
Aplicación de técnicas de machine learning para la predicción de los precios de la acción APPLE

Application of machine learning techniques for apple stock price prediction

Roberto Evaristo Broncano

<https://orcid.org/0000-0001-9010-3641>

roberto.evaristo@unmsm.edu.pe

Universidad Nacional Mayor de San Marcos. Lima, Perú

RECIBIDO: 03/07/2022 - ACEPTADO: 08/08/2022 - PUBLICADO: 20/09/2022

RESUMEN

La presente publicación tiene el objetivo de comparar el nivel de predicción de determinadas técnicas como series de tiempo y *machine learning* en el contexto del pronóstico del precio de la acción APPLE que tiene el mayor volumen de capitalización en el sector de la tecnología. Se ha descargado el histórico de precios cuyas fechas están entre el 01 de enero del 2019 y el 30 de junio del 2021 y se buscó medir el nivel de predicción teniendo como target los precios del día siguiente, al día 15 y al día 30, asimismo, se probó la información textual para poder medir su aporte en la mejora de la predicción. Se utilizaron los indicadores de ajuste Raíz del Error Cuadrático Medio (*RMSE*) y del Error Porcentual Absoluto Medio (*MAPE*), considerando el criterio de que en la medida que una técnica presente menor valor será la mejor técnica para cada uno de los escenarios propuestos. Respecto a los resultados que se obtuvo en la aplicación de las diversas técnicas para cada escenario, se encontró que la técnica de Machine Learning XGBOOST con parámetros tuneados mediante Cross Validation y el modelo de Regresión Lineal múltiple son las más útiles para la predicción del precio de la acción APPLE y la información textual también mejoró el nivel de predicción del precio de la acción APPLE.

Palabras clave: Series de tiempo; machine learning; predicción; precio de acción; bolsa de valores.

ABSTRACT

The objective of this publication is to compare the level of prediction of certain techniques such as time series and machine learning in APPLE stock price forecasting, which It has the largest capitalization volume in the technology sector. The price history is between January 1, 2019 and June 30, 2021, that it has been downloaded and it was used to measure the level of prediction having as a target the future price of the following day, on the 15th and on the 30th, also, other type of textual information were tested in order to measure their contribution to improving the prediction. The Root Mean Square Error (*RMSE*) and the Mean Absolute Percentage Error (*MAPE*) adjustment indicators were used, considering the criterion that to the extent that a technique has a lower value, it will be the best technique for each of the proposed scenarios. Regarding the results obtained in the application of the various techniques for each scenario, it was found that the Machine Learning XGBOOST technique with parameters tuned by Cross Validation and the Multiple Linear Regression model are the most useful for predicting the price of APPLE stock and textual information also improved the prediction level of APPLE stock price.

Keywords: Time series; machine learning; prediction; stock price; stock exchange.

I. INTRODUCCIÓN

Los movimientos de precios a corto plazo contribuyen en gran medida a la imprevisibilidad de las bolsas de valores, por lo que predecir con precisión las fluctuaciones de los precios en el mercado de valores es una gran ventaja económica para los inversores (Khare et al., 2017). La variabilidad en los precios de acciones se deben a diversos factores como son: (1) factores fundamentales que hace referencia al análisis financiero de las empresas que cotizan en la Bolsa de Valores, (2) factores técnicos que hace referencia al análisis de las tendencias de los precios de las acciones, (3) noticias de naturaleza financiera, (4) sentimiento del mercado que hace referencia a la perspectiva de los inversionistas, entre otros factores adicionales (Zinn, 2021).

La volatilidad representa un problema para los inversores porque no saben realmente si los precios subirán o caerán en un periodo futuro de cotizaciones lo cual impacta directamente en su rentabilidad, además, la seguridad en las decisiones de compra/venta de acciones representa una necesidad clara para los inversionistas, ya que ellos son los principales interesados en saber cómo se comportarán los precios de las acciones en las próximas horas, días, semanas, meses y años inclusive.

Dentro de la analítica financiera existen muchas formas de satisfacer la necesidad de pronosticar los precios futuros de las acciones, el presente trabajo lo aborda desde la aplicación de técnicas de series de tiempo y machine learning donde se estudian las tendencias de una forma más analítica con el empleo de los modelos estadísticos y las técnicas computacionales.

