Predicción del precio de acciones de la bolsa de valores estadounidense utilizando técnicas de aprendizaje automático basadas en datos de análisis técnico

1. Introducción

El mercado de valores se define como un mercado público donde empresas ponen a la venta acciones, que representan una parte de su valor total, y los inversionistas compran estas acciones esperando recibir un beneficio a corto, mediano o largo plazo en relación al comportamiento del precio de la acción de una empresa determinada. La importancia de este concepto reside en la gran capacidad que tiene de mover la economía de una nación, beneficiando en el proceso a los actores en este mercado, de los cuales -desde la visión de este proyecto- se cubren únicamente empresas e inversionistas.

Con el tiempo se desarrollaron distintas técnicas de análisis bursátil con bases matemáticas para identificar el comportamiento de una acción obteniendo así información de gran relevancia para establecer futuras estrategias de inversión sobre dicha acción. Los datos de estos tipos de análisis bursátil despertaron la curiosidad de los investigadores, y con el aumento de la capacidad de cómputo y la digitalización de los datos se pusieron a prueba estos procesos para la predicción de acciones haciendo uso del aprendizaje de máquina.

De esta manera, este proyecto se centra en analizar el comportamiento de algunas técnicas de aprendizaje automático aplicado al análisis técnico y fundamental, los cuales son dos tipos de análisis bursátil clásicos, para la predicción del precio de un conjunto de acciones de la bolsa de valores estadounidense.

El proyecto abarcó la selección de las fuentes de los datos de análisis técnico y fundamental, la preparación de los conjuntos de datos para la implementación de los modelos la identificación de variables o atributos influyentes en la predicción, en seleccionar tres técnicas de aprendizaje automático, La estimación de hiper-parámetros de los modelos a construir y evaluar el comportamiento de cada modelo.

2. Fundamentación teórica

2.1. Análisis bursátil

El análisis bursátil representa los métodos analíticos para obtener información relevante de un mercado de valores que pueda ayudar a la toma de decisiones de inversión. Se destacan principalmente dos tipos de análisis bursátil clásico, los cuales son: análisis técnico y análisis fundamental.

El análisis técnico es un tipo de análisis bursátil, que se encarga del estudio de los movimientos pasados de los precios con el objetivo de predecir los movimientos futuros de los precios a partir de los movimientos del pasado.

El análisis fundamental es el estudio del potencial financiero que tiene una compañía o empresa, este estudio se basa en datos históricos consignados en reportes periódicos; en el sector de la empresa y su posicionamiento en la industria, así como también influyen factores administrativos, dividendos y capitalización para poder identificar un potencial de crecimiento a futuro.

2.2. Predicción del mercado de valores

Predicción del mercado de valores: La predicción de acciones de la bolsa de valores es el proceso por el cual se determina el valor futuro de la acción de una compañía o varias compañías (portafolio). Este ha sido un tema ampliamente debatido, lo que ha generado distintas posiciones en el ámbito académico. En [1] se sostiene que el mercado de valores es predecible hasta cierto punto, cuando se tienen en cuenta comportamientos económicos y socio-económicos desde un punto de vista teórico-financiero.

2.3. Técnicas de aprendizaje automático utilizadas

Para este proyecto se decidió utilizar 3 técnicas de aprendizaje automático tanto para los datos de análisis técnico, así como para los datos de análisis fundamental. Las técnicas empleadas en los modelos fueron bosques aleatorios, Máquinas de vectores soporte y Perceptrón multicapa. Estas tres técnicas fueron seleccionadas con base a la información recopilada de distintos artículos en [2], donde destacan la participación de estas tres técnicas en la mayoría de los artículos revisados, siendo ANN y SVM los más utilizados, puesto que según [3][4], consiguen una generalización potencial mucho mayor que sus contrapartes.

2.4. Datos utilizados

Se utilizaron datos de 5 empresas pertenecientes al S&P 500 uno de los índices más importantes de estados unidos, a los datos obtenidos se les aplicaron algoritmos como análisis de componentes principales, ventana deslizante y estandarización, esto para generar distintos conjuntos de datos los cuales se compararon para observar cuales eran los que permiten un mejor desempeño para las tres técnicas de aprendizaje automático que se usaron (máquina de vectores de soporte, perceptrón multicapa y bosques aleatorios).

A partir de los datos obtenidos de cada uno de los análisis bursátiles tratados, se obtuvieron nuevas variables. Siendo en el caso del análisis técnico indicadores y en el análisis fundamental ratios. Estos se seleccionaron según la información disponible en internet acerca de estos indicadores y ratios eligiendo los que más utilizan los expertos.

