MÁSTER EN BIG DATA Y DATA SCIENCE

05MBID Estadística Avanzada

CURSO 2021-2022

ACTIVIDAD 1: Análisis de datos

Hecho por el estudiante Carlos de la Morena Coco

Introducción

Las criptomonedas son activos digitales que emplean un cifrado criptográfico para garantizar la titularidad y asegurar la integridad de las transacciones. La gran novedad de estos activos es que son completamente descentralizados, es decir, no es necesaria la existencia de ninguna entidad bancaria o gobierno que dé validez a cada transacción, o que tenga potestad suficiente como para incrementar la masa monetaria repentinamente causando inflación.

Que estos activos sean o no el futuro de las finanzas mundiales es un tema interesante a discutir por políticos o economistas, pero a día de hoy su futuro parece bastante incierto.

Hay una característica que, nos guste o no, es intrínseca a estos activos hoy en día, y es la alta volatilidad de su precio. Mientras que en el mercado tradicional movimientos de un 1% diarios se consideran grandes caídas o subidas, en el mercado de las criptomonedas es prácticamente habitual ver variaciones del 5% de un activo de un día a otro. Y esto las convierte un los activos perfectos para especular con ellos a corto plazo.

El único inconveniente es la imprevisibilidad de estos movimientos, debido a que el mercado de las criptomonedas es bastante joven y no está suficientemente estudiado.

De todas formas, los expertos más veteranos han observado tendencias en estos movimientos de precios, lo cual evicencia cierta previsibilidad de la cual se puede generar beneficio, comprando activos en zonas previas a grandes subidas y vendiéndolos antes de grandes bajadas.

Hay dos factores que tenemos que tener en cuenta a la hora de realizar este tipo de predicciones:

- El mercado de las criptomonedas sigue a bitcoin. Bitcoin es la primera criptomoneda, la que más tiempo lleva en el mercado, la que mayor confianza inspira en los inversores y la moneda de mayor capitalización de mercado. Cualquier movimiento que haga bitcoin se refleja en mayor o menor medida en el resto de criptomonedas (llamémoslas altcoins) debido al flujo de capital de una criptomoneda a otra por parte de los inversores después de cada movimiento.
- El factor más determinante en el precio de una criptomoneda es la confianza de los inversores, traducido en compras o ventas masivas. Es por ello que después de una gran subida haya una bajada (o corrección a la baja), pues los inversores venden sus posiciones para tomar beneficio y esto se refleja en el precio. Otras noticias relacionadas a las criptomonedas pueden influir en el precio, pero el factor que más nos va a ayudar a predecir el precio de las criptomonedas a corto plazo es el comportamiento de los inversores mayoritarios.

Elección de la base de datos de partida

Precisamente debido a que este mercado es demasiado joven, no hay demasiados datos por internet. Afortunadamente, en kaggle tenemos datasets con la información de los precios de las criptomonedas más importantes (con el precio de comienzo del día, de final de día, máximo y mínimo diarios y capitalización de mercado para cada criptomoneda desde su creación hasta verano de 2021). Son 5 atributos para cada criptomoneda. Para tener una percepción más amplia del flujo de capital de una a otra usaremos los datos de 10 criptomonedas, lo cual nos deja un total de 50 atributos.

Los datos se encuentran en la siguiente dirección: https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

Por desgracia, estos datos son incompletos a día de hoy, por lo que esta información la completamos con información extraída de la API de coinbase (https://api.pro.coinbase.com). De esta API con la versión gratuita no podemos extraer más de 6 meses de información, lo cual se ajusta perfectamente a nuestras necesidades.

Selección de dataset

De kaggle se ha descargado el dataset con el registro más completo de internet con los precios de las criptomonedas más importantes, de los cuales usaremos 10.

Para cada criptomoneda tenemos una entrada por cada día desde que se tienen registros del precio (en el caso de bitcoin hay 2991 entradas). De este modo tenemos suficientes datos para analizar.