El presente trabajo trata de predecir los precios de acciones como APPLE que cotizan en una bolsa de valores aplicando diversas técnicas como los modelos de series de tiempo y también las técnicas de machine learning como la regresión lineal múltiple y la técnica de regresión Extreme Gradient Boosting (XGBOOST). Las técnicas son utilizadas para predecir el precio en diversas ventanas temporales (predicción para el siguiente día, quince días y treinta días) y también considerando diversas fuentes de información como la cotización de precios *per se* y la información textual recolectada de *Twitter*.

Como parte de la organización del trabajo de investigación, en la sección 2 se describe los aspectos más relevantes de investigaciones relacionadas, fundamentos teóricos de las técnicas utilizadas, en la sección 3 se definen los objetivos de la investigación, en la sección 4 se detalla el método utilizado

para la aplicación de las técnicas de aplicación, en la sección 5 se muestran y comparan los resultados de la predicción de los precios de acciones, finalmente, en las secciones 6, 7 y 8 se detalla la discusión de los resultados, conclusiones y recomendaciones respectivamente.

II. FUNDAMENTOS TEÓRICOS

2.1. Antecedentes

En la literatura de la analítica financiera existe un aporte considerable de la aplicación de los modelos de predicción aplicados a diferentes ámbitos y problemas que puedan existir en el contexto de las finanzas. Para el presente informe se detallan algunos que se consideran relevante en línea con el trabajo de investigación.

Chen et al. (2020) compararon diferentes técnicas de modelamiento como regresión logística, análisis discriminante, random forest, SVM, XGBOOST, memoria de corto plazo a largo plazo para la predicción diaria de precios de Bitcoin. En su trabajo de investigación encontraron que en ventanas de largo plazo encontraron que las técnicas tradicionales lograron un mejor nivel de precisión a comparación de técnicas de aprendizaje automático (66% contra 65.3%), mientras que en la predicción de intervalos de cinco minutos la mejor predicción la obtuvieron las técnicas de aprendizaje automático (67.2%).

Naik y Mohan (2019), estudiaron la Red Neuronal Recurrente (RNN) con Memoria a Largo Plazo (LSTM) para pronosticar futuros rendimientos de las acciones de la Bolsa Nacional de Valores de India (NSE) donde concluyeron que el modelo RNN con LSTM propuesto superó el rendimiento a comparación de una Red Neuronal Artificial.

Mohan et al. (2019), estudiaron la correlación entre los artículos de noticias y el movimiento de los precios de las acciones con la hipótesis de una mejora en el nivel de predicción. Para ello se integraron con algoritmos como Máquinas Vectorial de Soporte (SVM), Regresión de Redes Bayesianas, Deep Learning y modelos de series temporales como ARIMA, RNN y Facebook Prophet. Para entrenar estos modelos de forecasting se usaron cotizaciones diarias de acciones que pertenecen al índice S&P500 durante cinco años, junto con más de 265,000 artículos de noticias financieras. Finalmente, encontraron mejores resultados con RNN y evidenciaron la existencia de una correlación entre la información textual y la dirección del precio de las acciones.

2.2. Bases Teóricas

Las series de tiempo son un tipo de datos que miden cómo cambia una variable a lo largo del tiempo. En un conjunto de datos de series de tiempo, la columna de tiempo no representa una variable *per se*, en realidad es una estructura primaria que puede usar para ordenar su conjunto de datos (Lazzeri, 2021). Se muestra la Figura N° 1 para tener mayor alcance de una serie de tiempo sobre los precios de una acción determinada.

Media Móvil. La técnica de promedio móvil aprovecha los errores de pronóstico anteriores en un enfoque de regresión para pronosticar las observaciones futuras en las próximas marcas de tiempo: cada observación futura puede considerarse como un promedio móvil ponderado de los errores de pronóstico anteriores (Lazzeri, 2021).

Modelo Autoregresivo. La autorregresión es un enfoque de pronóstico de series de tiempo que depende sólo de los resultados anteriores de una serie de tiempo, asume que las observaciones futuras en la siguiente marca de tiempo están relacionadas con las observaciones en marcas de tiempo anteriores a través de una relación lineal (Lazzeri, 2021).

