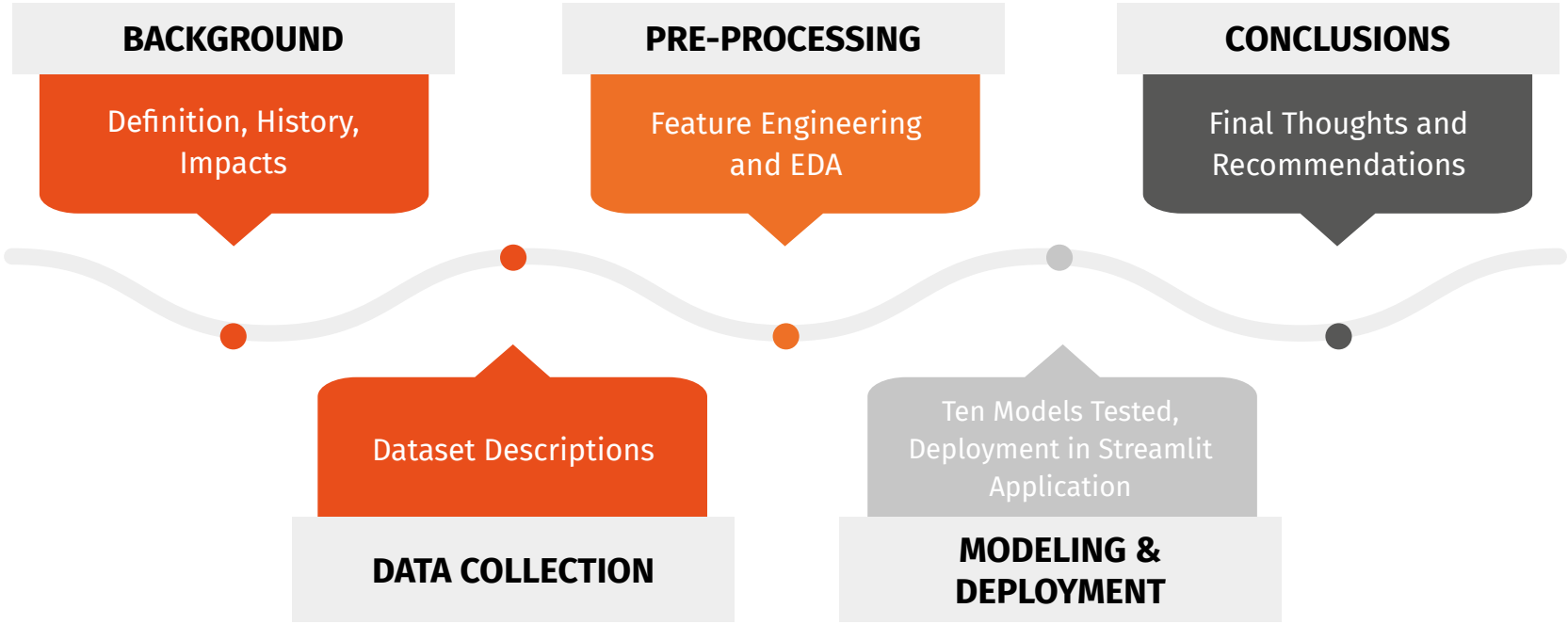
Phishing URL Detection

Katie Sylvia

Presentation Outline



BACKGROUND

What is phishing?

Phishing is a form of cybercrime in which a target is contacted via email, telephone, or text message by an attacker disguising as a reputable entity or person. The attacker then lures individuals to counterfeit websites to trick recipients into providing sensitive data.

The purpose of this project is to help individuals identify these phishing URLs in order to provide safer practices online.

Types of Phishing Tactics

Email

96% of phishing attacks arrive by email.

