Phishing URLs detection using machine learning

Prepared by: Ragad Alsmadi
Albatool Qanah
Lujayn Alghodran
Supervised by: Dr.Yazan Shboul

Table of Contents

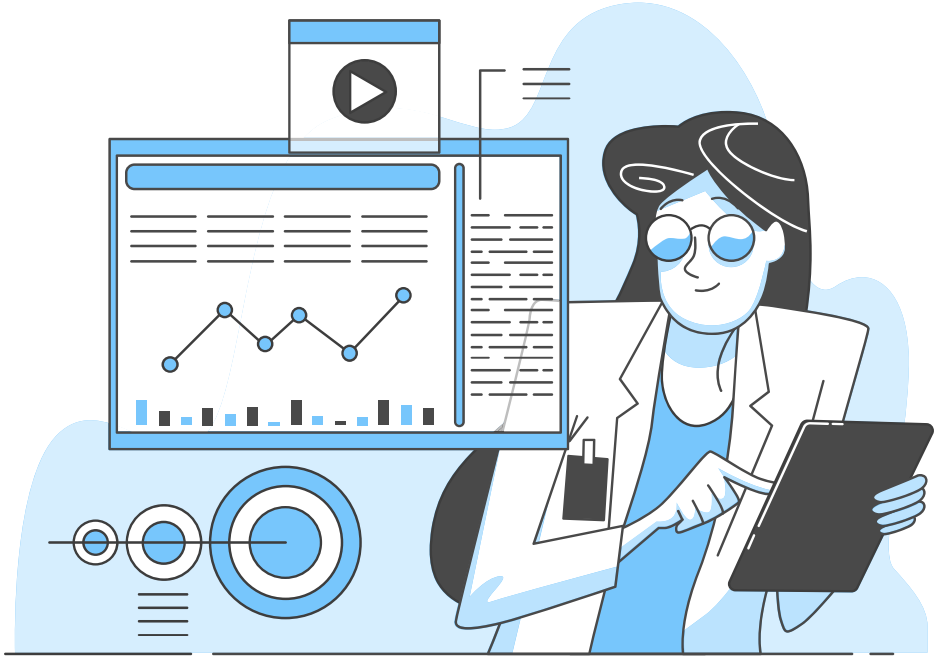
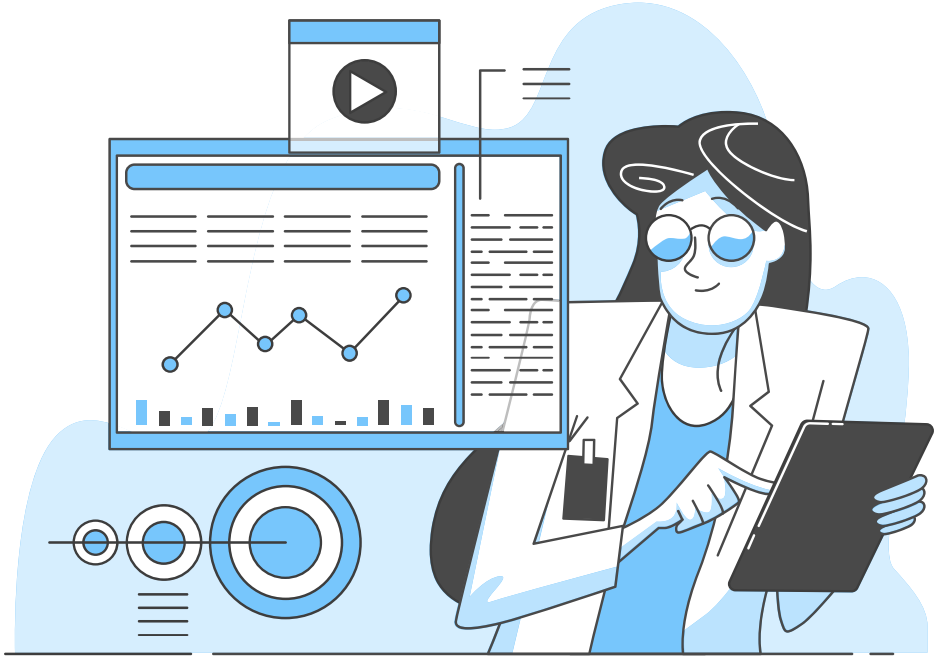


