Minería de datos para trading eficiente

Alan Barriga Rosales
Joel Sebastian Avalos González
Juan Pablo Hernández Lozano

IIMAS, UNAM, CDMX, Mexico
1 de Diciembre de 2021

***Resumen*—Dado la alta volatilidad del mercado de valores, resulta útil aplicar técnicas de aprendizaje de máquina y minería de datos, los cuales basádos en fundamentos e indicadores financieros, sean capaces de ayudar al inversor en la toma de decisiones para hacer "trading"más eficiente y optimizar sus ingresos. Es por eso que en este proyecto se implementa un pipeline de distintos algoritmos de inteligencia artificial, aplicados sobre los precios de distintas acciones y fundamentados en los indicadores técnicos más usados en el ámbito del mercado de valores. Aunque los resultados son modestos, se logran modelos capaces de predecir tendencias en la bolsa.**

***Palabras Clave*—Bolsa de valores - Aprendizaje de máquina - Minería de datos - Finanzas - Acciones - Predicción - SP 500 - Trading.**

I. Introducción

En los mercados de valores, el "trading" es la especulación sobre instrumentos financieros con el objetivo de obtener un beneficio económico. Los activos más comunes en el mundo del "trading"son las divisas, criptomonedas y acciones.El "trading" tiene diversas estrategias de inversión según el inversor, sin embargo, la mayor parte de inversores prefieren basar sus portafolio bajo parámetros matemáticos como el análisis técnico; o parámetros basados en la experiencia como el análisis fundamental para la aplicación de una estrategia concreta para operar.

Pese a que la intención del inversionista promedio puede ser descrita bajo el lema de çomprar barato y vender.[es]perando retornos de inversión positivos, fáciles y rápidos, el porcentaje de traders exitosos, capaces de solventar sus vidas con su actividad en el "trading", es muy pequeño. Incluso los pocos casos de éxito tienen la tarea aún más difícil de seguir generando ganancias a largo plazo. Esto es debido a que la bolsa es impredecible y no solo esta sujeta a factores económico-financieros; factores sociales como nuevas legislaciones de impuestos en las principales economías del mundo, guerras y embargos comerciales o incluso tweets de las personas más ricas del planeta alteran la dirección del mercado.

Dada la gran volatilidad de la bolsa de valores es conveniente hacer uso de herramientas computacionales que apliquen técnicas avanzadas de minería de datos y aprovechen un gran poder de procesamiento para poder hacer miles de análisis rápidos que nos ayuden a tomar mejores decisiones de inversión. Es por eso que proponemos implementar algoritmos de aprendizaje de máquina para poder implementar un programa computacional el cual se capaz de servir como auxiliar en la tarea de hacer "trading.[en] los distintos mercados de manera más eficiente. La eficiencia a la que queremos llegar no se limita a generar beneficios dado un activo, también comprende el monitorear acciones para distintas con el propósito de ampliar nuestro portafolio de inversión con la finalidad de minimizar el riesgo.

II. Preliminares

II-A. Objetivo

De manera simplificado, el objetivo se resume en: comparar, relacionar las distintas empresas para predecir el comportamiento de ellas mismas y tomar decisiones de inversión a corto plazo. De tal forma que se cumplan los siguientes criterios de evaluació:

- Disminución de incertidumbre en inversión.
- Mayor diversificación en el portafolio de inversión.
- Optimizar tiempos en toma de decisiones.

Es así que nuestros proyecto tiene un alcance práctico, útil para cualquier inversionista.

II-B. Marco estratégico

Como ya se ha señalado, la bolsa de valores es altamente volátil e impredecible por lo que suponer un algoritmo de inteligencia artificial capaz de adivinar correctamente la evolución de una serie financiera es una exageración de las capacidades de los modelos de aprendizaje de máquina y minería de la información. Aunque modelar y predecir comportamientos de series de tiempo es posible bajo estrictos fundamentos matemáticos, el modelo a aplicar en este proyecto debe de ofrecer indicadores de confianza lo suficientemente robustos para justificar su uso en un caso práctico.

