PHISHING WEBSITE DETECTION INTERNSHIP

Using EDA and Machine Learning Deployment

By YD Saritha

OBJECTIVE

- Detect phishing websites with high accuracy.
- Analyze features that influence phishing predictions.
- Deploy the model for real-world usability.
- Improve cybersecurity measures for individuals and organizations.

Work Flow

- Data Preprocessing (cleaning, feature selection).
- Model Training (choosing algorithms).
- Explainability Analysis (SHAP/LIME).
- Deployment (Flask app and Docker container).
- Monitoring and Maintenance.

Feature Analysis

- Top features influencing phishing detection.
- Visuals: Correlation heatmap, SHAP summary plot, or bar graph ,count plot, box plot of feature importance.

Model Results

Algorithms Explored
Logistic Regression
Random Forest
Decision Tree
SVM
Naive Bayes

Final Choice

- Random Forest

- Reason is Random Forest Classifier has best accuracy

Model Evaluation

- Random Forest:
- Accuracy: 0.9663167104111986
- precision recall f1-score support

- 0 0.96 0.97 0.97 1157
- 1 0.97 0.96 0.97 1129

- accuracy 0.97 2286
- macro avg 0.97 0.97 0.97 2286
- weighted avg 0.97 0.97 0.97 2286

Model Explainability

- SHAP used for visualize and analysis.
- Most influential features include URL length, HTTPS usage, and domain age.
- Ensures the model predictions are interpretable and align with domain knowledge.
- Builds trust in predictions and supports ethical AI deployment.

Deployment

- **Local Deployment:**
  - Flask-based application for predictions.
  - User interface to upload data or URLs.
- **Cloud Deployment:**
  - Scalable on AWS/GCP/Azure.
  - API available for integrations.
- **Portability:**
  - Dockerized application for consistent deployment.

THANK YOU