Comparative Analysis of Machine Learning Models for Phishing URL Detection

Ankita Shelke
Master of Science in Data science, Thompson Rivers University

Abstract:

The number of individuals utilising the Internet and the web daily has increased, along with the use of daily activities like online banking, shopping, and other services. This is leading to phishing attacks, in which online criminals steal user information or intrude on their networks on their behalf and harm resources. The objective of the phishing technique is to get or steal user credentials, including usernames, passwords, and credentials for financial operations. Because phishing evolves with technology, we must find a way to stop it by using anti-phishing measures. To avoid phishing attacks, I developed a number of machine learning models that will help researchers in detecting phishing URLs. For this machine learning models I have used logistic regression model, Regularization technique, decision tree and random forest algorithm and artificial neural network. The random forest model, which has an accuracy of 97.97% has the highest accuracy compared to other models.

Introduction:

Due to advancement in internet and cloud technology, people are making more online purchase and transactions. This growth has leads to unauthorize access of user's sensitive information. Phishing attack is one of the tricks through which cyber criminals fool people into clicking malicious link or handing over their personal and important information. Usually, these kinds of attacks are done via emails, text messages, or websites. Phishing websites looks like actual legitimate website and trick users into revealing their sensitive information, such as credit card numbers and login credentials. Due to similar URL to the legitimate website, the phisher convinces the user that they are on genuine website and trick them to enter the personal details. Phishing URLs may be distributed via email, social media, or any other messaging platforms. As per the 2020 Phishing Attack Landscape Report from Great horn (2020 Phishing Attack Landscape 2020), about 53 percent of cyber security professionals have stated that they have witnessed a spike in these attacks during COVID 19 Pandemic, and enterprises are facing about 1185 phishing attacks every month [1] as online transactions have increased in the covid time. It is very crucial for the user to verify for the legitimacy of the website before entering the sensitive information. This can be done by checking https tag means ensuring that website is using secure connection, misspelling in the portion of the website or uncommon domain extension. It is difficult to distinguish between phishing site from the normal site for an average person. Therefore, machine learning models are useful to detect phishing websites; these models are on a dataset of known phishing websites and legitimate websites, and then use this knowledge to predict whether a new website is a phishing site or not.

Data:

Name: Phishing Dataset for Machine Learning [2]
This dataset has been taken from Kaggle and There are no missing values in the dataset. It consists of a collection of legitimate, as well as phishing website instances. Each website is represented by the set of features that denote whether the website is legitimate or not. This dataset contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages, which were downloaded from January to May 2015 and from May to June 2017 [3].

The source for the phishing webpages is PhishTank and OpenPhish whereas the source for Legitimate webpage are Alexa and Common Crawl. This dataset is useful for machine learner researcher to build classification model which will help to prevent phishing attack.
Overall, it has total 49(48 feature + phishing/ legitimate) attributes and 10,000 observations in which class_Label is the response variable which represents 1 for phishing website and 0 for legitimate website. Figure 1 shows the count of phishing and legitimate website. This dataset will be analysed using various machine learning models to find out effective and important features to predict phishing websites.

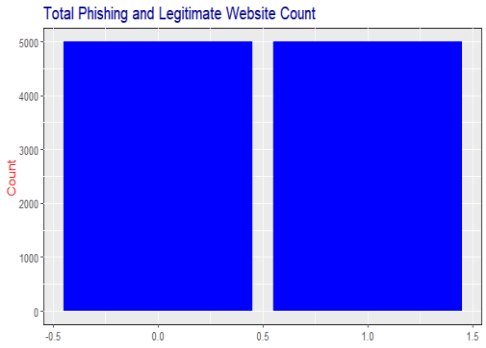

Figure 1: Count of legitimate and phishing URL

Methods:

I have implemented different methods and algorithms on this dataset to develop various machine learning models which will be important in predicting phishing website.

1) Logistic Regression:
   It is a classification algorithm which is used to predict the binary outcome on the basis of independent variables. Logistic regression does not model output variable Y directly, rather it predicts the probability that output variable Y belongs to a particular category. To predict the binary response using multiple predictors is achieved by log odds and is defined as

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + x_1\,\beta_1 + \cdots \quad x_n\,\beta_n$$

   $log\left(\frac{p(x)}{1-p(x)}\right)$: This is called log odds or logit.