3. Resultados

3.1. Primera etapa

En la primer etapa se pone prueba los distintos conjuntos de datos generados, con el fin de descartar los que tenían un peor desempeño y obtener el conjunto con el que se tenia menor error respecto a las métricas RMSE y MAE.

Esta etapa no se centra en la búsqueda de hiperparámetros. En este proceso de la primer etapa se generaron aproximadamente un total de 180 modelos de aprendizaje automático basados en cada una de las 3 técnicas seleccionadas alimentado con cada uno de los 8 conjuntos de datos para cada una de las 5 acciones.

Para poder hacer tener un punto de comparación para el resto de los modelos se construyo un modelo base. Este modelo predice el precio de mañana como si fuese el mismo precio de hoy.

Tanto para análisis técnico como para análisis fundamental en la primera etapa los conjuntos de datos con los cuales los modelos tuvieron un mejor desempeño fueron los que se muestran en Tabla 1. Por lo tanto, estos conjuntos de datos serán utilizados para entrenar y evaluar los modelos en la segunda etapa.

Tabla 1: Conjuntos de datos seleccionados según técnica. Análisis técnico y fundamental

Técnica	Conjunto
SVR	SXY
RF	SXY
MLP	SX

3.2. Segunda etapa

La segunda etapa respecto al desarrollo de los modelos consistió en realizar el entrenamiento y prueba de los mismos con los conjuntos de datos seleccionados de la etapa 1. Una vez se selecciona el conjunto de

datos se realiza una exploración de los hiper-parámetros mucho más exhaustiva, con rangos más amplios. Se utiliza el algoritmo GridSearchCV (de la librería de python Sci-kit Learn) para seleccionar los atributos que den un mejor resultado, respecto a las métricas utilizadas.

3.2.1. Análisis técnico

En la mayoría de las acciones no fue posible obtener modelos que sobrepasan la métrica RMSE del modelo base construido, como se puede observar en FIGURA, los resultados fueron muy similares entre una técnica y otra. Observando el detalle, el único modelo que en todas las acciones tuvo mejor comportamiento que el modelo base fue el modelo del SVR-SXY-T, seguido del RF-SXY-T, el cual en dos de las cinco acciones logró un mejor desempeño que el modelo base. Y dejando al modelo MLP-SX-T en último lugar, teniendo un error superior para la predicción en todas las acciones utilizadas.

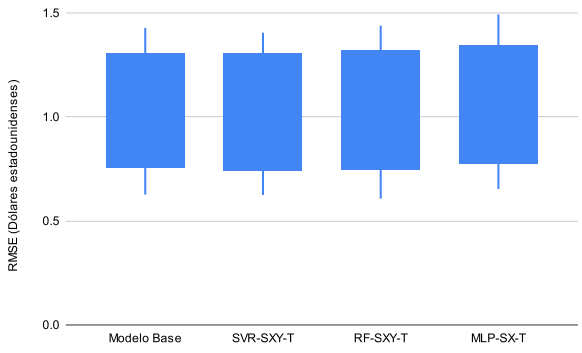

Figura 1: Análisis técnico: Diagrama de cajas del error obtenido (RMSE) por los modelos resultantes y el modelo base

El modelo que más se aproxima a los resultados del modelo base es SVR-SXY. Los otros dos modelos RF-SXY-T y MLP-SX-T, se encuentran ligeramente elevados uno del otro. Este comportamiento de los modelos con datos de análisis técnico puede ser debido a diversas razones. Una de ellas es la dificultad existente en predecir el precio, lo cual es mucho más complejo que predecir el movimiento del precio. Por otro lado, puede que los indicadores usados sean de utilidad para predecir la tendencia, momentum, volatilidad, volumen y la fuerza pero no sean tan efectivos en la ayuda de la predicción del precio. Por el lado de los modelos, visualmente se puede explicar el mejor desempeño del modelo SVR-SXY-T, dado que pueden observar similitudes visualmente con las "predicciones"del modelo base, donde básicamente se predice que el precio de mañana sería el mismo que el de hoy y no estaría haciendo realmente una predicción válida.

3.2.2. Análisis fundamental

Según los resultados, se observa que el RMSE muestra que, en promedio, el modelo SVR-SXY-T obtiene errores ligeramente menores de el modelo MLP-SX-T. Siendo el último lugar para RF-SXY-T con un promedio notablemente mayor al de los otros dos modelos.

