Detecting Phishing Links

Team 2

Group Members: Rand Alattar, Junhyung Han, Anantha Aksshaj Erigisetty, Ruth Tadesse, Mesfin Mekonnen, Ishika Maisha, Mohamed Reda Erradi

CRISP-DM Process

01

Business Understanding

02

Data Understanding

03

Data Preparation

04

Modeling

05

Evaluation

06

Deployment

01 Business Understanding

Business Question

How can businesses and organizations identify if a URL they have clicked on is part of a phishing attack?

Background

What is Phishing?

- Phishing is an attack designed to steal personal or sensitive information from individuals or businesses.
- This is typically done by imitating legitimate websites, people, or organizations.
- Our focus is on identifying phishing websites.

Why is it Important?

- Phishing attacks can lead to significant financial losses and breaches of sensitive data.
- As these attacks become more advanced, it's harder for individuals and businesses to recognize them.

Who Cares?

- Businesses: Protecting data and maintaining trust are critical for operations.
- Employees: Everyone, regardless of their role, can be targeted by these attacks.

How Can We Prevent Them?

- By analyzing patterns in suspicious URLs, we can develop tools to help detect phishing attempts and reduce their impact.

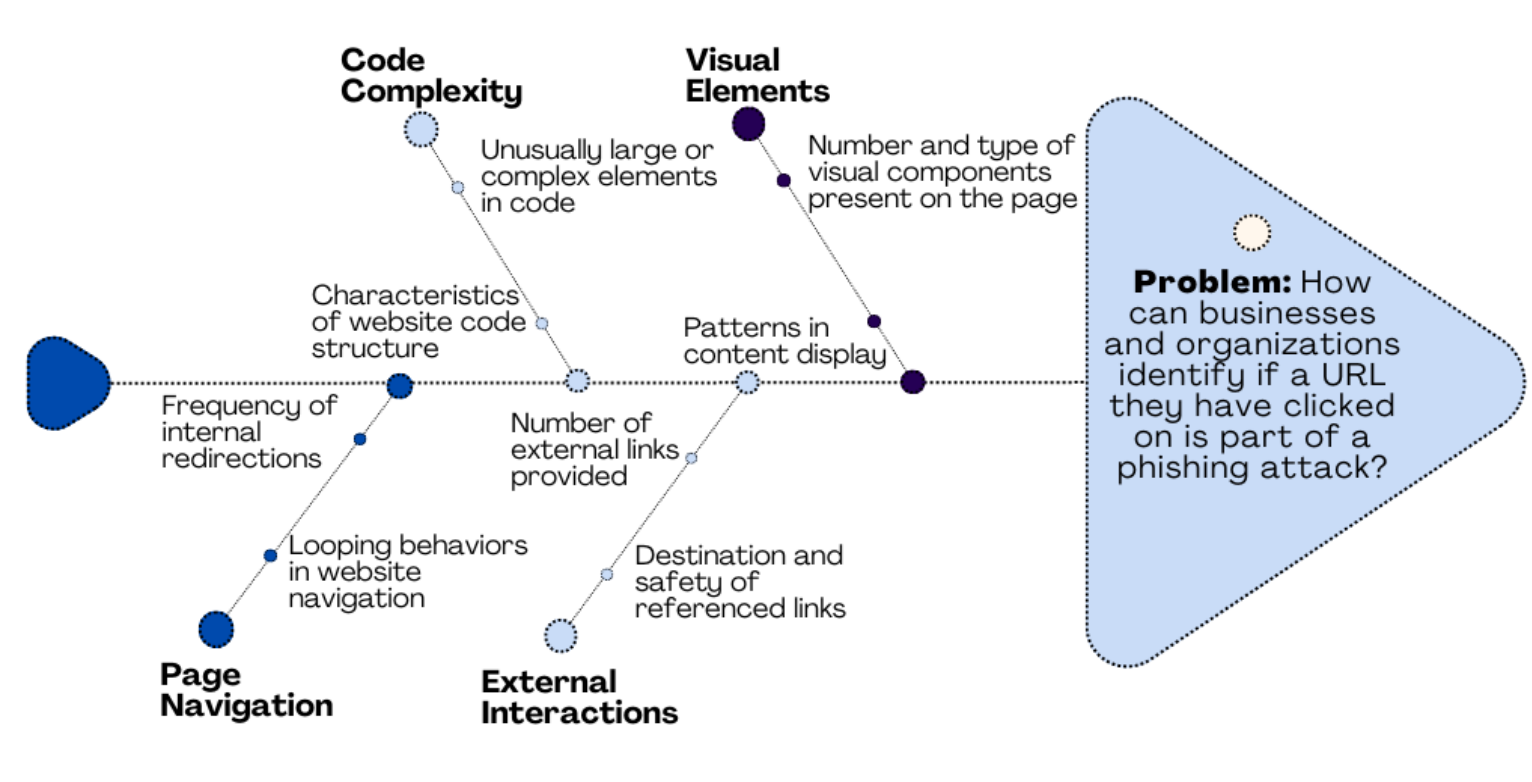
Why Our Business Question Matters

- While it would be ideal to avoid the clicking of a phishing link altogether, it is not always preventable.
- In a study done in 2012 by MIS Quarterly, most models made to detect phishing URLs fell below 70% accuracy.
  While technology improves over time, so does phishing techniques, causing issues to persist even to this day.
- Our study can be later using in collaboration with a model that identifies if a URL is phishing prior to individuals clicking on it.
- That way if any false positives occur in a different algorithm, our model can warn users to click out of the cite before it is too late.

