Applied Machine Learning

Suspicious URLs Detection

AGENDA

Suspicious URLs Case-Study

Data Understanding

Analysis Approach

Predictive Model Development

Suspicious URLs Case-Study

Suspicious URLs Detection

Essential steps towards secure web browsing

Business Perspective

Phishing emails, malicious URLs, sending fake text messages, and other methods are frequently used to carry out online illegal practices

it is essential to identify malicious URLs on the internet and add them to a blacklist to prevent user attacks

Running multiple classification ML models in order to achieve the best prediction to identify Suspicious URLs

Case-Study

This URL identification is a multi-classification problem, and this can be done by using various algorithms that are present in Machine Learning. We will be focusing on multiple Classification Algorithms. We will report the accuracies on each model and choose the best one for predictions.

Business Challenges

Traditional solutions are lacking stability in identifying and dealing with the un-ethical web browsing activities

The cost of gathering the URL data very high and requires a lot of compliance clearance

Technical Challenges

In-sufficient Labeled data The labeled data is not in abundance and identifying key variables is lacking

Lack of centralized analytical platform that can generate the required data for URL detection and provide results instantly

Data Understanding

Data Set Description

Mean values of the cell area

Lexical features can be considered as the textual properties of an URL like QueryLength, DomainLength, URL LetterCount, Length of the HostName etc. Lexical Features are lightweight in nature so, it takes less time for computation and due to its lightweight property, it is popular in the field of Machine Learning [15]. The Lexical features are extracted from an URL and it does not depend on any specific application like email, social networking websites etc. Since most of the Malicious or Spam URLs have a short life span, the features that are extracted will remain present and can be utilised to detect new incoming Malicious URLs even when the old Malicious URLs are unavailable

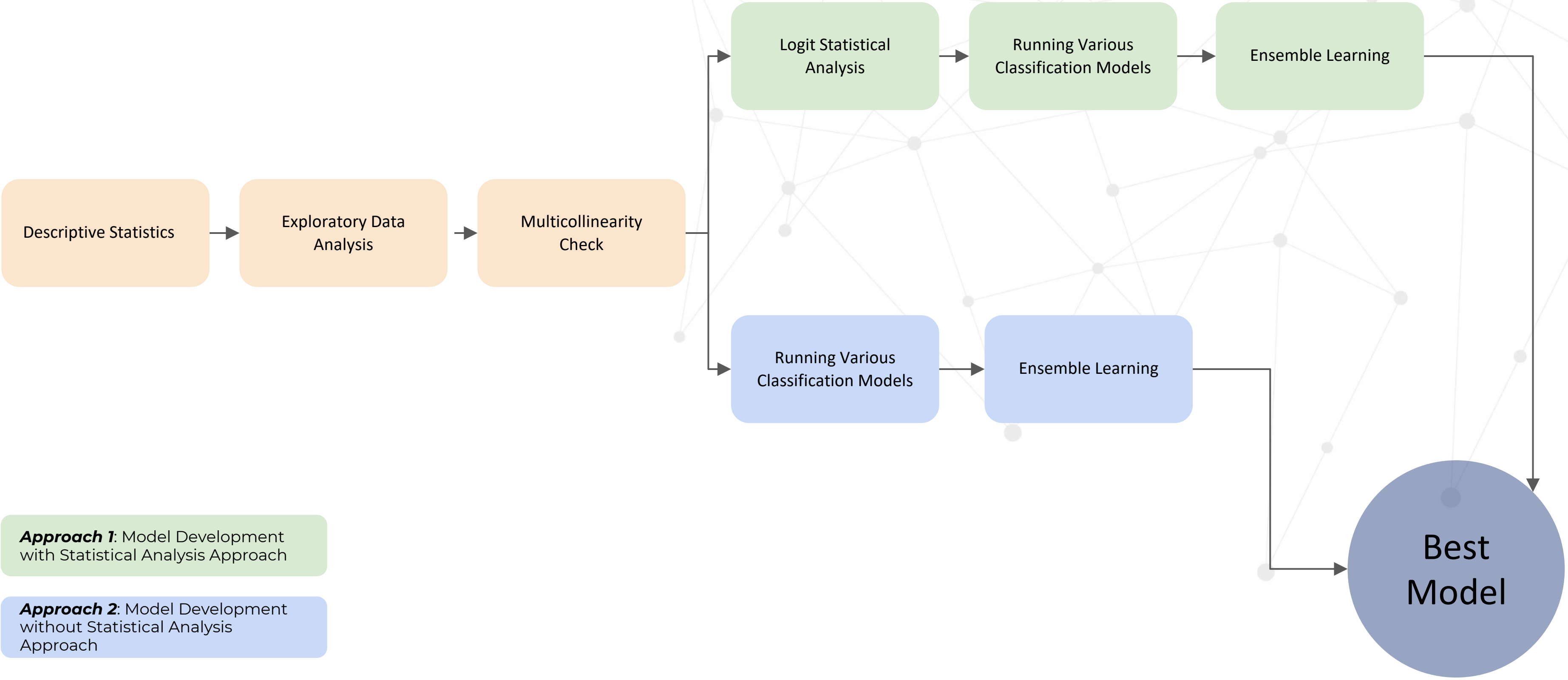
Lexical Features	Lexical Features
Query length	Directory DigitCount
Domain token count	File name DigitCount
Path token count	Extension DigitCount
Avgdomaintokenlen	Query DigitCount
Longdomaintokenlen	URL LetterCount
Avgpathtokenlen	Host LetterCount
Tld	Directory LetterCount
Charcompvowels	Filename LetterCount
Charcompace	Extension LetterCount
Ldl url	Query LetterCount
Ldl domain	LongestPathTokenLe ngth
Ldl path	Domain LongestWordLength
Ldl filename	Path LongestWordLength
Ldl getArg	Sub Directory LongestWordLength
Dld url	Arguments LongestWordLength
Dld domain	URL sensitiveWord
Dld path	URLQueries variable
Dld filename	SpcharUrl
Dld getArg	Delimeter Domain
UrlLen	Delimeter path
Domainlength	Delimeter Count