En cuanto a los datos, en este proyecto se acude a las fuentes en tiempo real de la información de las acciones tomadas en cuenta. También se toman los indicadores de riesgo financiero ofrecidos por la agencia de calificaciones estadounidense Standard Poor's para las 500 empresas con

mayor capitalización de mercado.

Por otra parte se le exige al proyecto ser entrenado por una variedad de modelos que nos ofrezcan un grado más amplió de estudio sobre los datos. Es importante resaltar que estos modelos se califican bajo diferentes métricas de evaluación para juzgar su capacidad de ser útiles en la práctica de análisis de tendencias de mercado.

Finalmente la metodología debe de estar basada en un marco teórico lo suficientemente robusto para validar el trabajo como veraz en la realidad. Es por eso que nuestros modelos se entrenan bajo una variedad de los indicadores técnicos más usados en los mercados financieros. Estos indicadores operan sobre el valor de las acciones, más en específico sobre el precio del cierre ajustado: y sobre el volumen de las mismas, que es un indicador de la fortaleza del mercado, ya que los mercados con un volumen en aumento tienden a presentar tendencias alcistas.

II-C. Definiciones

Dentro de los indicadores técnicos que sustentan nuestro análisis se encuentran:

- **Media Móvil:** cálculo utilizado para analizar un conjunto de datos en modo de puntos para crear series de promedios. Así las medias móviles son una lista de números en la cual cada uno es el promedio de un subconjunto de los datos originales. Una media móvil simple (Moving Average) es la media aritmética de los n datos anteriores.
- **Momento:** mide la tasa de subida o bajada de los precios de las acciones. Desde el punto de vista de la tendencia, el impulso es un indicador muy útil de fortaleza o debilidad en el precio de la emisión.
- **Disparidad:** Mide la posición relativa del precio de cierre más reciente de un activo a un promedio móvil seleccionado y reporta el valor como un porcentaje. Está definida por:

$$\text{Disparity} = \frac{\text{MP} - \text{MA}}{\text{MA} \times 100}$$

  donde MP es el precio del mercado y MA es la media móvil.
- **Bias:** El sesgo también se conoce como tasa de desviación, que es un índice técnico derivado del principio de media móvil. Su función es medir el grado de desviación entre el precio de las acciones y la línea de promedio móvil en el proceso de fluctuación.
- **PSY:** un indicador, es la relación entre el número de períodos crecientes sobre el número total de períodos. Refleja la compra poder en relación con el poder de venta. Si PSY está por encima del 50 %, indica que los compradores tienen el control. Igualmente, si está por debajo del 50 %, indica que los vendedores tienen el control. Si el PSY se mueve a lo largo del área del 50 %, indica el equilibrio entre los compradores y vendedores y por lo tanto no hay movimiento de dirección para el mercado.
- **On Balance Volume:** relaciona el volumen con los cambios de precio que han acompañado este volumen. se basa en el principio que los cambios del OBV preceden los cambios de precios. Según este principio, el aumento del volumen de balance significa que en el instrumento invierten los profesionales. Cuando más tarde el gran público empieza a invertir, el precio y los valores del indicador OBV empiezan a crecer bruscamente. Si el precio adelanta en su movimiento el indicador On Balance Volume, ocurre así llamada "la falta de confirmación". Se puede observarlo en el pico del mercado alcista o en la base del mercado bajista.
  Si el precio de cierre actual es más alto que el anterior, entonces:

$$\text{OBV}(i) = \text{OBV}(i-1) + \text{VOLUME}(i)$$

  Si el precio de cierre actual es más bajo que el anterior, entonces:

$$\text{OBV}(i) = \text{OBV}(i-1) + \text{VOLUME}(i)$$

  Si el precio de cierre actual es igual al precio anterior, entonces:

$$\text{OBV}(i) = \text{OBV}(i-1)$$

III. Desarrollo

III-A. Integración de la información

El proyecto empieza con la recolección de los datos financieros concernientes a las empresas más importantes en el ámbito de tecnología del planeta: Google, Tesla, Microsoft, Apple, Facebook, Amazon y Netflix. Estos datos son colectados en tiempo real mediante la API de Yahoo Finance de python. Las temporalidades definidas para los datos son las siguientes:

- Se manejaron ventanas de 1 hora en los precios de las acciones.
- El histórico de datos es de 2 semanas en el pasado a partir del momento de consulta a la API.
- Se pretende hacer predicciones para 1 día de actividad.

