

**DATA SCIENCE
MACHINE LEARNING**

DOCUMENTO EJECUTIVO

BITCOIN



CODERHOUSE



COMISION 31490

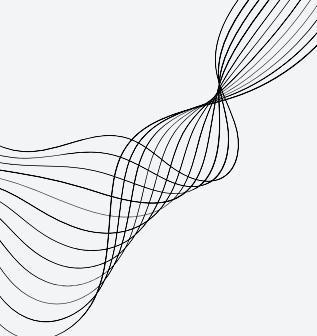


LUCIANO BIZIN



21 OCTUBRE 2022





1

Presentación

Problema de investigación



2

Etapas del proceso

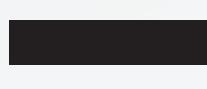
Estructura de desglose de trabajo



3

Data acquisition

Fuentes y proceso de adquisición



4

Data wrangling

Manipulación de datos



5

Data analysis

Univariado, bivariado y multivariado



6

Modelado (ML)

Modelos supervisados



7-8

Conclusiones

Resultados y próximos pasos



PRESENTACIÓN

PROBLEMA

El dinamismo de suba y baja del precio del BTC es multicausal. Por eso, todo intento de predicción de su valor futuro debe considerar diversas variables e interacciones.

Deabajo, se aplicarán diferentes modelos de ML al precio de cierre de BTC a fin de intentar predecir con grado de previsibilidad >50% su comportamiento futuro a +1 día

PREGUNTAS

I. ¿Cuáles son las variables que influyen en el precio de BTC en el corto plazo? ¿Cómo lo hacen?

II. ¿Son capaces los modelos supervisados de clasificación de machine learning -que se pueden ejecutar en computadoras domésticas- de predecir con una previsibilidad >50% la suba y/o baja del precio de este activo?

Objetivos general

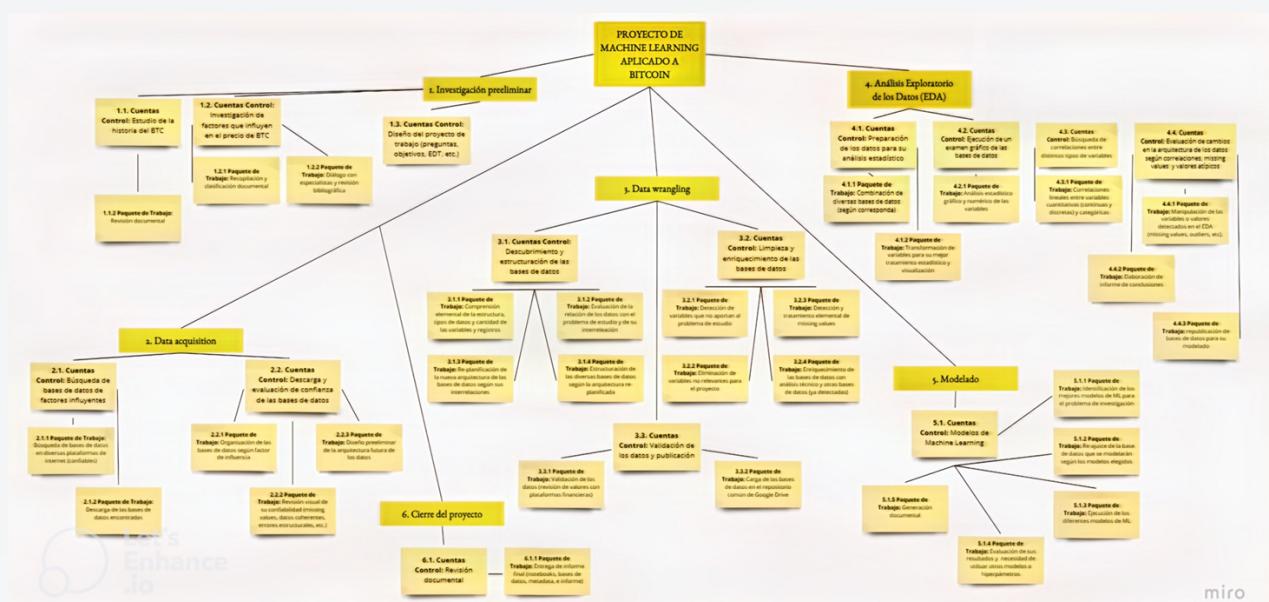
I. Diseñar diversos modelos de machine learning capaces de predecir con cierto grado de previsibilidad positiva (>50%) el dinamismo del precio de Bitcoin en el corto plazo, es decir, si el precio de este criptoactivo subirá o bajará en la temporalidad estudiada (+1 día).

