Primera entrega de proyecto

POR:

Samuel Gil Arboleda

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE

ANTIOQUIA FACULTAD DE

INGENIERÍA

MEDELLÍN 2022

1. Planteamiento del problema

El problema que se está abordando es un problema de clasificación, que tiene como objetivo predecir si un sitio web es considerado malicioso (phishing) o no, teniendo en cuenta características en la URL que normalmente tienen estos sitios. Este problema es de vital importancia debido a que estos sitios son difíciles de reconocer para gente que no tiene conocimientos en tecnología, por lo cual es necesario resolverlo mediante técnicas de aprendizaje de máquina debido a que reconocerlas manualmente podría conllevar mucho tiempo o ser un trabajo imposible, ya que la cantidad de páginas web existentes son prácticamente infinitas..

2. Dataset

El dataset a utilizar proviene de un repositorio de datos de machine learning llamado UCI (https://archive.ics.uci.edu/ml/datasets/phishing%20websites#). La base de datos de este proyecto es “Phishing Websites Data Set” [5], cuenta con 11055 muestras, y 31 variables, las cuales 30 son para explicar las características de una url de un sitio web, y la otra es la variable de salida “Result” la cual se encarga de decirnos si un sitio web es phishing o no.

Las variables presentes en el dataset pueden contener 2 o 3 de los siguientes valores: -1, 0 ó 1 , lo cual significa que cumplen con la condición propuesta en la variable, es sospechosa, y no la cumple, respectivamente. Además no hay datos faltantes en ninguna característica.

#	Nombre variable	Codificación
0	having_IP_Address	(categorical - signed numeric) : { -1,1 }
1	URL_Length	(categorical - signed numeric) : { 1,0,-1 }
2	Shortining_Service	(categorical - signed numeric) : { 1,-1 }
3	having_At_Symbol	(categorical - signed numeric) : { 1,-1 }
4	double_slash_redirecting	(categorical - signed numeric) : { -1,1 }
5	PrefixSuffix-	(categorical - signed numeric) : { -1,1 }
6	having_Sub_Domain	(categorical - signed numeric) : { -1,0,1 }
7	SSLfinal_State	(categorical - signed numeric) : { -1,1,0 }

8	Domain_registeration_length	(categorical - signed numeric) : { -1,1 }
9	Favicon	(categorical - signed numeric) : { 1,-1 }
10	port	(categorical - signed numeric) : { 1,-1 }
11	HTTPS_token	(categorical - signed numeric) : { -1,1 }
12	RequestURL	(categorical - signed numeric) : { 1,-1 }
13	URL_of_Anchor	(categorical - signed numeric) : { -1,0,1 }
14	Links_in_tags	(categorical - signed numeric) : { 1,-1,0 }
15	ServerFormHandler	(categorical - signed numeric) : { -1,1,0 }
16	Submitting_to_email	(categorical - signed numeric) : { -1,1 }
17	AbnormalURL	(categorical - signed numeric) : { -1,1 }
18	Redirect	(categorical - signed numeric) : { 0,1 }
19	on_mouseover	(categorical - signed numeric) : { 1,-1 }
20	RightClick	(categorical - signed numeric) : { 1,-1 }
21	popUpWidnow	(categorical - signed numeric) : { 1,-1 }
22	Iframe	(categorical - signed numeric) : { 1,-1 }
23	AgeofDomain	(categorical - signed numeric) : { -1,1 }
24	DNSRecord	(categorical - signed numeric) : { -1,1 }
25	web_traffic	(categorical - signed numeric) : { -1,0,1 }
26	PageRank	(categorical - signed numeric) : { -1,1 }
27	GoogleIndex	(categorical - signed numeric) : { 1,-1 }
28	LinksPointingToPage	(categorical - signed numeric) : { 1,0,-1 }
29	Statistical_report	(categorical - signed numeric) : { -1,1 }
30	Result	(categorical - signed numeric) : { -1,1 }

3. Métricas

Para evaluar el sistema se van a usar las siguientes métricas de evaluación: accuracy y f1 score. Siendo accuracy la medida principal.

Por otra parte, en cuanto a la métrica de negocio, se tiene interés en que las predicciones sean lo suficientemente confiables para saber si un sitio web es de pishing o no. Con esta información un navegador de internet podría evitar que sus usuarios entren a estos sitios maliciosos solo leyendo la página a la cual se dirigen.

4. Desempeño

Lo que se esperaría de un modelo de este tipo es obtener una predicción con bastante desempeño (más de un 80% de precisión) , porque no sería viable tener muchos falsos positivos, ya que bloquear constantemente las páginas de un usuario de un navegador podría hacer que este deje de usarlo. En un ambiente productivo sería usado como filtro para evitar que los usuarios entren a las páginas que sean sospechosas.

5. Bibliografía

- UCI Machine Learning Repository: Phishing Websites Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/phishingwebsites .