Lab 1:

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

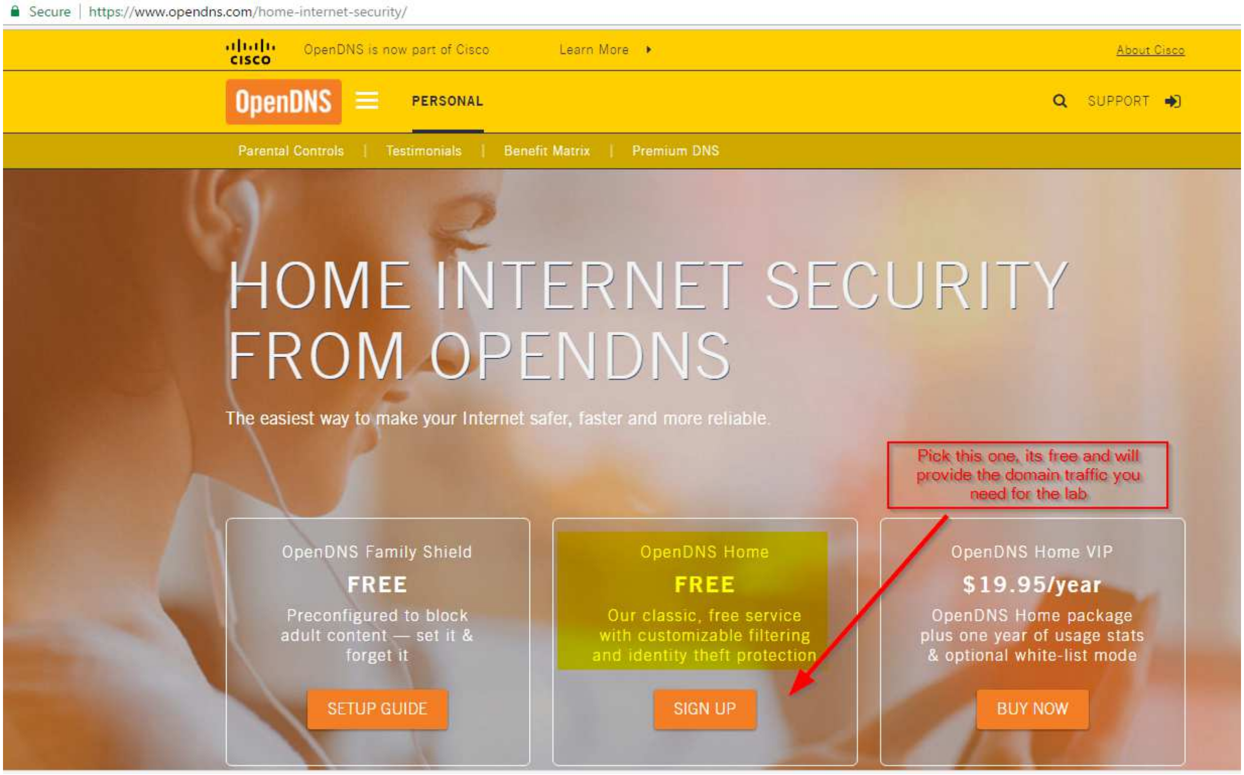
OpenDNS Account – http://www.opendns.com
Jupyter Notebook - http://jupyter.org/

A Vmware Virtual Machine is available for download on http://www.MLresearchLab.com. If you want to have everything pre-built with the sample data

- If you do not have VMware, VMware Player is free or you can convert it to Virtual Box which is also free

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.


Secure | https://www.opendns.com/home-internet-security/
OpenDNS is now part of Cisco
Learn More
About Cisco
OpenDNS
PERSONAL
SUPPORT
Parental Controls | Testimonials | Benefit Matrix | Premium DNS
SETUP GUIDE
SIGN UP
BUY NOW

HOME INTERNET SECURITY FROM OPENDNS

The easiest way to make your Internet safer, faster and more reliable.

Pick this one, its free and will provide the domain traffic you need for the lab

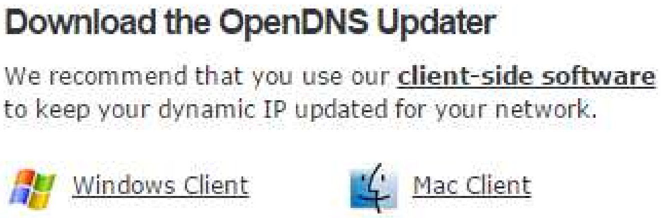
OpenDNS Family Shield
FREE
Preconfigured to block adult content — set it & forget it

OpenDNS Home
FREE
Our classic, free service with customizable filtering and identity theft protection

OpenDNS Home VIP
$19.95/year
OpenDNS Home package plus one year of usage stats & optional white-list mode

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



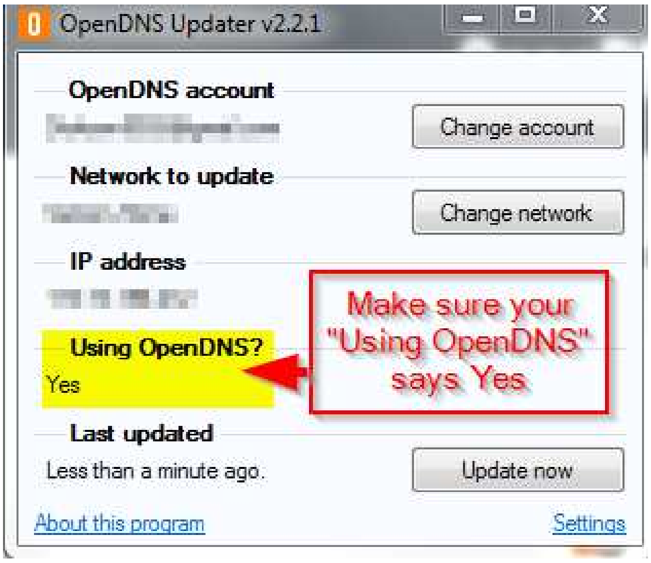
Download the OpenDNS Updater

We recommend that you use our **client-side software** to keep your dynamic IP updated for your network.

Windows Client Mac 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.



