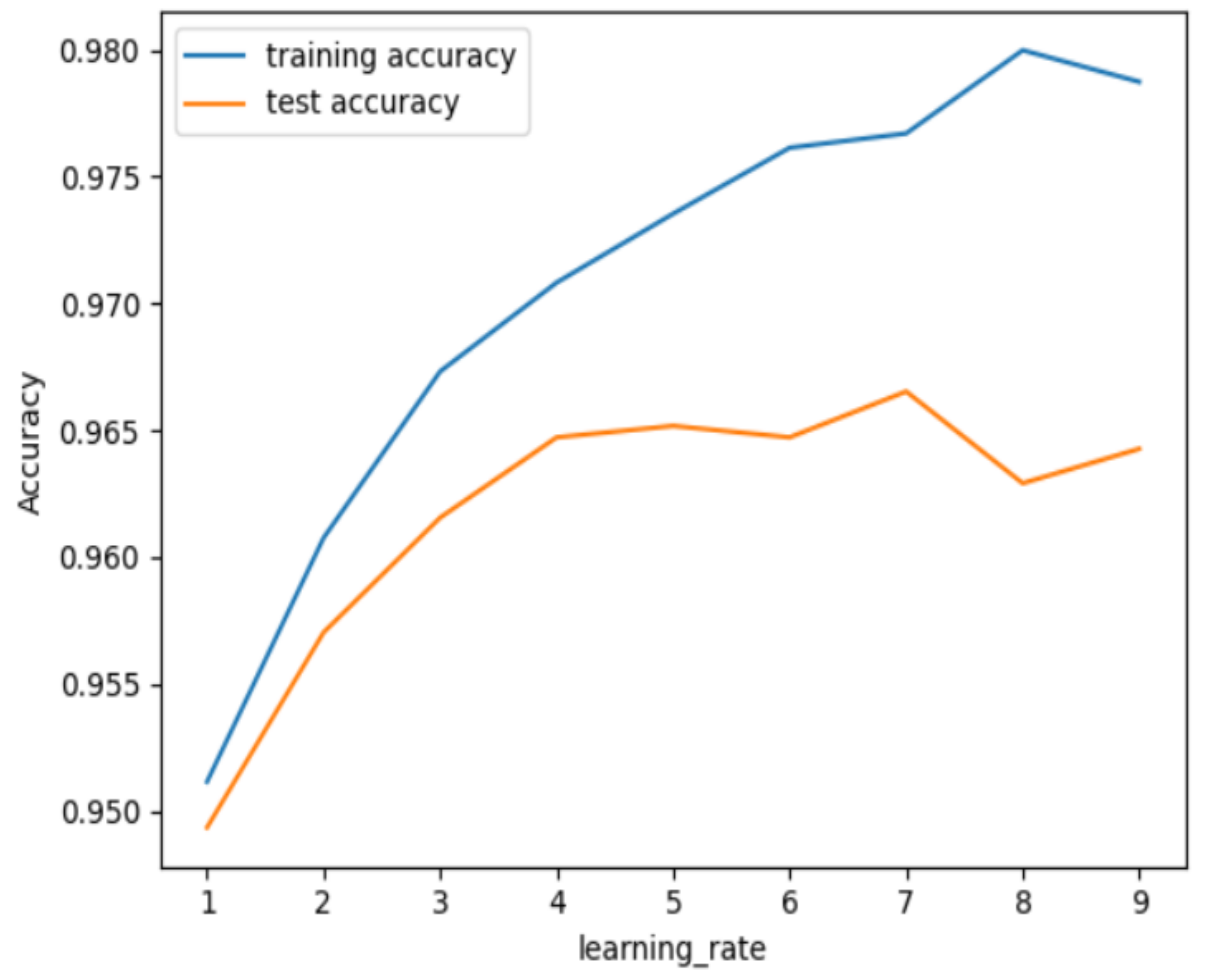
Conclusions

- The **XGBoost** model is a highly effective and reliable tool for detecting phishing websites, achieving 97% accuracy.
- The project successfully proves that features engineered directly from URL strings can be powerful **predictors of malicious intent**.
- The final model serves as a strong foundation for a practical **web security** application.

Future Enhancements

- **Integrate External Features:** Enhance accuracy by incorporating real-time API lookups for features like AgeofDomain (WHOIS) and DNSRecording.
- **Hyperparameter Tuning:** Systematically optimize the XGBoost model's parameters using GridSearchCV to potentially increase performance further.
- **Deployment:** Package the model as a public web application or a browser extension for real-world use.