Cybersecurity Suspicious Web Threat Interactions

Created by: Divya Krishnakumar

Type of project: Internship Project

Tools Used: Python, SQL

Submitted to: Unified Mentor Pvt. Ltd

Date: 24/07/2025

Executive Summary

The aim of the project is to detect anomaly or unusual behaviours in web traffic. The methodology followed in this project consists of the following steps – Data collection, Data wrangling, Data analysis, Data visualisation, Feature engineering, Model building, Model Evaluation.

Data is collected and read to a data frame and subjected to descriptive statistics to understand the data. The dataset provides web traffic details of suspicious activity which can be used by data scientists to detect threats. Each row of data set represents a stream of traffic to a web server. The columns were bytes_in, bytes_out, creation_time, end_time, src_ip, src_ip_country_code, protocol, response.code, dst_port, dst_ip, rule_names, observation_name, source.meta, source.name, time, detection_types. As a part of data wrangling the columns are renamed to relevant names. The columns with irrelevant data types are converted to relevant data types. Then the dataset is checked for null values and duplicates.

Data is analysed with SQL. Data is analysed for further insights with data visualisations. The transactions of unusual behaviour is found out. The columns Session_Duration and Average_Packet_Size are created through feature engineering for prepare for model development. A model is developed with IsolationForest algorithm and fitted and predicted with X. The results are visualised with scatter plot. A classification report is printed to evaluate the model. A confusion matrix is plotted to evaluate further. The results are stored in a .csv file.

Introduction

The project aims to identify patterns of unusual behaviour in the web traffic. This involves analysis of dataset to detect anomaly in the web traffic.

The key questions are:

- Which streams of web traffic are anomalous?
- From which IP's suspicious activities originated?
- Which are the 10 source IP's from where maximum times suspicious activities are found?
- What is the total number of bytes in for suspicious activity?
- What is the total number of bytes out for suspicious activity?
- What are the unique Src_IP_Country_code from where suspicious activities are found?
- What is the proportion of suspicious activities from different countries?
- What are the minimum bytes in and bytes out for suspicious traffic?
- What is the trend of suspicious traffic at different times?

Methodology

Data is collected and loaded to a data frame. The methodology used in this project has following steps: Data collection, Data wrangling, Data analysis, Data visualisation, Feature engineering, Model building, Model Evaluation.

The data collected is cleaned, analysed, visualized and subjected to machine learning algorithm, IsolationForest.

Data Collection: Data is collected and loaded to a dataframe. The data set consists of columns such as bytes_in bytes_out, creation_time, end_time, src_ip, src_ip_country_code, protocol, response.code, dst_code, dst_ip, rule_names, observation_name, source.meta, source.name, time, detection_types.

Data Wrangling

During this step the columns are renamed to relevant column names. The data types of columns are changed to relevant data types. The data set is checked for missing values and duplicates and found no missing values and duplicates. The data frame thus obtained is as follows:

	Bytes_in	Bytes_out	Creation_Time	End_Time	Src_IP	Src_IP_Country_code	Protocol	Response_Code	Dst_Port	Dst_IP	...	Source_Name	Detected_Event_Time	Detection_Types	SCT	Detected_Time	Suspicious_End_Time	Duration_of_suspicious_activity	Session_Duration	Average_Packet_Size	anomaly
0	5602	12990	2024-04-25 23:00:00+00:00	2024-04-25 23:10:00+00:00	147.161.161.82	AE	HTTPS	200	443	10.138.69.97	...	prod_webserver	2024-04-25 23:00:00+00:00	waf_rule	23:00:00	23:00:00	23:10:00	0 days 00:10:00	600.0	30.986667	0
1	30912	18186	2024-04-25 23:00:00+00:00	2024-04-25 23:10:00+00:00	165.225.33.6	US	HTTPS	200	443	10.138.69.97	...	prod_webserver	2024-04-25 23:00:00+00:00	waf_rule	23:00:00	23:00:00	23:10:00	0 days 00:10:00	600.0	81.830000	0
2	28506	13468	2024-04-25 23:00:00+00:00	2024-04-25 23:10:00+00:00	165.225.212.255	CA	HTTPS	200	443	10.138.69.97	...	prod_webserver	2024-04-25 23:00:00+00:00	waf_rule	23:00:00	23:00:00	23:10:00	0 days 00:10:00	600.0	69.956667	0
3	30546	14278	2024-04-25 23:00:00+00:00	2024-04-25 23:10:00+00:00	136.226.64.114	US	HTTPS	200	443	10.138.69.97	...	prod_webserver	2024-04-25 23:00:00+00:00	waf_rule	23:00:00	23:00:00	23:10:00	0 days 00:10:00	600.0	74.706667	0
4	6526	13892	2024-04-25 23:00:00+00:00	2024-04-25 23:10:00+00:00	165.225.240.79	NL	HTTPS	200	443	10.138.69.97	...	prod_webserver	2024-04-25 23:00:00+00:00	waf_rule	23:00:00	23:00:00	23:10:00	0 days 00:10:00	600.0	34.030000	0

