

## Memoria Hackaton. Reto DataScience

- Empiezo analizando los datasets. En el caso del train hay 133 valores nulos y decido sustituirlos por el método 'backfill', es decir, con los datos del día o fecha siguiente en que no haya valor nulo.
- Tanto en el dataset de train como el de test cambiamos el tipo de la columna 'Date' a formato fecha.
- Hay valores iguales a 0 en la columna 'Volume'. Decido eliminar estas filas, puesto que no tienen sentido sustituir dichos valores por una media o cualquier otra medida, prefiero trabajar con los datos que tengan la mayor exactitud posible. Pasamos a tener 4.828 filas.
- Utilizo el *candlestick* chart para visualizar la totalidad de los datos.
- Veo que las columnas 'Open', 'High', 'Low', 'Close' y 'Adj Close' están estrechamente relacionadas entre ellas.
- Aprovecho para crear nuevas columnas y deshacerme de otras:
  1. Me deshago de la columna 'Close', ya que es prácticamente igual que 'Adj Close'.
  2. Añado una nueva columna con la diferencia entre la columna 'Adj Close' y la columna 'Open'.
  3. Creo una nueva columna con la media móvil de 3 días de la columna 'Adj Close' llamada 'MA3'.
- Dibujo un boxplot para comprobar outliers, si bien es normal que los tenga por los picos que existen en la cotización, y que son perfectamente normales.
- Implemento dos modelos: Decision Tree y XGBoost
- Aunque hay una correlación alta entre las variables, compruebo que al eliminarlas y quedarme tan solo con una o dos, ninguno de los modelos mejora.
- Los resultados finales son:

RandomForest: F1 F1 0.57757

XGBoost: F1 0.60157

- Decido quedarme con el XGBoost por la versatilidad y facilidad que tiene a la hora de implementarse. Además, ha tenido un resultado superior en la métrica de evaluación.
- Si bien no son resultados óptimos. De hecho, son bastante bajos, hay que tener en cuenta que existen muchas más variables que pueden influir en el 'Target', y que no están contenidas en las variables que hemos entrenado. Sin embargo, es una primera aproximación a mejorar.