OpenDNS Updater v2.2.1

OpenDNS account
[redacted]

Network to update
[redacted]

IP address
[redacted]

Using OpenDNS?
Yes

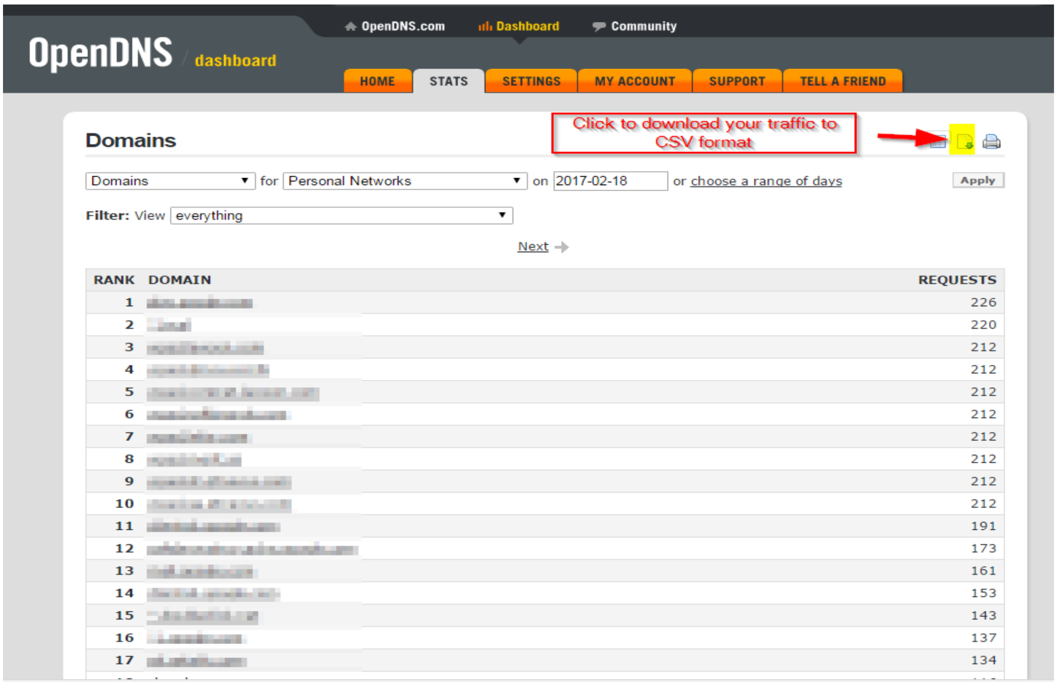
Last updated
Less than a minute ago.

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.



Domains

Click to download your traffic to CSV format

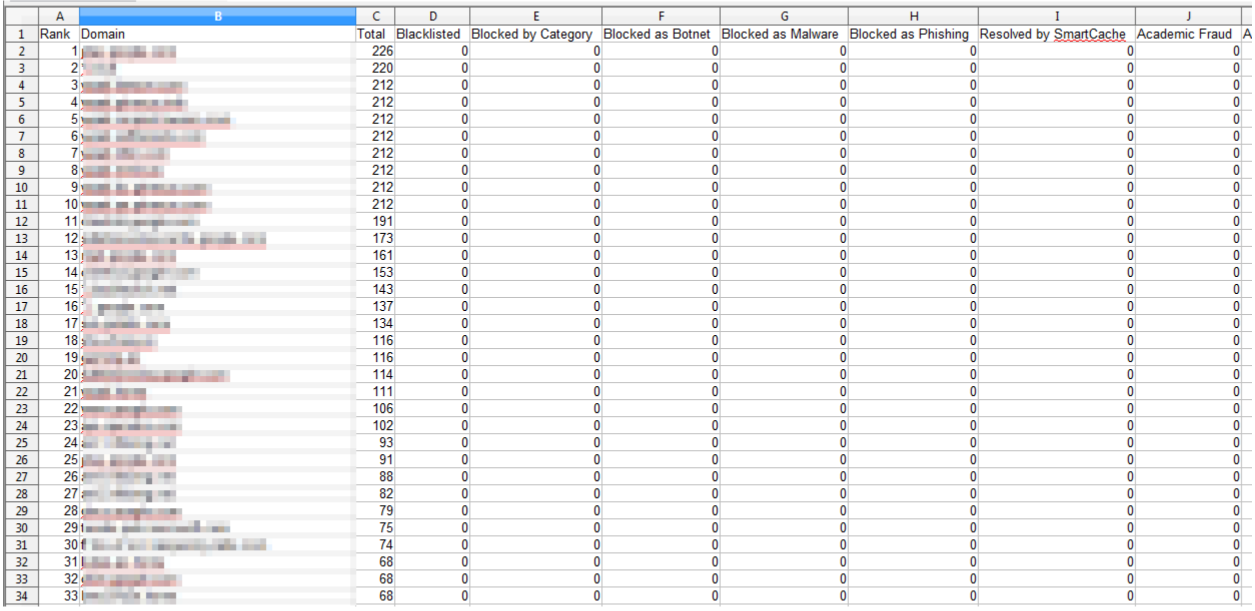
Domains for Personal Networks on 2017-02-18 or choose a range of days

Filter: View everything

| RANK | DOMAIN | REQUESTS |
|---|---|---|
| 1 | [redacted] | 226 |
| 2 | [redacted] | 220 |
| 3 | [redacted] | 212 |
| 4 | [redacted] | 212 |
| 5 | [redacted] | 212 |
| 6 | [redacted] | 212 |
| 7 | [redacted] | 212 |
| 8 | [redacted] | 212 |
| 9 | [redacted] | 212 |
| 10 | [redacted] | 212 |
| 11 | [redacted] | 191 |
| 12 | [redacted] | 173 |
| 13 | [redacted] | 161 |
| 14 | [redacted] | 153 |
| 15 | [redacted] | 143 |
| 16 | [redacted] | 137 |
| 17 | [redacted] | 134 |

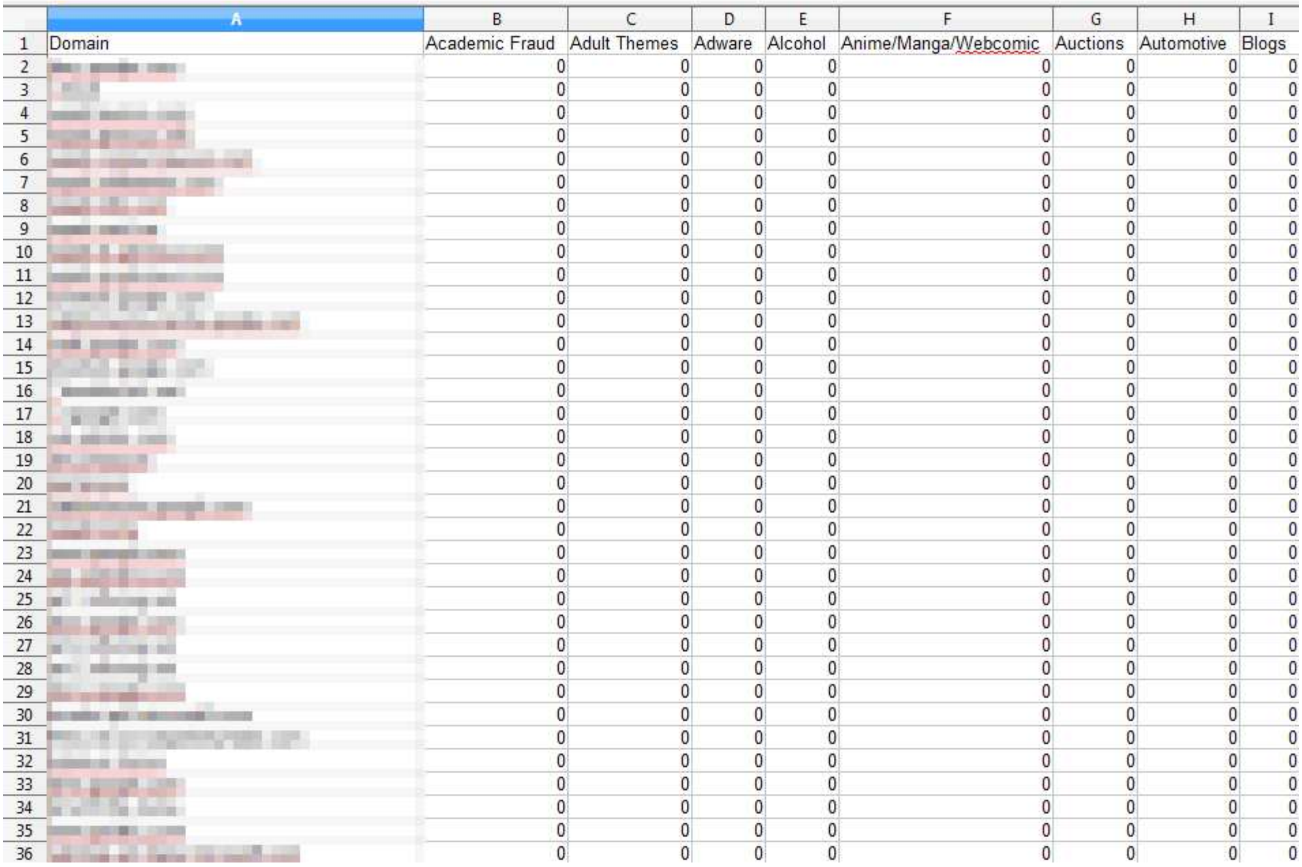
10. To download the DNS traffic data you will look for a button of a file download in the top right hand corner of the "Domains" Dashboard. (See screenshot from previous step)

