Hemal Mamtora
UIN: 730007890

CSCE 685 626: **Directed Studies** under Prof. Duncan Walker
Project Proposal: **WebSite Phishing Detection**

About the project:

An attacker uses a Website Phishing attack to create a false website that impersonates a legitimate entity. The attacker attempts to obtain credit card details, passwords, emails, and other sensitive information from users. To combat this, a variety of website phishing detection strategies have been investigated in the literature, including URL-based phishing detection, website image-based detection and content-based detection.

For this project, I focus on content-based phishing detection, where I create features based on the content of legitimate and forged websites to detect phishing.

Dataset exploration done: https://data.mendeley.com/datasets/n96ncsr5g4

About the dataset:

The dataset contains html content of legitimate and phishing websites. The data has been indexed using an index.sql file. There are a total of 80,000 samples. Out of which the 50,000 examples are legitimate and 30,000 examples are fake. The dataset is divided into 8 sections.

About the approach:

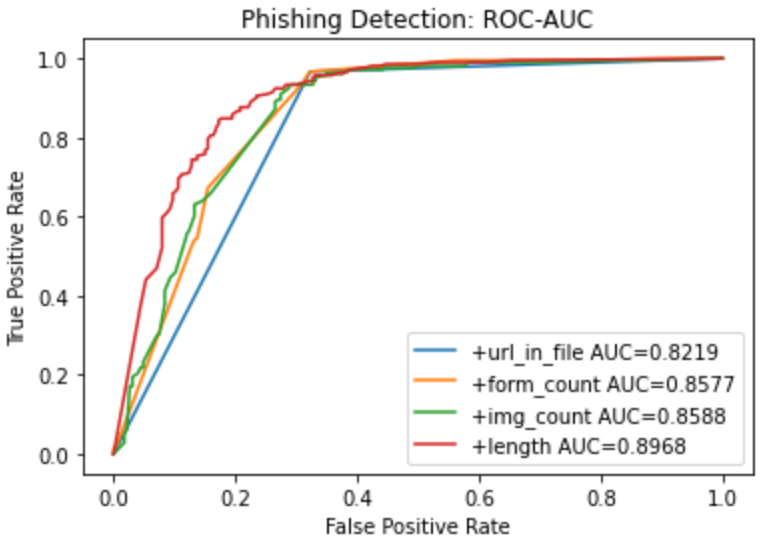
By using a smaller reference dataset which has features already extracted, I extracted similar features from the Phishing detection dataset. Right now I am using 1/8th of the dataset for experimenting my approach.

The features which I tried to extract are:

1. Old Features: (already experimented in literature)
    a. Url_in_file: Does the content contain URL in html file. This implies there is a link to home which shows some legitimacy and the website might not be a phishing site.
2. New Fertures: (new content based features for detection)
    a. Form_count: How many forms are present in the given html web page? If there are more forms, it might be a phishing attempt
    b. Img_count: Similarly, a lot of images might imply phishing attempt
    c. length : length of the file is an indicator of malicious code present on the website. If it is unusually long, it might show a phishing attempt.

Results:

Hemal Mamtora
UIN: 730007890



AUC for usage of 4 features: each time adding additional feature, increases AUC.

Next steps:

I would try out more old features and additional new features. I will check if the accuracy can be improved. I would run the code on the entire dataset (50,000 + samples). I would report the accuracy after introducing more features.

References:

[1] Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence, 104, 104347.
[2] Van Dooremaal, B., Burda, P., Allodi, L., & Zannone, N. (2021, August). Combining text and visual features to improve the identification of cloned webpages for early phishing detection. In The 16th International Conference on Availability, Reliability and Security (pp. 1-10)
[3] Yang, P., Zhao, G., & Zeng, P. (2019). Phishing website detection based on multidimensional features driven by deep learning. IEEE access, 7, 15196-15209.