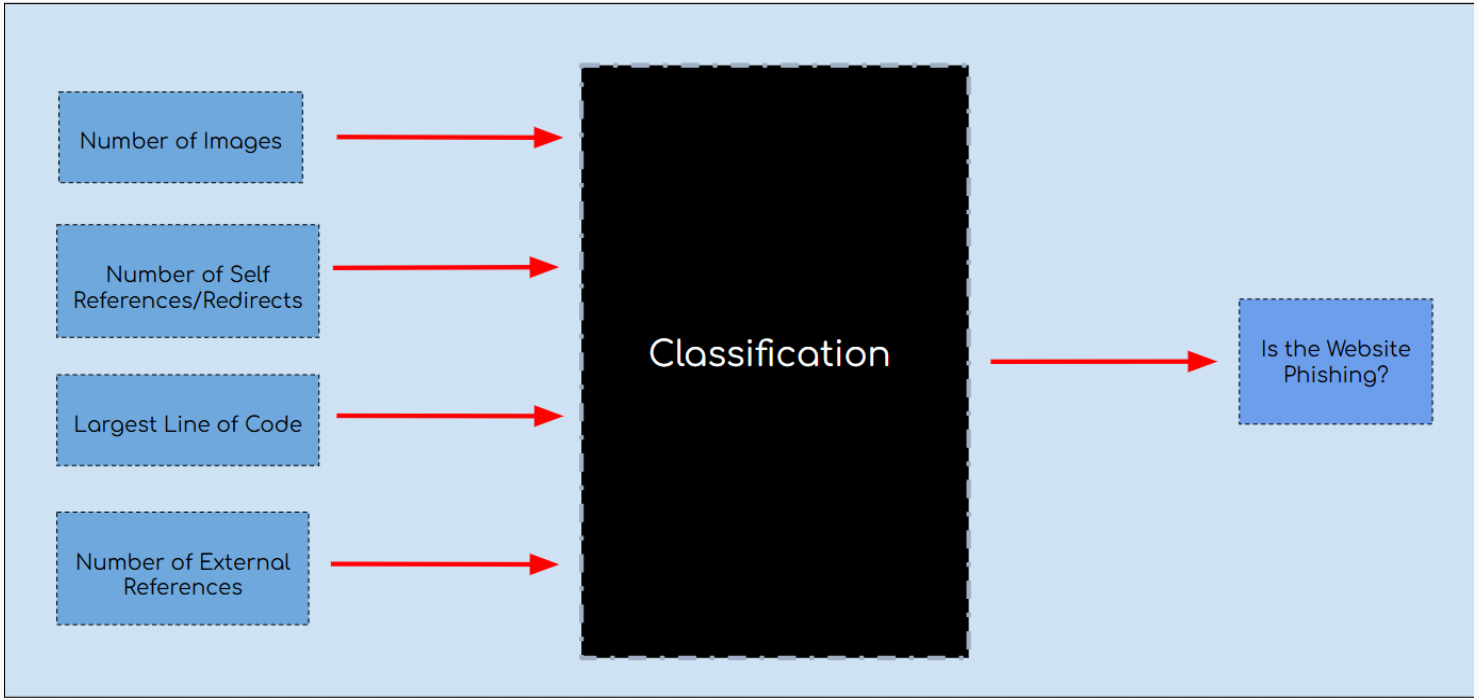
Analytic Problem

What are the most frequent features that appear with a phishing website in both its URL and its behavior with the user?

Fishbone


Code Complexity
Unusually large or complex elements in code
Visual Elements
Number and type of visual components present on the page
Characteristics of website code structure
Patterns in content display
Problem: How can businesses and organizations identify if a URL they have clicked on is part of a phishing attack?
Frequency of internal redirections
Number of external links provided
Looping behaviors in website navigation
Destination and safety of referenced links
Page Navigation
External Interactions

Black Box


Number of Images
Number of Self References/Redirects
Largest Line of Code
Number of External References
Classification
Is the Website Phishing?

02 Data Understanding

Dataset Description

- The dataset we used is PhiUSIIL_Phishing_URL
  - Includes 235,796 rows and 56 columns of data
  - Content:
    - URLs and features describing their structure and behavior (e.g., length, domain details).
    - Focused on identifying patterns related to phishing attempts.

Feature and Target Definitions

Features

- Number of Images
  - Counts the number of images on a website.
- Number of Self References / Redirects
  - Tracks how often a link redirects back to itself (when clicked on).
- Largest Line of Code
  - Counts the number of characters in the largest line of code on the website.
    - (Phishing websites can have excessive, unnecessary amounts.)
- Number of External References
  - Tracks how often a link redirects to websites outside the original link.

Target

- Label
  - Indicator in our dataset as to whether a website is phishing or not.

