Gotcha!: A Novel Approach to Phishing Detection Using Machine Learning

Dinorah García-Vásquez, Mustafa Eren, Hazel Ortega, and Elizabeth Cabrera

Prof. Jennifer Holst, CSCI 401, John Jay College

Introduction

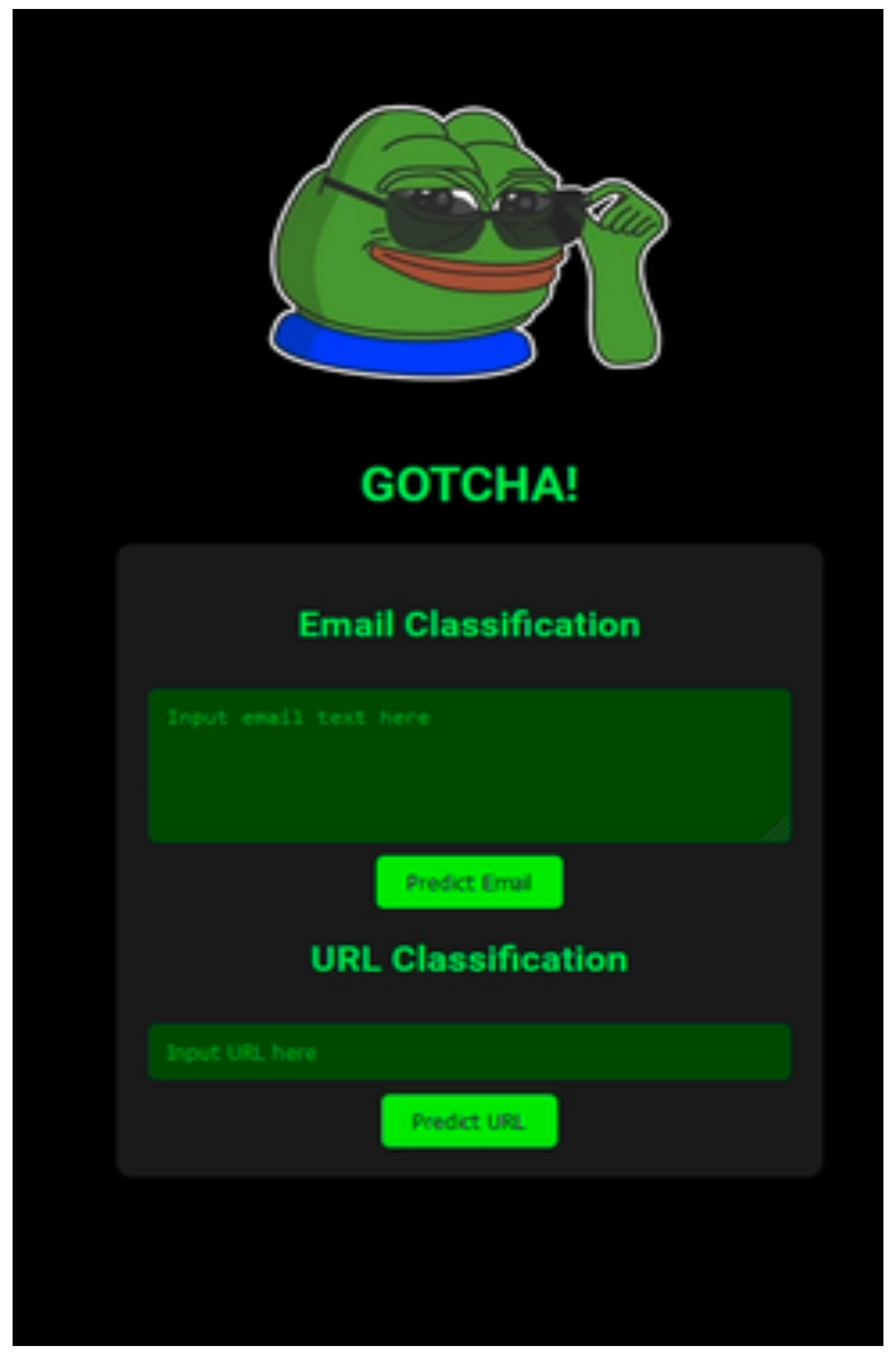
According to researchers Kapan and Gunal, for detecting phishing attacks, machine learning methods prove superior to basic blacklisting tactics. They excel in their ability to adjust to emerging attack patterns without the need for manual updates (Kapan & Sora Gunal, 2023).