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 |
| 2 | 1 | [redacted] | 226 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | [redacted] | 220 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 5 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 6 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 7 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 8 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 9 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 10 | [redacted] | 212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 11 | [redacted] | 191 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 12 | [redacted] | 173 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 13 | [redacted] | 161 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 14 | [redacted] | 153 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 15 | [redacted] | 143 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 16 | [redacted] | 137 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 17 | [redacted] | 134 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 18 | [redacted] | 116 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 19 | [redacted] | 116 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 20 | [redacted] | 114 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 21 | [redacted] | 111 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | 22 | [redacted] | 106 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | 23 | [redacted] | 102 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 24 | [redacted] | 93 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | 25 | [redacted] | 91 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 26 | [redacted] | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | 27 | [redacted] | 82 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | 28 | [redacted] | 79 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | 29 | [redacted] | 75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | 30 | [redacted] | 74 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | 31 | [redacted] | 68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | 32 | [redacted] | 68 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 34 | 33 | [redacted] | 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.

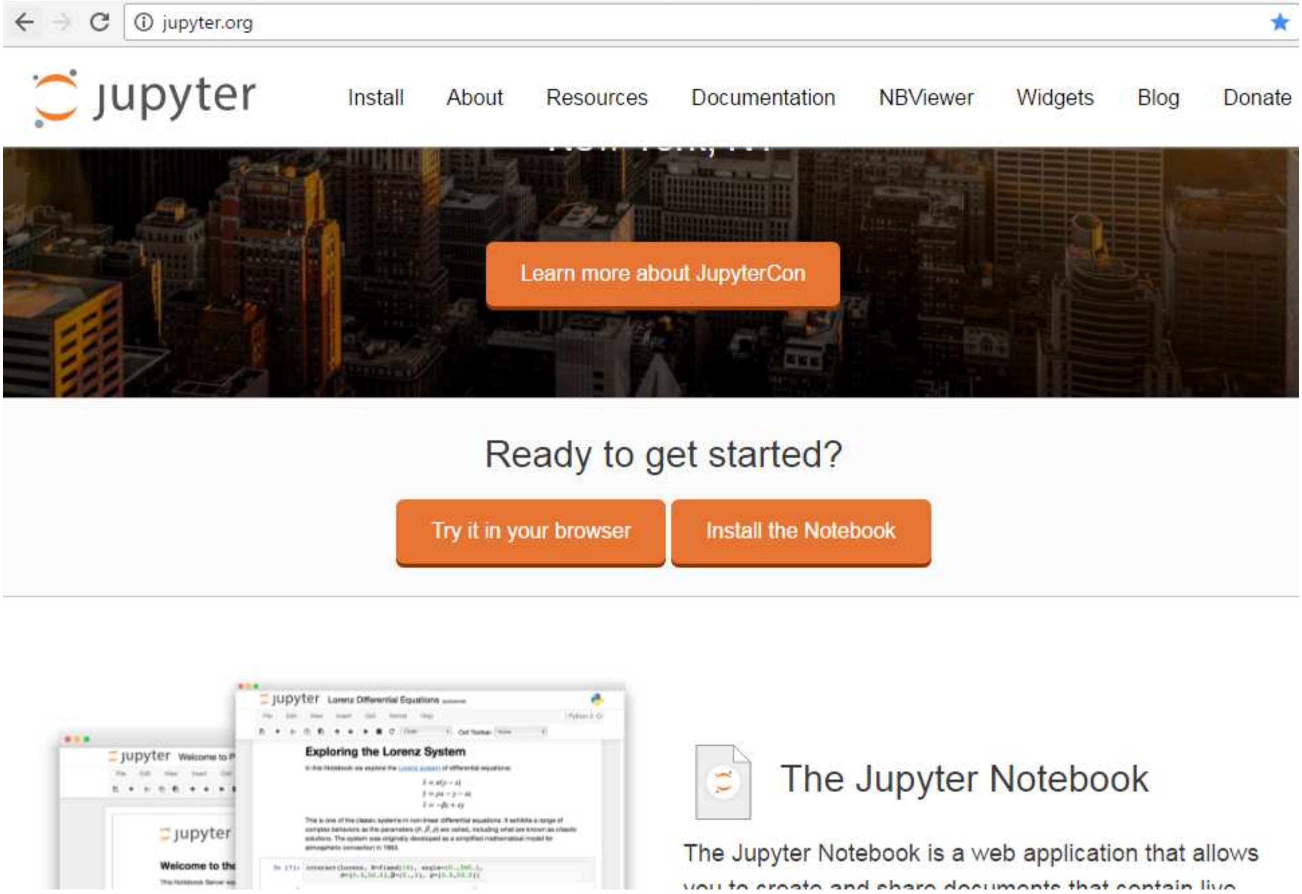
17. You should now have a “1” or “0” in the “WorkRelated” column for every row. Once confirmed, remove the “Domain” column because the “Domain” does not help the ML make a decision, the “Categories of website classifications” and “WorkRelated” columns help the ML make the decisions.

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

classifications” and a “1” or “0” in “WorkRelated” column.

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

20. Visit the Jupyter Notebook website: [http://jupyter.org/](http://jupyter.org/)



Ready to get started?