Table of Contents

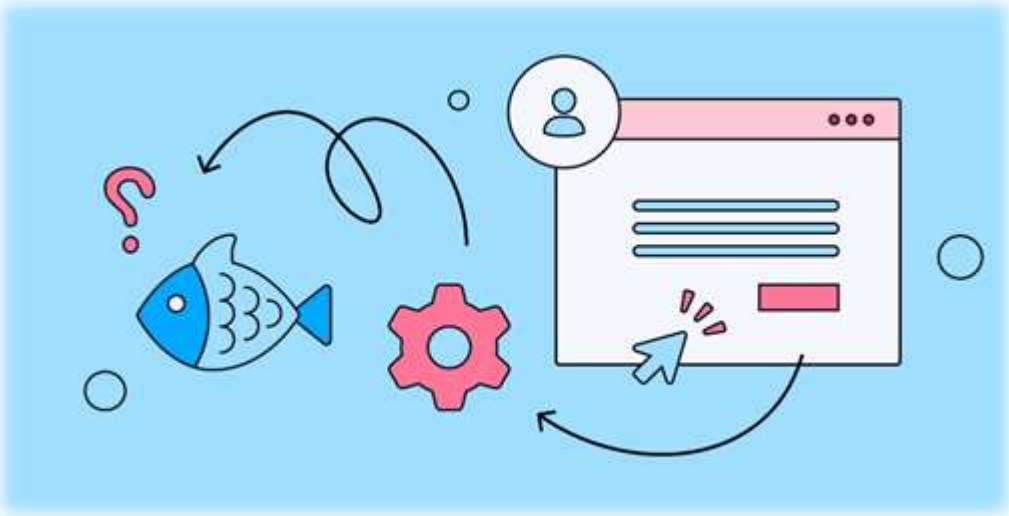


01

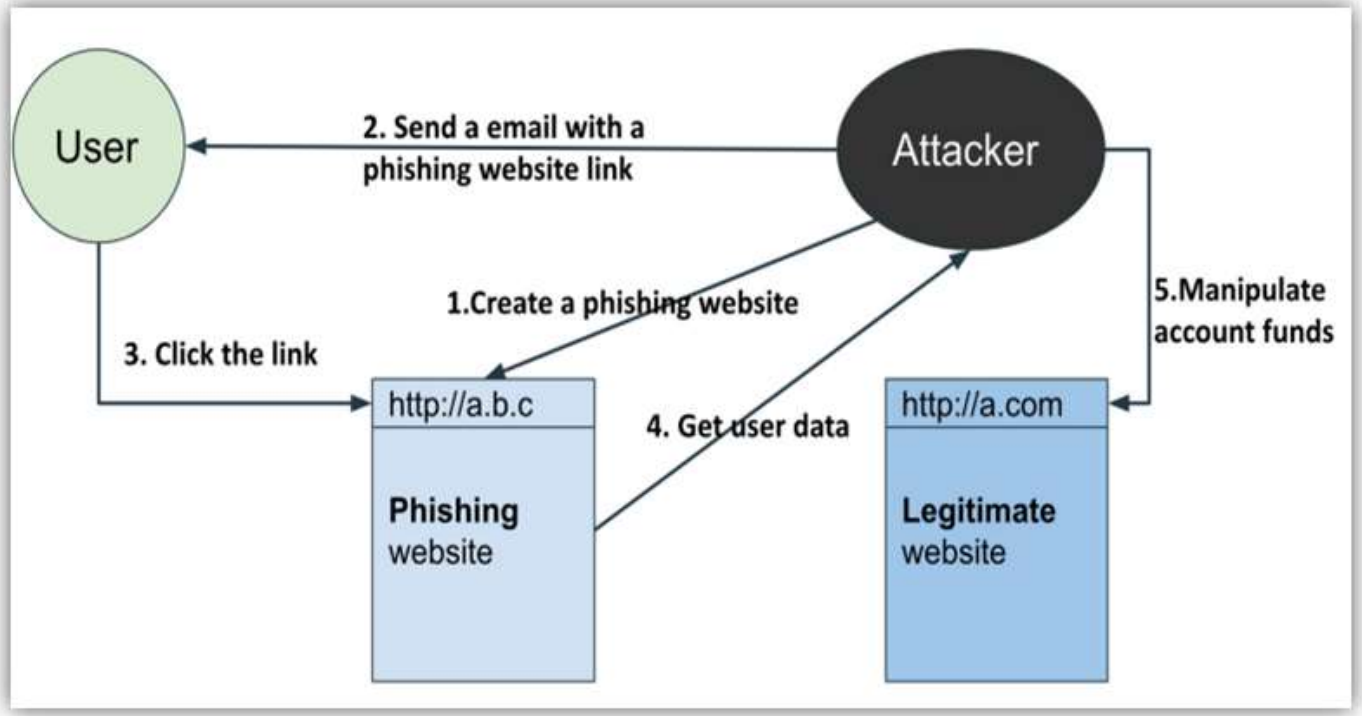
Introduction

What is phishing attack ?

Phishing is a type of social engineering attacks that takes place when a malicious website impersonates a legitimate website to gain sensitive information about user, such as passwords, account information, or MasterCard numbers.



Phishing Attack Lifecycle



Phishing URLs detection techniques

- **Blacklist**: a large database of URLs and web sites that have been known to be exploited by hackers to steal the User's sensitive information or to install harmful software on the User's device.
- Blacklists are not always successful and have a very low success rate in real-world applications; this is because of hackers' adaptability of new URLs and websites.

Phishing URLs detection techniques

- **Whitelist**: A whitelist is a list of Websites that have been approved for permitted access or privileged membership to enter a certain region of the computer world.
- Whitelisting restricts access to websites that are not on the whitelist.
- It prevents the execution of unknown or suspicious applications.

Phishing URLs detection techniques

- **Machine-learning**: In machine-learning technique, supervised learning algorithms are used to develop machine-learning models that categorize a given URL as phishing or not.
- To determine the effectiveness of each model, several algorithms are evaluated after being trained on a dataset.
- Any differences in the training data have a direct impact on the model's performance.
- This technique offers effective methods with good performance for phishing detection.

Phishing URLs detection techniques

- For identifying phishing websites, a variety of machine learning models are available, including Naive Bayes, Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, XGBoost classifiers, CatBoost classifiers, gradient boost classifier and others.
