CORRECCIÓN DE OUTLIERS (Anexo 2)

Ciertos algoritmos son sensibles a outliers.

Sensibles porque significa que la presencia o no de este tipo de datos puede influenciar **enormemente** el aprendizaje del algortimo y llevar a predicciones que no se ajusten a la realidad.

Familia	Algoritmo	Sensibilidad
Métodos geométricos	• Regresión lineal • Regresión logística	Si
Métodos de reglas	• Decision Tree • Random Forest • XGBoost	No
Métodos de vecindad	• K-NN • K-Means	Si
Métodos probabilísticos	• Naïve Bayes	No

Fuente: https://arsrinevetha.medium.com/ml-algorithms-sensitivity-towards-outliers-f3862a13c94d

DETECCION DE OUTLIERS

Ver slides
E2E Supervised Machine Learning – S24-S30
apartado 5

CORRECCIÓN DE OUTLIERS (Anexo 2)

Los modelos de la familia de **métodos de reglas y de vecindad son sensibles a los outliers**.

Existen diversos métodos para ajustar y corregir los outliers:

1. **Eliminación de registros:**
   - reconocemos el outlier como un error **random**
   - El número de filas afectadas es relativamente pequeño (<1% of the total dataset)

2. **Tranformación**

   1. **Logaritmica:**
   - Contrae los valores extremos, traiendo los outliers más cerca del grueso de la distribución
   - Aplicando la misma transformación también al target nos aseguramos una relación lineal entre ellas (necesario para la Regresión lineal)
   2. **Scaling** (más adelante en non-supervised methods)
   - Normalizar
   - Escala 0-1

3. **Imputación:** al igual que con los valores nulos, podemos aplicar las mismas técnicas a los outliers

TRANSFORMACIÓN LOGARÍTMICA (base 10)

Utilizaremos la siguiente función matemática.

- X es la variable con outliers que queremos transformar.
- Y es la variable transformada que utilizaremos en su lugar.

$$\mathbf{Log_{10}(x) = y}$$

equivale a

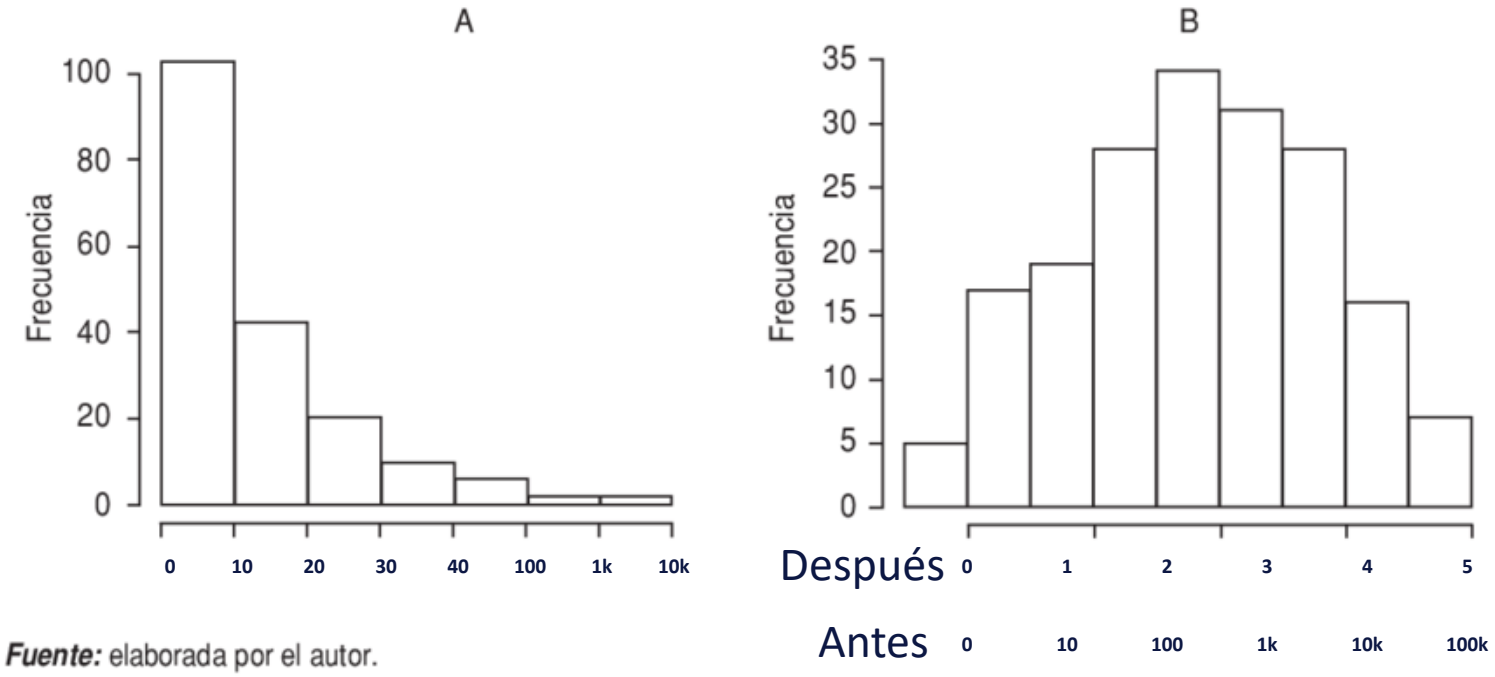
$$10^y = X$$

A que valor debo elevar el 10 para que me de 1 ?
10 elevado a 0 (y) es 1 (x)
A que valor debo elevar el 10 para que me de 10 ?
10 elevado a 1 es 10
A que valor debo elevar el 10 para que me de 100 ?
10 elevado a 2 es 100
A que valor debo elevar el 10 para que me de 1000 ?
10 elevado a 3 es 1000
A que valor (y) debo elevar el 10 para que me de X?

NOTA: si X es igual a 0, el $\log_{10}(0)$ no existe, por esta razón aplicaremos la transformación $\log_{10}(1+x)$

TRANSFORMACIÓN LOGARÍTMICA (base 10)

El efecto de la transformación es el siguiente:



Fuente: elaborada por el autor.

- Los valores extremos se agrupan en el centro de la distribución
- La distribución resultante es parecida a una distribución normal (aunque no exacta)

TRANSFORMACIÓN LOGARÍTMICA (base 10)

En resumen la transformación logarítmica nos ofrece las siguientes ventajas:

1. Transformación de la distribución en una distribución normal (para distribuciones sesgadas)

Lo que permite utilizar ciertas técnicas como el BoxPlot para detectar outliers

2. Eliminación de outliers ya que los trae más al centro de la distribución

Lo que permite utilizar ciertos algortimos como la Regresión lineal con mayor tranquilidad

3. Comparación de distribuciones en el Data Understanding utilizando técnicas de correlación lineal

Lo que facilita el proceso de análisis cuando tenemos muchas variables

NOTA: la transformación logarítmica dificulta la lectura de resultados y la explicación del modelo resultante