MACHINE LEARNING – LEARNING CYBERSECURITY

By

Akash Sarode

Author: Akash Sarode, akky_sanj@yahoo.com
CEH, CHFI, ACE, Cyber security professional with 5 + years of experience
Owner of the course: "Machine Learning: The Future"

Abstract

Today, in this world of technology we get to know something new every day. So one such technology which is growing day-by-day is Machine Learning. Now Machine learning can find its application in multiple industry but we would be focusing on "**Machine Learning in Cyber Security**". Today, a lot of malware is being created and utilized. To solve this problem, many security solutions are using Machine learning to detect and prevent malwares. Using such technology will prevent the industry in improving their line of defense. Therefore, I think that if we can identify whether machine learning is really worth utilizing and how is it used in Cybersecurity to enhance the security, we can improve our security in real world.

Table of Contents

Is It Really Learning?

We are living in the world of technology where technology moves swiftly. Not many could survive this speed and hence they are left behind. Technology plays an important part in our day-to-day life and one such technology which is changing the world in a tragic way is "Machine Learning". Be it your smartphones, online shopping, Facebook, Google ads, Automobile, everywhere machine learning is finding its way to improve and advance technology. And such aid improves user experience and working in a great deal. Our focus in this article would be more on Machine Learning in Cyber Security. So our aim for this article would be to understand -

"Is Machine Learning, Really Learning?"

If YES, How Is It Learning in Cyber security?

How is it used?

Is it Accurate?

Machine Learning & Artificial Intelligence

Most of the people must have heard various applications of Machine learning being applied. Also many must have come across a word "**Artificial Intelligence**" alongwith Machine Learning. So what is Artificial Intelligence AI? Artificial Intelligence is intelligence showed by machines which will mimic like human. So we can say that Machine Learning is way by which we can achieve Artificial Intelligence. Machine learning is about training a machine to learn on itself and Artificial Intelligence is once a machine is capable of learning on itself, it can imitate, work and react like human. Mostly people do get confused between these terms, but we should be clear about both these terms and their perspective. Our focus would be more on Machine Learning and its magic we are about to apply.

*"Computers are able to see, hear and learn. Welcome to the future." -**Dave Waters***

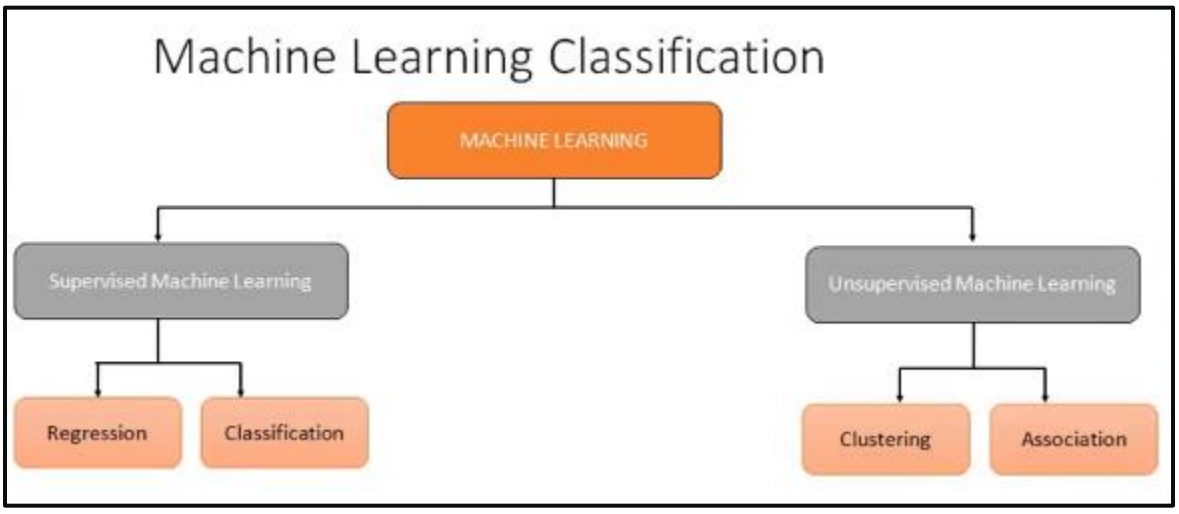
So it's the era where computers are able to perform these activities and we never could have imagined these 10-15 years back.