- In comparison to other techniques, this is a highly well-liked technique that has proven to be quite effective and accurate.

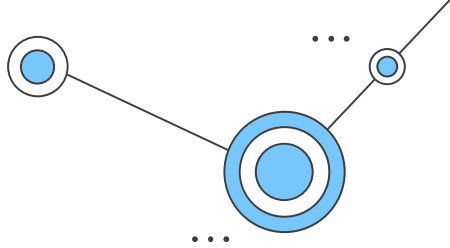
02
Problem statement

Problem statement

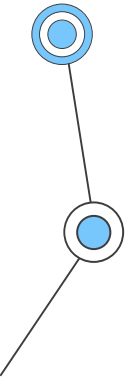
- Traditional detection techniques (whitelist , blacklist) are not enough to detect phishing urls. It needs to be updated continuously.
- To overcome this problem, we are using some of the machine learning algorithms in which will help us to identify the phishing websites based on the features present in the algorithm. By using these algorithms, we can be able to keep the user's personal credentials or sensitive data safe from attackers.

03
Purpose of the project

Purpose of the project



1. We propose using machine learning to solve the problems with traditional techniques for phishing detection.
2. Due to the ease with which sensitive data on phishing attack patterns can be obtained, the problem of phishing detection makes it a prime candidate for the use of machine learning methods.
3. The main idea is to utilize machine learning techniques on a dataset of phishing websites to develop a model that can be used to determine if a particular web page is a phishing page or a legitimate webpage in real-time.



04
Implementation

Implementation

The process of implementing the project:

- Search for a dataset containing phishing and legitimate websites.
- Divide the dataset into training and testing sets.
- Run selected machine learning models like SVM, Random Forest on the dataset.
- Run code of selected feature selection techniques like CFS, MI and ANOVA.
- Compare the obtained results for trained models and specify which is better.
- Run the feature extraction code, call the TrainModel code that contains the trained model, and deploy the project using Streamlit.

