ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

INSPIRATION

Phishing attacks have emerged as a significant and persistent threat in the digital world, targeting individuals, organizations, and even governments. These deceptive techniques used by cybercriminals aim to trick unwitting users into revealing sensitive information such as login details, financial details, or personal data.

Research shows that more than 48% of emails sent in 2022 are spam, with approximately 3.4 billion spam emails being sent every day. Globally, 323,972 internet users fell victim to phishing attacks in 2021. Considering that the average loss per phishing attack is $136, this means that $44.2 million was stolen by cybercriminals through phishing attacks in 2021.

As a notable example, in 2016, JP Morgan Chase, a major American bank, was the victim of a phishing attack. In this attack, some of the bank's employees opened a fake email and lost their login information to the attackers. As a result of this incident, the information of 76 million households and 7 million small businesses was compromised. Such attacks show that not only individuals but also large organizations can be targeted and cause serious security vulnerabilities.

To prevent phishing attacks, individuals and institutions need to be careful, receive awareness training, and constantly update their security measures. Additionally, paying attention to small details in emails and websites, not clicking on suspicious links, and not disclosing information in cases where security is not ensured are basic steps to protect against such attacks.

As a result, phishing attacks continue to pose a significant threat in the digital world, and it is of great importance for both individuals and institutions to be prepared against these threats. Education, awareness, and security measures play a critical role in reducing the effects of phishing attacks.

Techjury - Spam Statistics 2022

Statista - Spam e-mail volume

Statista - Phishing victims worldwide 2021

Security Boulevard - The Real Cost of Phishing

New York Times - JPMorgan Chase Hacking Affects 76 Million Households

ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

PROBLEM STATEMENT

Phishing attacks pose a major threat to online users, jeopardizing their privacy, financial security, and confidence in online interactions. Detecting and preventing phishing sites is challenging due to the ever-evolving tactics of cybercriminals and requires effective techniques to accurately distinguish between legitimate and malicious websites.

Current phishing detection methods are inadequate in the face of the changing strategies of cybercriminals, and therefore an improved approach is needed to detect phishing sites.

Therefore, there is a critical need to develop an advanced system that can detect phishing sites accurately and efficiently using a combination of advanced machine learning techniques, feature engineering, and behavioral analysis. By overcoming these challenges, the proposed methodology aims to improve the security of online users, protect their sensitive information, and create a safer digital environment.

INTRODUCTION

The aim is to contribute to the development of a safer digital environment by providing an advanced approach to detecting phishing sites. By accurately detecting and preventing phishing threats, the proposed model will increase the security and reliability of online interactions, preventing users from becoming victims of phishing attacks.

In the following sections, we discuss the relevant literature, present the methodology, describe the experiments and results, and evaluate the implications and future directions of the research. The aim of this study is to ensure the security of online users and to create a more robust security infrastructure in the digital world.

APPROACH

- Datasets containing phishing and legitimate websites is collected from open-source platform PhishTank.
- Write a code to extract the required features from the URL database.
- Analyze and preprocess the dataset by using EDA techniques.
- Divide the dataset into training and testing sets.
- Run selected machine learning and deep neural network algorithms on the dataset like Decision Tree, Random Forest, Multilayer Perceptrons, XGBoost and Support Vector Machines on the dataset.
- Write a code for displaying the evaluation result considering accuracy metrics.
- Compare the obtained results for trained models and specify which is better.

ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

PROCEDURE

1) All libraries required for the Project:

- TensorFlow
- NumPy
- Pandas
- SciKit-Learn

2) Understanding the content of datasets:

- Datasets containing phishing and legitimate websites are collected from the open-source platform PhishTank.
- This service provides a set of phishing URLs in multiple formats, like CSV, JSON etc. , that gets updated hourly. From this dataset, 5000 random phishing URLs are collected to train the machine learning models.
- The legitimate URLs are extracted from the open datasets of the University of New Brunswick. This dataset has a collection of benign, spam, phishing, malware, & defacement URLs. Out of all these types, the benign URL dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

