MINNETONKA RESEARCH

Using Machine Learning Algorithms to detect phishing scams through Domain Name Detection

Isaac Berlin

MINNETONKA PUBLIC SCHOOLS

Introduction

Cybersecurity has become increasingly more important and many aspects of life move to an online platform. One particularly difficult cybersecurity issue is phishing scams. This type of scam is when a scammer impersonates a trusted individual or entity and attempts to obtain sensitive information. More often than not this is done through fake hyperlinks to websites that look similar enough to the real thing to catch people off guard. Phishing is difficult to prevent because it contains both aspects of a typical cyber attack with malicious intent and a socially engineered scam that has happened long before the introduction of the internet. This project has attempted to take the social engineering factor out of phishing.

Methods

The algorithm that was created for this project was built in Python. The following API's were used.

- NumPy for basic calculations
- Pandas for database processing
- PyTorch for machine learning
- MatPlotLib for data visualization

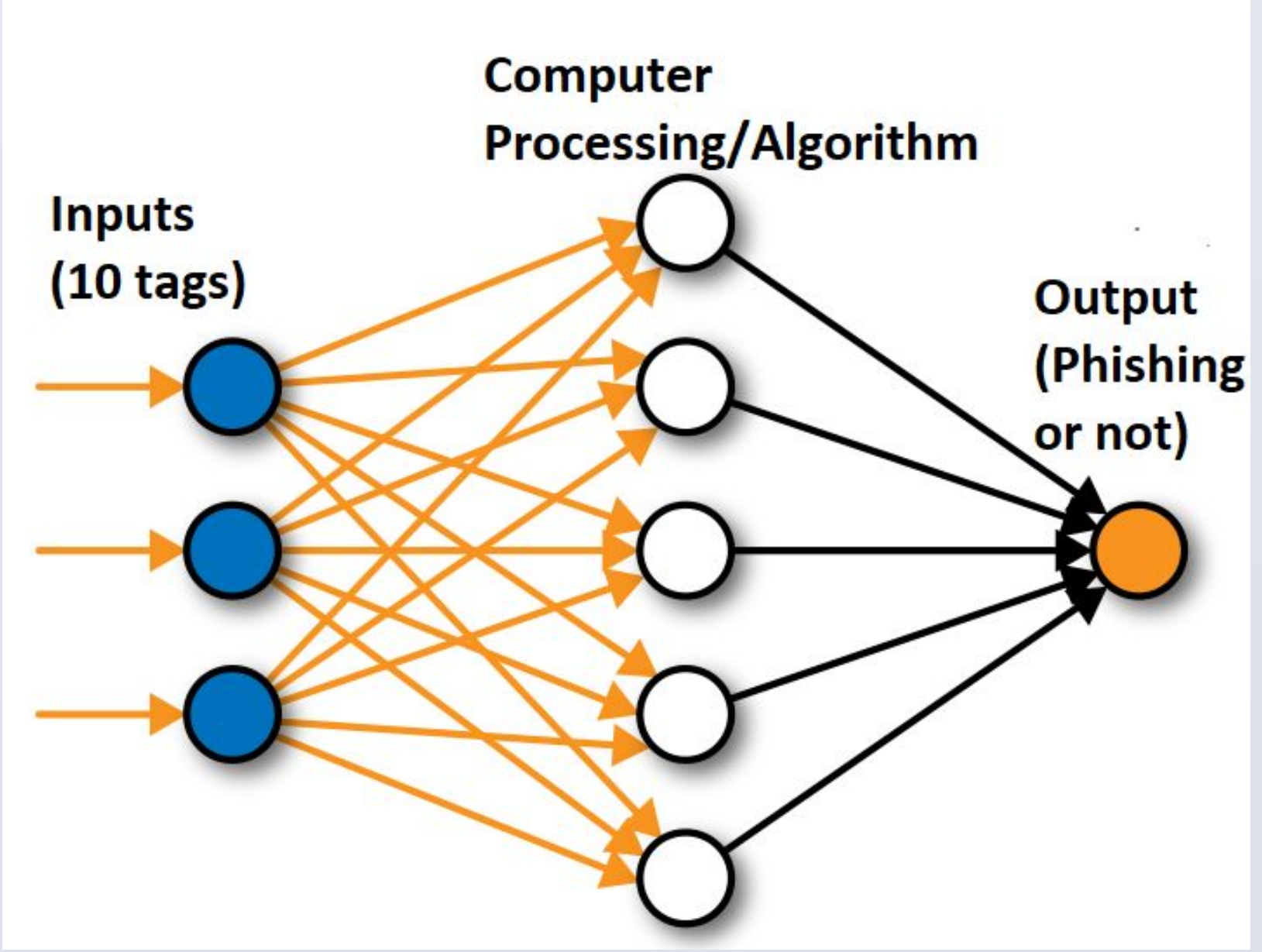
After all the imports are done the next step was to process the data. This was done by removing some unnecessary information that is unable to be processed by the machine learning algorithm and splicing the data into labels and tags.

