CMSI 533 End-to-End Data Science Project

Due April 9th 2024

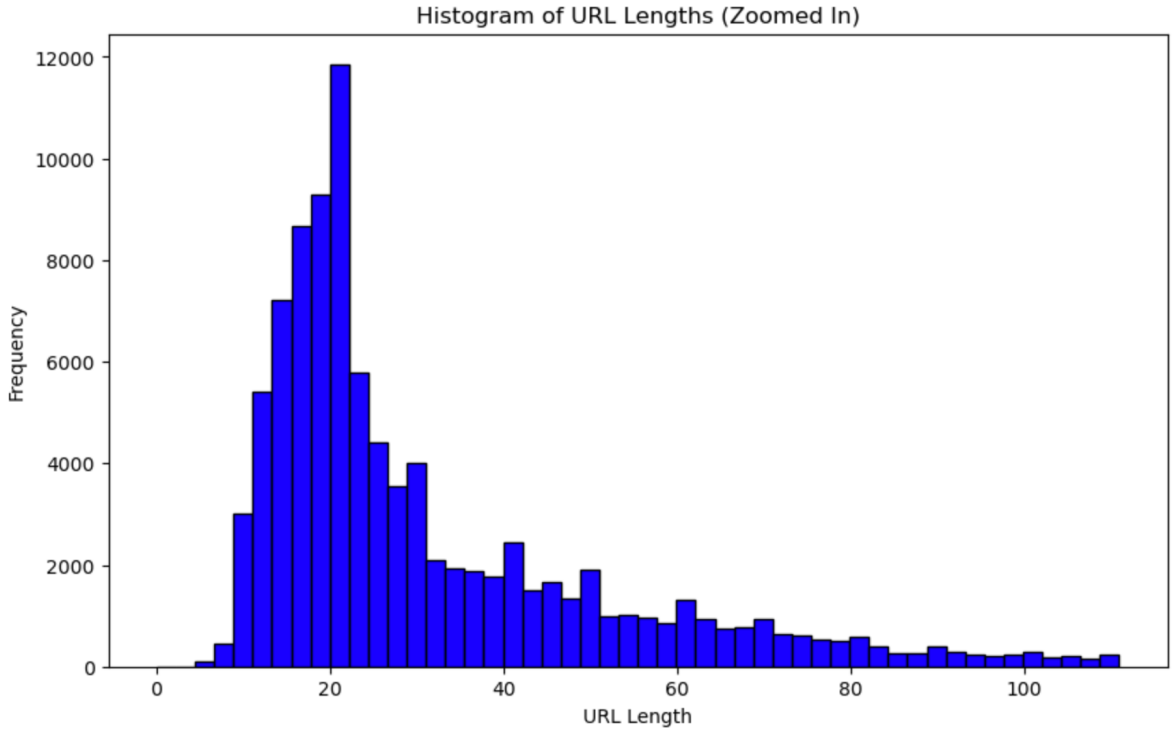
Data Description

- Web Page Phishing Dataset
- 100077 rows
- 20 columns
- Predicting whether a given website is phishing based on various features derived from its URL.

Descriptive Statistics

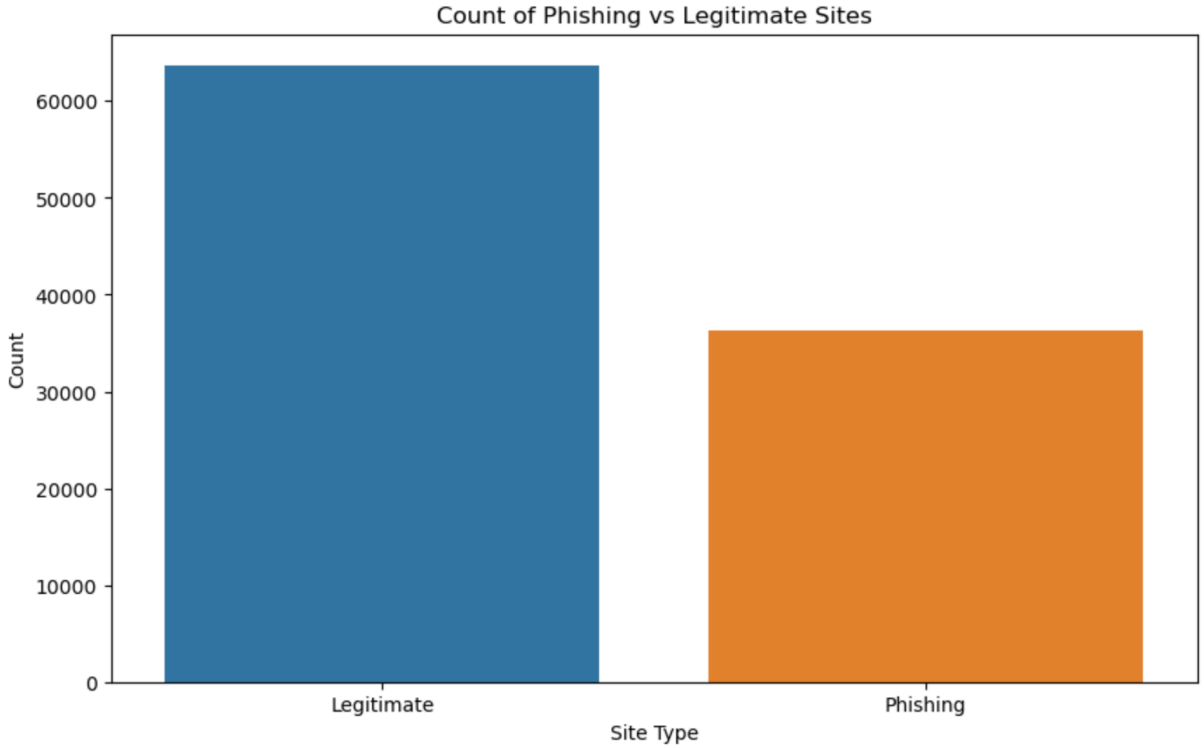
- Mean -> 39.18
  - This is the
- Median -> 24
  - This is the
- Mode -> 18
  - This is the
- Standard Deviation -> 47.97
  - This is the

