Website Phishing Detection

Andy Alvarez

6140523

CNT 6150

Introduction

Main Goal: Predict the legitimacy status of a website.

What is Phishing?

- An attacker impersonates a well-known entity to extract sensitive information through forms of communication often with malicious links and attachments
- Phishing "accounts for 90% of all data breaches,"(Cisco Umbrella, 2021).

Why make this prediction?

- Prevent users and businesses from losing millions of dollars
- Prevent users from having personal information leaked
- Prevent corporations from having proprietary information leaked
- Worst case scenario: Prevent attacks looking to collapse the economy

Objectives:

- ❑ Build predictive machine learning models.
- ❑ Build and optimize various predictive multi-layer perceptron models
- ❑ Assess and compare the performance of the models

FIU

Dataset

Features

Target

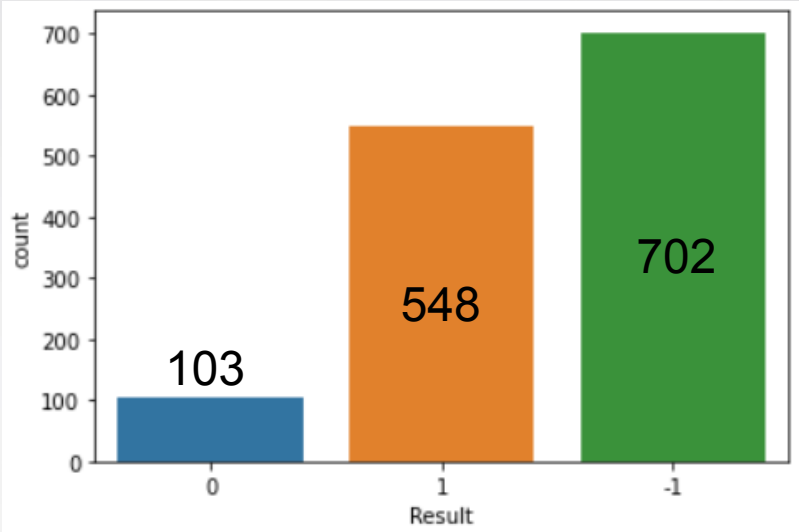
	SFH	popUpWidnow	SSLfinal_State	Request_URL	URL_of_Anchor	web_traffic	URL_Length	age_of_domain	having_IP_Address	Result
0	1	-1	1	-1	-1	1	1	1	0	0
1	-1	-1	-1	-1	-1	0	1	1	1	1
2	1	-1	0	0	-1	0	-1	1	0	1
3	1	0	1	-1	-1	0	1	1	0	0
4	-1	-1	1	-1	0	0	-1	1	0	1
...	...	...	...	...	...	...	...	...	...	...
1348	-1	-1	-1	-1	-1	-1	0	1	0	1
1349	-1	0	1	0	-1	0	0	1	0	-1
1350	-1	0	-1	-1	-1	0	-1	-1	0	1
1351	0	0	1	0	0	0	-1	1	0	1
1352	1	0	1	1	1	0	-1	-1	0	-1

Classification Problem:

- 1: Legitimate
- 0: Suspicious
- -1: Phishy

FIU

Dataset



How was the data found?

- A PHP web script was plugged to a browser
- Legitimate website were found on Yahoo
- Phishy websites were found on Phishtank

FIU

Pre-Processing

Classical Machine Learning Algorithms

- Random Forest Classifier

- The data can be fed directly as there are no NaN or Null values

```
Checking for NaN Values:
SFH                   False
popUpWidnow           False
SSLfinal_State        False
Request_URL           False
URL_of_Anchor         False
web_traffic           False
URL_Length            False
age_of_domain         False
having_IP_Address     False
Result                False
dtype: bool
```

```
Checking for Null Values:
SFH                   False
popUpWidnow           False
SSLfinal_State        False
Request_URL           False
URL_of_Anchor         False
web_traffic           False
URL_Length            False
age_of_domain         False
having_IP_Address     False
Result                False
dtype: bool
```

FIU

Pre-Processing

Deep Learning Algorithms

- Multi-Layer Perceptron
- Different architectures were tested to increase accuracy

