



Resumen

En este artículo, se investiga el comportamiento de los precios de Bitcoin a lo largo del tiempo y se exploran dos enfoques diferentes para la predicción: Random Forest y Redes Neuronales Recurrentes con unidades LSTM. El estudio se centra en la volatilidad de las criptomonedas, que ha generado escepticismo debido a su sistema descentralizado y la falta de control por parte de autoridades reguladoras. A medida que las criptomonedas ganan aceptación en el mercado, su alta volatilidad se ha convertido en un tema de preocupación. El objetivo principal de esta investigación es proporcionar una comprensión más profunda de cómo las técnicas de Random Forest y Redes Neuronales Recurrentes pueden aplicarse al análisis de criptomonedas con el fin de reducir esta volatilidad y predecir los valores de las criptomonedas.

Objetivos

El objetivo principal de este proyecto es analizar y predecir el comportamiento de los precios de Bitcoin a lo largo del tiempo, abordando la volatilidad característica de las criptomonedas. Para lograr este objetivo, se emplearán dos enfoques diferentes: Random Forest y Redes Neuronales Recurrentes (RNN) con unidades LSTM.

Fundamentos Teóricos

Series de Tiempo

Una serie de tiempo es un conjunto de datos que se recopila o registra en intervalos regulares a lo largo del tiempo.

Características de una Serie de Tiempo

1. **Movimientos a Largo Plazo o Tendencia.**
2. **Movimientos Cíclicos: Fluctuaciones que ocurren en patrones recurrentes, pero no necesariamente a intervalos fijos.**
3. **Movimientos Estacionales: Intervalos regulares a lo largo del tiempo.**
4. **Movimientos Irregulares o Aleatorios.**

Etapas Preliminares del Análisis

- **Experticia en el Ámbito de Aplicación.**
- **Definición de Objetivos.**
- **Recolección de Datos.**
- **Análisis Exploratorio.**
- **Representación de Componentes.**

Estacionariedad y Función de Autocorrelación

La estacionariedad se refiere a que las propiedades estadísticas de la serie se mantienen constantes a lo largo del tiempo. La función de autocorrelación es una herramienta esencial en este contexto, ya que permite comparar la señal con ella misma y revelar información valiosa sobre la estructura temporal de la serie.

Machine Learning

Árboles de Decisión

Los árboles de decisión son algoritmos versátiles utilizados en tareas de clasificación y regresión. Son especialmente útiles para capturar relaciones no lineales en los datos y son una excelente opción para principiantes en el campo del Machine Learning.

Bootstrap Aggregation (Bagging)

El algoritmo Bagging se utiliza para mejorar la estabilidad y precisión de los modelos, incluyendo árboles de decisión. A través de la combinación de múltiples modelos entrenados con muestras de datos seleccionadas con reemplazo, Bagging reduce la varianza y aumenta la precisión.

Random Forest

Bosques Aleatorios son una extensión de Bagging que combina múltiples árboles de decisión mediante la aleatorización de características. Esto produce un modelo robusto y de alto rendimiento que es eficaz en la clasificación y la regresión de series de tiempo.

Redes Neuronales Recurrentes

Modelado secuencial

Los modelos de aprendizaje automático o profundo, cuya entrada o salida es una secuencia, son conocidos como modelos secuenciales. Textos, audio, vídeo, series de tiempo y otros tipos de secuencias son ejemplos de datos secuenciales.

Redes neuronales recurrentes

Las Redes Neuronales Multicapa (MLP) son estáticas y, por lo tanto, son incapaces de procesar series de tiempo o secuencias, es decir, no son capaces de capturar dependencias en los datos. Es aquí donde aparecen las Redes Neuronales Recurrentes (RNN), que son una clase de red neuronal donde las conexiones logran capturar estas dependencias a lo largo de una secuencia. Se ha demostrado que las redes neuronales recurrentes tienen la llamada propiedad de aproximación universal, es decir, son capaces de aproximar arbitrariamente sistemas dinámicos no lineales, mediante la realización de funciones complejas de la sucesión de entrada a la sucesión de salida. Sin embargo, la arquitectura particular de una RNN determina cómo fluye la información entre las diferentes neuronas y su diseño correcto es crucial para la realización de un sistema de aprendizaje robusto.

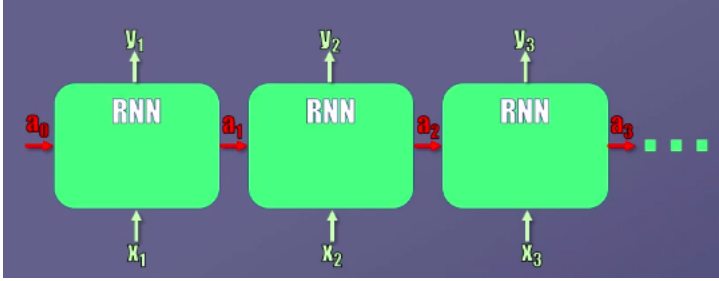


Figura 1: <https://www.youtube.com/@codificandobits>

Arquitectura

- **Uno para uno:** Una red neuronal tradicional puede ser vista como una RNN, y su arquitectura se llama uno para uno. Por ejemplo tenemos de entrada un único dato y de salida tenemos una secuencia.
- **Uno para muchos:** Este tipo de arquitectura es útil para, por ejemplo, la generación de música. La entrada es una secuencia y la salida es una categoría.
- **Muchos para muchos:**
 - Esta arquitectura es muy útil para máquinas de traducción.
- **Muchos para uno:** Este tipo de arquitectura es utilizado, por ejemplo, para la clasificación de sentimientos. A partir de una conversación entrenamos un sistema para ver si lo que quiere expresar la persona es positivo negativo o central.