   $X = (X_1 \ldots\ X_n)$: These are predictors.

   In this model, if x1 is increased by one unit then it changes the log odds by β_1.[6]
   Confusion Matrix: Confusion matrix is used to calculate the accuracy of any classification machine learning model. Figure shows the confusion matrix with TP, FP, FN and TN. “TN (True Negative) means, the number of benign websites discovered as benign websites. FP (False Positive) means the number of non-phishing websites that were mistakenly identified as phishing websites. TP (True Positive) means, the number of phishing websites that were correctly identified by the model. FN (False Negative) The number of phishing websites that were mistaken for legitimate websites. [7] Accuracy is given by following formula and figure 2 shows accuracy matrix

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

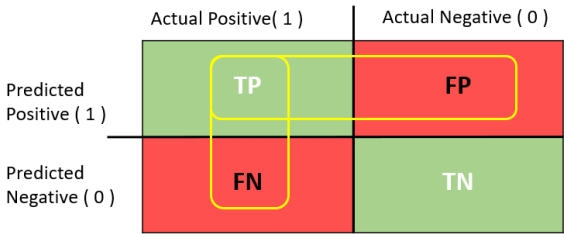

Figure 2: Accuracy Matrix [4]

2) Ridge:
It's also known as l2 regularisation. Due to the shrinking of the parameters, it is primarily used to prevent multicollinearity. By lowering the parameter coefficient, it lessens model complexity. This way, we decrease model complexity while keeping all variables in the model. Penalty is controlled by tuning parameter and it's optimal value is calculated by cross validation technique. It can't be used for variable selection. Cost function for ridge regression is as follows:

$$\min \left(\left| |Y - x(\theta)| \right|_2^2 + \lambda ||\theta||_2^2 \right)$$

3) Lasso:
It's also known as l1 regularisation and it is used for feature selection. So, it is generally used when for high dimension datasets. It also uses cross validation to find optimal tuning parameter for which penalty is less. It makes the coefficient absolute zero and hence preferred for variable selection. We can use the glmnet() package to build the regularized regression models. Cost function for ridge regression is as follows:

$$\min \left(\left| |Y - x(\theta)| \right|_2^2 + \lambda ||\theta||_1 \right)$$

4) Decision Tree:
It is one supervised machine learning algorithm that uses a tree-like structure to model decisions and their possible consequences. It is used for both classification and regression task. This algorithm creates a tree that predicts response variable by recursively splitting the data into smaller subsets based on the most significant feature at each step. Splitting will continue until there is no improvement in prediction accuracy. Decision tree uses entropy and information gain to decide which feature should be selected to split the data. Entropy is a measure of disorder or impurity in a node which is given by following formula [5]

$$E = \sum_{i=1}^{c} (-p_i log_2 p_i)$$

Information gain is a measure of how much information a feature provides about a class. It determines the order of attributes in the nodes of a decision tree. It is given by

$$Information\ Gain = E_{Parent} - E_{Child}$$

5) Random Forest:
It is one of the popular machine learning algorithms, used for classification, regression and other tasks. It constructs many individual decision trees at training. To make final prediction, prediction from all the trees are combined. For classification, majority

voting is considered whereas for regression mean of prediction is considered. This approach helps to reduce overfitting and to improve the accuracy and stability of the predictions. This algorithm is used in several industries like finance, mechanical and healthcare industry.

6) Artificial Neural Network:
It is inspired by human brain and it has same structure like brain. A neural network contains layers of interconnected nodes. It has input layer in which all the inputs provided by the programmer; hidden layer where computations are performed gives the output and output layer which gives the final output after series of computations.

