SPAM URL DETECTOR

TO CLICK OR NOT TO CLICK

DEVELOPER: KEJIN QIAN

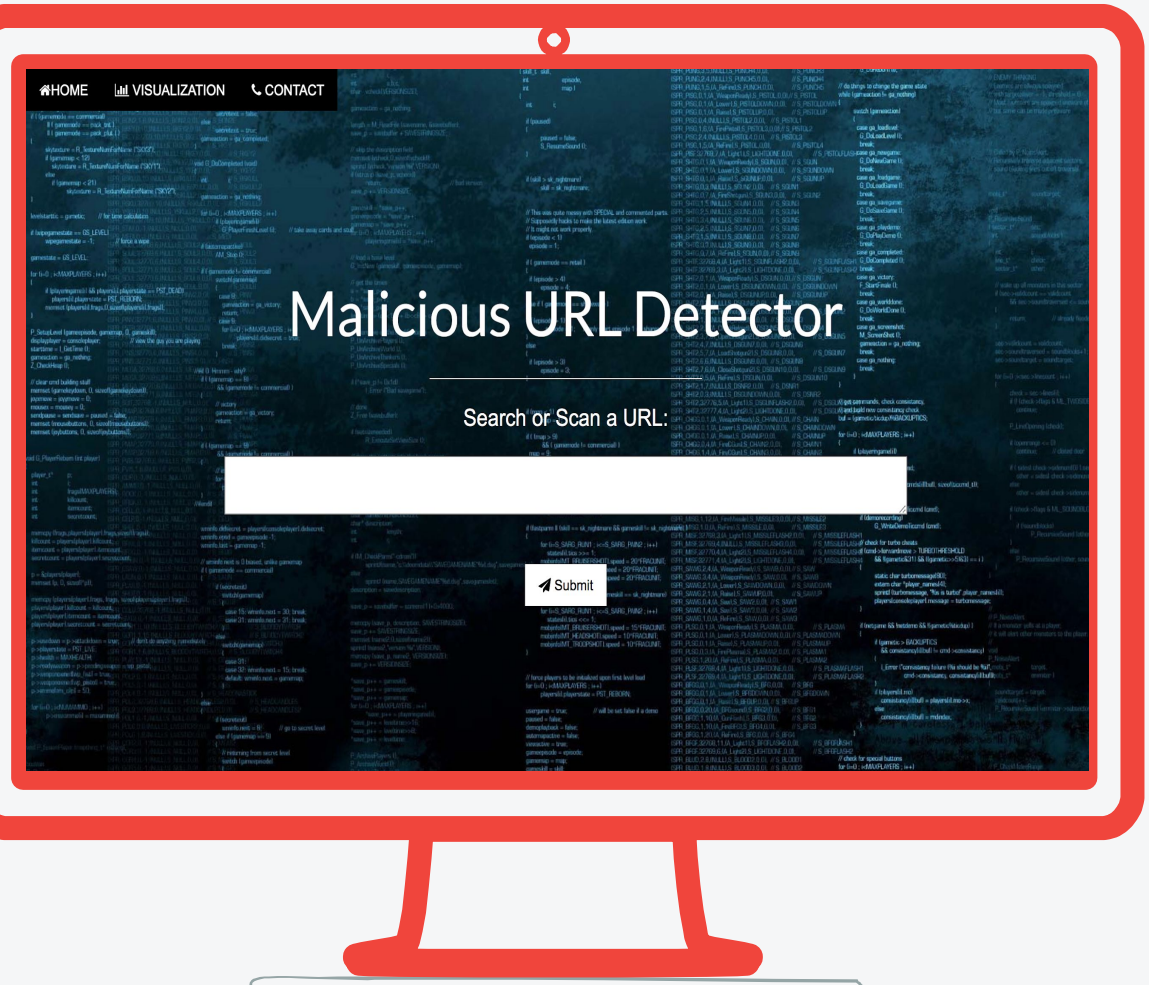
JUNE 10TH 2019

Motivation

- **What I experienced:**
  - More than 10 job posting/referral/job offer emails every day
  - all from different email addresses, no way to stop it
  - > 2000 spam emails in 6 months
- **What I need:**
  - Create a spam classifier using primarily the url string as features.
  - Search engines have limited crawl resources, so being able to identify a malicious url without retrieving the page content can save a lot of resource and reduce spam to users.

App Demo

http://3.18.102.108:3000/



DATA

01. MALICIOUS URL(13854)

- All the malicious/phishing URLs were collected from an anti-phishing site *Phishtank.com*

Recent Submissions

You can help! Sign in or register (free! fast!) to verify these suspected phishes.

