An investigation of social engineering security threat using supervised machine learning classification algorithm

ABSTRACT

The danger of network security is becoming increasingly a threat to both individuals and companies. This may lead to attack of network system and exploit human sensitivity. Therefore, it is a high-level importance to protect network security against attacks including phishing emails. Thus, this project analyzes a dataset of 10000x50 shape and understand the relationships and correlation of features. This extends to the use of plots to visualize the dataset features and observations more distinctly. This is not limited to visualization and getting intuition about the dataset but also preprocessing data for null values and imbalanced/balanced dataset checking. The main objective of the final project is to create a suitable machine learning model to predict either a random email is phishing or benign. The tools used in this project are popular packages such as sklearn, NumPy and Pandas. The progress to reach the most suitable model is by trying different techniques and machine learning supervised algorithm. This includes linear, logistic regression, K-Nearest-Neighboring, Decision Tree, Random Forests and Neural Network.

INTRODUCTION

Information security is a high-level priority for all companies and individuals. Human attacking, social engineering, is becoming increasingly a security issue for the offense that affects information systems. By definition, social engineering is the act of manipulating employees or a person to give up conferential information and to get access to accounts or highly confidential network security protocol. Therefore, people are receiving intelligent and deceptive emails in order to fool them to open a link or a suspicious document. Thus, this report explains the advantage of technology to predict whether an email is phishing or benign by using supervised machine learning classification algorithm.

PLAN

Machine Learning (ML) is a very large technology used in data science with a verity of algorithms to predict outputs. Initial assumption is that supervised learning, more specifically K-Nearest-Neighbor is a useful algorithm to predict the output either an email is phishing or benign. As a matter of fact, during bootcamp we have covered different algorithms not only to predict output but also to analyze and visualize data by the use of popular Python libraries such as NumPy, Panda, Sklearn and Matplotlib. Thus, the plan is to analyze the data and understand it, afterwards approach the main objective by using ML algorithm in order to predict spam emails.

PROGRESS

This section illustrates the procedure of both analyzing and predicting the output step by step and also commenting in each step. In fact, full code will be shared.

READING DATA

Table

Description automatically generated Figure 1 shows the first 5 observation of the data and the output named “Class Label” either class 0 or 1 which represents either benign or phishing email and it is considered as a binary class including the shape of the data which is 10,000x50.

Fig 1. Reading Dataset Using Pandas library

Graphical user interface, text, application

Description automatically generated

CHECKING NULL in DATASET

One of the important procedures in data analysis is to analyze data and make sure that the data does not have null information. Therefore, figure 2 shows that the data has zero null values for all observations, features and outputs.

Fig 2. Checking Null values

DESCRIBING DATA

Figure 3 explains the count, mean, min, max, standard diviation of the data and other more statistical values including percentiles. It has been analyzed the data and it has been found that the minimum length of Uniform Source Locator (URL) is 12 characters and the average or mean of the number of underscores is 0.3232 which indicates that the number of underscores is low. One important observation is that the mean of the output is 0.5 which illustrates that the class labeling is balanced between phising emails and benign emails.

Fig 3. Statistical information

Graphical user interface, table

Description automatically generatedGETTING INSIGHT ABOUT THE DATA

Graphical user interface, text, application, email

Description automatically generatedFigure 4 shows that several information about the data. For example, there are 322 missing title tag in the HTML source code. Additionally, it checks how many forms containing the contend of URL without HTTPS protocol. In addition, there are 2153 out of 10,000 matches of the most frequent domain hostname in the HTML source code with the URL of the provided

Fig 4. Sum of some features in the dataset

website. Moreover, there are 140 JavaScript

command in the HTML source code that disable the right click. Thus, this section illustrates a general insight of what is happening after analyzing the data and get as much information as possible.

CLEANING DATA

Table

Description automatically generatedCleaning data is an essential task that a data scientist must take into consideration during the data analysis. Cleaning data can be illustrated by giving examples such as removing duplicate rows or columns, removing null values which affects the machine learning model, or other more. To illustrate more, it has been found out by using Pandas library with

Fig 5. Describing a feature

describe method that the feature HttpsInHostname has

Graphical user interface

Description automatically generated with medium confidenceall zeros in all observations. Figure 5 highlights

the description of feature HttpsInHostname with all zeros numerical values which can be dropped from the data since it has no contribution to the output. Therefore, as a data science specialist, it is better to drop useless columns. Figure 6 shows the method of dropping a column within Place to be true in order to assign changes to the original data. An additional point that should be taken into consideration during

Fig 6. Dropping a column

Table

Description automatically generated with medium confidencecleaning data is that the first column of the dataset shows id feature which does not represent a valuable meaning, it acts as an index of a number of observations ranging from 1 to 10000. Thus, figure 7 shows a beneficial approach by dropping

Fig 7. Dropping indexing column

the column as well to have a dataset with

meaningful features that can predict the output either an email is phishing or a ham email.