05
Methodology

Methodology

01

Dataset

02

Train-test split

03

Machine Learning Models

04

Model comparison

05

Feature Selection

Dataset

- In this model, we have used a phishing dataset from Kaggle.
- A dataset is a table that contains information about phishing and legitimate websites.
- It contains 11,055 URLs. Each row has 30 features.
- Each features is associated with a rule. If the rule is met, the URL is phishing. If the rule does not meet, the URL is legitimate.
- The features have three distinct values:
    - ‘1’ if the rule is satisfied,
    - ‘0’ if the rule is partially satisfied,
    - ‘-1’ if the rule is not satisfied.

Dataset

The following components are used to detect and classify phishing websites:

1. Address Bar based Features :

    - Using the IP address (UsingIP)
    - Long URL to hide the Suspicious Part (LongURL)
    - Using URL shortening services TinyURL (ShortURL)
    - URLs having @ symbol (Symbol@)
    - Redirecting using // (Redirecting//)
    - Adding Prefix or Suffix Separated by (-) to the Domain (PrefixSuffix-)

Dataset

1. *Address Bar based Features (CONT)*:

    - Sub Domain and Multi Sub Domains *(SubDomains)*
    - HTTPs (Hyper Text Transfer Protocol with Secure Sockets Layer) *(HTTPS)*
    - Domain Registration Length *(DomainRegLen)*
    - Favicon *(Favicon)*
    - Using Non-Standard Port *(NonStdPort)*
    - The existence of HTTPS Token in the Domain Part of the URL *(HTTPSDomainURL)*

Dataset

2. Abnormal-Based Features :

   - Request URL (RequestURL)
   - URL of Anchor (AnchorURL)
   - Links in <meta>, <Script> and <Link> tags (LinksInScriptTags)
   - Server From Handler(SFH) (ServerFormHandler)
   - Submitting Information to Email (InfoEmail)
   - Abnormal URL (AbnormalURL)

Dataset

3. HTML and JavaScript Based Features:

    - Website Forwarding (WebsiteForwarding)
    - Status Bar Customization (StatusBarCust)
    - Disabling Right Click (DisableRightClick)
    - Using Pop-Up Window (UsingPopupWindow)
    - IFrame Redirection (IframeRedirection)

Dataset

4. Domain-Based Features:

- Age of Domain (AgeofDomain)
- DNS Record (DNSRecording)
- Website Traffic (WebsiteTraffic)
- Page Rank (PageRank)
- Google Index (GoogleIndex)
- Number of Links Pointing to Page (LinksPointingToPage)
- Statistical-Reports Based Feature (StatsReport)

Train-test split

- The dataset is splitted into two subsets: testing set and training set. the training dataset can be used with the algorithms and then used for detecting the phishing websites on testing dataset.

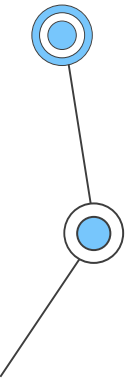
- In this model, A 20% of phishing dataset from Kaggle is used to test our model and then the other 80% of the dataset is used for model training.

Machine Learning Models

- This is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.
- This data set comes under classification problem, as the input URL is classified as phishing (0) or legitimate (1).
- The machine learning models used :
  - Gradient Boosting Classifier
  - Catboost Classifier
  - Multi-layer Perceptrons
  - XGBoost Classifier
  - Random Forest
  - Support Vector Machines

Machine Learning Models

- Decision Tree
- K-Nearest Neighbors
- Logistic Regression
- Naïve Bayes Classifier



Model comparison

- The models are evaluated, and the considered metric is accuracy.
- Here are the results of the accuracy of machine learning models before applying feature selection:

MI-Model	Accuracy	F1_Score	Recall	Precision	Time
gradient boosting classifier	0.974	0.977	0.994	0.986	1.139
catboost classifier	0.972	0.975	0.994	0.989	7.103
multi-layer perceptron	0.971	0.974	0.992	0.985	0.893
xgboost classifier	0.969	0.973	0.993	0.984	11.914
random forest	0.967	0.970	0.992	0.991	0.094
support vector machine	0.964	0.968	0.980	0.965	17.512
decision tree	0.961	0.965	0.991	0.993	0.031
k-nearest neighbors	0.956	0.961	0.991	0.989	2.150
logistic regression	0.934	0.941	0.943	0.927	0.079
naïve bayes classifier	0.605	0.454	0.2992	0.997	0.031

Model comparison

- Depending on the results above we see that the Gradient Boosting Classifier model provides better performance when compared to the other models.

