A Data Mining Approach to detect Website Phishing

Introduction:

The research study proposes a Data Mining Methodology to detect Website Phishing. Website Phishing is a form of cyber-crime that falls under the category of passive attacks. A web link or URL is said to be phished if it poses or claims to look familiar to some other website that has a wide visitor/customer base for the sole intention of breaching the privacy to personal data which is highly sensitive. For instance, Amazon is one of the leading websites in the E-Commerce domain. So, if someone randomly creates a fake E-Commerce Website Framework exactly identical to Amazon and is successful in rising it up the Google Search Rankings or linking it with related product websites, let say, there is a website for Phone Configuration Details which has a good Search Volume and if the ‘Order Now’ or ‘Buy Now’ button attached to the phone is linked to the website which is faking it to be as Amazon, it will be very hard to distinguish whether that particular redirecting website is Amazon’s sub-domain or not. The scenario can quite well be anticipated. The user who is the victim of Phishing attack continues the phone purchase with the phished website posing Amazon and, in the process, the attackers can get access to the payment credentials of that particular user. A similar phishing attack can happen with Net Banking portals attached in payment gateways of several E-Commerce websites as well but the possibility for the same is very minimal in recent times, as the popular E-Commerce are very cautious regarding payment gateways and links attached to them. But one way or the other, Website Phishing is still on the rise and is a very detrimental issue. In this research study, a Data Mining based methodology is proposed to detect the phished website links based on some of the very significant deciding features of a website such as Web-Page Ranking, Web-Page Status of Validity, Activity Duration, Length of the URL, Whether the link is a re-direct link or not, Presence of Special Characters such as dashes and ‘@’ and Number of sub-domains. Given all these deciding features as predictors, a Data Mining Model based on Classifications methods such as KNN, Logistic Regression and SVM is to be developed. Such a Data Mining Model can foretell a URL, whether it is phished or not given the above discussed predictors are input to the Model. This can immensely help the vigilance department of the Internet to keep track of the upcoming and newly registered domains and associated links whether they are phished or not. If found to be phished i.e., if the probability of the URL to be phished is on the higher-end, appropriate actions can be taken against the same, thereby saving a lot of Internet users from such fraud. Therefore, as Phishing is something that attacks by fooling the victim psychologically, it is a Passive Attack whereas Viruses and Malware are direct attacks, bull all falling under Cybercrime.

Related Works:

Li et al. [1] proposed an E-Commerce Fraud Detection Model on the basis of Information Fusion Technology that uses several information sources followed by a fusion process. A comparative analysis of the Information Fusion Technology (IFT) Model is done with the other Data Mining Techniques Logistic Regression Model and Support Vector Machine and the IFT Model had proven to be more accurate in Fraud Detection than the other techniques.

Somani and Balachandra [2] developed Data Mining Models to detect phished web links or URLs by employing classification models: C4.5, SVM, Random Forest, Treebag and GBM. Random Forest emerged as the best performing classifier among the 5 classifiers in detecting Phishing in website links obtaining the best accuracy of 97.21%.

Ganesan [3] proposed a Data Mining Model Development using the classification algorithms: Bagging, C4.5 and Random Forest Classifier by training the Model on the UCI Dataset containing Phished Website Links. The C4.5 Model has the highest classification accuracy of 90.8%.

Pandey and Chadawar [4] proposed a Data Mining Model in the form of a hybrid ensemble of weak learners that can be deployed in web browser extensions to detect malicious or illegitimate website links but at the same time, in terms of performance (Accuracy/Error Rate) it has failed to surpass Hybrid SVM-KNN Model and C4.5 Model.

Alsariera et al. [5] developed a Data Mining Method the best first induction process by incorporating best split at every stage of a Decision Tree Algorithm in order to detect Website Phishing. Post this ensemble modelling is performed in terms of Bagging and Boosting in order to elevate Model Performance.

Dhamdhere and Vanjale [6] propose a Data Mining Technique based on Random Forest classification model trained on data related to client interaction and how and what the client shares on the website and the nature of the website.

Patel [7] trained an ANN Model on the on selected set of features extracted from the 5000 website URLs in order to detect phished website links. The performance of the trained ANN Model is compared with Decision Tree, Random Forest and Support Vector Machine among which ANN Model is found to perform the best.

Aljofey at al. [8] developed an XGBOOST Classifier Model to detect Website Phishing by training it on URL character sequence data, hyperlink and textual content of the webpage. The principal advantage of the model is that, it is highly successful in terms predictive performance an accuracy of 98.48%. But the disadvantage of the model is that the model is an ensemble model and hence, can take very long times for re-training purposes.