III-A1. Variables totales: Es así que se logran colectar los siguientes datos para cada una de las empresas:

- Precio de apertura (`Open`)
- Precio de cierre (`closeHigh`)
- Precio más alto (`High`)
- Precio más bajo (`Low`)
- Precio de cierre ajustado (`Adj Close`)
- Volumen (`Volume`)

III-A2. Transformación de los datos: El procesamiento de los datos consiste pasar el precio de cierre ajustado y el volumen a nuestros indicadores técnicos. Son los datos procesados por los indicadores técnicos los que de evalúan en los modelos pero antes de este paso se elaboran los conjuntos de entrenamiento y prueba correspondientes y se hace un escalamiento de los datos.

III-A3. Analytical Base Table (ABT): El proceso descrito sobre el procesamiento e integración de los datos puede verse en la figura 1 a continuación:

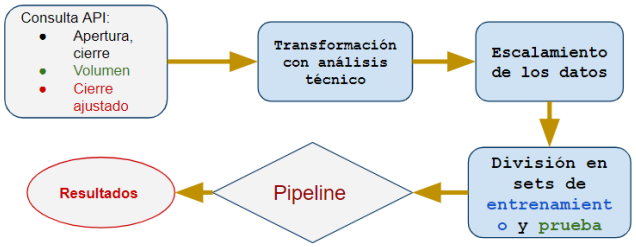

Figura 1: Flujo de los datos dentro del marco de trabajo.

III-B. Modelación

Como ya se ha mencionado, para darle mayor robustez al proyecto se ha decidido implementar un "pipeline.[el] cual integre distintos modelos de aprendizaje de máquina. Estos modelos se describen de manera concisa y superficial a continuación:

- **Regresión lineal:** Modelado estadístico para describir una variable de respuesta como función de una o varias variables predictoras.
- **Series de tiempo (Holt-Winters):** Movimientos a largo plazo en la información permiten la elaboración de pronósticos a corto plazo
- **Árboles de decisión:** Se fabrican diagramas de construcciones lógicas que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva.
- **Regresión logística:** Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores.
- **Redes neuronales:** Las redes neuronales se pueden utilizar para realizar predicciones sobre datos de series de tiempo. Detectan patrones en los datos de entrada y producen una salida libre de ruido.

Los modelos de regresión de árboles de decisión son en principio modelos más simples de aplicar, sin embargo corren el riesgo de trivializar el problema y generar un desempeño más pobre. Por otra parte el modelo de Holt-Winters para serires de tiempo es un algoritmo específico para modelar series financieras, pero al igual que las redes neuronales, su aplicación requiere de un mayor trabajo de código.

III-B1. Descripción de las técnicas o métricas de evaluación del modelo: Dentro de las métricas de evaluación usadas, el mejor indicador es el puntaje de precisión, el cuál nos indica de forma porcentual como qué tanto se acercan los resultados de las predicciones en comparación a un conjunto de pruebas. Por otra parte también se hace uso del error cuadrático medio que mide el promdio de las diferencias entre las predicciones y los datos reales.

