

**AI Tools & Techniques For Cybersecurity**

Year 2/3 (2024/25) Semester 4/6

**SCHOOL OF INFOCOMM TECHNOLOGY**

Diploma in Cyber Security & Digital Forensics

Assignment Writeup

|  |  |  |
| --- | --- | --- |
| Name | Student ID | Grade |
| Oh Wei Qin | S10257453H |  |
| Shane | S10259959E |  |

# Project Description

This report will document an in-depth research on a chosen cybersecurity dataset. All 7 steps in the data science lifecycle; business understanding, data mining, data cleaning, data exploration, feature engineering, predictive modelling and data visualization will be included.

This report will consist of creating a problem statement based on the dataset sourced. Research on findings together with analysis and evaluation will be carried out on the dataset by developing a machine learning model that will make predictions to assist in answering the problem statement.

# Problem Statement

How can Cybersecurity analyst develop effective and efficient prevention and control measures for cyber-attacks by analysing data such as known attack signature collected from logs of cyber-attacks.

# Dataset Sourced

The cybersecurity dataset was sourced from Kaggle. The data set sourced is a synthetic cyber dataset, it offers for a realistic representation of cyber-attacks and logs collected. The dataset contains 25 metrics and 40,000 records.

# Members Work Contribution

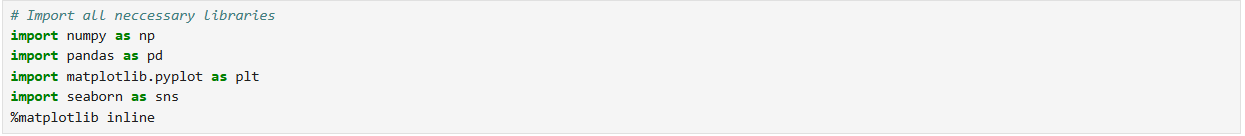
Wei Qin: Problem Statement, Dataset Sourced, Data Cleaning, Predictive Modelling, Findings

Shane: Project Description, Data Research, Exploratory Data Analysis, Model Evaluation, Conclusion

# Data Research

## Import libraries

Basic libraries such as numpy, pandas, matplotlib and searborn are imported



## Import csv file

df = pd.read\_csv(‘filename’)

Data frame is displayed to reveal information about the dataset.

A screenshot of a computer

Description automatically generated

Basic Analysis

1. Finding basic information of dataset

A screenshot of a computer

Description automatically generated

df.shape tells us how the size of the data such us the number or rows and columns that we are using for analysis. That way we can ensure that we are using the most efficient and effective methods to handle and analyse the data.

A screenshot of a computer

Description automatically generateddf.head.T allows shows us the first five rows of the data as a table; it also gives us the names of the columns and values associated which are being used in the dataset. This provides us with knowledge of what data we are dealing with and helps us think of ways to handle and analyse the data.

A screenshot of a computer

Description automatically generated

df.info again gives us the names of the columns that is present in the data set. Additionally, it provides the data type of each column, knowing the data type for each column is necessary as it will help us with the data cleaning process and let us know how the data should be handled.

A screenshot of a computer error

Description automatically generateddf.columns gives us the name of the columns that are in the dataset and further confirms the column names if any were missing.

A screenshot of a computer

Description automatically generateddf.describe gives us the statistics of the data that can be calculated. These statistics can provide us with crucial info such as mean, standard deviation, min, max etc. Knowing these statistics will help us summarize, interpret and validate the data.

1. Check for any missing/null values

A screenshot of a computer

Description automatically generated

df.isnull.sum provides us with critical information about the dataset we are using, it tells us the columns which have missing values or no values in it and provides us with the number of rows that the column has missing data. This will help us to clean our dataset further by removing the null values or using the mean or median values which was found in the describe command. Removing or replacing these null values make our dataset more reliable and accurate thus further improving on our analysis and modelling of the dataset.

Columns with null/missing values: ‘Malware Indicators’, ‘Alerts/Warnings’, ‘Proxy Information’, ‘Firewall Logs’, ‘IDS/IPS Alerts’.