The Jupyter Notebook

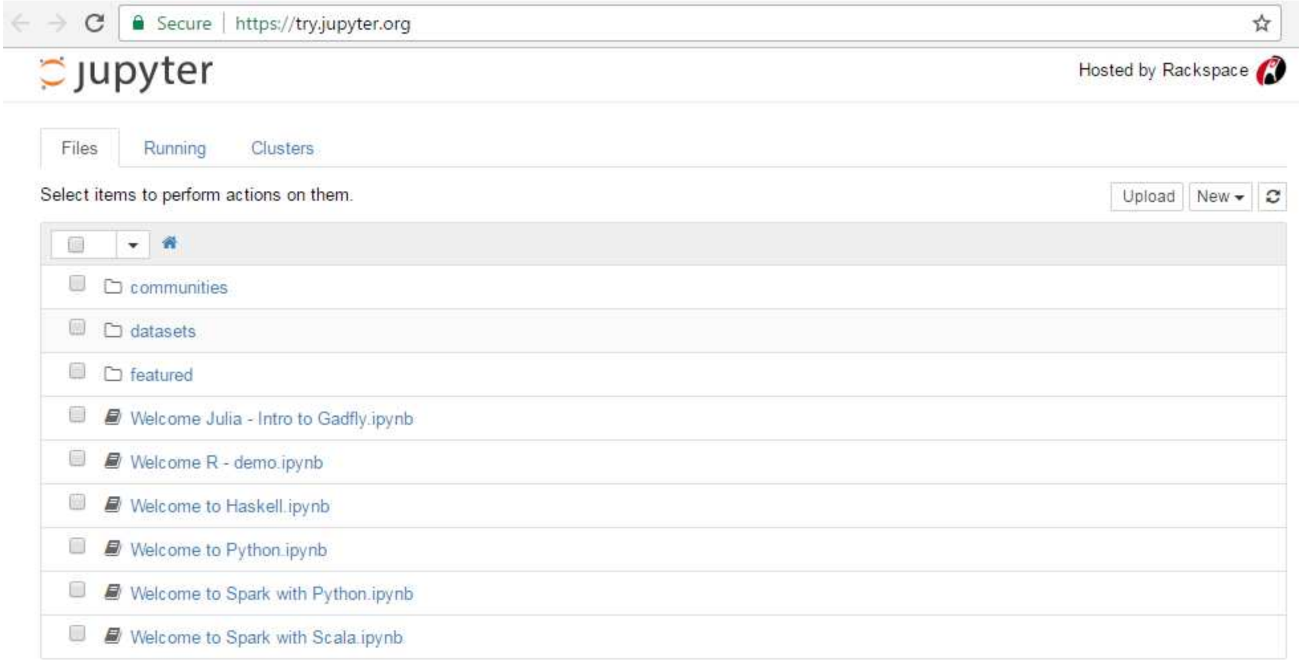
The Jupyter Notebook is a web application that allows

21. You have two choices you can choose to try it in your browser and for the lab exercise you can upload the data to the online site OR install Jupyter Notebook locally on your computer.

It is purely personal preference and if you want to upload your data onto a 3$^{rd}$ party site

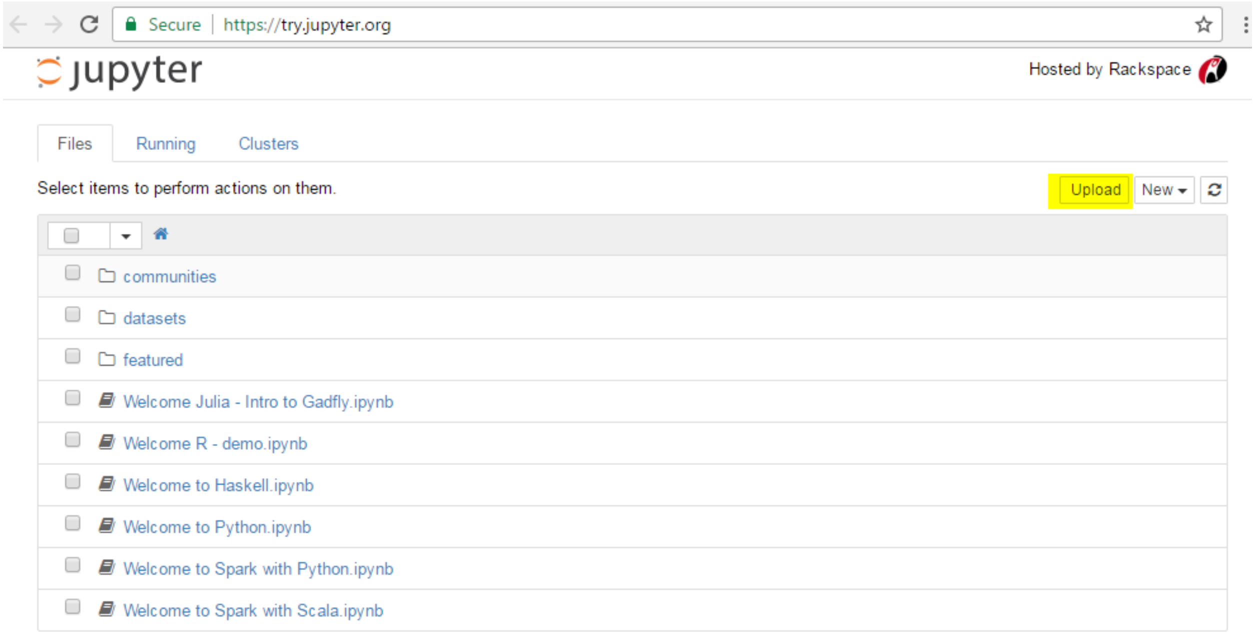
The instructions below work for both “Try it in your browser” or “Install the Notebook”
For those that use “Install the Notebook” replace the language “Jupyter website” with your localhost

22. If you click “Try it in your browser” or you started Jupyter from a local install from “Install the Notebook” you should be presented with the screen below