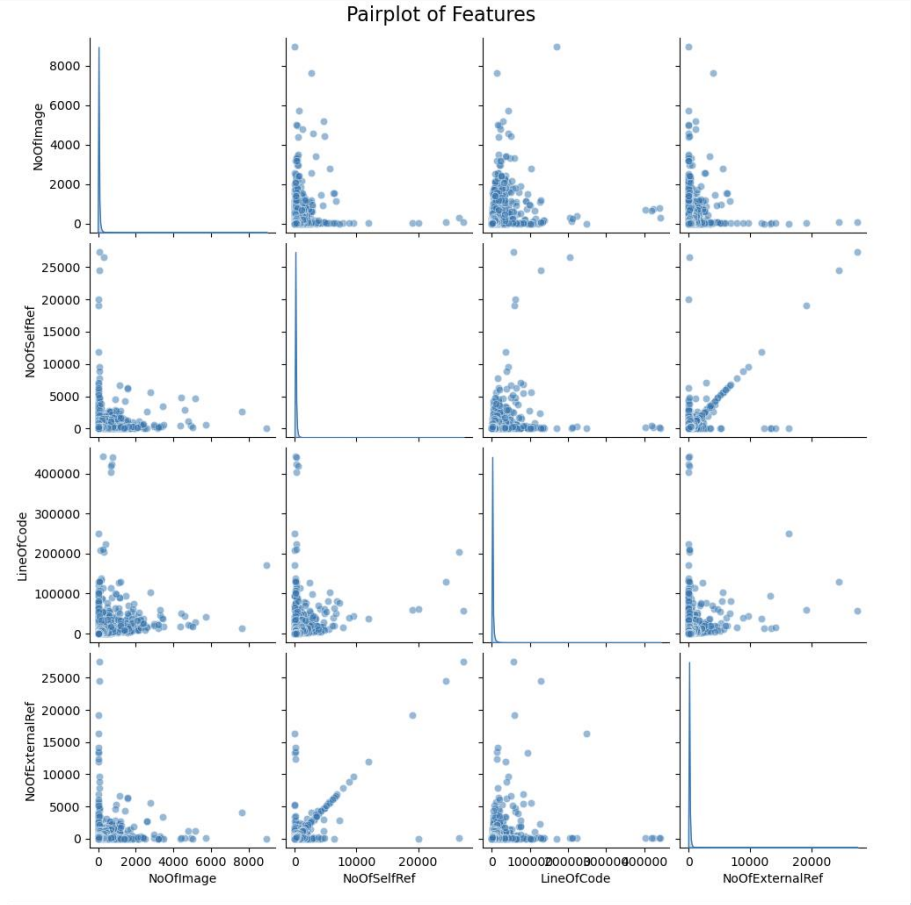
Feature Ranges and Data Types

Feature Name	Data Type	Range
Number of Images	Integer	(0, 8956)
Number of Self References	Integer	(0, 27397)
Largest Line of Code	Integer	(2, 442666)
Number of External References	Integer	(0, 27516)
Label (Target)	Boolean	(0,1)

Summary Stats

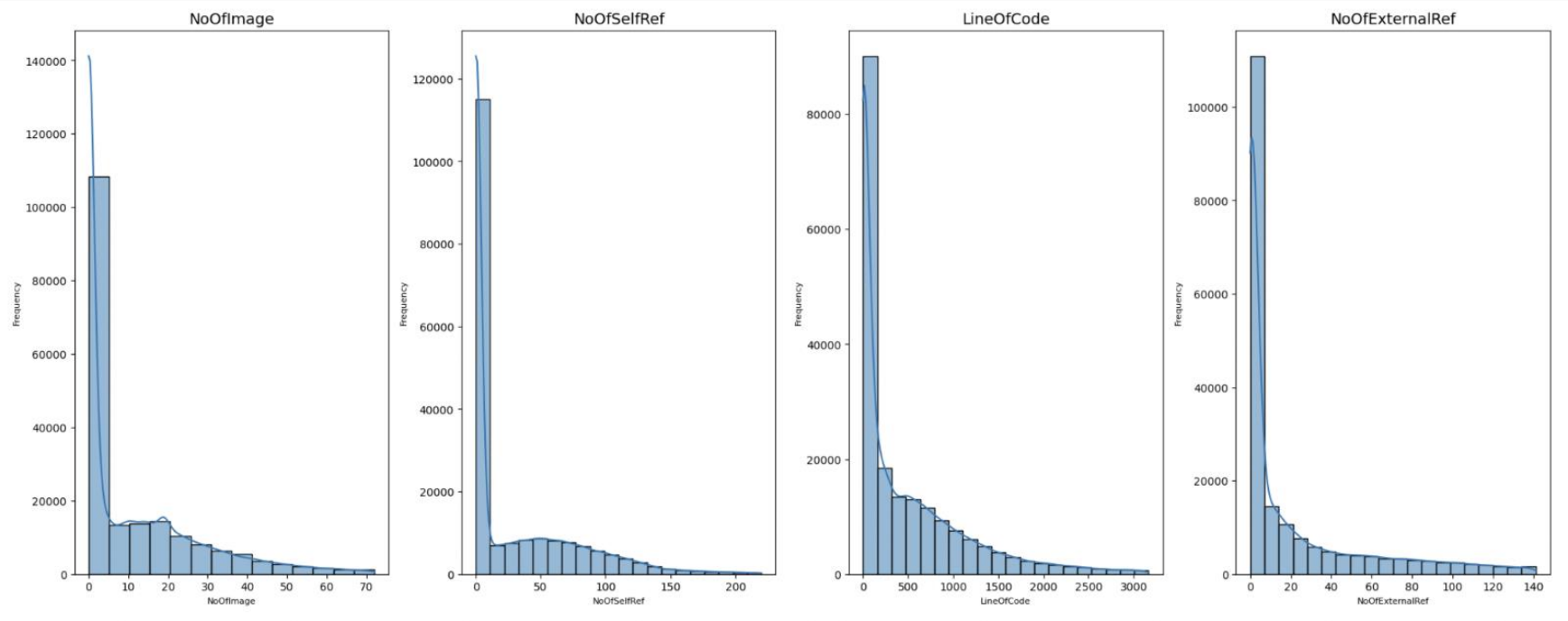
	NoOfImage	NoOfSelfRef	LineOfCode	NoOfExternalRef
Count	235795	23579	235795	235795
Mean	26.07	65.07	1141.9	49.26
Stdev	79.41	176.68	3419.95	161.02
Min	0	0	2	0
25%	0	0	18	1
50%	8	12	429	10
75%	29	88	1277	57
Max	8956	27397	442666	27516