Recent Transformation

Technology is changing rapidly. Most of the big companies do held some technical conferences where people do talk about recent & new technologies being discovered and how is it improving the IT industry. If you have walked around in any such conferences, you must have heard dozens of vendors/executives talking about AI and Machine Learning. Multiple products, applications do find their way to use machine learning to provide useful outcomes. This marks the importance of the technology and the way it's changing the world. So what is we try to understand this technology and decide whether it's really worth. Let's try and understand the technology and conclude the discussion with a decisive statement.

Understanding Machine Learning

At its simplest level, machine learning is defined as the ability for computers to learn without being explicitly programmed. To understand how ML works we first need to understand the fuel that makes ML possible: data. Today whatever we do, we create/generate data. So as the 21st century progress, there would be large amount of data in Exabytes and we need the machines to work for us to process huge data and provide us with valuable outcome. Machine Learning has various models which are developed using python or R language. These models are basically ways in which the algorithm inputs, processes and outputs the data. Types of machine learning includes Supervised and unsupervised machine learning. Various types of models are available such as Regression, Classification, Association, Clustering.



Deep Learning is also a type of Machine learning which tries to mimic human brain and predicts the outcome based on Artificial Neural networks.

The basic process in developing the machine learning algorithm is to identify the business problem and try to resolve it using machine learning. Machine Learning approach involves various components like dataset, model to built, etc. First step involves pre-processing data where we convert/process data

in the form which is understandable by machines. During the process of data pre-processing, we divide the dataset into training set and test set. Our dataset would consist of entries which will help our Machine learning model to learn.

Also, during this data pre-processing phase, we identify independent and dependent variables in our dataset. Basically, dependent variable is the actual outcome which we want to predict using machine learning model and the independent variables have relevance which will help in predicting dependent variable. After pre-processing data, we built our model by fitting the machine learning model with the training set. This simply means that the machine learning model will learn based on our training set which was a part of dataset.

So, our model has learned now, next step is to predict the test set results based on the learning of machine learning model and compare those results with actual value. You may observe that the predicted values and actual values will be approximately similar as the machine learning model has learned and it is now able to predict outcomes. Based on training set, machine has learned and it is now trying to predict output based on its learning.

Machine learning algorithms essentially build models of behaviors and use those models as a basis for making future predictions based on newly input data. This is the magic of Machine learning. And this is how machine learning works.

Cybersecurity

Enough of Machine Learning, let's understand a new term "**CYBERSECURTY**". So we all know what is cyber security and how important is cyber security in our present digital world. Cyber security comprises technologies, processes and controls that are designed to protect systems, networks and data from cyber attacks. Now-a-days, many organization do implement multiple high-end security products at its each layer of defense as a best practice for cyber security. Firewalls, IDS/IPS, Endpoint protection solution and many others are included in these security products stack. But modern day attackers are bypassing these security solutions and they are disrupting the business of an organization causing great loss for the company. So, we need more effective approach to improve our line of defense. We have discussed enough of Machine Learning technology. Wouldn't it be great if we can apply Machine Learning in Cyber Security.