Proposed Solution:

The following dataset is considered for developing the Data Mining Model for detecting Website Phishing: <https://www.kaggle.com/kunal4892/phishingandlegitimateurls>. It is compiled from various sources among which, PhishStorm is one of the prominent resources, <https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset>. It contains 95,911 records and 12 attributes among which many are fetched from the WHOIS API of Google. The attributes present in the dataset are discussed in detail as follows:

1. domain: It is the Website URL.
2. ranking: It is the web-page ranking. [ordered categorical or continuous feature]
3. valid: Whether the current status of the registered URL is valid (=1) or not (=0), obtained from

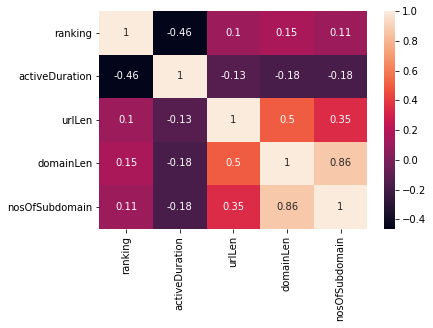
WHOIS API. [categorical feature]

1. activeDuration: Time since the URL is active from the time of registration, obtained from

WHOIS API. [continuous feature]

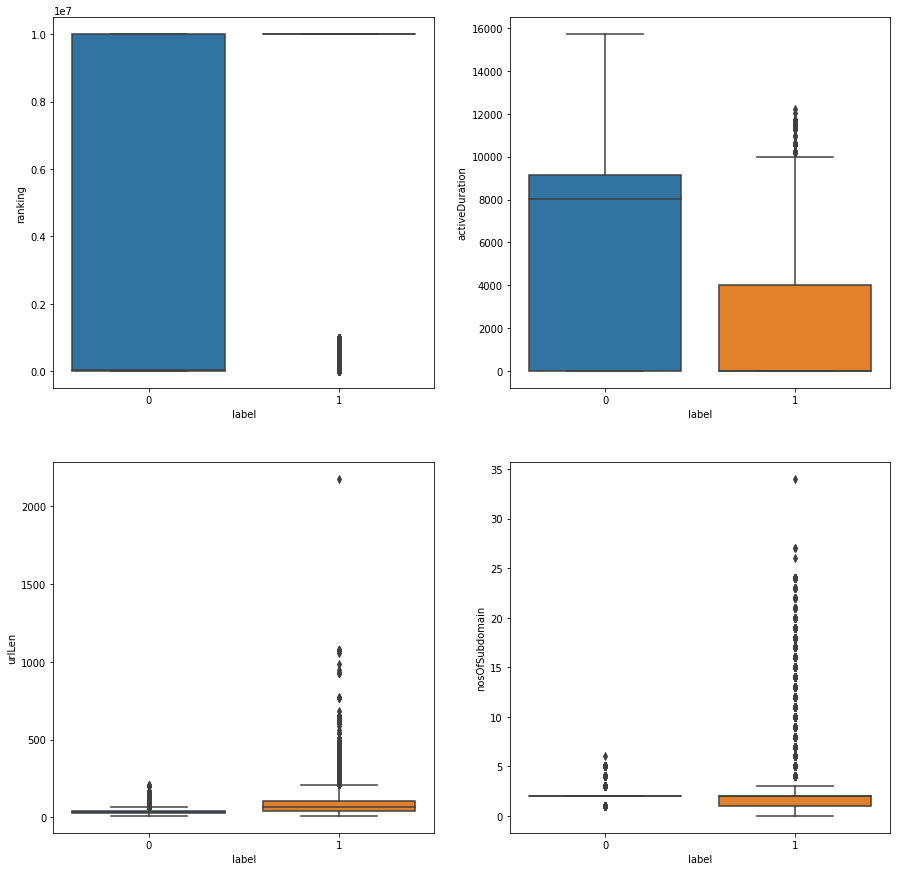
1. urlLen: It is the length of the URL. [continuous feature]
2. is@: Whether the link has ‘@’ character (=1) or not (=0). [categorical feature]
3. isredirect: Whether the link is a redirect link (=1) or not (=0). [categorical feature]
4. haveDash: Whether the domain name contains any dash (=1) or not (=0). [categorical feature]
5. domainLen: Length of the domain name. [continuous feature]
6. nosOfSubdomain: Number of sub-domains present in the URL. [continuous feature]
7. label: phished/spam (0), legitimate (1). [target label]

The Exploratory Data Analysis (EDA) consists of the following:

1. Bivariate Analysis to explore feature-to-feature correlation among continuous features: A Correlation Heatmap is generated taking all the continuous feature variables, ranking, activeDuration, urlLen, domainLen, nosOfSubdomain into account.

Here, there is a very high positive correlation of 0.86 between domainLen and nosOfSubdomain. In order to avoid multi-collinearity that can impact model interpretability, one of the above is dropped. From the heatmap, it can be seen domainLen also has a good correlation of 0.5 with urlLen whereas nosOfSubdomain does not have that strong positive or negative correlation with any of the other features, hence urlLen is dropped from the analysis.

1. Bivariate Analysis to explore feature-to-label relationship for continuous features: A dual axis Boxplot by taking the label in the x-axis and continuous feature in the y-axis is generated for each of the remaining continuous features, ranking, activeDuration, urlLen, nosOfSubdomain.



