CaixaBank
TECH

CAIXABANK TECH

Hackaton:
Data Science

Predicción precio del IBEX35

Aitor Larrinoa Rementería

28 de mayo del 2022

Índice

1. EDA

En esta primera aproximación he intentado comprender bien los datos y si había algún error en los mismos, corregirlo.

El entendimiento de los datos ha sido sencillo. Tenemos datos del precio máximo, mínimo, de cierre, etc. del IBEX35 en frecuencia diaria. Luego, lo primero que he decidido hacer es eliminar aquella información innecesaria y quedarme únicamente con aquellos datos ricos en información. El resultado de esto ha sido quedarme únicamente con tres columnas, la del precio de cierre ajustado, la fecha y la variable target. Realmente, las variables precio máximo, mínimo, apertura y cierre no nos proporcionan mayor información que la que nos proporciona la variable del precio de cierre ajustado. Luego, me quedo únicamente con un data frame de 3 columnas.

Además, como no existen fechas duplicadas, decido utilizar la variable fecha como el índice del data frame. El data frame resultante ha sido el siguiente:

Date	Adj Close	Target
1994-01-03	3654.496338	0
1994-01-04	3630.296387	1
1994-01-05	3621.196289	1
1994-01-06	NaN	0
1994-01-07	3636.396240	1

Figura 1: Data frame con el que comenzar a trabajar

El siguiente paso ha sido el tratamiento de nulos. La columna Adj Close contenía nulos, luego había que tratarlos. Analizando esos nulos, me he dado cuenta de que todos ellos tenian un valor en la variable target de 0, lo que significaba que la variable target no estaba definida. Luego, la idea que he llevado a cabo ha sido rellenar los nulos con el valor de la media entre el valor siguiente y el anterior. Una vez los nulos han sido modificados, también había que hacer lo mismo para la variable target. Luego así lo he hecho, he ido analizando qué valor debía introducir en la variable target para los nulos tratados y si debía cambiarlo, lo hacía.

Finalmente ya he conseguido tener el data frame que quería sin ningún nulo, luego, he comenzado la parte del modelado que explico, brevemente, a continuación.

2. Modelado

En primer lugar, antes de comenzar con el modelado hay que tener claro del tipo del problema que se está analizando. En este caso se trata de un modelo de **clasificación binaria**, puesto que se tiene que predecir si el precio del IBEX35 subirá de aquí a 3 días o no.

Luego, los algoritmos que he planteado son algoritmos de clasificación y son los siguientes:

- Random Forest
- XGBoost
- Light GBM

Los 3 son algoritmos de árboles de decisión. Para elegir el mejor de los 3 modelos he decidido hacer caso a las métricas AUC y $f1$. La primera de las métricas la he elegido puesto que me parece la métrica más completa para comparar modelos ya que es una métrica que no depende del valor del threshold y, por otro lado, he tomado el valor de la métrica f1 ya que es la que pide el ejercicio. Los resultados han sido los siguientes:

	RF	XGB	LGBM
f1	0.550584	0.600873	0.593043
AUC	0.529544	0.531490	0.519220

Figura 2: Resultados de las métricas para los 3 modelos

Luego, podemos observar que el modelo que mejores resultados nos proporciona tanto en f1 como en auc es el XGBoost. Ahora, con el fin de poder encontrar un modelo aún mejor, se ha optado por realizar una optimización de hiperparámetros.

Después de la optimización de hiperparámetros, el resultado ha sido un f1_score de 0.65, esto es, hemos mejorado un 0.05 el f1_score con respecto a la no optimización de hiperparámetros. Para terminar, se muestra a contnuación la matriz de confusión y la curva ROC del modelo resultante.

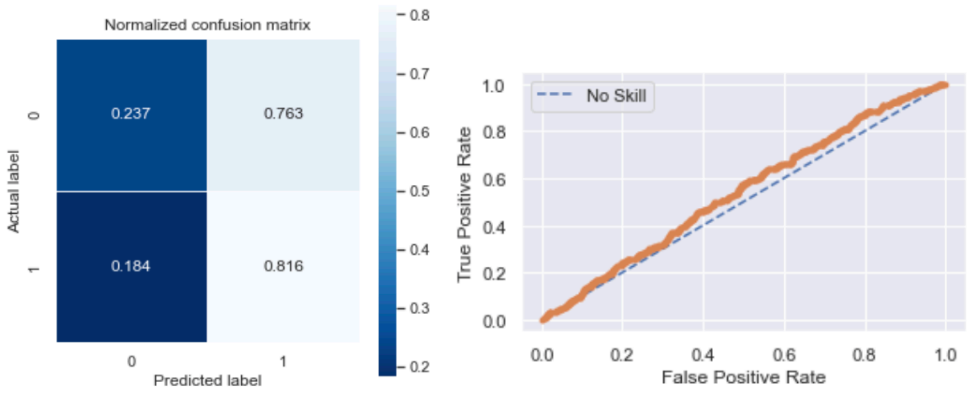

Figura 3: Matriz de confusión y crva ROC del modelo resultante