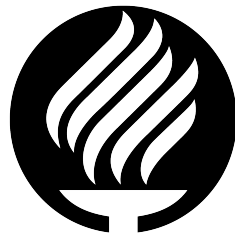


INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY
CAMPUS QUERÉTARO
DEPARTAMENTO DE COMPUTACIÓN Y MECATRÓNICA



Tecnológico
de Monterrey

Predicción del valor de criptomonedas usando el modelo híbrido
ARIMA-LSTM

por

[Alan Macías](#)

Proyecto Integrador para el Desarrollo de Soluciones Empresariales

Ingeniería
en
Sistemas Computacionales

Asesor: Dr. Benjamín Valdés
Co-Asesor: Dr. Alfonso Gómez

Santiago de Querétaro, Querétaro, México
07/05/2020

Abstract

El análisis de series temporales es utilizado comunmente para hacer proyecciones de precios, incluyendo el de las criptomonedas. Implementar modelos que hagan predicciones precisas no es una tarea sencilla, especialmente para mercados tan volátiles. Se desarrolla un modelo híbrido ARIMA-LSTM que es evaluado(RECM) contra otros modelos ya utilizados para predecir el precio de criptomonedas. Se obtuvieron resultados que demuestran que el modelo híbrido se aproxima mejor que los otros modelos al valor real de los precios.

Índice general

Abstract	I
1. Introducción	1
1.1. Contexto del problema	1
1.2. Motivación	1
1.3. Objetivo del trabajo	2
1.4. Estructura del trabajo	2
2. Estado del Arte	3
2.1. Contexto del Área	3
2.1.1. Series de tiempo	3
2.1.1.1. Definición	3
2.1.1.2. Visualización	4
2.1.1.3. Análisis y Pronóstico de series de tiempo	4
2.2. Trabajos Relevantes	5
2.2.1. ARIMA	5
2.2.1.1. Descripción del modelo	5
2.2.1.2. Pronósticos del precio de Bitcoin	5
2.2.2. LSTM	5
2.2.2.1. Descripción del modelo	5
2.2.2.2. Pronósticos del precio de Bitcoin	6
2.2.3. ARIMA-LSTM	6
2.2.3.1. Descripción del modelo	6
2.2.3.2. Pronósticos hechos con ARIMA-LSTM	7
2.3. Conclusión	8

Capítulo 1

Introducción

En este trabajo se presenta la investigación e implementación de un modelo híbrido de series temporales utilizado para hacer proyecciones del precio de criptomonedas (Bitcoin y Litecoin). El desarrollo del proyecto implica la implementación de dos modelos (ARIMA y LSTM) que se han utilizado actualmente para hacer proyecciones de criptomonedas y se evalúan los resultados obtenidos por dichos modelos contra un tercer modelo híbrido ARIMA-LSTM que a su vez se implementa.

1.1. Contexto del problema

Las criptomonedas son monedas digitales que operan de forma descentralizada, permitiendo que los usuarios lleven a cabo transacciones sin que éstas dependan de un intermediario. Las criptomonedas, y en especial Bitcoin que es la más popular, atraen cada vez más la atención haciendo que muchas personas opten por invertir en ellas y que ciertos negocios comiencen a aceptarlas como forma de pago [1]. Las criptomonedas se describen como rápidas, convenientes, libres de impuesto, e incluso revolucionarias [1], sin embargo, también son conocidas por su alta volatilidad debido a distintos factores como el tener un mercado relativamente pequeño, baja liquidez, gran cantidad de especulaciones respecto a su futuro, entre otros. Esta naturaleza volátil, hace que predecir los precios de criptomonedas sea una tarea complicada [2].

1.2. Motivación

La alta volatilidad de las criptomonedas hace que sean una inversión bastante riesgosa, sin embargo, actualmente se llevan a cabo más de 250, 000 transacciones de criptomonedas al día [3]. El desplome del valor de Bitcoin más reciente ocurrió el 12 de Marzo, reportando pérdidas de

prácticamente el 50 % para los inversionistas en tan solo un día [4]. A partir de datos como éstos, nace la motivación para realizar este trabajo, que busca explorar la posibilidad de acercarnos a proyecciones más precisas y así reducir el riesgo al invertir en criptomonedas.

1.3. Objetivo del trabajo

El objetivo de este trabajo es determinar si las proyecciones del precio de criptomonedas hechas por el modelo híbrido ARIMA-LSTM son más certeras que las de los modelos ya hechos anteriormente.

1.4. Estructura del trabajo

Este trabajo está conformado por 7 capítulos que detallan la investigación hecha y posteriormente la implementación de los modelos.

En el capítulo 2 se habla del Estado del Arte. Se explican conceptos relevantes para el proyecto en general, así como otras soluciones propuestas para la problemática.