## Data Cleaning

* Fill up missing or null values

A screenshot of a computer program

Description automatically generatedAs from the df.isnull.sum command, we have traced down the following columns as having null values. The columns being ’Alerts/Warnings’, ’Malware Indicators’, ’Proxy Information’, ’Firewall Logs’, and ’IDS/IPS Alerts’. Through our own analysis we have conclude that filling up the missing values and the null values would be the best course of action thus we have used the above method.

The code shown in the screenshot will check for any missing/null values; if found, the corresponding value will replace the empty value.

Confirm if missing and null values are cleaned

A screen shot of a computer

Description automatically generated

We then use the df.isnull.sum command once again to look through and ensure that no more null values are present.

* Check for duplicated values

A white box with black text

Description automatically generated

We then use df.duplicated.sum to check for any duplicated vaules. Checking for any duplicated values is important as it could affect the analysis and the accuracy of the results.

* Enhance “Device Information Column” to extract Devices and Operating Systems

A screenshot of a computer

Description automatically generated

The code shown above extract only the browser keyword from ‘Device Information’ and create a new column from it. This is to allow for better data visualization and analysis later as not all the content is important.

A screenshot of a computer program

Description automatically generated

A white card with red and blue text

Description automatically generated

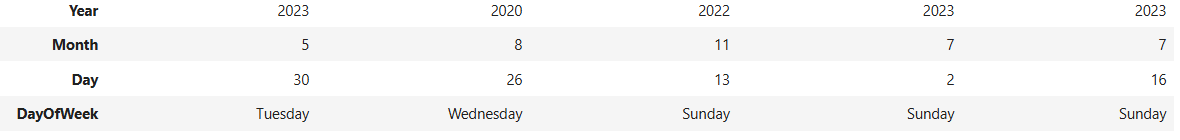
To further help in our analysis of this dataset, we have decided to find out more information about the device. (e.g. device used, device’s operating system, etc.)

The code shown above look for corresponding devices/OS in the values array; if found, the device/OS will be appended into a new column created ‘Device/OS’.

Knowing more about the devices that are being attacked and what their specifications are. We will be able to better grasp which devices are more vulnerable and pinpoint exactly what is to be fixed and present the findings with a clearer graph.

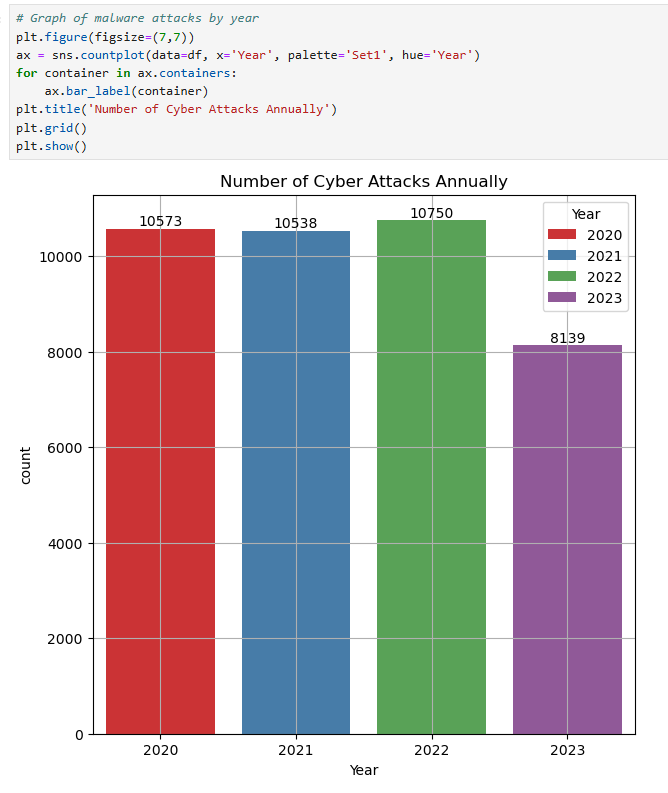
* Ensuring proper datetime format and creating individual columns for year, month, day and day of the week

A screen shot of a computer code

Description automatically generated We can further improve our analysis of this dataset by formatting the timestamps of the attacks and dividing them into thier own columns. This will allow us to see each years statistics and also see which months and days have the most attacks and use relate it to real world events that happen which could impact the number of attacks. This will also allow us to better visualize it in a graph.