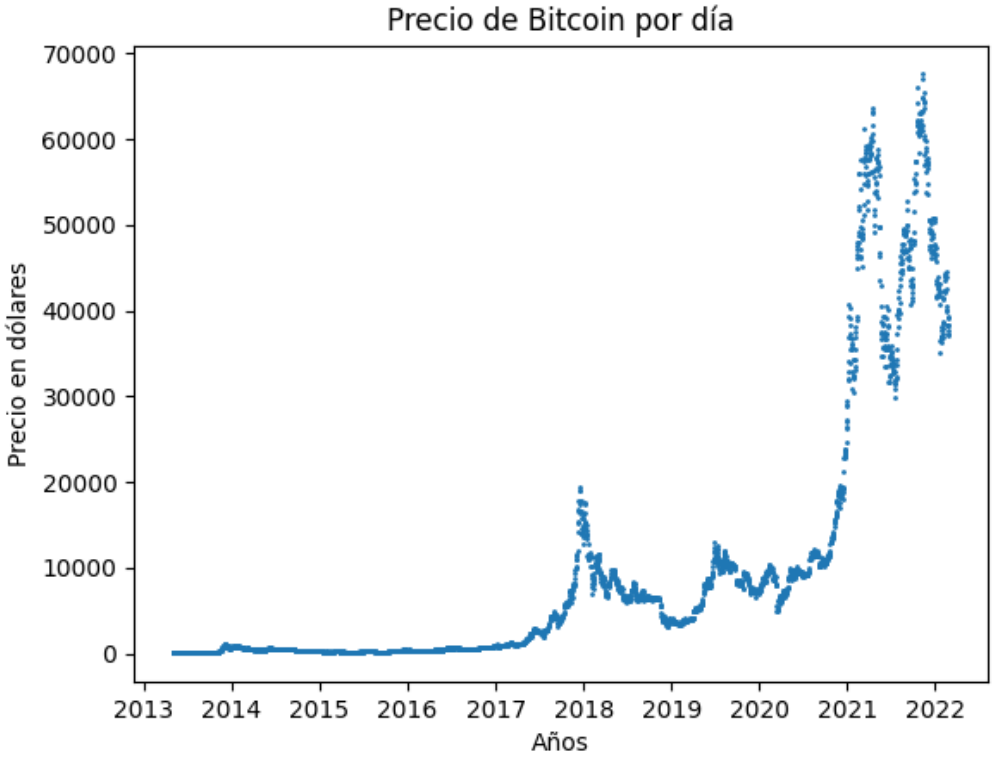
Además, hemos extraído las entradas complementarias para tener el precio de bitcoin hasta el día de hoy, 28 de febrero de 2022, obteniendo con ello otras 498 entradas.

Preparación de los datos

Para este caso, realizaremos un análisis con regresión lineal simple. Como parámetro Y usaremos el precio en dólares del bitcoin, y como parámetro X la fecha en la que de entrada se dio dicho precio. Al no poderse realizar operaciones matemáticas con las fechas, se decidió usar una serie numérica, de forma que el primer precio registrado se corresponde con la cifra 0, el segundo con la cifra 1, y así respectivamente con todos los datos del dataset.

Visualización y toma de decisiones

Al visualizar los datos, vemos que el gráfico tiene la siguiente forma



Como se puede apreciar, el gráfico a simple vista no parece una dependencia lineal clara, pero puede parecer una dependencia exponencial.
Y en este caso, podríamos usar estos datos igualmente para hacer una regresión lineal tomando el logaritmo.
Demostremos esto antes de ir más allá.
Una dependencia exponencial se vería de la siguiente manera:

$$F(n)=c^n*b$$

Mientras que una dependencia lineal se vería así:

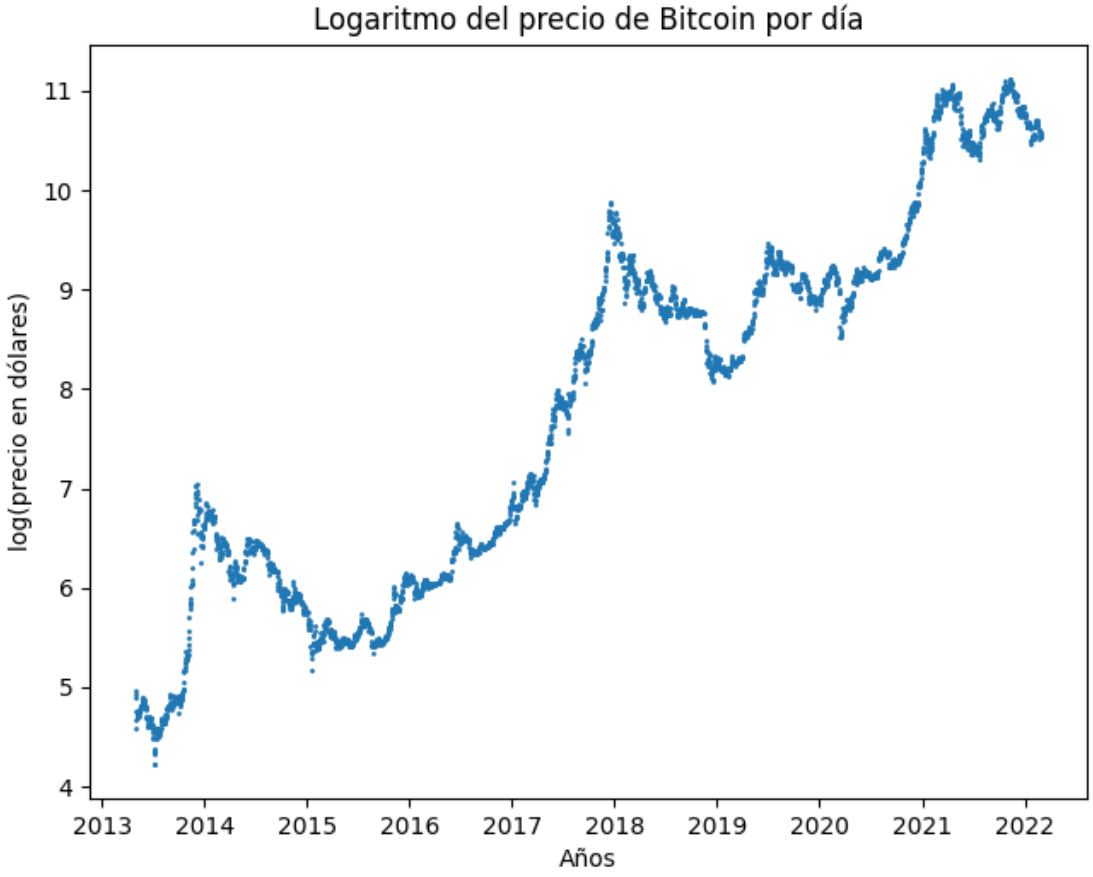
$$F(n)=a*n+b$$

Para pasar la dependencia lineal a exponencial, podemos tomar logaritmos de ambas partes

$$\widetilde{F}(n)=\log(F(n))=\log(c^n*b)=\log(c^n)+\log(b)=n*\log(c)+\log(b)$$

$$\widetilde{F}(n)=n*\log(c)+\log(b)$$

Donde ‘log(c)’ sería como ‘a’ para la ecuación lineal y ‘log(b)’ sería el equivalente de ‘b’.
Por lo cual vemos que tendría sentido analizar el gráfico de precios de bitcoin con una regresión lineal a pesar de haber dependencia exponencial.
Tomamos el logaritmo del precio y lo visualizamos:



Ahora la dependencia lineal parece más obvia, por lo que podemos seguir con nuestro análisis.

Planteamiento y análisis

Lo que nosotros queremos es aproximar el gráfico del logaritmo del precio de bitcoin a la ecuación de una recta. La ecuación de la recta se vería de la siguiente manera:

$$\check{y} = a * x + b$$

Donde $\check{y}$ sería el logaritmo del precio de bitcoin y x es el número del día. De modo que calculando a y b podríamos averiguar esta dependencia.

Podríamos describir el conjunto de puntos de esta gráfica como

y = x, lo cual nos posibilita el cálculo del error de nuestra aproximación con respecto a nuestra gráfica.

$$E(x)=\sum_{i=0}^{N}\left(y^{(i)}-\check{y}^{(i)}\right)^2$$

Podemos sustituir $\check{y}$ por el resultado de su ecuación, lo cual nos acabaría dando:

$$E(x)=\sum_{i=0}^{N}\left(y^{(i)}-a*x^{(i)}-b\right)^2$$

Para encontrar el error mínimo, simplemente tendríamos que calcular la derivada de esta ecuación por a y por b e igualarlas a 0.
Derivamos por a:

$$\frac{dE(x)}{d(a)}=\sum_{i=0}^{N}2*\left(y^{(i)}-a*x^{(i)}-b\right)*\left(-x\right)^{(i)}=0$$

Y por b:

$$\frac{dE(x)}{d(a)}=\sum_{i=0}^{N}2*\left(y^{(i)}-a*x^{(i)}-b\right)*\left(-1\right)=0$$

Obtenemos un sistema de dos ecuaciones para dos incógnitas, lo cual podemos resolver fácilmente.
Al resolverlo, obtenemos:

$$a=\frac{N*\sum_{i=0}^{N}y^{(i)}*x^{(i)}-\sum_{i=0}^{N}x^{(i)}*\sum_{i=0}^{N}y^{(i)}}{N*\sum_{i=0}^{N}x^{(i)}*x^{(i)}-\left(\sum_{i=0}^{N}x^{(i)}\right)^2}$$

$$b=\frac{\sum_{i=0}^{N} x^{(i)} * \sum_{i=0}^{N} y^{(i)} * x^{(i)} - \sum_{i=0}^{N} x^{(i)} * \sum_{i=0}^{N} y^{(i)}}{\left(\sum_{i=0}^{N} x^{(i)}\right)^2 - N * \sum_{i=0}^{N} x^{(i)} * x^{(i)}}$$

Aprovechando que tenemos información tanto de x como de y, los calculamos y presentamos el resultado.