ML approaches phishing detection dynamically; rather than avoiding known senders, ML can analyze the content of emails and classify them as phishing attempts, even if the sender is not classified as malicious.

Our research focuses on training a machine learning algorithm with phishing datasets. According to the Federal Trade Commission, phishing is a fraudulent scheme online, where scammers send emails that mimic reputable sources like internet service providers, banks, or mortgage companies. These emails request recipients to provide personal identifying information. (Federal Trade Commission, 2018).

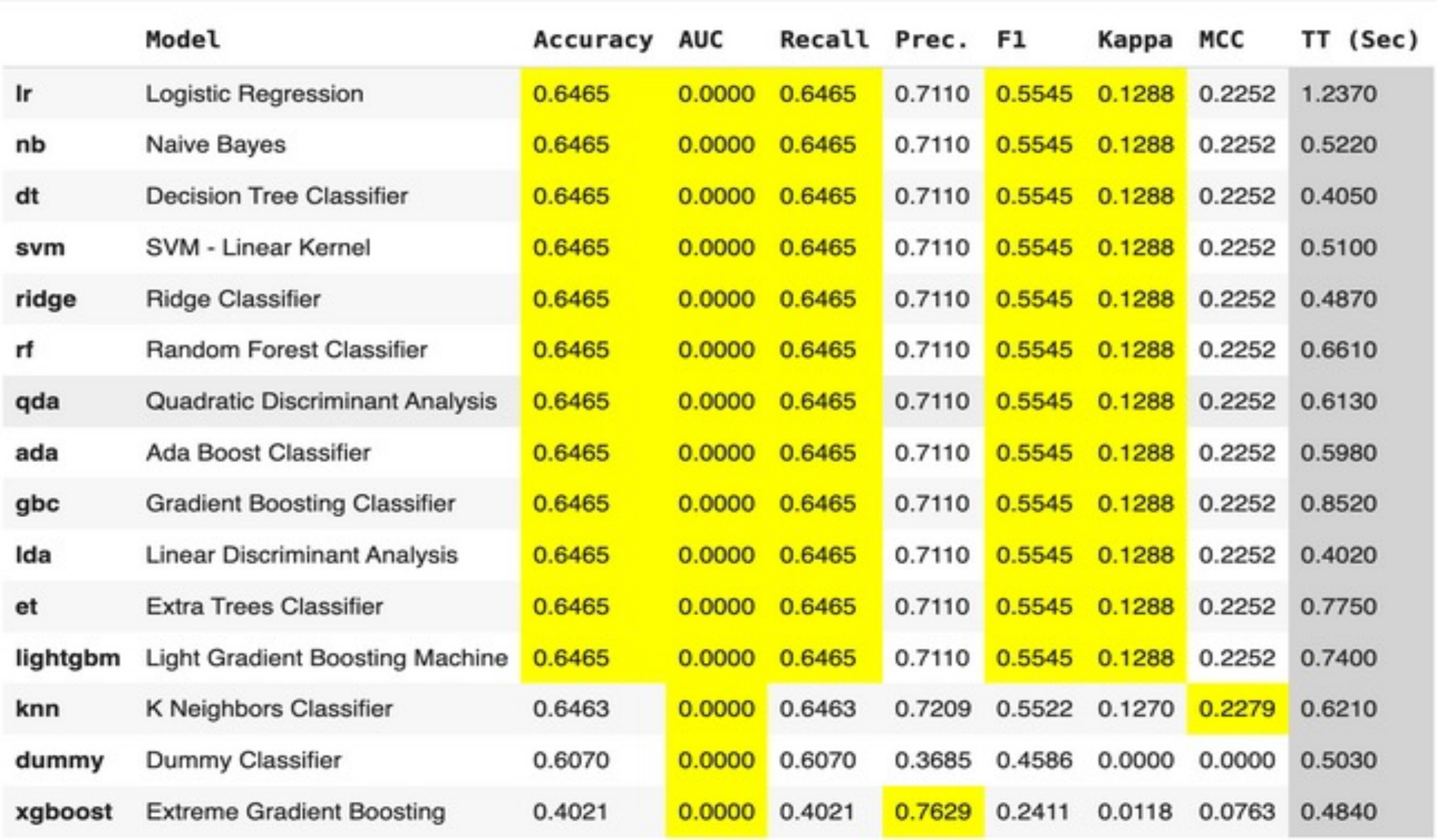
Objectives

Our goal is to enhance cybersecurity awareness among students by providing dynamically tailored solutions. Our phishing detection program will block attempts based on the content (text), regardless of the sender (who typically attempts phishing attacks from multiple different accounts).



Methodology

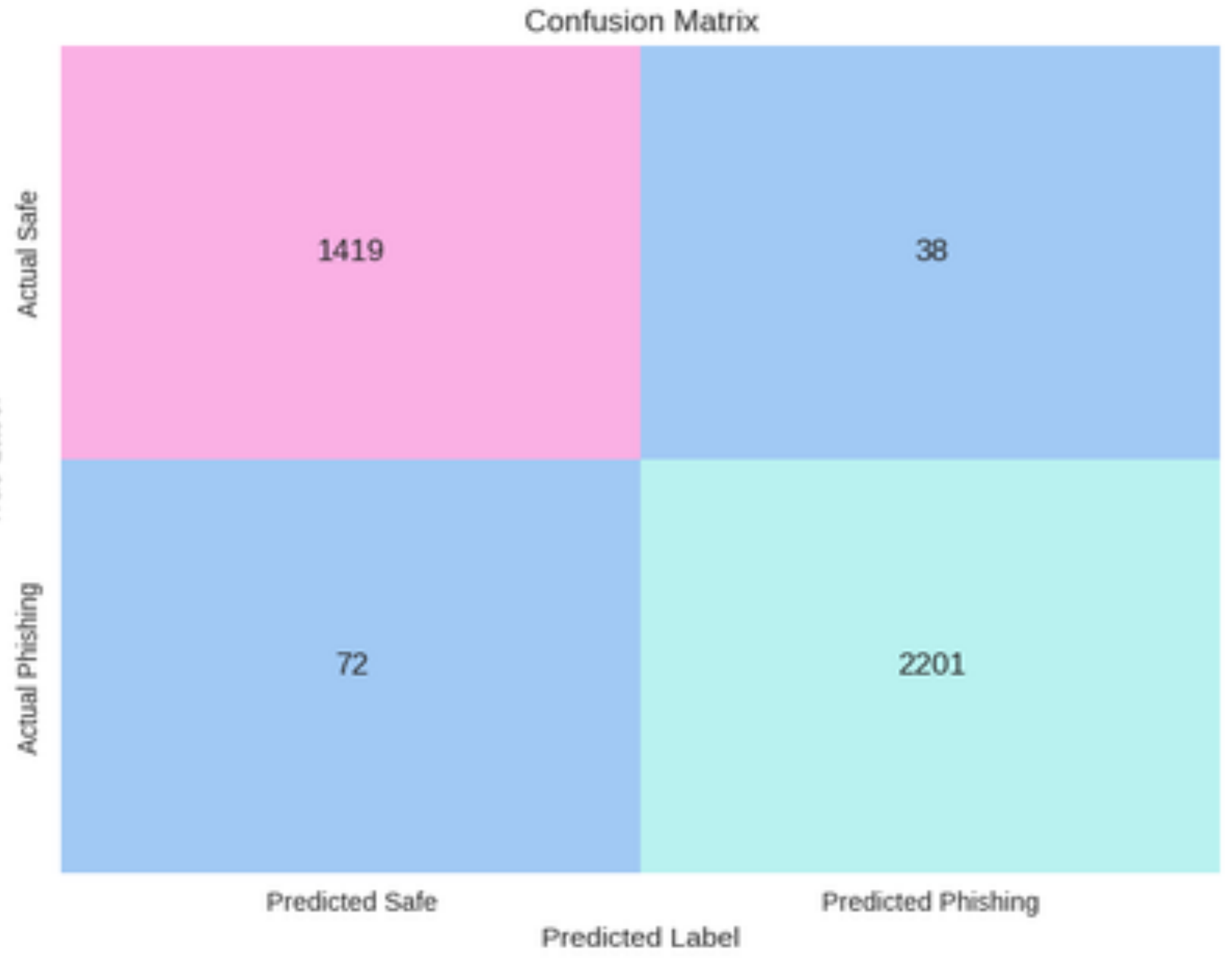
We began by selecting phishing datasets from Kaggle known for their high usability, encompassing malicious URLs and phishing text alongside legitimate emails. Utilizing these datasets, we trained a random forest classifier to discern between legitimate and malicious content. However, to ensure robustness, we also employed PyCaret to explore alternative classification algorithms.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	1.2370
nb	Naive Bayes	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.5220
dt	Decision Tree Classifier	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.4050
svm	SVM - Linear Kernel	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.5100
ridge	Ridge Classifier	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.4870
rf	Random Forest Classifier	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.6610
qda	Quadratic Discriminant Analysis	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.6130
ada	Ada Boost Classifier	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.5980
gbc	Gradient Boosting Classifier	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.8520
lda	Linear Discriminant Analysis	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.4020
et	Extra Trees Classifier	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.7750
lightgbm	Light Gradient Boosting Machine	0.6465	0.0000	0.6465	0.7110	0.5545	0.1288	0.2252	0.7400
knn	K Neighbors Classifier	0.6463	0.0000	0.6463	0.7209	0.5522	0.1270	0.2279	0.6210
dummy	Dummy Classifier	0.6070	0.0000	0.6070	0.3685	0.4586	0.0000	0.0000	0.5030
xgboost	Extreme Gradient Boosting	0.4021	0.0000	0.4021	0.7629	0.2411	0.0118	0.0763	0.4840

