Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER-M2

Área: Data Analysis y Big Data

Batir al índice de renta variable S&P 500 con la mejor selección de sectores.

Aplicaciones Prácticas de Machine Learning y Análisis Económico

---

Autor: Manuel García Guillén

Tutor: David García Agudiez

Profesor: Albert Solé Ribalta

Ciudad de México, 16 de mayo de 2024

Créditos/Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	Batir al índice de renta variable S&P 500 con la mejor selección de sectores.
Nombre del autor:	Manuel García Guillén
Nombre del colaborador/a docente:	Albert Solé Ribalta
Nombre del PRA:	David García Agudiez
Fecha de entrega (mm/aaaa):	07/2024
Titulación o programa:	Máster en Ciencia de Datos
Área del Trabajo Final:	M2.882 - TFM - Área 5 - Aula 1
Idioma del trabajo:	Español
Palabras clave	Deep Learning, Optimización de Carteras, Estrategia de Inversión

Dedicatoria/Cita

No nos atrevemos a muchas cosas porque son difíciles, pero son difíciles porque no nos atrevemos a hacerlas. - Séneca

Agradecimientos

A Oliva, por su apoyo y comprensión.

A mis padres, por enseñarme el valor de la perseverancia y la educación. Su ejemplo me ha guiado.

Resumen

Superar el rendimiento del índice S&P 500 de forma consistente es un reto complejo debido a una diversidad de factores como la eficiencia y la volatilidad del mercado, los costos de transacción, la selección adecuada de los activos y los tiempos correctos para la compra y la venta. En este trabajo, trataremos de solventar algunas de estas dificultades para conseguir una estrategia de inversión avanzada basada en el análisis de datos que iniciará con una extensa recopilación de datos provenientes de fuentes abiertas, como indicadores del mercado financiero de EE. UU., datos económicos significativos y características técnicas desarrolladas mediante ingeniería de datos. Optimizaremos la selección de variables críticas mediante los métodos más apropiados, buscando las combinaciones más efectivas para maximizar el desempeño de nuestros modelos predictivos, seleccionados de un conjunto de técnicas de machine learning y deep learning. Estas herramientas nos permitirán predecir con precisión la rentabilidad de los activos y elegir mensualmente los prospectos más prometedores. La estrategia se perfeccionará a través de la optimización de la cartera, basándonos en principios económicos sólidos y se validará en un entorno de backtesting, con el objetivo de demostrar la viabilidad de superar el índice S&P 500 de forma consistente, aplicando rigurosamente ciencia de datos y análisis económico.

Palabras clave: Deep Learning, Machine Learning, Optimización de Carteras, Estrategia de Inversión

Abstract

Consistently outperforming the S&P 500 index is a complex challenge due to a variety of factors such as market efficiency and volatility, transaction costs, proper asset selection, and the right timing for buying and selling. In this work, we will attempt to address some of these difficulties to achieve an advanced investment strategy based on data analysis, starting with an extensive collection of data from open sources, such as U.S. financial market indicators, significant economic data, and technical features developed through data engineering. We will optimize the selection of critical variables using the most appropriate methods, seeking the most effective combinations to maximize the performance of our predictive models, selected from a set of machine learning and deep learning techniques. These tools will allow us to accurately predict the profitability of assets and choose the most promising prospects on a monthly basis. The strategy will be refined through portfolio optimization, based on solid economic principles, and validated in a backtesting environment, with the goal of demonstrating the feasibility of consistently outperforming the S&P 500 index by rigorously applying data science and economic analysis.

Keywords: Deep Learning, Machine Learning, Portfolio Optimization, Investment Strategy

Índice general

Índice de figuras

Índice de cuadros

1. Introducción

En el actual mercado financiero, superar de forma consistente el rendimiento del índice S&P 500 se ha convertido en un gran desafio, influenciado por factores tales como la eficiencia del mercado, la volatilidad, los costos de transacción y la selección y el timing adecuado de los activos. Ante esta complejidad, este trabajo se propone desarrollar una estrategia de inversión avanzada, iniciando con una meticulosa recopilación de datos de fuentes abiertas, incluyendo indicadores del mercado financiero estadounidense, datos económicos relevantes y características técnicas obtenidas a través de ingeniería de datos.

Con el objetivo de maximizar el desempeño de nuestros modelos predictivos, se optimizará la selección de variables críticas utilizando métodos apropiados para identificar las combinaciones más efectivas. Este análisis se apoyará en un conjunto de técnicas de machine learning y deep learning, en este sentido se han utilizado varias opciones destacando el uso de XGBoost , redes neuronales LSTM e incluso el uso de redes Convolucionales previo paso de convertir las series de tiempo en imagenes, estas herramientas se han vuelto esenciales para la predicción precisa de la rentabilidad de los activos y la selección mensual de los prospectos más prometedores.

