MET INSTITUTE OF ENGINEERING, NASHIK

DATA MINING AND WAREHOUSING MINI-PROJECT

REPORT ON

"CLASSIFYING URL DATASETS"

SUBMITTED BY

Sanika_Gaikwad - A_13

Mohini_Deshmukh- A_39

Rutuja_Dhemse - A_42

Under the guidance

of

Mr.Vishal Patil

DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2021-22

Contents

1.Problem Statement

Consider a labeled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

2 .Abstract

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, we can build a classification model to categorize whether the URL applications as either safe or risky . Such analysis can help provide us with a better understanding of the data at large. In this project we use multiple classification models to analyze the outcome of URL based on Various Categories . Use suitable data preprocessing steps. We then compare performance of classification models to find which one is the best.

3. INTRODUCTION

We have been provided with the data regarding aspects of URL . The Data fields are
Based on their Length of URL, Presence Of Hyphens, their subdomain, Length Of Domain, Suspicious Activities in domain, their IP Address, etc.

Train Set Contains Various Records and then we predict or classify Normal or Phishing URL depending on their Length, Normal or Malicious URL based on their Dots.
We Used Various Classification Models:

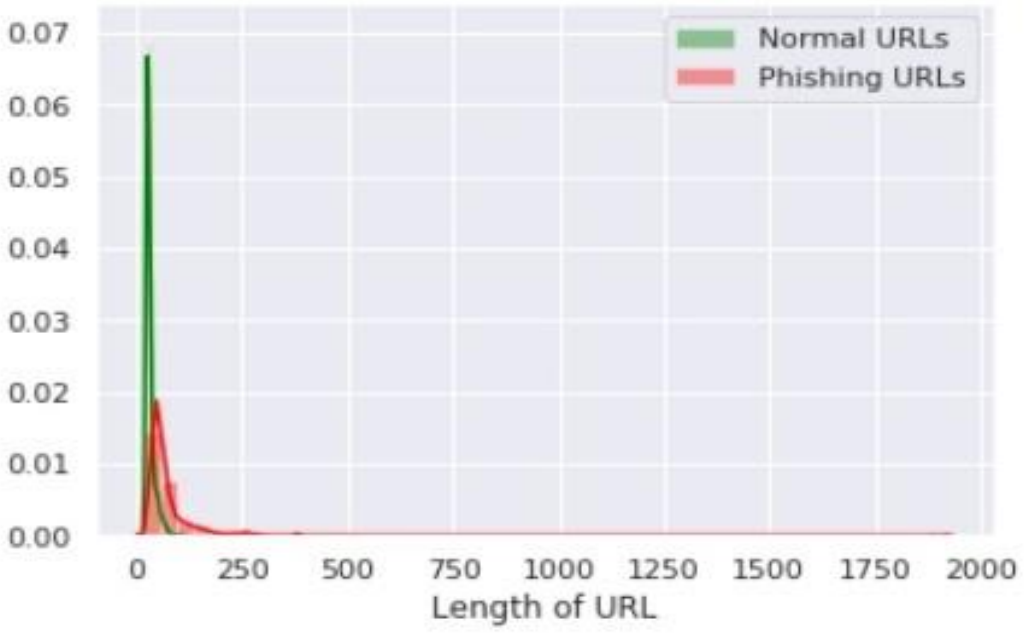
- Logistic Regression
- Gaussian Naive Bayes
- Decision Tree Classifier
- Random Forest Classifier
- AdaBoostClassifier

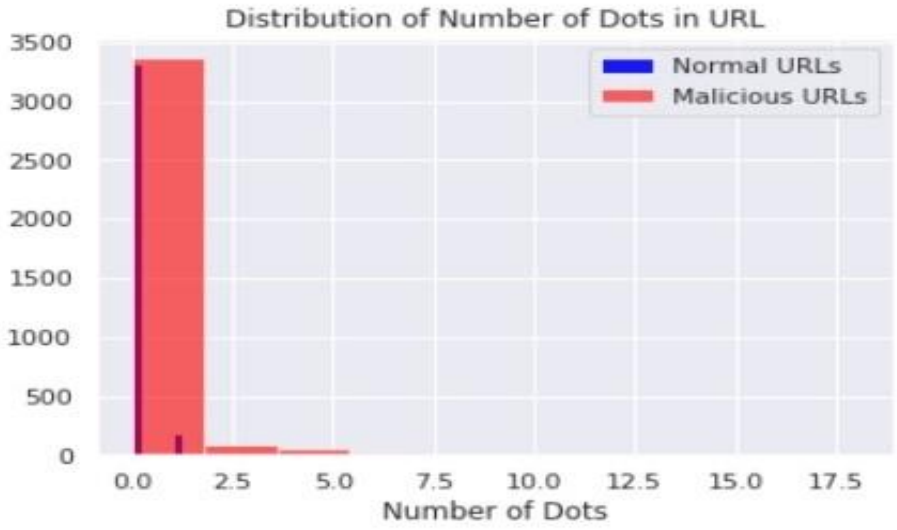
4.Objective:

