Universitat de les Illes Balears

Modelado predictivo de la dirección del precio del Bitcoin utilizando índices de mercado, análisis de sentimientos en Twitter e índices de popularidad por término mediante Google Trends

Trabajo final de màster entregado a la Universitat de les Illes Balears de acuerdo con los requisitos del **Màster Universitari en Anàlisi de Dades Massives en Economia i Empresa (MADM)**

Autor

Justo Herrero Amorós

Tutor

Pere Antoni Palmer Rodríguez

01/09/2018

Modelado predictivo de la dirección de la cotización del Bitcoin utilizando índices de mercado, análisis de sentimientos en Twitter e índices de popularidad por término mediante Google Trends

Justo Herrero Amorós
Tutor: Pere Antoni Palmer Rodríguez
Treball de fi de Màster Universitari en Anàlisi de Dades Massives en Economia i Empresa (MADM)
Universitat de les Illes Balears
07122 Palma de Mallorca
jherreroa23@gmail.com;pere.palmer@uib.es

Resumen

El propósito de este artículo es exponer el estudio realizado acerca de la predicción de la evolución de la cotización del Bitcoin. Se describe el proceso de obtención de datos mediante técnicas de *scraping* sobre diferentes fuentes de datos utilizadas: Binance, Twitter y Google Trends. Se han tenido en cuenta diferentes índices de mercado que rigen la criptomoneda, como son el volumen y precios de apertura o cierre, entre otros. Además, se han tratado índices de polaridad de sentimientos y popularidad a partir de Twitter y Google Trends respectivamente. A partir de los datos obtenidos, se han aplicado procesos de *feature engineering* y minería de textos que han permitido la obtención de información implícita sobre la original que complementa y potencia el conjunto de datos. El modelado de los datos se ha realizado utilizando técnicas estadísticas, como son redes neuronales del tipo *Long-Short Term Neural Network* u otras cómo por ejemplo *Gradient Boosting*.

Abstract

The purpose of this paper is to present a study conducted around the price development of Bitcoin. The process starts by collecting data using Scraping techniques on different data sources: Binance, Twitter and Google Trends. We have been considering different market indexes that govern the cryptocurrency, as they might be the volume and the opening or closing prices. In addition, the polarity indexes and popularity have been treated, respectively from Twitter and Google Trends. The data obtained has been analyzed using feature engineering and data mining techniques, and new information has been obtained complementing and improving original data. Data modeling has been done using neural networks like Long-Short Term or other such as Gradient boosting.

Palabras clave: Criptomoneda, Bitcoin, Binance, Twitter, Google Trends, Análisis de Sentimientos, *Scraping*, *Gradient Boosting*, *LSTM*, *Random Forest*, Regresión Logística, *Boosting*

1. Introducción

Los grandes avances tecnológicos en materias relacionadas con la computación y comunicaciones han permitido en poco tiempo un gran auge en el ámbito de las criptomonedas, de manera que en los últimos años su presencia en todos los medios ha aumentado de forma notable. La aparición del concepto *Blockchain*, también conocido como cadena de bloques, ha permitido la aparición entre otras monedas del Bitcoin, Ethereum y Litecoin, con la intención de acabar con algunos de los grandes problemas a los que se enfrenta el actual sistema financiero como son la economía sumergida y el blanqueo de capitales ya que como dijo Dee Hock, fundador de Visa: "*We live in the 21st century but are still using command and control organizational structures from the 16th century. Bitcoin is one of the best examples of how a decentralized, peer-to-peer organization can solve problems that these dated organizations cannot. Like the Internet, Bitcoin is not owned or controlled by any one entity, so it presents incredible opportunities for new levels of efficiency and transparency in financial transactions.*".

En realidad, las criptomonedas tienen un valor monetario (sean dólares, euros u otras monedas convencionales) que va fluctuando al igual que lo hacen los activos financieros en el mercado de valores. Sin embargo, cabe destacar que el comportamiento que tienen las criptomonedas difiere mucho al de un activo financiero corriente, pues la volatilidad y la especulación en torno a ellas es extrema. Lo más similar que podemos encontrar a día de hoy en el mercado de valores son los activos de capital riesgo cuyo índice de volatilidad es bastante elevado y las fluctuaciones que se producen son mayores.

Este punto es especialmente crítico a la hora de realizar procesos de aprendizaje estadístico pues los cambios que se pro-

ducen son extremos, con lo que en la mayoría de las veces se dificulta la obtención de un modelo estable. Aún así, existen indicadores técnicos utilizados en el proceso de análisis técnico que permiten anticiparse, en algunos casos, a la dirección que tomará la cotización del activo [4].