Esta estrategia se refinará mediante la optimización de carteras, basándose en principios económicos robustos y su viabilidad se demostrará en un entorno de backtesting.

El objetivo principal es evidenciar que es posible superar el rendimiento del índice S&P 500 de manera consistente, a través de la aplicación rigurosa de la ciencia de datos y el análisis económico.

1.1. Contexto

El objetivo de superar al índice S&P 500 nos lleva a explorar el uso de técnicas de Machine Learning y Deep Learning en el desarrollo de una estrategia de inversión que se actualiza

mensualmente. Este enfoque se centra en la identificación y selección rigurosa de atributos relevantes, la aplicación de modelos predictivos eficaces y la mejora continua de la estrategia de inversión a través de la optimización de la cartera. Nuestro propósito es gestionar y dirigir este proceso analítico completo con el objetivo de alcanzar rendimientos que superen al mercado.

1.2. Motivación Personal

Mi motivación para emprender este trabajo radica en la intención de aplicar de manera concreta los conocimientos adquiridos durante la maestría en un contexto práctico real. Elegí el ámbito financiero, un terreno menos familiar para mí, buscando no solo desafiar mis propias limitaciones y ampliar mis competencias, sino también explorar nuevas oportunidades profesionales.

1.3. Objetivos

Objetivo Principal: Desarrollar una estrategia de inversión que, mediante el análisis avanzado de datos y la aplicación de Ciencia de Datos, busque superar consistentemente el rendimiento del índice S&P 500.

- **Objetivos Secundarios**:
  - Recolectar y analizar datos históricos de precios de ETFs sectoriales y datos económicos fundamentales.
  - Identificar las mejores combinaciones de atributos para los modelos de predicción.
  - Implementar modelos de Machine Learning y Deep Learning para la predicción del rendimiento.
  - Optimización de la cartera.

1.4. Sostenibilidad, diversidad y desafíos ético/sociales

Este es un proyecto muy técnico, sin embargo si queriamos mostrar nuestro compromiso con una visión integral y responsable de la inversión, subrayando la necesidad de avanzar hacia la inclusión de criterios ESG y contribuir a los ODS mediante la innovación en estrategias de inversión. Lamentablemente, la falta de fuentes de datos con criterios ESG me impidieron añadirlo como factor discriminante en la selección de activos.

Sostenibilidad: La selección de ETFs está basada en aspectos de rentabilidad y riesgo; sin embargo, nos gustaría hacer un llamado para promover el desarrollo de fuentes de datos ESG, resaltando la importancia de considerar el impacto ambiental en las estrategias

de inversión. En nuestro proyecto analizaremos fuentes de datos reducidas que generaran poco impacto y utilizaremos métodos para la selección de variables para mejorar el rendimiento de nuestros modelos, apoyando la consideración de ODS relacionados con la acción climática (ODS 13) y la producción y consumo responsables (ODS 12).

Comportamiento Ético y Responsabilidad Social: Aunque en este trabajo predomina el análisis técnico, no se descuidan las implicaciones éticas, tales como el respeto por las leyes de privacidad y propiedad intelectual y el impacto sobre el empleo. En nuestro caso, todas las fuentes de datos que utilizaremos son públicas y toda las referencias estan debidamente citadas reconociendo la autoria y propiedad intelectual, es por esto que estamos alineados con el ODS relacionado con el trabajo decente y el crecimiento económico (ODS 8).

Diversidad, Género y Derechos Humanos: Reconocemos en el ámbito financiero la necesidad de la inclusión y equidad, en nuestro trabajo no utilizamos datos personales de ningún tipo, esto no quita que apoyemos objetivos como la igualdad de género (ODS 5) y la reducción de desigualdades (ODS 10).

1.5. Enfoque y metodología

Seguiremos la siguiente metodología de trading:

1. **Definición y Preparación de Datos:** Comenzaremos con una recopilación detallada de datos de ETFs y variables económicas, seguida de un proceso de limpieza y preprocesamiento para asegurar la calidad de los datos.

2. **Análisis Exploratorio y Selección de Características:** Emplearemos diferentes técnicas para identificar variables predictivas clave y reducir la dimensionalidad de los datos.

3. **Desarrollo de Modelos Predictivos:** Probaremos una variedad de modelos de aprendizaje automático y aprendizaje profundo para pronosticar los rendimientos de los ETFs, con un enfoque en la precisión y la robustez de las predicciones.