5 rows × 23 columns

Data Analysis with SQL results

Anomalous rows for Bytes_in and Bytes_out are displayed by finding IQR and using IQR rule. The dataset is then subjected to SQL analysis. The following are insights from SQL analysis.

Unique IP of origin of suspicious activity

```
[13]: %sql select distinct(Src_IP) from CloudWatch_Traffic_Web_Attack;
 * sqlite:///CyberSecurityDB
Done.
```

[13]:

Src_IP
147.161.161.82
165.225.33.6
165.225.212.255
136.226.64.114
165.225.240.79
136.226.77.103
165.225.26.101
155.91.45.242
165.225.209.4
147.161.131.1
136.226.67.101
94.188.248.74
165.225.213.7
165.225.8.79
136.226.80.97
192.241.230.19
198.235.24.81

Data Analysis with SQL results

10 source IP from where maximum times suspicious activities are found

```
%sql select Src_IP, count(Src_IP) as "Src_IP_Count" from CloudWatch_Traffic_Web_Attack group by Src_IP order by Src_IP_Count de
```

```
 * sqlite:///CyberSecurityDB
Done.
```

Src_IP	Src_IP_Count
165.225.209.4	29
165.225.26.101	28
155.91.45.242	28
136.226.67.101	28
147.161.131.1	21
165.225.240.79	18
136.226.77.103	17
147.161.161.82	16
165.225.212.255	15
94.188.248.74	14

Data Analysis with SQL results

Total of bytes in for suspicious activity

```
%sql select sum(Bytes_in) as 'Total Bytes received by the server for suspicious acitivities' from CloudWatch_Traffic_Web_Attack;
```

```
 * sqlite:///CyberSecurityDB
Done.
```