Feature Selection

- In machine learning, feature selection is the process of selecting a subset of relevant features from the data to improve the accuracy and efficiency of the model.
- We used some techniques for feature selection such as Correlation-based (CFS) , Mutual Information (MI), and ANOVA .

Feature Selection

Below are the results of the CFS after setting correlation coefficient values (0.8) with (26 features):

	ML Model	Accuracy	f1_score	Recall	Precision	Time
0	CatBoost Classifier	0.972	0.975	0.993	0.990	7.482
1	Gradient Boosting Classifier	0.970	0.973	0.991	0.986	1.108
2	XGBoost Classifier	0.967	0.971	0.992	0.985	1.032
3	Multi-layer Perceptron	0.967	0.971	0.990	0.984	14.306
4	Random Forest	0.967	0.970	0.994	0.988	0.106
5	Support Vector Machine	0.965	0.969	0.978	0.962	16.957
6	Decision Tree	0.959	0.964	0.990	0.993	0.030
7	K-Nearest Neighbors	0.955	0.960	0.991	0.989	1.710
8	Logistic Regression	0.932	0.940	0.946	0.926	0.052
9	Naive Bayes Classifier	0.600	0.443	0.286	0.996	0.030

Feature Selection

- Below are the results of the Ml after selecting a specific number of features (25 features):

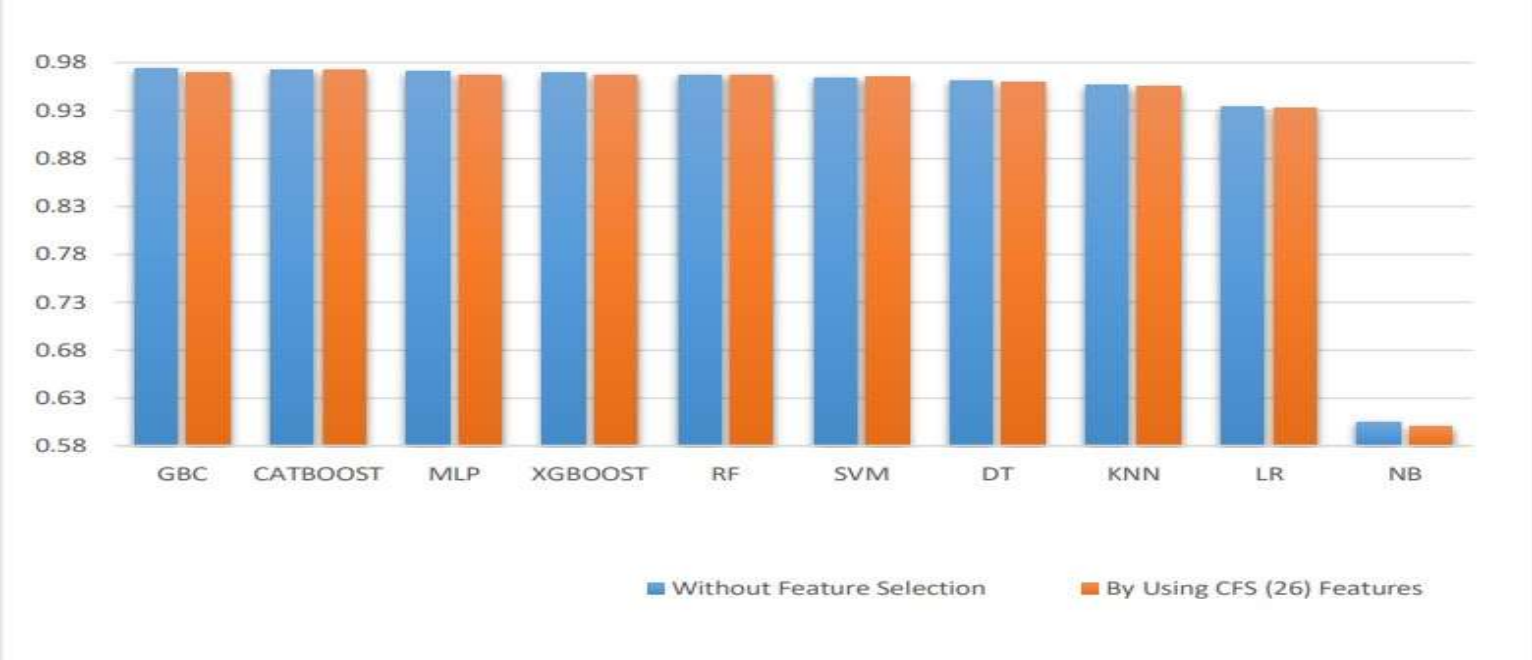
	ML Model	Accuracy	f1_score	Recall	Precision	Time
0	CatBoost Classifier	0.972	0.975	0.995	0.988	6.965
1	XGBoost Classifier	0.970	0.973	0.991	0.985	0.878
2	Gradient Boosting Classifier	0.967	0.971	0.995	0.983	0.958
3	Multi-layer Perceptron	0.967	0.970	0.996	0.980	11.611
4	Random Forest	0.965	0.969	0.992	0.990	0.079
5	Decision Tree	0.963	0.967	0.991	0.992	0.015
6	Support Vector Machine	0.961	0.966	0.978	0.961	13.873
7	K-Nearest Neighbors	0.955	0.960	0.990	0.989	2.181
8	Logistic Regression	0.932	0.940	0.947	0.927	0.047
9	Naive Bayes Classifier	0.598	0.438	0.284	0.998	0.016