	phish_id	url	phish_detail_url	submission_time	verified	verification_time	online	target
0	6557033	http://u1047531.cp.regruhosting.ru/acces-inges...	http://www.phishtank.com/phish_detail.php?phis...	2020-05-09T22:01:43+00:00	yes	2020-05-09T22:03:07+00:00	yes	Other
1	6557032	http://hoysalacreations.com/wp-content/plugins...	http://www.phishtank.com/phish_detail.php?phis...	2020-05-09T22:01:37+00:00	yes	2020-05-09T22:03:07+00:00	yes	Other
2	6557011	http://www.accsystemprblemhelp.site/checkpoint...	http://www.phishtank.com/phish_detail.php?phis...	2020-05-09T21:54:31+00:00	yes	2020-05-09T21:55:38+00:00	yes	Facebook
3	6557010	http://www.accsystemprblemhelp.site/login_atte...	http://www.phishtank.com/phish_detail.php?phis...	2020-05-09T21:53:48+00:00	yes	2020-05-09T21:54:34+00:00	yes	Facebook
4	6557009	https://firebasestorage.googleapis.com/v0/b/so...	http://www.phishtank.com/phish_detail.php?phis...	2020-05-09T21:49:27+00:00	yes	2020-05-09T21:51:24+00:00	yes	Microsoft

[2]:
```
# veri setinin yolu
data0 = pd.read_csv(r"C:\Users\Yekta\Desktop\phishing-attack-project\dataset\url-dataset.csv")
data0.head()
```

[2]:

	Domain	Have_IP	Have_At	URL_Length	URL_Depth	Redirection	https_Domain	TinyURL	Prefix/Suffix	DNS_Record	Web_Traffic	Domain_Age	Domain_End	iFrame	Mouse_Over
0	graphicriver.net	0	0	1	1	0	0	0	0	0	1	1	1	0	0
1	ecnavi.jp	0	0	1	1	1	0	0	0	0	1	1	1	0	0
2	hubpages.com	0	0	1	1	0	0	0	0	0	1	0	1	0	0
3	extratorrent.cc	0	0	1	3	0	0	0	0	0	1	0	1	0	0
4	icicibank.com	0	0	1	3	0	0	0	0	0	1	0	1	0	0

3) Feature Extraction:

The below-mentioned category of features are extracted from the URL data:

a) Addressed Bar-based features

- In this category, 9 features are extracted.

b) Domain-based Features

- In this category, 4 features are extracted.

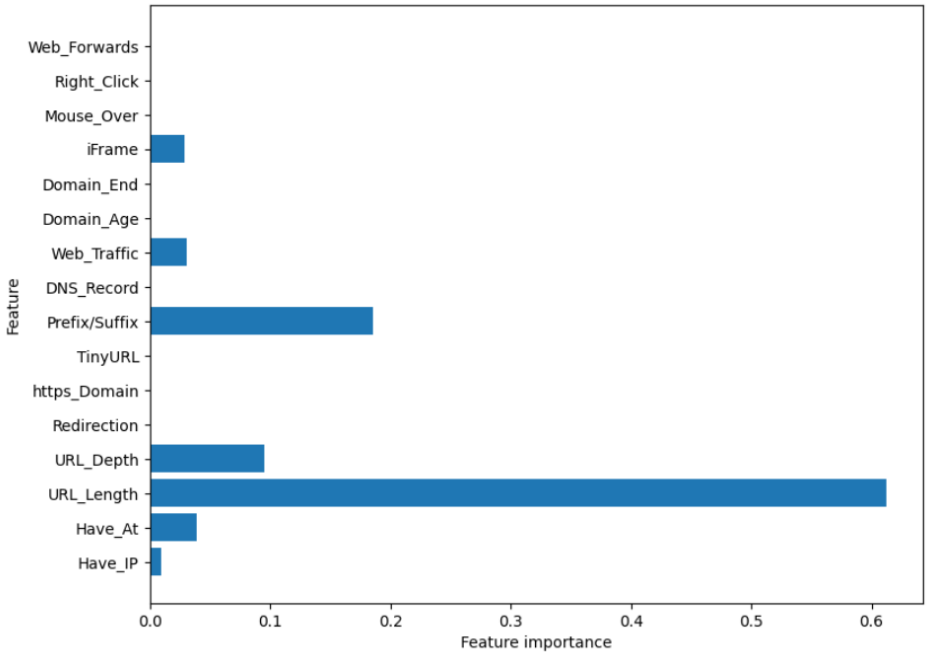
c) HTML & Javascript-based Features

- In this category, 4 features are extracted.