Modelo ARMA. Un método ARMA consta de dos partes: (a) Una autorregresión, (b) Un modelo de media móvil. Los modelos ARMA proporcionan el modelo lineal más eficiente de series de tiempo estacionarias, ya que son capaces de modelar el proceso desconocido con un número mínimo de parámetros (Zhang et al. 2015).

Modelo ARIMA. Los modelos de media móvil autorregresiva integrada (ARIMA) se consideran un

desarrollo de los modelos de media móvil autorregresiva (ARMA) más simples e incluyen la noción de integración.

La Media Móvil Autorregresiva (ARMA) y la Media Móvil Integrada Autorregresiva (ARIMA) presentan muchas características similares: sus elementos son idénticos, en el sentido de que ambos aprovechan una autorregresión general AR (p) y un modelo de media móvil general MA (q). Las principales diferencias entre los métodos ARMA y ARIMA son las nociones de integración y diferenciación.

Por otro lado, en línea con el contexto de la investigación, se abordan los fundamentos teóricos de los modelos de machine learning.

Machine Learning se considera como una rama de la informática que tiene por objetivo estudiar teorías, algoritmos y aplicaciones de sistemas que aprenden como los humanos (Sugiyama, 2015). Generalmente se clasifica en (a) aprendizaje supervisado, donde el algoritmo es entrenado con datos de entrada preetiquetados para lograr converger al mejor clasificador posible $f: X \rightarrow Y$ y de esta manera predecir etiquetas para otros datos; también está el (b) aprendizaje no supervisado donde no se tiene una etiqueta predefinida y se construyen los modelos a través del análisis de similitudes en los datos de entrada de tal forma que se obtengan grupos relevantes diferenciados (Fernandes y Antonelli, 2018). Además, respecto a los modelos supervisados, se tiene a la regresión cuando se predicen variables cuantitativas y clasificación cuando se predicen variables cualitativas (Hastie et al., 2009). Las técnicas que se aplicarán en el presente trabajo de investigación corresponden a las técnicas de regresión.

Figura N° 1
Serie de Tiempo de un Precio de Acción.



Regresión lineal múltiple. Para Hastie et al (2009), en la matriz de variables se tiene la matriz de k variables predictoras identificadas como $X = (X_1, X_2, \dots, X_k)$ con la que se quiere predecir una variable respuesta Y (\hat{Y}) considerando la estimación de los coeficientes de regresión $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ entonces el modelo de regresión lineal múltiple tiene la forma:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^k X_j \hat{\beta}_j$$

La estimación de los coeficientes de regresión β se realiza normalmente mediante el método de mínimos cuadrados, que implica elegir los coeficientes β que logren minimizar la suma de cuadrados residuales $RSS(\beta)$

$$RSS(\beta) = \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^k x_{ij} \beta_j \right)^2$$

Xtreme Gradient Boosting (XGBoost). Fue propuesto por Chen y Guestrin (2016). Para (Zhang y Zhan, 2017), el algoritmo XGBoost es una biblioteca optimizada de refuerzo de gradiente distribuida diseñada para ser altamente eficiente, flexible y portátil. Proporciona un impulso de árbol paralelo que resuelve muchos problemas de ciencia de datos de una manera rápida y precisa. La idea es combinar cientos de árboles simples con baja precisión para construir un modelo más preciso. Cada iteración generará un nuevo árbol para el modelo. El siguiente algoritmo fue tomado de Zhang y Zhan (2017):

Input: I , conjunto de instancias del nodo actual

Input: d , dimensión de variables

$gain \leftarrow 0$

$G \leftarrow \sum_{i \in I} g_i$

$H \leftarrow \sum_{i \in I} h_i$

Para $k = 1$ **hasta** m **hacer**

$G_L \leftarrow 0, H_L \leftarrow 0$

Para j **en** Ordenado (I , por X_{jk}) **hacer**

$G_L \leftarrow G_L + g_j, H_L \leftarrow H_L + h_j$

$G_R \leftarrow G - G_L, H_R \leftarrow H - H_L$

$score \leftarrow \max(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda})$

Fin

Fin

Por otro lado, respecto a la utilización de los indicadores de validación se han utilizado dos mediciones que normalmente son aplicados para comparar la eficacia de la predicción que son RMSE y MAPE.

