

Machine Learning

Preprocesado

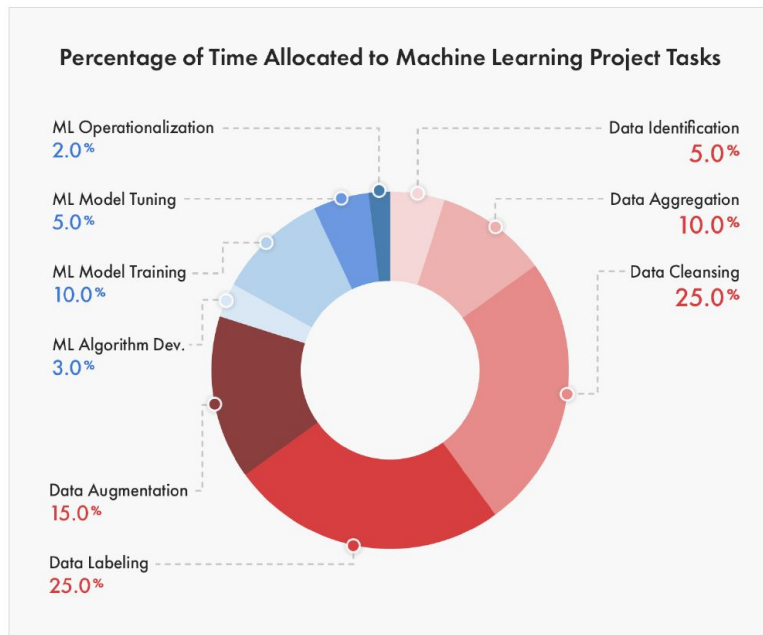
Christian Oliva Moya
Luis Fernando Lago Fernández

Auditoría de datos

- En la fase de construcción del dataset...
 - Tengo fotos de peces. ¿Quién nos dice si son sardinas o salmonetes?
 - Tengo una serie de activos financieros. ¿Cómo sé cuándo comprar o vender?
 - Tengo una serie de tweets relevantes. ¿Cómo sé si tienen sentimiento positivo o negativo?
 - Tengo fotos de una cámara de tráfico. ¿Quién dice si hay un atasco o no?

Auditoría de datos

- En la fase de construcción del dataset hay que tener en cuenta el **etiquetado**.
 - Proceso mediante el cual un conjunto de ejemplos son clasificados manualmente para luego entrenar un algoritmo de ML.
 - **Gasto económico importante**



Auditoría de datos

- En la fase de construcción del dataset...
 - Tengo identificadores.
 - El DNI de mis clientes, el customerID de mis clientes, etc.
 - Tengo valores nominales, sin relación numérica que representan categorías.
 - El sector de una empresa, el sexo de mi cliente, etc.
 - Tengo valores ordinales, atributos que se pueden ordenar pero cuya distancia no es relevante.
 - Día del mes de una transacción, año de compra de un producto, etc.

Auditoría de datos

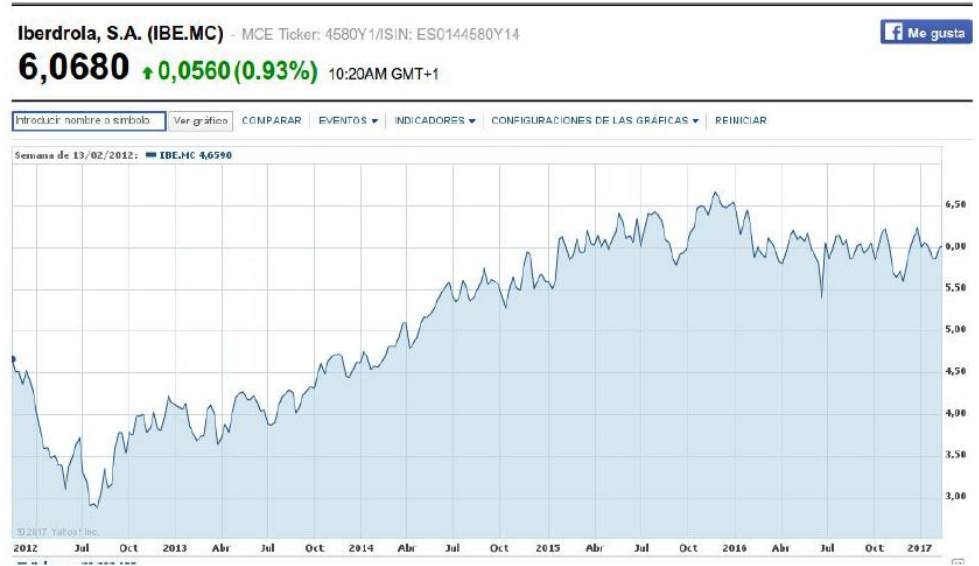
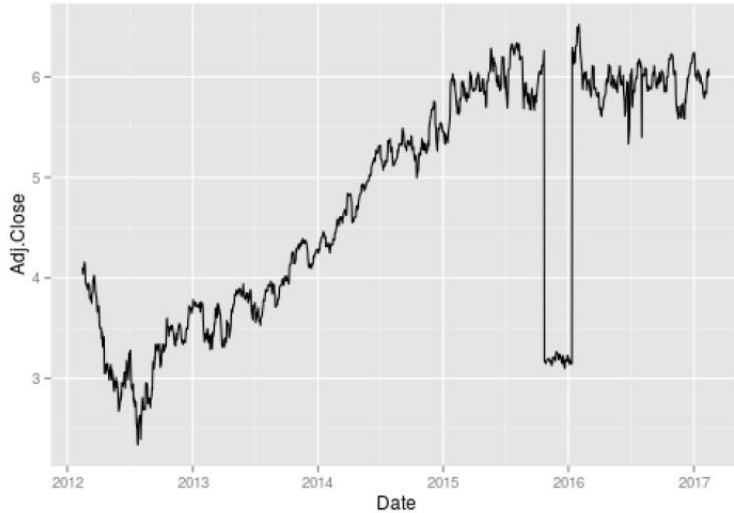
- En la fase de construcción del dataset...
 - Tengo valores numéricos, donde sí tienen sentido las distancias, la media, etc.
 - Salario de los trabajadores, valor de cierre de un activo, etc.
 - Tengo series temporales
 - Tengo imágenes
 - Tengo texto
 - Tengo grafos