Sin embargo, al observar con mayor amplitud los resultados de todos los modelos como se muestra en Figura 2 es posible generar mayor claridad sobre los resultados que han tenido cada uno de los modelos. Se visualiza la caja generada por los resultados del modelo MLP-SX-T como la que se concentra en valores menores que los demás, aunque teniendo un valor máximo muy prominente. Sin embargo, este dato puede ser tratado como un dato atípico, por lo que sería la opción correcta escoger el modelo MLP-SX-T como el que mejores resultados sobre métricas de error respecta. Con respecto a los otros dos modelos construidos, se puede declarar al SVR-SXY-T como un modelo que genera menores errores que el modelo RF-SXY-T.

Es posible observar que los resultados han sido positivos para los datos de análisis fundamental. Como se muestra en la Figura 2 las cajas se mantienen por debajo del modelo base, lo cual es, a priori, un buen indicador del desempeño de estos modelos con datos de análisis fundamental.

Cabe la posibilidad que el uso de estos datos fundamentales haya potenciado la capacidad de predicción de los modelos, dada la información mucho más re-

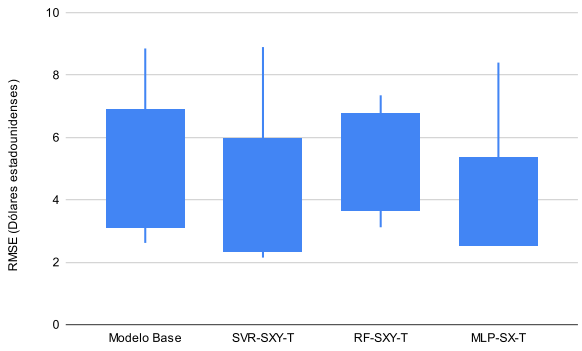

Figura 2: Análisis fundamental: Diagrama de cajas del error obtenido (RMSE) por los modelos resultantes y el modelo base

lacionadas a la realidad de la empresa que dan los ratios fundamentales. También es posible que, debido a la poca cantidad de datos como para entrenar modelos para estas técnicas, el conjunto de pruebas no sea lo suficientemente grande para tener mayor confiabilidad en las métricas de error presentadas.

3.3. Backtesting

El proceso de backtesting consistió en poner a prueba las acciones PEP e IBM, las cuales representan la acción que mejor y peor desempeño tuvieron en las predicciones respectivamente. Para esto se aplica la estrategia de comprar bajo y vender alto. El criterio para valorar los resultados obtenidos del backtesting se encuentra en las ganancias en dólares estadounidenses obtenidas a lo largo de dos trimestres, desde 2020-09-30 hasta 2021-03-31, donde para análisis técnico representa cada día de este periodo de tiempo y para el fundamental representa 3 periodos, es decir, se tienen 3 fechas. Se utilizan los tres modelos construidos en la segunda etapa para realizar las predicciones.

Se hace uso de una gráfica que describe las acciones tomadas por el algoritmo de backtesting implementado. Las tres variables que se encuentran en la gráfica cambian de valor dependiendo de la acción tomada si es compra o venta.

3.3.1. Backtesting para análisis técnico

Los resultados que se aprecian en la Tabla 2 muestran que el modelo que mejor desempeño obtuvo en el backtesting, es RF con una ganancia de $7,54 (siete dolares con cincuenta y cuatro centavos) en PEP y de $25,67 (Veinticinco con sesenta y siete centavos) en IBM. Mientras que el modelo que implementa SVR tuvo el peor desempeño de los tres modelos utilizados. Esta es una situación a destacar, dado que según los resultados mostrados hasta ahora con las métricas utilizadas (RMSE y MAE) el modelo de máquinas de vectores de soporte es el que presenta un mejor comportamiento en las predicciones, según las métricas. Sin embargo, estas métricas no evidencian unos mejores resultados en el backtesting. Este fenómeno podría deberse a lo descrito con anterioridad en el análisis de resultados, donde se decía que la predicción que realiza el modelo de SVR es similar a lo que propone el modelo base.