ID	URL	Submitted by
6071061	https://www.lepanierdurousillon.fr/modules/protec...	PhishReporter
6071058	http://smitahav.aba.ae/aba/cust0mersmanagementdepa...	PhishReporter
6071057	http://smitahav.aba.ae/aba/cust0mersmanagementdepa...	PhishReporter
6071055	https://aqualimpa.com.br/usaa/	cleanmx
6071054	https://aqualimpa.com.br/usaa/contact%20info.html	cleanmx
6071053	https://aqualimpa.com.br/usaa/Security%20Question%...	cleanmx
6071052	https://aqualimpa.com.br/usaa/USAA%20_%20Welcome%2...	cleanmx

02. BENIGN URL(RANDOM SUBSET 13854)

- All the benign URLs were collected from a link index database *The Majestic Million*
- The dataset contains domains with most referring subnets and IPs
- Problem: All urls here have much shorter url length and short domain length compared to the malicious urls collected

Position	+/-	Domain	TLD	TLD Rank	Referring Subnets	Referring IPs
1		google.com	.com	1	486,832 586	2,977,019 383
2		facebook.com	.com	2	472,757 496	2,990,331 1,750
3		youtube.com	.com	3	432,363 620	2,449,031 1,487
4		twitter.com	.com	4	422,069 433	2,408,780 805
5		linkedin.com	.com	5	314,181 543	1,397,821 440
6		microsoft.com	.com	6	313,387 516	1,226,059 88
7		instagram.com	.com	7	310,896 437	1,517,932 1,069
8		wikipedia.org	.org	1	292,174 429	1,204,439 215
9		apple.com	.com	8	288,995 768	1,056,695 82

03. SCRAPED URL FROM GITHUB(RANDOM SUBSET 20000)

- To offset the impact of URL length in classification, a list of scraped URLs with various length were collected from an existing GitHub repo.

```
divorcesource.com/ds/main/site-map-1457.shtml,0
cancertutor.com/discussions/how-can-i-suggest-an-alternative-treatment-for-a-family-member,0
byu.edu/about-byu/accreditation,0
twitter.com/intent/tweet?url=www.thecalifornian.com/picture-gallery/news/2017/03/07/photos-salinas-sanctuary-city-demonstrationlawsuit-announcement/98883962/&text=PHOTOS%20Salinas%20sanctuary%20city%20demonstration/lawsuit%20announcement&via=salnews,0
technologyevaluation.com,0
gurl.com/category/your-life,0
articlesbase.com/art-and-entertainment-articles/my-top-10-all-time-black-and-white-american-films-before-1962-1974249.html,0
wikipilipinas.org/index.php?title=Sabas%2C_Ang_Barbaro,0
ipsnews.net/primary-region/global,0
lazada.co.id/sanken-official-store,0
fastweb.it/adsl-fibra-ottica/rete-mobile/?from=menu_home,0
php100.com/sjhass/c8bvkf2p,0
bleep.com/label/605-infin,0
ecocontrolsystem.com.br/ecocontrolsystem.com.br/mecs.html,0
brunel.ac.uk/about/chls,0
perl.com/pub/2002/10,0
nmgnews.com.cn/system/2016/11/26/012198481.shtml,0
```

Features Generated

01. Extracted from URL string

- # of dots . appeared
- # of soft hyphens - appeared
- # of @ appeared
- # of double slashes
- URL length
- Contains IP?
- # of Queries
- # of subdomain
- presence of Suspicious_TLD
- File extension

02. WHOIS server

- Create_age(months)
- expiry_age(months)
- update_age(days)
- Country
- Domain length

Sample WHOIS output:

https://www.lepanierduroussillon.fr/modules/protectmyshop/lgn/

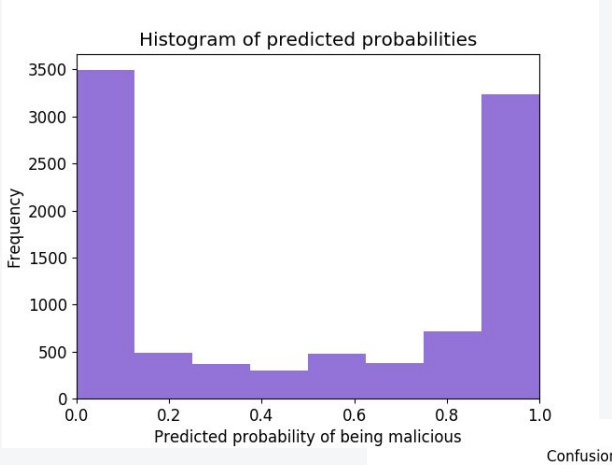
```
domain:       lepanierduroussillon.fr
status:       ACTIVE
hold:         NO
holder-c:     LPDR49-FRNIC
admin-c:      OVH5-FRNIC
tech-c:       OVH5-FRNIC
zone-c:       NFC1-FRNIC
nsl-id:       NSL93567-FRNIC
dsl-id:       SIGN1263276-FRNIC
registrar:    OVH

Expiry Date: 2019-12-25T23:07:39Z
created:     2014-10-29T17:30:51Z
last-update: 2018-12-25T23:34:07Z
source:      FRNIC

country:      FR

phone:        +33 8 99 70 17 61

e-mail:       tech@ovh.net
```

