MLresearchLab.com
Lab 1: Random Forest

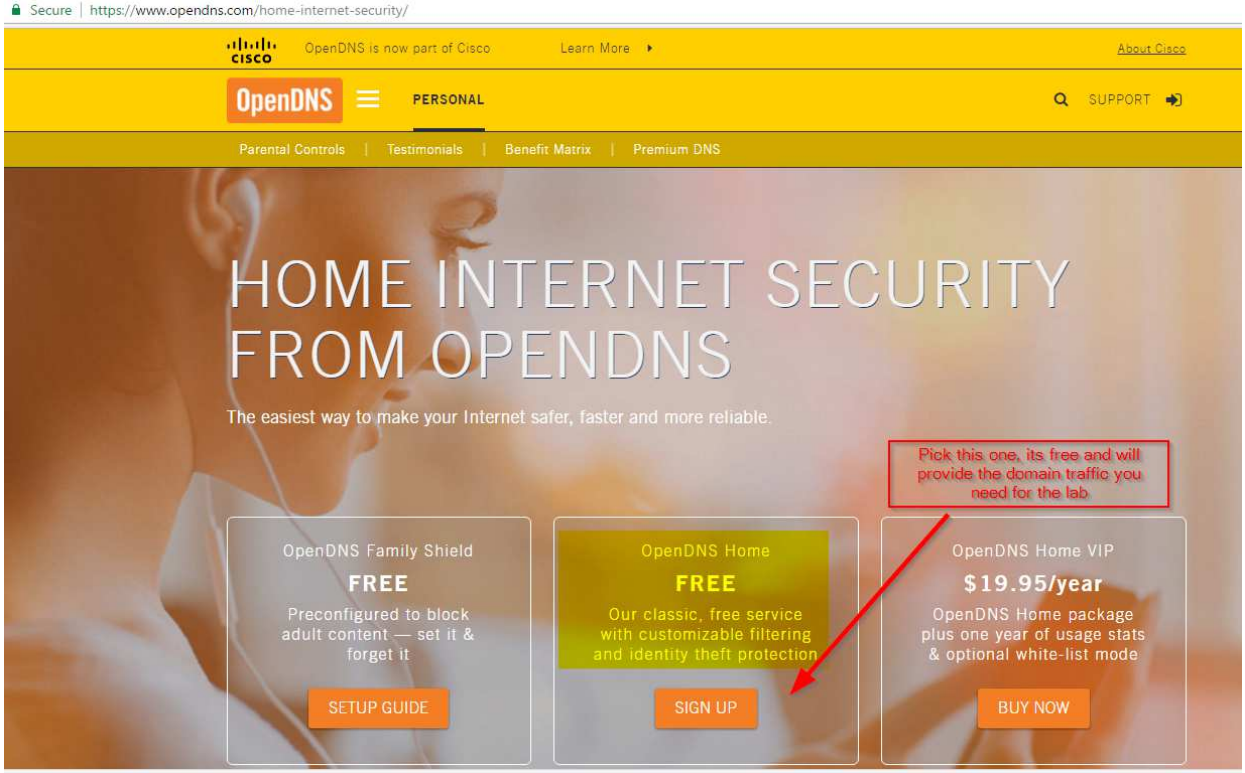
Resources Requred:
Linux Host – Tested on Ubuntu
Windows / MAC Host – For OpenDNS Client

OpenDNS Account – http://www.opendns.com
Jupyter Notebook - http://jupyter.org/

Virtual Machine provided if you want to have everything pre-built with data on GitHub

Goal: Utilize machine learning (ML) random forest classifier to analyze OpenDNS (www.opendns.com) logs of user domain requests to determine if a user has been visiting work related sites.