Telephone

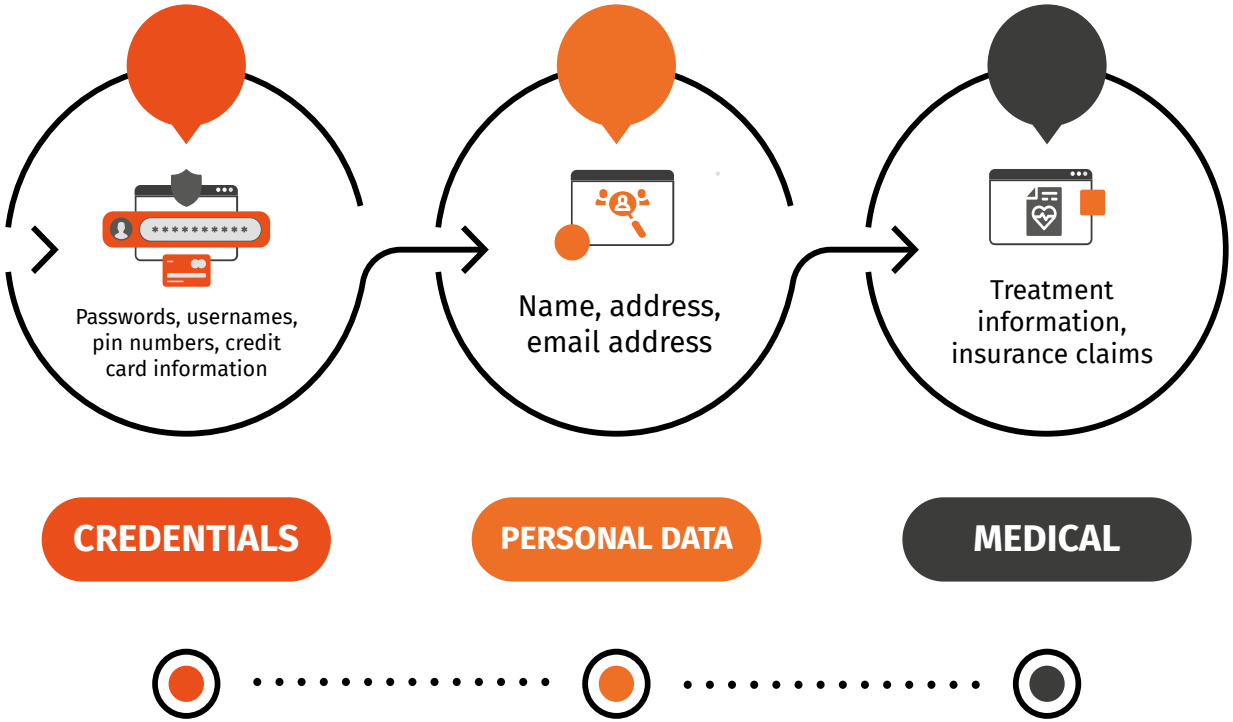
3% of phishing attacks is done over the telephone. This is also known as *vishing*.

Text Message

1% of phishing attacks is done via text message. This is also known as *smishing*.