Total bytes out for suspicious acitivity

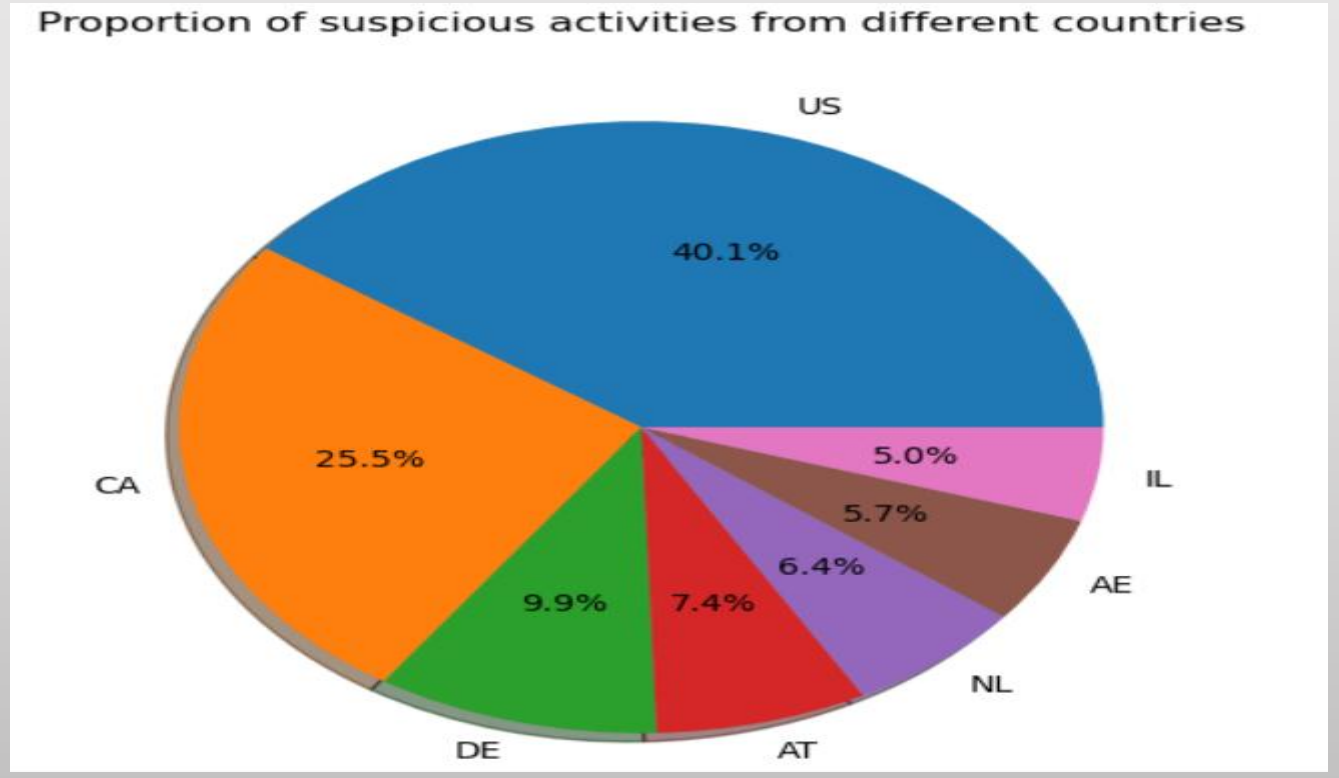
```
%sql select sum(Bytes_out) as 'Total Bytes sent from the server for suspicious activity' from CloudWatch_Traffic_Web_Attack
```

```
 * sqlite:///CyberSecurityDB
Done.
```

Total Bytes sent from the server for suspicious activity
23844310

Exploratory Data Analysis and Visualisation

Proportion of suspicious activities from different countries



Exploratory Data Analysis and Visualisation

What are the minimum bytes in and bytes out for suspicious traffic?

	Bytes_in	Bytes_out	Creation_Time	End_Time	Src_IP	Src_IP_Country_code	Protocol	Response_Code	Dst_Port	Dst_IP	Rule_Names
213	80	44	2024-04-26 08:50:00+00:00	2024-04-26 09:00:00+00:00	65.49.1.72	US	HTTPS	200	443	10.138.69.97	Suspicious Web Traffic
276	40	264	2024-04-26 09:50:00+00:00	2024-04-26 10:00:00+00:00	192.241.205.18	US	HTTPS	200	443	10.138.69.97	Suspicious Web Traffic