- To understand data preprocessing .
- To perform classification on dataset and predict Cross Validation Score Of test datasets.

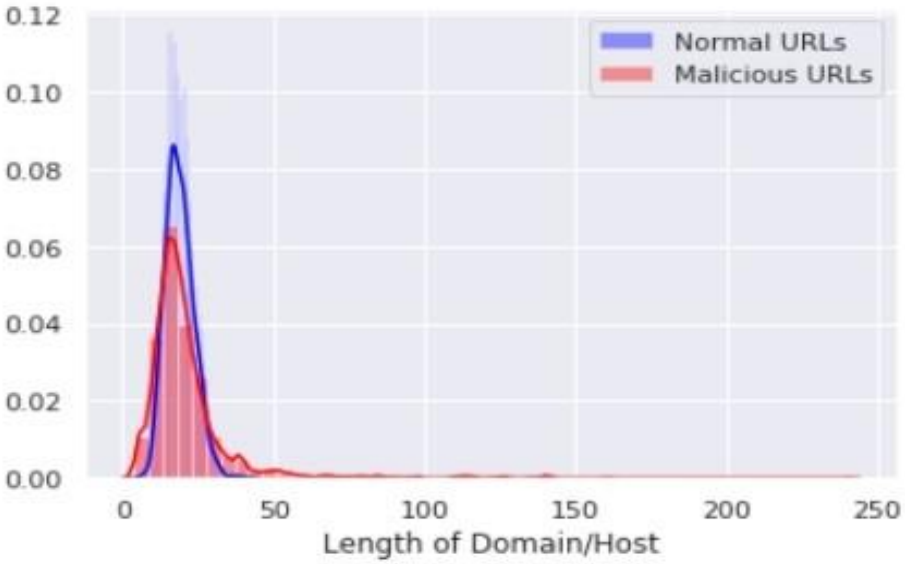
5.Test Cases:

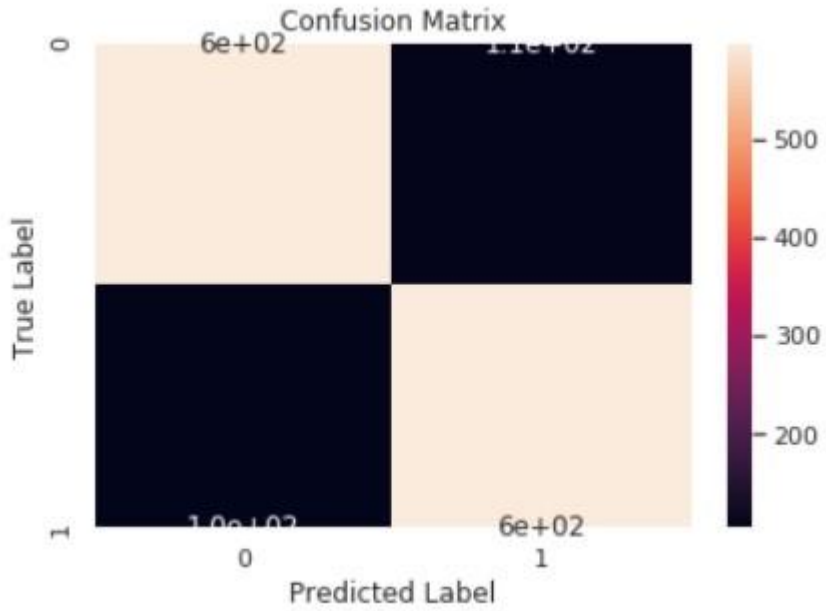
	url	no of dots	presence of hyphen	len of url	presence of at	presence of double slash	no of subdir	no of subdomain	len of domain	no of queries	is IP	presence of Suspicious_TLD	presence of suspicious domain	labe
	erical-methods.com/	0	1	33	0	0	1	1	25	0	0	0	0	
	v.eshinejewelry.com/	0	0	29	0	0	1	1	21	0	0	0	0	
	ıonroephoto.com/w/	0	0	34	0	0	2	1	24	0	0	0	0	
	http://bit.ly/2iffhNV	0	0	21	0	0	1	0	6	0	0	0	0	
	w.tableandvine.com/	0	0	28	0	0	1	1	20	0	0	0	0	