Por otra parte, las métricas de Recall y F1 score nos dan una mejor idea del comportamiento de los datos catalogados como verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) o falsos negativos (FN),

La métrica de exhaustividad (Recall) nos va a informar sobre la cantidad que el modelo de machine learning es capaz de identificar. Dada por la siguiente ecuación:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

El valor F1 se utiliza para combinar las medidas de precision y recall en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

$$\text{F1} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

GOOGLE	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.5527	0.6190	0.6180	0.6190	0.5714	0.5294	-2.07519
MSE	0.4523	0.3810	0.3820	0.3810	0.4286		
Recall	0.2	0.25	0.10	0.666	0.6363		
F1	0.1426	0.619	0.1820	0.5	0.6087		

AAPL	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.5055	0.5069	0.5238	0.5238	0.5166	0.5588	-1.4673
MSE	0.4945	0.4931	0.4762	0.4762	0.4834		
Recall	1.0	0.333	0.4166	0.5833	0.666		
F1	0.5517	0.4	0.5	0.5833	0.5217		

MSFT	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.5305	0.5055	0.5714	0.5027	0.5541	0.5490	-2.0740
MSE	0.4695	0.4945	0.4285	0.4973	0.4459		
Recall	0.6	0.1538	0.5833	0.7777	0.6250		
F1	0.5217	0.2666	0.6086	0.5384	0.5000		

AMZN	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.5238	0.5238	0.5240	0.4286	0.5714	0.4901	-5.3884
MSE	0.4761	0.4230	0.4761	0.5714	0.4285		
Recall	0.2500	0.2500	0.4285	0.3333	0.3000		
F1	0.2857	0.3400	0.3750	0.4545	0.4000		

Figura 2: Métricas de evaluación de los modelos para algunas empresas

FB	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.5678	0.5696	0.5500	0.6000	0.5196	0.5454	-2.8868
MSE	0.4322	0.4304	0.4500	0.5555	0.4804		
Recall	0.1250	0.5000	0.7143	0.5555	0.3750		
F1	0.2000	0.4400	0.5263	0.5555	0.3999		

NFLX	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.5089	0.5410	0.6589	0.5670	0.5444	0.5858	-0.9152
MSE	0.4804	0.4590	0.3411	0.4330	0.4556		
Recall	1.0	0.75	0.6366	0.5000	0.6000		
F1	0.6923	0.5714	0.6086	0.5555	0.6315		

TSLA	Regresion Logistica	Kernel PCA	Tree Decision PCA	Kernel PCA sklearn	Tree Decision sin PCA	Holt Winter	Regresion Lineal
Score	0.4607	0.5678	0.5571	0.6214	0.5767	0.5454	-0.5511
MSE	0.5393	0.4322	0.4429	0.3786	0.4233		
Recall	0.6666	0.1818	0.4545	0.2857	0.6666		
F1	0.4444	0.3070	0.5882	0.5882	0.5000		

Figura 3: Métricas de evaluación de los modelos para algunas empresa

III-B2. Modelos:

IV. Resultados

IV-A. Conclusión de los modelos utilizados

Como podemos observar, de manera general no podemos establecer un modelo preferido para el caso de las clasificaciones aunque el que nunca obtuvo la mayor eficiencia fue el de la regresión logística, algo esperado ya que notenía el perfilamiento idóneo para el problema. Sin embargo, es importante mencionar que para las series de tiempo el modelo de regresión lineal tiene un desempeño malo por lo que decidimos quedarnos únicamente con el modelo de Holt-Winters, además de las razonas mencionadas con anterioridad en el trabajo.

Por otro lado, podemos observar que **Kernel PCA** logra tener un buen margen en comparación con el resto de modelos, y a pesar de que los árboles de clasificación tenían la posibilidad de trivializar el problema, tuvieron un desempeño competitivo.

IV-B. *Resultados finales*

Dadas las observaciones mencionadas con anterioridad es importe remarcar la importancia de ”diversificar”los métodos para clasificar y las series de tiempo, ya que además de encontrar el mejor score nos pueden servir como un segundo apoyo para portar alguna decisión o poner en duda la veracidad de nuestro resultado óptimo.

IV-C. *Enfoque más matemático*

V. Conclusiones

Aunque predecir el comportamiento de series financieras no es tarea fácil hemos llegado a resultados interesantes. Partimos del análisis técnico, donde las métricas utilizadas nos permiten agilizar el proceso de toma de decisiones por sí solas sin necesidad de un análisis de aprendizaje de máquina. Esto significa que un inversor con experiencia, puede consultar los resultados que estos indicadores ofrecen para realizar su toma de decisiones.