EXCTRACT FEATURES AND OUTPUT

Table

Description automatically generatedFigure 8 shows the advantage of using Pandas library in order to slice a data frame to have y as an output with class label and x to be features of dataset containing 47 different features to have an interesting problem to be solved. As a matter of fact, the built-in function used is iloc which helps in slicing a data frame by index, another built-in function can be useful is loc function which also helps in slicing the dataset by column names instead of indices. Thus, now we are able to deal with both features and outputs separately and also, we can combine

Fig 8. Extracting features and labels

them together to have a full dataset.

VISUALIZING RELATIONSHIP BETWEEN DATA SERIES

Chart, bar chart

Description automatically generatedThis section includes some of the analysis in the dataset because these plots obtained from seaborn and matplotlib helps to visualize the data using different types of plots and it also helps to examine the relationships between features. Figure 9 visualizes the relationship between the number of underscores in a random email with the URL length using bar plot from matplotlib library. In fact, we can see that

Fig 9. Bar graph of Underscore vs URL Length

Chart, scatter chart

Description automatically generatedthe maximum URL length is obtained when

the number of scores is 0 and the minimum URL length is approximately 125 which matches 16 underscores. Additionally, figure 10 shows a scatter graph between two main features which are the percentage of external hyperlinks and percentage of URL external resource in HTML source codeChart, box and whisker chart

Description automatically generated.

Fig 10. Scatter plot of ExtHyperlinc vs ResourceURL

LINEAR REGRESSION ALGORITHM

Graphical user interface, text, application

Description automatically generated The output, class label, is clearly a binary class which indicates that linear regression is not a suitable algorithm to be implemented. Even though output is not numerical, it is better to try to use a linear regression for the sake of practice and draw a conclusion. The linear regression model is implemented via sklearn library with both Linear Regression class and R squared function. By

Fig 11. R squared of Linear model

definition, R squared is a statistical measurement

Graphical user interface, text, application, Word

Description automatically generatedthat represents the proportion of the variance for

the output by the features in a regression model.

To illustrate, it has been defined an object of Linear Regression model then fitting both features and output in order to get the parameters w which are called the coefficients of the features. In addition,

Fig 12. R squared of Linear model after filtering

it has been defined a predicted output variable to

predict the output by giving the function the features.

Graphical user interface

Description automatically generatedTherefore, it is applicable to compare the predicted values with the existing output. Figure 11 shows the R squared value which indicates 0.692 roughly. In other words, 69.2% of the independent variables can explain the dependent variable which is either an email is phishing or benign. As a matter of fact, the output of a linear regression is numerical while the required output is only two states either 1 to show a phishing email or 0 to show a benign email. Thus, by filtering the predicted output to have 1 for numerical values above 0.5, and 0 for numerical values less than 0.5. Figure 12 illustrates that by filtering the output, the R squared become better by 6%. Seaborn library is also a tool that shows the correlation between variables in a heatmap which helps the analyst to see which feature is highly correlated with. Figure 13 shows a portion of the heatmap in x-axis all features and output and in y-axis is also all columns.

Fig 13. Heatmap of all features and class label relationship

Fig 13. Heatmap of all features and class label relationship

LOGISTIC REGRESSION ALGORITHM

Chart

Description automatically generatedLogistic regression is a supervised machine learning model that classifies an output with a categorical form. Compared to linear regression, logistic regression gives better predictions for both binary class such as the required output of this final project or a multiclass with categories more than two. Therefore, this section will show the steps and the explanation of implementing logistic regression to predict either an email is phishing or not. Figure 14 shows a scattering plot of hostname length

Fig 14. Scatter plot of hostname vs class label

Chart, box and whisker chart

Description automatically generatedindependent variable versus the class label output of either a spam email or not by

using matplotlib tool which is a library for

plotting. As a matter of fact, in the previous algorithm implementation in linear regression, it has been used all the dataset as training which is not a practical approach. Therefore, in this algorithm, most likely 80% of the dataset will be used for training and 20% will be used for testing the logistic regression algorithm.

Fig 15. Relation of URL Length and Class Label

One additional essential point that must be

Graphical user interface, text, application, email

Description automatically generatedtaken into consideration is sigmoid function. Sigmoid function is given by , where x is a numerical value from (-infinity, infinity). In fact, the range of sigmoid function is [0,1]. Thus, any numerical value will output a value between zero and one which makes logistic regression to be a useful analysis in binary classification. Figure 15 shows the relation between urlLength feature and the output. It is observed that it is difficult to know the class label by using only the urlLength which indicates that there are lots of urlLength that share the same

Fig 16. Accuracy of Logistic Regression Model

number with both binary classes.

