PROYECTO MICROSOFT MALWARE

MACHINE LEARNING CANVAS

Autor: Carlos Cabañó Muñoz

Proyecto: MICROSOFT MALWARE Fecha: 15/11/2020 Versión:1.0

NUEVOS DATOS Y REENTRENAMIENTO

Por ahora no se cuenta con datos nuevos para poder reentrenar el modelo, pero se podrían recabar en futuras actualizaciones de la base de datos.

El modelo se reentrenará cada vez que el lote de datos nuevos llegue a las mil muestras.

PREDICCIÓN (ON / OF)

La predicción del modelo se ha realizado teniendo en cuenta diferentes métricas: accuracy, precision/recall y AUC.

Finalmente se presentarán los resultados de la AUC por ser los que tienen un valor más alto.

PROPUESTA DE VALOR

El problema radica en una base de datos con un gran número de variables que, dado su carácter altamente ilegible y críptico, dificulta la comprensión de su relación con el target por métodos normales .

La intención del proyecto es crear un modelo de predicción de detección de virus para ordenadores con sistema operativo Windows.

El objetivo es conocer la probabilidad de detección de virus a través de diversas métricas y observar las variables que mejor expliquen el caso con un árbol de decisión.

ORÍGENES DE DATOS

Tenemos una base de datos con las detecciones de virus en todo el mundo en ordenadores con sistema operativo Windows.

TAREA DE ML

El proyecto se encuadra dentro de la clasificación supervisada, ya que la información está etiquetada, a la que aplicaremos un árbol de decisión por tratarse de un target con información categórica (sí/no).

EVALUACIÓN EN SERVICIO Y ALM

Se seguirá utilizando la métrica AUC con la curva ROC para evaluar el modelo una vez en producción.

Como las versiones de los productos que se definen en las variables explicativas pueden actualizarse con frecuencia, se reentrenará el modelo cada 1000 nuevos datos que vayan entrando.

MÉTRICA DE EVALUACIÓN

La métrica que se ha seguido para la evaluación del modelo es la AUC ejecutada con el algoritmo XGBoost.

El valor mínimo esperado estaría entre el 60% y el 70%, lo que coincidiría con las expectativas para el negocio.

ATRIBUTO

Los atributos disponibles son más de 80 variables que van desde la versión del SO o el procesador, hasta información regional, como el país de procedencia.

DEFINICIÓN DEL PERÍMETRO Y TARGET

Se cuenta con un dataset de 500000 registros con más de 80 variables, entre las que se incluye el target.

El target se ha definido en la variable HasDetections del DF. Esta variable es numérica, pero de carácter booleano (sí/no).

USO DEL MODELO, TOMA DE DECISIONES Y EXPLICABILIDAD

El modelo realizado es prescriptivo, pues trata de predecir la probabilidad de detección de virus dadas ciertas variables. Podrá servir como referencia para implementar la idea de negocio, pero no será excesivamente determinante, dado que la probabilidad de acierto del modelo no llega al 75%.

Los resultados se entregarán en formato Notebook, con las visualizaciones de las variables más importantes y las métricas del modelo debidamente explicadas.