Por otra parte, el uso de programas computacionales nos permiten monitorear y diversificarnos en varias acciones, minimizando así el riesgo de inversión de nuestros portafolios. Aunque se trabajaron con acciones de un relativamente fácil manejo, dado que están entre las empresas más ricas del mundo y cuentan con distintos indicadores de confianza (SP 500), los resultados de este proyecto pueden ser generalizados a otro tipo de activos como divisas, criptomonedas.

Aunque varios modelos parecen dejar las predicciones a la suerte, otros sí logran resultados aplicables. Más en específico, el modelo de redes neuronales parece ser bueno dado que acierta el precio de las acciones pero no en la ventana de tiempo prevista. Por tanto, una mejor implementación de este modelo podría explorar tomar un histórico de datos más amplio para poder ajustar la temporalidad de la inversión a un solo día. Por otra parte, el mejor modelo de la serie de financiera está dado por Holt-Winters. Este resultado era predecible pues el modelo Holt-Winters se especializa en el ámbito de series financieras.

Finalmente, este proyecto sirve de base para trabajos futuros donde se mejoren algunos parámetros en los modelos y se complementen que indicadores técnicos más complejos.

Referencias

[1] Examples(scikit-learn), https://scikit-learn.org/stable/auto$_{e}xamples/index.html{c}lassification$

[2] Classifier comparison(scikit-learn) https://scikit-learn.org/stable/auto$_{e}xamples/{c}lassification/plot_{c}lassifier_{c}omparison.html{s}phx - glr - auto - examples - {c}lassification - plot - {c}lassifier - comparison - py$

[3] Decision Tree Regression(scikit-learn), https://scikit-learn.org/stable/auto$_{e}xamples/tree/plot_{t}ree_{r}egression.html{s}phx - glr - auto - examples - tree - plot - tree - regression - py$

[4] Principal components analysis(PCA)(scikit-learn), https://scikit-learn.org/stable/auto$_{e}xamples/decomposition/plot_{p}ca_{3}d.html{s}phx - glr - auto - examples - decomposition - plot - pca - 3d - py$

[5] Kernel PCA (scikit-learn), https://scikit-learn.org/stable/auto$_{e}xamples/decomposition/plot_{k}ernel_{p}ca.html{s}phx - glr - auto - examples - decomposition - plot - kernel - pca - py$

[6] Examples(scikit-learn), https://scikit-learn.org/stable/auto$_{e}xamples/index.html{e}xamples$

[7] Model Selection(scikit-learn), https://scikit-learn.org/stable/auto$_{e}xamples/index.html{m}odel - selection$

[8] Plotting Cross-Validated Predictions (scikit-learn),

https://scikit-learn.org/stable/auto$_{e}xamples/model_{s}election/plot_{c}v_{p}redict.html{s}phx - glr - auto - examples - model - selection - plot - cv - predict - py$

