Tal Shafir
Email: Talshaf93@gmail.com

Phishing detection research task - Perception Point

The Task

Research and implementation of an algorithm that detects phishing URLs labeled as Microsoft and confirms for legitimate URLs.

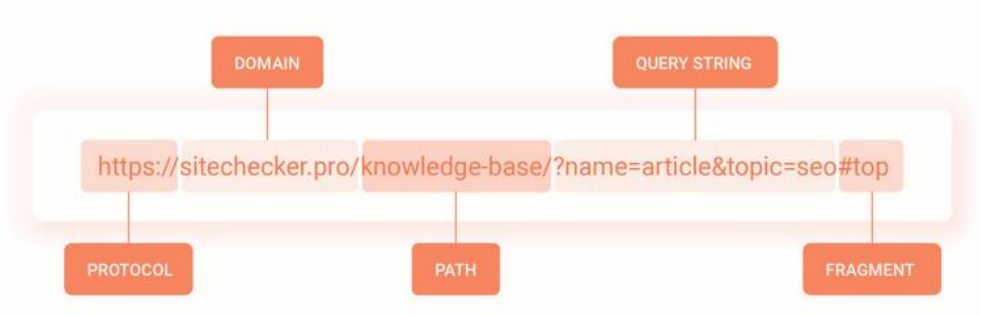
Research

Most studies show the features that can identify a phishing URL are categorized to 4 sections:

- URL structure features
- URL lexical features (heavy models – can be more than 1M features)
- URL domain features
- URL destination page features

URL's structure:

To better understand how to detect a phishing URL I started by getting better understanding of the URL's structure, according to most of the research I done online, the URL should be divided to 5 sections:



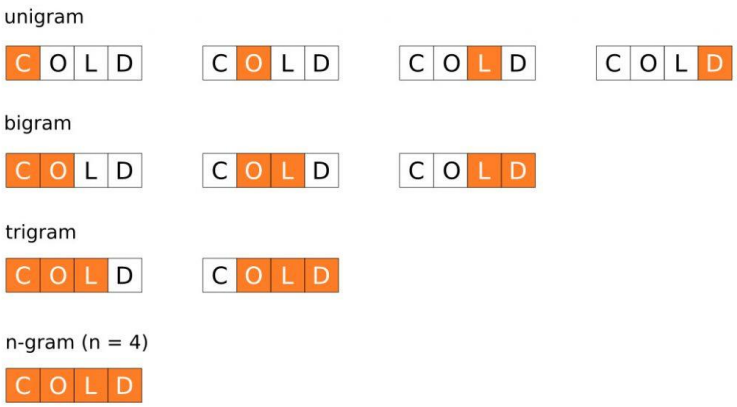
* In the project, using a function from urllib library the **protocol**, **domain**, **path**, **query**, and **fragment** were extracted from thed URL and respective columns were created.

For each section in the URL, we can find marks and patterns that indicates on a phishing structed URL. Therefore, for each section I checked for its length and the quantity of specific characters.

Tal Shafir
Email: Talshaf93@gmail.com

Lexical features:

I have found that in advanced studies more complex algorithms from the field of NLP have been used to extract features from the URL relating to lexicographic strings, checking individual letters, pairs, triplets and so on within the URL can give a best indication of malicious URL.



One of the methods I have seen is the **n-gram** method by which we run on a labeled dataset and create features using the information available for each URL.

URL domain features:

The information regarding the URL's domain can give us a better understanding if the URL can be a scam:

- Protocol – http/https (although today more and more sophisticated phishing URLs are labeled secure with https://).
- Traffic – more traffic a domain has the more it is likely to be safe.
- Survival time of domain (domain age) – usually phishing domains are likely to be younger than legitimate domains.

