ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

A Multi-Modal Hierarchical Attention Model for Phishing Threat Intelligence: An Explanatory Framework

Dr.Malathi.P
Department of MCA
Dhanalakshmi Srinivasan College of Engineering and Technology

Ms.Reshma.A
Department of MCA
Dhanalakshmi Srinivasan College of Engineering and Technology

Abstract: Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media system use spoofed emails form legitimate companies and agencies to enable users to use fake websites to divulge financial details like username and passwords. We have identified different features related to legitimate and phishy websites and collected 1353 different websites from difference sources. Phishing websites were collected from Phishtank data archive which is a free community site where users can submit, verify, track and share phishing data. The legitimate websites were collected from Yahoo and starting point directories using a web script developed in PHP. The PHP script was plugged with a browser and we collected 548 legitimate websites out of 1353 websites. There is 702 phishing URLs, and 103 suspicious URLs. The main Scope is to detect the phishing website Prediction, which is a classic text classification problem with a help of machine learning algorithm. It is needed to build a model that can differentiate between phishing website OR not.

I. INTRODUCTION

Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights
from data across a broad range of application domains.

Data Scientist:

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data.

EXISTING SYSTEM

Existing CTI for phishing website detection methods can be divided into three types: lookup systems, fraud cuebased methods, and deep representation-based methods. The lookup system detects a phishing website by "looking up" the website URL against a blacklist of phishing URLs and an alarm is raised when the website's URL appears in the list. The blacklists are classifiers (e.g., SVM, decision tree) and novel machine learning methods (e.g., statistical learning theory based methods, genre tree kernel methods and recursive trust labeling algorithm) have been devised to detect phishing websites. Similarly, website traffic based fraud cues requires to analyze the website traffic within a period of time, making them hard to meet the real-time detection requirement.

DISADVANTAGES OF EXISTING SYSTEM

- It takes more time to make the transfer learning if we want to change some features and train the model.
- They are not mentioning the Accuracy of the model.
- The performance metrics like recall F1 score and comparison of machine learning algorithm is not done.
- The performance is not good and its get complicated for other networks.

PROPOSED SYSTEM

The proposed model is to build a machine learning model for anomaly detection. Anomaly detection is an important technique for recognizing fraud activities, suspicious activities, network intrusion, and other abnormal events that may have great significance but are difficult to detect.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

The machine learning model is built by applying proper data science techniques like variable identification that is the dependent and independent variables. Then the visualization of the data is done to insights of the data.

The model is build based on the previous dataset where the algorithm learn data and get trained different algorithms are used for better comparisons. The performance metrics are calculated and compared.

ADVANTAGES OF PROPOSED SYSTEM

- The anomaly detection can be automated process using the machine learning.
- The Accuracy level of Machine Learning Algorithm Model is Calculated.
- Performance metric are compared in order to get better model.

MODULES:

1. Data Pre-processing
2. Data Analysis of Visualization
3. Comparing Algorithm with prediction in the form of best accuracy result
4. Deployment Using Flask

MODULES DESCRIPTION

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset.

Module Diagram

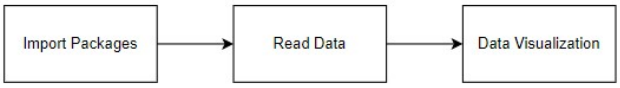
Given Input Expected Output

Input : data
Output : removing noisy data

Exploration data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

Module Diagram



Given Input Expected Output

Input: data
Output: visualized data

Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare.

Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data.

Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

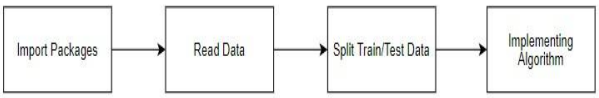
- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

- The independent variables should be independent of each other. That is, the model should have little.

Module Diagram



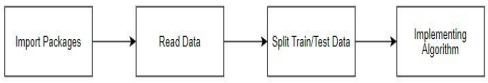
Given Input Expected Output
Input : data
Output : getting accuracy

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.
The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 & 2

Module Diagram



Given Input Expected Output
Input : data
Output : getting accuracy

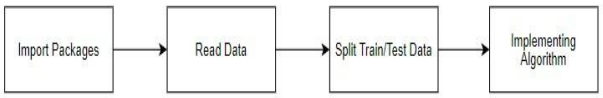
Decision Tree Classifier

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Module Diagram



Given Input Expected Output
Input : data
Output : getting accuracy

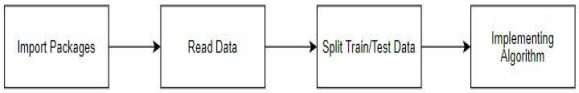
Naive Bayes algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms.Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.
Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

Module Diagram



Given Input Expected Output
Input: data
Output: getting accuracy
Deployment Using Flask

Flask (Web FrameWork):

Flask is a micro web framework written in Python. It was created by Armin Ronacher of Pocoo, an international group of Python enthusiasts formed in 2004. According to Ronacher, the idea was originally an April Fool's joke that was popular enough to make into a serious application.

any extensions *you think you need. Also you are free to build your own modules. Flask depends on two external libraries: the Jinja2 template engine and the Werk zeug WSGI toolkit.*

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

Sample Screens

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of phishing website or not.

REFERENCES

- P.Prakash, M.Kumar, R.R.Kompella, and M.Gupta, "PhishNet : Predictive Blacklisting to Detect Phishing Attacks," 2010.
- Bradley Barth,"SOC teams spend nearly a quarter of their day handling suspicious emails,"

  https://www.scmagazine.com/home/email-security/soc-teams-spend-nearly-a-quater-of-their-day-handling-suspicious-emails.2021.
- Crane Hassold, "Employee-Reported Phishing Attacks Climb 65% Clobbering SOC Teams,"

  https://www.agari.com/email-security-blog/employee-reported-phishing-attacks-soc/.2020.
- A. Y. Fu, W. Liu, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD), "IEEE Trans. Dependable Secur. Comput., vol. 3, no.4,pp.301-311,2006,doi: 10.1109/TDS.2006.50.