RMSE. Es la raíz del promedio de los cuadrados del error de cada artículo en el periodo i y también se utiliza para comparar la precisión de diferentes métodos de pronóstico. El resultado está en las unidades originales de la información histórica. (Gálvez, 2016)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

MAPE. Es la media de los errores porcentuales en valor absoluto, no considera el signo del error sólo la magnitud. El MAPE es una de las medidas más utilizadas a nivel mundial, pero no se recomienda para la selección de un método de pronóstico porque presenta sesgos que favorece a los pronósticos que están por debajo de los valores reales. (Gálvez, 2016)

$$MAPE = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{Y_i}}$$

III. OBJETIVOS

3.1. Objetivo principal

Aplicar y comparar las técnicas de series de tiempo y machine learning para la predicción del precio de acciones de la Bolsa de Valores de New York.

3.2. Objetivos específicos

Comparar las técnicas de machine learning entre la regresión lineal múltiple y xgboost desde la predicción del precio de la acción Apple.

Evaluar diversas fuentes de información desde la perspectiva de la mejora del nivel de predicción de los precios de acciones.

IV. MATERIALES Y MÉTODOS

Como parte de los materiales, se ha trabajado con las siguientes herramientas:

Google Colaborative, donde se tiene disponible de manera libre Jupiter notebook que permite codificar en python a través de un servicio en la nube. Esta herramienta se utilizó para la descarga de información, tratamiento de la información, aplicación de las técnicas de modelamiento y deploy de modelos.

YahooFinance, se utilizó como una fuente de información del histórico de los precios, pues en la página se publican diariamente los precios obtenidos de acciones y otros activos financieros de manera

libre. Se descargó los precios históricos de la acción APPLE.

Openblender, fue útil para la obtención de información de tweets de los términos APPLE y Bloomberg.

Como parte del método se tomó como referencia a Lazzeri (2021), quien propuso un marco ágil e iterativo para ofrecer una solución de pronóstico de series de tiempo de manera eficiente. En la Figura N° 2 se muestran las diversas etapas que forman parte de un proyecto de predicción de series temporales, las etapas son: (a) entendimiento del negocio y definición de las métricas de rendimiento, (b) obtención de los datos, (c) análisis exploratorio de los datos, (d) pre-procesamiento de los datos e ingeniería de variables, (e) construcción y selección del modelo, (f) implementación del modelo y (g) aceptación de la solución de pronóstico.

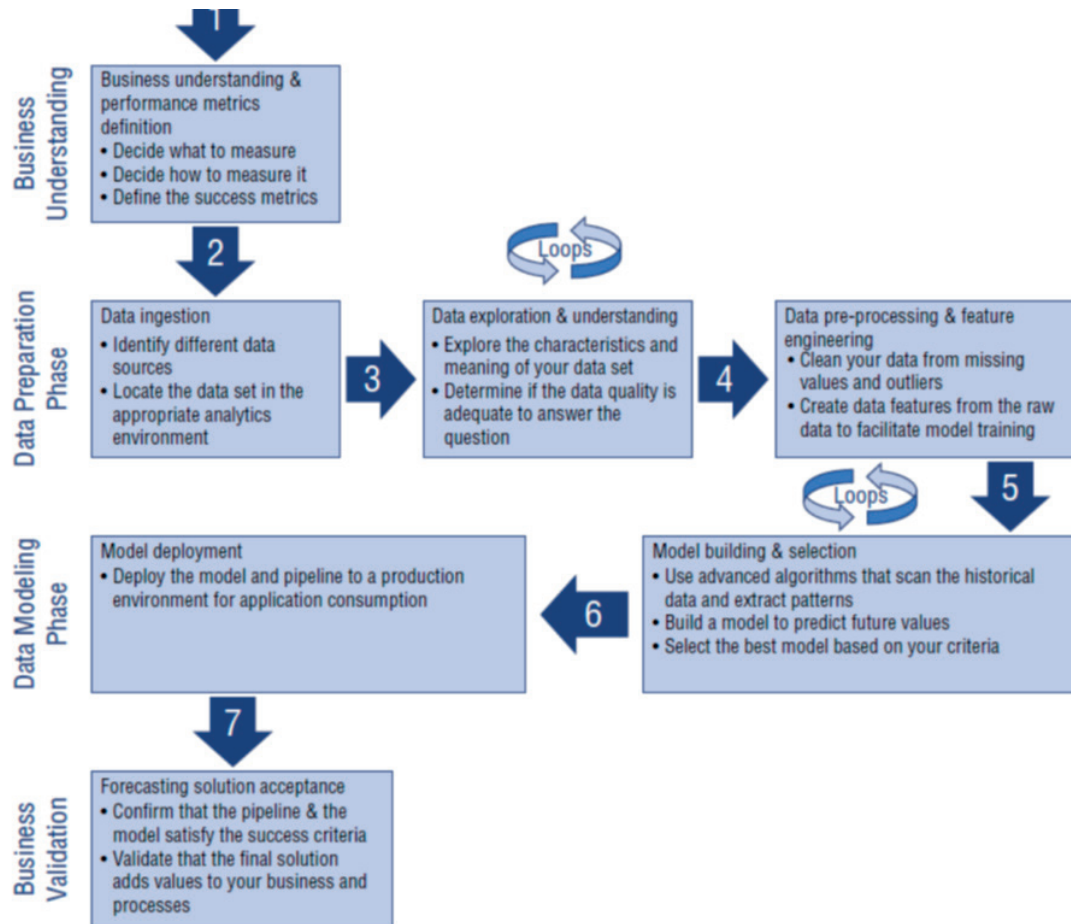
Entendimiento del negocio y definición de las métricas de rendimiento. El presente trabajo se