Distribution of Number of Dots in URL
Normal URLs
Malicious URLs
3500
3000
2500
2000
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Dots

[80]:

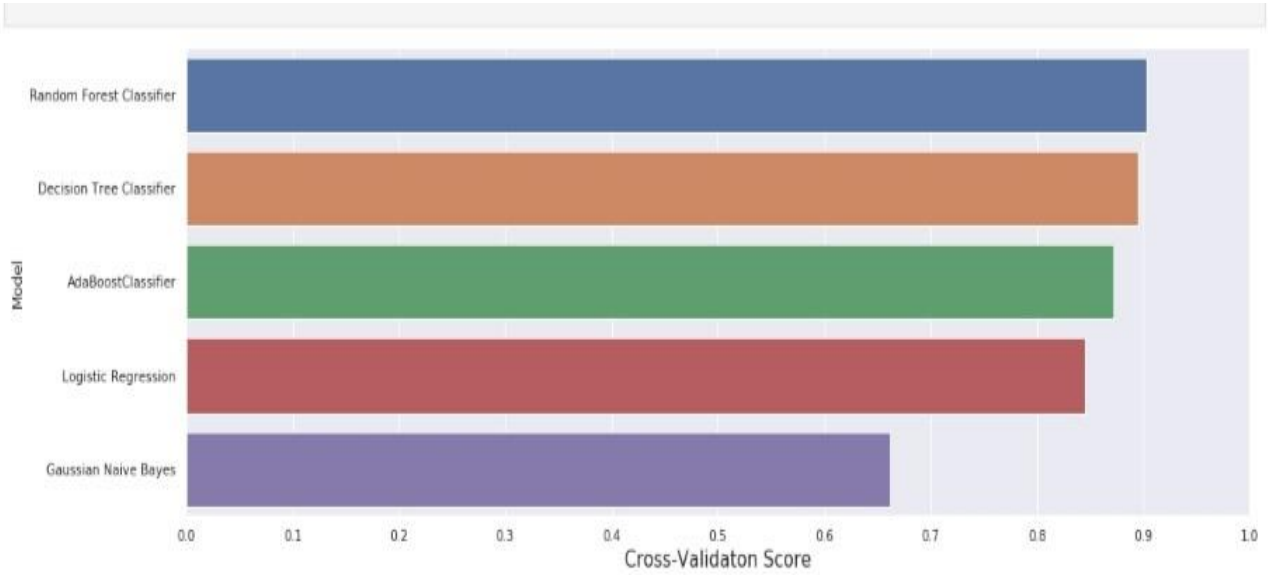

Normal URLs
Malicious URLs
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
Length of Domain/Host



2...

	Model	True Positive	False Positive	True Negative	False Negative	Accuracy	Cross-Validation
0	Logistic Regression	596	108	597	105	0.848506	0.845310
1	Gaussian Naive Bayes	695	9	293	409	0.702703	0.662166
2	Decision Tree Classifier	634	70	632	70	0.901138	0.895443
3	Random Forest Classifier	631	73	639	63	0.903272	0.903272
4	AdaBoostClassifier	621	83	630	72	0.889758	0.872333

6.Result:


Random Forest Classifier
Decision Tree Classifier
AdaBoostClassifier
Logistic Regression
Gaussian Naive Bayes
Model
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Cross-Validaton Score

7. Conclusion:

We have analyzed the URL dataset and performed data pre-processing steps. We have experimented multiple classification models and found out the best performer among them. We have then used this model to make predictions on the test dataset