

CatchPhish

An ML Approach to URL Phishing Detection

Omar Kreidie

February 25th, 2025



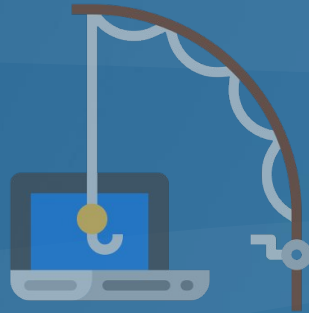
The Problem



- Cybercrime is on the rise.
- Small to Medium Businesses (SMB's) account for 43% of all cyber attacks.
- 95% of all cyber breaches are attributed to human error.



What is Phishing?

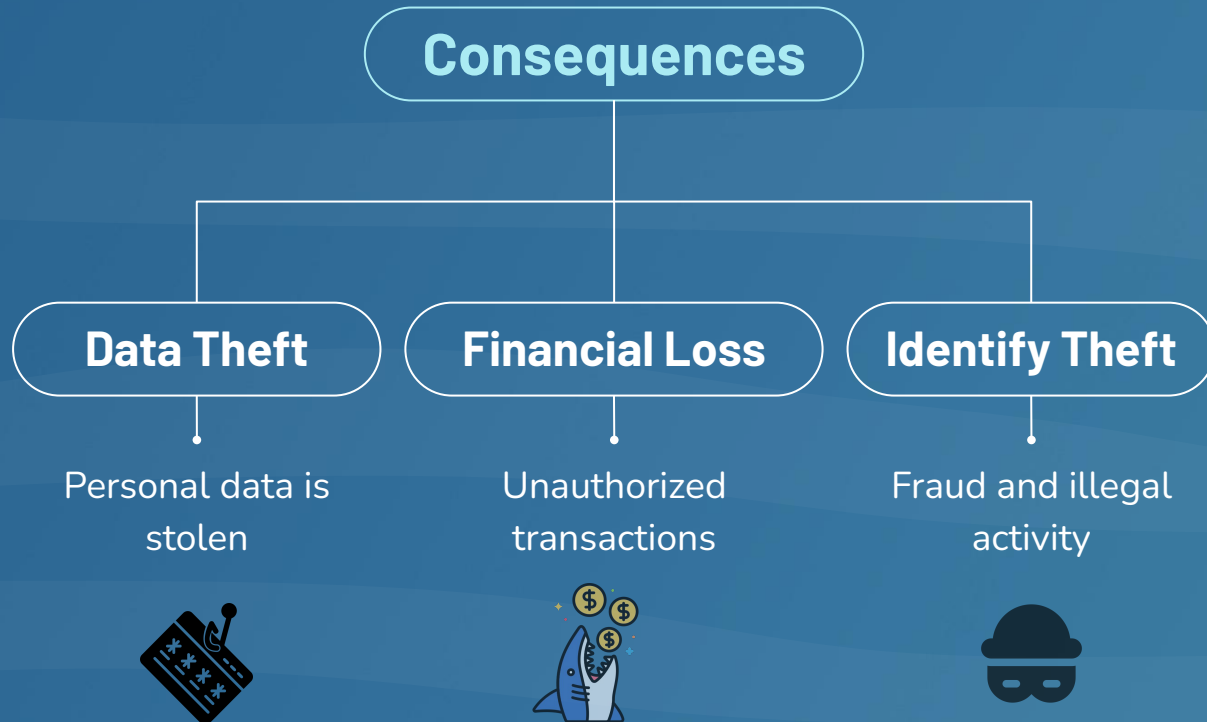


Don't We Already Have a Solution For This?

- More layers more protection
- SMB's struggle to adopt effective security
 - Lack of education
 - Cost
 - Optimism Bias
- SMB's are the best phish!



Impact of a Phishing Attack



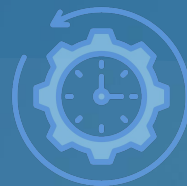
The Impact of The Solution

- Financial Protection.
- Preservation of Reputation and Customer Trust.
- Reduce Downtime and Improve Productivity.



The Data Science Solution

- Relying on humans is outdated & ineffective.
- Building a model for the URL.
- The 3 Part Process
 - Data Wrangling + Preliminary EDA
 - Data Pre-processing + EDA
 - Predictive Modelling



The Dataset

- Found on UCI Machine Learning Repository.
- The dataset contains 235,795 rows and 56 features.
- Data Types
 - Object
 - Int64
 - Float64