- Data needed to be **One-Hot Encoded**
- Get dummies from pandas was used

	SFH_-1	SFH_0	SFH_1	popUpWidnow_-1	popUpWidnow_0	popUpWidnow_1	SSLfinal_State_-1	SSLfinal_State_0	SSLfinal_State_1	Request_URL_-1	...	web_traffic_-1	web_traffic_0	web_traffic_1	URL_Length_-1	URL_Length_0	URL_Length_1	age_of_domain_-1
866	1	0	0	1	0	0	1	0	0	1	...	0	0	1	0	1	0	1
976	0	1	0	1	0	0	1	0	0	0	...	1	0	0	1	0	0	0
738	0	0	1	0	0	1	1	0	0	0	...	1	0	0	0	0	1	0
1235	0	0	1	0	1	0	0	0	1	1	...	0	1	0	0	0	1	1
246	0	1	0	1	0	0	1	0	0	0	...	1	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
898	0	0	1	0	0	1	0	0	1	1	...	0	1	0	0	0	1	0
78	0	0	1	0	1	0	0	0	1	0	...	0	1	0	1	0	0	0
404	1	0	0	1	0	0	1	0	0	1	...	0	0	1	0	1	0	0
600	0	0	1	0	1	0	0	0	1	0	...	0	0	1	1	0	0	0
1060	1	0	0	1	0	0	0	0	1	1	...	0	1	0	0	1	0	1

1014 rows × 25 columns

	-1	0	1
866	0	0	1
976	0	0	1
738	1	0	0
1235	1	0	0
246	0	0	1
...	...	...	...
898	1	0	0
78	1	0	0
404	0	0	1
600	1	0	0
1060	0	0	1

Model Building and Result Analysis

Random Forest Classifier

- Classification reports for different values of n-estimators were tested

n = 1

	precision	recall	f1-score	support
-1	0.91	0.92	0.91	169
0	0.87	0.84	0.85	31
1	0.89	0.89	0.89	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 50

	precision	recall	f1-score	support
-1	0.91	0.92	0.92	169
0	0.87	0.84	0.85	31
1	0.90	0.89	0.90	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 100

	precision	recall	f1-score	support
-1	0.91	0.92	0.92	169
0	0.87	0.84	0.85	31
1	0.90	0.89	0.90	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 200

	precision	recall	f1-score	support
-1	0.91	0.92	0.92	169
0	0.89	0.81	0.85	31
1	0.89	0.90	0.90	139
accuracy			0.90	339
macro avg	0.90	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 400

	precision	recall	f1-score	support
-1	0.91	0.92	0.92	169
0	0.87	0.84	0.85	31
1	0.90	0.89	0.90	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 500

	precision	recall	f1-score	support
-1	0.91	0.92	0.92	169
0	0.87	0.84	0.85	31
1	0.90	0.89	0.90	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 800

	precision	recall	f1-score	support
-1	0.91	0.92	0.91	169
0	0.87	0.84	0.85	31
1	0.89	0.89	0.89	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

n = 1000

	precision	recall	f1-score	support
-1	0.91	0.92	0.91	169
0	0.87	0.84	0.85	31
1	0.89	0.89	0.89	139
accuracy			0.90	339
macro avg	0.89	0.88	0.89	339
weighted avg	0.90	0.90	0.90	339

Objectives:

- While there were some differences in some precisions, recalls, and f1-scores, their averages were virtually identical
- Accuracy was the same for all variations of n-estimators
- 90% accuracy was achieved with Random Forest Classifier

Model Building and Result Analysis

MLP 1 :

- 2 Dense Layers with hidden units set at 64
- 1 Dense Layer with hidden units set at 3
- 2 Dropout Layers of 0.2
- Activation Function: --
- Batch Size: 32
- Optimizer: RMS Prop
- Learning Rate: 0.0005
- Epochs: 20