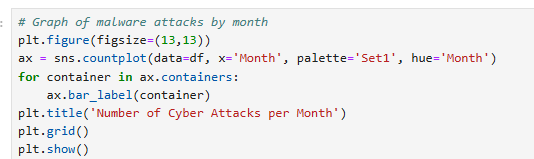
## Exploratory Data Analysis

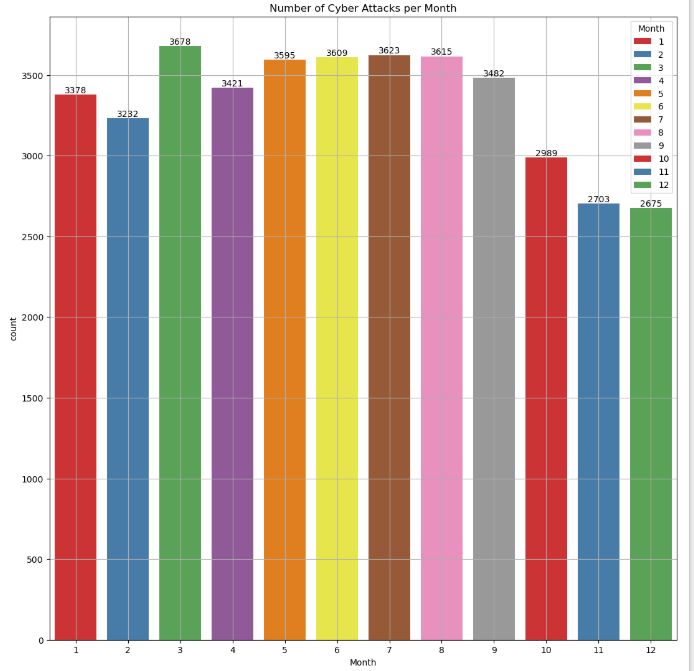
### Annual, Monthly and Daily Cyber Attacks Graph



Using the cleaned data frame that we have created. We start by analysing the number of attacks that happen each year, we have concluded that a count plot would be the best plot to represent this data. Count plot allows for easy comparison of number of attacks between each year.

From the graph above we can see that the attacks between 2020 to 2022 have stayed in about the same range of 10,500 –11,000, with 2022 having the greatest number of attacks at 10,750. This can be inferred that during 2022, cyber-attacks are more prevalent, and security is not sufficient. However, in 2023 it dropped to 8139. The reduce in attacks could be due to more people being educated about safety in the cyberspace or better security measures.



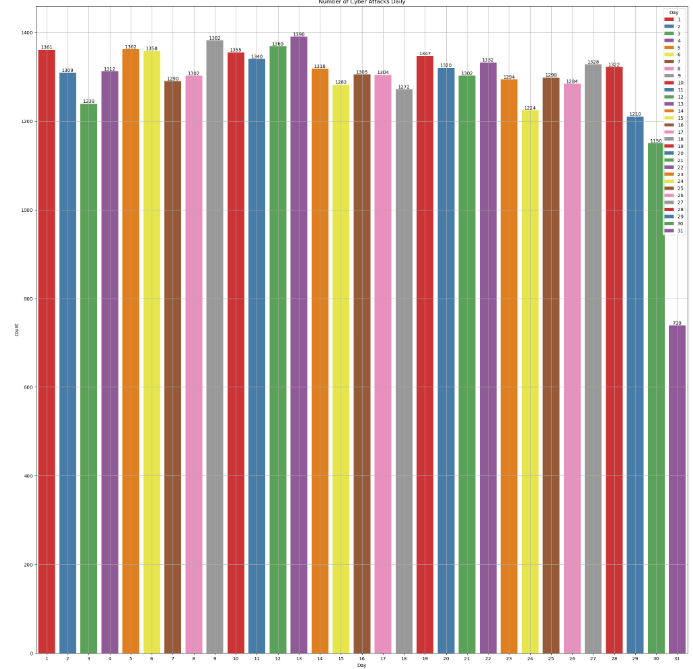


We analysed the number of attacks that occur every month. We have decided that the best representation on this data would be through a count plot. The count plot allows us to easily compare the number of attacks that occur for each month.