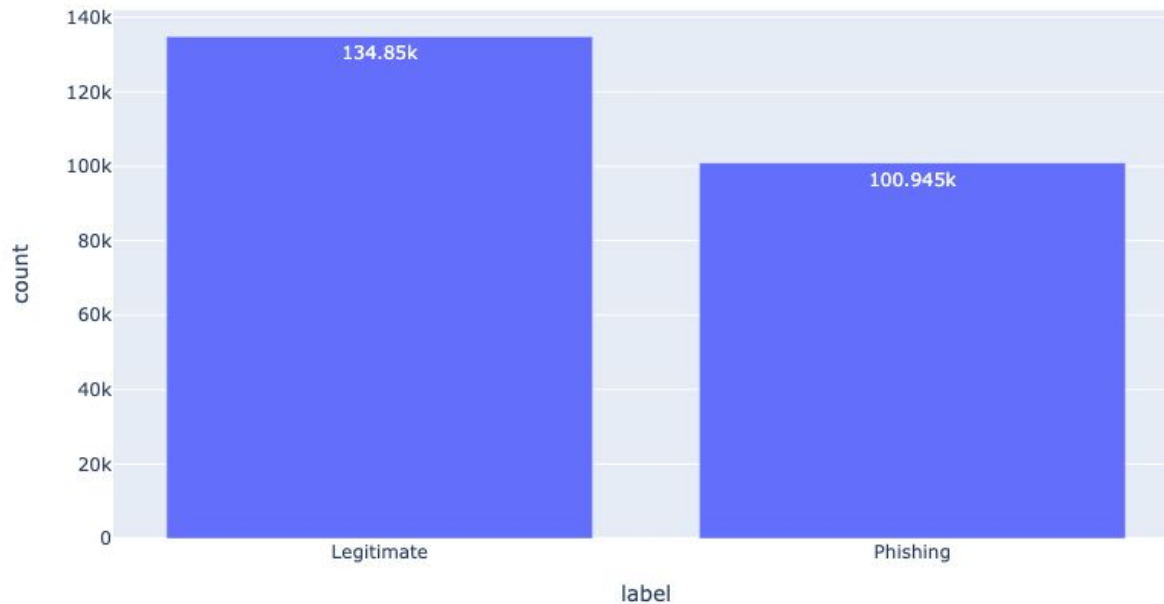
Data Cleaning & Feature Engineering

- 56 Features, 21 features where collected, 35 engineered.
- The Dataset has 0 null values & 0 duplicate values.
- Need to understand the 56 features before the full EDA.



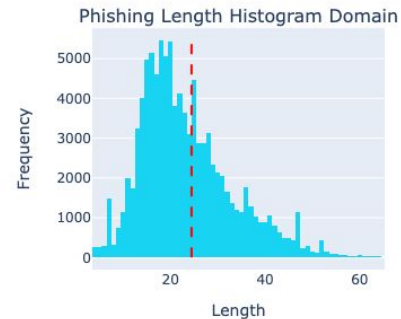
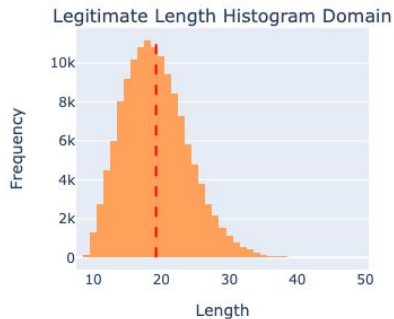
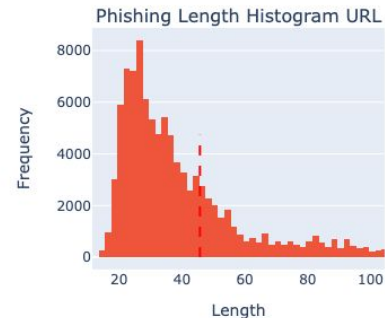
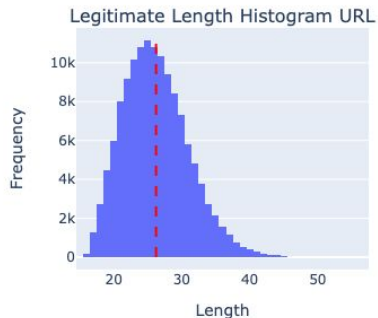
Understanding The Data

Number of Phishing vs Real URL



Findings from Preliminary EDA

Histograms of URL and Domain Lengths for Legitimate and Phishing Classes



Findings from Preliminary EDA

- Building an efficient model will be crucial.
- There will be a significant amount of collinearity
- Need to understand the 56 features before the full EDA.

What's Next?

1. Data Collection

Finding and committing to the dataset

Handling null + duplicate values.
Analyzing Relationships.

2. Data Wrangling + Prelim EDA

3. EDA + Baseline Modeling

Finding the features with the highest predicting power. Building a Prelim Logistic Model

Try different classification models. Use Random Forest for feature importance.

4. Advanced Modeling

References

- [1]<https://smallbiztrends.com/small-business-cybersecurity/>
- [2]<https://www.ibc.ca/news-insights/news/small-businesses-are-underestimating-their-cyber-risk-despite-increased-threats>
- [3]<https://www.forbes.com/sites/edwardsegal/2022/03/30/cyber-criminals/>

Thank You

