

Predicción del precio diario de Cierre y dirección del Bitcoin mediante Modelos de Regresión Lineal Múltiple y Logísticos.

Pablo Martinez Angerosa, Vanessa Alcalde

Resumen

En 2008 Satoshi Nakamoto presentó una tecnología revolucionaria llamada Bitcoin. En evolución constante, el Bitcoin ya forma parte del interés diario de inversores y el mundo de la tecnología. Aquí presentamos una investigación que mediante la comparación de distintas técnicas de Machine Learning (Linear Models, Ridge, GAM, Logistic) establece a los modelos de regresión lineal y logística como la opción candidata para la predicción del precio de Bitcoin.

1 Introducción

Satoshi Nakamoto en 2008 presentó una tecnología revolucionaria[3], que cambiaría la historia de la humanidad por siempre, mediante la publicación de un artículo que describía un sistema P2P (peer-to-peer) de dinero digital, llamado Bitcoin.

El Bitcoin no es un simple cambio de modalidad, es un cambio de paradigma, que ha cobrado un impulso tal que, ha dejado de ser un tema de entusiastas en la criptografía a ser una realidad diaria, una opción de inversión real y parte de las noticias. Asimismo es el impulsor de una nueva tecnología llamada la cadena de bloques [4] (Blockchain) que es el fundamento tecnológico de las criptomonedas, y que

actualmente esta revolucionando y ampliando las posibilidades del mundo tecnológico mas allá de las criptos y el ecosistema fintech. [1].

Esta moneda digital, fue creada en su propia arquitectura como una red descentralizada, un libro abierto de balance contable, donde todas las transacciones son públicas y verificadas mediante un proceso criptográfico realizado por los nodos de la misma red (miners), sin la necesidad de una casa centralizadora o un tercer interesado como agente de control y validación[3].

El objetivo de este paper de investigación es encontrar un modelo de predicción de la dirección y el precio diario de cierre del Bitcoin mediante la utilización de modelos de Aprendizaje Au-

tomático.

Para esto se utilizaron modelos de regresión dado que la variable de interés precio es continua y también se utilizaron modelos de regresión logística para la predicción de la dirección del precio que es una variable categórica. Para los modelos de regresión continua se testearon técnicas de Regresión Lineal Múltiple, Regresión Ridge y Modelos Aditivos Generalizados. Estos modelos representan distintos grados de flexibilidad del algoritmo de predicción.

Existen actualmente diversos papers para la predicción del precio del Bitcoin, pero comparado con predicciones tradicionales de Stock Market esta sigue siendo un área inexplorada con mucho potencial y margen de investigación. La lectura de estos papers fue fundamental para la elección de variables y la selección de las técnicas aplicadas [12].

En la segunda sección de este paper detallamos la obtención y tratamiento de los datos, en la tercer sección describimos los modelos utilizados, en la cuarta sección presentamos los resultados y finalmente en la quinta sección mostramos las conclusiones de esta investigación y un breve enfoque de trabajos a futuro.

2 Datos

2.1 Descripción de los datos

Para la construcción de la base de datos utilizamos los datos diarios del opening price, closing price, volumen transactions, provistos por Cryptodatadownload¹ que

¹<https://www.cryptodatadownload.com>

mantiene actualizada las bases de precios diarios de las principales criptomonedas y los principales exchanges. Para esta investigación se utilizaron los datos del exchange Coinbase² que es considerado uno de los exchanges más seguro y estable por la comunidad cripto, el cual permite la comercialización y respaldo de criptomonedas como Bitcoin, Ethereum y cuya sede se encuentra basada en USA.

Se utilizaron los registros diarios en dolares americanos (USD) de las criptomonedas Bitcoin(BTC), Ethereum(ETH), Litecoin(LTC), Bitcoin Cash(BCH), Ethereum Classic(ETC), Chainlink(LINK) y Augur (REP).

