Phishing Site Predictor

Yashwanth Sai Tirukkovalluru

And

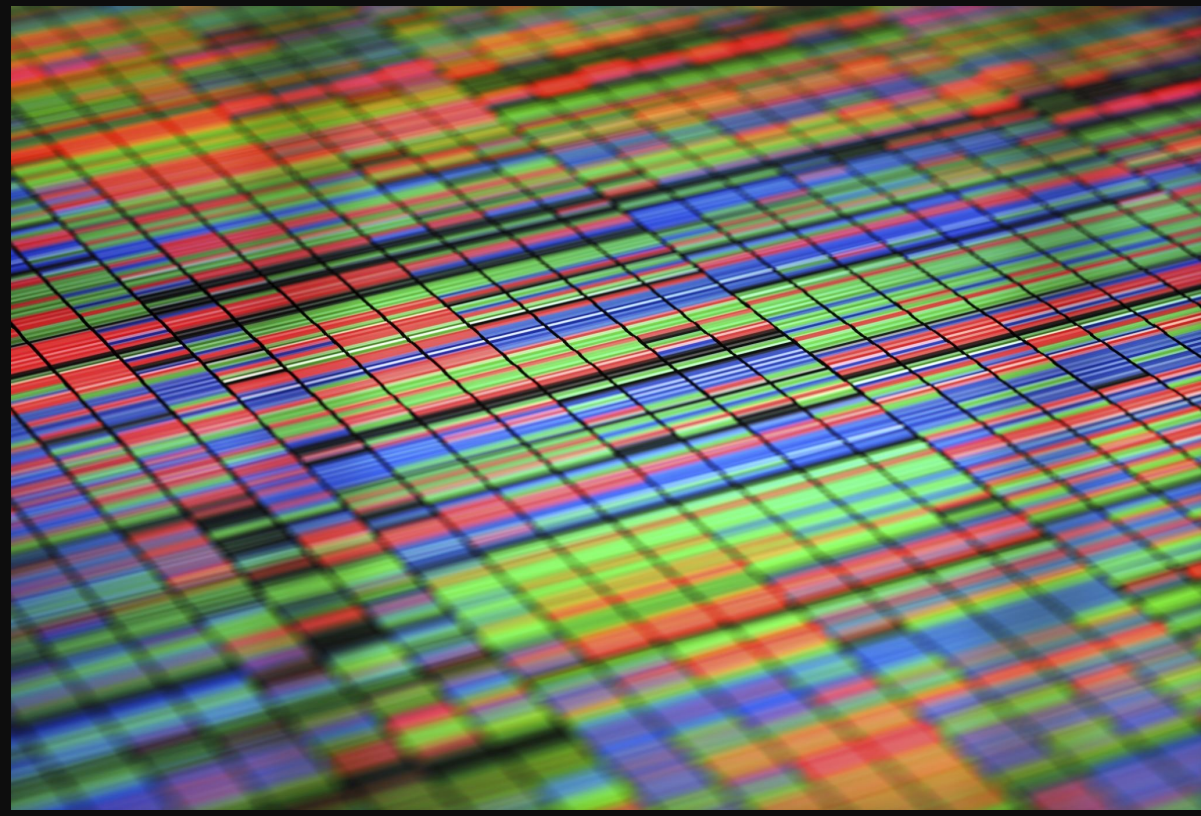
Sandipta Subir Khare

Libraries used

- Pandas
- NumPy
- Seaborn
- Matplotlib
- Sklearn
- Nltk
- Wordcloud
- Bs4
- Selenium

Dataset

https://www.kaggle.com/datasets/taruntiwarihp/phishing-site-urls



Regexp Tokenizer

- A tokenizer that splits a string using a regular expression, which matches either the tokens or the separators between tokens.

	URL	Label	tokenized
54491	www.iadis.org/icwi2006/	good	[www, iadis, org, icwi]
236406	sabres.nhl.com/club/player.htm?id=8468700	good	[sabres, nhl, com, club, player, htm, id]
272271	allposters.com/-st/Arthur-Lismer-Posters_c3195...	good	[allposters, com, st, Arthur, Lismer, Posters,...
453956	tvguide.com/tvshows/home-improvement/100210	good	[tvguide, com, tvshows, home, improvement]
209732	loveofthegameproductions.com/modules.php?name=...	good	[loveofthegameproductions, com, modules, php, ...

	URL	Label	text_tokenized	text_stemmed
10834	solonegocios.com.do/templates/beez/secure_Aktu...	bad	[solonegocios, com, do, templates, beez, secur...	[solonegocio, com, do, templat, beez, secur, a...
546442	sevegep.com/dfmine/	bad	[sevegep, com, dfmine]	[sevegep, com, dfmine]
474387	youtube.com/watch?v=Sr43EG0ksFM	good	[youtube, com, watch, v, Sr, EG, ksFM]	[youtub, com, watch, v, sr, eg, ksfm]
353728	homeinfomax.com/Public/County_Records/TX-Texas...	good	[homeinfomax, com, Public, County, Records, TX...	[homeinfomax, com, public, counti, record, tx,...
422741	reuters.com/article/2011/10/07/us-banks-europe...	good	[reuters, com, article, us, banks, europe, idU...	[reuter, com, articl, us, bank, europ, idustr, u]