Lexical Features	Lexical Features
PathLength	NumberRate URL
SubDirLen	NumberRate Domain
FileNameLen	NumberRate DirectoryName
This.fileExtLen	NumberRate FileName
ArgLen	NumberRate Extension
PathurlRatio	NumberRate AfterPath
ArgUrlRatio	SymbolCount URL
ArgDomanRatio	SymbolCount Domain
DomainUrlRatio	SymbolCount Directoryname
PathDomainRatio	SymbolCount FileName
ArgPathRatio	SymbolCount Extension
Executable	SymbolCount Afterpath
IsPortEighty	Entropy URL
NumberofDotsinURL	Entropy Domain
ISIpAddressInDomain Name	Entropy DirectoryName
CharacterContinuityR ate	Entropy Filename
LongestVariableValue	Entropy Extension
URL DigitCount	Entropy Afterpath
Host DigitCount	URL Type (Output)

Analysis Approach

Model Development Approach

Stage wise approach to build the best model



Approach 1: Model Development with Statistical Analysis Approach

Approach 2: Model Development without Statistical Analysis Approach

Predictive Model Development

Model Goals & Objectives

Setting up the goals for model development

Correctly classifying Target Variable: URL Type: Benign (0), Defacement (1), Phishing (2), Malware (3), Spam (4)

Developing Classification with minimum features

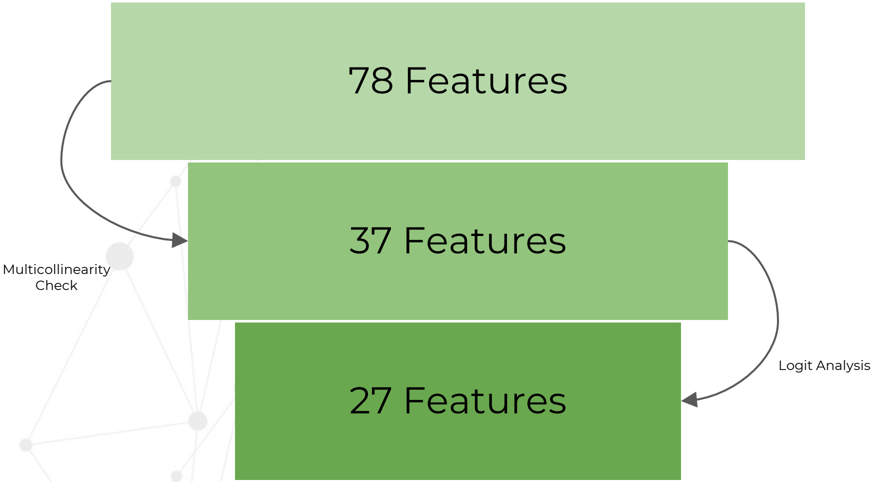
Identifying a ML model with the best accuracy