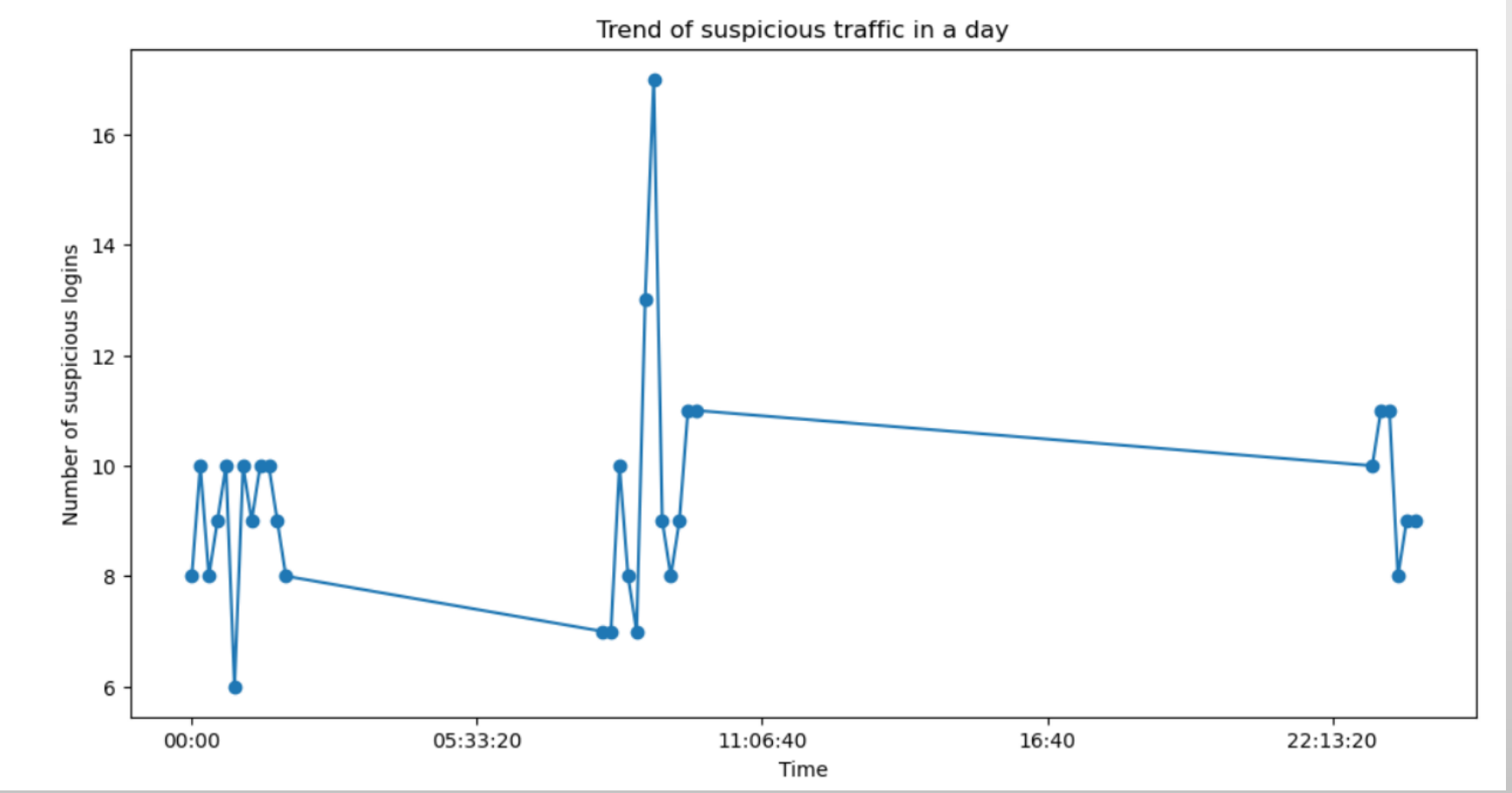
Exploratory Data Analysis and Visualisation

Trend of suspicious traffic

```
SCT
00:00:00     8
00:10:00    10
00:20:00     8
00:30:00     9
00:40:00    10
00:50:00     6
01:00:00    10
01:10:00     9
01:20:00    10
01:30:00    10
01:40:00     9
01:50:00     8
08:00:00     7
08:10:00     7
08:20:00    10
08:30:00     8
08:40:00     7
08:50:00    13
09:00:00    17
09:10:00     9
09:20:00     8
09:30:00     9
09:40:00    11
09:50:00    11
23:00:00    10
23:10:00    11
23:20:00    11
23:30:00     8
23:40:00     9
23:50:00     9
dtype: int64
```