Histogram of Numeric Column

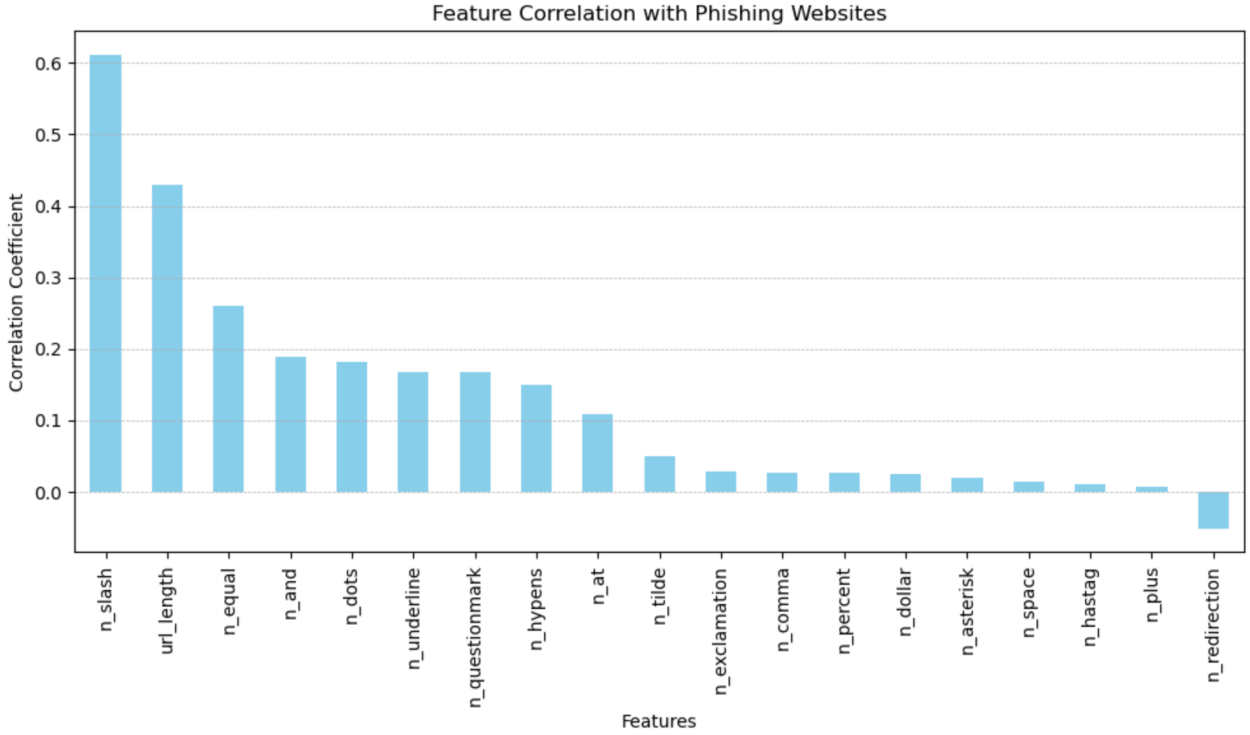


Data Visualization

Count of Phishing vs Legitimate Sites



Feature Correlation with Phishing Websites



Write a paragraph about what your 2 visualizations mean in context

The count plot illustrates an almost balanced dataset. This is beneficial for training machine learning models as it suggests that the dataset is not heavily skewed toward one class, which could lead to a biased predictive model. The balance implies that accuracy can be a reliable metric for evaluating model performance, although precision, recall, and the F1-score will provide a more abstract view, especially in a security context where false negatives (phishing sites misclassified as legitimate) can have serious implications.

The bar plot of feature correlation with the phishing label, indicates which characteristics are most associated with phishing websites. The top-ranked features, such as the number of slashes and the URL length, have the highest correlation coefficients and are therefore strong candidates for predictive features in a model. Features with lower correlation may not contribute as much predictive power and could potentially be excluded to streamline the model.

Machine Learning

- Predictor Value: Phishing
- Features:
  - url_length: The length of the website's URL.
  - n_dots: The count of . characters in the URL.
  - n_hypens: The count of - characters in the URL.
  - n_slash: The count of / characters in the URL.
  - n_questionmark: The count of ? characters in the URL.
  - n_equal: The count of = characters in the URL.
  - n_at: The count of @ characters in the URL.
  - n_exclamation
  - N_space
  - N_tilde
  - N_comma
  - N_plus
  - N_asterik
  - N_)hashtag
  - N_dollar
  - N_percent
  - N_redirection

Machine Learning

Accuracy: 0.86 – Great

```
# Create and train the logistic regression model
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

# Predictions and evaluation
predictions = model.predict(X_test)
print(confusion_matrix(y_test, predictions))
print(classification_report(y_test, predictions))
```

Model Evaluation

Create new features from the existing data that could better capture the nuances between phishing and legitimate sites. For example, consider the ratio of special characters to the total URL length or the presence of certain keywords.