En el capítulo 3 se describe más a detalle la problemática actual y del porqué es importante seguir buscando soluciones alternas.

En el capítulo 4 se habla a profundidad de la propuesta de solución, explicando tanto los aspectos y conceptos relevantes para su desarrollo, como el funcionamiento de los modelos a implementar.

En el capítulo 5 se define la metodología de evaluación que se utiliza, así como la justificación para la elección de la misma.

En el capítulo 6 se muestran los resultados obtenidos al evaluar cada modelo. Se aprecian visualmente las proyecciones de precios hechas por cada modelo y los precios reales para compararlos.

En el capítulo 7 se presentan las conclusiones basadas en los resultados obtenidos, se retoman puntos importantes del trabajo y se plantean posibles trabajos futuros teniendo ya una base.

Capítulo 2

Estado del Arte

En este capítulo se describen conceptos relevantes en el área. Posteriormente se hace mención a los modelos ARIMA y LSTM, que se han implementado para hacer pronósticos del precio de las criptomonedas, así como los resultados que obtuvieron. Analizar estos trabajos permite encontrar un área de oportunidad a la cual se puede contribuir.

2.1. Contexto del Área

2.1.1. Series de tiempo

2.1.1.1. Definición

Una serie de tiempo es un conjunto secuencial de datos medidos típicamente en tiempos sucesivos. Las medidas que se toman durante cierto periodo se organizan en orden cronológico [5]. Se define matemáticamente una serie de tiempo como:

$$\{X_1, X_2, \dots, X(t)\}, t = 1, 2, 3, \dots$$

t = Tiempo transcurrido

$X(t)$: Valor del dato medido durante un tiempo en específico.

Las series de tiempo se denominan univariantes si los registros que se tienen a lo largo del tiempo son de una sola variable, pero si los registros son de más de una variable entonces se dice que es multivariante. [5, 6] Un ejemplo de serie de tiempo univariante es el medir la temperatura de alguna máquina de forma periódica cada minuto, por otro lado si se midiera tanto la temperatura como la vibración y las revoluciones, se estaría hablando de una serie de tiempo multivariante.

2.1.1.2. Visualización

Para visualizar los datos de una serie de tiempo, generalmente éstos se grafican trazando las observaciones con respecto al tiempo. Esto permite tener una primer aproximación y entendimiento sobre el cómo se comporta la variable a lo largo del tiempo.

La figura 2.1 muestra la visualización de una serie de tiempo, donde se aprecian los precios de Bitcoin (en dólares) desde Enero del 2015 hasta Septiembre del 2018. El eje x representa el tiempo, y el eje y el precio de la criptomoneda.



FIGURA 2.1: Visualización de precios históricos de Bitcoin [11].

2.1.1.3. Análisis y Pronóstico de series de tiempo

El pronóstico de series de tiempo consiste en recolectar datos históricos de alguna variable de interés para posteriormente ser analizados y desarrollar un modelo que comprenda las relaciones que existen en los datos y que posiblemente no sean notorias. El modelo después es utilizado para hacer pronósticos del comportamiento que se espera que tenga la variable recolectada [7, 8].

2.2. Trabajos Relevantes

2.2.1. ARIMA

2.2.1.1. Descripción del modelo

De todos los métodos que existen actualmente para hacer pronósticos de series de tiempo, el modelo autorregresivo integrado de promedio móvil (ARIMA) es uno de los más utilizados. Este modelo tiene una característica relativamente sencilla a grandes rasgos, y que a su vez representa una deficiencia: ARIMA asume que los valores futuros y los valores históricos se relacionan linealmente, cuando la realidad es que la mayoría de los datos de series temporales presentan una relación no lineal [9].

2.2.1.2. Pronósticos del precio de Bitcoin

El modelo ARIMA es una elección muy popular para realizar pronósticos de criptomoneda, especialmente a corto plazo. Como se describe en [10], se implementó este modelo utilizando los datos históricos de Bitcoin desde Mayo del 2016, hasta finales de Marzo del 2018. Los datos obtenidos contenían información sobre el precio de apertura y cierre, el alto y bajo de cada día, entre otros [10]. Para preprocesar los datos, se removieron todas las columnas a excepción del precio de cierre, y se utilizó este dato porque es el que genera más interés para los inversionistas [11].

Para implementar el modelo ARIMA, se necesitan definir 3 valores (p , d , q) utilizados para la construcción del mismo, y ya que seleccionar los valores que encajan mejor en el modelo no es una tarea sencilla, el desarrollo se realizó en el lenguaje de programación R, que cuenta con una función que encuentra los valores óptimos de forma automática [10].