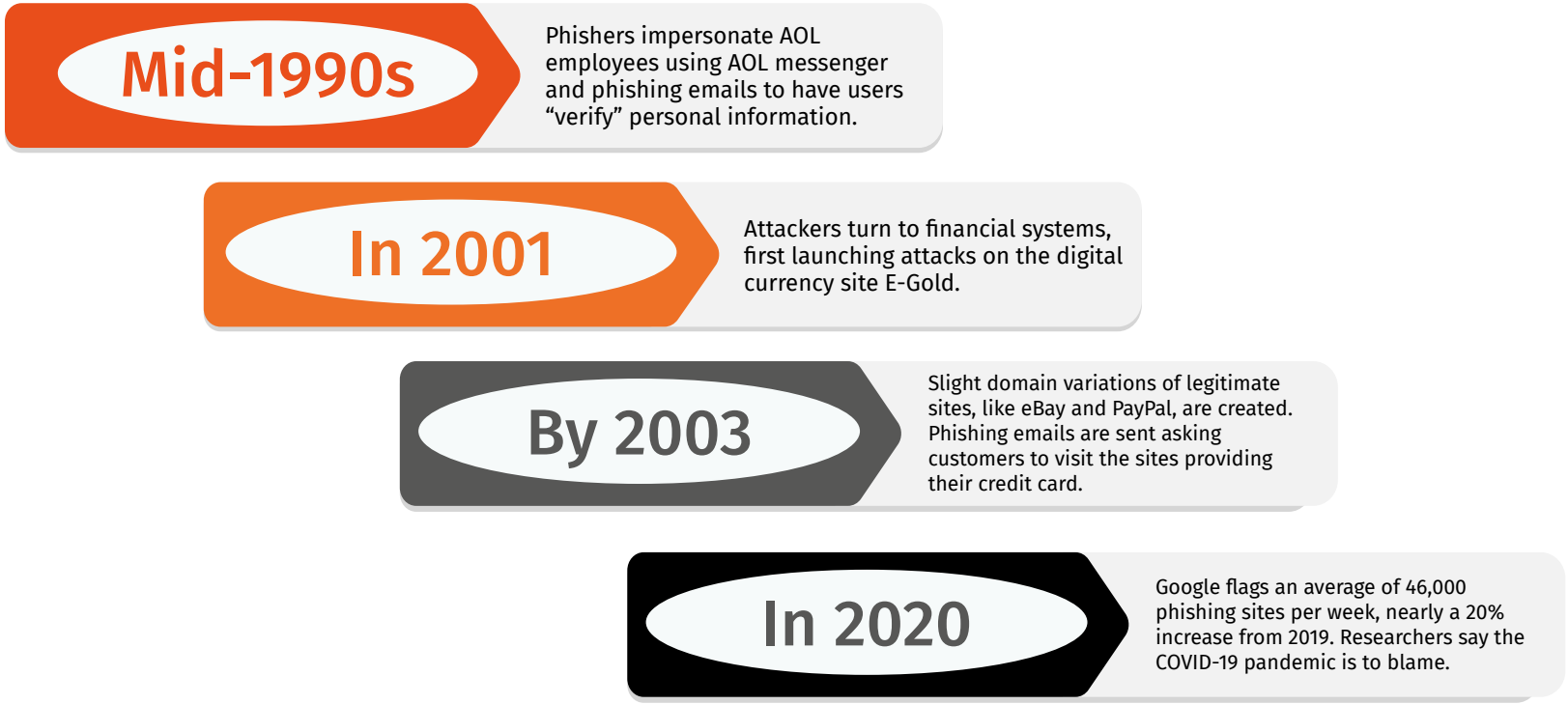
Source: www.tessian.com

Top Three Types of Data



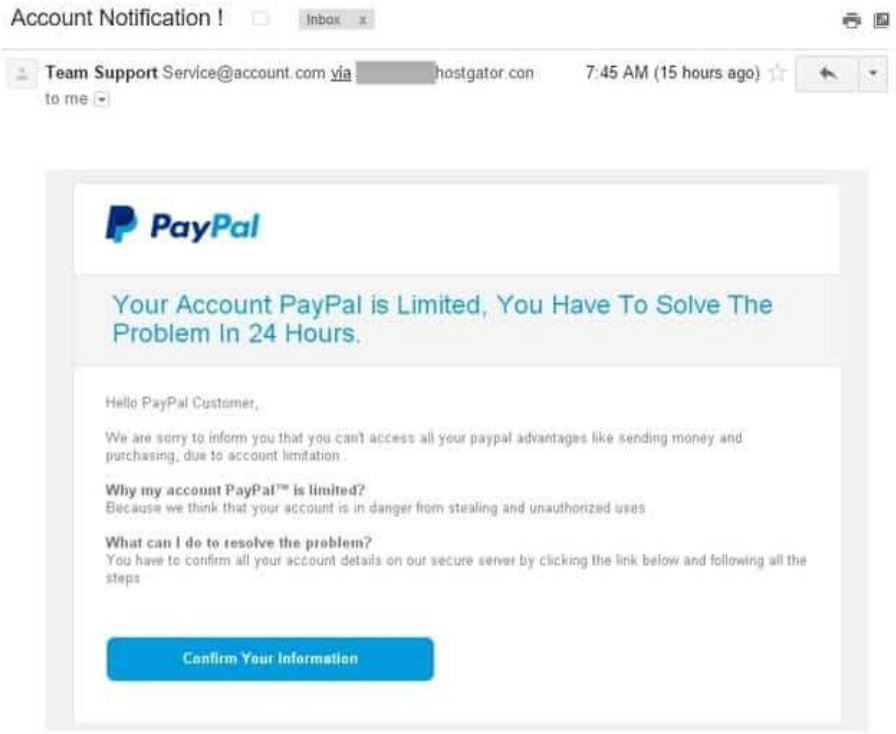
Source: Verizon's 2021 Data Breach Investigations Report