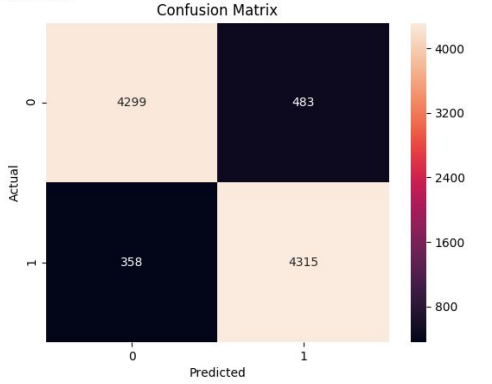
Model - Random Forest

- Tuned and compared 5 models: Random Forest, Decision Tree, AdaBoost, Gradient Boosting, Logistic Regression

- Selected Random Forest with number of trees = 150 as the final model

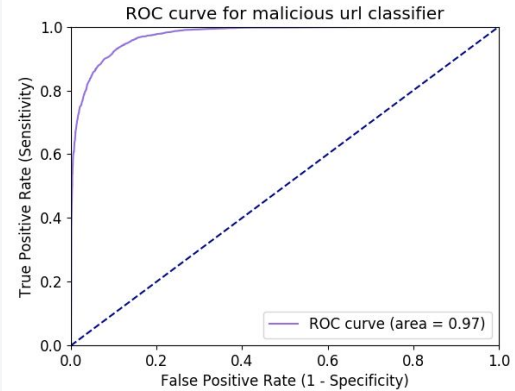
Random Forest	n_estimators = 150
Classification Accuracy	0.9110523532522475
False Positive	0.10100376411543287
False Negative	0.07661031457307939
F1 Score	0.9112026185196916
AUC	0.9738898669783425





Success Criteria

- Successfully build a Machine Learning predictive model that dynamically classifies user-provided URL strings into Malicious/Benign class with a **F-score greater than 0.85** and **AUC higher than 0.7**.
- Get above 50% of returning users.

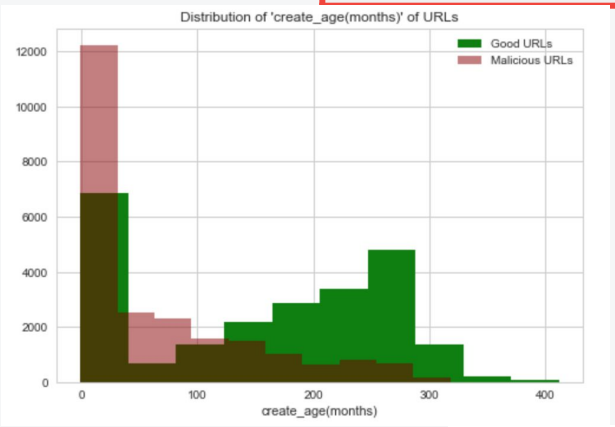


Insights

Feature	Importance
File extension	0.282621
Create age (months)	0.212760
Length of domain	0.138897
Number of subdomain	0.081531
Expiry age (months)	0.074433
Update age (days)	0.064525

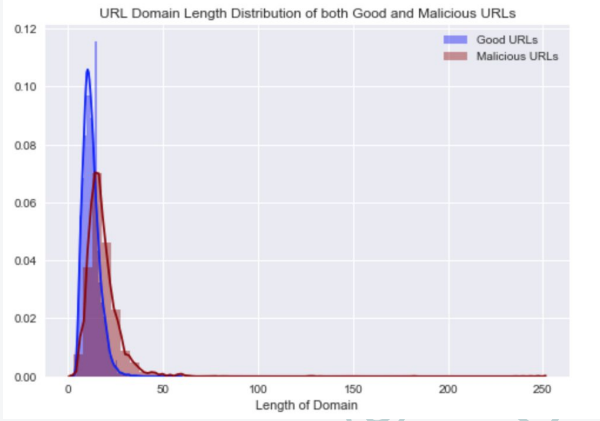
- ## Create Age(Months):

  Malicious urls are mostly likely to have shorter create age(0 ~ 20 months). And the frequency decreases as the create_age(months) increases. While more than 60% of the good urls have create_age longer than 100 months.



- ## Domain Length:

  The domain length of malicious urls has a larger average and variance. Its values ranging from 0 to 250. Whereas good urls tend to have a shorter and more steady domain length.



- ## Number of Subdomain:

  All urls with number of subdomains greater than 2 are malicious, according to the dataset.

Insights

- ## Punctuations:

  All urls with double slashes, soft hyphen, dots or @ in domain are highly likely to be malicious.
  Although those punctuation features do not stand out in the feature importance calculation as the number of relevant urls is too small, we should pay extra attention to them.

- ## Existence of IP:

  Users should be careful to urls which use an IP address as an alternative of the domain name.
  From my dataset:
  There are 595 urls which use IP address as domain name and 594 of them are malicious.

Thank You

Developer: Kejin Qian

Master of Science in Analytics

Northwestern University

https://github.com/kejin-qian/Malicious-URL-Classifier

https://www.linkedin.com/in/kejin-qian/

Evanston, IL, US

Email: kejinqian2019@u.northwestern.edu