* 1. Mostly Phished Links or Spam Links have very high ranks (ranking behind) but there are some exceptional pretentious links that may have very good web-page ranks and are the most mis-leading and dangerous.
  2. Phished Links are very rarely having an active duration greater than 10,000
  3. Legitimate Links mostly have 2 sub-domains but there may be exceptions of 1 or within the range of 2-6. On the other hand, Phished Links have a higher possibility to have more than 2 sub-domains and even though rarely, some of them have even more than 6-35 sub-domains.

Application of Data Mining:

Train-Test Split: The dataset is randonly split in the ratio of 75-25 such that 75% of the Data forms the Training Set and 25% of the Data forms the Test Set.

Post this, the feature variables are scaled in the Training Set by fitting the Standard Scaler Object and the same is used to transform the Test Set.

The following are the Classification methods with which the Data Mining Model is trained:

* kNN Classifier: This algorithm uses the intuition of Feature Space Geometry in which each and every data instance is a point in a n-dimensional feature space (n = no. of features). Any test data-point is assigned the class to which majority of the k-nearest neighbors of the data-point belongs to. 5-Fold Grid Search Cross Validation is used to tune the hyper-parameter, n\_neighbors (k) starting from 2 to 10 for obtaining the best hyper-parameter. 9 nearest neighbors is obtained as the best hyper-parameter. So, the kNN Model is trained using 9 nearest neighbors on the Training Set and the performance analysis is done on the Test Set.
* Logistic Regression: This algorithm generates a decision boundary separating data-points into the classes in a n-dimensional feature space where n is the number of features. 5-Fold Grid Search Cross Validation is used to tune the hyper-parameter, C for values, 0.01, 0.1, 1, 10, 100 forming the hyper-parameter space for obtaining the best hyper-parameter. C = 0.01 is obtained as the best hyper-parameter. So, the Logistic Regression Model is trained using C = 0.01 and default Scikit-Learn hyper-parameters on the Training Set and performance analysis is done on the Test Set.
* Support Vector Machine (SVM): This is a kernel-based algorithm that transforms the feature space accordingly and post this, the decision boundary is generated via training that separates the data-points mapped in the new feature space. 5-Fold Grid Search Cross Validation is used to tune the hyper-parameter, C for values, 0.1, 1, 10 and kernel for kernels, ‘linear’ and ‘rbf’, forming the hyper-parameter space for obtaining the best hyper-parameter. C = 10 is obtained as the best hyper-parameter. So, the SVC Model is trained using C = 10, kernel as Radial Basis Function (RBF) and default Scikit-Learn hyper-parameters on the Training Set and performance analysis is done on the Test Set.

Performance Analysis of Data Mining Model:

The Data Mining Models developed using the aforementioned classification methods are evaluated using the following performance metrics: Accuracy, Precision, Recall, F1-Score and Area Under the ROC Curve. A comparative analysis of the Data Mining Models is done and tabulated in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Performance Metrics** | **k-Nearest Neighbor** | **Logistic Regression** | **SVM Classifier** |
| Accuracy | **0.92** | 0.87 | 0.90 |
| Precision | **0.92** | 0.88 | 0.91 |
| Recall | **0.94** | 0.90 | 0.93 |
| F1-Score | **0.93** | 0.89 | 0.92 |
| Area Under ROC Curve | **0.97** | 0.94 | 0.96 |

Table 1. Comparative Analysis among the Data Mining Methodologies

The K-Nearest Neighbour Classifier turned out to be the best classifier among the 3 classification algorithms attempted upon for Phishing Detection in terms of Accuracy, Precision, Recall, F1-Score and Area Under ROC Curve. Also, KNN is the simplest model thereby obeying the Occam’s Razor Principle as well.

Bibliography

[1] Li, J., 2022. E-Commerce Fraud Detection Model by Computer Artificial Intelligence Data Mining. *Computational Intelligence and Neuroscience*, *2022*.

[2] Somani, M. and Balachandra, M., 2022. Detection of phishing websites using data mining tools and techniques. *International Journal of Advanced Intelligence Paradigms*, *22*(1-2), pp.167-183.

[3] Ganesan, S., 2022. Detection of Phishing Websites Using Classification Algorithms. In *Cyber Security and Digital Forensics* (pp. 129-141). Springer, Singapore.

[4] Pandey, A. and Chadawar, J., Phishing URL Detection using Hybrid Ensemble Model.

[5] Alsariera, Y.A., Balogun, A.O., Adeyemo, V.E., Tarawneh, O.H. and Mojeed, H.A., 2022. Intelligent tree-based ensemble approaches for phishing website detection. *J. Eng. Sci. Technol*, *17*, pp.563-582.

[6] Mhaske-Dhamdhere, V. and Vanjale, S., 2022. Classification and Performance Evaluation of Phishing Email or URL Using Random Forest. In *ICCCE 2021* (pp. 797-802). Springer, Singapore.

[7] Patel, J., 2022. Phishing URL Detection using Artificial Neural Network. *International Journal of Research in Engineering, Science and Management*, *5*(4), pp.47-51.

[8] Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q. and Wang, Y., 2022. An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, *12*(1), pp.1-19.