[9] SP500(investing), https://mx.investing.com/indices/us-spx-500

[10] yfinance, https://pypi.org/project/yfinance/

Anexo

V-A. *Metodo Holt Winters*

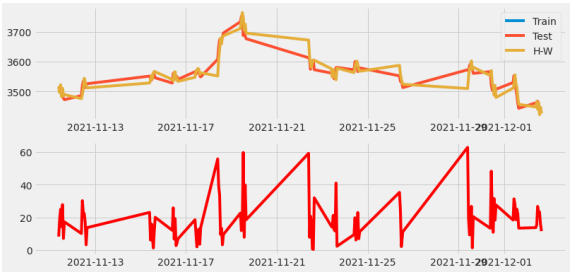

Figura 4: Empresa Amazon aplicando el metodo Holt Winters

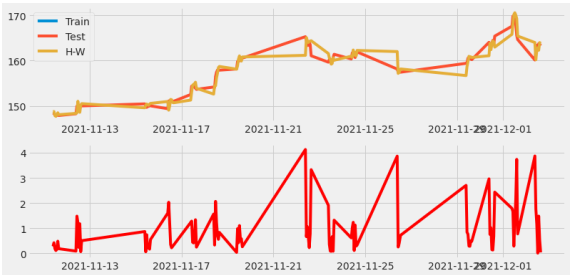

Figura 5: Empresa Apple aplicando el metodo Holt Winters

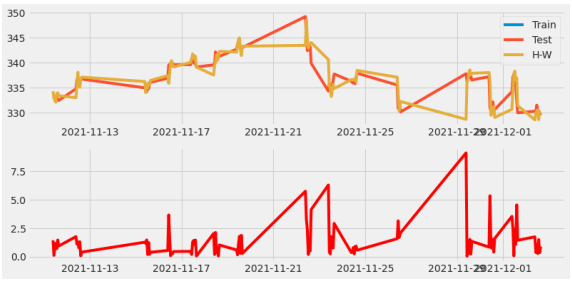

Figura 6: Empresa Microsoft aplicando el metodo Holt Winters

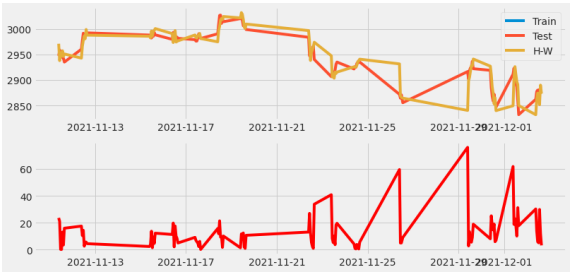

Figura 7: Empresa Google aplicando el metodo Holt Winters

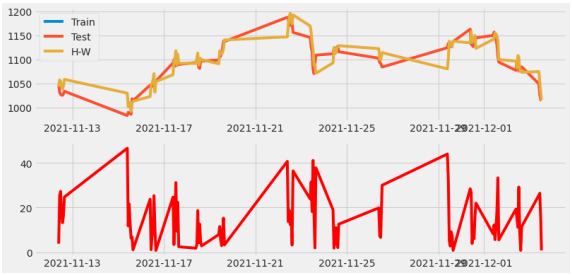

Figura 8: Empresa Tesla aplicando el metodo Holt Winters

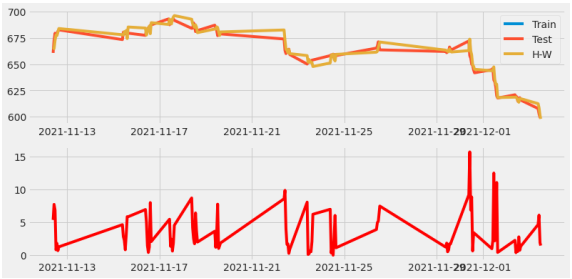