A día de hoy existen aplicaciones con una fuente de datos lo suficientemente completa y sobre las que se han extraído indicadores técnicos útiles a la hora de predecir la cotización utilizando redes neuronales u otros métodos como, por ejemplo, se puede observar en la web de Neurobot (https://neurobot.trading) . La propuesta de este estudio, es determinar el impacto social sobre la cotización de las criptomonedas a partir de un muestreo sobre los tuits publicados en Twitter para poder determinar el sentir general de las personas y conocer si los inversores tienen un sentimiento de compra / venta, ya que no olvidemos que la cotización de las criptomonedas se rige por las leyes de oferta / demanda de manera muy semejante a como lo hacen los mercados financieros convencionales.

Además, se ha tomado en consideración la información de otra fuente de datos que permite complementar la actividad de un activo: su índice de búsquedas en Google. De esta manera podemos hacer un seguimiento en tiempo real de la popularidad de una criptomoneda y ver la incidencia que tiene este concepto en su cotización.

2. Proceso de obtención de datos, tratamiento de datos y *feature engineering*

El proceso de obtención de datos es uno de los puntos críticos pues de la información que recopilemos dependerá, en gran medida, la capacidad del modelo para obtener mejores resultados. Como ya se había anticipado previamente, el proceso de obtención de datos tiene tres fuentes diferentes: Binance, Twitter y Google Trends.

El tratamiento de los datos y la obtención de nueva información implícita en el conjunto de datos es también un punto fundamental para complementar el conjunto de datos inicial. En la figura 1 se ilustra gráficamente el proceso de obtención y tratamiento de los datos aplicado durante el estudio.

2.1. Obtención de datos

Para la obtención de datos se han tenido en cuenta diferentes tipos de datos dependiendo de la naturaleza de los mismos. En primer lugar, se tiene en cuenta la información de mercado, es decir, índices de mercado tales como:

- Cotización de apertura, cierre, máxima y mínima.
- Volumen total, del activo base y del activo de cotización.
- Número de operaciones.

Estos indicadores son básicos para cualquier criptomoneda, al igual que para un activo financiero, pues detallan la información directamente relacionada con su operativa y ofrecen el estado de la misma en cualquier momento. No obstante, se puede extraer nueva información a partir de los indicadores expuestos anteriormente, cómo por ejemplo índices de volatilidad, o la diferencia porcentual de precios por lapso de tiempo, entre otros, como veremos más adelante. Además, un punto interesante y muy recurrente en este tipo de entornos es la utilización de lo que se denominan indicadores en análisis técnico. Algunos de estos indicadores emiten señales que permiten anticipar en muchos casos la dirección que va a tomar la cotización en un momento dado, por lo que se han tenido en cuenta a la hora de confeccionar el conjunto de datos.

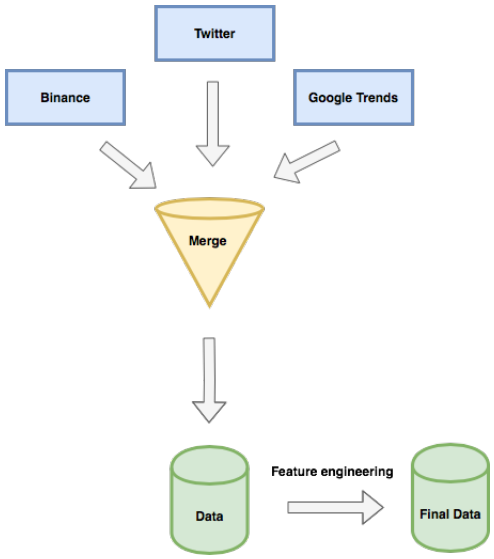

Figura 1: Descripción gráfica del proceso de obtención y tratamiento de datos.

En segundo lugar, Twitter ha servido como fuente de datos. A partir de la API, se ha realizado un *streaming* en base a una serie de términos relacionados con Bitcoin y las criptodivisas. De esta manera, se han obtenido los tuits en tiempo real para poder saber todo aquello que se está comentando y poder analizar la información que proporcionan las personas. De acuerdo con [3], una manera de analizar los tuits es mediante el análisis de sentimientos, ya que nos permite saber si la percepción de ese tuit es algo positivo o negativo. Por ello, se ha querido comprobar la influencia de la percepción social sobre la dirección que toma la cotización en cada instante de tiempo. Por lo que respecta al indice de polaridad, hay que decir que este toma valores comprendidos en el rango [-1, 1] donde los valores negativos representan comentarios con mala percepción mientras que los valores positivos representan comentarios con una buena percepción, para cada instante de tiempo [6].