4. **Optimización de Carteras:** Aplicaremos algoritmos de optimización para elegir la combinación óptima de ETFs, equilibrando el rendimiento esperado y el riesgo.

5. **Validación y Backtesting:** Evaluaremos los modelos y la estrategia de trading usando backtesting en datos históricos, realizando ajustes según sea necesario para afinar el rendimiento.

1.6. Planificación

Mostramos el diagrama de Gantt donde describimos todas las tareas necesarias para completar el Trabajo Final de Maestria, con los rombos indicamos las fechas de entrega de los diferentes trabajos que completan el semestre.

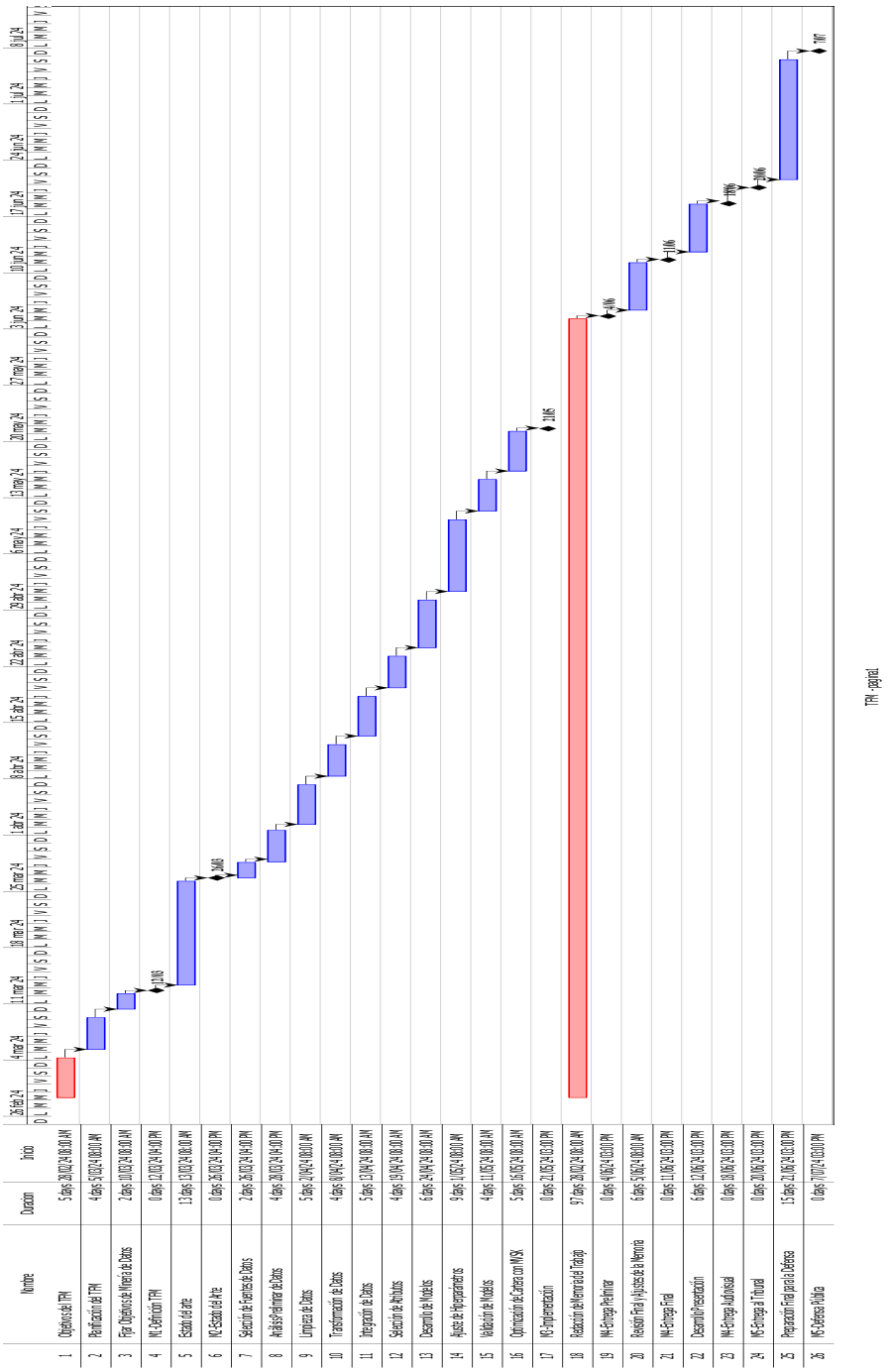

Figura 1: Diagrama Gantt del Trabajo Fin de Maestría

1.7. Resumen de los productos del proyecto

Los productos que pretendemos conseguir al final de este proyecto son los siguientes:

- **Modelos Predictivos:** Modelos de **machine learning** y **deep learning** ajustados para predecir la rentabilidad de los ETFs, fundamentales en la estrategia de selección de activos.
- **Selección de Características:** Un conjunto de técnicas para identificar las variables más influyentes en la rentabilidad de los activos.
- **Optimización de Carteras:** Un marco de trabajo basado en conceptos económicos para optimizar la cartera minimizando el riesgo y aumentando el retorno.
- **Sistema de Backtesting:** Un entorno para evaluar retrospectivamente la estrategia de inversión, asegurando su robustez y viabilidad.
- **Documentación Metodológica:** Una guía que detalla el proceso de investigación y desarrollo.

1.8. Breve descripción de los demás capítulos del informe

Este trabajo se organiza en capítulos que abordan distintas facetas del desafío de superar el rendimiento del índice S&P 500 a través de una estrategia de inversión basada en análisis avanzado de datos.

A continuación, se proporciona una breve descripción de cada uno:

Capítulo 3: Métodos y Recursos. Detallará las técnicas y herramientas utilizadas en la investigación. Se explicará la metodología para la recopilación y procesamiento de datos, los procesos utilizados para la selección de atributos, así como los enfoques de machine learning y deep learning empleados y finalmente, se examinarán las técnicas utilizadas para la optimización de la cartera.

Capítulo 4: Resultados. Presentará los hallazgos del estudio. Se analizará el desempeño de los modelos predictivos desarrollados y se evaluará la eficacia de la estrategia de inversión propuesta mediante backtesting.

Capítulo 5: Conclusiones y Trabajo Futuro. Resumirá los puntos clave de la investigación, discutirá las implicaciones de los resultados obtenidos y sugerirá direcciones para futuras investigaciones en este campo.

Capítulo 6: Glosario. Ofrecerá definiciones de términos técnicos y conceptos clave utilizados a lo largo del informe para facilitar su comprensión.

Anexos. En caso de ser necesario, se incluirá material complementario que respalde la investigación, como códigos de programación, tablas de datos detalladas y gráficos adicionales.

2. Estado del Arte

Gestión Clásica de Carteras

La gestión de carteras experimentó un cambio radical en 1952 con la publicación de la Teoría Moderna de Carteras (MPT por sus siglas en Ingles) por Harry Markowitz. Este enfoque innovador introdujo la cuantificación del riesgo, medido a través de la desviación estándar y su relación intrínseca con el retorno. Por primera vez, se establecieron principios matemáticos para la selección de activos no correlacionados, enfocados en la minimización del riesgo. La MPT destacó tres conceptos fundamentales: la importancia de la diversificación para reducir el riesgo, la dualidad entre riesgo y retorno y la utilización de la frontera eficiente para la selección óptima de carteras. Estos principios han guiado la investigación en inversión cuantitativa, enfocándose en explorar y expandir estos conceptos clave. Sin embargo, la MPT no está exenta de críticas y limitaciones, particularmente en lo que respecta a sus supuestos de normalidad de los retornos y dependencias a largo plazo. La complejidad de los mercados financieros y sus implicaciones para la gestión lo podemos ver en el estudio [1] donde los autores, profundizan en la naturaleza de los mercados financieros, sugiriendo que las propiedades de memoria larga y persistencia en los precios de los activos pueden desafiar los supuestos tradicionales de la MPT. Markowitz no solo cuantificó el riesgo, sino que también estableció las bases para distinguir entre el riesgo específico y el riesgo sistémico. La diversificación aparece como una estrategia esencial para mitigar el riesgo específico, permitiendo a los inversores minimizar el impacto de eventos adversos al invertir en una variedad de activos de distintos sectores. Esta práctica equilibra las pérdidas potenciales con ganancias en otras áreas de la cartera, estabilizando así el rendimiento general. Los beneficios de esta estrategia se han demostrado en múltiples ocasiones, como revela el análisis de Francesco Guidi y Mehmet Ugur [2], su estudio, que abarca desde septiembre de 2007 hasta junio de 2013, incluyendo la crisis financiera de 2008, muestra que las carteras diversificadas superan a las no diversificadas, independientemente de si se aplican estrategias de cartera igualmente ponderadas o de varianza mínima. Este hallazgo señala la efectividad de la diversificación no solo en términos de rentabilidad ajustada al riesgo, sino también en la rentabilidad media no ajustada. Harry Markowitz también fue pionero en abordar la dualidad entre riesgo y retorno, introduciendo un marco cuantitativo que permite a los inversores maximizar el retorno para un nivel de riesgo dado o minimizar el riesgo para un retorno esperado, a través de la construcción de la frontera eficiente. Esta relación fundamental ha sido el foco de

