CS 459 – Introduction to Machine Learning

ومن يؤت الحكمة
فقد أوتي خيرا كثيرا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

Assignment No. 4

Site Phishing Classification

Submitted To:

Dr. Ayaz Hussain

Submitted By:

Muhammad Nasir Khurshid - 04071913019

1. Introduction

In this project, my goal is to develop a predictive model to determine if a website is a phishing site or not. I analyzed a comprehensive dataset containing information about websites and their classification. The dataset includes various features like website characteristics and domain information. By using these features, we can uncover patterns and insights that will help accurately predict whether a site is phishing or legitimate.

To achieve this, I evaluated the performance of different machine learning models such as Logistic Regression, Neural Networks, Decision Trees, Random Forest, and Naïve Bayes. Each model provides a unique approach to solving the prediction problem, allowing us to understand their strengths and weaknesses in the context of identifying phishing websites.

Furthermore, I examined the effects of implementing Principal Component Analysis (PCA) on the dataset. PCA aims to reduce the complexity of the data by condensing its dimensions, while still capturing the crucial information necessary for accurate predictions. I analyzed the optimal number of components that preserve a substantial amount of the data's variance, allowing us to assess the potential advantages of dimensionality reduction in our customer churn prediction task.

2. Dataset

The dataset consists of several features that provide valuable information for determining whether a website is phishing or legitimate. These features include:

1. "having_IP_Address": Indicates whether the website has an IP address associated with it.
2. "URL_Length": Represents the length of the URL, which can be an indicator of suspicious or lengthy URLs often used in phishing attempts.
3. "Shortening_Service": Identifies whether a URL shortening service has been utilized, as these services are sometimes employed to hide the true destination of a link.
4. "having_At_Symbol": Indicates whether the website's URL contains the "@" symbol, which is uncommon and potentially indicative of phishing.
5. "double_slash_redirecting": Identifies if the website uses a double slash redirection technique, which may be used to obfuscate the actual destination of the URL.
6. "Prefix_Suffix": Represents the presence of prefixes or suffixes in the website's URL, which can be common in phishing URLs.
7. "having_Sub_Domain": Indicates whether the website uses subdomains, which can be an important factor in phishing detection.
8. "SSLfinal_State": Represents the SSL/TLS certificate status of the website, indicating whether it is secure or not.

9. "Domain_registeration_length": Indicates the duration for which the domain has been registered, as shorter registration periods are often associated with suspicious websites.
10. "Favicon": Identifies the presence of a favicon, which is a small icon displayed in the browser's address bar. Its absence can be indicative of phishing attempts.

Other features include "port," "HTTPS_token," "Request_URL," "URL_of_Anchor," "Links_in_tags," "SFH," "Submitting_to_email," "Abnormal_URL," "Redirect," "on_mouseover," "RightClick," "popUpWidnow," "Iframe," "age_of_domain," "DNSRecord," "web_traffic," "Page_Rank," "Google_Index," "Links_pointing_to_page," "Statistical_report," and "Result." These features encompass a wide range of characteristics and statistics that play a role in assessing the legitimacy of a website in the context of phishing prediction.

3. Data Splitting

To train and evaluate the models, the dataset was split into a training set and a testing set using a 70:30 ratio. This ensures that 70% of the data is used for training the models, while 30% is used for evaluating their performance.

4. Model Selection

Following models were chosen to predict whether a website is phishing website or not:

1. Logistic Regression
2. Neural Network (MLPClassifier)
3. DecisionTrees
4. RandomForest
5. Naïve Bayes (GuassianNB)

5. Model Evaluation

Performance metrics, such as precision and recall, were utilized to evaluate the models. These metrics offer valuable insights into the overall performance of the models and their effectiveness in accurately predicting phishing sites.

6. Results & Analysis

After analyzing the obtained outcomes, I have assessed the performance of various classifiers in predicting whether a site is phishing or not. The precision and recall scores have been employed as metrics to gauge the predictive capabilities of the models.

6.1. Without PCA

Classifier	Precision	Recall
LogisticRegression	.919526627218935	.9283154121863799
MLPClassifier	.9466019417475728	.931899641577061
DecisionTreeClassifier	.9290865384615384	.9235364396654719
RandomForestClassifier	.9564164648910412	.9438470728793309
GaussianNB	.9956709956709957	.2747909199522103

6.2. With PCA

Classifier	Precision	Recall
LogisticRegression	.9161747343565525	.927120669056153
MLPClassifier	.9396863691194209	.9307048984468339
DecisionTreeClassifier	.8846625766871166	.8614097968936678
RandomForestClassifier	.9314903846153846	.9259259259259259
GaussianNB	.9038461538461539	.8984468339307049

The results offer insights into each classifier's performance. Precision score indicates the proportion of accurately predicted cases among the predicted instances, while recall score measures the proportion of correctly predicted cases among the actual instances. Evaluating both precision and recall provides a comprehensive assessment of the models, as they assess different aspects of their performance.

7. Conclusion

The results suggest that the performance of the classifiers varies depending on the metric used (precision or recall) and the utilization of PCA. Considering both precision and recall is crucial for a comprehensive evaluation of the models. When examining the precision scores, the MLPClassifier with PCA demonstrated the highest precision (0.9396), indicating a relatively greater proportion of accurately predicted cases among the predicted instances. This implies that the MLPClassifier model with PCA could be the optimal classifier for maximizing precision in phishing site prediction.