URL destination page features:

The highest threat of a phishing URL comes from its destination page where than can cause multiple functions that are commonly used be phishing links:

- IFrame redirection
- Checks the effect of mouse over on status bar
- Checks the status of the right click attribute
- Number of pages forwarding
- Html content

Tal Shafir
Email: Talshaf93@gmail.com

Data Collection

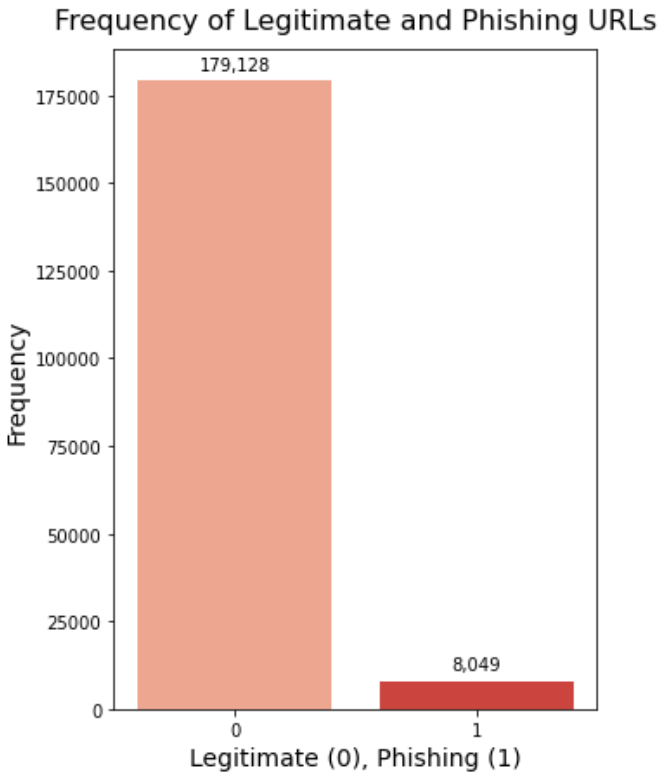
For this task I had to look for both phishing and legitimate URLs datasets, Phishing URLs were pulled from PhishTank and Microsoft site and legitimate URLs were pulled from Alexa top million sites dataset.

Over all I gathered:

```
phishing urls: 8053
legit urls: 537394
```

in this section I tried filtering the entire dataset so it will better suit Microsoft purposes, I dumped from the dataset every URL that didn't contain a Microsoft key word, but after running the whole process this method gave much worse result, therefore I concentrated on a global URL distribution.

Because the dataset was very unbalance in reference to the suggested number of phishing URLs are actually exist on the internet (~4%), I dumped a big number of legitimate URLs to create this balance:



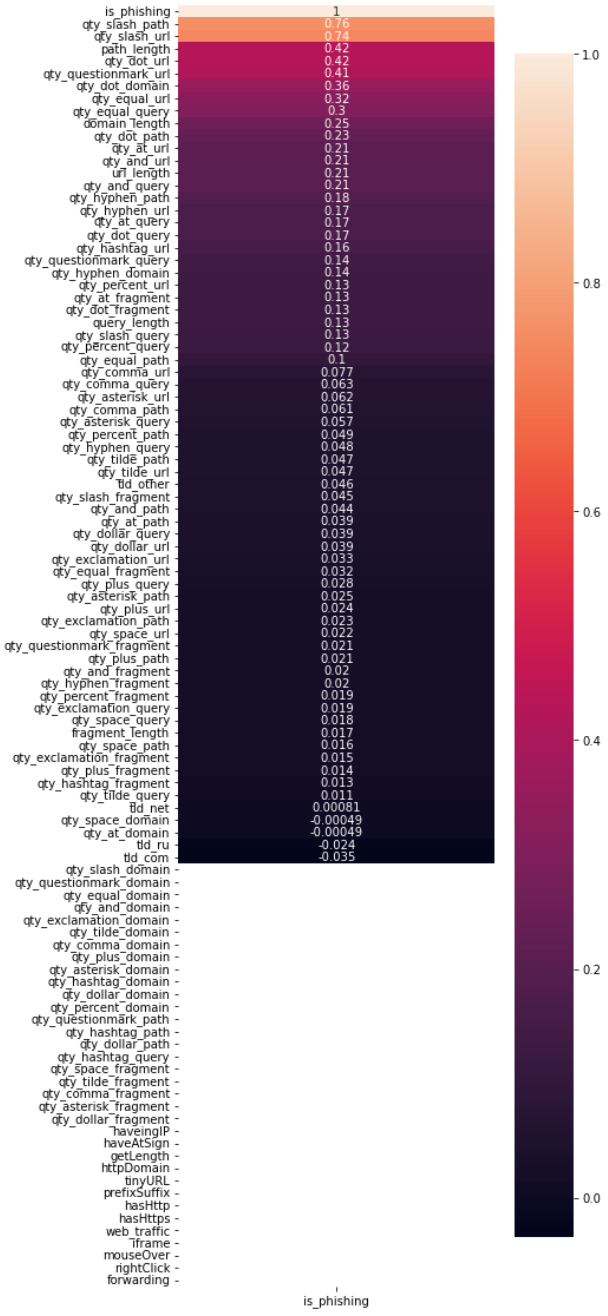
Tal Shafir
Email: Talshaf93@gmail.com

Pre-Processing

In this section I performed a number of actions to create a learnable dataset we can use in the learning process:

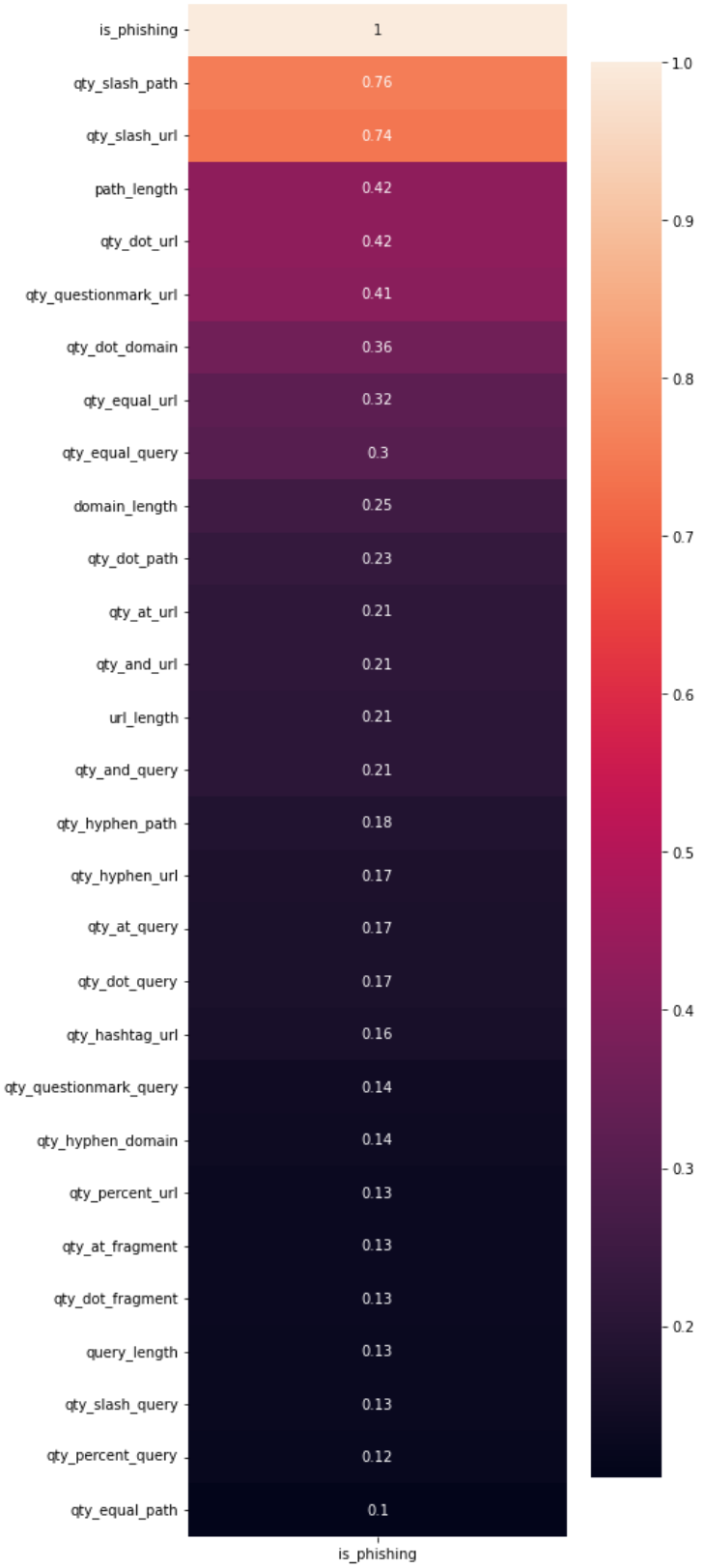
- Data cleaning
- Features extraction
- Features relevance measurement