Objetivos específicos

- I. Detectar los factores económicos que afectan el precio de BTC.
- II. Analizar las variables de los factores detectados a fin conocer el modo mediante el cual influencian la variación del precio de este criptoactivo.
- III. Diseñar diversos modelos de machine learning capaces de predecir con cierto grado de previsibilidad positiva (>50%) a corto plazo (+1 día), el movimiento alcista o bajista del precio de BTC.

ETAPAS DEL PROCESO

Estructura de Desglose de Trabajo (EDT)



Etapa 1. Investigación preliminar

Etapa 2. Data acquisition

Etapa 3. Data wrangling

Etapa 4. Análisis Exploratorio de los Datos

Etapa 5. Modelos de Machine Learning

Etapa 6. Cierre del proyecto

DATA ACQUISITION

METALES

Oro, plata,
cobre, platino,
paladio, aluminio

INDICADORES TÉCNICOS

Tendencia, volumen,
ciclicidad, momento y
volatilidad

INDICADORES DE EEUU

Bonos 10 años, FOREX,
índice de precio al
consumidor y bolsas

PATRONES DE VELAS JAPONESAS

Subida, bajada,
reversión de precios,
cambio de tendencia

DATA WRANGLING

Nota importante: en la distintas instancias de manipulación de datos se fueron reduciendo la cantidad de variables de las bases de datos y se tomaron decisiones sobre los valores NaN (ver "Entrega final.docx" y/o "Metadata.docx"). Con respecto a los outliers se evitó eliminarlos debido a que resulta arbitrario hacerlo en una serie de tiempo del tipo de BTC: ¿qué período de tiempo es el correcto para considerar un dato como un outlier? ¿Con respecto a qué otro valor lo es (por ejemplo un valor de 0.1 en 2010 es un outlier con respecto a los 62 mil pesos de 2021, o viceversa)?.

1

BTC_metals (v1)

BTC_metals_fv (v2)

Precio de BTC + Precio de metales

2

BTC_ind_trend (v1)

BTC_ind_trend_ET (v2)

BTC_ind_trend_ET_fv (v3)

Precio de BTC + indicadores de tendencia

3

BTC_ind_mom (v1)

BTC_ind_mom_ET (v2)

BTC_ind_mom_ET_fv (v3)

Precio de BTC + indicadores de momento

4

BTC_US (v1)

BTC_US_fv (v2)

Precio de BTC + indicadores de EEUU

5

BTC_multiple_indicators (v1)

BTC_multiple_indicators_fv (v2)

Precio de BTC + ciclicidad, regresiones, volatilidad y volumen

6

BTC_various_indicators_fv (v1)

Precio de BTC + variables de alta correlación

DATA ANALYSIS



UNIVARIADO

- Cantidad de variables
- Tipos de variables
- Missig values
- Histogramas
- Profilings

BIVARIADO

- Relaciones entre variables
 - Foco en [Price]
 - Foco en [Percentage_difference]
 - Foco en [Target]
- Análisis de correlaciones
- Estacionariedad y autocorrelación
- Gráficos (violin, heatmaps, histogramas)

MULTI-
VARIADO

- Relaciones multidimensionales
- Discriminación de variables
- Identificación de futuras bases de datos
- Detección de potenciales estrategias de trading

→ Resultados más relevantes

- Se detectaron las variables con mayor correlación en cada data set.
- La serie no es estacionaria y la autocorrelación responde mejor a -1 día.
- Se descartaron las variables que menos correlación tenían con [Price], [Target] y [Percentage_difference].
- Se delinearon las estrategias necesarias para el modelado de los data set.

MODELADO (ML - CLASIFICACIÓN)

Nota importante: la variable [Target] de cada base de dato refleja la subida (1) y bajada (0) del precio de BTC con manipulación de shift -1 para transformar un problema de regresión en uno de clasificación (para más información, revisar "Entrega final.docx" y/o "Metadata.docx").