Por último, se ha utilizado Google Trends teniendo en cuenta, al igual que en el caso de Twitter, una serie de términos relacionados sobre los cuales podemos capturar el índice de popularidad, que toma valores de [0,100] donde valores próximos a 0 indican niveles de popularidad nula mientras que

valores cercanos a 100 indican niveles de popularidad máxima [5]. El estudio que determinaba una cierta correlación entre el precio del Bitcoin y el índice de popularidad que ofrece Google a través del concepto Bitcoin fue realizado en 2013 por Kristoufek [7].

Notése que dicha información ha sido considerada en un lapso de tiempo de 1 minuto, lo cual permite tener una cantidad de registros considerable y una baja granularidad en los datos. Por lo que respecta al período de captación de datos, existen dos vertientes claramente diferenciadas. Una de ellas está relacionada con los indicadores de mercado y Google Trends los cuales se capturan una única vez cada hora, mientras que por otra parte la captación de Twitter debe realizarse de manera continua en tiempo real.

2.2. Tratamiento de datos

A partir de la información almacenada con diferentes orígenes, se ha realizado una agrupación que permite tener un único conjunto de datos sobre el que operar. En el caso de la información relativa a las fuentes de Binance y Google Trends no es necesario realizar ningún tipo de tratamiento inicial, pues debido a que es información numérica puede guardarse directamente, sin embargo, no ocurre lo mismo con la información que proviene de Twitter.

Debido a que los tuits constituyen una fuente de datos textual, es necesario aplicar tratamientos de minería de textos. Entonces sobre cada uno de ellos se han aplicado siguientes tratamientos antes de realizar el análisis de sentimientos:

- Eliminación de URL's.
- Eliminación de hashtags.
- Eliminación de menciones.
- Eliminación de tags de retuit.
- Eliminación de saltos de línea.

Gracias a estos tratamientos, obtenemos una información textual simplificada que facilita el proceso de análisis eliminando así ruidos potenciales que terminen por distorsionar la muestra. Para cada tuit procesado se obtiene un índice de polaridad que posteriormente es agrupado utilizando la media aritmética para obtener un valor sobre el lapso de tiempo escogido, en este caso de 1 minuto.

Finalmente, se realiza una agrupación de los datos obtenidos de las diferentes fuentes, para constituir una fuente de datos única sobre la que poder trabajar.

2.3. *Feature engineering*

A partir del conjunto de datos completo, se puede extraer información implícita que puede ayudar a completar el conjunto de datos inicial. La parte de los datos que permite la aplicación de técnicas para la obtención de nuevos indicadores es la referente a la de índices de mercado. Mediante la utilización de estos índices podemos obtener la diferencia porcentual de precios por lapso de tiempo, que marca de una manera sencilla si la cotización ha aumentado o disminuido, en primera instancia, y que además nos permite saber la diferencia que existe respecto al momento de tiempo inmediatamente anterior.

$$Dif(\,\%) = \frac{Close(t) - Open(t)}{Open(t)}$$

Por otra parte, podemos obtener la volatilidad que se entiende como la diferencia entre el valor más alto y el más bajo del precio de apertura del periodo.

$$Volatility = \frac{Max(t) - Min(t)}{Open(t)}$$

El punto más interesante, lo encontramos en la utilización de los indicadores técnicos muy recurrentes en el conocido análisis técnico que ofrecen una visión complementaria a partir de los índices más básicos. De esta manera, podemos obtener información con un cierto nivel de sofisticación que ayudan a la predicción, en muchos casos, de la dirección del precio que tomará el activo.

Algunos de los índices estadísticos obtenidos y de especial interés han sido:

- *Moving average*: permite obtener nuevos valores a partir del promedio de un subconjunto de los datos originales. Nótese que M representa el valor específico de la serie escogida de tamaño n.

$$SMA = \frac{p_M + p_{M-1} + ... + p_{M-(n-1)}}{n}$$

- *Relative strength index (RSI)*: permite obtener un índice de la fuerza que ejerce el precio a partir de la comparativa entre los movimientos al alza o a la baja de los precios de cierre.

$$RSI = 100 - \frac{100}{1 + RS}$$

  donde RS es:

$$RS = \frac{SMMA(U, n)}{SMMA(D, n)}$$

  y SMMA es *smoothed or modified moving average*.
- *Bollinger brands*: se obtienen a partir de la media móvil (tanto simple como exponencial) sobre el precio de cierre la cual es envuelta por dos bandas que se obtienen de añadir y sustraer, para la superior y inferior respectivamente, al valor de la media K desviaciones estándar, donde K generalmente es igual a 2.