Feature Selection

- Below are the results of the ANOVA after selecting a specific number of features (25 features):

	ML Model	Accuracy	f1_score	Recall	Precision	Time
0	CatBoost Classifier	0.972	0.975	0.994	0.988	7.447
1	Gradient Boosting Classifier	0.970	0.974	0.993	0.985	1.071
2	XGBoost Classifier	0.967	0.971	0.991	0.986	0.868
3	Random Forest	0.966	0.969	0.992	0.989	0.092
4	Support Vector Machine	0.962	0.967	0.978	0.961	17.058
5	Multi-layer Perceptron	0.961	0.965	0.978	0.991	13.332
6	Decision Tree	0.960	0.964	0.990	0.993	0.032
7	K-Nearest Neighbors	0.957	0.962	0.990	0.989	2.286
8	Logistic Regression	0.933	0.940	0.946	0.927	0.056
9	Naive Bayes Classifier	0.602	0.449	0.289	0.996	0.016

Feature Selection

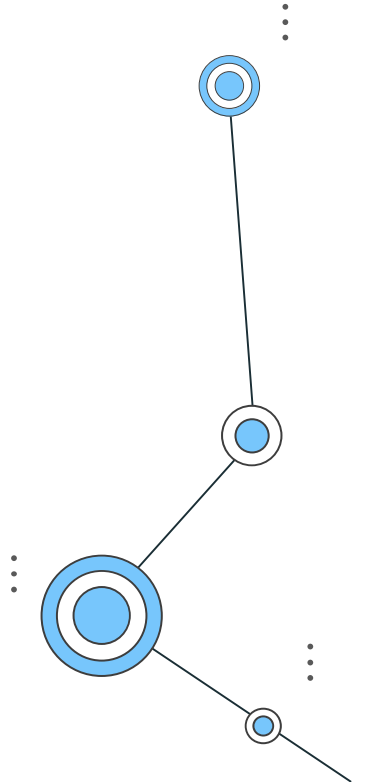


- Based on the chart above the results of the accuracy before feature selection was better. Therefore, we decided to keep the dataset as it is to get the best accuracy and performance.

06
Conclusions

Conclusions

- Phishing is a huge threat to the security and privacy of users personal credential.
- Phishing detection is critical point to focus on to reduce the possibility of phishing attacks.
- We then chose the best algorithm based on its performance and designed a streamlit tool for detecting phishing websites.
- The tool allows end users to easily check if the URL is phishing or legitimate.
- Our intent in the future is to make the tool integrated into systems or an extension on web browsers.



Thanks!

Any questions?