MÁSTER EN BIG DATA Y DATA SCIENCE

05MBID Minería de datos

CURSO 2021-2022

ACTIVIDAD 1: Propuesta de proceso KDD completo

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 de 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

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.

Descripción de las características del dataset

Vamos a usar el un dataset para cada una de las monedas a analizar, y estas son Bitcoin, Ethereum, Litecoin, Dogecoin, Cardano, Polkadot, Monero, Solana, BinanceCoin y CryptocomCoin.
Los parámetros más importantes son:

- Date – Fecha de este registro concreto. Está en formato YYYY-MM-DD. Tipo datestamp.
- High – Precio máximo de una criptomoneda un día concreto. Tipo float.
- Low – Precio mínimo de una criptomoneda en concreto. Tipo float
- Open – Precio de una criptomoneda concreta al comienzo del día. Tipo float.
- Close – Precio de una criptomoneda concreta al final del día. Tipo float.
- MarketCap – Precio total de toda la masa monetaria de una criptomoneda en un día. Tipo float.

Preparación de los datos

Dado que tenemos los datos en 10 datasets distintos, vamos a juntar todos en un nuevo dataset.
La fecha será común para todos los atributos, ya que obviamente se repite para todas las criptomonedas. El resto de campos mencionados anteriormente serán incluidos para cada criptomoneda con distinto nombre, siendo rasgo distintivo en símbolo de la criptomoneda en cuestión separado por una barra baja.
Por ejemplo, el símbolo de bitcoin es BTC. El campo de nuestro dataset que nos indique el precio de apertura del bitcoin un día concreto será Open_BTC. Y así con cada criptomoneda.

Limpieza de datos

Afortunadamente, este dataset no tiene datos perdidos, así que no vamos a tener que eliminar ninguna entrada. Sí que es verdad que

hay partes en las que, al no empezar a existir todas las criptomonedas en el mismo tiempo en concreto, habrá datos de una criptomoneda y no de otra, pero para este caso simplemente centraremos nuestro análisis en las existentes.
De los datos de máximo y mínimo diario para cada moneda podemos ver la variación diaria y sacar un porcentaje de cuánto ha variado cada criptomoneda en un día, lo cual puede darnos pistas del comportamiento del precio a corto plazo.

Posibles modelos a utilizar

Usando el cheatsheet de scikit-learn y dado que se trata de un modelo regresivo, el modelo que más probablemente acabemos usando será Lasso o Elastic-Net.

Información adicional y posibles pasos futuros

La dificultad de este problema reside en la necesidad de analizar los datos en diferentes conjuntos de distinta cantidad de días, saber hacerlo escalonadamente y saber reflejar tanto el comportamiento del precio de cada criptomoneda por separado como el comportamiento de cada criptomoneda con respecto a bitcoin y a las demás. Este no es el típico problema en el que metemos varios objetos en la lista x, varias respuestas en la lista y, entrenamos y probamos el modelo, sino que vamos a tener que pensar cómo organizar este aprendizaje. Además debemos tener en cuenta que la masa monetaria de cada criptomoneda aumenta a lo largo del día, de modo que la variable que nos va a dar el precio con mayor exactitud es la capitalización de mercado, pero de igual moda podemos considerar que el aumento de masa monetaria a lo largo

del día es despreciable y utilizar los datos de precios de cada moneda para obtener la variación diaria.

Hay muchas posibles combinaciones y vamos a probarlas todas.