1

Decision tree
(x7 bases de datos)
GridSearchCV /
RandomizedSearchCV

2

Random forest
(x7 bases de datos)
GridSearchCV /
RandomizedSearchCV

3

K-Nearest Neighbor (KNN)
(x7 bases de datos)
GridSearchCV /
RandomizedSearchCV

4

Logistic regression
(x7 bases de datos)
GridSearchCV /
RandomizedSearchCV

5

Support Vector Machine (SVM)
(x7 bases de datos)
GridSearchCV /
RandomizedSearchCV

6

Boosting models
(x3 bases de datos)
GridSearchCV /
RandomizedSearchCV
LeaveOneOut

RESULTADOS Y PRÓXIMOS PASOS

MEJOR MODELO
73% ACCURACY

Mejor modelo ---> HistGradientBoostingClassifier
Base de datos ---> BTC_multiple_indicators_fv

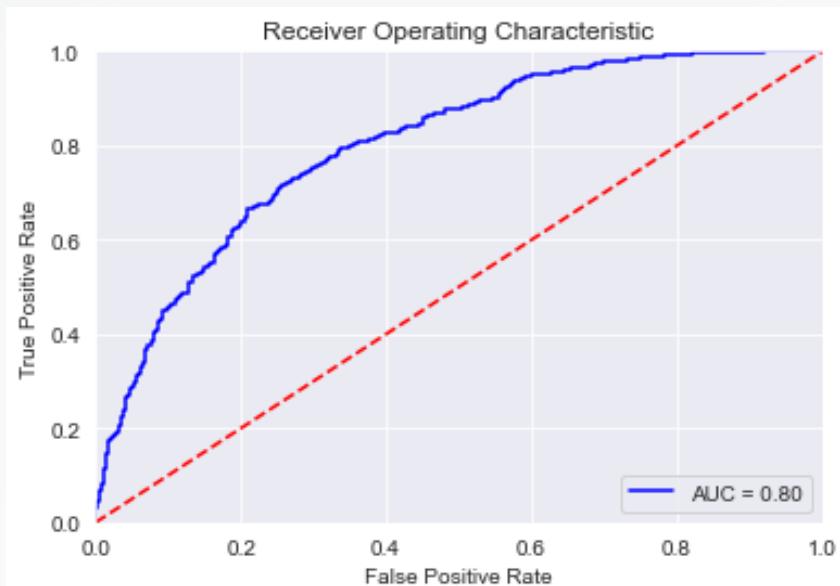
- *Resultado de accuracy (testing): 0.7294*
- *Resultado de precisión es: 0.7115*
- *Resultado de recall es: 0.7115*
- *Resultado del F1 Score es: 0.7115*
- *Resultado del AUC es: 0.8009*

Parámetros del modelo --->

- 'warm_start': True,
- 'random_state': 2537,
- 'max_leaf_nodes': 15,
- 'max_iter': 100,
- 'loss': 'auto',
- 'learning_rate': 0.1,
- 'l2_regularization': 0.1

Matriz de confusión --->

(VN) [521 - 178] (FP)
(FN) [178 - 439] (VP)



RESULTADO Y PRÓXIMOS PASOS

Conclusiones generales

- En general, los modelos que dieron mejores métricas fueron los Decision Tree, Random Forest e HistGradientBoostingClassifier.
- Los resultados obtenidos sugieren que los modelos de ML poseen como cota superior de predicción de la subida y bajada del precio de BTC en el corto plazo (+1 día) entre 65-73% (accuracy [testing]).
- Los modelos de boosting no mejoraron considerablemente los resultados de los modelos, con excepción del mejor modelo de todos, el HistGradientBoostingClassifier ()�.
- Se cree que el 30% faltante del accuracy de los modelos para una predicción perfecta se debe entender mediante los aportes de variables relacionadas con un modelo económico de tipo comportamental (noticias, análisis de sentimientos, etc.) y de variables relacionadas con la blockchain de BTC (dificultad de minado, cantidad de BTC minados, movimiento de ballenas, etc.).

Próximos pasos

- Modelar los data set con modelos de deep learning.
- Construir otras bases de datos con variables de sentimientos y métricas de la blockchain de BTC.
- Extender el análisis a otros activos financieros menos volátiles.
- Analizar las correlaciones entre diversos criptoactivos.
- Desarrollar modelos de ML para determinar la predicción de estrategias de trading ganadoras (BTC y otros activos financieros).
- Proceder a realizar nuevos modelos de ML con valores de [Target] = 1, cuando [Percentage_difference] sea mayor distintos valores positivos.