Pair Plot


Pairplot of Features
NoOfImage
NoOfSelfRef
LineOfCode
NoOfExternalRef
0
2000
4000
6000
8000
0
5000
10000
15000
20000
25000
0
100000
200000
300000
400000
0
5000
10000
15000
20000
25000

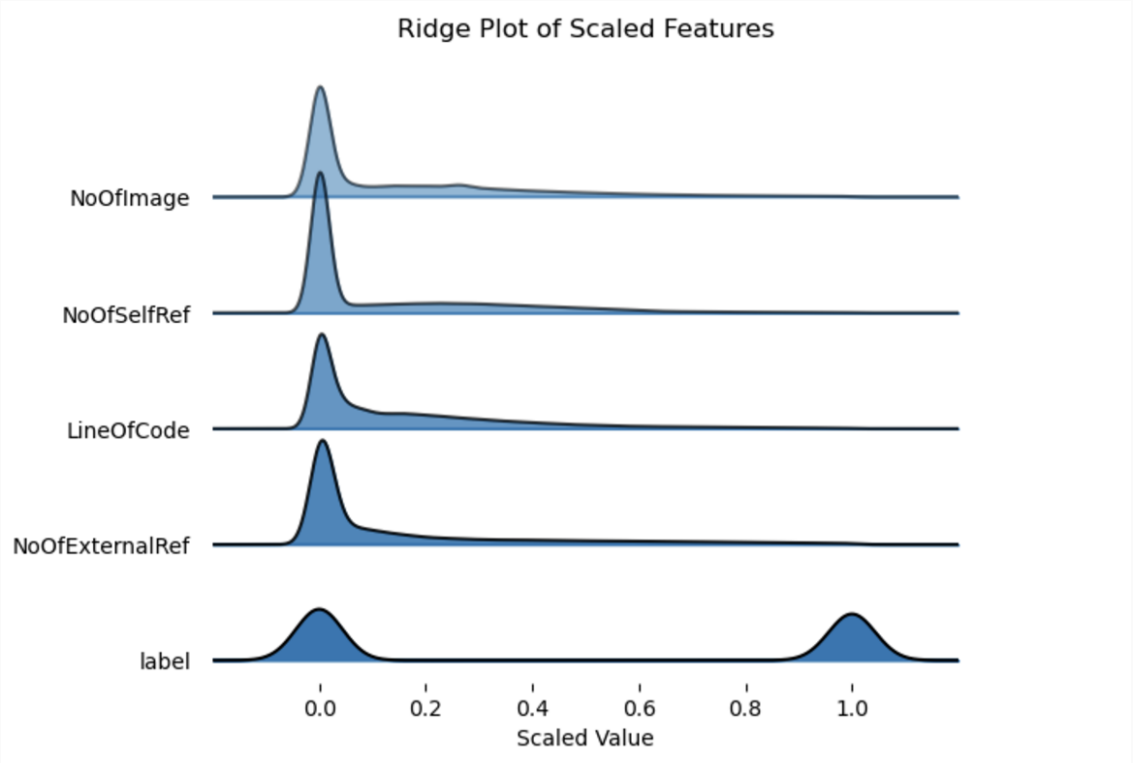
03 Data Preparation

Non-Normal Distributions



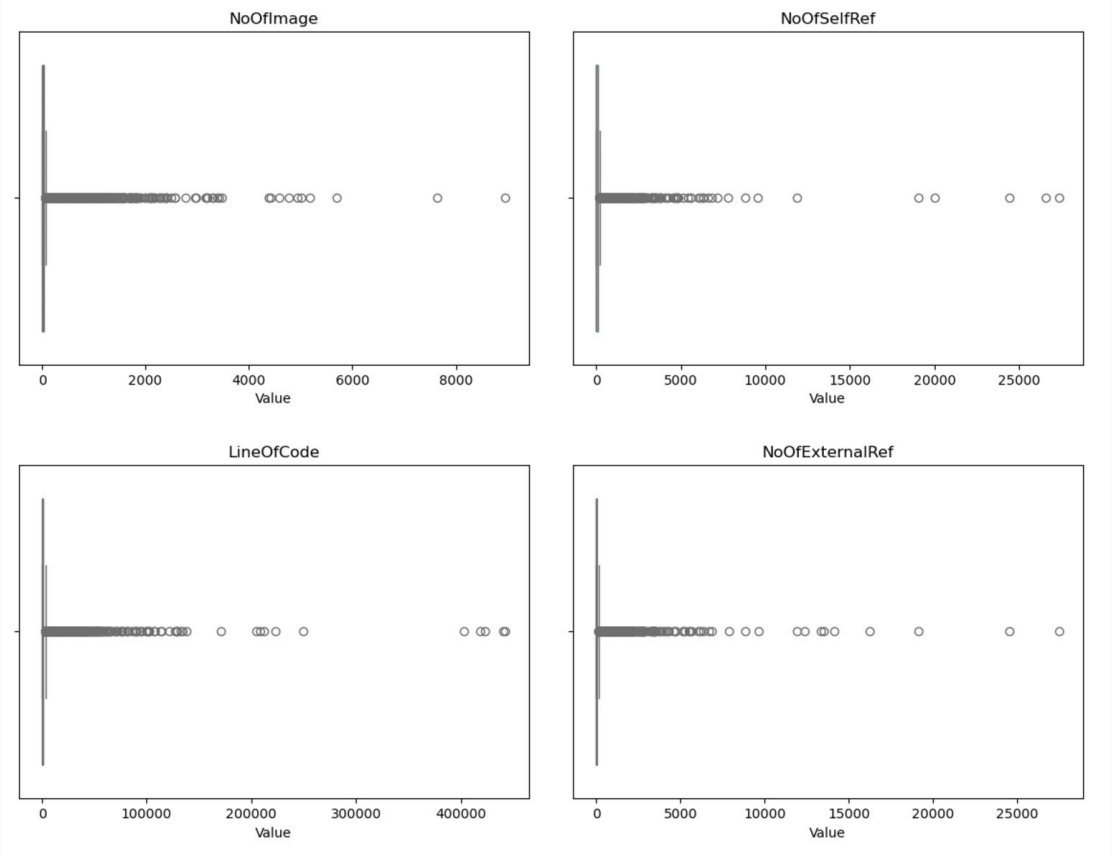
This figure has outliers removed for visibility

