DATA ANALYTICS
ECS 784P

Coursework 1

Using Machine Learning techniques to detect Phishing in webpages.

- Semil Halani
(210147012)

ABSTRACT

This paper is based on the findings of predictions made to detect phishing webpages; Phishing is a method of social engineering attack often used to steal personal data of a user or an entity, including login credentials and credit card numbers using deceptive websites and e-mails. Support Vector Classifier, Extra Trees Classifier as well as Logistic Regression have been utilized here as machine learning methodologies. This dataset consists of 10,000 rows and around 50 columns, and has features which are correlated with websites and their characteristics to specifically be a Phishing website. We assess the models, and are explained why they were specifically used and then, present the final results. Data management and data cleansing has also been discussed as well as exploratory data analysis of the dataset has been done in this report. We will also discuss and make some conclusions based on the analysis and study of the machine learning model. A literature review of related works is also presented here to expand the scope of the discussion for the reader. An evaluation of Support Vector Classifier, Extra Trees Classifier and Logistic Regression is also conducted to identify pros and cons of both models and more specifically the accuracy and precision of each models. Traditionally detecting phishing webpages was really tough and depended a lot on the keen and observation skills of the user, however, the recent advancements in Machine Learning and Data Science techniques can now give a quantitative value to the likelihood or risk of someone accessing a phishing website. Extra Trees Classifier and Logistic Regression, both are examples of Supervised Learning Algorithms The findings are presented visually for ease in comprehending for the user.

KEYWORDS

Phishing detection, logistic regression, decision tree classifier, machine learning in cyber security

INTRODUCTION

This project will focus on the process of predicting a phishing website using Extra Trees Classifier and Logistic Regression models. These models will attempt to predict if a specific website is a phishing website or not. Extra Trees Classifier will create decision trees based on the data to output a classification of whether a website is or is not a phishing website. This is a real- world problem and the report findings could be used in the real-life applications, given, there needs to be industrially levelled research and optimization of algorithms, but we attempt to try and create a stepping stone for this type of data and issue. Some important characteristics of a website used to judge if it is a phishing website or not are, SubdomainLevel, PathLevel, UrlLengthRT, ExtMetaScriptLinkRT etc.

According to the National Cyber Security Centre, “Phishing is when attackers attempt to trick users into doing 'the wrong thing', such as clicking a bad link that will download malware, or direct them to a dodgy website.” Phishing can be done via text messaging, social media, or by phone, however the term 'phishing' is mainly used to describe attacks that arrive by email directing users to the attacker’s website. Phishing emails can reach millions of users directly, and hide amongst the huge number of daily emails that busy users receive. Attacks like this, can lead to the attacker installing malware (for example ransomware), sabotage systems, or steal intellectual property and money from the victim's device.

In 2021, it was estimated by RiskIQ that businesses worldwide lose $1,797,945 per minute due to cybercrime and the average breach costs a company $7.2 per minute. *IBM*'s 2021 research into the cost of a daa breach ranks the causes of data breaches according to the level of costs they impose on businesses. According to IBM, *Phishing* ranks as the second most expensive cause of data breaches, a breach caused by phishing costs businesses an average of $4.65 million. Nevertheless, that's not the only way Phishing can lead to a costly breach; attacks using compromised credentials were ranked as the fifth most costly cause of a data breach (averaging $4.37 million).

LITERATURE REVIEW

Here, in this section, a literature review and related discussions will be discussed. The work in discussion will be then evaluated with respect to the research problem being examined.

The first book in consideration for the discussion is the *Machine Learning Approaches In Cyber Security Analytics* by Thomas, Tony et al. and a report from the *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* with the title "Phishing detection using Machine Learning Techniques" by Santhi H, Supraja, Basi Reddy A and Sailaja G. It discusses about the social and financial implications of Phishing and all the losses and repercussions of not being able to detect phishing in the initial stage. It also discusses about some solutions for the same like, Authentication, Web Application Security, anti-phishing toolbars and so on. It also discusses about some machine learning techniques which can be used in order to detect Phishing before becoming a victim to it and thereby, prevent it from happening.