After completing the data cleaning and features extraction I calculated the correlation of each feature to the label:



Tal Shafir
Email: Talshaf93@gmail.com

We can see a huge number of features that are on low correlation rate and some of thew are non-correlated at all. Therefore, I dumped all the features that are ranked correlation under 0.1:



Tal Shafir
Email: Talshaf93@gmail.com

Models Selection and Evaluation

For this dataset I tested few relevant modules:

- Stochastic Gradient Descent Classifier
- Logistic Regression
- Support Vector Machine
- AdaBoost
- Gradient Boost
- Decision Tree Classifier
- Bagging Classifier
- K-Nearest Neighbors Classifier
- Extra Trees Classifier
- Random Forest Classifier

I spited the dataset into Train and Test data frames and calculated for each model its evaluation metrices. The total majority of the models hade high accuracy rate and good performance over all matrices, that I assumed, is because of the imbalanced data and the fact that the legitimate URLs dataset is not diverse and very unique, plus, the majority of the features didn’t work.

Tal Shafir
Email: Talshaf93@gmail.com

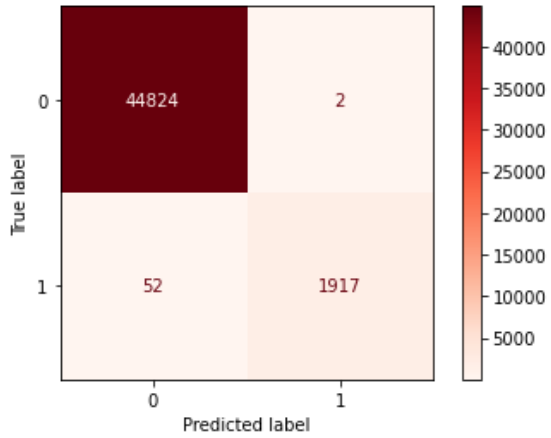
To complete the task, I choose Random Forest model to be trained over my dataset, I performed hyper parameter tuning for better results and calculates the overall model evaluation metrices over the tested data:

```
Fitting 3 folds for each of 10 candidates, totalling 30 fits
Best Parameters: {'n_estimators': 400, 'min_samples_split': 5, 'min_sam
ples_leaf': 1, 'max_features': 'auto', 'max_depth': 50, 'bootstrap': Fa
lse}
```

```
Training Score: 0.9992947813822285
Testing Score: 0.9988460305588204
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	44826
1	1.00	0.97	0.99	1969
accuracy			1.00	46795
macro avg	1.00	0.99	0.99	46795
weighted avg	1.00	1.00	1.00	46795

```
[[44824     2]
 [   52  1917]]
```



Tal Shafir
Email: Talshaf93@gmail.com

Conclusion

This task can be solved in a more professional and in-depth way with regard to research, data collection, features extraction and in any other aspect. Given more time for the task I would choose to go in the direction of developing a neural network that combines the different types of features and especially in the direction of developing lexical features. The existing model of course does not satisfy the requirement as it is mainly built on features related to the URL structure, which in this case, cannot work well as long as the data for legitimate URLs does not satisfied and contains only the domain.

Resources

Data Collection:

1. Phishtank.com
2. https://docs.microsoft.com/en-us/azure/devops/organizations/security/allow-list-ip-url?view=azure-devops&tabs=IP-V4#azure-devops-import-service
3. https://github.com/v2ray/domain-list-community/blob/master/data/microsoft
4. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip

Research:

1. Intelligent phishing url detection using association rule mining. S. Carolin Jeeva & Elijah Blessing Rajsingh
2. Phishing URL Detection Through Top-Level Domain Analysis: A Descriptive Approach. Orestis Christou, Nikolaos Pitropakis, Pavlos Papadopoulos, Sean McKeown and William J. Buchanan
3. Hybrid Rule-Based Solution for Phishing URL Detection Using Convolutional Neural Network. Youness Mourtaji , 1 Mohammed Bouhorma , 1 Daniyal Alghazzawi , 2 Ghadah Aldabbagh , 3 and Abdullah Alghamdi 2
4. What's in a URL: Fast Feature Extraction and Malicious URL Detection. Rakesh Verma Avisha Das
5. Youtube.com