Tabla 2: Análisis técnico: Resultado del backtesting para las distintas técnicas

Análisis Técnico	Acción	Inicial	Ganancia	Valor final
RF	PEP	$1000	$7,54	$1007,54
RF	IBM	$1000	$25,67	$1025,67
SVR	PEP	$1000	$3,41	$1003,41
SVR	IBM	$1000	$6,17	$1006,17
MLP	PEP	$1000	$4,79	$1004,79
MLP	IBM	$1000	$20,67	$1020,67

3.3.2. Backtesting para análisis fundamental

Los resultados que se aprecian en la Tabla 3 muestran que el modelo que mejor desempeño obtuvo en el backtesting, es RF con una ganancia de $10,59 (Diez con cincuenta y nueve centavos) en PEP y $14,74 (Catorce con setenta y cuatro centavos) en IBM. Siendo, en este caso, SVR el segundo lugar y MLP en ter-

Tabla 3: Análisis fundamental: Resultado del backtesting para las distintas técnicas

Análisis Fundamental	Acción	Inicial	Ganancia	Valor final
RF	PEP	$140	$10,59	$150,59
RF	IBM	$140	$14,74	$154,74
SVR	PEP	$140	$4,92	$144,92
SVR	IBM	$140	$14,74	$154,74
MLP	PEP	$140	$0	140
MLP	IBM	$140	$0	140

cer lugar generando $0 dólares en ganancias, puesto que según la predicción que se dio el precio iba en caída, por lo que no sería rentable comprar. Sin embargo, en realidad el precio de estas dos acciones no se comportó así, por lo tanto, la predicción dada por el modelo es errónea. Como se pudo ver gráficamente en los resultados, el modelo de MLP suele quedar levemente por debajo del precio real y esto puede causar que se vea siempre la predicción como pérdidas.

4. Discusión y conclusiones

4.1. Discusión

Durante el desarrollo de este proyecto se tuvo en cuenta un factor muy importante, el cual es la composición o formato del conjunto de datos. En este caso se quiso realizar una exploración sobre los atributos, teniendo de esta forma un conjunto de atributos seleccionados manualmente y otro conjunto con los componentes resultantes de la aplicación del algoritmo PCA. Por otro lado, también se exploró el formato o estructura en la que se presentan los datos, teniendo la forma natural de los datos, en la cual se tiene el precio del periodo actual para predecir el del periodo siguiente, y la forma en la que se le aplica una ventana deslizante al conjunto de datos.

Según los resultados que se obtuvieron en este proyecto, ninguna de estas dos aproximaciones adicionales al estado natural de los datos generó errores menores en comparación. Por lo tanto, estos fueron descartados en la primera etapa.

Los resultados obtenidos con los modelos finales sostienen las afirmaciones encontradas en la literatura, la cual menciona que en primer lugar, la técnica de aprendizaje automático que genera un menor error en la predicción son las ANN (Artificila Neurla Networks) en genera, siendo esto comprobable en el caso del análisis fundamental donde MLP se destacó sobre las demás técnicas. Y en segundo lugar, la literatura sitúa a la técnica de SVR, la cual resultó la mejor para los datos de análisis técnico.

Los dos enfoques seleccionados (análisis técnico y fundamental) también sugieren distintas formas en las que se puede tratar el mismo problema de la predicción de acciones, salvando la diferencia entre periodos. Y se muestra en este proyecto que los datos de análisis fundamental también son competitivos en este ámbito, puesto que los datos técnicos son los más utilizados según la literatura. Por ende, es beneficioso tener en cuenta los datos de análisis fundamental a la hora de abordar temas de esta disciplina.

4.2. Conclusiones

En primer lugar se logra concluir para este proyecto que los conjuntos con variables calculadas, así como ratios e indicadores, generan un error menor como se evidencia en los resultados de los modelos en comparación con los demás conjuntos de datos.

La técnica con la que se consiguieron mejores modelos con datos de análisis técnico fue el Regresor de Vectores de Soporto con un RMSE promedio de $1.07 y un MAE promedio de $0.5. Por otro lado, para el análisis fundamental la técnica que mejor desempeño tuvo fue el Perceptrón Multicapa con un RMSE promedio de $4.75 y un MAE promedio de $3.1.

Por último, se evidencia en un backtesting bastante superficial, que los modelos que implementan Bosques Aleatorios consiguieron mejores retornos de inversión. Este resultado puede estar sesgado debido a factores como la cantidad de datos procesados.

Referencias

[1] C. Chen, W. Dongxing, H. Chunyan, and Y. Xiaojie, "Exploiting social media for stock market prediction with factorization machine," *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*, vol. 2, pp. 142–149, 2014.

[2] I. K. Nti, A. F. Adekoya, and B. A. Weyori, *A systematic review of fundamental and technical analysis of stock market predictions*, vol. 53. Springer Netherlands, 2020.

[3] L. A. Teixeira and A. L. I. De Oliveira, "A method for automatic stock trading combining technical analysis and nearest neighbor classification," *Expert Systems with Applications*, vol. 37, no. 10, pp. 6885–6890, 2010.

[4] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of ARIMA and artificial neural networks models for stock price prediction," *Journal of Applied Mathematics*, vol. 2014, pp. 9–11, 2014.