Moreover, the other extracts to be considered are from the book "*A Machine Learning Based Approach For Phishing Detection Using Hyperlinks Information"* by Jain, Ankit Kumar, and B. B Gupta, from the *International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019)* with the title "Detection of Phishing Websites using Machine Learning Approach" by Kahksha and Sameena Naaz, from the *International Journal of Recent Technology and Engineering (IJRTE)* with the title "Phishing Websites Detection Using Machine Learning" by R. Kiruthiga and D. Akila and a Survey of Machine Learning-Based Solutions for Phishing Website Detection by Tang L. and Mahmoud Q. in *Machine Learning and Knowledge Extraction*. All of these literary works talk about various techniques and machine learning models which can be used for the detection of phishing websites. Some of these machine learning techniques are Decision Tree, Random Forest, Generalized Additive Model, Bayesian Additive Regression Trees, Logistic Regression, Support Vector Machines, Neural Networks and so on.

DATA MANAGEMENT

This dataset was found from Kaggle. In this dataset, we have 50 attributed including the target attribute and 10,000 data instances. This makes our dataset with the shape of 10000x50. The phishing website is marked as 1 whereas the one which is not a phishing website is marked 0. Similarly, all the data features have their values assigned and assorted.
If there are missing values in the dataset, then they could have been solved by the following methods:

- use the mode value for discrete data and mean value for continuous data,
- eliminate the rows with missing data as our dataset is quite large and removal of some values wouldn't make a noticeable impact on our decision model, or
- use KNN or K-means method to replace the missing values, etc.

However, our dataset had no missing values. Thus, the need for this part of pre-processing the data was eliminated.

EXTERNAL LIBRARIES USED:

Numpy: A library that has support for large multidimensional arrays and matrices, mainly used for high-level math functions for operations on these arrays and matrices.

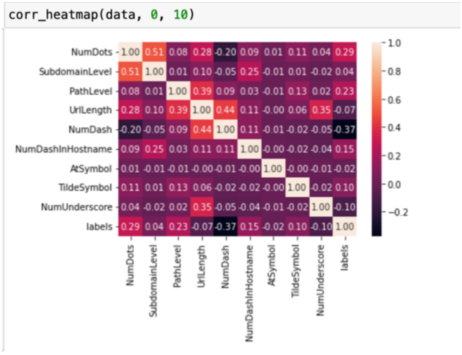
Pandas: This is a popular data framework library in Python, with the help of which one can manipulate and present data.

Matplotlib: A library for visualisation of data

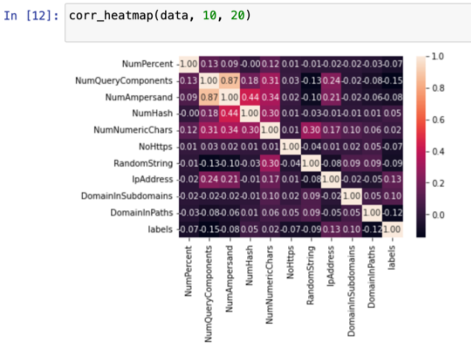
Scikit-learn: A machine learning library in Python, here used for the implementation of the algorithms as it is easy and straightforward to implement.

EXPLORATORY DATA ANALYSIS:

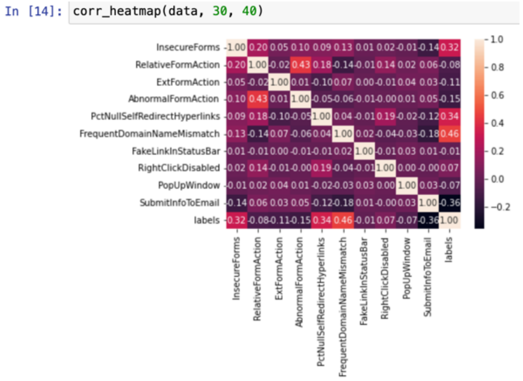
Applying the correlation heatmap function, we get the following outputs:



```
corr_heatmap(data, 0, 10)
```

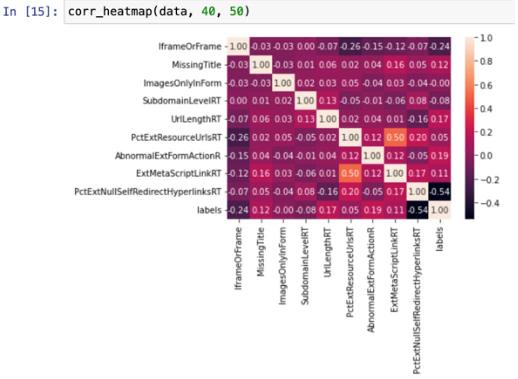


```
In [12]: corr_heatmap(data, 10, 20)
```



```
In [14]: corr_heatmap(data, 30, 40)
```