Scaled Features

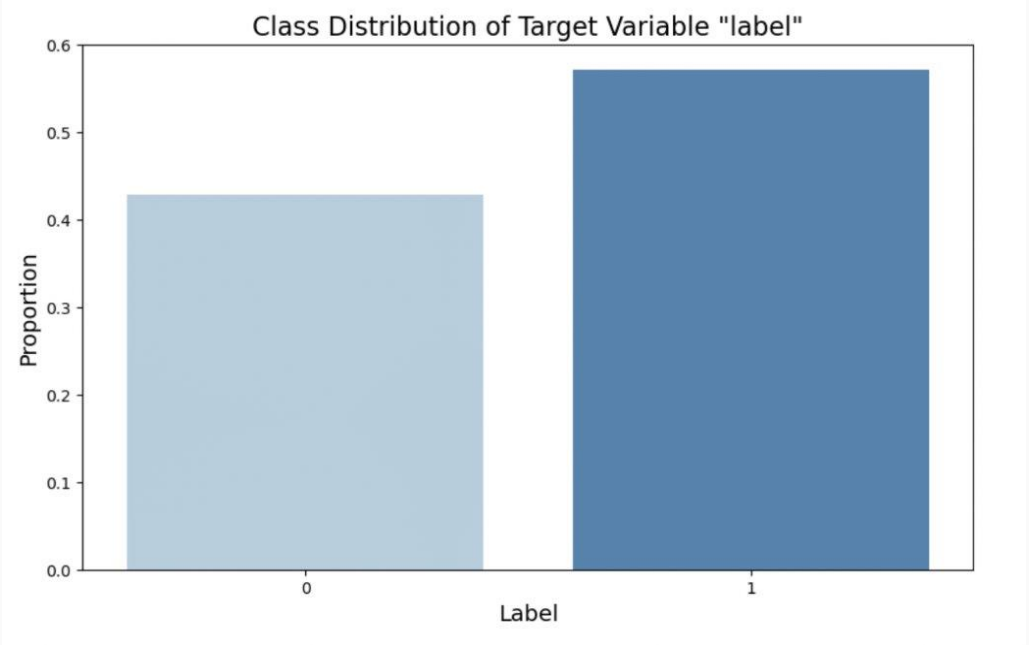


This figure has outliers removed for visibility

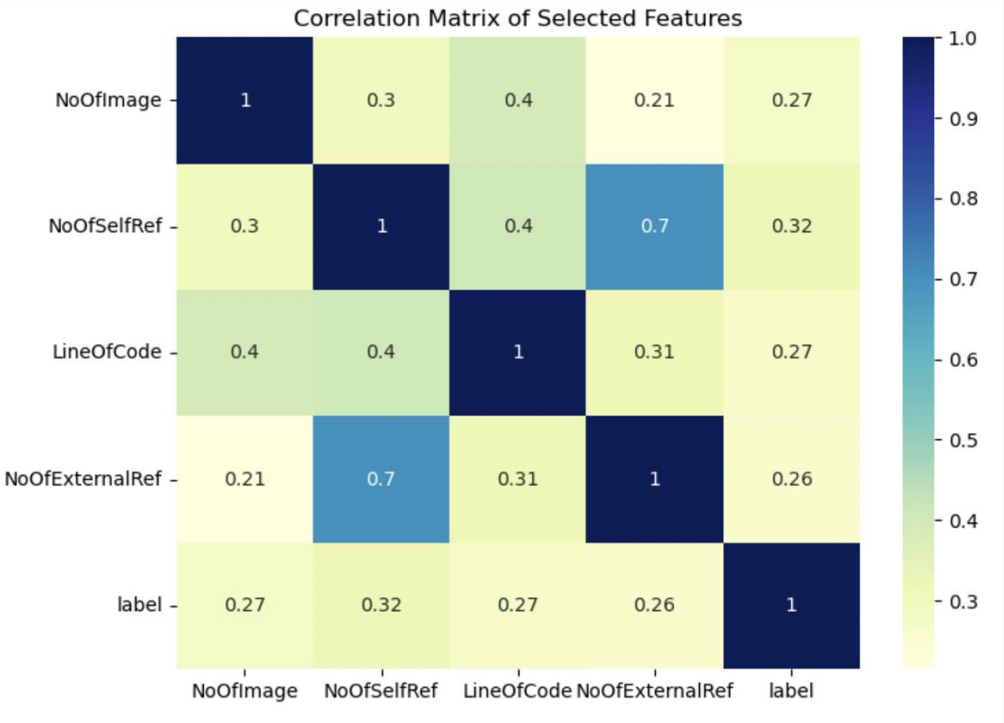
Outliers


NoOfImage
0
2000
4000
6000
8000
Value
NoOfSelfRef
0
5000
10000
15000
20000
25000
Value
LineOfCode
0
100000
200000
300000
400000
Value
NoOfExternalRef
0
5000
10000
15000
20000
25000
Value