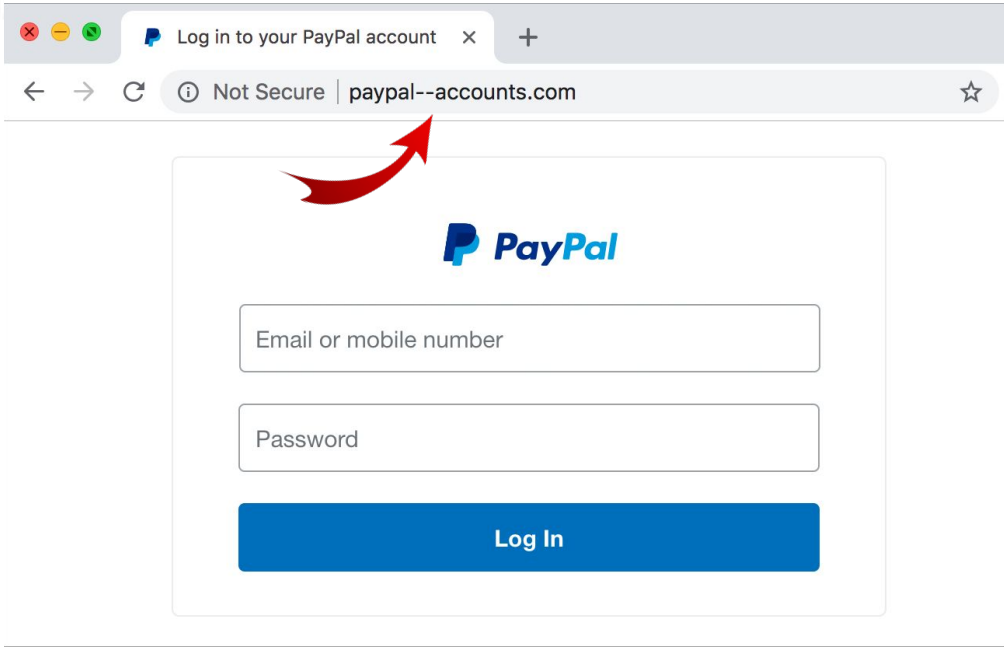
History of Phishing



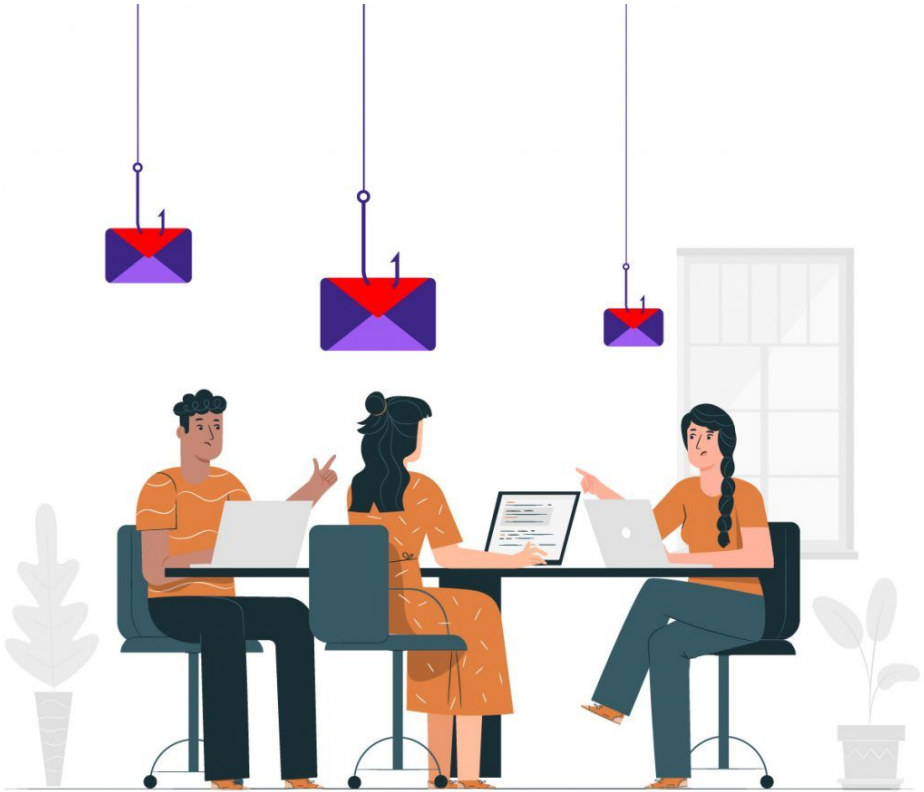
Source: www.knowbe4.com

Paypal


Account Notification !
Inbox x
Team Support Service@account.com via hostgator.con
7:45 AM (15 hours ago)
to me
PayPal
Your Account PayPal is Limited, You Have To Solve The Problem In 24 Hours.
Hello PayPal Customer,
We are sorry to inform you that you can't access all your paypal advantages like sending money and purchasing, due to account limitation .
Why my account PayPal™ is limited?