El gradiente que desaparece o explota

A medida que el algoritmo de retropropagación a través del tiempo se desarrolla en una RNN, los gradientes pueden volverse cada vez más pequeños debido a que muchos de estos valores son menores que 1, lo que hace que se aproximen a cero, lo cual afecta el algoritmo en el sentido de que la actualización de los pesos no se altera y, por lo tanto, no podemos encontrar un mínimo de la función. Esto se conoce como el gradiente que desaparece (vanishing gradient).

También, a medida que el algoritmo de retropropagación a través del tiempo se desarrolla, los gradientes pueden volverse demasiado grandes debido a que muchos de estos valores son mayores que 1, lo que hace que las actualizaciones de los pesos se vean afectadas y el algoritmo de descenso del gradiente diverja. Esto se conoce como el gradiente que explota (exploding gradients).

Redes de gran memoria de corto plazo o Long Short-Term Memory (LSTM)

La arquitectura Long Short-Term Memory (LSTM) es una de las arquitecturas más utilizadas debido a su funcionamiento superior en el modelado preciso de dependencias de datos tanto a corto como a largo plazo. LSTM intenta resolver el problema del gradiente que se desvanece no imponiendo ningún sesgo hacia las observaciones recientes, pero mantiene un error constante que fluye hacia atrás en el tiempo. Una celda LSTM está compuesta por cinco componentes no lineales diferentes, que interactúan entre sí de una manera particular. El estado interno de la celda es modificado solo a través de interacciones lineales. Esto permite que la información se propague de manera suave a través del tiempo, con la consecuencia de un mejoramiento en la capacidad de memoria de la celda. Las celdas LSTM protegen y controlan la información de las celdas a través de tres puertas, las cuales son implementadas por una función sigmoide y multiplicación uno a uno. Para controlar el comportamiento de cada puerta, se entrenan un conjunto de parámetros con el descenso del gradiente, para así poder resolver una tarea objetiva.

Metodología

Las metas específicas son las siguientes:

1. Investigar y adquirir conocimientos sobre el algoritmo Random Forest, las Redes Neuronales Recurrentes y su aplicación en el análisis de criptomonedas.
2. Recopilar datos históricos del precio de Bitcoin, obtenidos de Yahoo Finance.
3. Construir modelos predictivos utilizando el algoritmo Random Forest y Redes Neuronales Recurrentes con unidades LSTM para predecir los precios futuros de Bitcoin.

Resultados

Regresión del Bosque Aleatorio:

- Buen rendimiento en la predicción de precios de Bitcoin.
- Dificultades en predecir precios extremos, especialmente cuando superan los 60,000 dólares.
- Importancia destacada de las variables OHLC del precio de Bitcoin del periodo anterior.
- Cambios en la importancia de las variables a lo largo del tiempo, reflejando la evolución del mercado.

LSTM:

- Sensible a la inclusión de variables redundantes, con mayor precisión al usar un conjunto seleccionado de variables.
- Disminución de la precisión al predecir precios por encima de los 60,000 dólares, más evidente en el Periodo 2.
- Mayor precisión en el Periodo 1, sugiriendo menor correlación con los mercados tradicionales y mayor aleatoriedad en los precios.
- La precisión disminuye al agregar más periodos anteriores, respaldando la hipótesis de eficiencia de mercado.

Comparación:

- La regresión del bosque aleatorio muestra un rendimiento equivalente o superior a LSTM en la predicción del precio de Bitcoin.
- Importancia de seleccionar variables relevantes para mejorar la precisión en ambos modelos.
- Ambos modelos presentan desafíos en la predicción de precios extremos y muestran una relación inversa entre la precisión y el número de periodos anteriores considerados.

Referencias

- [1] Carrión Pérez, Rafael. (2020). *Predicción de Precios de Criptomonedas con Random Forest*, <https://oa.upm.es/62842/>, UPM.
- [2] Chen, Junwei. (2023). *Analysis of Bitcoin Price Prediction Using Machine Learning*. MPDI, <https://www.mdpi.com/1911-8074/16/1/51>
- [3] Abhijit, Roy. (2020). *A Dive Into Decision Trees*. towardsdatascience, <https://towardsdatascience.com/a-dive-into-decision-trees-a128923c9298>
- [4] Journal of Risk and Financial Management. MDPI. <https://www.mdpi.com/journal/jrfm>
- [5] Yahoo! en español. <https://espanol.yahoo.com/>.
- [6] Nakamoto, Satoshi. 2008. *Bitcoin: un sistema de efectivo electrónico entre pares*. Disponible en línea: <https://bitcoin.org/bitcoin.pdf>