Simon E. Sanchez Viloria

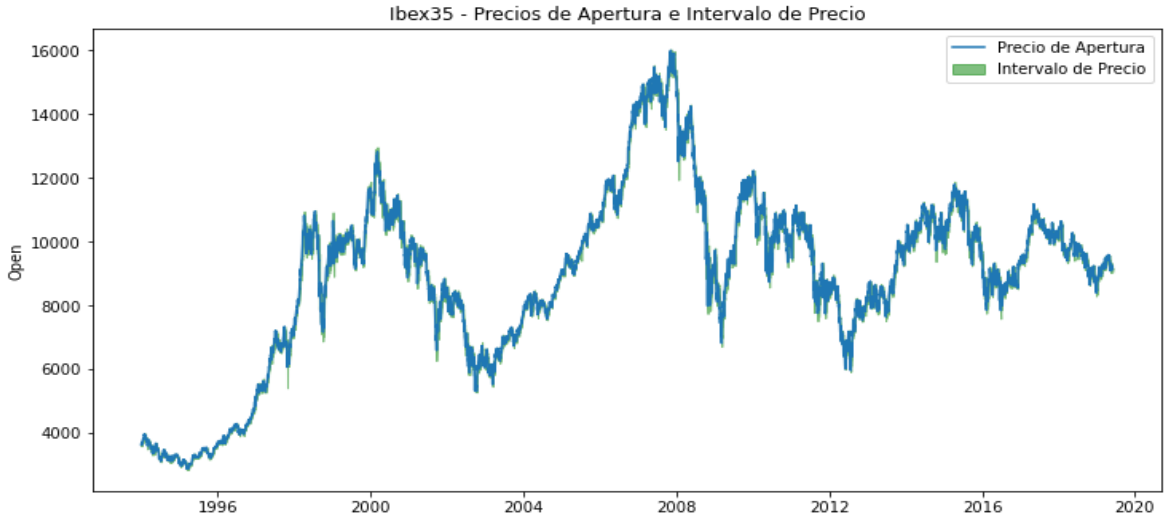
28 de Mayo de 2022

CaixaBank Tech Hachaton 2022

Reporte de Entrega del Reto de Data Science

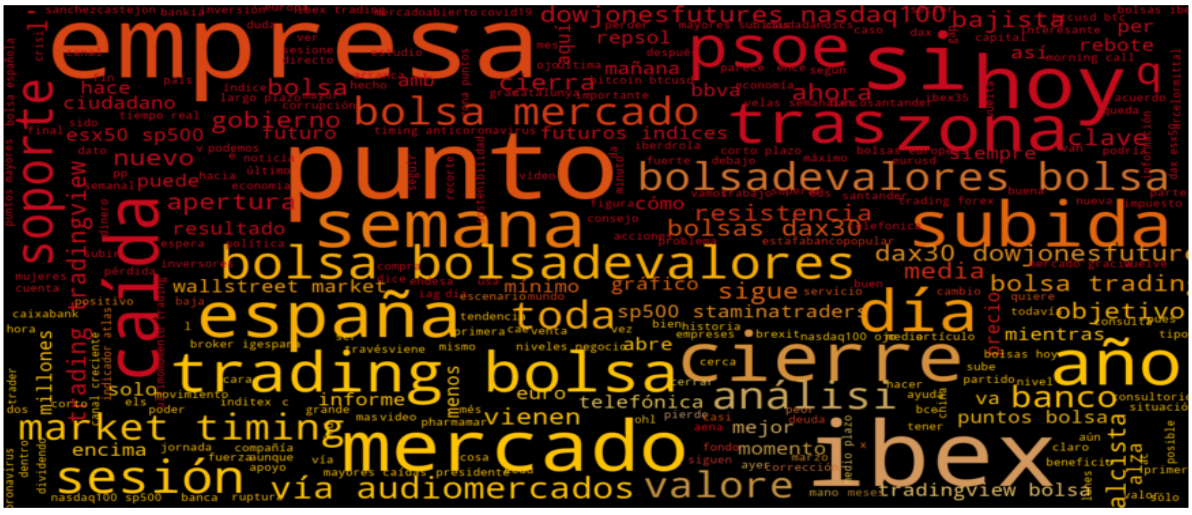
Introducción

El objetivo del reto se trataba de desarrollar un modelo de predicción binario que permita predecir la variable de salida (target) de si el precio de cierre del IBEX35 en 3 dias seria superior o inferior al precio de cierre actual. Los datos de entrada dados para el reto son datos historicos de apertura, cierre, maximo y bajo del IBEX35 en un periodo de 25 años de 1994 a 2019.



Tambien hemos recibido un conjunto de datos que continene alrededor de 9800 tweets publicos hechos con el hashtag #IBEX35 que han recibido más de dos likes y de dos retweets desde 2015.

Nube de Palabras del corpus de tweets:



Metodología

He utilizado una metodología de Machine Learning usando modelos basados en arboles (gradient boosting) para predecir la variable de salida. Este tipo de modelos suele funcionar bien con datos tabulares que asumen que cada observacion es independiente de las demas. Para modelar la temporalidad de los datos, lo usual es valerse de realizar transformaciones que representen informacion de una variable en una ventana de tiempo (como lag, rolling).

Como mi principal interes estaba en modelar usando procesado de lenguaje natural, he decidido entrenar solo usando la serie de datos de entrada desde 2015 hasta 2019 (para así incorporar los datos de tweets en cada día). La meta era obtener una serie de features que contengan informacion relevante del texto de cada tweet y que el modelo aprenda la relacion entre el precio de cierre del IBEX35 y esto.

Para eso he usado la libreria Spacy de python. Primero eliminando stopwords y lematizando para dejar solo palabras relevantes y luego obteniendo los word embeddings de los tweets de cada día. Los word embeddings pueden ser usados directamente como variables pero como el numero de dimensiones de los word embeddings dados por Spacy es alto (300 features), he usado Principal Component Analysis (PCA) sobre ellos para dejar solo 10 variables (componentes) que expliquen estos embeddings.

Finalmente, tambien he decidido añadir dos variables indicando el numero de tweets y el numero de tokens (palabras) que se han tweeteado con el hashtag #Ibex35 en cada día y creado variables lag a cada una de ellas de los ultimos 3 y 7 dias anteriores con el proposito de capturar informacion acerca de eventos donde el IBEX35 se encuentre en tendencia.

Resultados

Mi entrega final ha sido conseguida con un modelo l*ightgbm* (gradient boosting machine) con todas las variables que he contruido por medio de los metodos mencionados arriba. A continuacion presento una tabla con los resultados conseguidos despues de evaluar cada metodo de entrenamiento en un conjunto de evaluacion de 200 dias (de la serie desde 2015):

Metodo	**F1 en conjunto de test**
lightgbm sin feature engineering (dejando los datos como vienen dados)	0.52
lightgbm con feature engineering (variables temporales + lag y rolling)	0.535
lightgbm con word embeddings	0.542
lightgbm con feature engineering + word embeddings	0.584
lightgbm con feature engineering + word embeddings + variables basadas en numero de tweets y tokens por día	0.63