1. Create an OpenDNS account by visiting www.opendns.com and click personal to reach the web page below.



2. Click the orange “Sign Up” button and follow the instructions to create an accont.

3. Once you have the account created, download the client

4. Install the client on your test machine or Virtual Machine (VM).

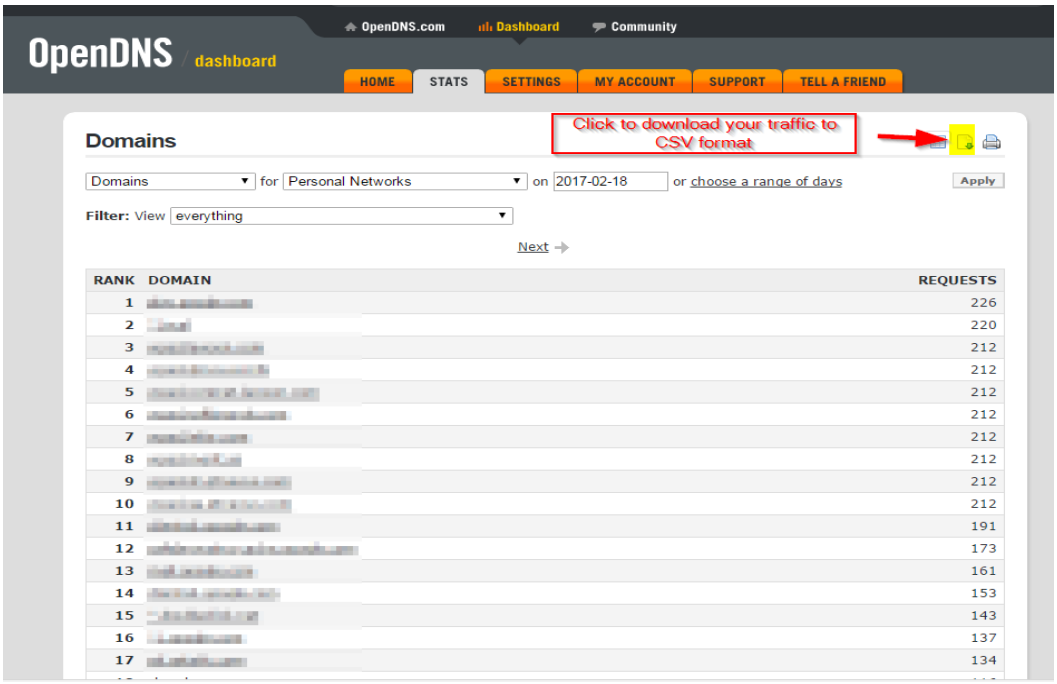
5. Once the client is installed it will ask for the same credentials you signed up under, so enter those in and you should see the screen shot below.

6. Make sure the client under “Using OpenDNS” says “Yes” which means now your DNS traffic is being forwarded to the OpenDNS servers.

7. Browse random websites to generate traffic on the host that has the OpenDNS client, its best to visit at least 50 websites or more

8. Once you have visited 50 or more websites log back into OpenDNS you should see traffic in your dashboard. (If there is no traffic it might take some time to populate so try back a few hours later or a day later)

9. If you have traffic in your dashboard you are now ready to begin to download the traffic as data for the ML lab. Click “Stats” the orange tab in the center of the dashboard then “Domains” on the left pane and you should see something similar to the screenshot below.



10. To download the DNS traffic data you will look for a button of a file download in the top right hand corner of the “Domains” Dashboard. (See screenshot from previous step)

11. Upon Success you should end up with a comma separated value (CSV) file on your computer, go ahead and open it with your favorite editor

	A	B	C	D	E	F	G	H	I	J	
1	Rank	Domain	Total	Blacklisted	Blocked by Category	Blocked as Botnet	Blocked as Malware	Blocked as Phishing	Resolved by SmartCache	Academic Fraud	A
2	1		226	0	0	0	0	0	0	0	
3	2		220	0	0	0	0	0	0	0	
4	3		212	0	0	0	0	0	0	0	
5	4		212	0	0	0	0	0	0	0	
6	5		212	0	0	0	0	0	0	0	
7	6		212	0	0	0	0	0	0	0	
8	7		212	0	0	0	0	0	0	0	
9	8		212	0	0	0	0	0	0	0	
10	9		212	0	0	0	0	0	0	0	
11	10		212	0	0	0	0	0	0	0	
12	11		191	0	0	0	0	0	0	0	
13	12		173	0	0	0	0	0	0	0	
14	13		161	0	0	0	0	0	0	0	
15	14		153	0	0	0	0	0	0	0	
16	15		143	0	0	0	0	0	0	0	
17	16		137	0	0	0	0	0	0	0	
18	17		134	0	0	0	0	0	0	0	
19	18		116	0	0	0	0	0	0	0	
20	19		116	0	0	0	0	0	0	0	
21	20		114	0	0	0	0	0	0	0	
22	21		111	0	0	0	0	0	0	0	
23	22		106	0	0	0	0	0	0	0	
24	23		102	0	0	0	0	0	0	0	
25	24		93	0	0	0	0	0	0	0	
26	25		91	0	0	0	0	0	0	0	
27	26		88	0	0	0	0	0	0	0	
28	27		82	0	0	0	0	0	0	0	
29	28		79	0	0	0	0	0	0	0	
30	29		75	0	0	0	0	0	0	0	
31	30		74	0	0	0	0	0	0	0	
32	31		68	0	0	0	0	0	0	0	
33	32		68	0	0	0	0	0	0	0	
34	33		68	0	0	0	0	0	0	0	