Determine the features that are highly associated with Non-Benign URLs

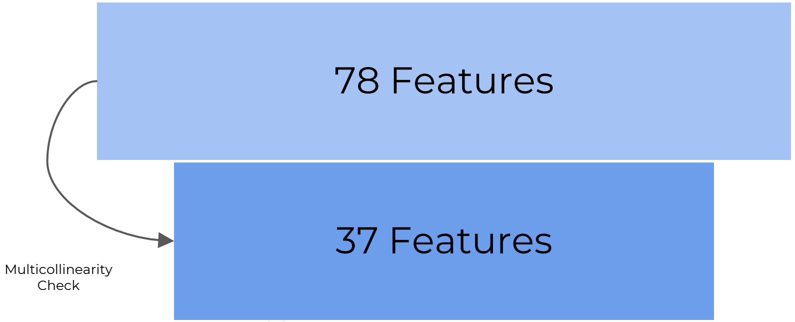
Feature Engineering

Approach 1 vs Approach 2

Approach 1



Approach 2



Approach 1 ML Model Summary

Comparing all the model performance

KNN

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	7818
1	0.93	0.94	0.93	1957
2	0.91	0.83	0.86	1906
3	0.94	0.94	0.94	1640
4	0.96	0.92	0.94	1692
accuracy			0.95	15013
macro avg	0.94	0.93	0.93	15013
weighted avg	0.95	0.95	0.95	15013

Hyperparameters: K = 3

Accuracy:
95.23%

Decision Tree

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.95	0.90	6190
1	0.78	0.51	0.62	1608
2	0.61	0.67	0.64	1523
3	0.40	0.18	0.25	1301
4	0.59	0.74	0.66	1388
accuracy			0.75	12010
macro avg	0.65	0.61	0.61	12010
weighted avg	0.73	0.75	0.73	12010

Hyperparameters:
max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_split': 2

Accuracy:
74.92%

Logistics Regression Multinomial Model

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.96	0.92	6190
1	0.73	0.74	0.73	1608
2	0.69	0.72	0.70	1523
3	0.64	0.39	0.48	1301
4	0.81	0.75	0.78	1388
accuracy			0.81	12010
macro avg	0.75	0.71	0.72	12010
weighted avg	0.80	0.81	0.80	12010

Hyperparameters: C: 0.1, Penalty: L2

Accuracy:
81.27%

Approach 1 ML Model Summary

Comparing all the model performance

Naive Bayes

Classification Report :

	precision	recall	f1-score	support
0	0.86	0.82	0.84	6190
1	0.42	0.85	0.56	1608
2	0.67	0.23	0.34	1523
3	0.66	0.16	0.25	1301
4	0.42	0.61	0.50	1388
accuracy			0.65	12010
macro avg	0.61	0.53	0.50	12010
weighted avg	0.70	0.65	0.64	12010

Hyperparameters: priors: None
'var_smoothing': 1e-09

Accuracy:
65.2%

Support Vector Machine

	precision	recall	f1-score	support
0	0.82	1.00	0.90	6190
1	0.99	0.78	0.87	1608
2	0.97	0.51	0.67	1523
3	1.00	0.92	0.96	1301
4	1.00	0.84	0.92	1388
accuracy			0.88	12010
macro avg	0.95	0.81	0.86	12010
weighted avg	0.90	0.88	0.87	12010

Hyperparameters: C=0.1, gamma=0.01,
kernel=rbf

Accuracy:
88%

Approach 1 ML Model Summary

Comparing all the model performance

Ensemble Learning: Random Forest

	precision	recall	f1-score	support
0	0.92	0.99	0.96	6190
1	0.95	0.86	0.90	1608
2	0.81	0.83	0.82	1523
3	0.97	0.73	0.84	1301
4	0.95	0.93	0.94	1388
accuracy			0.92	12010
macro avg	0.92	0.87	0.89	12010
weighted avg	0.92	0.92	0.92	12010

Hyperparameters: {'criterion': 'entropy',
'max_depth': 8,
'max_features': 'auto',
'n_estimators': 500}

Accuracy: 91.56%

Ensemble Learning: Voting Classifier

	precision	recall	f1-score	support
0	0.90	1.00	0.94	6190
1	0.96	0.90	0.93	1608
2	0.93	0.73	0.82	1523
3	0.99	0.90	0.95	1301
4	1.00	0.88	0.94	1388
accuracy			0.93	12010
macro avg	0.96	0.88	0.91	12010
weighted avg	0.93	0.93	0.93	12010