Exploratory Data Analysis and Visualisation

Trend of suspicious traffic

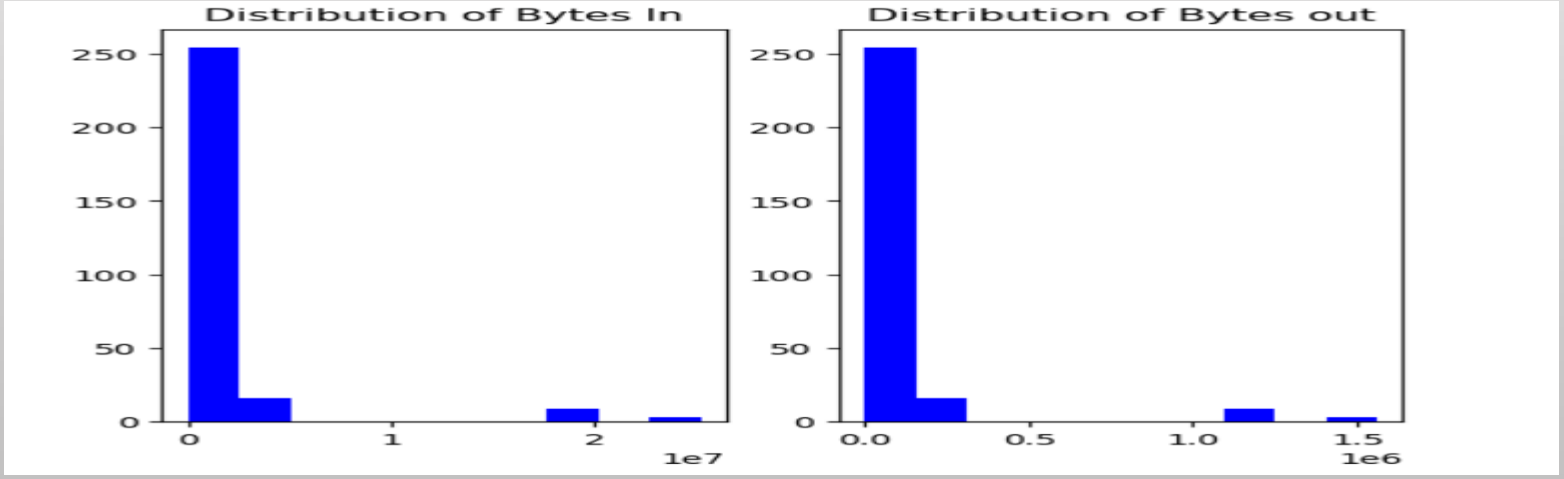


Exploratory Data Analysis and Visualisation

Duration of suspicious activity

```
0   0 days 00:10:00
1   0 days 00:10:00
2   0 days 00:10:00
3   0 days 00:10:00
4   0 days 00:10:00
Name: Duration_of_suspicious_activity, dtype: timedelta64[ns]
```