Github_link:
https://github.com/Ankita918/DASC-5420-Project--Phishing-URL-Prediction

Result and Analysis:

The optimal model to detect fraudulent websites will be determined by comparing the performance of the various classifiers and algorithms used to train the phishing URL detection model. The dataset is split into train and test data, and each model has been fitted to the train data after it is evaluated on the test data. On the basis of train data containing 48 features, a logistic regression model was trained, and its accuracy has been found to be 94.88%. According to this model's output, around 24 features are the most significant for the response variable which means these are the most important features for predicting phishing URL.
This dataset has singularity, it means there is at least one feature which is correlated with another feature. To address this problem, I used regularization technique which reduces the dimensionality of the dataset and improve the performance of the model. Figure 3(a) shows the cross-validation curve and we can see that the fitted ridge model, has selected all 47 features to explain the response variables. It did not do feature selection, rather it shrank the coefficients of the parameters which helped avoid multicollinearity. The ridge model's accuracy is 91.96%. Tuning parameter lambda (λ) is used to control the penalty of the model and the optimal value of lambda is calculated using 10-fold cross validation technique. The optimal value of lambda for ridge is 0.027. Figure 3(b) shows that as lambda increases, the coefficient if the parameter decreases and it also shows than important features for the response variable.
In the case of the Lasso model, it chose 43 features that are crucial to detecting phishing URLs. The minimal lambda is calculated using 10-fold cross validation, and it is equal to 0.00019. Figure 4(a) is the cross-validation curve in which first vertical dotted line shows the minimum value of lambda for which penalty is minimum whereas second vertical line shows the value of lambda(lambda.1se) that gives the most regularized model such that the cross-validated error is within one standard error of the minimum. From Figure 4(b), we can state that in the case of the lasso, as lambda increases, the coefficient of parameters drops and, for some parameter, reaches zero, making it less significant for the model. So, it selected 43 important features for the model. Accuracy for this model is 94.84%.