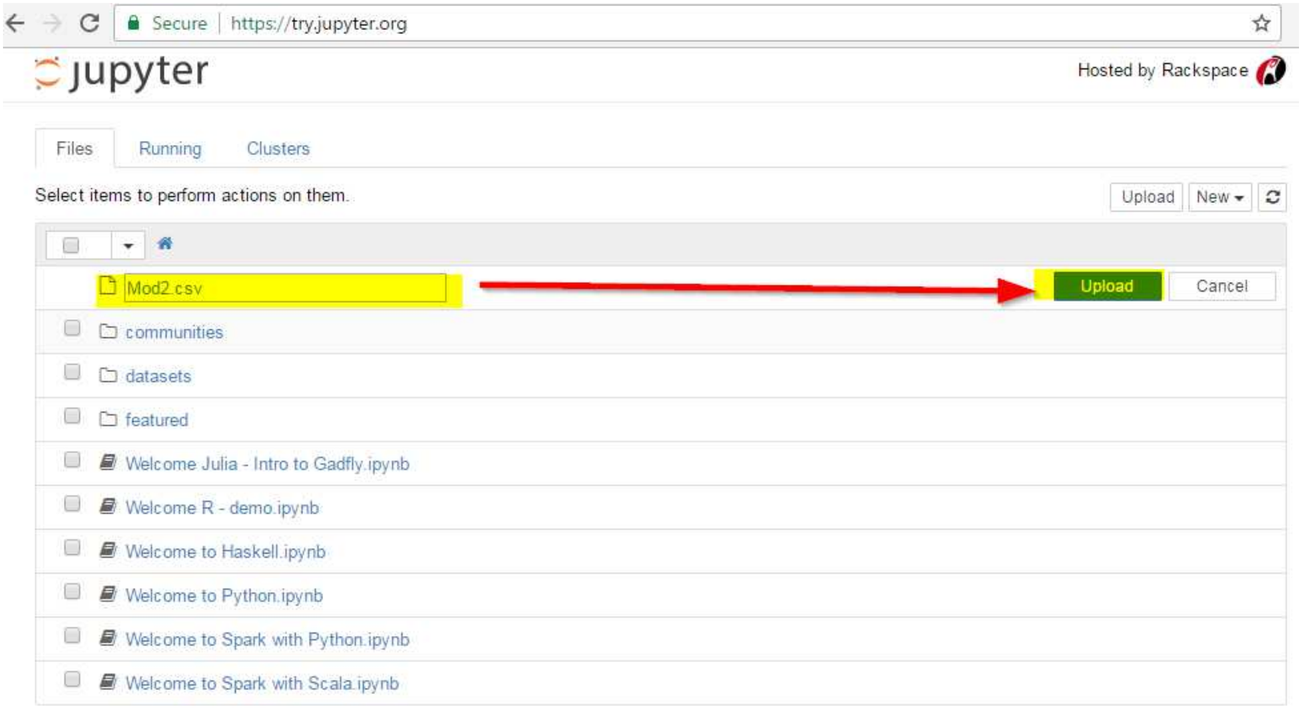
Select items to perform actions on them.

23. First thing is to upload your data to the Jupyter website or use the test sample data provided “Mod2.csv” by going to the top right hand corner and clicking “Upload”

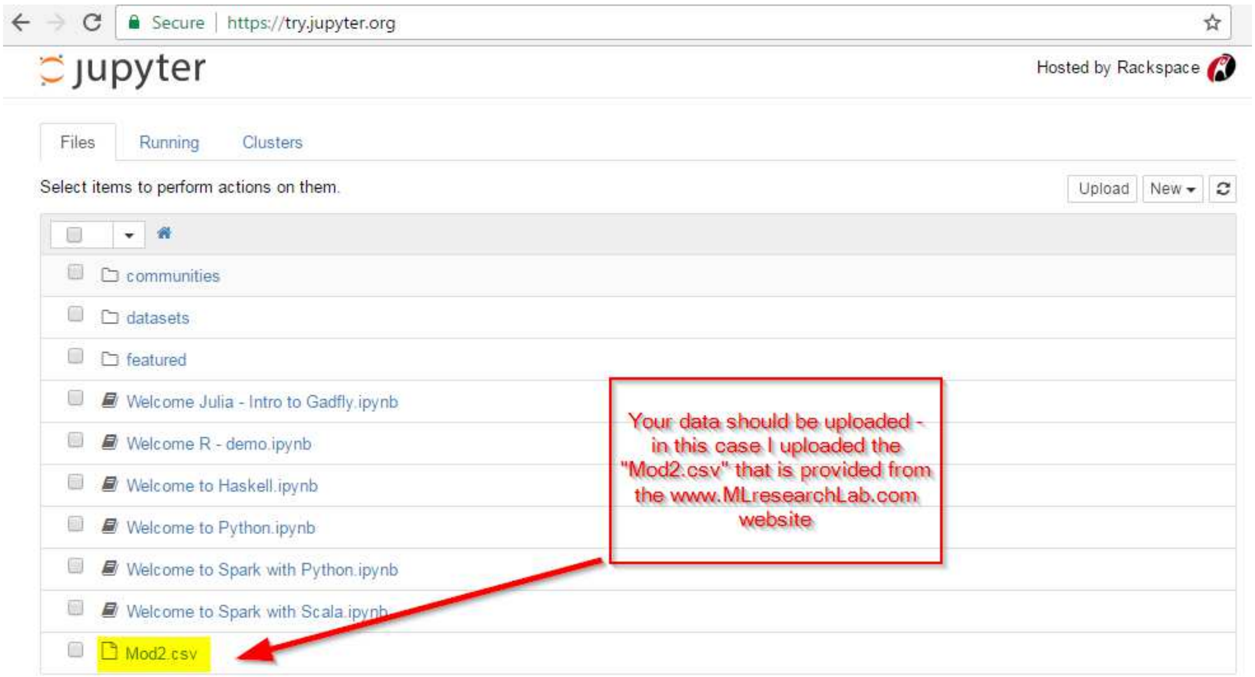


Select items to perform actions on them.

24. The Jupyter website will ask for the file location, once supplied you should see the file ready to be upload as seen in the screenshot below

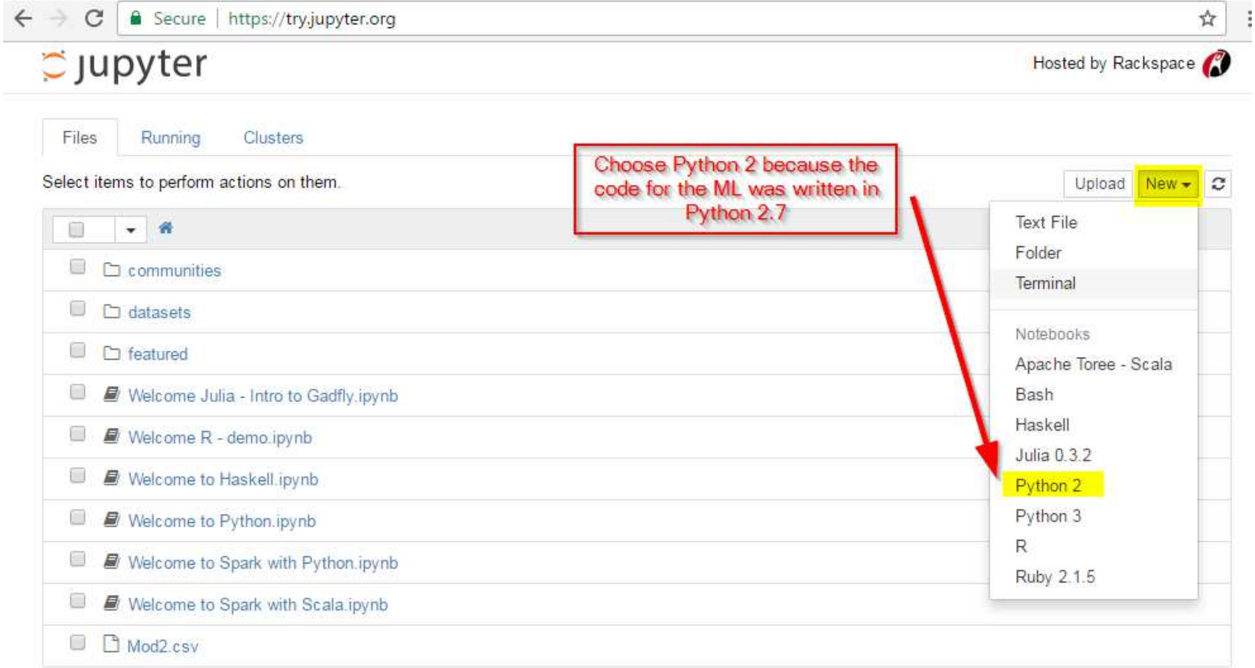

Secure | https://try.jupyter.org
jupyter
Hosted by Rackspace
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Mod2.csv
Upload
Cancel
communities
datasets
featured
Welcome Julia - Intro to Gadfly.ipynb
Welcome R - demo.ipynb
Welcome to Haskell.ipynb
Welcome to Python.ipynb
Welcome to Spark with Python.ipynb
Welcome to Spark with Scala.ipynb