From the graph, we can see that March has the most attacks that happen at 3678 and December has the least attacks that happen at 2673. Additionally, we can see that for January and February the average number of attacks is at around 3300 but it increases to average of 3500 for March to September, the average then drops to 2800 attacks from October to December.

This tells us that attacks are more likely to happen during the middle months rather than the staring or ending months. This could be due to solutions being found to counter the attacks for the starting and later months and as it reaches March new methods to attacks are found and thus the number of attacks increases.





We analysed the attacks the occur for each day. We decided that a count plot would be best for the comparison of each day.

From the graph above we can conclude that every day has about the same number of attacks. Thus, there we should keep vigilant each day. The only outlier being the 31st, however that is only due to 6 months having 31 days.

### Number of Cyber Attacks sorted by Attack Types Graph

A screenshot of a computer

Description automatically generatedWe analysed the type of attacks that happen to pinpoint the attacks that we should focus on. We have concluded that a count plot would be best for the comparison between each type of attack.

As seen from the graph above, the attack which is most frequent is DDoS at 13428. However, the other two types of attacks are not far behind with Malware being at 13307 and Intrusion being at 13265. This tells us that we need to emphasize on DDoS a little more than the two which can be due to less efficient security measures.

### Top Device/OS Targeted Graph

A screen shot of a graph

Description automatically generatedWe analysed the number of attacks for each type of device/operating system (OS). We have concluded that the best way to compare the data is using a count plot. The results of this comparison would help us know which devices/OSs we will need to focus on.

From the graph above, we can see that Windows OS is the most targeted by threat actors at 17953 attacks followed by Linux at 8840 then Macintosh at 5813 then iPod at 2656 and iPhone, Android and iPad all being at around 1600. This would tell us that we will have to further focus security measures on Windows, Linux and Macintosh as they might have more security vulnerabilities uncovered, and patches are not updated quick enough.

### Different Metrics against Attack Type Graph

A loop is used to iterate through the different columns to plot a count graph showing the different attack type based on the chosen metric.

A screenshot of a computer screen

Description automatically generatedWe analysed the different protocols and the types of attacks the attackers try on them. We have decided to make use the count plot to compare the data. We can use the results and findings of this graph to further understand if any protocols might need to be changed or strengthen in security.

Using ‘Attack Type’ as the base, different metrics are used to compare using a count plot visualization to understand how different features may affect the number of attacks of each type.

From the graph above, we can see that all the protocols have similar number of attacks being carried out on them and the number for each attack type is also similar. This shows that every protocol is equally vulnerable to every attack type. From this we can conclude that every protocol might need strengthening and that every protocol will need controls against each attack type.

A graph of data and information

Description automatically generated with medium confidenceWe have analysed the packet type of each attack. We have used a count plot to represent this data for comparison. The results will tell us the packet type which we have to put in controls against to lessen the attacks that occur.

Form the graph above, we can see that both packet types have similar counts to them. This tells us that controls need to be put in place for both packet types as every packet type is equally likely to be attacked.

A graph of data with numbers and a number

Description automatically generated with medium confidenceWe analysed the traffic type of the attacks. We decided on a count plot for this comparison. The results would tell us the traffic type that attackers are more likely to use in their attacks and we can put more control over the ones which are used more.

From the graph above, we can see that the three traffic types have similar counts. This tells us that attackers are likely to use any of the traffic types and that controls need to be put in place for all three and disregarding even one can led to the network being exploited through the less secure traffic type.

A graph of malware indicators

Description automatically generatedWe analysed the malware indicators for the attacks recorded. We used a count plot to compare the indicators. The results will show whether the attacks that happen are being detected by the system put in place.

Form the graph above, we can see that the count for detected and not detected are similar. This tells us that the system put in place to detect the attacks need improvement as half the attacks do not have an indicator of compromise detected which is a clear sign of a weak security system. It needs to be recalibrated to be able to detect majority of the attacks.