enfoca en determinar los precios futuros de las acciones con el fin de satisfacer una necesidad de los inversores que es predecir con mayor certidumbre los precios de tal forma que les permita tomar mejores decisiones logrando una mayor rentabilidad por su inversión. Para ello se están utilizando dos métricos de rendimiento, que exclusivamente se enfocan en el rendimiento de la capacidad predictiva de los modelos utilizados (RMSE y MAPE); sin embargo, es posible utilizar otros tipos de métricas desde la perspectiva de rentabilidad generado, valoración cualitativa de los inversores sobre el uso de las técnicas de predicción, entre otros; los cuales son útiles desde una perspectiva de modelo de negocio, pero, dado el enfoque de la publicación, la publicación se enfocó en resultados técnicos principalmente.

Obtención de los datos. Para esta parte del proyecto, se definieron tres fuentes de información que serán utilizadas para evaluar su utilidad en la

Figura N° 2

Etapas para la Predicción de Series Temporales



Nota. Tomado de "Machine learning for time series forecasting with Python", (Lazzeri, 2021). Pag.7

mejora del nivel de predicción: (a) el primero son las cotizaciones de los precios históricos y (b) la información textual generada y almacenada con los tags de APPLE y Bloomberg.

Cotizaciones en los precios históricos. Desde Yahoo Finance se puede descargar información de la fecha del día de la cotización, precio de apertura con que se inicia el día, precio más alto alcanzado durante el día, precio más bajo alcanzado durante el día, precio al final del día de cotización y el volumen negociado durante el día (ver Tabla 1).

De esta fuente es posible calcular las variables que serán útiles para la aplicación de las técnicas, primero, es posible construir los diferentes targets

que serán utilizados como variable a predecir y también se pueden calcular las variables predictoras que permitan entender el comportamiento histórico de los precios calculando diversos indicadores.

Información textual. Desde OpenBlender es posible obtener información almacenada de Twitter de Apple donde se tiene el texto publicado y que a través del análisis de sentimientos fueron procesados para ser utilizados como variables predictoras para la predicción del precio de la acción Apple (ver Figura N° 3).

Análisis exploratorio de los datos. Implica resumir las características principales de un conjunto

Tabla 1

Estructura de información descargada de cotizaciones de precios.

Campo	Descripción
Date	La fecha de la cotización
Open	El precio de apertura de la acción en el día
High	El precio más alto de la acción en el día
Low	El precio más bajo de la acción en el día
Close	El precio de cierre de la acción en el día
Volume	Cantidad de acciones que cambiaron de manos

Figura N° 3

Muestra de información textual de Twitter.

re_tweeter	links	author	text	hashtags	reply_count	associated_tweet	timestamp
0	[https://t.co/Nh00Auckzq]	Bloomberg Markets	JUST IN: Australia is prepared to pay the econ...				1627018000
1	[https://t.co/V6Y0uW43a5]	Bloomberg Markets	As she leaves office, Angela Merkel, hailed fo...				1627017790

re_tweeter_x	links_x	author_x	text_x	hashtags_x	reply_count_x	associated_tweet_x	timestamp_x
0	[https://t.co/vvFjA]oAk]	Bloomberg	Related Cos. was sued by a group of prospectiv...				1627016588
1	[https://t.co/mhMDiIOWt]	Bloomberg	Airbnb provided money and support after a New				1627016407

de datos, incluido su tamaño, precisión, patrones iniciales en los datos y otros atributos.