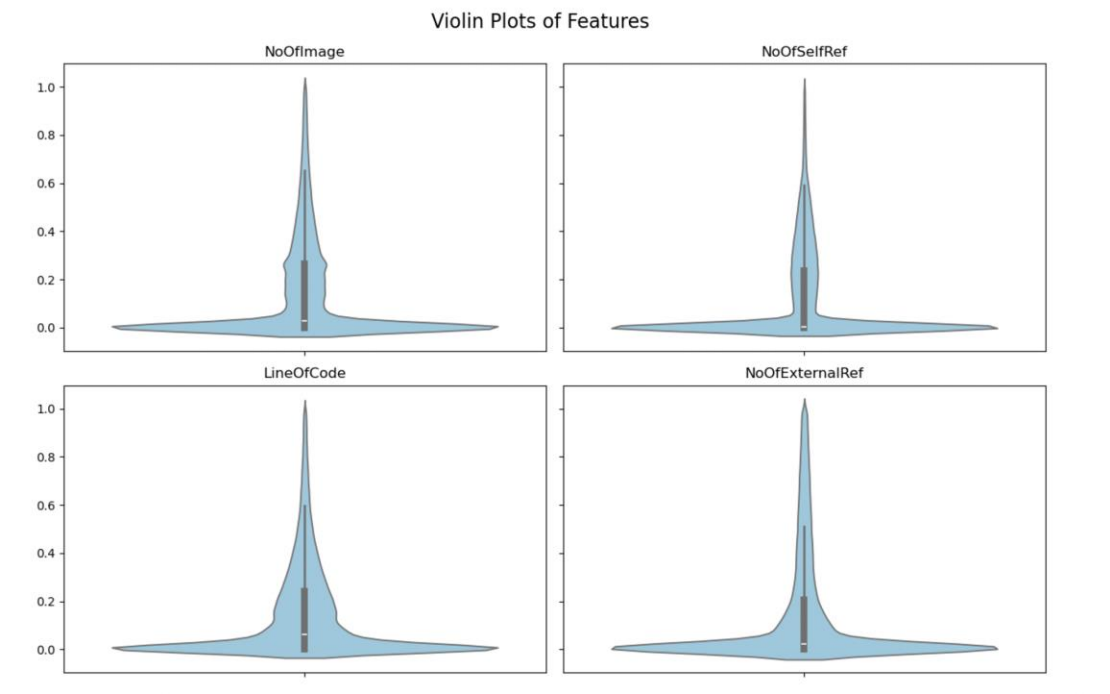
Class Imbalance



Correlation Matrix



Violin Plot



This figure has outliers removed for visibility

Data Preparation Process Summary

1. Dataset Exploration
    - Reviewed data to understand features, distributions and patterns
    - No missing values were found in the dataset
2. Outlier Detection
    - Calculated the IQR to determine the outliers that are in each feature.
        - Values greater than Q3 + (1.5*IQR) and values less than Q1 - (1.5*IQR) were considered outliers
        - Decided to keep these outliers since they were feasible.
3. Class Imbalance
    - Divided testing and training data using the Stratified K-Fold method.
4. Non-Normal Data
    - Histograms revealed heavily skewed distributions.
    - Normalized data using Min-Max Scaling and saved to a new data frame.
5. Categorical Data
    - Dropped unnecessary columns.

04 Modeling

Pros and Cons Modeling Algorithms

We chose a classification algorithm and considered four model types. Here are some pros and cons listed based on the needs of our data.

- Random Forest/Decision Tree
  - Can handle outliers, however overfits easily.
- Neural Network
  - Good for non-linear data and non-normal distributions, however, is moderately sensitive to outliers.
- Naive Bayes Classifier
  - Robust to outliers and has high scalability, however, does not handle non-normal distributions well.
- Support Vector Machines
  - Can handle non-linear data, however, the performance depends on hyperparameters and is sensitive to outliers.

We decided to test all of these algorithms to see which would perform best with our data.

Model Approaches We Tried (Testing)

Algorithm	Best Parameters
Support Vector Machine	C [1.0], kernel [rbf], gamma [scale], probability [True]
Neural Network	activation: ['relu'], alpha: [0.0001], hidden_layer_sizes: [(50, 30)]
Naive Bayes	var_smoothing [1e-08]

Testing Metrics for Phishing Class:

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	96.25%	94.50%	96.89%	95.68%
Neural Network	98.65%	98.68%	98.16%	98.42%
SVM	98.02%	97%	98%	98%

The results of the Random Forest model was not included since we had issues with overfitting

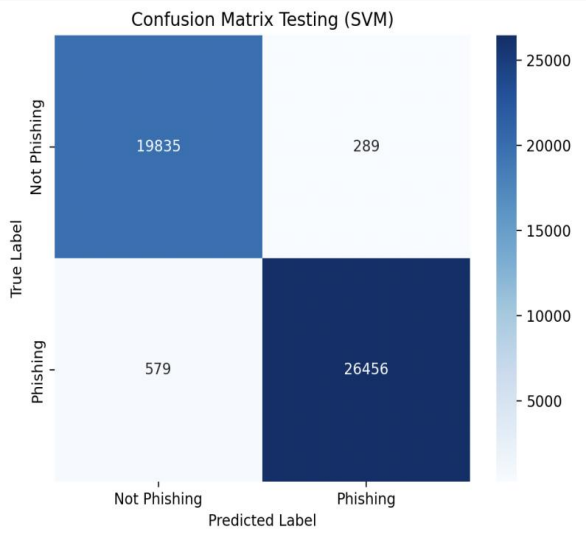
Metric Definitions

True Positive	False Positive
• A link is phishing and is identified as phishing • Good to have high values on!	• A link is phishing and is identified as legitimate • Dangerous to have high values on!
True Negative	**False Negative**
• A link is legitimate but is identified as phishing	• A link is legitimate and is identified as legitimate