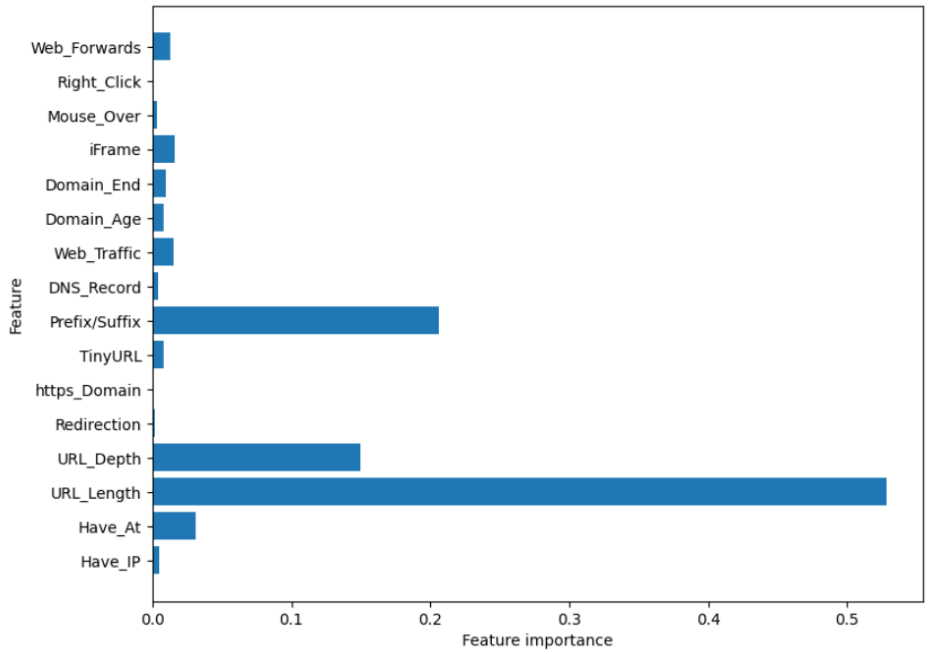
ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

Decision Tree and Feature Selection:



Random Forest and Feature Selection:



DETECTING/PREVENTING PHISHING ATTACKS PROJECT

4) Build and train the model:

Before starting the ML model training, the data is split into 80-20, i.e., 8000 training samples & 2000 testing samples. From the dataset, it is clear that this is a supervised machine-learning task. This data set comes under a classification problem, as the input URL is classified as phishing (1) or legitimate (0).

The supervised machine learning models (classification) considered to train the dataset in this project are:

- **Decision Tree**
- **Random Forest**
- **Multilayer Perceptrons**
- **XGBoost**
- **Support Vector Machines**

Accuracy Rates of Models before 50 Epoch

	ML Model	Train Accuracy	Test Accuracy
0	Decision Tree	0.813	0.814
1	Random Forest	0.816	0.817
2	Multilayer Perceptrons	0.869	0.857
3	XGBoost	0.869	0.852
4	SVM	0.801	0.808

Accuracy Rates of Models after 50 Epoch

	Model	Mean Accuracy	Standard Deviation
0	SVM	0.80900	1.110223e-16
1	MLP	0.83675	3.833080e-03
2	XGBoost	0.84350	1.110223e-16
3	Random Forest	0.84618	1.080555e-03
4	Decision Tree	0.84429	2.467793e-04

ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

Accuracy Rates of Models after 5-Flod Cross Validation

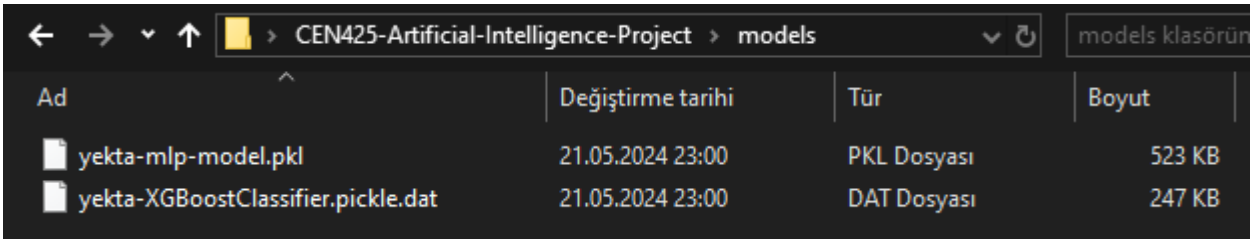
	Model	Accuracy Mean	Accuracy Std
0	SVM	0.8180	0.005477
1	MLP	0.8511	0.009351
2	XGBoost	0.8581	0.009876
3	Random Forest	0.8587	0.008165
4	Decision Tree	0.8574	0.009002

