CatchPhish

An ML Approach to URL Phishing Detection

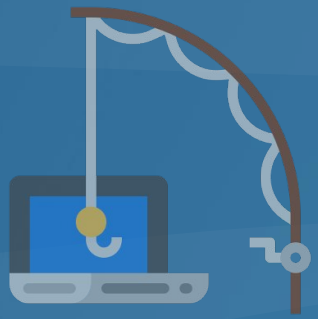
Omar Kreidie

April 26th, 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

Consequences

Data Theft

Personal data is stolen

Financial Loss

Unauthorized transactions

Identify Theft

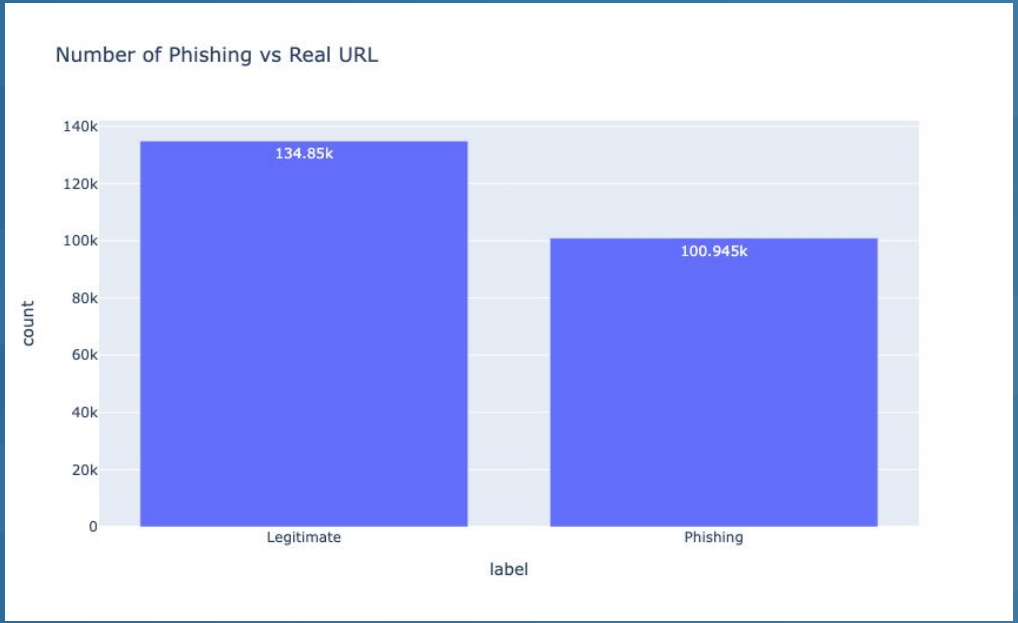
Fraud and illegal activity

The Impact of The Solution

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

Understanding The Data

- The dataset contains 235,795 rows and 56 features.
  - 50 Feature Engineered by Data Engineer
    - Clean Dataset (no nulls or duplicates)



Preliminary Model

Logistic Regression

```
Accuracy: 0.9980491528658368
Precision: 0.9982205086379476
Recall: 0.9983685576566556
F1 Score: 0.9982945276583123

Confusion Matrix:
[[20141    48]
 [   44 26926]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     20189
           1       1.00      1.00      1.00     26970

    accuracy                           1.00     47159
   macro avg       1.00      1.00      1.00     47159
weighted avg       1.00      1.00      1.00     47159
```

Preprocessing

Variance Thresholding

Correlation Analysis

Intercorrelation Analysis

Feature Importance

Findings From EDA

Findings From EDA

- A lot of inter collinearity
  - Total features reduced 10
- A lot of data leakage (target leakage)

What's Next?

1. Data Collection

Finding and committing to the dataset

2. Data Wrangling + Prelim EDA

Handling null + duplicate values. Analyzing Relationships.

3. EDA + Baseline Modeling

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

4. Advanced Modeling

Find the best model without target data leakage. Use NLP techniques build models using tokenized URL's

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