Because we think that your account is in danger from stealing and unauthorized uses
What can I do to resolve the problem?
You have to confirm all your account details on our secure server by clicking the link below and following all the steps
Confirm Your Information


Log in to your PayPal account
Not Secure
paypal--accounts.com
PayPal
Email or mobile number
Password
Log In

2021 Phishing Statistics

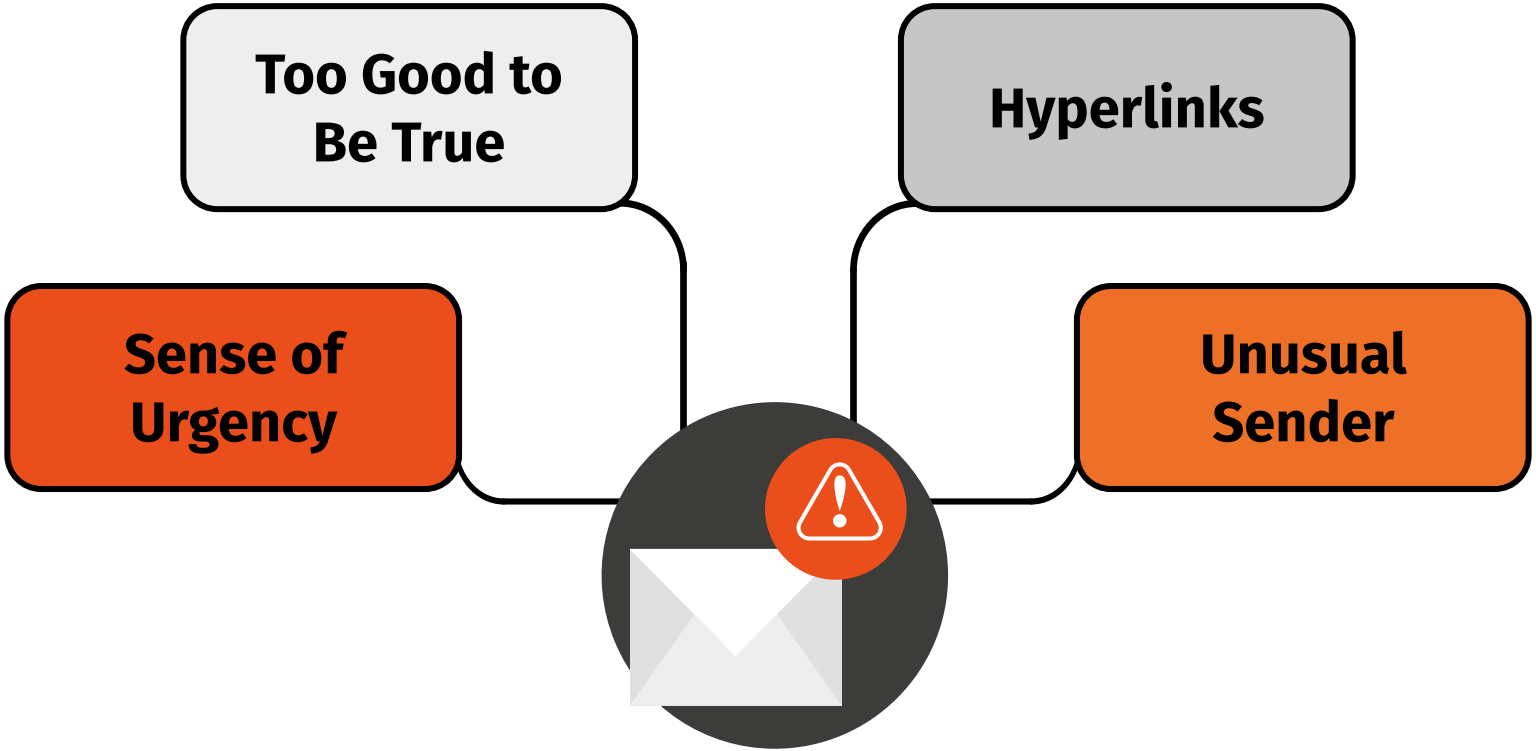


30% of phishing emails are opened by users, and **12%** of these targeted users click on the malicious link or attachment.