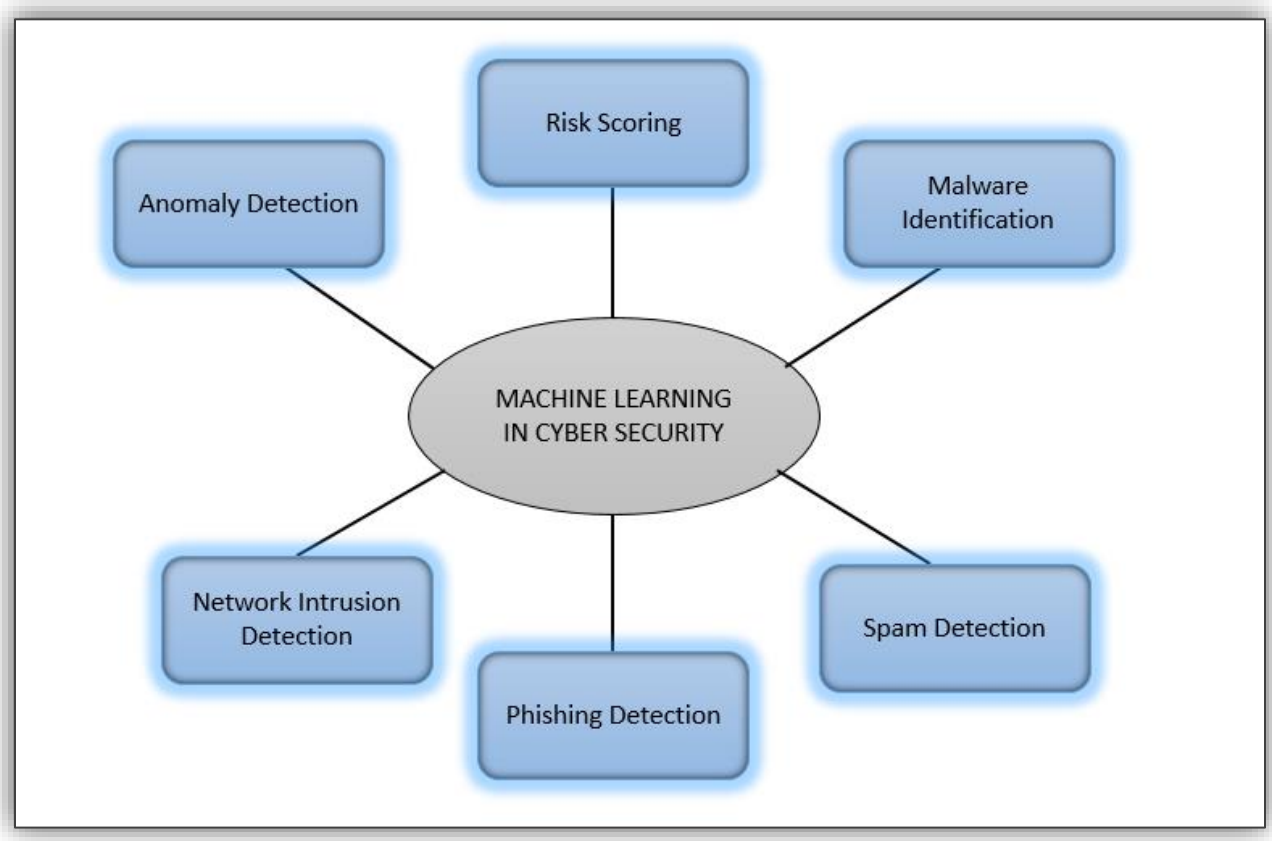
So here, we have a business problem of identifying attacker and Machine Learning is our solution to the problem.

- ***Akash Sarode***

If we can apply the Machine Learning technology to Cybersecurity, can we achieve our goal which is to secure our machines. The combination seems to be interesting and let's try to analyze and predict if such blending can add addition to our lives by improving our security.

Machine Learning + Cyber Security

Machine Security is of utmost importance for any organization and we need to blend in both these technologies together i.e. Machine Learning and Cyber Security to achieve this. There are multiple ways in which machine learning can be used in cyber security.



Anomaly detection, malware identification, Spam detection, Phishing detection, network intrusion detection, and many more are various application of Machine Learning in Cyber Security. So our problem is to understand whether Machine Learning is really worth learning data and if yes, how is it learning and how can it

be applied to cyber security. To answer all these questions, we will practically demonstrate an application of ML in cyber security and will also confirm about the accuracy of the approach.