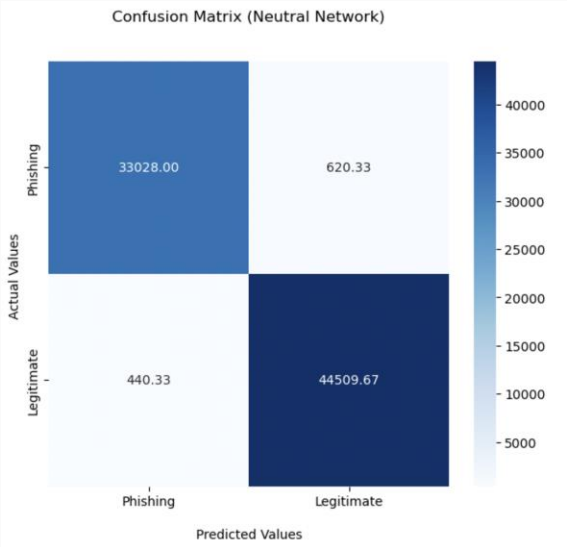
The values of the True Positive and False Positives are what will help make out model accurate to our business problem!

Correlation Matrix of All Three Models

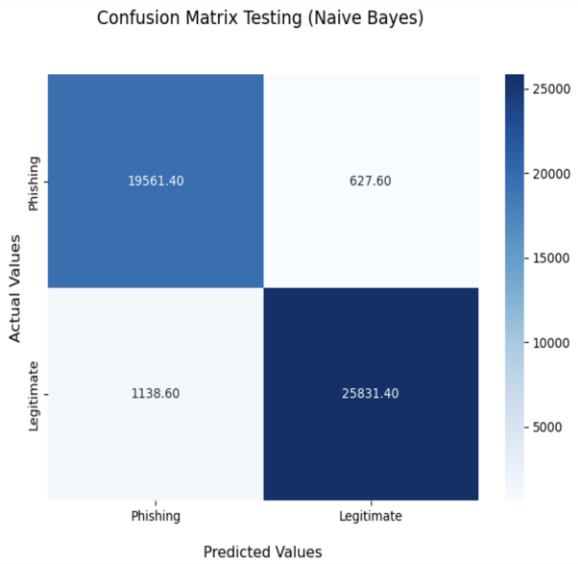
SVM



Neural Network



Naïve Bayes



Main Model: Support Vector Machine

Hyper-Parameters Definitions and Settings (SVM)

C: [1.0]

- A regularization parameter C allows SVM to handle overfitting better in high-dimensional spaces

Kernel: ['rbf']

- Specifies the kernel type for mapping data into higher dimensions.

Gamma [scale]

- Determines the influence of training examples. Avoids overfitting by having too high of an influence.

Probability: [True]

- Enables predictive probabilities in the model.

SVM Model Results (Testing)

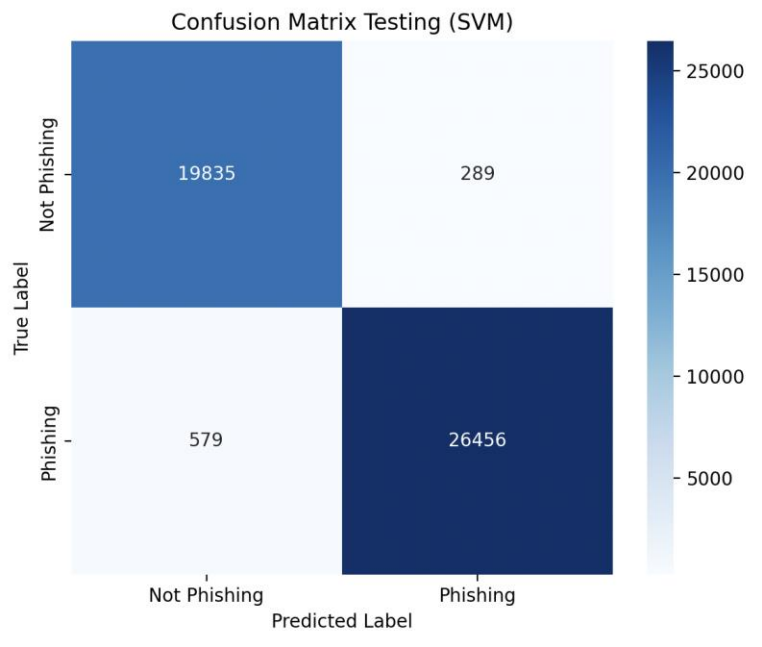
Metric	Precision	Recall	F1-score	Support
0 (Phishing)	0.97	0.99	0.98	0.98
1 (Legitimate)	0.99	0.98	0.98	0.98
Accuracy			0.98	0.98
Macro Average	0.98	0.98	0.98	0.98
Weighted Average	0.98	0.98	0.98	0.98