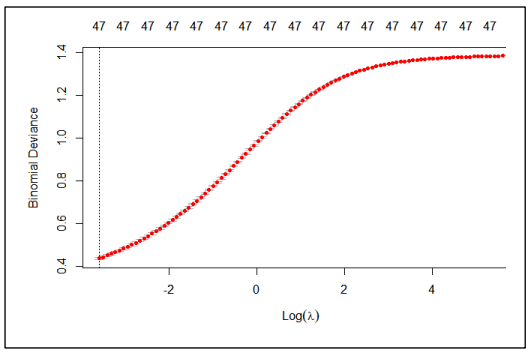

Figure 3(a): CV in Ridge

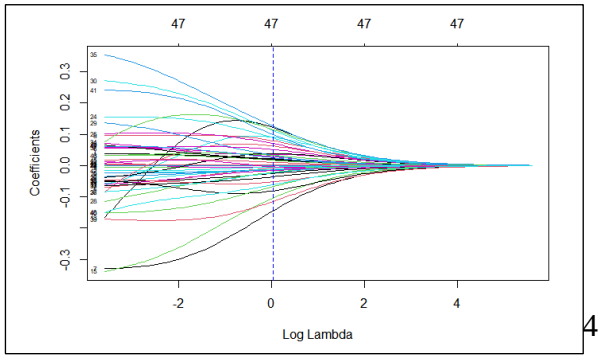

Figure 3(b): Coefficient vs lambda

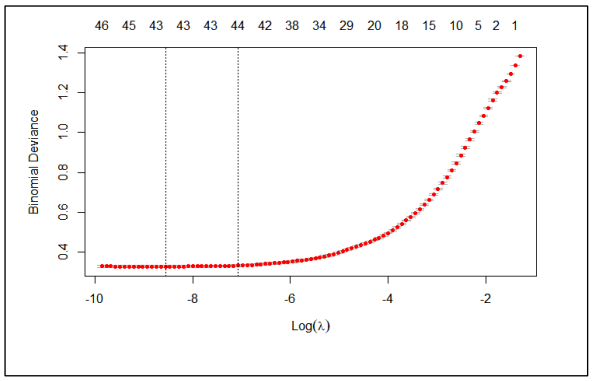

Figure 4(a): CV in Lasso

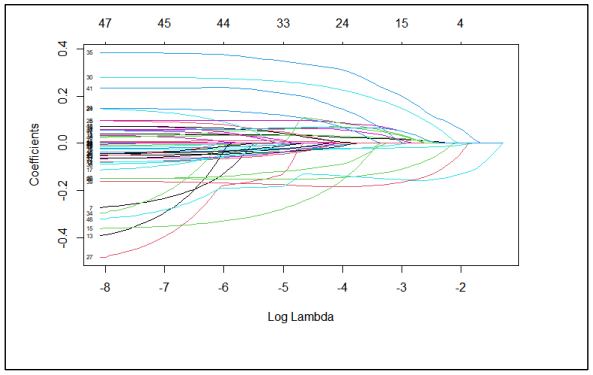

Figure 4(b): Coefficient vs lambda

Random forest model gives the accuracy of 97.97%. Figure 5 shows the graph between error and number of trees which helps to tune the parameter for better accuracy. Top 10 important variables in which improves the accuracy of the model can be seen in the figure 6
Decision tree algorithm is used to train the model on train data and evaluated its performance using cross validation and pruning and made prediction on the test data to find out whether the URL is phishing or not. The accuracy of the model is 94.6 % and figure shows the final decision tree which used four variables to split the data.

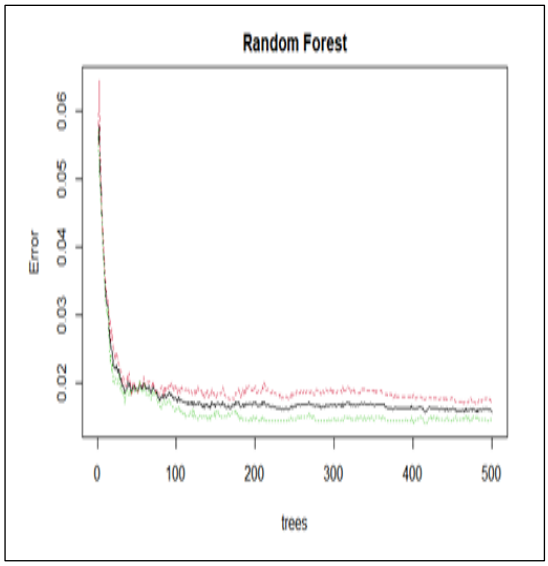

Figure 5: RF-Error vs Trees

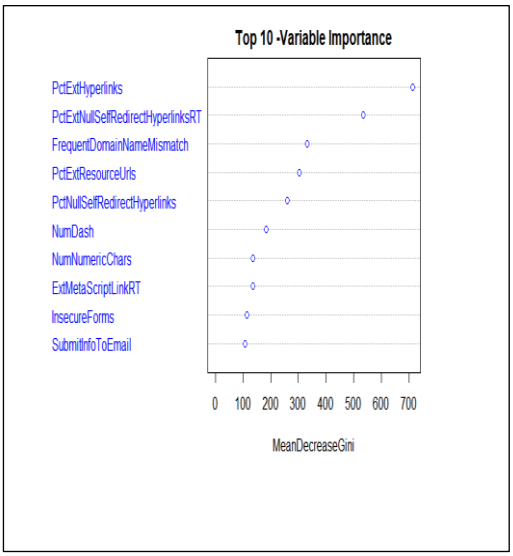

Figure 6: Importance Variable

In case of neural network, in order to reduce computing complexity and assess the model's accuracy, I chose 1000 observations from 10,000 tuples in the study of artificial neural networks. The accuracy for the neural network is 92% on the basis of 1000 observations. Figure7 shows the neural network for the phishing data with input layer, 2 hidden layers and a output layer.

