Modelo Predictivo de Phishing

Grupo N°8
Alumnos:

- Lucas Kiriyama
- Juan Cruz Camacho
- Tobias Pucci
- Brian Slavkin

Descripción

En la actualidad la problemática del phishing es uno de los principales métodos para la obtención de información personal y confidencial a partir de un engaño a través de una página web. Mediante el uso de un dataset que reúne características de distintas páginas web clasificadas entre legítimas o engañosas (phishing) buscamos explorar y elegir el mejor modelo predictivo utilizando técnicas estadísticas y de ciencia de datos. Se trata de un problema de clasificación binaria o discriminación supervizado, por lo que tenemos una gran cantidad de modelos para ensayar.

Dataset

Los links adjuntos poseen la descripción del paper que explica la generación del set de datos y el sitio de descarga de los mismos. El dataset original cuenta con una gran cantidad de variables las cuales no nos sirven en su totalidad (por ejemplo, por tener nula varianza), por lo que para poder trabajar con este realizaremos una limpieza de los datos. Dentro de las variables podemos encontrar muchas que son contadores de caracteres del link del dominio web, del URL, del directorio, el tiempo que tardó en activarse el dominio entre otras, las cuales son numéricas. Además de la variable binaria que nos explica si es o no es un caso de phishing.

1. https://www.sciencedirect.com/science/article/pii/S2352340920313202#bib 0004
2. https://data.mendeley.com/datasets/72ptz43s9v/1

Objetivo

Buscamos obtener un modelo que generalice bien antes nuevas observaciones, permitiendo detectar si una página tiene contenido legítimo o si este es un método de phishing web usando la menor información posible detectar.

Este servicio podría usarse como plugins (complementos) en buscadores web y en servicios de email o mensajería para prevenir a los usuarios el acceso a páginas potencialmente peligrosas. Una posibilidad es informar la probabilidad de que dicha página pueda tratarse de una actividad ilegítima a traves de una predicción discreta.

Planteo y diferenciación con lo existente:

Al ser un problema tan directo ya existen análisis realizados con herramientas de Machine Learning para predecir si un url es una estafa de phishing o no. De cualquier manera, no encontramos ninguno realizado con el dataset (ni las variables específicas) que propusimos anteriormente debido a que este se publicó en diciembre de 2020. Además, uno de los factores que nos interesa saber es si es posible detectar con exactitud si un url es phishing o no únicamente utilizando el url, es decir, sin contar con variables que deben generarse de una manera que no sea partiendo del string del url. Si esto es así, podría construirse una simple aplicación que permita ingresar el url del cual uno tenga dudas y nos entregue la probabilidad de que estemos en frente de una estafa.

De ser posible esto, podríamos construir nuestro propio dataset partiendo de links que estamos seguros que son phising (realizando un scraping desde PhishTank > Phish Search) y otros que no lo sean y mejorar el funcionamiento y la exactitud de nuestro modelo.

También consideramos que sería enriquecedor realizar un análisis sobre la interpretación de las variables (por ejemplo partiendo de una regresión logística) para tratar de entender qué caracteres (o palabras, etc) son más frecuentes en urls de phishing y dejar en claro qué patrones pueden detectarse a simple vista para saber que no debemos abrir este link.

Proponemos plantear diferentes modelos de clasificación (vistos en clase y otros nuevos), realizar la comparativa entre ellos y entre los dataset con y sin variables no provenientes del url. Creemos que el trabajo integral como lo planteamos no se encuentra en internet y su realización será enriquecedor para nosotros.