También se utilizaron los datos diarios provistos en Google Trends³ que reflejan el grado de interés diario de la búsqueda de la palabra clave Bitcoin en el buscador de Google, siendo este el principal motor de búsqueda en la actualidad. El índice que provee este dataset mide el interés relativo de una búsqueda en una región dada y un intervalo de tiempo determinado. Según las explicaciones de Google en su web un valor de 100 significa un pico en la popularidad del término buscado, un valor de 50 significa que el término es medianamente popular. Un resultado de 0 significa que no hay búsquedas suficientes en ese término.

2.2 Preparación de los datos

El dataset contiene 232 días comenzando en el día 22/1/2020 hasta el 9/9/2020.

Todas las bases provistas por Crypto-

²<https://www.coinbase.com>

³<https://trends.google.com/trends>

datadownload se encuentran en formato de archivo .CSV y no requirieron mayor preparación ya que no existían datos faltantes y el formato es ampliamente utilizado.

Para armar la variable de Google Trends fue necesario la implementación recursiva de la descarga de la base en distintos frames de tiempo para una unión posterior de los datos. Esto es debido a que la información diaria provista por Google tiene un marco de ventana de dos meses.

La base se estructuró de modo de tener un conjunto de entradas X y salidas Y con una dependencia temporal. La variable de predicción Y corresponde al *Close* del precio del Bitcoin en el tiempo n . Las entradas correspondientes a X se encuentran en diversos tiempo del pasado.

Todos los *Open* de precios se encuentran en el mismo tiempo n que se quiere predecir, ya que para realizar la predicción es información que existe en ese momento.

Todas las variables *Volume* correspondientes a las distintas monedas corresponden al tiempo $n - 1$ ya que es información que no existe a la hora de predecir el momento n .

También se crearon variables llamadas *lag*, para el precio de *Close* y *Volume* de Bitcoin. Los números utilizados para nombrar las variables de *lag*, representan el tiempo del pasado. Por ejemplo la variable *btc_close_lag3* representa el precio del *Close* del Bitcoin 3 días previos al momento n . [9]

En la base se incluyó una variable cualitativa que refleja la dirección del precio. Para obtener la dirección se resta el precio del *Close* del Bitcoin y el precio del *Open* del Bitcoin en el momento n para cada una

de las entradas de la base de datos. Esta variable cualitativa que toma valores *Up* y *Down* es la variable a predecir mediante un modelo de regresión logística.

3 Los modelos

Los cuatros modelos utilizados en esta investigación incluyen regresión lineal múltiple, regresión Ridge, *GAM* polinómico y un modelo de regresión logística. Todos estos modelos se utilizaron exclusivamente desde una perspectiva de predicción del precio *Close* del Bitcoin por lo que los modelos que obtuvieron una mejor performance no son necesariamente los mejores modelos para hacer inferencia y determinar la causa y efecto entre la variables independientes y la explicada.

Un modelo de regresión lineal múltiple es un modelo lineal en los parámetros en el cual la variable de respuesta, Y , es determinada por un conjunto de variables independientes, las variables explicativas (matriz X). Se busca el hiperplano que mejor ajuste a los datos. Los parámetros β_i para el modelo se obtienen por mínimos cuadrados.

Para el caso de la regresión de Ridge, el modelo es ajustado incorporando un factor de penalización, λ , en la función de pérdida. El factor se obtiene mediante cross-validation.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d + \epsilon_i \quad (1)$$

En cuanto a Modelos Aditivos Generalizados *GAM* (1) nosotros nos enfocamos en la regresión polinómica, esta incorpora

potencias a las variables explicativas. El modelo es lineal en los parámetros y se estiman mediante mínimos cuadrados. Un mayor grado implica mayor flexibilidad en el modelo.