A graph with numbers and a bar

Description automatically generated with medium confidenceWe analysed the alert/warnings that the system gives. We have decided on using a count plot for the data. The results will tell us on how well the security system put in place is and whether it is providing ample warning if an attack is happening on the network.

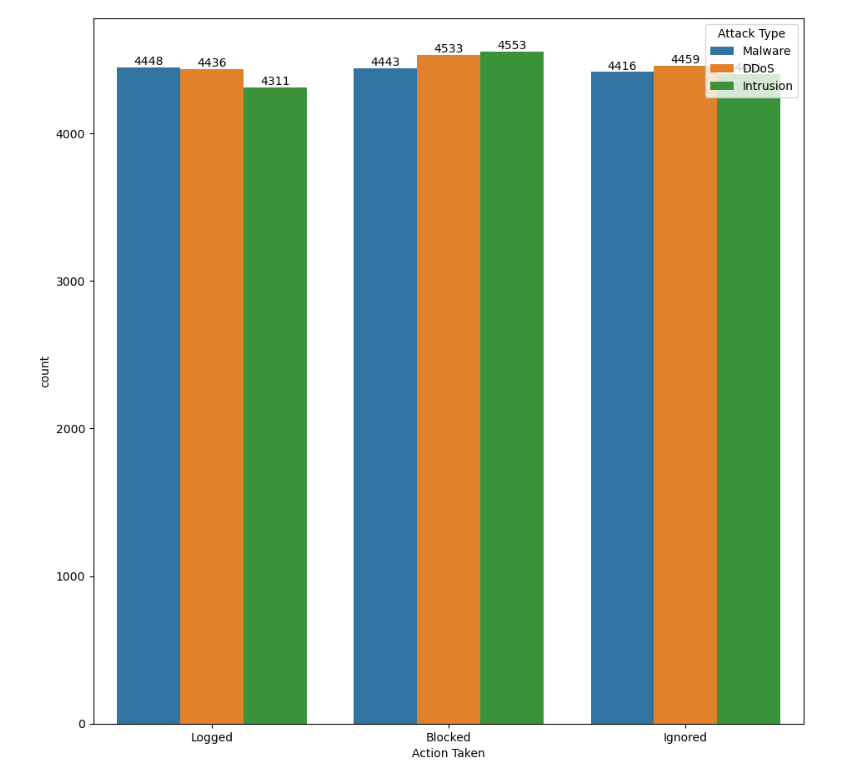
From the graph above, we can see that the counts for yes and no is similar. This indicates that the system is bad as it only warns or alerts the user of half of the attacks that happen to the network. This would need improvement as we will want at least 95 percent of the attacks that happen to the network to be notified to the relevant people.

A graph of multiple colored bars

Description automatically generated

We analysed the attack signature of the attack. We have used a count plot to represent the data for comparison. The results will tell us what attack signatures attacks like to use and thus we can focus more on the more frequent attack signature.

From the graph above, we see that both the attacks signatures A and B have similar counts to them. This tells us that the attackers do not have an attack signature which they are more likely to use. Efficient controls will be needed to be put in place to detect and distinguish each attack signatures to recognise what attack is taking place and the measures to deter them.

We analysed the Action Taken by the security system. We have chosen count plot to represent the data. The results will tell us how the security system is dealing with the attacks.

From the graph, we can see that the counts for each action taken against the attacks are about the same. However, about one-third of attacks are ignored which shows that the security system deemed those as safe, indicating the security system still has loopholes that needs to be fixed. This is to ensure that all attacks are logged or blocked.

A graph of multiple colored bars

Description automatically generated with medium confidenceWe analysed the severity level of the attacks. We have chosen count plot to represent the results. The results would give us information on the attacks that happen and would let us know if the attacks that happen can be ignored or needs immediate attention.

From the graph above, we can see that the severity of the attacks that happen have all similar counts. This would indicate that the security system and controls put in place are not adequate as the number of high and medium severity attacks are too high. It would be best if the high and medium severity level could be reduced as low as possible. Thus, controls and the security system need to be improved on.

A graph of different colored bars

Description automatically generated with medium confidenceWe analysed the number of attacks that happen in each network segment. We have chosen count plot to represent the results. The results will tell us which segment in the network segment is most venerable and needs more attention.

From the graph above, we can see that all three segments have similar counts. This tells us that all three network segments are being attacked by attackers and that there is a need to focus on all three segments as each network segment is equally likely to be attacked.

A graph of data with numbers and a few logs

Description automatically generated with medium confidenceWe analysed firewall logs to see if the attacks have been logged or not. We have chosen a count plot to represent the results. The results will tell us whether the firewall is recording down the attacks are occurring.

From the graph above, we can see that the firewall is logging at least half of the attacks. This indicates that the firewall is not catching onto almost half of the attacks and could pose as a major concern in security as half of the attacks are able to go past the firewall in place. It would be best if the firewall could record at least 95 percent of the attacks. Better firewall rules and practices needs to be implemented.

A graph of data with numbers and a number of data

Description automatically generated with medium confidenceWe analysed the IDS/IPS alerts that the system gives. We have decided on using a count plot for the data. The results will tell us on how well the security system put in place is and whether it is providing ample warning if an attack is happening on the network.

From the graph above, we can see that the counts for no alerts and alert data are similar. This indicates that the system is bad as it only warns or alerts the user of half of the attacks that happen to the network and about half the attack goes unnoticed. This would need improvement as we will want at least 95 percent of the attacks that happen to the network to be notified to the relevant people.

A graph of a number of bars

Description automatically generated with medium confidenceWe analysed log sources to see if the attacks have been logged by what. We have chosen a count plot to represent the results. The results will tell us whether logs are being recorded.

From the graph above, we can see that both the server and firewall logs are working. This tells us that the server and firewall logs are both being used and is a good sign as if there are any attacks that might have slipped through it could be captured by another log source.

A graph of different colored bars

Description automatically generatedWe analysed the browser that is being targeted. We have chosen a count plot to represent the results. The results will tell us which browser is targeted more by attackers and which we will need to pay more attention to.

From the above graph, we can see that Mozilla browsers is more prone to attacks than Opera browsers which shows that there might be more software vulnerabilities in Mozilla browser. This would mean that more controls will need to be put in place to harden the Mozilla browsers.

### Top Device/OS Targeted sorted by Browser Graph

A screenshot of a computer screen

Description automatically generatedWe analysed the number of attacks on targeted devices/OS and further categorized them for the browsers they use. We have chosen a count plot to represent the results. The results will show us if the browser being used for the device/OS could be the leading problem to why the device/OS is being targeted.

From the graph above, we can see that the Mozilla browser is being targeted more than on multiple devices/OSs than the Opera browser. This would indicate that the Mozilla browser is less secure than the Opera browser and that controls need to be put in place to harden the Mozilla browser and prevent attacks on Windows from occurring as often.

### Most Attacked Destination Ports Graph

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedWe analysed the ports that are being targeted by the attackers. We created a data frame to store the data and the counts of attacks on each port. We then used a bar plot to represent the results. The results of the data frame would tell us the vulnerable ports being exploited by the attackers, we can then use this information to disable the ports if they are not important or make it more secure if the port is needed.

From the graph above, we can see that the ports 34117, 7508, 30804, 37248 and 57508 all have attacks on them and have at least 5 attacks on each of them. This is a sign that the ports are more vulnerable and targeted and needs immediate attention. They can fix this by disabling the port if they are not needed or create firewall policies stopping any malicious data from being sent into or out of the ports.

## Feature Engineering

### Correlation Heatmap Plot

We take the columns with values that are numeric and plot a heatmap the show the correlation between features.

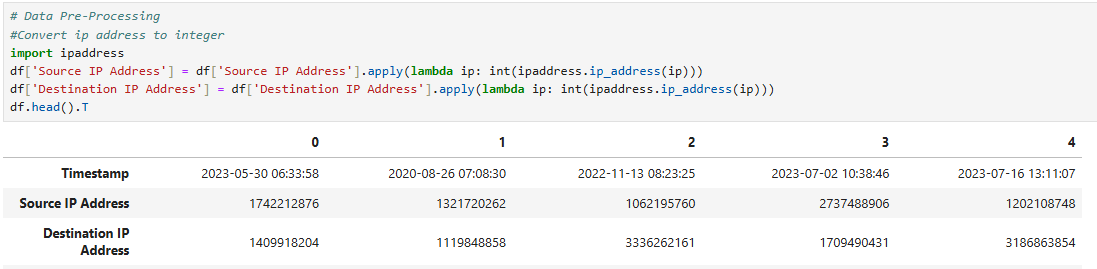
A screenshot of a computer

Description automatically generatedWe used a correlation heatmap to see the correlation between each feature. The correlation graph could help us to find out if the features are corelated. If the values were closer to 1, it would indicate that the attackers are using specific patterns when carrying out an attack.

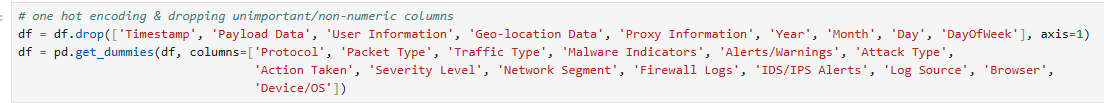
From the graph above, we can see that many of the features have very weak correlations as the many of the values is close to 0. This would suggest that the attacks could be caused by more complex factors.

### Data Preprocessing

(One hot encoding, Get\_Dummies etc.)



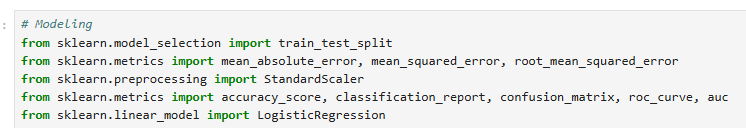
Change the data format of ‘IP Address’ to integer so that there are no dots between numbers to let the data type of IP address to be numeric for modelling.



We drop columns that are non-numeric and not important for our model so that the data frame will only be left with numeric and important features for modelling. This will ensure that only required features are used in our model.

## Predictive Modelling

### Importing necessary libraries



Import libraries needed to develop a machine learning model and model evaluation.

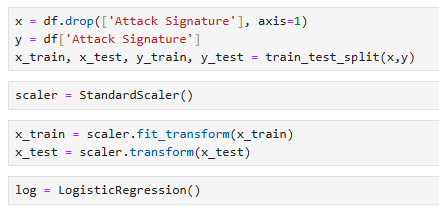
* Convert data type of ‘Attack Signature’ to integer ‘0’ or ‘1’



We import Label Encoder library to covert a categorical variable to a numerical variable. Changing the values to 0 (Known Pattern B) or 1 (Known Pattern A) will help in a better model and evaluation of the model.

### Logistic Regression

#### Create training and testing data splits

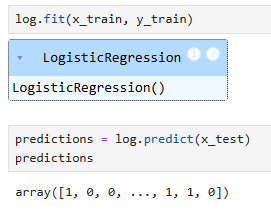


Y being the dependent variable is chosen as ‘Attack Signature’ while x being the independent variables will be all columns except ‘Attack Signature’

We chose ‘Attack Signature’ as we want to find out if other metrics can allow the model to predict known pattern type of attacks that will happen. This will allow cyber analyst or responder to be able to develop efficient and effective measures to detect and respond the recognized attacks.

x-train and x-test is scale using StandardScaler with fit\_transform to create a suitable training dataset.

#### Create Model Instance



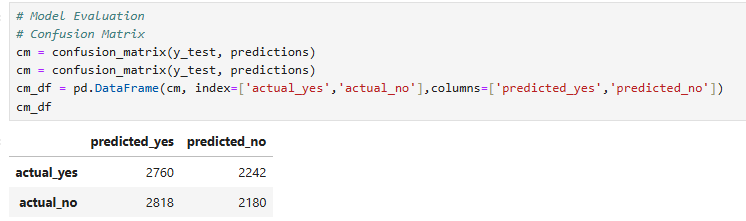
We used log.fit to create a logistic regression model instance that will train the dataset that was provided.