InsecureForms 1.00 0.20 0.05 0.10 0.09 0.13 0.01 0.02 -0.01 -0.14 0.32
RelativeFormAction 0.20 1.00 -0.02 0.43 0.18 -0.14 -0.01 0.14 0.02 0.06 -0.08
ExtFormAction 0.05 -0.02 1.00 0.01 -0.10 0.07 0.00 -0.01 0.04 0.03 -0.11
AbnormalFormAction 0.10 0.43 0.01 1.00 -0.05 -0.06 -0.01 -0.00 0.01 0.05 -0.15
PctNullSelfRedirectHyperlinks 0.09 0.18 -0.10 -0.05 1.00 0.04 -0.01 0.19 -0.02 -0.12 0.34
FrequentDomainNameMismatch 0.13 -0.14 0.07 -0.06 0.04 1.00 0.02 -0.04 -0.03 -0.18 0.46
FakeLinkInStatusBar 0.01 -0.01 0.00 -0.01 -0.01 0.02 1.00 -0.01 0.03 0.01 -0.01
RightClickDisabled 0.02 0.14 -0.01 -0.00 0.19 -0.04 -0.01 1.00 0.00 -0.00 0.07
PopUpWindow -0.01 0.02 0.04 0.01 -0.02 -0.03 0.03 0.00 1.00 0.03 -0.07
SubmitInfoToEmail -0.14 0.06 0.03 0.05 -0.12 -0.18 0.01 -0.00 0.03 1.00 -0.36
labels 0.32 -0.08 -0.11 -0.15 0.34 0.46 -0.01 0.07 -0.07 -0.36 1.00
InsecureForms
RelativeFormAction
ExtFormAction
AbnormalFormAction
PctNullSelfRedirectHyperlinks
FrequentDomainNameMismatch
FakeLinkInStatusBar
RightClickDisabled
PopUpWindow
SubmitInfoToEmail
labels
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



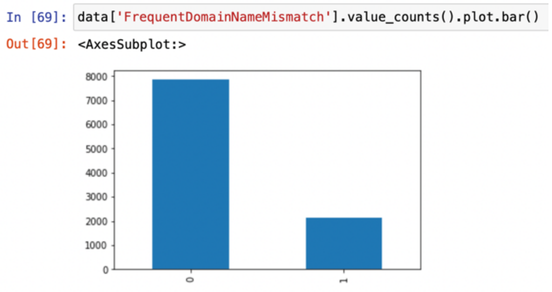
```
In [15]: corr_heatmap(data, 40, 50)
```

IframeOrFrame 1.00 -0.03 -0.03 0.00 -0.07 -0.26 -0.15 -0.12 -0.07 -0.24
MissingTitle -0.03 1.00 -0.03 0.01 0.06 0.02 0.04 0.16 0.05 0.12
ImagesOnlyInForm -0.03 -0.03 1.00 0.02 0.03 0.05 -0.04 0.03 -0.04 -0.00
SubdomainLevelRT 0.00 0.01 0.02 1.00 0.13 -0.05 -0.01 -0.06 0.08 -0.08
UrlLengthRT -0.07 0.06 0.03 0.13 1.00 0.02 0.04 0.01 -0.16 0.17
PctExtResourceUrlsRT -0.26 0.02 0.05 -0.05 0.02 1.00 0.12 0.50 0.20 0.05
AbnormalExtFormActionR -0.15 0.04 -0.04 -0.01 0.04 0.12 1.00 0.12 -0.05 0.19
ExtMetaScriptLinkRT -0.12 0.16 0.03 -0.06 0.01 0.50 0.12 1.00 0.17 0.11
PctExtNullSelfRedirectHyperlinksRT -0.07 0.05 -0.04 0.08 -0.16 0.20 -0.05 0.17 1.00 -0.54
labels -0.24 0.12 -0.00 -0.08 0.17 0.05 0.19 0.11 -0.54 1.00
IframeOrFrame
MissingTitle
ImagesOnlyInForm
SubdomainLevelRT
UrlLengthRT
PctExtResourceUrlsRT
AbnormalExtFormActionR
ExtMetaScriptLinkRT
PctExtNullSelfRedirectHyperlinksRT
labels
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

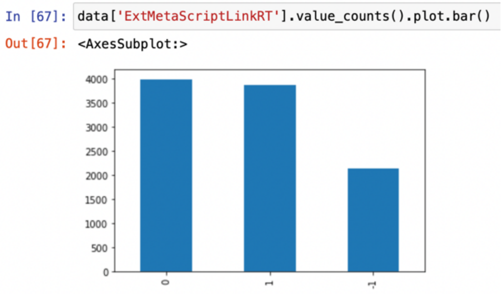
From these graphs, we come to a conclusion that there is only one column over here that has some correlation which has a negative effect with labels which is, PctExtNullSelfRedirectHyperlinksRT.