Un modelo de regresión logística es un modelo lineal generalizado con una variable explicada Y binaria, donde $Y \sim B(1, p)$, donde p es la probabilidad de que ocurra el evento y la función de enlace es la función link. Este se utilizó para predecir la dirección diaria del precio del *Close* del Bitcoin.

$$\begin{aligned} btc_close_i = & \beta_0 + \beta_1 btc_close_lag1_i + \\ & \beta_2 btc_close_lag2_i + \beta_3 btc_close_lag4_i + \\ & \beta_4 btc_vol_lag1_i + \beta_5 btc_vol_lag3_i + \\ & \beta_6 btc_vol_lag4_i + \beta_7 btc_trend_i + \\ & \beta_8 ltc_open_i + \varepsilon_i \end{aligned} \quad (2)$$

En la ecuación (2) se presenta el modelo de regresión lineal múltiple que es el que logró mejor performance de predicción entre todos los modelos evaluados. La variable explicada *Close* se predice por una combinación lineal de 3 *lags* de *Close* (*lag1*, *lag2*, *lag4*), 3 *lags* del *Volume* (*lag1*, *lag3*, *lag4*), la variable correspondiente a la tendencia relativa de búsqueda del termino “Bitcoin” en el buscador de Google y el precio del *Open* de la criptomoneda Lite Coin (*LTC*).

Para encontrar el modelo con mejor performance se construyó un algoritmo de fuerza bruta que evalúa secuencialmente una a una todas las posibles combinaciones de variables para generar un modelo de predicción. Este modelo se corrió

en distintas combinaciones fijadas manualmente (por intento y error) de posibles grupos de variables, ya que el orden de todas las combinaciones es de 2^n para n opciones de variables. Y en $n > 20$ el algoritmo se vuelve incomputable.

$$(min)RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - f_i)^2} \quad (3)$$

Para evaluar la performance de los distintos modelos de predicción se utilizó la medida de error *RMSE* como se ve en la ecuación (3) donde y_i representa las predicciones y f_i los valores reales. Se busca minimizar el *RMSE*.

Para la predicción logística de la dirección del precio *Close* del Bitcoin se utilizan las mismas variables que el modelo presentado en la ecuación (2) cambiando la variable y_i explicada por las direcciones del *Close*.

El dataset se dividió aleatoriamente, en una muestra de entrenamiento del 30% y una muestra de testeo del 70%.

4 Resultados

El objetivo de este paper es evaluar la predicción del precio *Close* del Bitcoin mediante algoritmos de Machine Learning.

La Tabla 1, muestra la performance de cada uno de los modelos seleccionados por el algoritmo de fuerza bruta para las distintas técnicas empleadas (Regresión Lineal Multiple, Regresión Ridge, GAM Poly). En este caso el modelo seleccionado para la técnica clásica de Regresión Lineal

Tabla 1: Resultados de RMSE para el precio *Close* de Bitcoin

Algoritmo de predicción	RMSE
Regresión Lineal Múltiple	291.971
Regresión Ridge	302.808
<i>GAM</i> Poly 2	295.932
<i>GAM</i> Poly 3	307.602
<i>GAM</i> Poly 4	311.713

Múltiple es el que logra mejores resultados. Esto concuerda con la literatura existente, y a la vez la tabla muestra, como la flexibilización del algoritmo, no necesariamente arroja mejores resultados.

En cuanto a la predicción logística de la dirección del precio *Close* del Bitcoin, el modelo seleccionado logra un grado de sensibilidad del 56% en la tendencia alcista. Es decir si el algoritmo predice una suba, este acierta el 56% de las veces. En términos de especificidad el modelo muestra una performance del 60,53%. Se probaron modelos más complejos con interacciones resultantes del algoritmo genético GLMULTI, pero no mostraron una mejor performance.

5 Conclusiones

Estos resultados sugieren que modelos de predicción basados en regresión lineal múltiple pueden proponerse como solución a la predicción del precio del Bitcoin y los modelos de regresión lineal logística pueden ser una solución para la predicción de la dirección del precio.