Then, log.predict is used to predict the results based on the x-test data which is the testing dataset created.

(1 being known pattern A and 0 being known pattern B)

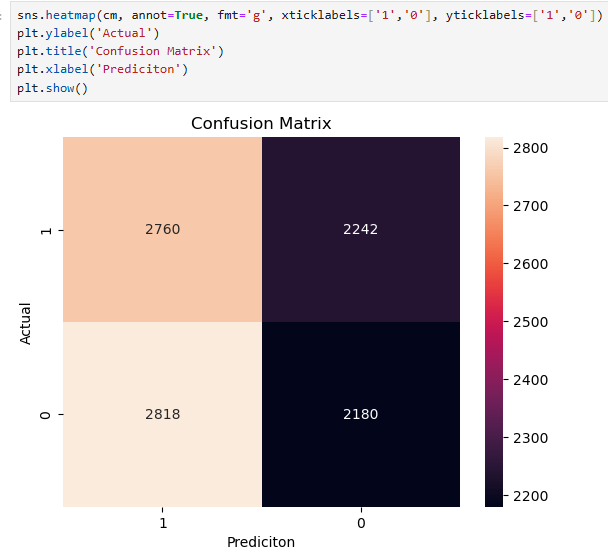
## Model Evaluation

* Creating a Confusion Matrix and Heatmap Visualization

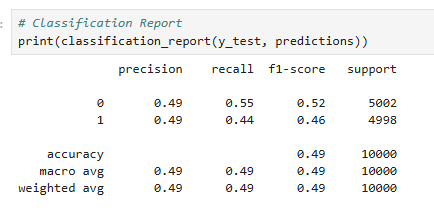


Using the confusion\_matrix library imported earlier, we created a confusion\_matrix dataframe to show the values of true and false positives and true and false negatives.

Using the results, we can infer that the number of true and false positives and number of true and false negatives are around the same which shows that the performance of the model is not that good.

Heatmap representation of the Confusion Matrix.

### Classification Report



Using the classification report library imported earlier, we can find out the precision, recall and f1-score of the logistic regression model. Through these values, we can know how accurate or precise the results are and can infer that the performance of the model is not good enough as the values are around 0.5 only.

# Findings

From the analysis of this dataset, the cyber analyst would have gained more insights on the devices/OSs which attackers like to target, furthermore providing information on features such as ports, browser, attack signature, traffic type, packet type, attack type, protocols etc. that the attackers prefer using on the network. This information will be crucial for the cyber analyst to come up with effective prevention and control measures as they will be more equipped with what to focus more on, what might be causing the vulnerability which the attackers are exploiting, what attack patterns they will need to find solutions to, what type of data and packets to set up secure policies for and much more.

The finding of this data set would be that for the companies’ cyber analyst, they will need to pentest the above-mentioned ports and other areas which have been attacked or bypass. They will want to focus on hardening the Mozilla browser, looking more into the Windows operating system which is the most used OS on the network, the detection system would also need to be improved to have at least a 95 percent detection rate when attacks happen, they would also want to look into putting in more secure firewall policies for the different protocols and traffic types so that more attacks can be blocked.

Based on the model created and tested and analysis of the performance, we can tell that it is not sufficient to predict the attack signature of cyber-attacks as almost half the results is false positives and negatives. We can infer through the research and modelling that there are other variables affecting what attack pattern will be used for attacks as the features used only affect a part of what attack pattern is used.

# Conclusion

In conclusion, from this report we have found out that analyzing cyber-attacks as a data set and creating graphs to represent different results of the data can help cyber analyst find out and analyse vulnerable places that need improvement. It can also allow cyber analyst and responder to figure out efficient and effective prevention and controls measures to combat against cyber-attacks to reduce the occurrence and impact of cyber-attacks.

To better counter cyber-attacks, more secure measures and controls to respond to attacks needs to be implemented and more in-depth data needs to be collected to develop and train a model with good performance and result.

# Appendix

The csv file of the dataset used will be attached for submission.

# References

1. *Cyber security attacks*. (2025, January 21). Kaggle. <https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks/data>