97% of the users are unable to recognize a sophisticated phishing email.

Source: Verizon's 2021 Data Breach Investigations Report

Common Features of Phishing Emails



Source: www.phishing.org

Common Subject Lines in Phishing Emails

(Q4 of 2020)

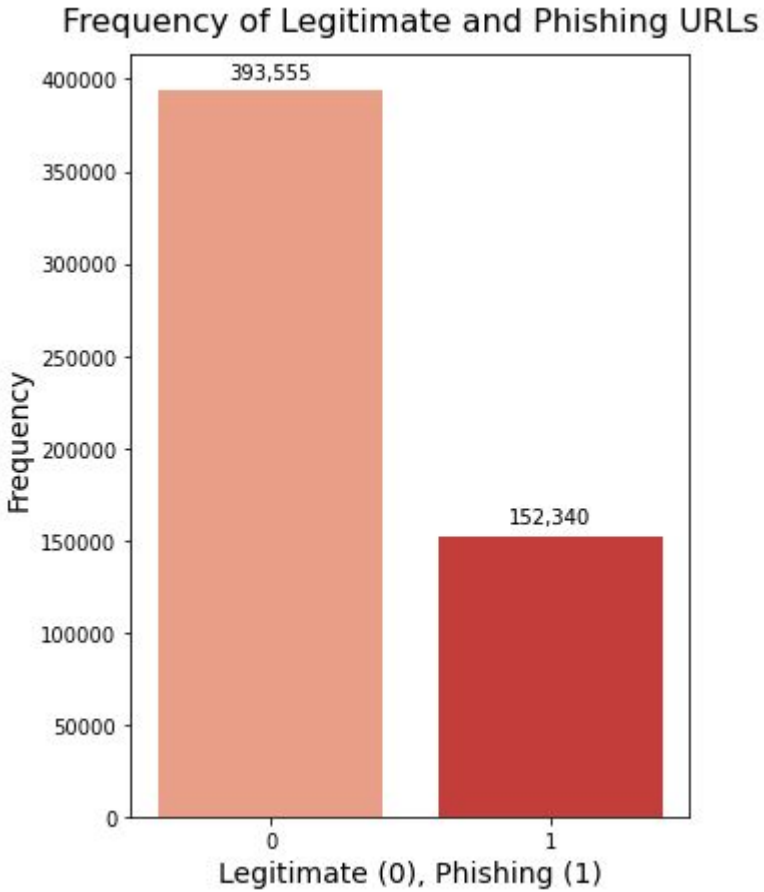
1. COVID-19 Remote Work Policy Update
2. Changes to your health benefits
3. Zoom: Scheduled Meeting Error
4. Twitter: Security alert: new or unusual Twitter login
5. Google Pay: Payment sent
6. Stimulus Cancellation Request Approved
7. Company Policy Notification: COVID-19 - Test & Trace Guidelines
8. Vacation Policy Update

Source: www.knowbe4.com

According to the FBI, phishing incidents nearly doubled in frequency, from 114,702 incidents in 2019, to 241,324 incidents in 2020. The increase in remote work could be to blame.

As the internet becomes a major mode for economic transactions and communications, online trust and cybercrimes have increasingly become an important area of study.