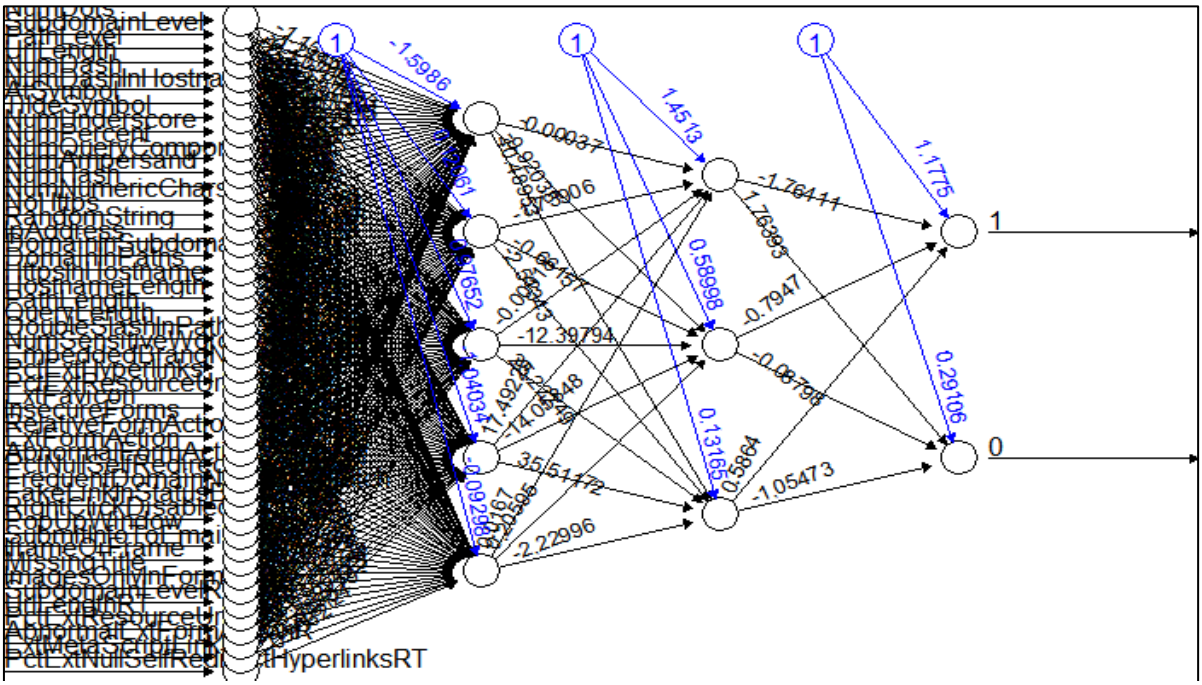

Figure 7: Neural Network for this Dataset

Table 1: Comparison of all models:

Classification algorithm	Accuracy
Logistic Regression	94.88%
Ridge	91.96%
Lasso	94.84%
Decision Tree	94.6%
Random Forest	97.97%
Artificial Neural Network	92%

Discussion:

Cybersecurity is becoming a challenge for internet users as a result of the technology's rapid progress. It might be quite challenging to recognise phishing URLs, but machine learning algorithms make it possible to create a machine learning model that can distinguish between authentic and fraudulent URLs. This analysis was carried out by building different models using logistic regression, ridge regression, lasso regression, decision tree, random forest and neural network and compared these models on the basis of accuracy. Random forest has the highest accuracy among all models.

References:

1. GreatHorn. 2020 Phishing Attack Landscape [Internet]. info.greathorn.com. [cited 2023 Apr 16]. Available from: https://info.greathorn.com/report-2020-phishing-attack-landscape
2. Phishing Dataset for Machine Learning [Internet]. www.kaggle.com. Available from: https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning
3. Choon Lin Tan. Phishing Dataset for Machine Learning: Feature Evaluation. mendeley [Internet]. 2018 Mar 24;1. Available from: https://data.mendeley.com/datasets/h3cgnj8hft/1
4. Precision, Recall, F1 스코어 등의 모델 평가 방법 [Internet]. Soon's Blog. 2022 [cited 2023 Apr 16]. Available from: https://meme2515.github.io/machine_learning/performance_measurement/
5. Entropy and Information Gain to Build Decision Trees in Machine Learning [Internet]. Engineering Education (EngEd) Program | Section. Available from: https://www.section.io/engineering-education/entropy-information-gain-machine-learning/
6. Ziegler A. An Introduction to Statistical Learning with Applications. R. G. James, D. Witten, T. Hastie, and R. Tibshirani (2013). Berlin: Springer. 440 pages, ISBN: 978-1-4614-7138-7. Biometrical Journal. 2015 Dec 7;58(3):715–6.
7. Bhavani PA, Chalamala M, Likhitha PS, Sai CPS. Phishing Websites Detection Using Machine Learning [Internet]. papers.ssrn.com. Rochester, NY; 2022. Available from: