0921 SupML Canvas Michael Woitass

DEFINICIÓN DEL PROBLEMA DE NEGOCIO

PROPUESTA DE VALOR

Definición del problema:

Se trata de indicar la probabilidad de que un Windows PC esté infectado con malware. La predicción se basa en 84 valores facilitados sobre el PC, el resultado consiste en clasificar la máquina en malware detectado o no.

Objetivos:

- Acelerar el diagnóstico de un PC, usando un algoritmo de Machine Learning
- Averiguar si hay indicadores muy significativos en relación al malware
- Encontrar tanto posibles factores que favorecen la existencia de malware como factores que aumentan la protección del equipo
- Sacar conclusiones sobre posibles relaciones entre malware detectado y grupos de indicadores como
  versiones de software,
  - nivel de seguridad del equipo,
  - componentes de hardware,
  - ubicación del PC,
  - tipo de usuario.

USO DEL MODELO, TOMA DE DECISIONES Y EXPLICABILIDAD

El modelo sirve para realizar una predicción sobre la detección de malware en un PC, el resultado consiste en una indicación categórica, sí o no.

Se entrega el resultado con el desglose de las variables que se han usado para realizar la predicción. Estas variables son un subset de los 84 valores que se aplican para medir el equipo. Se requiere conocimiento técnico de ellas para sacar conclusiones sobre acciones mitigantes y posibles causas.

El informe del resultado incluye también información sobre el algoritmo de clasificación aplicado y su rendimiento en términos de accuracy, precision, recall.

HIPÓTESIS DE APRENDIZAJE

ORIGEN DE DATOS

Competición Microsoft Malware Prediction en Kaggle (https://www.kaggle.com/c/microsoft-malwareprediction), muestra de 500.000 registros del dataset original.

Cada fila contiene datos sobre un único PC, los valores de las columnas corresponden a características del software Microsoft Defender.

TAREA DE MACHINE LEARNING

Se trata de un problema de Clasificación Supervisada.
El dataset de origen proporciona atributos así como el target, que contiene valores 1 o 0.

ATRIBUTOS

El dataset contiene 84 atributos. Se usa una selección de estos atributos para el algoritmo de Machine Learning, obteniendo esa selección mediante:
- eliminación de variables sin ninguna relevancia
- anáisis del target, distribución relativa entre valores 1 (malware sí) y 0 (malware no)
- selección de variables significativas en relación al target.

Se determina si una variable es significativa de la siguiente manera:
Se calcula para cada uno de los valores de una variable si la distribución de los registros correspondientes es parecida a la distribución de 1 y 0 en el target. Para ello se aplica un porcentaje de deviación, p.ej. un 3%, para fijar un margen ente la distribución del valor y la distribución del target. Si la distribución del valor supera este margen, se considera la variable significativa, ya que aporta información para la clasificación.

PERÍMETRO Y TARGET

La columna “HasDetections” define el target, todos los registros del dataset tienen uno de los valores 1 o 0 en esta característica.

Se parte el dataset completo en dos conjuntos, uno para desarrollo y uno para validación. El criterio para seleccionar los registros de validación es la versión 10.0.17134.228 del sistema operativa, que es la penúltima revisión del último build de la versión 10 de Windows en el dataset y cubre un 16% del total de registros.

Se particiona el conjunto de desarrollo de forma aleatoria en 70% de registros para entrenar el modelo (train) y 30% para probarlo (test).

VALIDACIÓN Y DEPLOYMENT

PREDICCIÓN

El modelo incluye los datasets del desarrollo y validación, filtrados por una selección de atributos, así como un algoritmo de Clasificación Supervisada.

Opciones de selección de atributos:
- Todas las variables númericas
- Variables seleccionadas manualmente por su nombre que parece relevante para determinar malware
- Variables más significativas según el cálculo del margen en relación al target

Opciones de algoritmo:
- Decision Tree
- Random Forest
- Gradient Boosting

Se realiza la predicción online, y se guardan los resultados en un dataframe que permite comparar las características de cada uno de los modelos.

MÉTRICA DE EVALUACIÓN (EN DESARROLLO)

Se aplican las siguientes métricas:
- Accuracy
- F1 Score
- AUC-ROC Score

Valores mínmos esperados son un valor por encima del 0,60 en Accuracy y F1 Score así como un valor por encima del 0,65 en AUC-ROC Score.

NUEVOS DATOS Y REENTRENAMIENTO

Debido a la dinámica actual en la seguridad informática es recomendable actualizar el dataset original cada tres meses, reentrenar el modelo y comparar la evaluación con las anteriores para mantener o aumentar los ratios obtenidos.

EVALUACIÓN EN SERVICIO Y ALM

Se propone efectuar periódicamente un análisis de la comparativa entre los resultados, con el fin de revisar la selección de variables, del algoritmo y de los scores y de conseguir optimizar el funcionamiento del modelo.