2022

MACHINE LEARNING

Hecho por: Diego Puchol Candel

Descripción del caso de uso

En este documento se hará un seguimiento de forma ordenada al problema que se plantea en el siguiente punto. Se evaluará su posible solución y como se ha enfrentado hasta ahora la empresa ante este problema. Con que criterio se llevará a cabo y lo que se espera una vez implementada esta mejora.

1 ¿Cuál es el problema?

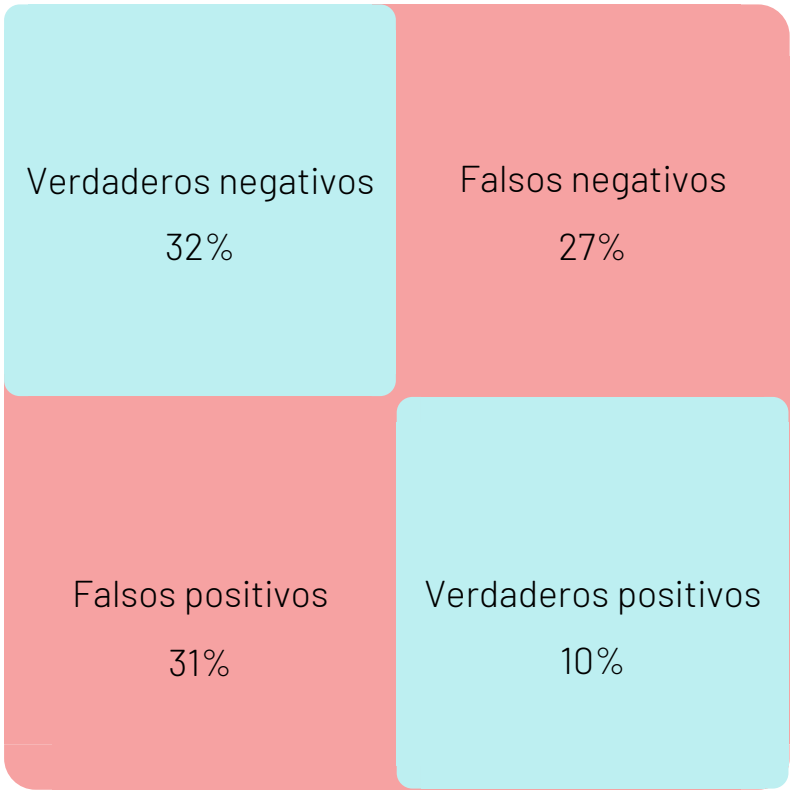
El gran riesgo que se corre en la empresa al no estar protegidos frente a los URL's maliciosos.
Enfrentarnos a un problema como lo sería un malware en cualquier dispositivo de la empresa dependiendo del objetivo del atacante podría generar un gran impacto económico incluso legal si hablamos de llegar a revelarse información confidencial violando el derecho de privacidad de nuestros clientes.

2 ¿Cómo se está afrontando ahora?

Actualmente contamos con el uso de Norton como antivirus.
Diferentes reglas que nos protegen en cierta medida de posibles ataques. Tales como la prohibición de conectar pendrives sin previa autorización a cualquier dispositivo de la empresa.
Prohibido hacer de uso personal cualquier ordenador perteneciente a la compañía.

Matriz de confusión

Falsos negativos: dejó vía libre pensando que no era ataque, cuando realmente se trataba de uno.
Falsos positivos: se envió notificación de ataque cuando no lo era.
Verdaderos positivos: ataques frenados por nuestra seguridad.
Verdaderos negativos: ataques recibidos y sin poder haber sido frenados.



3 Acción que buscamos poder hacer para solucionar el problema.

El objetivo sería combatir el riesgo al que se expone la empresa frente a URL's maliciosos. Intentando cerrar todas las puertas a posibles ataques o infiltraciones, protegiendo así la información de la empresa y todos los datos con los que trata.
El modelo elegido sería un Decision Trees aprovechando que es fácil de entender, fácil de combinar con otros modelos y muy bueno para hacer combinaciones de modelos y teniendo siempre cuidado de no cometer overfitting.

4 KPIs – Indicadores de negocio

Un indicador que nos muestre cuantas URL's maliciosas a detectado en el último mes.
Un indicador que nos muestre cuales de esos URL's detectados realmente eran maliciosos.
Un indicador de ataques recibidos por medio de URL's maliciosas.
Otro que nos informe de los ataques frenados gracias a nuestro detector de URL's maliciosos.

5 ¿Cuáles son los mínimos que se esperan de este caso de uso?

Un mínimo esperado es la poca diferencia entre los resultados de mis dos primeros indicadores. Es importante que nuestro modelo sea lo más exacto posible al momento de detectar URL's maliciosos.
Se espera una incrementación en el número de ataques frenados gracias a nuestro modelo.

6 Validación: ¿Qué criterio se va a usar para decidir si la solución es aceptable?

Se necesita que al ver el indicador de URL's marcados como maliciosos no sea mayor a 27% más a la hora de ver los que realmente eran maliciosos no superen el margen de error de 15%.

7 Experimentación: ¿Cómo vamos a corroborar el funcionamiento?

Para evaluar el funcionamiento de mi modelo de machine learning se usará la métrica de **Recall** que nos ayudará informando sobre la cantidad en este caso de los enlaces maliciosos detectados.
Con una frecuencia mensual para asegurar que todo va bien.

8 Productivización: ¿Qué salida debe tener la solución que se desarrolle?

La salida será presentada en un fichero de texto apoyada por gráficos de barras para hacer más fácil de entender cualquier aspecto a cualquier persona que deba leerlo.

9 Equipo de trabajo

Identificación de personas colaboradoras

Se encargarán un grupo de 3 personas que verificarán y analizarán los enlaces maliciosos detenidos por nuestro modelo. También se encargarán de proteger como equipo de blue team de los ataques realizados por dichos enlaces.

10 Detalle del caso de uso

Detalle funcional

Se tendrá en cuenta en todo el proceso del modelo que el valor de la varianza tendrá que ser más alta que el BIAS, sabiendo que tratamos con un árbol de decisión.
Cualquier URL malicioso detectado tendrá que sumarse al conjunto de URL's maliciosos previamente preparado para usarse como ejemplo de comparación.
El modelo se tendrá que actualizar cada 6 meses con nuevos URL's maliciosos encontrados en la red y así mantener actualizado nuestro modelo con cualquier avance en este tipo de ataques.

11 Identificación de orígenes de datos

Se hará uso de diccionarios para que con combinaciones de palabras se compruebe si los URL's son maliciosos, verificar si la página es "https" y no "http" y por último se pueden crear comparativas con URL's que ya sabemos que son maliciosas.

12 Desarrollo del caso de uso

Puntos intermedios o seguimiento

Cabe recalcar que todos podemos ser victimas de algunos URL's maliciosos por lo que no estaría de más reforzar nuestro modelo haciendo saber a nuestros empleados que hacer a la hora de enfrentarse a uno.

13 Aporte esperado por Big Data

En 2022, el 75% de las empresas de todo el mundo experimentaron ataques de phishing.
Aumentan un 46% los ataques de Phishing exitosos en el Último Año, según el Informe State of the Phish 2022 de Proofpoint.
Por lo que decidimos incrementar la seguridad en todos los aspectos pero sobre todo donde más recibíamos ataques que era por medio de enlaces maliciosos .