Accuracy Rates of Models after Chi-square Feature Selection

	Model	Train Accuracy Mean	Train Accuracy Std	Test Accuracy Mean	Test Accuracy Std
0	SVM	0.807025	0.002288	0.8050	0.006083
1	MLP	0.810800	0.001895	0.8084	0.009074
2	XGBoost	0.826550	0.002058	0.8223	0.010225
3	Random Forest	0.827100	0.002206	0.8234	0.009452
4	Decision Tree	0.827100	0.002206	0.8228	0.009179

5) Save the model:

- Save the model and calculate the training and testing accuracy.



Introduction to the Project Report

This project aims to develop a model using various machine learning algorithms to detect phishing attacks. Phishing is a type of malicious attack that tricks users into revealing sensitive information. In this study, five different algorithms were used to detect phishing URLs: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), XGBoost, Random Forest, and Decision Tree. Additionally, model performance was evaluated using 5-fold cross-validation and feature selection methods.

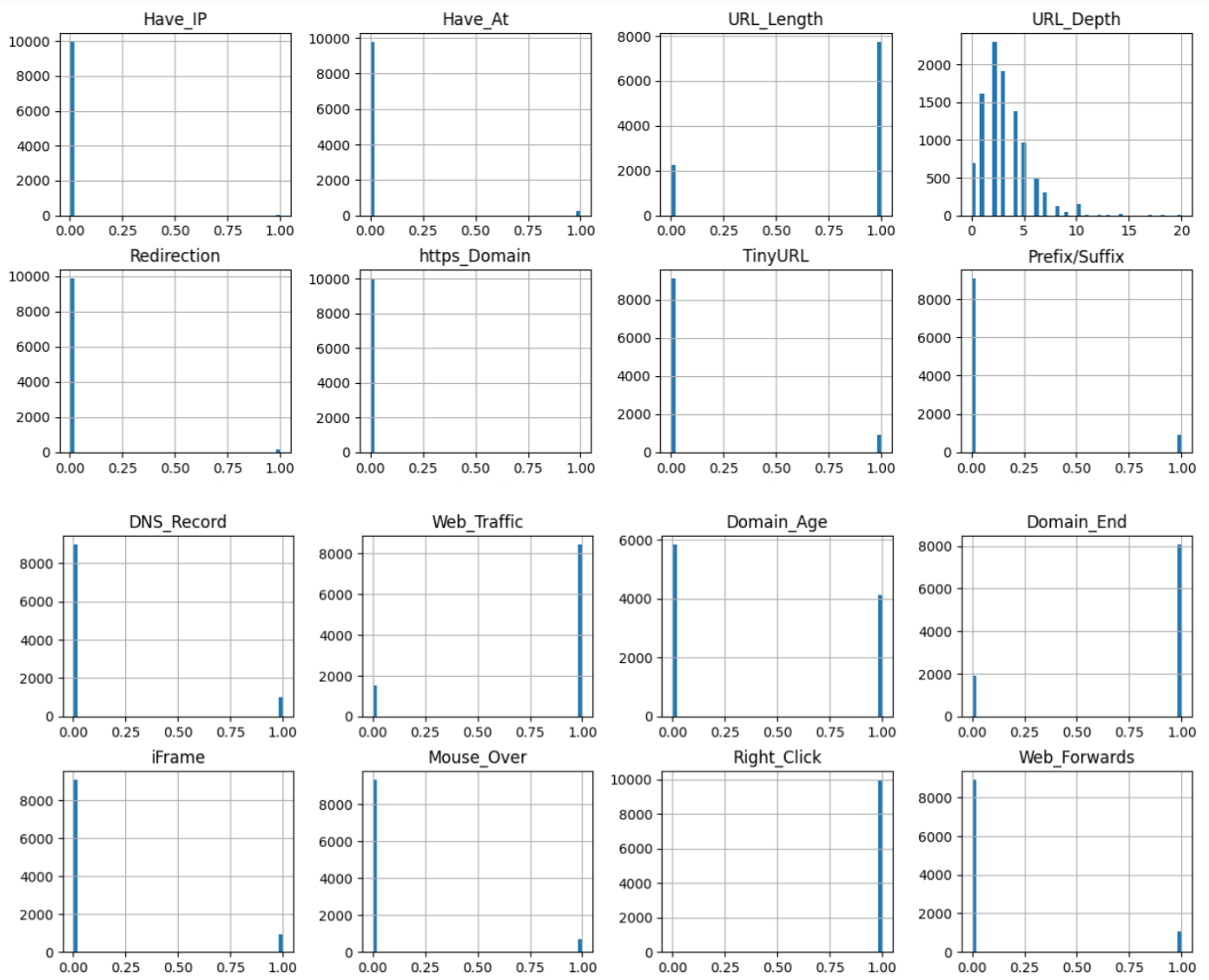
ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