Moreover, standardScaler object provided by sklearn tool helps in scaling the input features to have a mean of approximately zero and standard deviation to be 1. Thus, it is now applicable to apply logistic regression. One important point is that cross validation is an essential concept to use for better results. By definition, cross validation is resampling process to test the accuracy of a machine learning algorithm by training and testing different samples. In fact, this procedure can be used for different values of the parameter c of the logistic regression. Additionally, model selection provided by sklearn library helps in using cross validation. As a result, the accuracy score has become much better compared to linear regression. The accuracy of logistic regression is 94% as shown in figure 16. All in all, logistic regression indicates a better accuracy compared to linear regression which is not a surprising information due to the fact that logistic regression can deal with binary class.

K NEAREST NEIGHBOR MACHINE LEARNING ALGORITHM

A picture containing text

Description automatically generatedText

Description automatically generatedK Nearest Neighbor or KNN for short is considered one of the simplest supervised machine learning algorithms. This section explains the use of KNN algorithm by using sklearn library. First of all, there has been a training and splitting data in order to use some of the data to train the model and the other part as testing of the model. Figure 17 explains how to define a KNN model using sklearn and also fitting the model to obtain the predicted output and also it shows the percentage accuracy of the

Fig 17. KNN Model

model which is 86.55%. One important

terminology that must be taken into consideration is cross validation. Be definition, cross validation is a method used in applied machine learning to test the ability or the skill of a model in predicting future data. Figure 18 shows different ranges of K and using 10 cross validation which enhances the accuracy by 0.3%.

Fig 18. Tunning Parameter

DECISION TREE & RANDOM FORESTS MACHINE LEARNING ALGORITHM

Graphical user interface, text, application

Description automatically generatedDecision Tree is a machine learning algorithm which represents a flowchart with binary decisions with different nodes and leaves. This section will illustrate the use of decision tree and see the accuracy percentage of this popular algorithm. Sklearn is a useful library that can be used in order to define the decision tree classifier. Figure 19 shows the accuracy of predicting

Fig 19. Decision Tree Accuracy

the class label by using decision tree algorithm.

On the other hand, random forest is considered to be one of the most popular machine learning algorithms that perform better compared to decision tree. By definition, random forest is simply a wide range of decision tree with Graphical user interface

Description automatically generated with medium confidenceselecting some of the features, this is known as bootstrapping. In addition, by constructing

Fig 20. Random Forests Accuracy

different decision tree, it is applicable to feed the test

data to all decision trees and get the final results of each decision tree. The final perdition of random forest algorithm is achieved by aggregation which is simply voting or getting the maximum class perdition. Since random forest can solve the problem of scalability of dataset compared to decision tree, this illustrates the fact that random forest is much more stable in variance when dataset is changed. As a matter of fact, random forest performs well compared to decision tree. Figure 20 shows the percentage of accuracy obtained by implementing random forest algorithm.

NEURAL NETWORK

Text

Description automatically generated

Neural Network (NN) is a subset of machine learning and considered to be the main heart of deep learning. The name is extracted from human brain connection and the behavior is mimicking the way neurons in human brain send signals to another neurons. This section will illustrate the use of neural network on training data and test the accuracy of the algorithm. The tool used to implement neural network is Keras. Keras is a deep learning API written in Python. Figure 21 shows the deep learning model constructing using keras library. There is one

Fig 21. Neural Network Model

hidden layer with an activation function of sigmoid

Text

Description automatically generated with medium confidenceranging from zero to one value. In addition, the input layer takes the feature values of shape (10000,47) and being fed to the neural network algorithm. Moreover, the hidden layer consists of 5 units. One of the famous optimizers in neural network is adam. Therefore, it has been used adam

Fig 22. Neural Network Accuracy & Loss

optimizer in compiling the algorithm with

loss binary cross-entropy. As a matter of fact, optimizers in deep learning decrease the error function and increase the efficiency or accuracy of the model. Additionally, binary crossentropy is a loss function used in binary classification such as the intended problem. To illustrate, the intended problem of this project is a binary output either 0 or 1 where 0 is a safe email and 1 is a phishing email. Figure 22 shows the accuracy of 97.69% and loss of 0.0637.

CONCLUSION

To sum up, this report analyzed a dataset of binary classification obtained from Kaggle with a shape of 10000x50. The dataset represents different observations with class labeling that shows a random email is phishing or benign. It has been analyzed the dataset and visualize relationship between features of the dataset including plotting with class label output. It has also been examined different machine learning algorithms to reach a goal where the model is applicable to predict an email is safe or phishing. It has been used linear regression, logistic regression, KNN supervised machine learning algorithm, decision tree, random forests, and neural network. It has been obtained a result that random forests outperforms other used machine learning algorithm in this report with an accuracy of 98% in predicting the class label.