Pre-procesamiento de los datos e ingeniería de variables. En esta etapa los científicos de datos limpian el conjunto de datos de valores atípicos y datos faltantes y crean características adicionales con los datos sin procesar para alimentar sus modelos de aprendizaje automático. En la Tabla 2 y Tabla 3 se muestra el listado de las variables calculadas del tipo dependiente e independiente.

Adicionalmente, por el lado de los tweets, se recurrió al siguiente procedimiento:

- Se removieron las puntuaciones del Tweet.
- Se removieron los *stopwords* como artículos, pronombres, etc.

- Se obtuvieron las palabras raíces.
- Se obtuvo un score de sentimiento mediante el uso de TextBlob y se generaron los indicadores de polaridad y subjetividad (ver Figura N° 4).

Construcción y selección del modelo. Las técnicas de modelamiento que se han utilizado para la predicción del precio de las acciones son:

ARIMA: Técnica estadística de serie de tiempo.

Regresión Lineal Múltiple: Técnica estadística de predicción.

XGBOOST Regression: Técnica de Machine Learning.

Los indicadores que se han utilizado para evaluar la capacidad predictiva de los modelos son:

Tabla 2
Variables Dependientes para la Aplicación de las Técnicas de Machine Learning.

Tipo de variable	Variable input	Variable calculada
Dependiente	Precio de la acción (P)	Precio en el instante T+1
Dependiente	Precio de la acción (P)	Precio en el instante T+15
Dependiente	Precio de la acción (P)	Precio en el instante T+30

Tabla 3
Variables Predictoras para Evaluar el Comportamiento de Precios.

Tipo de variable	Variable input	Variable calculada	Fórmula
Independiente	Precio	Precio en el instante T	$X_1 = Precio_T$
Independiente	Precio	Variación diaria	$X_2 = \frac{Precio_T - Precio_{T-1}}{Precio_{T-1}}$
Independiente	Precio	Variación quincenal	$X_3 = \frac{Precio_T - Precio_{T-15}}{Precio_{T-15}}$
Independiente	Precio	Variación mensual	$X_4 = \frac{Precio_T - Precio_{T-30}}{Precio_{T-30}}$
Independiente	Precio	Variación semestral	$X_5 = \frac{Precio_T - Precio_{T-120}}{Precio_{T-120}}$
Independiente	Precio	Variación anual	$X_6 = \frac{Precio_T - Precio_{T-360}}{Precio_{T-360}}$
Independiente	Precio de S&P500	Precio S&P500 en el instante T	$X_7 = Precio_{SP500}_T$

Figura N° 4
Tratamiento del análisis de sentimiento.

fecha_hora	only_date	only_time	reviews.text	senti_score	polarity	subjectivity
2021-07-23 05:26:40	2021-07-23	05:26:40	australia prepar pay econom price china trade ...	(0.0, 0.0)	0.000000	0.000000
2021-07-23 05:23:10	2021-07-23	05:23:10	leav offic angela merkel hall pioneer global l...	(0.06666666666666667, 0.36666666666666667)	0.066667	0.366667

RMSE: Mide la cantidad de error que hay entre dos conjuntos de datos.

MAPE: es una desviación porcentual absoluta media. Es una medida de precisión de predicción de un método de predicción en estadística.

Las fases de **Implementación del modelo y aceptación de la solución de pronóstico** no se han detallado en el presente informe porque implicaría la puesta en producción de un modelo de negocio y dado el enfoque de la publicación, el alcance de los resultados obtenidos fue hasta la comparación de la capacidad predictiva de los modelos.