$$BBUpper = MA + K\sigma$$

$$BBLower = MA - K\sigma$$

- *MACD*: conocida como la Media móvil de Convergencia / Divergencia funciona a partir de tres componentes: MACD que corresponde con la diferencia entre dos medias móviles, una calculada a corto plazo (generalmente

12 periodos) y otra a largo plazo (26 periodos). El componente señal corresponde al promedio móvil exponencial del MACD usando 9 periodos. El último componente es el histograma, que se corresponde a la diferencia entre la MACD y la señal.

$$MACD = EMA(12) - EMA(26)$$

$$Signal = EMA(MACD, 9)$$

$$Histogram = MACD - Signal$$

Como vemos, algunos de estos indicadores tienen en cuenta aspectos como la volatilidad y el efecto de los cambios tan drásticos que se suceden en la cotización de las criptomonedas como el Bitcoin.

Debemos destacar que, sobre la variable objetivo se ha aplicado una conversión a tipo binaria, de manera que para valores negativos la variable toma como valor un 1, para valores positivos toma como valor un 0. De esta forma, transformamos el problema a uno de tipo clasificación, donde predecimos si la cotización sube o baja. Por ello, utilizaremos modelos de clasificación con este criterio y otros con el valor diferencial porcentual del precio. De esta manera se permite trabajar con problemas de diferente naturaleza y ver cual es el enfoque que presenta el mayor desempeño y es más adecuado para los datos recogidos.

3. Análisis de los datos

Un punto fundamental es el conocimiento sobre el conjunto de datos pues de ello dependerá, en parte, el desempeño que podamos conseguir a partir del modelo de datos. En primer lugar, si analizamos la variable objetivo podemos observar la distribución que toman las variaciones de la cotización del Bitcoin:

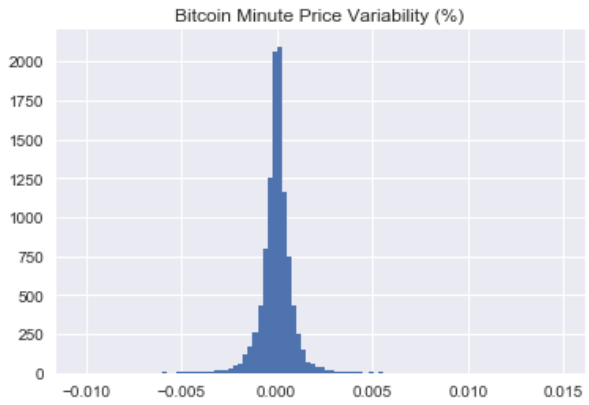

Figura 2: Distribución de la variabilidad de la cotización del Bitcoin.

Como podemos obervar en la figura 2, la variación de la cotización del Bitcoin minuto a minuto parece seguir una distribución normal, formando una campana prácticamente simétrica.

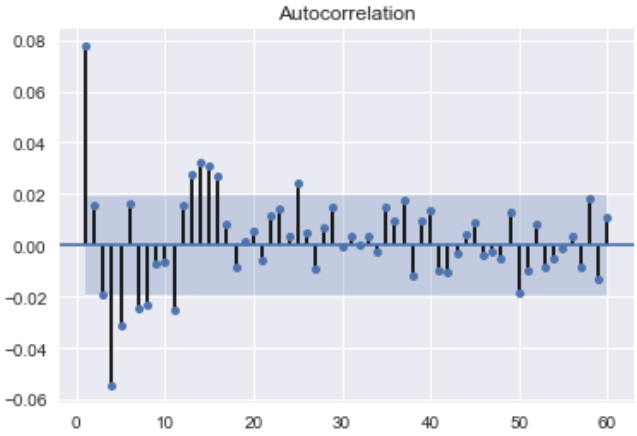

Figura 3: Test de autocorrelación de la serie sobre los primeros 60 minutos.

Otra parte fundamental en el análisis de una serie temporal es comprobar la autocorrelación existente de la misma para detectar si existe algún tipo relación entre el actual diferencial de precio respecto al anterior o futuros diferenciales. Como se aprecia en la figura 3, el diferencial de precio está correlacionado en un lapso de tiempo de hasta 16 minutos, donde vemos que hay un valor a tener en cuenta en el caso de 25 minutos como caso aislado. Además, se ha analizado para periodos más amplios de tiempo donde se ha visto que la autocorrelación estaba marcada por tramos que parecen seguir la tendencia de estacionalidad diaria de la figura 4.

