Tarea 4 - Computación Científica y Ciencia de los Datos
Modelamiento estadístico de una criptomoneda

Prof: Pablo Román

4 de Julio 2024

1 Objetivos

Desarrollar un modelo estadísticos Bayesianos de un fenómeno altamente incierto y comprobar su ajuste a los datos utilizando computación probabilística. El fenómeno a analizar son los precios al cierre del día y volumen de transacciones por día en número de bitcoin.

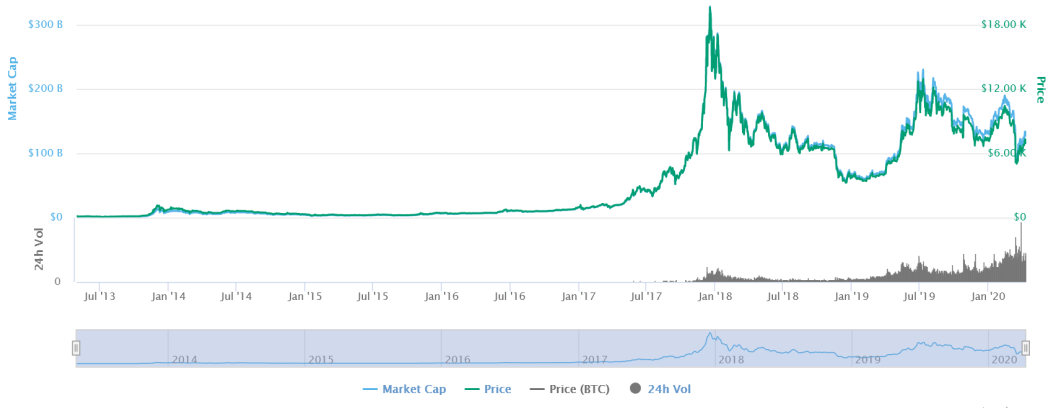

Figure 1: Bitcoin Historical Dataset (Fuente: Kaggle)

2 Contexto

Las cryptomonedas han adquirido una notable importancia en los mercados monetarios (`https://en.wikipedia.org/wiki/Cryptocurrency`). Se ha visto un constante incremento en su precio y esto ha generado un movimiento especulativo de inversores de riesgo que apuestan por la compra y venta de esta moneda. Por tanto existe un interés en desarrollar métodos que puedan predecir el precio futuro de estas monedas. Por ejemplo el concurso G-Research Crypto Forecasting (`https://www.kaggle.com/competitions/g-research-crypto-forecasting`) ofrecía 125.000 US$ en premios por el mejor algoritmo de predicción de precios. No es menor el desafío ya que se considera la evolución de este tipo de cambio volátil e impredecible. Teniendo en cuenta que se transan mas de 40 billones de dolares diarios en este tipo de monedas, tan solo lograr describir sus fluctuaciones es valioso. Claramente influyen muchas variables exógenas en su precio: políticas, rumores, el clima, etc; pero puede ser posible asignar a este fenómeno un modelo probabilístico Bayesiano (ejemplo `https://arxiv.org/pdf/2308.01013`). A pesar que dicho ejemplo es complejo, la data disponible es simple y susceptible a un primer análisis mas simple.

Para esta tarea se requiere utilizar el conjunto de datos de transacciones de Bitcoin (`https://www.kaggle.com/datasets/prasoonkottarathil/btcinusd?resource=download`), generar un modelo Bayesiano probabilístico, y analizar cuán fielmente describe a los datos.

3 Programación Probabilística

La programación probabilística representa en el lenguaje a variables aleatorias como objetos de primera clase (propios del lenguaje). Eso significa que tiene la habilidad de componer dichos objetos de manera construir modelos mas complejos. Adicionalmente puede construir objetos que representan verosimilitudes de forma de representar a los datos de entrada. Estos lenguajes tienen un método automático de calcular las distribuciones de probabilidad posterior bajo la forma de un muestreo empírico de ella. Con esto pueden calcularse varianzas y valores medios que de otra forma no podrían calcularse. El costo finalmente es de tiempo ya que los métodos de Montecarlo son costosos. El lenguaje de programación en esta evaluación es Python utilizando la librería PyMC (`https://www.pymc.io/welcome.html`).

4 Datos

Vimos que los datos a utilizar están disponibles en el sitio de Kaggle:
(`https://www.kaggle.com/datasets/prasoonkottarathil/btcinusd?resource=download`)
Corresponden a las transacciones desde 2017 de bitcoin agregada por minuto, hora, dia. Dichos datos contienen 9 columnas:

- Unix: Corresponde al tiempo Unix (timestamp) del registro
- Date: Igual que la anterior pero en formato fecha hora
- Symbol: Esta columna es redundante solo dice que es moneda bitcoin
- Open: el precio de bitcoin a la apertura del periodo.
- High: el precio más alto registrado durante ese período.
- Low: el precio más bajo registrado durante ese período.
- Close: El precio al cierre del período.
- Volume BTC: Cantidad de bitcoin transados en el período.
- Volume USD: Monto en dolares totales transados en el período.

La transacción de la moneda bitcoin tiene una alta frecuencia temporal, por este motivo se agrega por minuto, hora y día.

5 Experimentos a realizar

Debe concebir un conjunto de datos por ud mismo que sea demostrativo para la implementación que debe realizar. No se permite el uso de loop de Python salvo que lo justifique. Aquí no se pretende predecir sino modelar los datos.

1. (1 pto) Documente en forma clara y prolija en jupyter notebook la resolución de sus tarea y experimentos realizados. Describa en detalle su enfoque utilizado e interprete sus resultados. Limite su documentación, aproximadamente a 8 hojas carta sin gráficos. Ejemplifique mediante gráficos. Un gráfico bien desarrollado vale más que mil palabras (o nulo si no es autoexplicativo).

2. (3.5 pts) Utilizando el archivo BTC-Daily.csv seleccione un subconjunto interesante de datos. Podría ser entre marzo y diciembre 2018, es decir que muestre algo de estabilidad. Desarrolle un modelo probabilístico en PyMC que represente lo más fielmente posible a los datos en la columna *close*. Fundamente cuán cercano es su modelo a los datos observados.

3. (1.5 pts) Realice el mismo experimento anterior en algún otro segmento de los datos (que se vea estable) y compare ambas distribuciones. Puede decirse que su modelo se ajusta mejor a este nuevo segmento?

4. (1 pts BONUS) Cree un modelo para incluir segmentos que contengan grandes caídas o subidas en los montos y que sea lo más preciso posible. Puede estimar la frecuencia de las grandes perturbaciones del mercado de Bitcoin?

6 Entrega

Debe subir su Jupyter Notebook en classroom.