Auditoría de datos

- En la fase de construcción del dataset tengo multitud de información que tengo que **procesar**.
- Realizamos un análisis exploratorio y de visualización para identificar el preprocesado necesario para tener un dataset de calidad.
 - Identificar errores en los datos
 - Tratar valores omitidos
 - Tratar outliers
 - Identificar distribuciones y asimetrías

Preprocesamiento

- Errores de entrada (mis datos vs datos reales):



Preprocesamiento

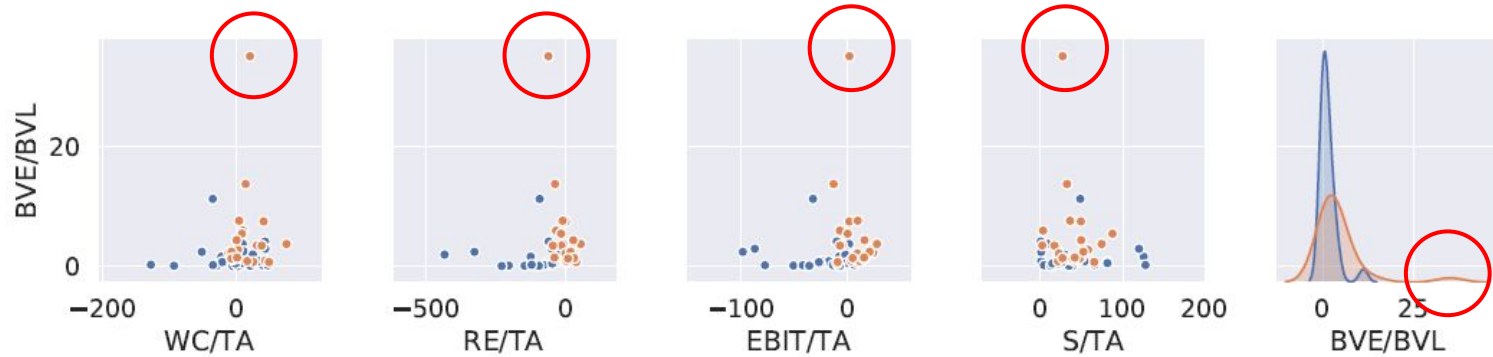
- Tratar valores omitidos (NaN):
 - Consideramos eliminar la instancia (dropna por filas)
 - Imputación de un valor que altere lo menos posible la información
 - Media
 - Moda
 - Valor fijo: 0 por ejemplo

Preprocesamiento

- Tratar outliers:
 - Se suele considerar su eliminación para reducir la distorsión de los modelos
 - Típicamente buscamos que los datos sigan una distribución normal
 - Son outliers aquellos que, por ejemplo, superan $\pm 3 * \text{std}$ (1% de los datos)

Preprocesamiento

- Tratar outliers:



Preprocesamiento

- Procesar variables nominales (categóricas):
 - Construimos variables dummies (one-hot)
 - Conseguimos que cada categoría sea equidistante a las demás