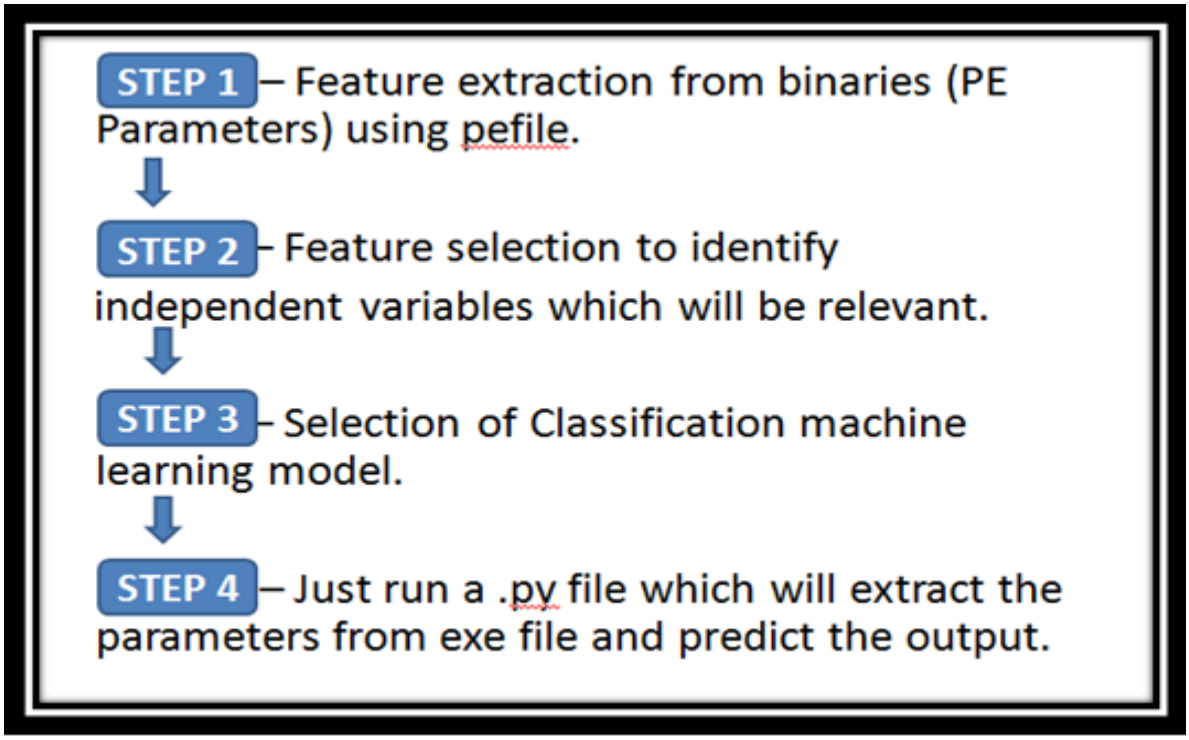
Let's start with Malware detection (Anti-Malware solution) using MACHINE LEARNING.

Build Anti- Malware using Machine learning

Let's start Machine Learning in Cyber security! We will build a Machine Learning based Anti-malware solution by using classification model. We can code this Machine learning application in python or R language. Traditional AV or anti malware solution works on signatures. These signature based techniques can be easily bypassed. Polymorphic malware bypass these traditional detection solutions very easily. Malware detection is a classification problem. Machine learning learns the dataset and then based on its learning, It predicts : malware/not malware. Supply more and more latest dataset to improve machine learning model performance.

Aim would be to identify whether a given binary is legitimate or malicious. Our dataset consists of n number of entries of executables (exe) and its properties. These exe will be malware executable as well as legitimate windows exe files. Once, our Machine Learning model learns based on our supplied dataset, it will predict the outcome i.e. classify the given input file as Malware or legitimate. This would be fun.

Let's look at the steps which we need to follow to build our Anti-Malware solution using ML:-



STEP 1

Our first step is to extract all the PE parameters of binary file using pefile. Now we will be using a large dataset which will be useful in training in classification Machine Learning model. So, we would include both legitimate as well as malicious binary files in our dataset. Our dataset would consist of all the PE parameters of legitimate and malicious files. Some of these parameters are Name, md5, Machine, SectionsMeanEntropy, ResoucesNb, FileAlignment and many more. So all these data is stored in a large csv file.

Here we are with large CSV file ready for playing!

STEP 2

So now, all those parameters which we extracted are all independent variables and there would be 1 dependent variable which would be an identification variable whether binary is Malware/ Not Malware. Next step is feature selection to identify independent variables which would be relevant for differentiating

legitimate binaries from malware. There are some algorithms which have been developed to identify the most interesting features and reduce the dimensionality of the data set. One such algorithm is Tree-based feature selection. Kindly refer below website to understand more on this algorithm - http://scikit-learn.org/stable/modules/feature_selection.html#tree-based-feature-selection

```
#Anti-Malware using Machine Learning

# Importing the libraries
import pandas as pd
import numpy as np
import pickle
import sklearn.ensemble as ske
from sklearn import cross_validation, tree, linear_model
from sklearn.feature_selection import SelectFromModel
from sklearn.externals import joblib
from sklearn.metrics import confusion_matrix

# Importing the dataset
data = pd.read_csv('data.csv', sep='|')
X = data.drop(['Name', 'md5', 'legitimate'], axis=1).values
y = data['legitimate'].values

print('Researching important feature based on %i total features\n' % X.shape[1])

# Feature selection using Trees Classifier
fsel = ske.ExtraTreesClassifier().fit(X, y)
model = SelectFromModel(fsel, prefit=True)
X_new = model.transform(X)
nb_features = X_new.shape[1]
```