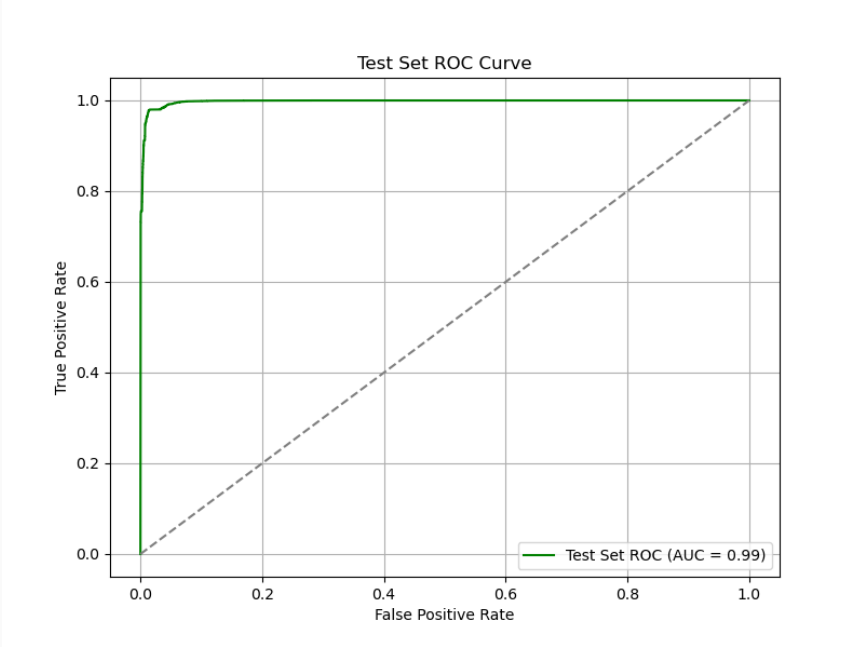
SVM Model Results (Training)

The performance of the model was evaluated using 5-fold stratified cross-validation. The average metrics across the folds are as follows:

Metric	Value
Precision	0.979490
Recall	0.980651
F1-score	0.980024
Support	22,636.5

Correlation Matrix and ROC Curve for SVM





05 Evaluation

Interpretation

Business Question: "How can businesses and organizations identify if a URL they have clicked on is part of a phishing attack?"

Answer:
Businesses can teach employees to identify phishing URLs by looking for certain characteristics:

- Number of Self-Redirects: Phishing sites often redirect back to themselves multiple times, hiding their true intentions..
- Largest Line of Code: Long and inefficient code is a common sign of phishing. This can be through "inspecting" a website's code, used to determine if preventive measures are needed.
- Unusual Use of Images: Too many or too few images suggest an unprofessional, suspicious website
- Number of External References: Links to unrelated external pages are often used in phishing, and can link to more dangerous websites.

How the Model Helps:
Since the SVM achieved high accuracy using these features, it confirms that these traits are reliable indicators for identifying phishing URLs. This reinforces that focusing on these features can help businesses better understand and detect phishing attacks.

06 Deployment

Deployment

- A browser extension can be used on company technology, alerting and blocking suspicious websites from access.
  - While training can be done to warn employees of suspicious websites and proper steps to take after clicking one, preventitive measures offer an extra layer of protection.
- As new phishing techniques are invented, new detection methods must be adopted as well.
  - This can be done by assessing the model's ability to identify a phishing URL. If it starts underperforming, more testing and reassessing methods can keep it up to date.
  - The only way to match the unpredictable progress of technology is regularly gathering more information on phishing websites, improving the model.

Conclusion

- What We Did
  - Analyzed URLs to classify them as phishing or legitimate, addressing the threat of phishing attacks.
  - Tested three models: Naive Bayes, Neural Network, and SVM.
  - Focused on features like number of images, self-references, largest line of code, and external references.
- What Worked
  - SVM achieved over 98% accuracy, with high precision, recall, and F1 scores.
  - Steps like handling outliers, scaling features, and addressing class imbalance improved results.
- What Didn't Work
  - Some features, like URL similarity, caused overfitting and were removed.
  - The study only analyzed URL-level features, missing website content or user behavior.
  - Our old business question was changed since we realized that it no longer aligned with our features. We reassessed and revisited our business understanding section to fix it.
- Next Steps
  - Expand the study to include website content and user interaction data.

Bibliographical references

- George Mason University. (2023, January 26). Phishing - Information Technology services. Information Technology Services. https://its.gmu.edu/working-with-its/it-security-office/phishing/
- Phishing. (2022, September 13). Federal Trade Commission. https://www.ftc.gov/business-guidance/small-businesses/cybersecurity/phishing
- Prasad, A. & Chandra, S. (2024). PhiUSIIL Phishing URL (Website) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.1016/j.cose.2023.103545.
- Protect yourself from tech support scams. (n.d.). Microsoft Support. https://support.microsoft.com/en-us/windows/protect-yourself-from-tech-support-scams-2ebf91bd-f94c-2a8a-e541-f5c800d18435
- 1. Supervised learning — scikit-learn 0.22 documentation. (2019). Scikit-Learn.org. https://scikit-learn.org/stable/supervised_learning.html
- What is phishing? (n.d.). https://support.microsoft.com/en-us/windows/protect-yourself-from-phishing-0c7ea947-ba98-3bd9-7184-430e1f860a44#:~:text=Phishing%20(pronounced%3A%20fishing)%20is,that%20pretend%20to%20be%20legitimate
- Wright, Ryan T., et al. “Research Note: Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance.” Information Systems Research, vol. 25, no. 2, 2014, pp. 385–400. JSTOR, http://www.jstor.org/stable/24700179. Accessed 14 Nov. 2024.

We Hope You Enjoyed!

This course was very beneficial for us! We gained a deep understanding of the steps involved in Data Mining and feel much more confident in our knowledge and skills. I hope that the work we put in reflects what we learned.