12. In your editor and remove the following columns
“Rank”, “Total”, “Blacklisted”, “Blocked by Category”, “Blocked as Botnet”, “Blocked as Malware”, “Blocked as Phishing”, and “Resolved by SmartCache” as shown below

	A	B	C	D	E	F	G	H	I
1	Domain	Academic Fraud	Adult Themes	Adware	Alcohol	Anime/Manga/Webcomic	Auctions	Automotive	Blogs
2	[redacted]	0	0	0	0	0	0	0	0
3	[redacted]	0	0	0	0	0	0	0	0
4	[redacted]	0	0	0	0	0	0	0	0
5	[redacted]	0	0	0	0	0	0	0	0
6	[redacted]	0	0	0	0	0	0	0	0
7	[redacted]	0	0	0	0	0	0	0	0
8	[redacted]	0	0	0	0	0	0	0	0
9	[redacted]	0	0	0	0	0	0	0	0
10	[redacted]	0	0	0	0	0	0	0	0
11	[redacted]	0	0	0	0	0	0	0	0
12	[redacted]	0	0	0	0	0	0	0	0
13	[redacted]	0	0	0	0	0	0	0	0
14	[redacted]	0	0	0	0	0	0	0	0
15	[redacted]	0	0	0	0	0	0	0	0
16	[redacted]	0	0	0	0	0	0	0	0
17	[redacted]	0	0	0	0	0	0	0	0
18	[redacted]	0	0	0	0	0	0	0	0
19	[redacted]	0	0	0	0	0	0	0	0
20	[redacted]	0	0	0	0	0	0	0	0
21	[redacted]	0	0	0	0	0	0	0	0
22	[redacted]	0	0	0	0	0	0	0	0
23	[redacted]	0	0	0	0	0	0	0	0
24	[redacted]	0	0	0	0	0	0	0	0
25	[redacted]	0	0	0	0	0	0	0	0
26	[redacted]	0	0	0	0	0	0	0	0
27	[redacted]	0	0	0	0	0	0	0	0
28	[redacted]	0	0	0	0	0	0	0	0
29	[redacted]	0	0	0	0	0	0	0	0
30	[redacted]	0	0	0	0	0	0	0	0
31	[redacted]	0	0	0	0	0	0	0	0
32	[redacted]	0	0	0	0	0	0	0	0
33	[redacted]	0	0	0	0	0	0	0	0
34	[redacted]	0	0	0	0	0	0	0	0
35	[redacted]	0	0	0	0	0	0	0	0
36	[redacted]	0	0	0	0	0	0	0	0

13. Remember the purpose of this lab is not to determine if the domains your user visited is malicious or not (it is beyond the scope of this lab). Rather that if they visit sites that are not Work Related, so you should end up with the “Domains” and “Categories of websites classifications”(e.g. Academic Fund, Adult Themes, Adware) in your CSV file.

14. To prepare your data, remove rows that have a “0” in all “Categories of websites classifications”. These rows do not help the ML and should be removed because they do not have a “1” in any of fields.

15. Now the data inside the CSV needs to be trained for the ML engine so you need to create a new column called “WorkRelated” at the end of the CSV file. So you should now have”Domains”, “Categories of websites classifications”, and “WorkRelated” as columns inside the CSV file.

16. You must now research the domain that is listed in the “Domain” column and if you determine that domain is work related put a “1” in the “WorkRelated” column else wise put a “0” in the “WorkRelated” column.

17. You should now have a “1” or “0” in the “WorkRelated” column for every row. Once confirmed, remove the “Domain” column because the “Domain” does not help the ML make a decision, the “Categories of website classifications” and “WorkRelated” columns help the ML make the decisions.

18. Your CSV file should only have the “Categories of website classifications” and “WorkRelated” columns and each row should have a “1” in one of the “Categories of website classifications” and a “1” or “0” in “WorkRelated” column.

19. Now the data is ready to be ingested into the ML.

20. Install Jupyter Notebook following the instructions on the website

21. Once Jupyter Notebook is installed and running you can go ahead and copy the CSV to the folder where Jupyter Notebook.

22. Create a new Jupyter Notebook and copy the code from the text file to the Jupyter Notebook.

23. Run the code inside Jupyter Notebook