The Database contained 12 different columns for each row. The columns are

- domain of the website (removed)
- page ranking
- if there is an IP address in the link
- if the website is valid according to the url registration
- how long the site as been active
- length of the url
- if the link has an @ symbol
- if the url has a dash
- if the url has double dashes possibly indicating a redirect
- length of the domain name
- number of subdomains
- whether the domain is a phish or not (label)

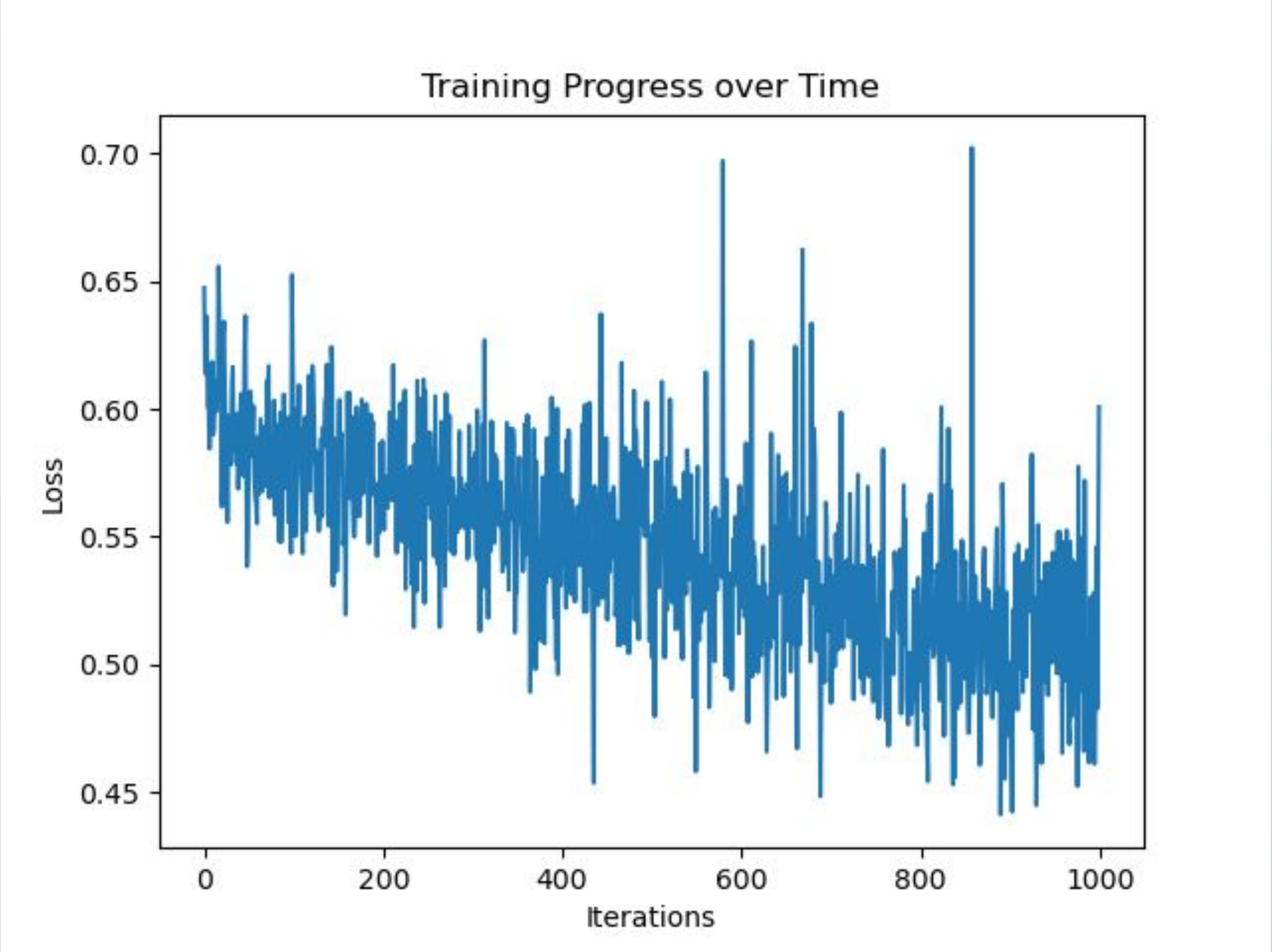
The machine learning algorithm calculates accuracy using a loss function. To put it simply, the loss function is a mathematical measurement of how far from correct the algorithm's prediction is. Learning is signaled when the loss function decreases.