Por otro lado, en la figura 4, vemos la estacionalidad de la serie de forma diaria, donde como dato a destacar observamos que la mayor caída del precio se produce generalmente sobre las 17:15h y el mayor incremento se produce de cara a las 21:10, aproximadamente. Este comportamiento suele ser habitual en el mercado de criptomonedas, pues debido a las grandes fluctuaciones que se producen durante el día, cada uno de los cambios extremos viene acompañado por lo que en términos de mercado se entiende como una corrección, de ahí a que observemos esa curva entre las 16:00h y las 00:00h. No obstante, vemos que en la mayor parte del día, el valor de la cotización va fluctuando de manera reducida y constante respecto al periodo anteriormente señalado.

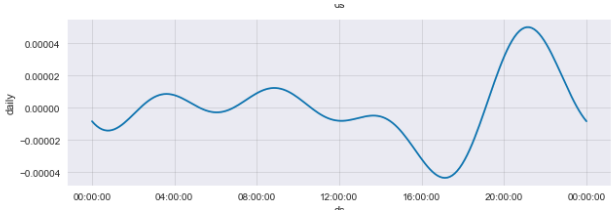

Figura 4: Estacionalidad de la serie diaria.

Este tipo de movimientos en los mercados, especialmente

en el mundo de las criptomonedas, tienen lugar debido a la utilización de sistemas inteligentes por parte de los grandes inversores que, a la postre, son los que dominan la tendencia del mercado.

Gracias a la inteligencia artifical, el análisis masivo de datos y los diferentes mecanismos existentes a día de hoy es posible llevar a cabo una interacción de forma automática. Mediante APIs, es posible gestionar una cartera de criptomonedas y tener un absoluto control sobre los activos propios. Es por ello que a través de una recopilación de los diferentes *inputs*, su posterior tratamiento y el conjunto de reglas y acciones a tomar, se genera un patrón de trabajo que afecta finalmente al comportamiento del mercado. Dicho funcionamiento pretende simular el comportamiento humano realizado desde un punto de vista sistemático por lo que, finalmente, para conocer como funciona el mercado sería conveniente utilizar técnicas de ingeniería inversa.

Al producirse una serie de movimientos, de forma sistemática, el mercado se comporta de manera similar diariamente. Este mismo comportamiento se detecta al analizar la variación anual que sufre el precio del Bitcoin donde también se aprecian patrones similares a fin de año y principio del año siguiente donde se producen subidas muy fuertes acompañadas de fuertes correciones de hasta el $100\,\%$ respecto al precio máximo del periodo.

Otro aspecto fundamental en este tipo de entornos es la entrada en el mercado de las diferentes zonas geográficas, donde durante el transcurso del día aquí en Europa se produce la entrada a *trading* de otras regiones como Estados Unidos o China las cuáles tienen una cuota de mercado considerable, sobretodo en el continente asiático, lo que genera mayores oscilaciones en la cotización de las monedas.

Además de lo anteriormente expuesto, también hay que tener en cuenta que no siempre se pueden utilizar comportamientos pasados para predecir el futuro pues las condiciones en cada momento son diferentes debido a diferentes factores como podría ser la aparición de la regulación y legislación por parte de los diferentes países.

4. Modelado de datos

Una vez preparado el conjunto de datos, como se ha detallado en el punto anterior, nos encontramos en disposición de utilizar técnicas estadísticas que nos permitan alcanzar el objetivo propuesto. Los principales modelos que se han tenido en cuenta son los siguientes:

- *LSTM (Long Short-Term Memory)* [2]
- *Gradient Boosting Regressor* [1]
- *Gradient Boosting Classifier* [1]
- *Random Forest* [1]
- Regresión Logística [1]

Nótese que se incluyen tanto redes neuronales, como modelos regresivos y de clasificación en función de como esté expresada la variable objetivo, que recordemos que podía representar un diferencial porcentual respecto al precio predecesor o simplemente si en ese instante de tiempo había subido o bajado el precio (0 ó 1) en función del signo de ese diferencial.

A continuación se detallan todas las especificaciones utilizadas para cada uno de los modelos de datos junto las pertinentes consideraciones al respecto.

4.1. *LSTM (Long Short-Term Memory)*

La estructuración de la red se ha realizado mediante la utilización de dos capas ocultas *LSTM* con un número de 256 neuronas por capa. Se han tenido en cuenta otros valores en el número de neuronas partiendo de las 64 neuronas por capa. Sin embargo, los mejores resultados se han obtenido con 256 neuronas por capa asumiendo una repercusión directa y lógica en el coste computacional. Además, se han probado diferentes configuraciones en cuanto al número de capas intermedias. Finalmente se ha determinado que no había una mejora evidente.