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_42 (Dense)             (None, 64)                1664

dropout_30 (Dropout)         (None, 64)                0

dense_43 (Dense)             (None, 64)                4160

dropout_31 (Dropout)         (None, 64)                0

dense_44 (Dense)             (None, 3)                 195

=================================================================
Total params: 6,019
Trainable params: 6,019
Non-trainable params: 0
_________________________________________________________________
```

Activation Function	Accuracy
ReLu	88.8%
Sigmoid	85%
Tanh	87.6%

Results:

✓ ReLu was the best performing activation function

✓ Did not outperform Random Forest Classifier

Model Building and Result Analysis

MLP 2 :

- 2 Dense Layers with hidden units set at 100
- 1 Dense Layer with hidden units set at 3
- 2 Dropout Layers of 0.2
- Activation Function: ReLu
- Batch Size: 50
- Optimizer: RMS Prop
- Learning Rate: 0.0005
- Epochs: 20

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 100)	2600
dropout_2 (Dropout)	(None, 100)	0
dense_4 (Dense)	(None, 100)	10100
dropout_3 (Dropout)	(None, 100)	0
dense_5 (Dense)	(None, 3)	303

Total params: 13,003
Trainable params: 13,003
Non-trainable params: 0

Accuracy: 89.1%

MLP Architecture	Accuracy
MLP 1	88.8%
MLP 2	89.1%

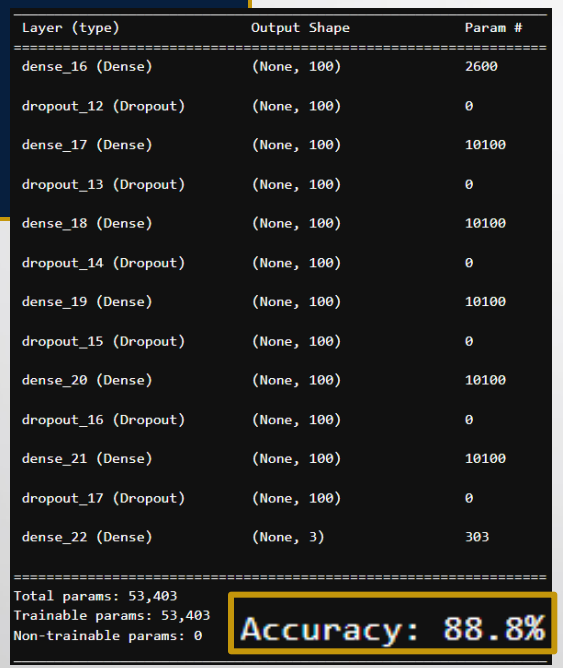
Results:

- ✓ MLP 2 outperformed MLP 1
- ✓ Increasing the units for the dense layer improved performance
- ✓ MLP 2 did not outperform Random Forest Classifier

Model Building and Result Analysis

MLP 3 :

- 6 Dense Layers with hidden units set at 100
- 1 Dense Layer with hidden units set at 3
- 2 Dropout Layers of 0.2
- Activation Function: ReLu
- Batch Size: 32
- Optimizer: RMS Prop
- Learning Rate: 0.0005
- Epochs: 20



Layer (type)	Output Shape	Param #
dense_16 (Dense)	(None, 100)	2600
dropout_12 (Dropout)	(None, 100)	0
dense_17 (Dense)	(None, 100)	10100
dropout_13 (Dropout)	(None, 100)	0
dense_18 (Dense)	(None, 100)	10100
dropout_14 (Dropout)	(None, 100)	0
dense_19 (Dense)	(None, 100)	10100
dropout_15 (Dropout)	(None, 100)	0
dense_20 (Dense)	(None, 100)	10100
dropout_16 (Dropout)	(None, 100)	0
dense_21 (Dense)	(None, 100)	10100
dropout_17 (Dropout)	(None, 100)	0
dense_22 (Dense)	(None, 3)	303

Total params: 53,403
Trainable params: 53,403
Non-trainable params: 0

Accuracy: 88.8%

MLP Architecture	Accuracy
MLP 1	88.8%
MLP 2	89.1%
MLP 3	88.8%

Results:

- ✓ MLP 3 performed worse that MLP 2
- ✓ 6 was too many layers, thus data was overfitted
- ✓ MLP 2 did not outperform Random Forest Classifier

Model Building and Result Analysis

MLP 4 :

- 4 Dense Layers with hidden units set at 200
- 1 Dense Layer with hidden units set at 3
- 2 Dropout Layers of 0.2
- Activation Function: ReLu
- Batch Size: 64
- Optimizer: RMS Prop
- Learning Rate: 0.0005
- Epochs: 20

Layer (type)	Output Shape	Param #
dense_11 (Dense)	(None, 200)	5200
dropout_8 (Dropout)	(None, 200)	0
dense_12 (Dense)	(None, 200)	40200