Hyperparameters: ('lr1', regularized_lr), ('svc', rbf), ('knn', 3)

Accuracy: 93%

Approach 2 ML Model Summary

Comparing all the model performance

KNN

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	7818
1	0.93	0.94	0.93	1957
2	0.91	0.82	0.87	1906
3	0.95	0.94	0.95	1640
4	0.96	0.93	0.95	1692
accuracy			0.95	15013
macro avg	0.94	0.93	0.93	15013
weighted avg	0.95	0.95	0.95	15013

Accuracy: 0.9531739159395191

Hyperparameters: K = 3

Accuracy:
93.75%

Decision Tree

	precision	recall	f1-score	support
0	0.85	0.96	0.90	6190
1	0.75	0.60	0.67	1608
2	0.61	0.69	0.65	1523
3	0.84	0.18	0.29	1301
4	0.59	0.77	0.67	1388
accuracy			0.77	12010
macro avg	0.73	0.64	0.64	12010
weighted avg	0.78	0.77	0.75	12010

Hyperparameters:
max_depth': 5, 'max_leaf_nodes': 10,
'min_samples_split': 2

Accuracy:
77%

Logistics Regression Multinomial Model

	precision	recall	f1-score	support
0	0.90	0.97	0.93	6190
1	0.79	0.75	0.77	1608
2	0.70	0.76	0.73	1523
3	0.75	0.47	0.58	1301
4	0.82	0.80	0.81	1388
accuracy			0.84	12010
macro avg	0.79	0.75	0.76	12010
weighted avg	0.83	0.84	0.83	12010

Hyperparameters: C: 0.1, Penalty: L2

Accuracy:
84%

Approach 2 ML Model Summary

Comparing all the model performance

Naive Bayes

Classification Report :

	precision	recall	f1-score	support
0	0.85	0.85	0.85	6190
1	0.46	0.82	0.59	1608
2	0.78	0.21	0.33	1523
3	0.65	0.34	0.45	1301
4	0.53	0.74	0.62	1388
accuracy			0.69	12010
macro avg	0.65	0.59	0.57	12010
weighted avg	0.73	0.69	0.68	12010

Hyperparameters: priors: None
'var_smoothing': 1e-09

Accuracy:
69%

Support Vector Machine

	precision	recall	f1-score	support
0	0.76	1.00	0.86	6190
1	1.00	0.69	0.81	1608
2	0.96	0.40	0.56	1523
3	1.00	0.90	0.95	1301
4	1.00	0.69	0.82	1388
accuracy			0.84	12010
macro avg	0.94	0.74	0.80	12010
weighted avg	0.87	0.84	0.82	12010

Hyperparameters: C=0.1, gamma=0.01,
kernel=rbf

Accuracy:
84%

Approach 2 ML Model Summary

Comparing all the model performance

Ensemble Learning: Random Forest

	precision	recall	f1-score	support
0	0.93	0.99	0.96	6190
1	0.96	0.86	0.91	1608
2	0.80	0.85	0.83	1523
3	0.98	0.75	0.85	1301
4	0.95	0.92	0.93	1388
accuracy			0.92	12010
macro avg	0.92	0.88	0.90	12010
weighted avg	0.93	0.92	0.92	12010

Hyperparameters: {'criterion': 'entropy',
'max_depth': 8,
'max_features': 'auto',
'n_estimators': 500}

Accuracy: 92%

Ensemble Learning: Voting Classifier

	precision	recall	f1-score	support
0	0.89	1.00	0.94	6190
1	0.98	0.87	0.92	1608
2	0.93	0.72	0.81	1523
3	0.99	0.90	0.94	1301
4	0.99	0.88	0.93	1388
accuracy			0.92	12010
macro avg	0.95	0.87	0.91	12010
weighted avg	0.93	0.92	0.92	12010

Hyperparameters: ('lr1', regularized_lr), ('svc', rbf), ('knn', knn)

Accuracy: 92%

Model Recommendation

Predicting malignant and benign

KNN or Voting Classifier

- The overall accuracy between 93% to 95%
- The number of independent variables required for predicting accurately is less, just 27 out of 78 variables
- Voting Classifier gives a boost to weak models and give combined results with better accuracy

Thank You