Por otro lado, tenemos la ventana temporal utilizada sobre el conjunto de datos donde, como ya es sabido, este tipo de red tiene un mejor desempeño cuanto mayor es dicha ventana temporal, ofreciendo resultados más discretos para ventanas de tiempo más cortas. El tamaño temporal óptimo encontrado durante el estudio ha sido de 25 minutos. Por lo que respecta al optimizador de la red, se han probado diferentes según las especificaciones que ofrece *Keras* donde se ha determinado que el que mejor se adapta a la naturaleza del problema es *RSMProp* además de atender a la recomendación que incluye el propio manual.

La función de activación utilizada ha sido *tahn*, después de probar la lineal, sigmoidal y combinación de ambas. En lo que a número de iteraciones se refiere hay que destacar que el modelo convergía alrededor de las 50 y 100 iteraciones.

4.2. *Gradient Boosting Regressor*

Para el modelo de *Gradient Boosting Regressor* se ha determinado que la mejor configuración para el conjunto de datos es aquella compuesta por un total de 128 estimadores, con profundidad máxima de 32, un mínimo 16 particiones por muestra y un ratio de aprendizaje del 0.05. Para valores más elevados de los parámetros mencionados anteriormente, hemos visto como el modelo tiende al sobreajuste o *overfitting*.

4.3. *Gradient Boosting Classifier*

Por lo que se refiere al modelo *Gradient Boosting Classifier* al igual que para el caso de *Gradient Boosting Regressor* se han utilizado diferentes configuraciones de parámetros y se ha determinado que la configuración mediante la cual se ha obtenido un mejor desempeño la que se corresponde con un total de 64 estimadores, con profundidad máxima de 16, un mínimo de 8 particiones por muestra y un ratio de aprendizaje del 0.001. Para valores más elevados de ratio de aprendizaje

se producía *overfitting*, por ello se ha tenido que ajustar principalmente dicho parámetro.

4.4. *Random Forest*

Otro de los modelos que se han tenido en cuenta para la clasificación es el de *Random Forest*. Mediante la utilización de funciones de optimización se ha obtenido que la mejor configuración es: una profundidad máxima de 7, un mínimo de muestras por hoja de 12 y un mínimo de muestras por partición de 2.

4.5. Regresión Logística

Es seguramente el modelo más simple de todos y el que peor se ajusta a la naturaleza de los datos, pero se ha incluído para compararlo con aquellos que están más orientados a la predicción de índices de mercado.

5. Resultados

Una vez expuestos los diferentes modelos utilizados y sus respectivas configuraciones, se presentan a continuación los resultados obtenidos durante el estudio. Los resultados han sido diferenciados por criterio según las métricas asociadas al tipo de modelo:

- Redes neuronales
- Regresión
- Clasificación

5.1. Criterios de redes neuronales

Neuronas	***MAE***	***MSE***
64	0.4058	0.5067
128	0.3851	0.4866
256	0.3336	0.4388

Cuadro 1: Resultados del ajuste sobre el conjunto de datos de entrenamiento *LSTM*.

Del modelo de redes neuronales de tipo *LSTM*, vemos como los resultados no son tan óptimos como se esperaba pues, los errores obtenidos no son despreciables. Sin embargo, a medida que se aumenta el número de neuronas, tanto el *MAE* como el *MSE* disminuyen. No obstante, la posibilidad de que se produzca *overfitting* se incrementa.

5.2. Criterios de regresión

Learning rate	***Score* (%)**	***MAE***	***MSE***
0.01	57.46	0.000001	0.000343
0.05	83.76	0.000001	0.000251
0.1	83.76	0.000001	0.000251

Cuadro 2: Resultados del ajuste sobre el conjunto de datos de entrenamiento *Gradient Boosting Regressor*.

Del modelo *Gradient Boosting Regressor* vemos como ofrece un *score* bastante elevado, pero sus bajos errores, denotan un *overfitting* latente en el ajuste. Si bien es cierto que para la predicción sobre el conjunto de test, los resultados son mejor de lo esperado.

5.3. Criterios de clasificación

Estimadores	**Precisión (%)**	***Recall***	***F1-Score***
64	88	84	84
128	93	92	92
256	96	95	95

Cuadro 3: Resultados del ajuste sobre el conjunto de entrenamiento *Gradient Boosting Classifier*.

En el cuadro 3 vemos las diferentes configuraciones en cuanto al número de estimadores utilizados en las pruebas realizadas para el entrenamiento del modelo. Observamos que a medida que se aumenta el número de estimadores los resultados mejoran. No obstante, como ya se ha comentado anteriormente, el ratio de aprendizaje utilizado es de 0.001 que es la recomendación que hacen desde *Keras* ya que utilizando otros valores hemos observado que se produce un *overfitting* llegando, en algunos casos, al 100 % de precisión. No obstante, vemos que la ganancia entre 128 y 256 no es demasiado significativa, por lo que en términos de cómputos y eficiencia nos interesa utilizar la configuración de 128 estimadores.