DATA COLLECTION

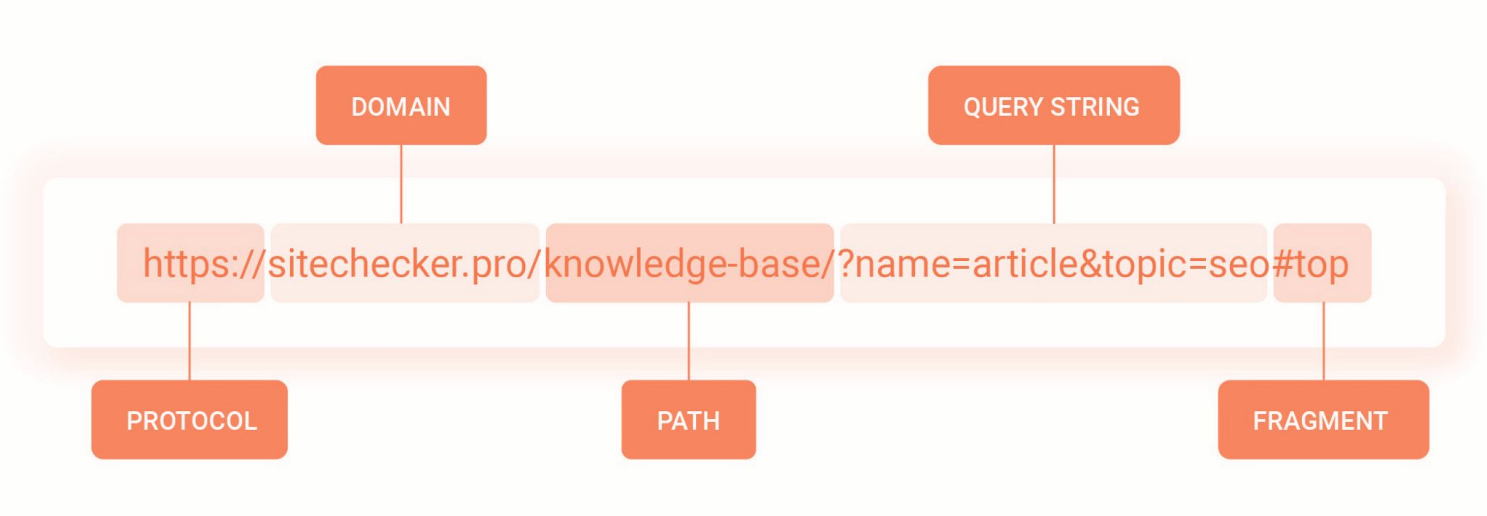


Data was collected from two separate datasets. Phishing URLs were pulled from websites such as PhishTank and OpenPhish and legitimate URLs were pulled from websites such as Alexa and Common Crawl.

The were 545,895 instances in total with a 72.1% baseline.

PRE-PROCESSING

Feature Extraction



Using a function from urllib library, protocol, domain, path, query, and fragment were extracted from the URL and respective columns were created. The protocol column was dropped as more sophisticated phishing URLs are labeled secure with https://.

Feature Extraction

Length of URL, domain, path, query, and fragment are extracted.

Quantity of specific characters in URL, domain, path, query, and fragment are extracted. These characters include:

-	=	!	+	$
.	@	~	*	%
?	&	,	#	space

65 Total Features Used in Model

MODEL SELECTION & EVALUATION

Models Tested

- Stochastic Gradient Descent Classifier
- Logistic Regression
- Support Vector Machine
- AdaBoost
- Gradient Boost
- Decision Tree Classifier
- Bagging Classifier
- K-Nearest Neighbors Classifier
- Extra Trees Classifier
- Random Forest Classifier

Baseline: 72.1%

GridSearchCV and RandomizedSearchCV tools were used to optimize the highest-scoring result.