dropout_9 (Dropout)	(None, 200)	0
dense_13 (Dense)	(None, 200)	40200
dropout_10 (Dropout)	(None, 200)	0
dense_14 (Dense)	(None, 200)	40200
dropout_11 (Dropout)	(None, 200)	0
dense_15 (Dense)	(None, 3)	603

Total params: 126,403
Trainable params: 126,403
Non-trainable params: 0

Accuracy: 89.7%

MLP Architecture	Accuracy
MLP 1	88.8%
MLP 2	89.1%
MLP 3	88.8%
MLP 4	89.7%

Results:

- ✓ MLP 4 outperformed all other MLPs
- ✓ 4 layers performed better than 2 and 6 layers
- ✓ MLP 2 did not outperform Random Forest Classifier

Model Building and Result Analysis

MLP 4 :

- 4 Dense Layers with hidden units set at 200
- 1 Dense Layer with hidden units set at 3
- 2 Dropout Layers of 0.2
- Activation Function: ReLu
- Batch Size: 64
- Optimizer: --
- Learning Rate: 0.0005
- Epochs: 20

Layer (type)	Output Shape	Param #
dense_11 (Dense)	(None, 200)	5200
dropout_8 (Dropout)	(None, 200)	0
dense_12 (Dense)	(None, 200)	40200
dropout_9 (Dropout)	(None, 200)	0
dense_13 (Dense)	(None, 200)	40200
dropout_10 (Dropout)	(None, 200)	0
dense_14 (Dense)	(None, 200)	40200
dropout_11 (Dropout)	(None, 200)	0
dense_15 (Dense)	(None, 3)	603

Total params: 126,403
Trainable params: 126,403
Non-trainable params: 0

Optimizer	Accuracy
RMS Prop	89.7%
Adam	91.4%
SGD	91.7%
Adagrad	90.9%

Results:

- ✓ MLP 4 outperformed all other MLPs and Random Forest
- ✓ Optimizer ‘Stochastic Gradient Descent’ performed the best

Final Result

MLP 4 :

- 4 Dense Layers with hidden units set at 200
- 1 Dense Layer with hidden units set at 3
- 2 Dropout Layers of 0.2
- Activation Function: ReLu
- Batch Size: 64
- Optimizer: SGD
- Learning Rate: 0.0005
- Epochs: 20

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_11 (Dense)             (None, 200)               5200

dropout_8 (Dropout)          (None, 200)               0

dense_12 (Dense)             (None, 200)               40200

dropout_9 (Dropout)          (None, 200)               0

dense_13 (Dense)             (None, 200)               40200

dropout_10 (Dropout)         (None, 200)               0

dense_14 (Dense)             (None, 200)               40200

dropout_11 (Dropout)         (None, 200)               0

dense_15 (Dense)             (None, 3)                 603

=================================================================
Total params: 126,403
Trainable params: 126,403
Non-trainable params: 0
_________________________________________________________________
```

MLP Architecture	Accuracy
Random Forest	90%
MLP 1	88.8%
MLP 2	89.1%
MLP 3	88.8%
MLP 4 (RMS Prop)	89.7%
MLP 4 (Adam)	91.4%
MLP 4 (SGD)	91.7%
MLP 4 (Adagrad)	90.9%

Conclusion

- ✓ Website legitimacy statuses were able to be predicted with over 90% accuracy
- ✓ MLP (Deep Learning Algorithm) outperformed Random Forest (Classical Machine Learning Algorithm)
- ✓ Performance Boosters for this Application:
  - 5 Total Dense layers favoring higher values for hidden units (200 in this case)
  - ReLu Activation Function
  - SGD Optimizer