- **First, we loaded the dataset from the Kaggle**
- **The data we took for prediction was already classified into two categories good and bad URLs.**
- **We used Regexp Tokenizer and Snowball**
- **We used snowball steamer basically it gives the root words.**
- **We made a visualizion bad and good URLs**

Visualization of common words from URLs

- Good urls

Bad urls


text_sent
appl
checker
nobel
skype
onlin
info
dghjdgf
bin
f
co
paypal
org
printakid
dtype
cce
theme
dca
serviciosbi
Length
d
Name
herf
content
webscrc
americanexpress
uptoday
thewhiskeydreg
cycgi
cg
iclod
en
uk
search
widescreen
wp
login
ffb
cgi
object
c
s
mail

Chrome driver

WebDriver tool use for automated testing of webapps across many browsers. It provides capabilities for navigating to web pages, user input and more

There are some malicious sites to grab the internal URLs we used chrome driver. Then we turned them into a dataframe.

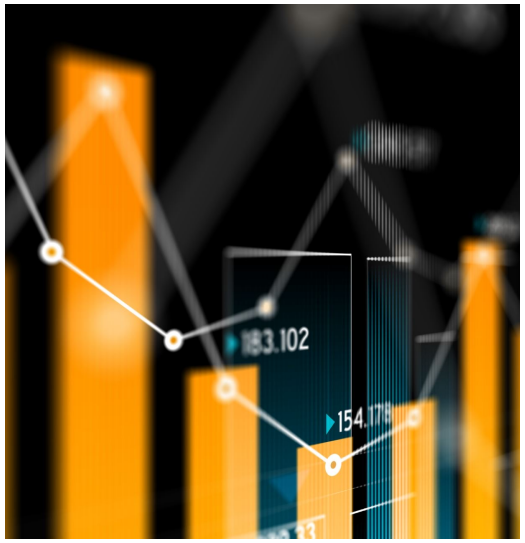
Use the BeautifulSoup library to extract only relevant hyperlinks for Google, i.e. links only with '<'a'>' tags with href attributes.

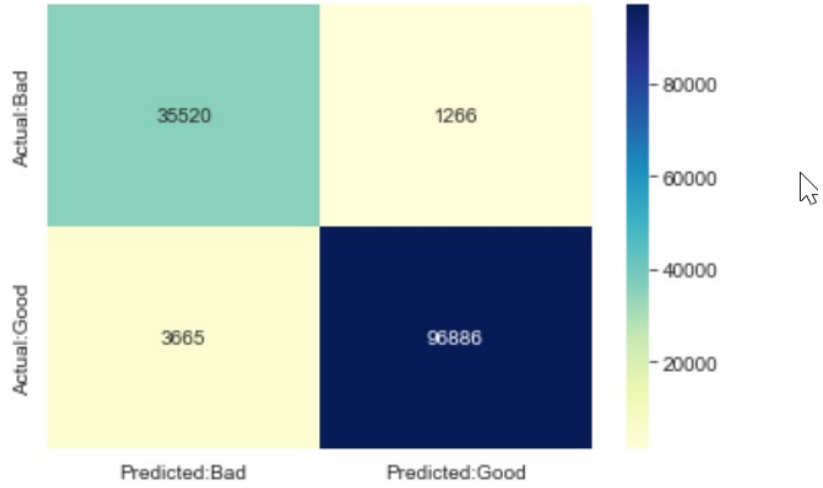
BeautifulSoup

We applied BeautifulSoup to scrape the words from the internal URLs.

Logistic Regression

- Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.





sklearn pipeline using Logistic Regression

- The pipeline is **a series of functions that the data is passed through, cumulating in the logistic regression model**. In the pipeline, numeric values are first scaled to a z-score using the StandardScaler() function.

Training Accuracy : 0.9786558060624889
Testing Accuracy : 0.9640956188062 94

CLASSIFICATION REPORT

	precision	recall	f1-score	support
Bad	0.91	0.97	0.94	36786
Good	0.99	0.96	0.98	100551
accuracy			0.96	137337
macro avg	0.95	0.96	0.96	137337
weighted avg	0.97	0.96	0.96	137337

Deployment

- **Unicorn**
- **Fastapi**
- **Joblib**

- **What is FastAPI used for?**
- **FastAPI is a Python framework and set of tools that enables developers to use a REST interface to call commonly used functions to implement applications.**

Name Description

features * required

(query)

https://www.youtube.com/

Execute Clear

Responses

Curl

```
curl -X 'GET' \
  'http://127.0.0.1:8000/predict/{feature}?features=https%3A%2F%2Fwww.youtube.com%2F' \
  -H 'accept: application/json'
```

Request URL

```
http://127.0.0.1:8000/predict/{feature}?features=https%3A%2F%2Fwww.youtube.com%2F
```

Server response

Code	Details
200	Response body

```
[
  "https://www.youtube.com/",
  "This is not a Phishing Site"
]
```

Downl

References

- https://www.youtube.com/watch?v=zKNXHluHneU&t=339s
- https://www.youtube.com/watch?v=-ykeT6kk4bk&t=607s
- https://www.youtube.com/watch?v=yIYKR4sgzI8
- https://www.youtube.com/watch?v=pqNCD_5r0IU
- https://ieeexplore.ieee.org/document/8250278
- https://link.springer.com/chapter/10.1007/978-981-33-4299-6_12


THANK
YOU