# CaixaBank Tech
# x
# NUWE

Chernysh, Antón

anton_chernysh@outlook.es

28 de mayo de 2022

# Índice

# 1. Introducción

El reto trata de un proyecto de Data Science donde el objetivo es realizar un modelo predicativo con la intención de predecir en precio del Ibex35. Para ello, los concursantes disponemos del precio histórico con frecuencia de vela de 1 día y varios tweets que contienen los tweets públicos que poseen el hashtag Ibex35 desde el año 2015 que han recibido más de dos likes y de dos retweets.

Para ello, se realizó un análisis de los datos en busca de patrones y características que ayudaran a reducir la búsqueda en el proceso de selección y entrenamiento de modelos. Una vez visualizados los datos y las estadísticas, se han extraído ratios (indicadores técnicos) que permiten extraer información más compleja de los datos y así poder ayudar al modelo a predecir mejor.

Finalmente se ha realizado una búsqueda del modelo más adecuado para el problema, junto al escalado y normalización de los datos.

En los siguientes apartados se explica más en detalle cada paso.

# 2. Análisis exploratorio de los datos

Durante la exploración de los datos, se ha descubierto que faltaban algunas fechas. Para esto hay que saber que la bolsa sólo esta abierta en los días laborales, por lo que los fines de semana y los festivos esta permanece cerrada y por tanto no vamos a tener datos. Una vez descartadas esas fechas, se han identificado unas pocas que no son festivos pero que no estaban. Al ser unos pocos se ha asumido que a lo mejor hay algún motivo externo que no se ha tenido en cuenta y se ha continuado con el análisis asumiendo que los datos estaban bien.

Por otro lado, se ha calculado de manera manual el Target ya que disponemos de todo el histórico. Una vez calculado, se han identificado algunos casos donde estos valores no coinciden, es decir, en algunos casos el precio en 3 días ha aumentado pero el Target es igual a cero. Al ser pocos casos y hablando con los moderadores, se ha seguido el análisis asumiendo que los Target estaban bien.

# 3. Preprocesado de datos

Antes de empezar con el preprocesado, se han interpolado las filas de datos sin datos y se ha asegurado que no hay filas duplicadas. Además, se ha visto que la variable Volumen contiene valores muy por debajo del resto antes de año 2011. Es por ello que se ha decidido descartar los datos previos a esas fechas. Aunque parece que estemos cometiendo un error descartando la mitad de los datos, estos no tienen porque ser útiles. El motivo de esto es que los patrones caducan y las personas que operan en la bolsa cambian también, es decir, es posible que los datos más antiguos no contengan ninguna información útil que nos permita predecir el futuro próximo.

## 3.1. Aumentado de datos

Para el aumentado de datos se han añadido los siguientes campos (Week cos, Year cos), que son creados a partir de cada fecha aplicándoles el coseno en un rango de 7 y 365 días respectivamente. Estas columnas permiten representar la **periodicidad y continuidad** de

las fechas, de manera que quede representado que el primer día del año/semana, va seguido del último.

Por último, hemos añadido varios **indicadores técnicos** que permiten añadir información más compleja sobre la situación de la bolsa.

## 3.2. División de datos en train y validation

Para llevar a cabo la división, se ha hecho uso de la estrategia *TimeSeriesKFold* de *scikit-learn*. Esta estrategia nos permite evaluar nuestro modelo contra varios trozos de datos en diferentes rangos de fechas cogiendo en cada fold más datos nuevos. Y después de obtener resultados de varios experimentos con distintos datos nos permite ver si es capaz de generalizar bien.

## 3.3. Escalado de datos

Una vez tenemos la división de datos, se escalan y se normalizan todas las columnas menos el target.

# 4. Selección de modelo

Antes de hablar del modelo, mencionar que se ha seguido una estrategia diferente a la ofrecida al principio. A pesar de que se ha empezado realizando un modelo clasificatorio, se ha visto que este no consigue aprender suficiente y no consigue superar el 51 % de aciertos. Es por ello que se ha realizado una predicción del valor absoluto del precio de cierre en 3 días y posteriormente se ha sacado el target de si el precio sube o baja.

Durante la selección de modelo se han probado muchas arquitecturas neuronales y todos los árboles disponibles de la librería *scikit-learn*. Después de realizar varios experimentos se ha visto que los algoritmos menos complejos como los árboles consiguen unos resultados más constantes y llegan a superar el 51 % de accuracy.