Also, it can be noted that InsecureForms shows that as the value is higher so the probability of being a phising site PctNullSelfRedirectHyperlinks shows the same positive correlation as InsecureForms FequentDomainNameMismatch shows that it has medium linear correlation in positive direction SubmitInfoToEmail seems to indicate that sites that ask users to submit their details to emails seems to be more high probability for phising

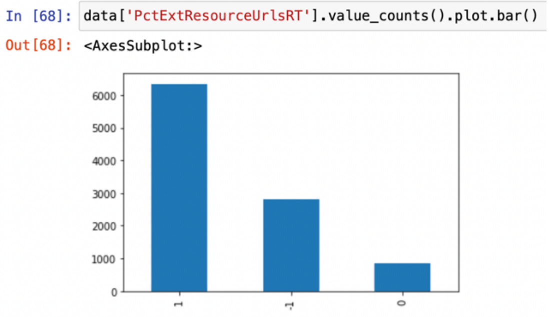
Upon visualising a few other attributes, we get the following results which helps us further in better understanding of the problem at hand:



```
In [69]: data['FrequentDomainNameMismatch'].value_counts().plot.bar()
Out[69]: <AxesSubplot:>
```



```
In [67]: data['ExtMetaScriptLinkRT'].value_counts().plot.bar()
Out[67]: <AxesSubplot:>
```



```
In [68]: data['PctExtResourceUrlsRT'].value_counts().plot.bar()
Out[68]: <AxesSubplot:>
```

MACHINE LEARNING MODELS:

Here, we have used 2 machine learning models, namely Logistic Regression, Extra Trees Classifier and Support Vector Classification for detecting the phishing websites. Logistic regression is a process of modeling the probability of a discrete outcome when we are given an input variable. The most common logistic regression models a binary outcome, which we can use as classifying between 0 and 1. Logistic regression, even though it has *regression* in its name, is a classification model rather than a regression model. It is a simple and more efficient method for binary and linear classification problems. It is a very easy classification model to realize and achieves very good performance with linearly separable classes. On the other hand, Extra Trees Classifier implements a meta estimator that fits a lot of randomized decision trees (also called extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. When it comes to Vector Support Classification (SVC), the implementation is based on libsvm. The fit time for that scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples. For large datasets we prefer using LinearSVC or SGDClassifier instead, possibly after a Nystroem transformer. Also, the multiclass support is handled according to a one-vs-one scheme.

ACCURACY

Considering the accuracy for Logistic Regression, it was the maximum for when the number of features were 42, which was 94.75%. On the other hand, when using the Support Vector Classification (SVC), we get an accuracy of 94.05%.

CONCLUSION

It is believed that that the project aims have been achieved and the set objective aims have been clearly identified in the report along with evidence. The results motivate future work to explore inclusion of additional variables to the data set, which might improve the predictive accuracy of classifiers. For instance, analysing email headers has proved to improve the prediction capability and decrease the misclassification rate of classifiers. In addition, we will explore developing an automated mechanism to extract new features from raw phishing emails in order to keep up with new trends in phishing attack.

REFERENCES

- Jain, Ankit Kumar, and B. B Gupta. *A Machine Learning Based Approach For Phishing Detection Using Hyperlinks Information*.

- *International Journal of Innovative Technology and Exploring Engineering*, 2019. Phishing Detection using Machine Learning Techniques. 8(12S2), pp.73-78.

- Thomas, Tony et al. *Machine Learning Approaches In Cyber Security Analytics*. Springer, 2020.

- https://www.sciencedirect.com/topics/computer-science/logistic-regression

- Tang, L. and Mahmoud, Q., 2021. A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *Machine Learning and Knowledge Extraction*, 3(3), pp.672-694.

- Ahmed, k. and Naaz, S., 2019. Detection of Phishing Websites Using Machine Learning Approach. *SSRN Electronic Journal*,.

- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

- Kalaycı, T., 2018. Comparison of Machine Learning Techniques for Classification of Phishing Web Sites. *Pamukkale University Journal of Engineering Sciences*, 24(5), pp.870-878.

- *International Journal of Recent Technology and Engineering*, 2019. Phishing Websites Detection using Machine Learning. 8(2S11), pp.111-114.

- https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

- D, Akila, and Kiruthiga R. *Phishing Websites Detection Using Machine Learning*. 2019.