Results

Surprisingly, our findings revealed that logistic regression consistently outperformed the random forest algorithm in terms of accuracy. Logistic regression tends to be better at figuring out if something is a phishing attempt because it looks at certain features in a straightforward way. It's like when you're trying to decide if a message is real or fake based on specific clues you know about. This method works well even if there's some confusing information in the messages. On the other hand, random forest, while powerful, can sometimes get confused by too much information and make mistakes.



Conclusions

We found that preserving special characters during data cleaning is crucial for detecting phishing emails and malicious URLs effectively. Additionally, we uncovered the concerning frequency of phishing attempts targeting John Jay College students, occurring nearly weekly. To enhance detection accuracy further, we will be exploring Natural Language Processing (NLP) due to its superior ability to understand human language nuances compared to traditional machine learning.

References

Datasets

Subhadeep Chakraborty. (2023). Phishing Email Detection [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/6090437

Sid321axn. https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset

References

Federal Trade Commission. (2018, October 31). Phishing Scams. Federal Trade Commission. https://www.ftc.gov/news-events/topics/identity-theft/phishing-scams

Kapan, S., & Sora Gunal, E. (2023). Improved Phishing Attack Detection with Machine Learning: A Comprehensive Evaluation of Classifiers and Features. Applied Sciences, 13(24), 13269. https://doi.org/10.3390/app132413269

Acknowledgements

“Gotcha!: A Novel Approach to Phishing Detection Using Machine Learning” research is part of our capstone project for CSCI 401 with Professor Jennifer Holst.