numerosas investigaciones posteriores, que han continuado explorando y validando la teoría de Markowitz en diversos contextos de mercado. Por ejemplo, podemos destacar el análisis comparativo realizado por Wang [3], donde se examinan y comparan modelos de gestión de carteras, incluyendo el Modelo de Media-Varianza de Markowitz. Este trabajo subraya la importancia de equilibrar riesgo y retorno, demostrando cómo diferentes enfoques pueden optimizar este equilibrio para lograr carteras eficientes.

Gestión Cuantitativa de Carteras

La consolidación y expansión de la gestión cuantitativa tras la introducción de la Teoría Moderna de Portafolios, ha sido marcada por avances significativos, como por ejemplo el Modelo de Valoración de Activos de Capital (CAPM) propuesto por William F. Sharpe en 1964 [4] y que se destaca como un hito fundamental, al introducir una medida sistemática del riesgo de mercado, beta y su relación con el retorno esperado. Este modelo proporcionó una base cuantitativa para evaluar el rendimiento ajustado al riesgo, influenciando profundamente la gestión de carteras y la valoración de activos. Mas adelante, investigadores como Fama y French en 1993 [5] ampliaron el entendimiento de los factores de riesgo en los rendimientos de las acciones, introduciendo un modelo multifactorial que complementa la visión de Markowitz. Este enfoque ha sido esencial en el desarrollo posterior de estrategias de inversión cuantitativas más sofisticadas. Posteriormente, la llegada de la era digital facilitó la implementación y evaluación de nuevas estrategias de inversión cuantitativas, gracias a la disponibilidad de grandes conjuntos de datos y avances tecnológicos, este entorno permitió a los gestores cuantitativos explorar y aplicar modelos predictivos y estrategias sistemáticas de inversión con gran eficiencia. Además, Estos avances transformaron la gestión de carteras, como se evidencia en el impacto del trading de alta frecuencia analizado por Brogaard [6], resaltando cómo estas innovaciones han redefinido las prácticas de inversión y su impacto en la calidad del mercado de valores.

Factores Alfa

La gestión cuantitativa de carteras, desde sus inicios con el MPT de Markowitz y con su posterior revisión de Sharpe en el modelo CAPM, han trazado un camino hacia una comprensión más profunda del rendimiento de los activos financieros en relación con su riesgo. Sharpe [4] introdujo el concepto de beta, estableciendo un vínculo entre el riesgo de mercado y los retornos esperados, una idea fundamental en el campo de las finanzas.

Sin embargo, fue Stephen A. Ross quien, en 1976 [7], desafió la perspectiva tradicional, argumentando que el rendimiento de los activos está influenciado por múltiples factores de riesgo, no solo por el riesgo de mercado. Este enfoque abrió nuevas vías para explorar y entender

los factores alfa, aquellos que explican los rendimientos excesivos ajustados al riesgo.

La investigación en este campo dio un salto significativo con el trabajo de Eugene F. Fama y Kenneth R. French en 1992 [5] con su modelo multifactorial, que incluía el tamaño de la empresa y el valor de mercado como factores adicionales, amplió la comprensión de cómo diversos aspectos, más allá del riesgo de mercado, pueden influir en los retornos de las acciones.

En años mas recientes, la contribución de Robert Novy-Marx [8] ha sido importante. Su estudio, destaca la rentabilidad bruta como un predictor robusto del rendimiento de las acciones, añadiendo una nueva dimensión a la selección de activos basada en factores alfa.

Esta evolución desde el CAPM hasta las teorías contemporáneas refleja una búsqueda constante por modelos más precisos y explicativos que abarquen la complejidad de los mercados financieros. La llegada del machine learning marca un nuevo capítulo en esta búsqueda, abriendo nuevas vías para la optimización de carteras y la gestión de riesgos, continuando la tradición de innovación en la gestión de inversiones.

La Irrupción del Machine Learning, Modelos

El desarrollo tecnológico ha desembocado de forma natural en la adopción del machine learning en la gestión de inversiones, apoyándose en una sólida base cuantitativa iniciada por el MPT de Markowitz y desarrollada con modelos teóricos como el CAPM, el análisis multifactorial de Fama y French y otras muchas investigaciones, quedo clara la importancia de factores adicionales al riesgo de mercado, preparando el terreno para un análisis más complejo y detallado.