Distribution of bytes in



Feature Engineering

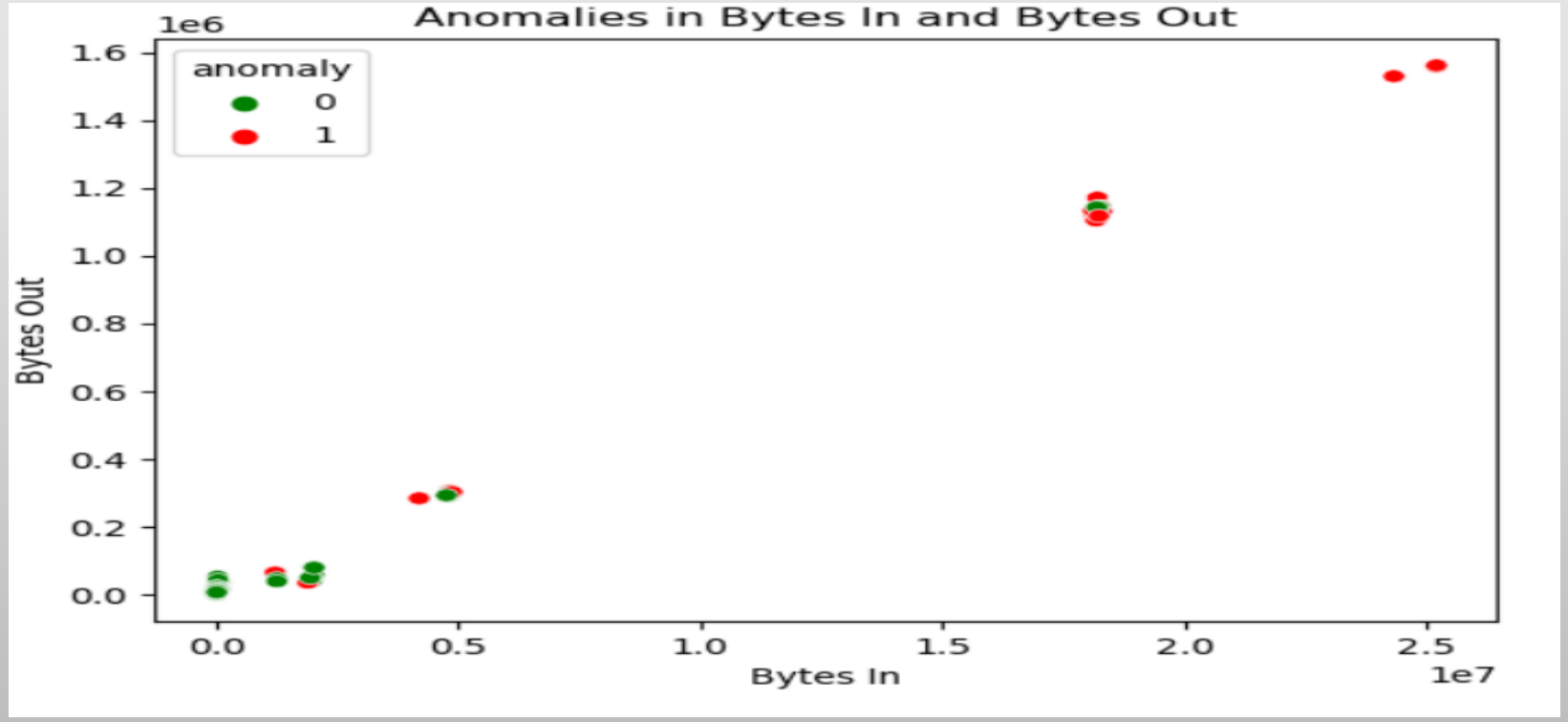
Session duration and average packet size are the newly created features. Session duration is the difference between session end time and session creation time. Average packet size is the sum of bytes_in and bytes_out divided by session duration.

	Session_Duration	Average_Packet_Size
0	600.0	30.986667
1	600.0	81.830000
2	600.0	69.956667
3	600.0	74.706667
4	600.0	34.030000

Model Building

A model is developed with machine learning algorithm IsolationForest. The model is initialised, fitted and predicted with train data set to predict unseen data.

The results of model development is visualised with a scatter plot. The scatter plot is as follows:



Model Evaluation

The model is evaluated with displaying a classification report.

The accuracy score of the model is 91%. The classification report is as follows:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	267
1	0.13	0.13	0.13	15
accuracy			0.91	282
macro avg	0.54	0.54	0.54	282
weighted avg	0.91	0.91	0.91	282

Findings and Implications

- Total bytes in for suspicious activity is 338228034.
- Total bytes out for suspicious activity is 23844310.
- The maximum bytes in for suspicious traffic is 25207794.
- The minimum bytes out for suspicious traffic is 44.
- Duration of suspicious activity is 10 minutes.
- For 267 non anomalous activities 15 suspicious activities could occur.
- The suspicious activities are during 9:00 hours to 10:00 hours and 23:00 hours to 23:50 hours.
- The maximum times the suspicious activities are from the unique id 165.225.209.4 and number of times is 29.
- The country codes of maximum suspicious activities are from US, CA.

Conclusion

The anomaly is detected in the web traffic by inspecting the number of bytes transferred. If the bytes transferred is unusually high it indicates suspicious activity. The web traffic of suspicious activity indicates that the attacks could be bot attacks. There are about 37 to 40 suspicious attempts. These suspicious attempts from same IP address have huge bytes transferred indicating bot attacks. The maximum of bot attacks are from the country code US.

From the predictive analysis we can conclude that for every 280 activities there could be 15 suspicious activities.

Future implications and suggestions:

- The suspicious attack can be identified and prevented by using defense bots which work round the clock, scanning systems and networks. It will raise alerts or take automated actions if any suspicious activity is detected.
- AI powered bots can be used to identify and prevent anomalous threats.