A futuro se plantea la posibilidad de utilizar otros algoritmos de Machine Learning, como SVM, Random Forest, Boost-

ing, Árboles de clasificación y regresión y Redes Neuronales.

También evaluar nuevas variables para el modelo, incluyendo algunas como Twitter Sentiment Analysis del Bitcoin, y otras variables que se sugieren en la literatura como el índice de dificultad del minning para Bitcoin.

Así mismo se plantea la posibilidad de utilizar un algoritmo más avanzado de selección de variables basado en la tecnología de algoritmos genéticos o reinforcement learning donde se efectuó la optimización del RMSE en la búsqueda del mejor modelo dentro del algoritmo.

Bibliografía

- [1] Sangoi, F. *Bitcoin. ¿Una revolución monetaria?* Universidad de Buenos Aires.
- [2] Azim Muhammad Fahmi, Noor Azah Samsudin, Aida Mustapha, Nazim Razali, Shamsul Kamal Ahmad Khalid. (2018). *Regression based Analysis for Bitcoin Price Prediction*. Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia.
- [3] Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*.
- [4] Carlos Dolader Retmal et al. *La blockchain: fundamentos, aplicaciones y relación con otras tecnologías disruptivas*, Universitat Politècnica de Catalunya.

- [5] Karame, G., Huth, M., Vishik, C. (2020). *An overview of blockchain science and engineering*. R. Soc. Open Sci. 7: 200168. <http://dx.doi.org/10.1098/rsos.200168>
- [6] Shah, D., Zhang, K. (2014). *Bayesian regression and Bitcoin*. Laboratory for Information and Decision Systems Department of EECS, Massachusetts Institute of Technology.
- [7] Ferdiansyah, Siti Hajar Othmanb, Raja Zahilah Raja Md Radzic, Deris Stiawan. (2019). *A Study of Bitcoin Stock Market Prediction: Methods, Techniques and Tools*.
- [8] Wheatley, S., Sornette, D., Huber, T., Reppen, M., Gantner, RN. (2019). *Are Bitcoin bubbles predictable? Combining a generalized Metcalfe's Law and the Log-Periodic Power Law Singularity model*. R. Soc. open sci. 6: 180538. <http://dx.doi.org/10.1098/rsos.180538>
- [9] Uras, N., Marchesi, L., Marchesi, M., Tonelli, R. (2020) *Forecasting Bitcoin closing price series using linear regression and neural networks models*. Department of Mathematics and Computer Science, University of Cagliari.
- [10] Asante Gyamerah, S. (2020). *On forecasting the intraday Bitcoin price using ensemble of variational mode decomposition and generalized additive model*. Pan African University, Institute for Basic Sciences, Technology, and Innovation. <https://doi.org/10.1016/j.jksuci.2020.01.006>
- [11] McNally, S. (2016). *Predicting the price of Bitcoin using Machine Learning*. School of Computing National College of Ireland.
- [12] Kristoufek, L. (2015). *What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis*. PLoS ONE 10(4): e0123923. doi:10.1371/journal.pone.0123923
- [13] Rajua, S., Mohammad, A. (2020). *Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis*. Computer Science, International Islamic University Malaysia.
- [14] Burnie, A., Yilmaz, E. (2019). *Social media and bitcoin metrics: which words matter*. R. Soc. open sci. 6: 191068. <http://dx.doi.org/10.1098/rsos.191068>
- [15] Garcia, D., Schweitzer, F. (2015). *Social signals and algorithmic trading of Bitcoin*. R. Soc. open sci. 2: 150288. <http://dx.doi.org/10.1098/rsos.150288>
- [16] Matta, M.; Lunesu, I. and Marchesi, M. (2015). *The Predictor Impact of Web Search Media On Bitcoin Trading Volumes*. Università degli Studi di Cagliari.

- [17] Hakim, R. (2020) *Bitcoin pricing: impact of attractiveness variables*. Sao Paulo School Of Economics.
<https://doi.org/10.1186/s40854-020-00176-3>