La capacidad para analizar grandes volúmenes de datos estructurados y no estructurados y para descubrir patrones complejos en el comportamiento del mercado, ha propiciado la llegada del machine learning. Ya desde 1997, Lawrence [9] aplica redes neuronales para predecir los precios en el mercado de valores, superando a los métodos de la época. En el 2001, Cao y Tay [10] utilizaron SVM´s para predecir el índice de precios diarios S&P 500, demostrando que los SVM´s tenían mejor desempeño y eran menos costosos de entrenar.

La investigación avanzó con los estudios de Mao et al. [11] y Bollen, Mao y Zeng [12] en 2011, que demostraron cómo el análisis de datos de redes sociales puede predecir movimientos de mercado con una gran precisión, desafiando la Hipótesis del Mercado Eficiente y destacando el valor predictivo de los datos en tiempo real. Estos estudios abrieron nuevas vías para la gestión de inversiones, integrando el análisis de sentimientos y tendencias emergentes con los principios de la inversión cuantitativa.

En 2013, Takeuchi y Lee [13] utilizaron Deep Learning para capturar una versión mejorada del efecto momentum en las acciones, entregando un retorno anualizado significativamente superior durante el período de prueba. La selección de variables críticas utilizando algoritmos

inspirados en procesos naturales como genéticos o de medusas entre otros han sido explorados en muchas investigaciones, como en los trabajos de Huang [14] y Kuo [15], mejorando los rendimientos de inversión. Además, Hu, Ruilin, Luo y Tianyang [16] emplearon un modelo integrado XGBoost-LSTM para realizar predicciones precisas de manera eficiente, combinando la potente capacidad de selección de características de XGBoost con la habilidad predictiva de los modelos LSTM. Sezer y Murat [17] avanzaron aún más al utilizar Redes Neuronales Convolucionales, transformando las series de tiempo en imágenes para mejorar la predicción de movimientos del mercado.

Siguiendo esta trayectoria de innovación, el estudio de Ali Shavandi y Majid Khedmati [18] introduce un marco de aprendizaje profundo por refuerzo multi-agente para el comercio algorítmico en los mercados financieros. Este enfoque, que aprovecha la inteligencia colectiva de múltiples agentes especializados en diferentes marcos temporales, demuestra una mejora significativa en la identificación y explotación de oportunidades de mercado, superando a estrategias de negociación convencionales y adaptándose de manera eficaz a las dinámicas del mercado.

La integración del machine learning en la gestión de inversiones ha supuesto un cambio en las estrategias de inversión, aprovechando tanto datos estructurados tradicionales como fuentes de información no estructurada, ofreciendo nuevas opciones para la optimización de carteras y la gestión de riesgos en un entorno de mercado cada vez más complejo.

Estratégia Rotacional

La irrupción del machine learning en la gestión de inversiones ha supuesto un cambio en la forma en que se analizan y se toman decisiones en los mercados financieros. Desde las primeras aplicaciones de redes neuronales para predecir precios de acciones hasta el uso de modelos avanzados como XGBoost y LSTM para capturar dinámicas de mercado complejas, el machine learning ha demostrado ser una herramienta poderosa para mejorar las estrategias de inversión. Esta evolución tecnológica, fundamentada en modelos teóricos robustos como el CAPM y el análisis multifactorial de Fama y French y posteriores, ha preparado el terreno para un análisis más detallado y complejo de los mercados financieros, un ejemplo destacado de la aplicación de machine learning en estrategias de inversión es el estudio llevado a cabo por Praphutikul y Limpiyakorn [19], este estudio investiga la selección de acciones para trading rotacional utilizando el algoritmo de XGBoost, enfocándose en datos de Tailandia y utilizando veintisiete factores que abarcan categorías como valor, crecimiento, momentum, liquidez, calidad, dividendos y tamaño. La investigación revela que los factores técnicos tienen una influencia muy importante en los movimientos de precios para la selección de acciones mensuales, mientras que los factores fundamentales predominan en la tendencia de cambio de acciones en la selección trimestral. La estrategia propuesta en el estudio, que consiste en rotar posiciones entre diversas

acciones basándose en su puntuación relativa, demuestra la importancia del machine learning para optimizar las decisiones de inversión. Este enfoque no solo supera los índices de referencia en términos de retorno ajustado al riesgo, sino que también maneja eficazmente los costos asociados con el trading frecuente, destacando la relevancia de los factores técnicos y fundamentales en la selección de acciones a corto y largo plazo. La integración del machine learning en este tipo de estrategias representa un gran avance, ofreciendo una metodología sólida para la selección de acciones que puede superar consistentemente a los índices de referencia, este enfoque subraya la importancia de adaptar la estrategia al horizonte temporal de inversión y valida el enfoque de rotación basado en el aprendizaje automático como una estrategia viable y eficaz para el mercado de valores.