Modelo	**Precisión (%)**	***Recall***	***F1-Score***
Random Forest	77	76	76
Regresión Logística	64	64	64

Cuadro 4: Resultados del ajuste sobre el conjunto de entrenamiento (Clasificación).

Como podemos observar del cuadro 4, el modelo *Random Forest* ofrece unos resultados ligeramente superiores a los que se han obtenido con la Regresión Logística que de largo es el peor modelo de todos. Finalmente, vemos que el modelo con mejores prestaciones para la clasificación es *Gradient Boosting Classifier*.

Nota: Se han realizado pruebas con un modelo adicional y no mencionado anteriormente: *Boosting* + Regresión Logística el cuál ha ofrecido unos resultados similares a los obtenidos con el modelo *Random Forest*, por ello se ha decidido no incluirlo.

6. Conclusiones

Del estudio previamente detallado se deslizan una serie de conclusiones a destacar. Hemos visto como la alta volatilidad que padece la cotización de las criptomonedas dificulta la obtención de un modelo estable y que retorne una predicción lo más ajustada posible a la realidad.

Es por ello que la aplicación de técnicas de *feature engineering* son de vital importancia para obtener nuevos índices que aporten información adicional e implícita sobre el conjunto de datos original. En esa dirección, sería de especial interés incorporar nuevos índices y conceptos, como por ejemplo el retroceso de Fibonacci [9] u otros como la Teoría de las Ondas de Elliot [8] para identificar los patrones que se producen y anticiparlos.

En cuanto al modelado de datos, hay que destacar que el reducido conjunto de datos del que se disponía a la hora de realizar el estudio es un factor condicionante. Concretamente, el periodo de recogida de datos fue de diez días de forma ininterrumpida acumulando más de 14.000 muestras en total. Nótese que es un lapso de tiempo insuficiente pues apenas cubre un tercio de un mes y, por tanto, no recogerá toda la variabilidad que sería deseable para el entrenamiento del modelo. Para la obtención de mejores resultados, sería deseable que en una futura puesta en producción del proyecto, hubiese una retroalimentación contínua en tiempo real a medida que los datos son almacenados. De esta manera, el modelo de datos tendría a su disposición nuevo conocimiento que permitiría realizar sucesivos ajustes con el fin de mejorar el rendimiento del mismo. Aun así, algunos de los modelos utilizados como *Gradient Boosting Classifier* y *LSTM* han ofrecido unos resultados notables y quedan como propuesta clara para seguir trabajando en ellos.

No obstante, hay que destacar que sería deseable realizar una variante sobre la propuesta de modelado de datos inicial utilizada en el presente estudio, pues mediante la combinación de los diferentes modelos utilizados, a partir de técnicas conocidas como *ensembling*, sería posible mejorar los resultados obtenidos durante el estudio. Mediante la combinación de modelos es posible abarcar un mayor espacio de hipótesis, proporcionando así una mayor flexibilidad y ayudando a su vez a reducir la posibilidad de obtener un *overfitting* sobre el conjunto de entrenamiento cubriendo de esta manera un número de casos mayor del que se podría obtener utilizando un único modelo.

Por otra parte, una vez profundizado en mayor medida sobre el estudio y habiendo obtenido un modelo de datos a partir del cual poder obtener unos resultados consistentes y fiables, podría surgir como proyecto anexo, uno en el cual se especifique un sistema apoyado por inteligencia artificial que sea capaz de administrar y gestionar una cartera de criptomonedas, además de realizar operaciones de compra - venta de manera autónoma en base a los futuros valores de cotización del Bitcoin, por ejemplo. Sin embargo, cabe destacar que surge una casuística nueva debido a las comisiones que se generan a la hora de realizar operaciones. Es por ello que, sobre el conjunto de reglas y acciones a tener en cuenta se debería añadir la casuística generada por la comisión que supone la operación en cuestión y comprobar la viabilidad de la operación para no incurrir en pérdidas.

Desde el punto de vista comercial, se podría ofrecer como producto un *dashboard*, también conocido como cuadro de mandos, mediante el cual ofrecer información detallada sobre el estado del mercado, entorno a una criptomoneda o criptomonedas, que facilite la toma de decisiones a partir de predicciones e indicadores de compra - venta. Inclusive, yendo más allá, se podría desarrollar una plataforma que funcionase como gestor de carteras de manera que se permitiese el control centralizado desde una herramienta única y sobre la que poder operar mediante llamadas a APIs.