So after applying Tree-based feature selection, algorithm selected some features which would be relevant in classification of malware files.

STEP 3

Next step is selection of classification model. There are multiple classification machine learning models available such as Logistic regression, K Nearest neighbors, SVM, Naïve Bayes, Decision Tree, Random forest model. We can use any one of these classification model. If want, we can also compare multiple classification models, based on score method being applied on the machine learning model. We could just save the classifier in a folder which would be helpful in next step.

```
#Algorithm comparison
algorithms = {
        "DecisionTree": tree.DecisionTreeClassifier(max_depth=10),
        "RandomForest": ske.RandomForestClassifier(n_estimators=50),
    }

results = {}
print("\nNow testing algorithms")

# Fitting Classification algorithms to the Training set
for algo in algorithms:
    clf = algorithms[algo]
    clf.fit(X_train, y_train)
    score = clf.score(X_test, y_test)
    print("%s : %f %%" % (algo, score*100))
    results[algo] = score

winner = max(results, key=results.get)
print('\nWinner algorithm is %s with a %f %% success' % (winner, results[winner]*100))

# Save the algorithm and the feature list for later predictions
print('Saving algorithm and feature list in classifier directory...')
joblib.dump(algorithms[winner], 'classifier/classifier.pkl')
open('classifier/features.pkl', 'w').write(pickle.dumps(features))
print('Saved')

# Predicting the Test set results
y_pred = clf.predict(X_test)
```

STEP 4

Now, our model has learned based on input dataset. Next step involves running a python script along with an input attribute of a binary file which we want to test. These python script will extract the parameters from input file and it will load the classifier which has been saved and predict the outcome whether the input binary file is malicious or legitimate.

```
(py27) C:\Users\akash.sarode\Desktop\Anti-malware>python checkpe.py notepad.exe
The file notepad.exe is legitimate

(py27) C:\Users\akash.sarode\Desktop\Anti-malware>python checkpe.py metasploit.exe
The file metasploit.exe is malicious

(py27) C:\Users\akash.sarode\Desktop\Anti-malware>python checkpe.py FreeISOBurner.exe
The file FreeISOBurner.exe is malicious

(py27) C:\Users\akash.sarode\Desktop\Anti-malware>
```

So, this is how Machine Learning is applied in CyberSecurity. For a demo session on this application, please refer:-

https://www.youtube.com/watch?v=Kf9VD1os_pY

Accuracy of Machine Learning

Machine Learning is an algorithm and its developed so that machines can learn on themselves without being explicitly programmed. So I won't say that there won't be any false positives but we need to provide the model with latest, datasets just so that it can learn on some more additional values and can improve its accuracy.

Road Ahead

In this article, we came across two terms which when combined can do wonders. As we progress in 21st century, there could be more such approaches of blending these technologies to mitigate our problems. Identify such kind of business problems in your organization and try to solve those problems with the help of Machine Learning. Applying Machine Learning to Cyber Security will not only improve our cyber security but also allow us to progress towards the more efficient technology.

References

- www.kaggle.com – *Place to do data science projects.*

- Machine Learning and security: Protecting systems with Data and Algorithms – *A book by Clarence Chio and David Freeman.*

- https://www.researchgate.net/publication/283083699_Applications_of_Machine_Learning_in_Cyber_Security

- www.analyticsvidhya.com

- www.kdnuggets.com – *Resources for Machine Learning and Cyber Security.*

- Introduction to Machine Learning with Applications in Information Security – *Book by Mark Stamp*

Ways to connect

https://twitter.com/akky2892

https://www.youtube.com/channel/UCzJnWB-dI8DrLGM-LZ-4FWw

https://in.linkedin.com/in/akash-sarode-b0193091