Dataset

The dataset used consists of URLs with various features. These features are:

- URL length
- Domain information
- IP address usage
- Using HTTPS
- symbols in URL
- Email content
- IFrame redirects
- Google indexing status
- Page rank
- DNS information



ALİ YEKTA CURA – BATUHAN AŞAN
21.05.24

DETECTING/PREVENTING PHISHING ATTACKS PROJECT

Data Preprocessing

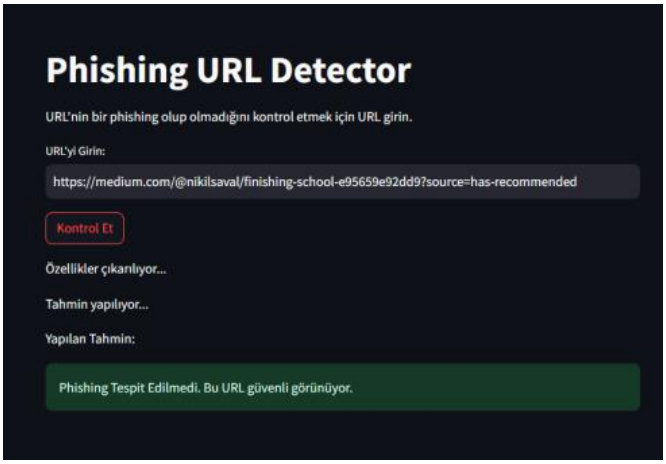
In the data preprocessing stage, missing data were completed, categorical data were converted to numerical values and the data were normalized. These steps are necessary for machine learning algorithms to work efficiently.

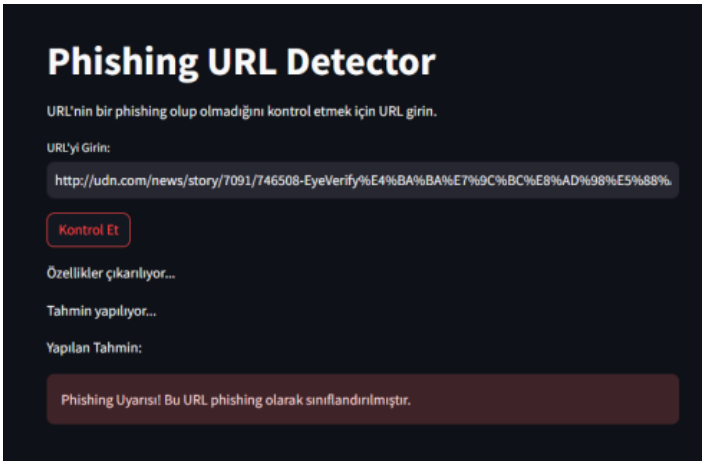
Model Training and Evaluation

Models were trained and evaluated using five different machine learning algorithms. Model performances were evaluated using the 5-fold cross-validation method and the results were presented in tables.

Results and Evaluation

In evaluations made using five different algorithms, the highest accuracy rates were obtained with XGBoost and Random Forest. Feature selection increased model performance and made the model run faster. The results obtained in this project demonstrate the effectiveness of machine learning algorithms in detecting phishing attacks.





How does the website work?

- By changing the path of the dataset, the Jupyter Notebook file on which the model is trained will be run step by step from the beginning.

- Then, the libraries and models required in the "application" file and in the Jupyter Notebook content will be installed via the console with the "pip install" command.

- After all the requirements of the code content are prepared, open the command line and type "streamlit run yekta-application.py" and the website will be opened on the local host.

NOTE: If you encounter an error, check the libraries, models and file paths!