Además, hay que destacar que la realización de un estudio de mercado de una forma más exhaustiva permitiría ampliar la base de conocimiento a la hora de modelar y operar. Una vía para ello sería mediante la aplicación de técnicas de ingeniería inversa a partir de las cuáles poder determinar el comportamiento y conocer el funcionamiento de los sistemas inteligentes a manos de grandes inversores que son, en última instancia, los que terminan modificando el estado en el que se encuentra el mercado.

Gracias a la estructura de APIs que se encuentran disponibles a día de hoy de los diferentes *exchanges*, es posible analizar un espectro de indicadores muy amplio, entre los que se incluyen los umbrales de cotización sobre los que existen operaciones de compra / venta programada por los inversores. De esta forma, es posible comprobar la cantidad de capital que hay de entrada / salida en función de la cotización que alcance la criptomoneda.

Apéndice A. Modelo de negocio

Como se ha comentado anteriormente, el desarrollo de una aplicación que centralice la gestión de carteras y la predicción del los diferentes activos del mercado tiene un claro interés comercial. A continuación se detalla de forma muy directa el análisis realizado a partir del modelo Canvas que permite evaluar los puntos clave de un proyecto en el ámbito empresarial. A partir de la figura 5, destacaremos los puntos más relevantes del análisis:

- Propuesta de valor
- Segmentos de mercado
- Socios clave
- Fuente de ingresos
- Estructura de costes

Figura 5: Modelo Canvas del proyecto.

Apéndice A.1. Propuesta de valor

La propuesta de valor está basada en la previsión de la cotización de la moneda, un sistema de recomendaciones en base a la previsión de la cotización y situación de la cartera. Además de ofrecer un seguimiento histórico de la rentabilidad de la misma. Partiendo de los puntos referenciados anteriormente sería interesante dotar de inteligencia artificial al sistema para ofrecer un servicio premium mediante el que se automatice el proceso de gestión.

Apéndice A.2. Segmentos de mercado

Los segmentos de mercado están divididos en dos grupos claramente diferenciados: personas con nivel adquisitivo bajo y medio, y los grandes inversores. Por ello, la oferta de servicios deberá ser diferenciadora para ajustarse a las necesidades de cada uno de ellos.

Apéndice A.3. Socios clave

Los socios clave principalmente serán las diferentes casas de intercambio que trabajan con criptodivisas: Binance, Bittrex, Coinbase, Okex.

Apéndice A.4. Fuente de ingresos

En el punto de segmentos de mercado se había anticipado que las principales fuentes de ingresos vienen definidas por el sistema de suscripciones (*Pay-Per-Service*) y la publicidad incluida en la plataforma.

Apéndice A.5. Estructura de costes

La estructura de costes está marcada por el mantenimiento habitual en cualquier proyecto de ámbito tecnológico, el asesoramiento de imagen para crear el concepto de marca y la escalabilidad del número de usuarios para hacer frente a la potencial demanda creciente.

Referencias

[1] Api reference — scikit-learn 0.19.2 documentation. *http://scikit-learn.org/stable/modules/classes.html*, 2018.

[2] Keras documentation. *https://keras.io*, 2018.

[3] Shuib Basri Aliza Sarlan, Chayanit Nadam. *Twitter Sentiment Analysis*. Universiti Teknologi PETRONAS, Perak, Malaysia, 2014.

[4] Jack Strauss Guofu Zhou Yingzi Zhu Andrew Detzel, Hong Liu. Bitcoin: Learning, predictability, and profitability via technical analysis. *Washington University in St. Louis*, 2018.

[5] Eric Ghysels Christian Conrad, Anessa Custovic. *Long- and Short-Term Cryptocurrency Volatility Components: A GARCH-MIDAS Analysis*, volume 11. University of North Carolina, Chapel Hill, NC 27599, USA, 2018.

[6] Prabhakaran R Saravanan S Vinoth M Dr.Balasaravanan.K, Bharathi Bhaskaran R. Twitter sentiment analysis. *SIGCSE Bull.*, 119(10), 2018.

[7] Ladislav Kristoufek. *BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era*. Charles University in Pra-

gue, Opletalova 26, 110 00, Prague, Czech Republic, EU, 2013.

[8] W.Z.A.W. Muhamad M.H. Zakaria A.M. Desa M.F. Ramli, A.K. Junoh and Mahyun A.W. Elliott wave pattern recognition for forecasting gbp/usd foreign exchange market. *Universiti Malaysia Perlis.*, 2018.

[9] Nabeeha Zulfiqar Rana Zafarullah Shaker, Muzaffar Asad. *Do Predictive Power of Fibonacci Retracements Help the Investor to Predict Future? A Study of Pakistan Stock Exchange*, volume 4. University of Central Punjab, 2018.