V. RESULTADOS

Se observa que el precio de la acción Apple, entre dic-2019 y jun-2021, en general presenta una tendencia creciente pero con determinados cambios en la tendencia en determinados periodos de la ventana, donde se presentan caídas y luego recuperaciones. Además, se evidencia la presencia de una perturbación aleatoria en el precio dado que hay fluctuaciones diarias, lo cual valida la premisa

de la incertidumbre que se tiene en estos tipos de mercados financieros. En esa línea, se busca utilizar modelos predictivos para entender el comportamiento del precio y pronosticar precios futuros (ver Figura N° 5).

En la predicción de los precios de Apple, se evaluó el nivel de ajuste tanto en la información de entrenamiento y validación, también se consideraron dos escenarios: (a) predicción considerando sólo los precios históricos y (b) predicción considerando precios históricos más información textual de Twitter.

Predicción del precio de Apple para el día siguiente (ver Tabla 4)

RMSE: Se encontró que al analizar sólo la información histórica se obtuvo la mejor precisión utilizando el modelo ARIMA, seguido por el modelo XGBOOST.

MAPE: Se encontró que al analizar la información histórica más la información textual se obtuvo la mejor precisión utilizando el modelo XGBOOST.

Figura N° 5
Evolutivo del Precio de Apple



Tabla 4

Comparación de la Precisión de las Técnicas Aplicación a la Predicción de Un Día.

Fuente	Modelo	Train		Test	
		RMSE	MAPE	RMSE	MAPE
Precios históricos	ARIMA	3.58	2.31	2.76	1.81
	Regresión	3.08	2.31	5.04	3.24
	XGB	0.21	0.15	2.78	1.8
Precios históricos y Twitter	Regresión	2.99	2.25	5.12	3.3
	XGB	0.17	0.12	2.78	1.76

Predicción del precio de Apple para el día 15
(ver Tabla 5)

RMSE: Se encontró que al analizar la información histórica más la información textual se obtuvo la mejor precisión utilizando el modelo XGBOOST.

MAPE: Se encontró que al analizar la información histórica más la información textual se obtuvo la mejor precisión utilizando el modelo XGBOOST.

Predicción del precio de Apple para el día 30
(ver Tabla 6)

RMSE: Se encontró que al analizar sólo la información histórica se obtuvo la mejor precisión utilizando el modelo ARIMA, seguido por el modelo XGBOOST.

MAPE: Se encontró que al analizar sólo la información histórica se obtuvo la mejor precisión utilizando el modelo ARIMA, seguido por el modelo XGBOOST.

VI. DISCUSIÓN

En primer lugar, se ha evidenciado una variabilidad de los precios de la acción Apple, lo cual sustento la necesidad de encontrar un mejor modelo para predecir el precio futuro de la acción. La presencia de incertidumbre también está alineada con los textos de finanzas donde se comenta que las bolsas de valores son dinámicas y profundas.

En segundo lugar, con respecto a la aplicación de la técnica XGBOOST como un modelo alternativo de machine learning, se evidenció que generó mayor nivel de predicción en determinados escenarios de los considerados, lo cual está alineado con otras publicaciones realizadas, donde se consideran que este tipo de modelos por lo general obtienen mejores resultados que las técnicas tradicionales. Sin embargo, es preciso mencionar que entre train y test se observa una tendencia al *overfitting* debido a que hay diferencias importantes en los niveles de los indicadores de predicción, lo cual es importante que haya una validación frecuente de los modelos con otras ventanas de validación más recientes.

En tercer lugar, la información textual de twitter resultó significativa en la predicción del precio de al día 15, mediante la aplicación de modelos de XGBOOST. Lo cual se alinea con los resultados obtenidos por Mohan et al. (2019) por el grado de correlación entre los textos y los precios de acciones.

VII. CONCLUSIONES

Los modelos de machine learning como la regresión lineal múltiple o la técnica de Extreme Gradient Boosting (XGBOOST) resultaron ser modelos eficaces en cuanto a la capacidad predictiva de los precios de la acción Apple que cotiza en la Bolsa de Valores de New York. Lo cual abre la posibilidad de poder utilizar otros tipos de técnicas en torno al machine learning.

Tabla 5
Comparación de la Precisión de las Técnicas Aplicación a la Predicción de Quince Días.