25. Your data should now be upload and you can verify by seeing if the file shows up in the listing as shown in the screenshot below


Secure | https://try.jupyter.org
jupyter
Hosted by Rackspace
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
communities
datasets
featured
Welcome Julia - Intro to Gadfly.ipynb
Welcome R - demo.ipynb
Welcome to Haskell.ipynb
Welcome to Python.ipynb
Welcome to Spark with Python.ipynb
Welcome to Spark with Scala.ipynb
Mod2.csv

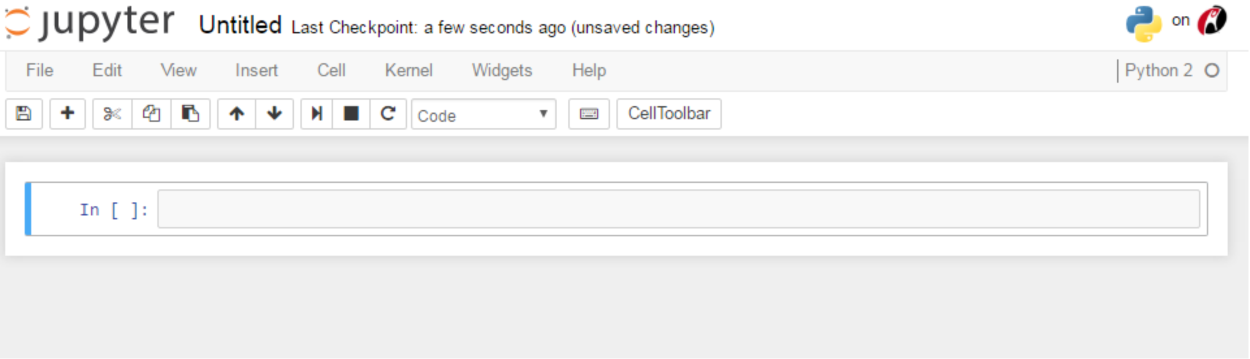
Your data should be uploaded - in this case I uploaded the "Mod2.csv" that is provided from the www.MLresearchLab.com website

26. Now that the data has been uploaded, the next step is to upload the actual ML code so on the top right hand corner click “New” and click “Python 2”