Las pruebas se realizaron pronosticando los precios de cierre desde el 13 de Diciembre del 2017, hasta el 27 de Marzo del 2018, y posteriormente evaluando los precios arrojados por el modelo contra los precios reales de las mismas fechas utilizando la raíz del error cuadrático medio (RECM). La evaluación resultó en un error de 593.80 [10].

2.2.2. LSTM

2.2.2.1. Descripción del modelo

Las Long Short-Term Memory Networks (LSTM) son un tipo de redes neuronales compuestas por una capa de entrada, una o más capas escondidas, y una capa de salida. Lo que diferencia

a las LSTM de las redes neuronales convencionales, es la estructura de sus capas escondidas que les permite recordar información por largos periodos de tiempo [12, 13]. Dentro de la capa escondida de una LSTM se encuentran celdas de memoria, que a su vez cuentan con 3 tipos de compuertas: La compuerta de entrada, la de salida y la de olvido. A grandes rasgos estas compuertas mantienen y ajustan el estado de la celda, es decir, la información que ésta contiene [12].

Este modelo a diferencia del anterior, es capaz de ajustarse a relaciones no lineales en las secuencia de datos, lo que permite generar pronósticos más confiables cuando son a plazos más largos [13].

2.2.2.2. Pronósticos del precio de Bitcoin

Wu, C.-H., et al. [12] implementaron una LSTM utilizando los precios de Bitcoin desde el inicio del 2018, hasta el 28 de Julio del mismo año. El preprocesamiento de los datos fue muy similar al mencionado previamente, quedándose únicamente con el precio de cierre para el entrenamiento de los modelos y los pronósticos a generar.

Para realizar las pruebas se utilizaron los últimos 71 registros de las series de tiempo, teniendo un total de 138 muestras para entrenar al modelo. La raíz del error cuadrático medio fue una de las metodologías de evaluación usadas, resultando en un error total de 247.33 [12].

2.2.3. ARIMA-LSTM

2.2.3.1. Descripción del modelo

El modelo híbrido ARIMA-LSTM combina ambos modelos aprovechando las ventajas que cada uno ofrece, buscando lograr pronósticos más precisos. Primero se utiliza el modelo ARIMA para ajustar las relaciones lineales en los datos y posteriormente el modelo LSTM se encarga de las relaciones no lineales [9].

La secuencia de pasos para hacer pronósticos de series de tiempo con el modelo ARIMA-LSTM se representa en la figura 2.2. A grandes rasgos, los pasos son los siguientes:

1. Entrenar el modelo ARIMA con los datos históricos de las series de tiempo.
2. Obtener la secuencia residual resultante.
3. Ajustar el modelo LSTM utilizando los residuos de ARIMA.
4. Pronosticar con el modelo ARIMA del paso 1.

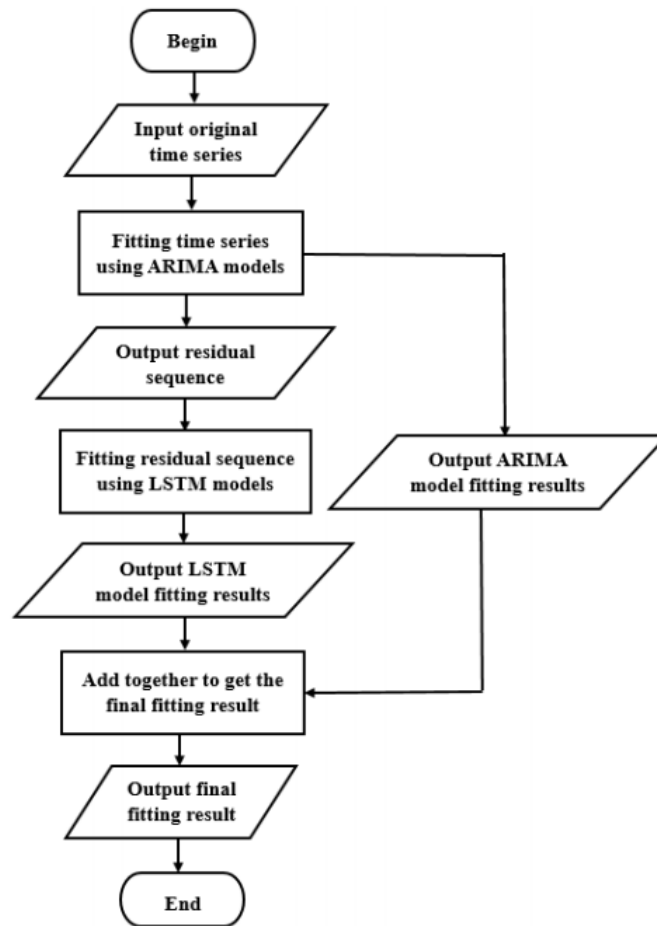


FIGURA 2.2: Diagrama de flujo del modelo ARIMA-LSTM [14].