Fuente	Modelo	Train		Test	
		RMSE	MAPE	RMSE	MAPE
Precios históricos	ARIMA	11.79	7.57	8.1	5.21
	Regresión	7.84	5.97	14.15	9.25
	XGB	0.18	0.12	8.57	5.53
Precios históricos y Twitter	Regresión	7.72	5.87	14.04	9.22
	XGB	0.13	0.07	8.8	5.54

Tabla 6
Comparación de la Precisión de las Técnicas Aplicación a la Predicción de Treinta Días.

Fuente	Modelo	Train		Test	
		RMSE	MAPE	RMSE	MAPE
Precios históricos	ARIMA	22.67	14.67	10.76	6.84
	Regresión	10.51	8.87	16.74	10.43
	XGB	0.19	0.11	6.11	3.62
Precios históricos y Twitter	Regresión	10.45	8.64	16.04	9.95
	XGB	0.14	0.07	6.33	3.79

La información textual se evidenció que tiene un aporte para la mejora de la predicción del precio de la acción Apple que cotiza en la Bolsa de Valores de New York. En esa línea es importante que se tenga un almacén de la información textual que se genera diariamente en el mundo para poder utilizarla de la mejor forma y encontrar *insights* para la predicción del precio de una acción.

VIII. RECOMENDACIÓN

Las recomendaciones que pueden obtenerse de la presente publicación se pueden abordar desde diferentes perspectivas.

Modelos alternativos. Tal como se mostró en los antecedentes de la investigación, es posible aplicar otros tipos de modelos como las redes neuronales (perceptrón multicapa, redes neuronales convolucionales, redes neuronales recurrentes, memoria de corto a largo plazo), deep learning o inclusive otras técnicas de regresión como regresión ridge, lasso, elastic net.

Fuentes de información. Este tipo de problemas de predicción de precios no sólo se da en las acciones de la Bolsa de Valores de New York, sino también en otros mercados financieros como en la Bolsa de Valores de Lima donde se negocian acciones peruanas, con lo cual también puede aplicarse las técnicas propuestas. Asimismo, en el antecedente también se precisó una aplicación al bitcoin, lo cual abre la posibilidad aplicarlo a otros tipos de activos financieros, como el mercado de las divisas que también presentan variabilidad en los tipos de cambios, entre otros.

Modelos de negocio. A partir del conocimiento de las técnicas de predicción es posible implementar una app que permita brindar información analítica a los inversores, de tal forma que se les pueda sugerir en qué activos se pueden invertir conociendo los pronósticos de los precios futuros.

IX. REFERENCIAS

- [1] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.
- [2] Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics.
- [3] Fernandes, R., & Antonelli, M. (2018). Machine Learning: A Practical Approach on the Statistical Learning Theory. Springer. doi:<https://doi.org/10.1007/978-3-319-94989-5>
- [4] Gálvez, T. (9 de Agosto de 2016). Linkedin. Obtenido de ¿Cómo medir la precisión de los pronósticos?: <https://es.linkedin.com/pulse/c%C3%B3mo-medir-la-precisi%C3%B3n-de-los-pron%C3%B3sticos-tom%C3%A1s-g%C3%A1lvez>
- [5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- [6] Khare, K., Darekar, O., Gupta, P., & Attar, V. (2017). Short term stock price prediction using deep learning. 2017 2nd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT).
- [7] Lazzeri, F. (2021). Machine learning for time series forecasting with python. Indiana: Wiley & Sons.
- [8] Mohan, S., Mullanpudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. (2019). Stock Price Prediction Using News Sentiment Analysis. IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService).
- [9] Naik, N., & Mohan, B. (2019). Study of Stock Return Predictions Using Recurrent Neural Networks with LSTM. International Conference on Engineering Applications of Neural Networks.
- [10] Sugiyama, M. (2015). Introduction to statistical machine learning. Morgan Kaufmann.
- [11] Zhang, L., & Zhan, C. (2017). Machine learning in rock facies classification: an application of XGBoost. Qingdao, China: International Geophysical Conference.
- [12] Zinn, D. (2021, Setiembre 27). Here's What Causes a Stock's Price to Go Up or Down — And How to Protect Yourself From Market Volatility. Retrieved from INVESTING: <https://time.com/nextadvisor/investing/what-causes-stock-prices-to-go-up-and-do>

Fuentes de financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.

Contribución del Autor

Roberto Evaristo Broncano (Investigación, redacción, coordinador, autor principal, administración del proyecto)