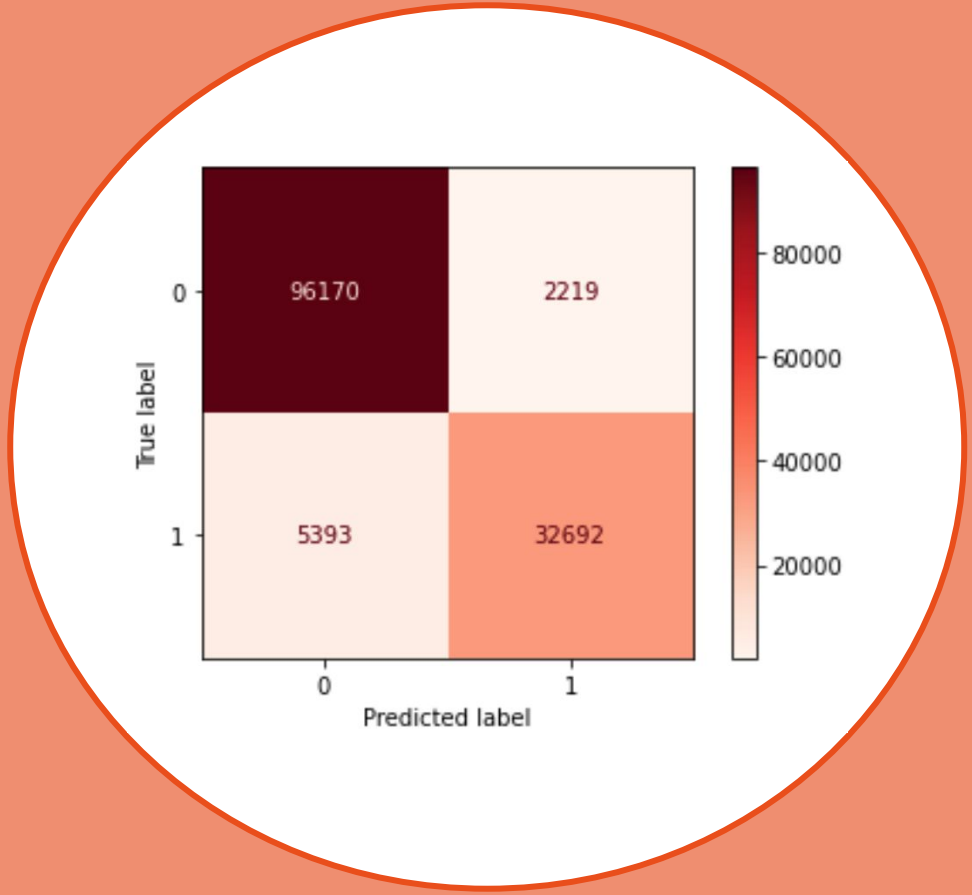
Once the best model was determined, hyperparameter tuning continued to optimize our model.

Model Selection

MODEL	TRAINING SCORE	TESTING SCORE	USED FOR DEPLOYMENT
k-Nearest Neighbors	94.8%	93.2%	✗
Decision Trees	97.6%	94.3%	✗
Extra Trees	97.9%	94.4%	✗
Random Forest	97.0%	94.5%	✓

Model Evaluation

Accuracy	94.5%
Recall	85.8%
Specificity	97.7%
Precision	93.6%



MODEL DEPLOYMENT

main - Streamlit

localhost:8501

Select a Page:

Home

Fishing for Phishers

Image by My TechDecisions

About

Phishing is a form of cybercrime in which a target is contacted via email, telephone, or text message by an attacker disguising as a reputable entity or person. The attacker then lures individuals to counterfeit websites to trick recipients into providing sensitive data. The purpose of this application is to help individuals identify these phishing URLs in order to provide safer practices online.

Prediction

Is the URL legitimate or phishing?

Enter URL here:

www.paypal-com.uni.me/update/security/profile/index.php

Make Prediction

The model predicts there is a 60.11% chance this is a phishing URL.

Made with Streamlit

CONCLUSIONS

How to Avoid Phishing Attacks

1

STAY INFORMED

Learn about new phishing techniques that are being developed to avoid falling prey to one.

2

THINK BEFORE YOU CLICK

Never click on hyperlinks without examining the hidden URL.

3

UTILIZE ‘FISHING FOR PHISHERS’

When in doubt, use the ‘Fishing for Phishers’ app to verify the authenticity of a website.

Thank you!

Any questions?

References

"The 2021 Ponemon Cost of PHISHING Study: Proofpoint US." *Proofpoint*, 19 Aug. 2021, www.proofpoint.com/us/resources/analyst-reports/ponemon-cost-of-phishing-study.

KnowBe4. "What Is Phishing?" *Phishing*, www.phishing.org/what-is-phishing.

"Phishing Statistics (Updated 2021): 50+ Important Phishing Stats." *Tessian*, 17 May 2021, www.tessian.com/blog/phishing-statistics-2020/.

"2021 DBIR Master's Guide." Verizon Business, www.verizon.com/business/resources/reports/dbir/2021/masters-guide/.

"History of Phishing." Cofense, 28 May 2021, cofense.com/knowledge-center/history-of-phishing/.

Slide Template: slidesgo.com