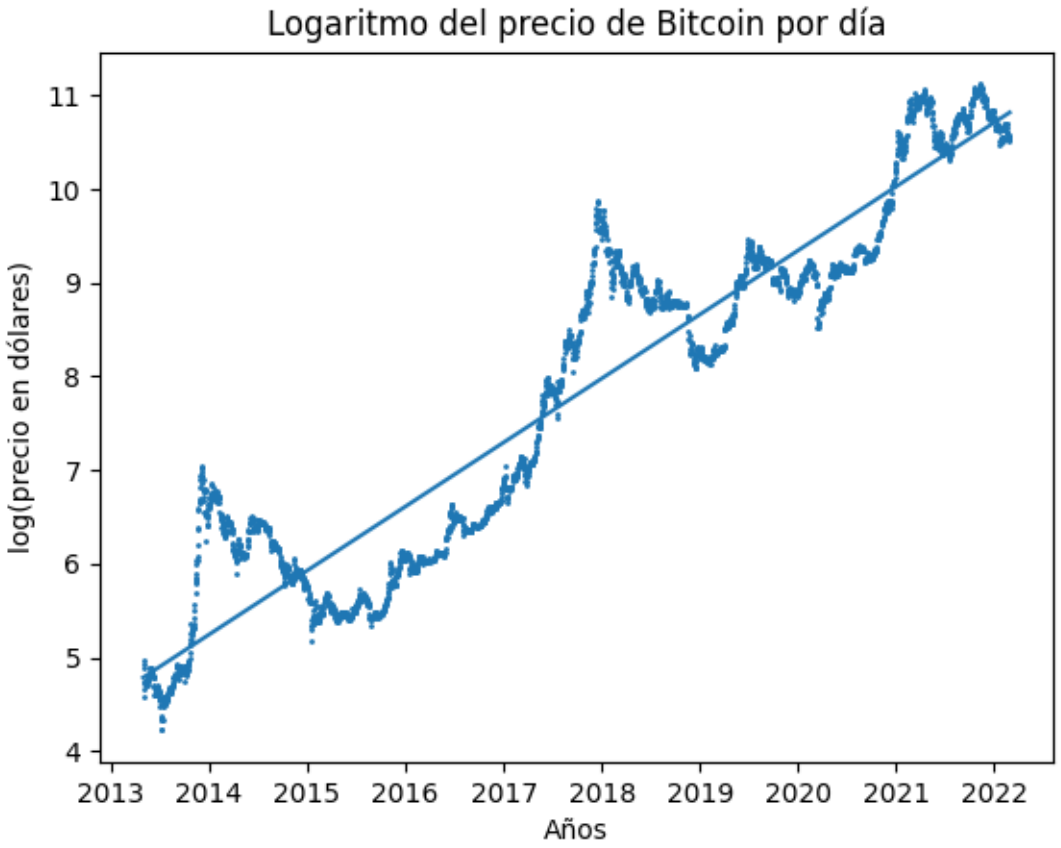
Con nuestros datos, obtenemos que a = 0.00186841872394541 y b = 4.7881864340241815.

Con lo que obtenemos la siguiente relación:

$$\ln\left(Precio\,de\,bitcoin\right) = 0.00186841872394541 * \left(numero\,de\,día\right) + 4.7881864340241815$$

Aplicación del modelo y análisis de resultados

Al dibujar el gráfico de la recta obtenida anteriormente sobre los puntos de nuestro dataset obtenemos lo siguiente.



Como vemos, nuestra recta sigue perfectamente el movimiento del logaritmo del precio del bitcoin a lo largo del tiempo.

Para evaluar la exactitud de nuestros resultados utilizaremos el error cuadrático medio.

Este error se calcula con la siguiente fórmula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Donde SS_{res} es la suma de cuadrados residual:

$$SS_{res} = \sum_{i=0}^{N} \left(y^{(i)} - \check{y}^{(i)} \right)^2$$

Y SS_{tot} es la suma de cuadrados total:

$$SS_{tot} = \sum_{i=0}^{\infty} \left(y^{(i)} - \bar{y}^{(i)} \right)^2$$

En este caso, hemos obtenido una suma de cuadrados residual de 1265.17 y una suma de cuadrados total de 11950.32.
Con estos datos, obtenemos un error cuadrático medio de 0.895. Una recta ideal tendría un error cuadrático medio de 1, por lo que este error cuadrático de prácticamente 0.9 nos indica que la correlación lineal que estábamos investigando es correcta.

Conclusión

Hemos aprovechado un dataset con unos datos que a primera vista parecen aleatorios, los hemos modificado para realizar un análisis estadístico y, gracias a los conocimientos adquiridos durante el curso, hemos conseguido encontrar una dependencia lineal que nos puede ser muy útil en el futuro.