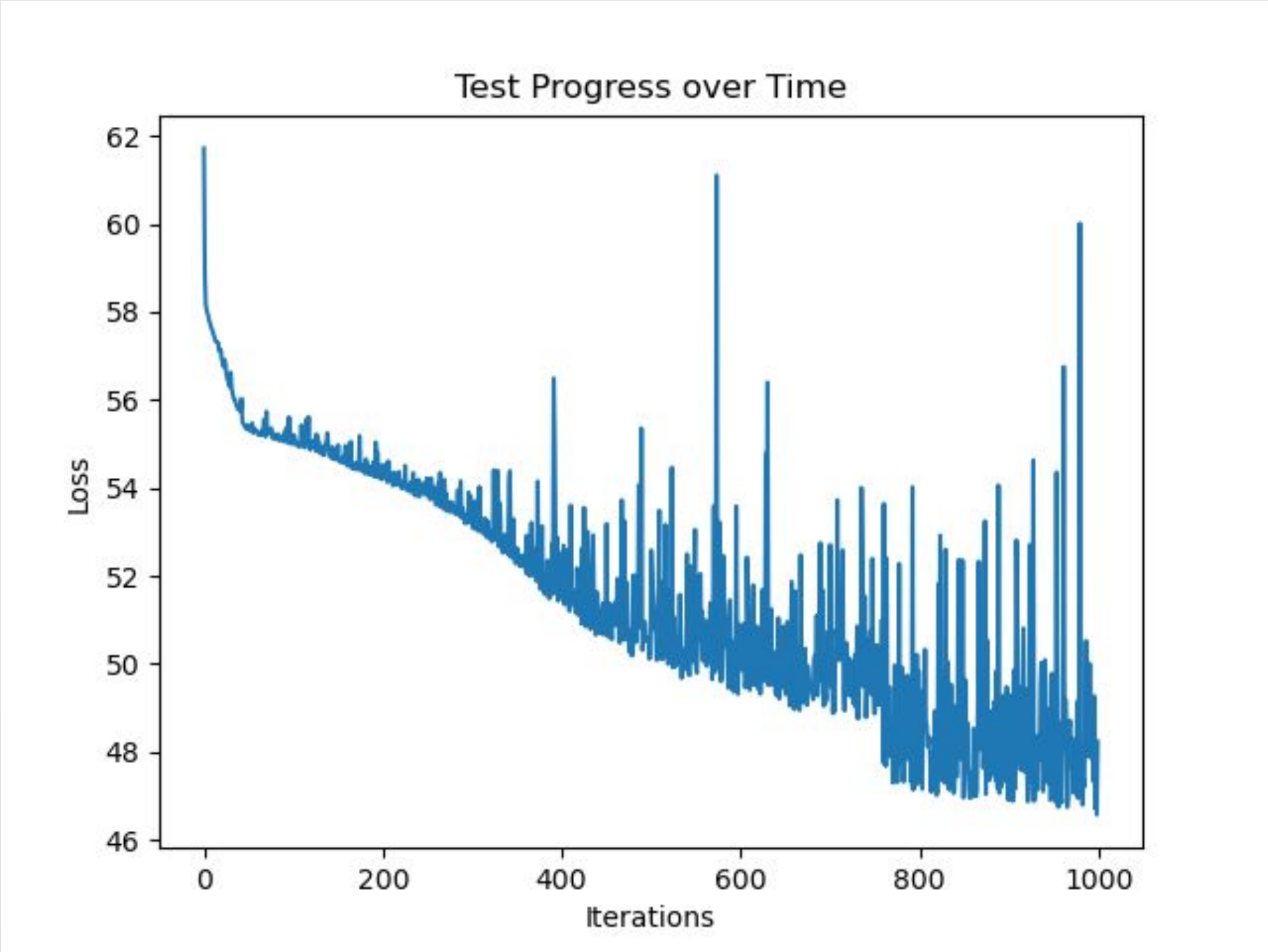
The image below is meant to represent how the algorithm processes the data through both its training and testing stage. The names may be self explanatory but the only real difference is that the algorithm does not learn from the results of the testing stage.



Data

The following two graphs are the output from running the algorithm at 1000 iterations. They show a downward curve in the loss.





Results

As shown in the data, the results for this study are promising. It is clearly shown that there is a downward trend in loss for both of the graphs and in the graph of the test loss there is a spike in learning at around 100 iterations and at around 750 iterations,

Summary

This study is an overall success. It shows that the machine learning algorithm can detect phishing based on facts of information about the domain hyperlink. While this is a success there are steps that can be taken to improve this study. While the dataset worked marvously there is always the want to have a larger dataset with better tags or to include more information about the content of the email and not just the hyperlink. If given the chance it is always better to attempt to create a new dataset rather than use an already created one (despite the creation of a new dataset taking years at best). The tags of information hopefully added would be as follows.

- country of origin of the email
- date sent
- number of spelling errors
- If the email had a similar template to any big businesses (attempting to impersonate)

Refrences

Chang, Jeong Ho; Bergholz, André . 2008. Improved Phishing Detection using Model-Based Features. The Fifth Conference on Email and Anti-Spam

Nagariya, Aman. 2020. Phishing websites Data Classifying Phishing websites from Legitimate ones. Kaggle

Sahingoz, Ozgur Koray; Buber, Ebubekir; Demir, Onder; Banu. 2019. Diri Machine learning based phishing detection from URLs. Expert Systems with Applications Volume 117, Pages 345-357