Figura 9: Empresa Netflix aplicando el metodo Holt Winters

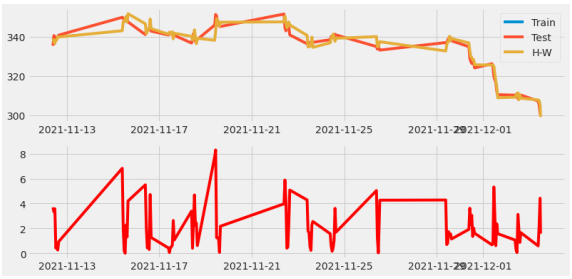

Figura 10: Empresa Facebook con su error

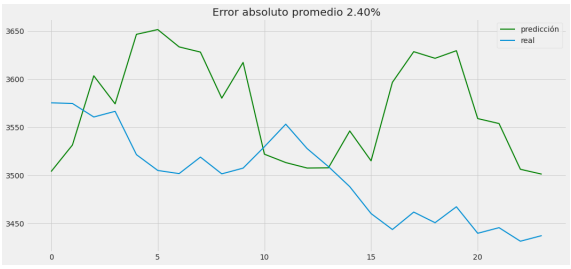

Figura 11: Empresa Amazon con su error

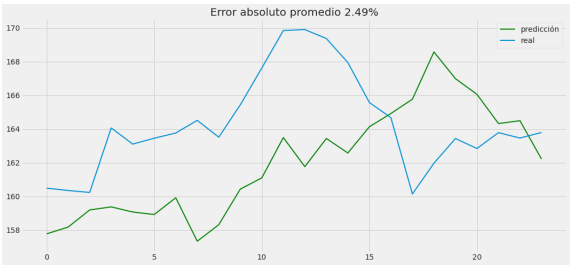

Figura 12: Empresa Apple con su error

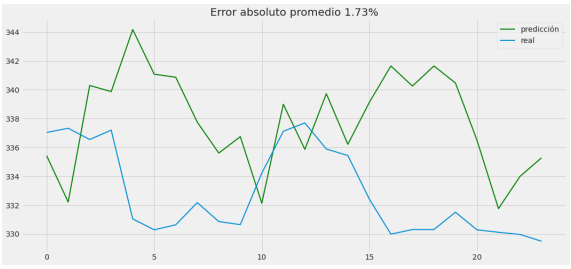

Figura 13: Empresa Microsoft con su error

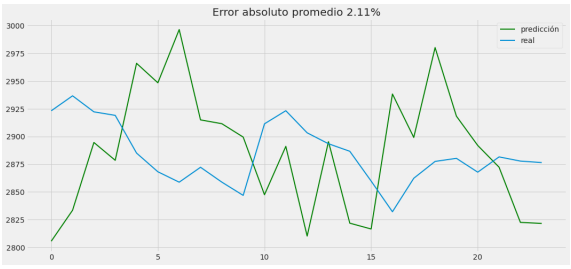

Figura 14: Empresa Google con su error

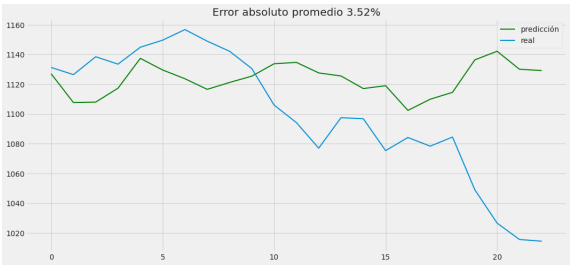

Figura 15: Empresa Tesla con su error

V-B. Error

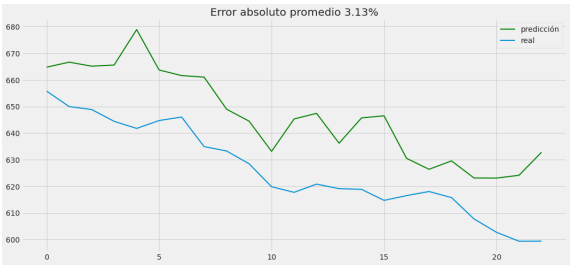

Figura 16: Empresa Netflix con su error