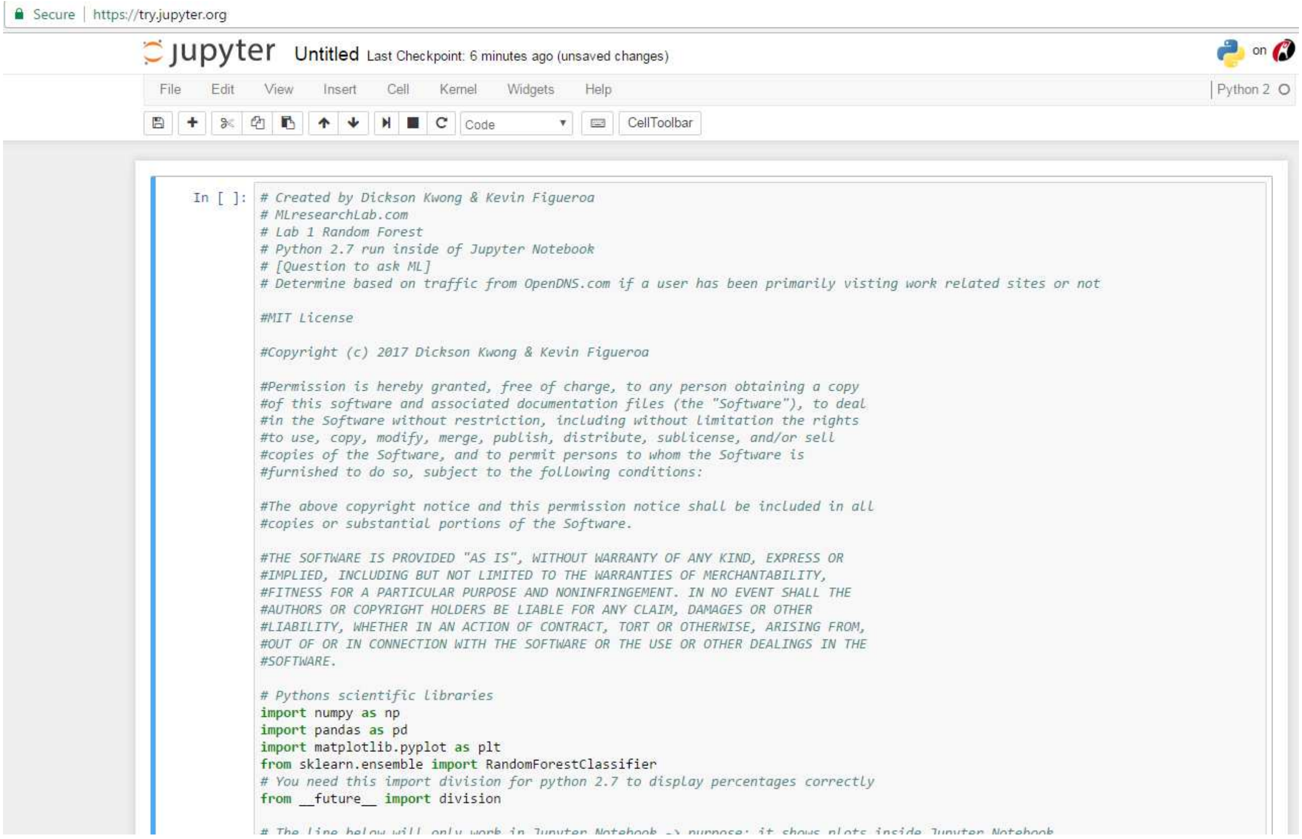


Choose Python 2 because the code for the ML was written in Python 2.7

27. You should be presented with a blank notebook as shown below

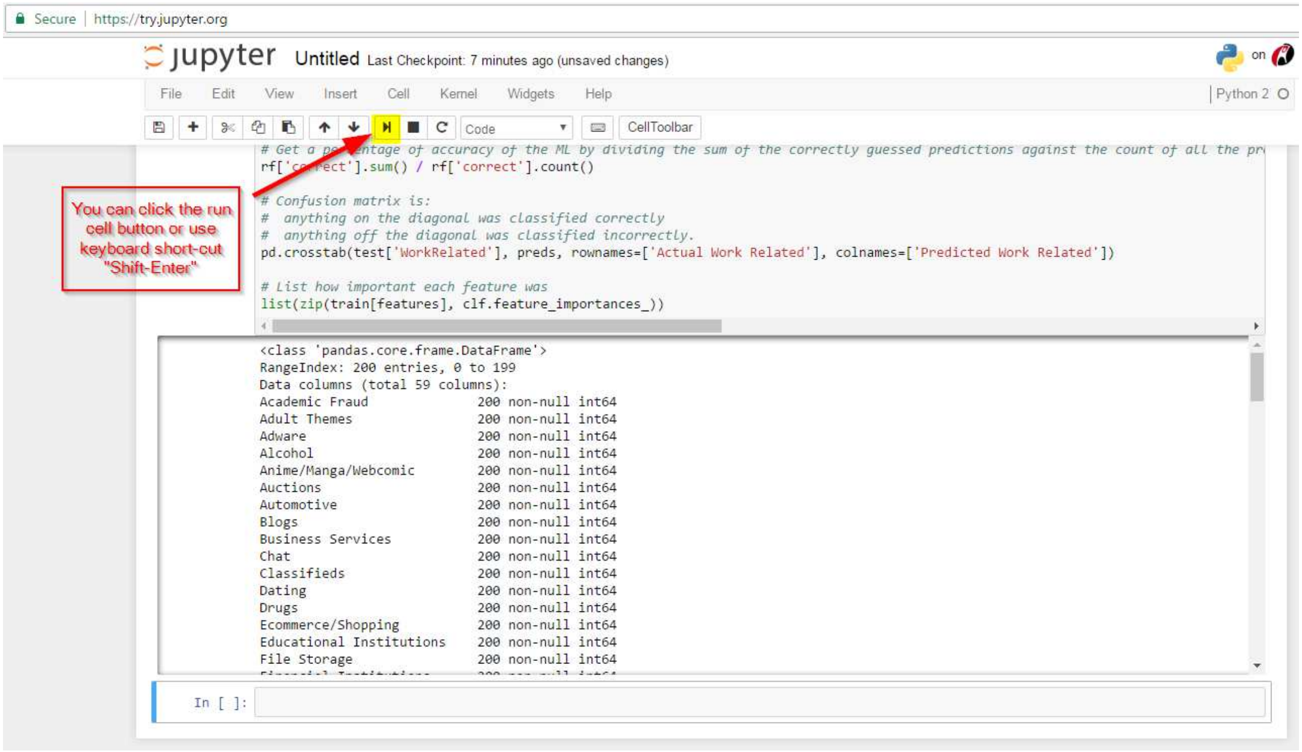


28. Open the “Lab1-Code.py” code from the Github link on www.MLresearchlab.com and Copy the entire contents and paste into your notebook as shown below



```
# Created by Dickson Kwong & Kevin Figueroa
# MLresearchLab.com
# Lab 1 Random Forest
# Python 2.7 run inside of Jupyter Notebook
# [Question to ask ML]
# Determine based on traffic from OpenDNS.com if a user has been primarily visting work related sites or not

#MIT License

#Copyright (c) 2017 Dickson Kwong & Kevin Figueroa

#Permission is hereby granted, free of charge, to any person obtaining a copy
#of this software and associated documentation files (the "Software"), to deal
#in the Software without restriction, including without limitation the rights
#to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
#copies of the Software, and to permit persons to whom the Software is
#furnished to do so, subject to the following conditions:

#The above copyright notice and this permission notice shall be included in all
#copies or substantial portions of the Software.

#THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
#IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
#FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
#AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
#LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
#OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
#SOFTWARE.

# Pythons scientific libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
# You need this import division for python 2.7 to display percentages correctly
from __future__ import division

# The line below will only work in Jupyter Notebook -> purpose: it shows plots inside Jupyter Notebook
```

29. Run the entire cell by either clicking the button on the top center of the screen or using the menu dropdown “Cell” → “Run Cell Below” or you can use the keyboard shortcut “Shift-Enter”. If your data file is correctly formatted and prepared you should see results as shown below with no errors.

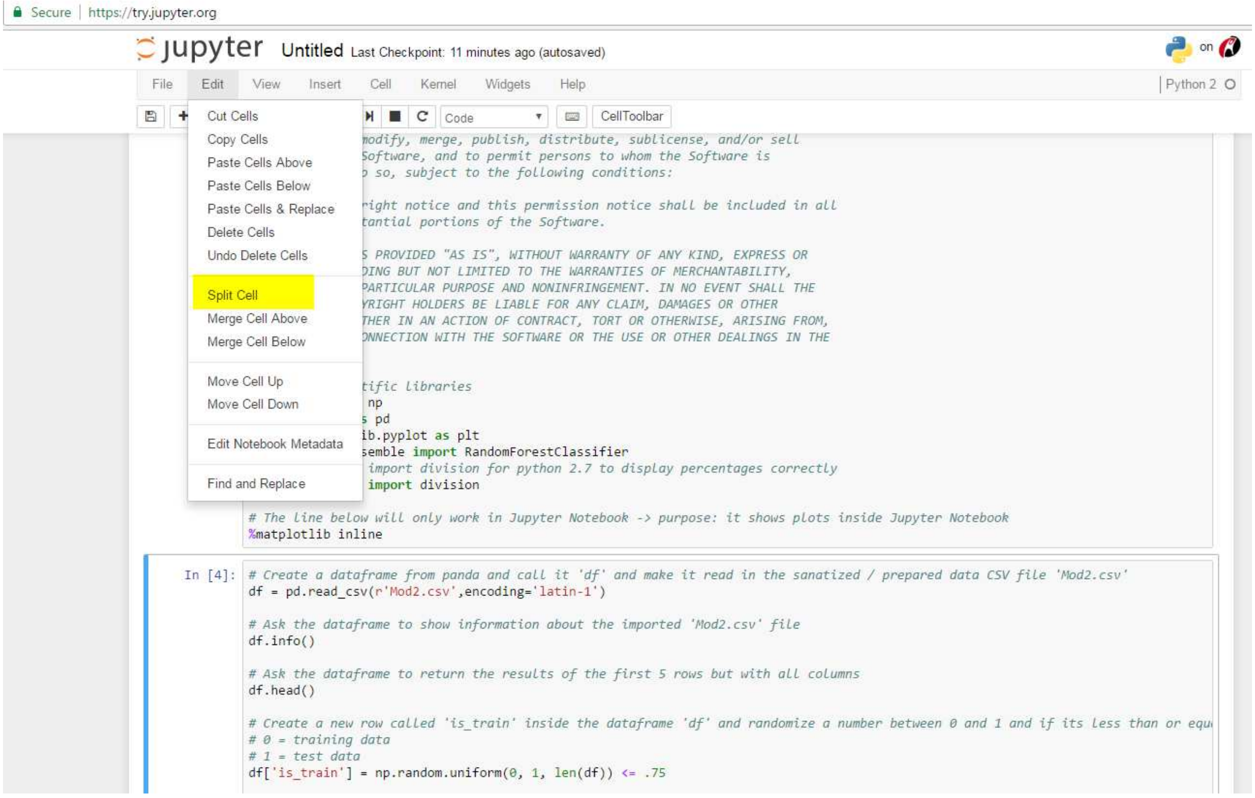

Secure | https://try.jupyter.org
jupyter Untitled Last Checkpoint: 7 minutes ago (unsaved changes)
on
File Edit View Insert Cell Kernel Widgets Help
Python 2
Code
CellToolbar
In [ ]:

You can click the run cell button or use keyboard short-cut "Shift-Enter"

```
# Get a percentage of accuracy of the ML by dividing the sum of the correctly guessed predictions against the count of all the pr
rf['correct'].sum() / rf['correct'].count()

# Confusion matrix is:
#  anything on the diagonal was classified correctly
#  anything off the diagonal was classified incorrectly.
pd.crosstab(test['WorkRelated'], preds, rownames=['Actual Work Related'], colnames=['Predicted Work Related'])

# List how important each feature was
list(zip(train[features], clf.feature_importances_))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 59 columns):
Academic Fraud              200 non-null int64
Adult Themes                200 non-null int64
Adware                      200 non-null int64
Alcohol                     200 non-null int64
Anime/Manga/Webcomic        200 non-null int64
Auctions                    200 non-null int64
Automotive                  200 non-null int64
Blogs                       200 non-null int64
Business Services           200 non-null int64
Chat                        200 non-null int64
Classifieds                 200 non-null int64
Dating                      200 non-null int64
Drugs                       200 non-null int64
Ecommerce/Shopping          200 non-null int64
Educational Institutions    200 non-null int64
File Storage                200 non-null int64
```

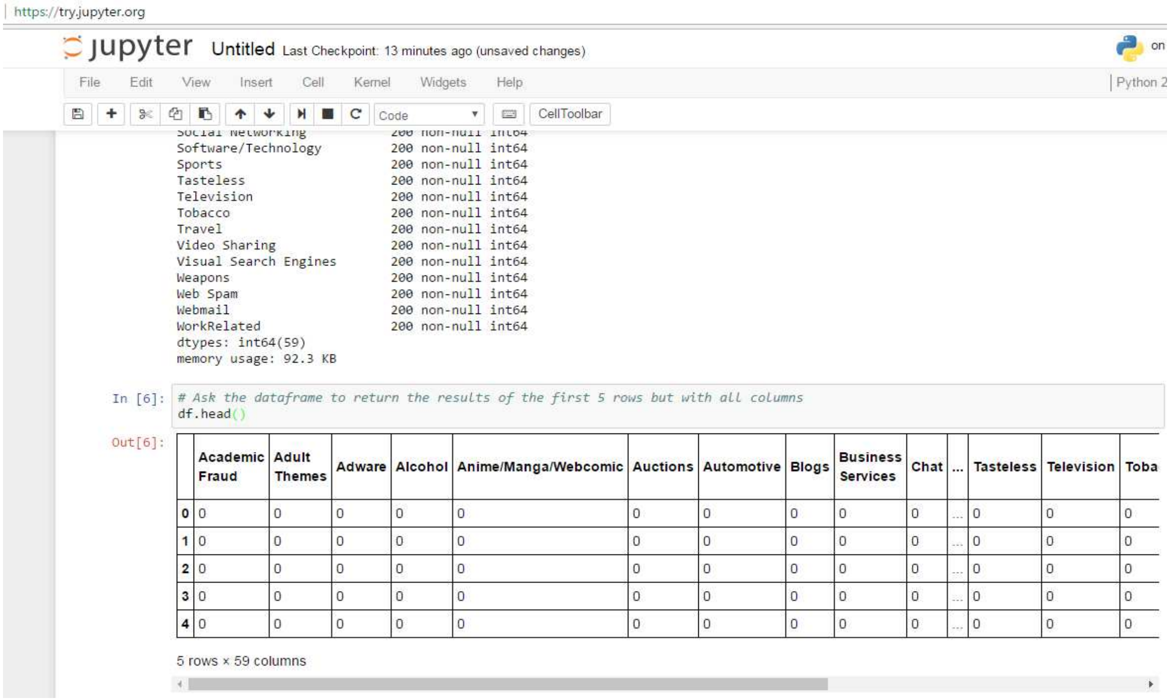
30. Now the results are limited because it only shows the first result and the last result what you want to do is start splitting lines of code into “Cells” so that you can see the individual results of each line of code. You split code into cells by click “Edit” → “Split Cells”



```
modify, merge, publish, distribute, sublicense, and/or sell
Software, and to permit persons to whom the Software is
o so, subject to the following conditions:

right notice and this permission notice shall be included in all
tantial portions of the Software.

S PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
DING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
YRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
THER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
ONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE

tific libraries
 np
s pd
ib.pyplot as plt
semble import RandomForestClassifier
 import division for python 2.7 to display percentages correctly
 import division

# The line below will only work in Jupyter Notebook -> purpose: it shows plots inside Jupyter Notebook
%matplotlib inline
```

```
In [4]: # Create a dataframe from panda and call it 'df' and make it read in the sanatized / prepared data CSV file 'Mod2.csv'
df = pd.read_csv(r'Mod2.csv',encoding='latin-1')

# Ask the dataframe to show information about the imported 'Mod2.csv' file
df.info()

# Ask the dataframe to return the results of the first 5 rows but with all columns
df.head()

# Create a new row called 'is_train' inside the dataframe 'df' and randomize a number between 0 and 1 and if its less than or equ
# 0 = training data
# 1 = test data
df['is_train'] = np.random.uniform(0, 1, len(df)) <= .75
```

31. Once you start splitting code into “Cells” you start to see individual results such as below when you run that “Cell”



```
Social Networking            200 non-null int64
Software/Technology          200 non-null int64
Sports                       200 non-null int64
Tasteless                    200 non-null int64
Television                   200 non-null int64
Tobacco                      200 non-null int64
Travel                       200 non-null int64
Video Sharing                200 non-null int64
Visual Search Engines        200 non-null int64
Weapons                      200 non-null int64
Web Spam                     200 non-null int64
Webmail                      200 non-null int64
WorkRelated                  200 non-null int64
dtypes: int64(59)
memory usage: 92.3 KB
```

```
In [6]: # Ask the dataframe to return the results of the first 5 rows but with all columns
        df.head()
```

Out[6]:

| | Academic Fraud | Adult Themes | Adware | Alcohol | Anime/Manga/Webcomic | Auctions | Automotive | Blogs | Business Services | Chat | ... | Tasteless | Television | Toba |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 |

5 rows × 59 columns

32. If you encounter problems or are unable to run the code please do not hesitate to email “Info@MLresearchLab.com” or on Twitter ML Research Lab@ML_Research_Lab

    There is Vmware image (Link on www.MLresearchLab.com) of a local install with Jupyter installed with the sample data all ready to go if you do not want to upload your data online and you want to have a quick environment to learn from.

    Questions / Comments / Concerns please do not hesitate to reach out to us we want to help the security community to grow and understand ML

    Thank you for trying out the lab and more labs to come!