3. Contenido y Métodos

En este trabajo, he intentado desarrollar una estrategia de inversión rotacional basada en ETFs con el objetivo de superar al índice S&P 500 consistente en evaluar mensualmente una cartera de ETFs seleccionados entre los que presentan mejores oportunidades de rendimiento según la predicción de modelos de machine learning.

Para ello, recopile datos de precios de ETFs y les añadi datos económicos adicionales provenientes de FRED y algunas métricas técnicas, como osciladores y momentos, para capturar diversos aspectos del mercado y su comportamiento. Utilizamos técnicas de selección de características para identificar de forma dinámica las variables más influyentes para cada ETF y cada periodo temporal, evaluandolas mediante varios métodos estadísticos y aprendizaje automático.

Implementamos modelos de regresión, como RandomForest, XGBoost y LightGBM, para predecir los rendimientos de cada ETF, en cada periodo. Aqui deje abierta la posibilidad de calcular los mejores hiperparámetros para cada ETF en cada ciclo temporal o seleccionar los hiperparametros mas seleccionados para cada modelo.

Cada mes, seleccionamos los ETFs cuyas predicciones superaban al S&P 500 y eligiendo los tres mas prometedores para incluir en la cartera. Utilizamos técnicas de optimización para distribuir el capital de inversión entre los ETFs seleccionados maximizando el retorno ajustado al riesgo.

Realizamos un backtester, que nos permite realizar backtesting para las multiples combinaciones posibles que soporta nuestro código y de esa forma, seleccionar el mejor conjunto de técnicas posibles para cumplir nuestro principal objetivo, superar el índice S&P 500

Definición y Preparación de Datos

En este proyecto, debemos superar al S&P 500 mediante el Trading Quantitativo con las siguientes ETF´s

IYR: Sector Inmobiliario.

VOX: Sector Comunicaciones.

XLB: Sector Materiales.

XLE: Sector Energético.

XLF: Sector Financiero.

XLI: Sector Industrial.

XLK: Sector Tecnológico.

XLP: Sector Consumo Básico.

XLU: Sector Público.

XLV: Sector Sanitario.

XLY: Sector de Consumo Discrecional.

SPY: Índice Standard and Poors 500. Será nuestro benchmark.

Para conseguir este objteivo, Hemos utilizado varias fuentes de datos, entre los que incluimos datos de yfinance de yahoo, datos economicos de FRED y también hemos construido una serie de indicadores técnicos que nos permitiran seleccionar los mejores atributos para la toma de decisión de nuestros modelos.

Describiremos brevemente estos atributos.

3.0.1. Métricas Económicas

- **GDP (Producto Interno Bruto)**: Mide el valor total de todos los bienes y servicios producidos dentro de una economía durante un periodo específico, reflejando la salud económica general.
- **CPIAUCSL (Índice de Precios al Consumidor para Todos los Consumidores Urbanos)**: Representa el cambio promedio en los precios pagados por los consumidores urbanos por un mercado de bienes y servicios, siendo un indicador primario de la inflación.

- **FEDFUNDS (Tasa de Fondos Federales)**: Es la tasa de interés a la que las instituciones depositarias prestan fondos mantenidos en la Reserva Federal a otras instituciones de depósito de forma overnight.

- **UNRATE (Tasa de Desempleo)**: Muestra el porcentaje de la fuerza laboral que está desempleada y busca activamente trabajo, siendo un indicador clave de la salud laboral de la economía.

- **BOPGSTB (Balanza de Pagos - Bienes y Servicios)**: Indica la diferencia entre las exportaciones e importaciones de bienes y servicios de un país, reflejando la posición comercial del país en el mercado global.

- **PPIACO (Índice de Precios al Productor para Todos los Productos)**: Mide el cambio promedio a lo largo del tiempo en los precios de venta recibidos por los productores domésticos de bienes y servicios, indicando tendencias inflacionarias desde la perspectiva de la producción.

- **UMCSENT (Índice de Sentimiento del Consumidor de la Universidad de Michigan)**: Evalúa la confianza, las condiciones económicas y las expectativas de los consumidores, siendo un predictor de la disposición de los consumidores a gastar.