País
Alemania
Francia
España

→

pais_DE	pais_FR	pais_ES
1	0	0
0	1	0
0	0	1

Preprocesamiento

- Eliminar falsos predictores:
 - Nuestro dataset puede contener variables altamente correlacionadas con la clase
 - ¿Tenemos esa información en un escenario real?
 - Si la respuesta es NO, hay que eliminar ese falso predictor

Preprocesamiento

- Balanceo de clases:
 - Nuestro dataset puede tener datos sobre clases desbalanceadas, imagina un dataset donde un 80% de veces se toma la decisión de compra y un 20% la decisión de venta
 - Si tu modelo dice siempre “compra”, tendrás un acierto de un 80%
 - ¿Qué hacemos?
 - Oversampling de la clase minoritaria (data augmentation)
 - Undersampling de la clase mayoritaria (balanceamos a mano reduciendo el dataset)

Preprocesamiento

- Normalización:
 - Como ya hemos visto con KNN, muchos algoritmos de ML requieren que todos los atributos tengan una escala similar.

$$x_{norm} = \frac{x - \mu}{\sigma}$$

Preprocesamiento

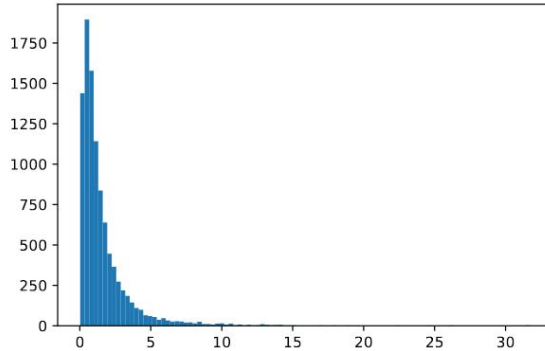
- Escalado:
 - Como alternativa, podemos escalar los datos para que estén en el rango [a, b]

$$x_{scaled} = a + (b - a) \frac{x - min}{max - min}$$

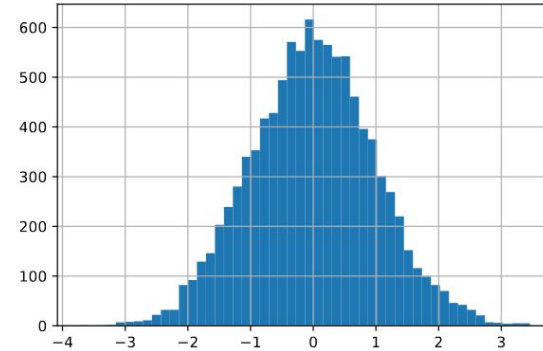
Preprocesamiento

- Logaritmo cuando tienes una distribución log-normal:

X

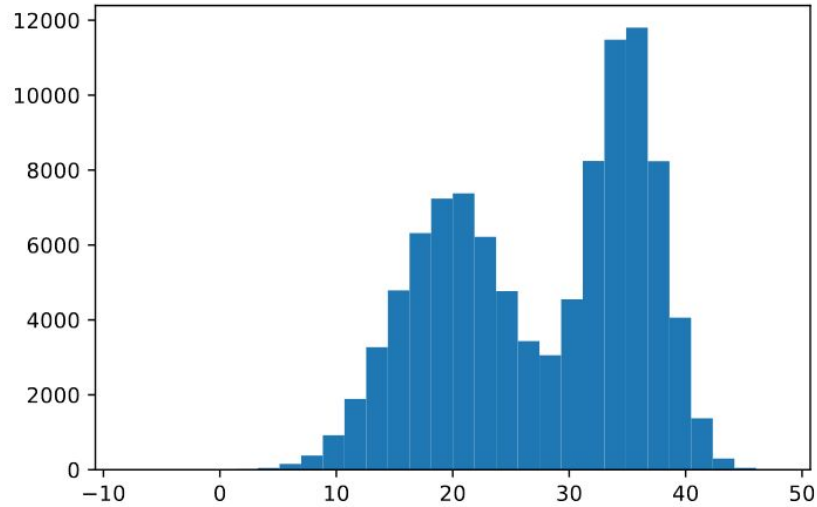


$\log(X)$



Preprocesamiento

- Otras distribuciones: Buscar un clustering previo que ayude a diferenciar grupos



Preprocesamiento

- Generar nuevas características (atributos)

Puede ser interesante combinar o transformar atributos siguiendo **un conocimiento experto**

- Trabajar con retornos en lugar de precios de un activo
- Trabajar con la productividad de un empleado en lugar del número de clientes que tenga
- Trabajar con espectrogramas en lugar de una señal de audio sin procesar