Athena (Android Malware Detection)

Mateo Aguirre Duque

Departamento de ingeniería de sistemas

Universidad de Antioquia

Inteligencia Artificial para las Ciencias e Ingenierías

Raul Ramos P.

Abril 23 de 2023

Avances

El avance de estas semanas en el proyecto ha consistido en lograr tener una primera versión de un modelo que me permita darle solución al problema planteado en la entrega anterior. Hasta ahora lo que se ha construido es el primer análisis de los datos tal cual los entrega kaggle, la primera versión del preprocesado de los datos y un primer modelo (random forest classifier) obteniendo unas métricas muy buenas pero que a su vez preocupan.

Exploración de datos

En la exploración de datos se encontró que la variable objetivo es una variable categórica con 4 posibles valores (Android_Adware, Android_Scareware, Android_SMS_Malware y Benign), para el proyecto no es necesario tanta información ya que el objetivo es solo saber si existe o no malware en el dispositivo. Luego de analizar la variable objetivo de forma individual se realizó una revisión de la correlación lineal de todas las variables contra la variable objetivo, en la revisión de estas correlaciones se detectó que ninguna de las variables posee una correlación de interés con la variable objetivo, pero sí poseen correlaciones fuertes entre ellas. Algunas de las variables que más correlaciona tienen son:

- Fwd IAT Min - Flow IAT Max
- Flow Duration - Flow IAT Max
- Bwd Packet Length Max - Max Packet Length

Después de revisar las correlaciones se analizó la cantidad de valores faltantes en el dataset, para mi sorpresa el dataset de kaggle estaba bastante completo y solo 5 registros tenían

problemas de valores faltantes, estos registros tienen múltiples columnas en Nan pero no representaban ni un 1% de los datos.

Preprocesado

Para poder trabajar en el procesado fue necesario realizar cambios en el dataset original ya que este no cuemple el requisito de tener un 5% de valores faltantes, para poder cumplir dicho requisito se optó por crear un método que coge 13 columnas del dataset y reemplaza valores de forma aleatoria para simular los datos faltantes.

El llenado de los datos se realizó con 2 métodos diferentes, uno basado en los cuartiles y la distribución normal, y otro basado únicamente en la distribución normal. Los cuantiles se usan con el fin de quitar posibles valores outliers y mejorar el llenado con valores aleatorios a partir de la distribución normal.

Después del llenado de los datos faltantes se realizó una reducción del problema con el fin de remover las 4 categorías de la variable objetivo y establecer una escala binaria que indica si el registro posee malware o no lo posee.

Modelado

La primera metodología que se implementó fue el random forest classifier el cual se entrenó con 10 estimadores y el 70% de los datos. El modelo se evaluó con las funciones de precision y de recall, las dos medidas dieron como resultado que el modelo está por encima del 90%.

Algo preocupante en este punto es que en esta primera iteración no se realizó una limpieza de datos outliers ni se modificaron parámetros específicos del modelo, esto nos da una idea de que realmente el modelo quedó sobre ajustado y se prendió los datos de memoria, este punto será el de partida de la segunda iteración.

Video

https://youtu.be/qC9Fpeldgpg