5. Pronosticar con el modelo LSTM del paso 3.
6. Sumar ambos resultados para obtener el pronóstico final [14].

2.2.3.2. Pronósticos hechos con ARIMA-LSTM

Wang, Z., & Lou, Y. propusieron un modelo de series temporales ARIMA-LSTM para pronosticar el nivel de agua promedio de una estación hidrológica ubicada en el río Chuhe. Para ponerlo a prueba, desarrollaron los modelos de forma independiente y de forma híbrida para ver si en conjunto conseguían mejores resultados. Sus experimentos mostraron que las ventajas de ambos modelos se complementan, y logran pronósticos más precisos [9].

Liu, J., Tan, X., & Wang, Y. desarrollaron el modelo ARIMA-LSTM para pronosticar el envejecimiento del software para servicios en la nube. Su propuesta (muy similar a la anterior) buscaba comprobar si se podían obtener mejores resultados usando el modelo híbrido. Para comprobarlo, obtuvieron registros diarios del Centro de Datos de Google Cloud, teniendo acceso al uso del CPU y la memoria de más de 1600 máquinas. Después de evaluar los tres modelos, la raíz del

error cuadrático medio de ARIMA-LSTM fue la menor de todas, concluyendo que se puede reducir el error y hacer mejores pronósticos con el modelo híbrido [14].

2.3. Conclusión

Actualmente existen varios métodos utilizados para pronosticar el precio de criptomonedas. Los trabajos previamente mencionados muestran el uso de distintos modelos de series temporales, así como las evaluaciones de los mismos. El punto más interesante a notar es el resultado que se obtiene al comparar el modelo híbrido ARIMA-LSTM contra ellos mismos de forma individual, ya que se han hecho pronósticos con un menor margen de error.

A pesar de los resultados favorables, el modelo ARIMA-LSTM no se ha utilizado comúnmente para hacer pronósticos de series de tiempo, y en particular, no se ha utilizado para pronosticar el precio de criptomonedas. A partir de esto se detecta un área de oportunidad, ya que es posible que el modelo híbrido también sea una mejor alternativa que ARIMA y que LSTM, misma que no se ha puesto a prueba aún.

Bibliografía

- [1] M. Rahouti, K. Xiong and N. Ghani, "Bitcoin Concepts, Threats, and Machine-Learning Security Solutions,in IEEE Access, vol. 6, pp. 67189-67205, 2018.
- [2] Asante, S. Are Bitcoins price predictable? Evidence from machine learning techniques using technical indicators. 2019. pp. 2.
- [3] Beck et al. Sensing Social Media Signals for Cryptocurrency News . 2019. pp. 1-2.
- [4] Bovaird, C. (2020). Bitcoin Lost Roughly 50 % Of Its Value In A Day. Mayo 4, 2020, de Forbes Sitio web: <https://www.forbes.com/sites/cbovaird/2020/03/12/bitcoin-lost-roughly-50-of-its-value-in-a-day>
- [5] Adhikari, R; Agrawal R. An Introductory Study on Time Series Modeling and Forecasting. pp. 12-13.
- [6] John H. Cochrane, "Time Series for Macroeconomics and Finance", Graduate School of Business, University of Chicago, spring 1997.
- [7] G.P. Zhang, "A neural network ensemble method with jittered training data for time series forecasting", Information Sciences 177 (2007), pp: 5329–5346.
- [8] G.P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing 50 (2003), pp: 159–175.
- [9] Wang, Z., & Lou, Y. (2019). Hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).
- [10] Yenidogan, I., Cayir, A., Kozan, O., Dag, T., & Arslan, C. (2018). Bitcoin Forecasting Using ARIMA and PROPHET. 2018 3rd International Conference on Computer Science and Engineering (UBMK).
- [11] Anupriya, & Garg, S. (2018). Autoregressive Integrated Moving Average Model based Prediction of Bitcoin Close Price. 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT).

-
- [12] Wu, C.-H., Lu, C.-C., Ma, Y.-F., & Lu, R.-S. (2018). A New Forecasting Framework for Bitcoin Price with LSTM. 2018 IEEE International Conference on Data Mining Workshops (ICDMW).
 - [13] S. Tandon, S. Tripathi, P. Saraswat and C. Dabas, "Bitcoin Price Forecasting using LSTM and 10-Fold Cross validation," 2019 International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2019, pp. 323-328.
 - [14] Liu, J., Tan, X., & Wang, Y. (2019). CSSAP: Software Aging Prediction for Cloud Services Based on ARIMA-LSTM Hybrid Model. 2019 IEEE International Conference on Web Services (ICWS).