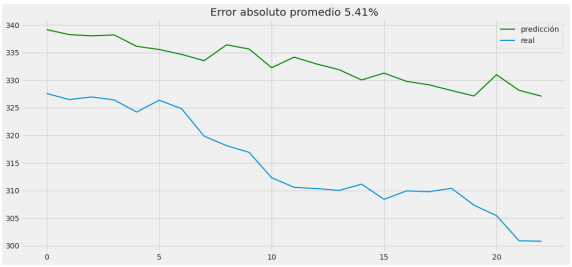

Figura 17: Empresa Facebook con su error

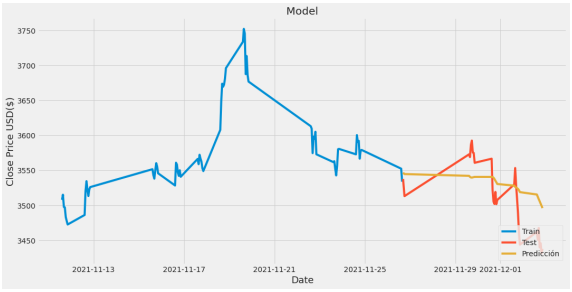

Figura 18: Empresa Amazon con su predicción

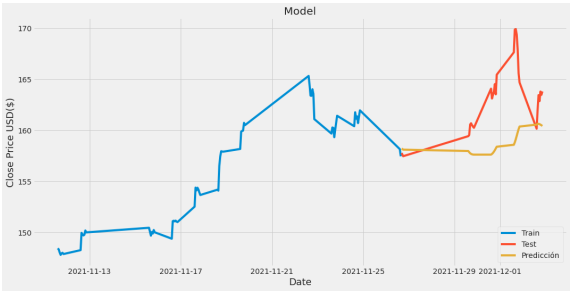

Figura 19: Empresa Apple con su predicción

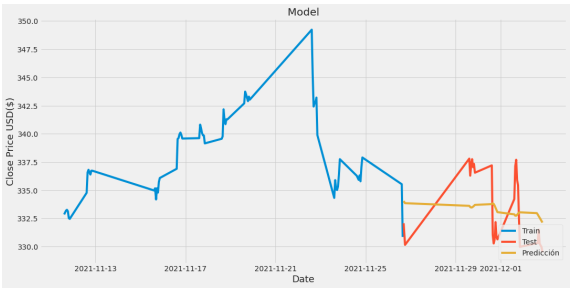

Figura 20: Empresa Microsoft con su predicción

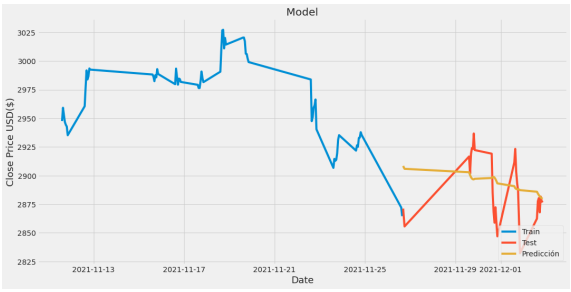

Figura 21: Empresa Google con su predicción

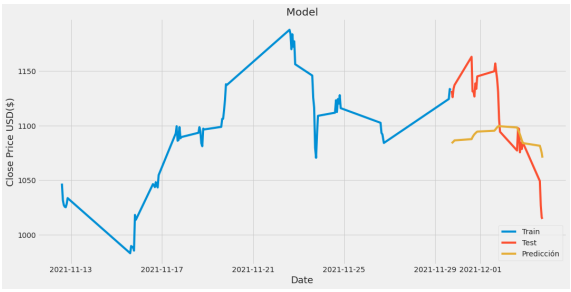

Figura 22: Empresa Tesla con su predicción

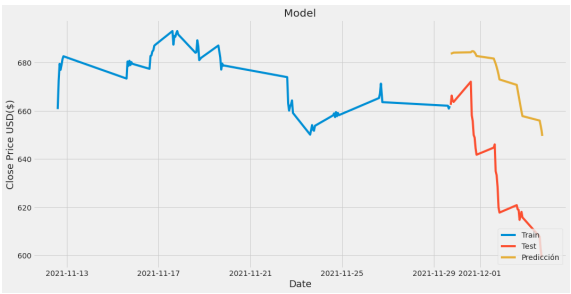

Figura 23: Empresa Netflix con su predicción

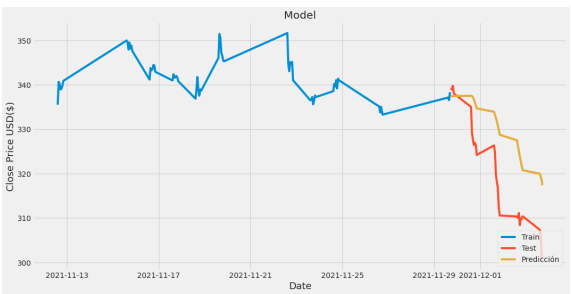

Figura 24: Empresa Facebook con su predicción