- **T10Y2Y (Diferencial de Tasas entre el Tesoro a 10 años y el Tesoro a 2 años)**: Mide la diferencia entre los rendimientos de los bonos del Tesoro a 10 años y a 2 años, utilizado como un indicador de las expectativas económicas futuras y la forma de la curva de rendimientos.

- **TB3MS (Tasa de los Bonos del Tesoro de EE.UU. a 3 meses)**: Refleja la tasa de rendimiento de los bonos del Tesoro a tres meses, proporcionando una visión de las expectativas de la política monetaria a corto plazo.

3.0.2. Indicadores Técnicos

- **Volatilidad Histórica**: Calculada como la desviación estándar de los retornos de un activo, escalada para reflejar una base anualizada, proporcionando una medida de la incertidumbre o el riesgo asociado con el precio del activo.

- **Skewness y Kurtosis**: Estas estadísticas describen la asimetría y la agudeza de la distribución de retornos de un activo, respectivamente, indicando la probabilidad de obtener resultados extremos.

- **Indicadores de Momento (RSI, MACD, PPO, ROC, ADX)**: Herramientas analíticas que evalúan la velocidad o fuerza del movimiento de precios de los activos para identificar condiciones de sobrecompra o sobreventa, tendencias emergentes y potenciales puntos de reversión.

- **Bandas de Bollinger**: Proporcionan niveles de precios relativos altos y bajos basados en desviaciones estándar de una media móvil simple del precio, útiles para identificar sobrecompra o sobreventa.

- **Oscilador de Precios**: Muestra la diferencia entre dos medias móviles, identificando la dirección del impulso y potenciales cambios en la tendencia del mercado.

- **Volumen de Balance (OBV)**: Utiliza el volumen de comercio para confirmar la tendencia de los precios, asumiendo que los cambios en el volumen pueden preceder a los cambios en el precio.

- **Índice de Distribución de Acumulación (ADL)**: Combina el precio y el volumen para determinar la presión de compra o venta detrás de movimientos de precios, ayudando a confirmar tendencias o advertir de reversiones.

- **Índice de Canal de Mercancías (CCI)**: Este índice compara el precio actual de un activo con su precio promedio en un periodo dado, normalizado por la desviación típica de ese precio promedio, para identificar ciclos de precios y condiciones extremas de sobrecompra o sobreventa.

Todos estas métricas son integradas en un único dataset que posteriormente pasamos a un selector de atributos, de donde obtenemos para cada ciclo mensual y cada ETF las mejores combinaciones de atributos para maximizar el rendimiento de nuestros modelos. El rango temporal desde el que recuperamos estos datos es desde 2010 hasta 2024, de forma que desde 2010 hasta 2020 los usamos para entrenar los modelos y a partir de 2020 y hasta 2024 los utilizamos para realizar las pruebas mediante backtesting, eso si, en cada ciclo añadimos el nuevo mes en training.

El rendimiento lo calculamos con la siguiente formula:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right) \tag{1}$$

donde r_t es el rendimiento logarítmico en el tiempo t, P_t es el precio de cierre en el tiempo t y P_{t-1} es el precio de cierre en el tiempo $t-1$.

Para la predicción retrasaremos los atributos dependientes un periodo temporal, para que el rendimiento futuro del periodo t, queden en el mismo nivel que las variables de t-1 y podamos predecir con los modelos de regresión.

Selección de Atributos

Para la selección de atributos, hemos comparado cuatro métodos diferentes, buscando el selector que diera mejores rendimientos en nuestros modelos, esta fase es muy importante, ya que realizo una selección de atributos por mes y ETF.

SHAP (SHapley Additive exPlanations)

SHAP es un método basado en la teoría de juegos que determina la importancia de cada característica en un modelo de machine learning. Este método asigna a cada característica un valor SHAP, que indica cuánto contribuye al cambio en la predicción en comparación con una predicción base.

Causalidad con causalml

Causalml es una biblioteca que implementa técnicas avanzadas de aprendizaje para estimar el impacto causal de las características en los resultados. Utilize modelos como LGBM y BaseSRegressor, esto nos ayuda a identificar qué características tienen un efecto causal sobre el resultado.

SelectKBest

SelectKBest es un método de selección de características que se basa en pruebas estadísticas para elegir los mejores k atributos. Es un enfoque de filtrado que evalúa la importancia de las características basándose en una función de puntuación específica y es efectivo para reducir la dimensionalidad del espacio de características a las más significativas para el modelo.

Estático

En Estático, he medido el impacto de utilizar las caracteristicas mas importantes obtenidos por estos métodos de forma fija o estática, sin calcularlos en cada ciclo mensual y por ETF, para comprobar que incluso así, el método dinámico da mejores resultados.

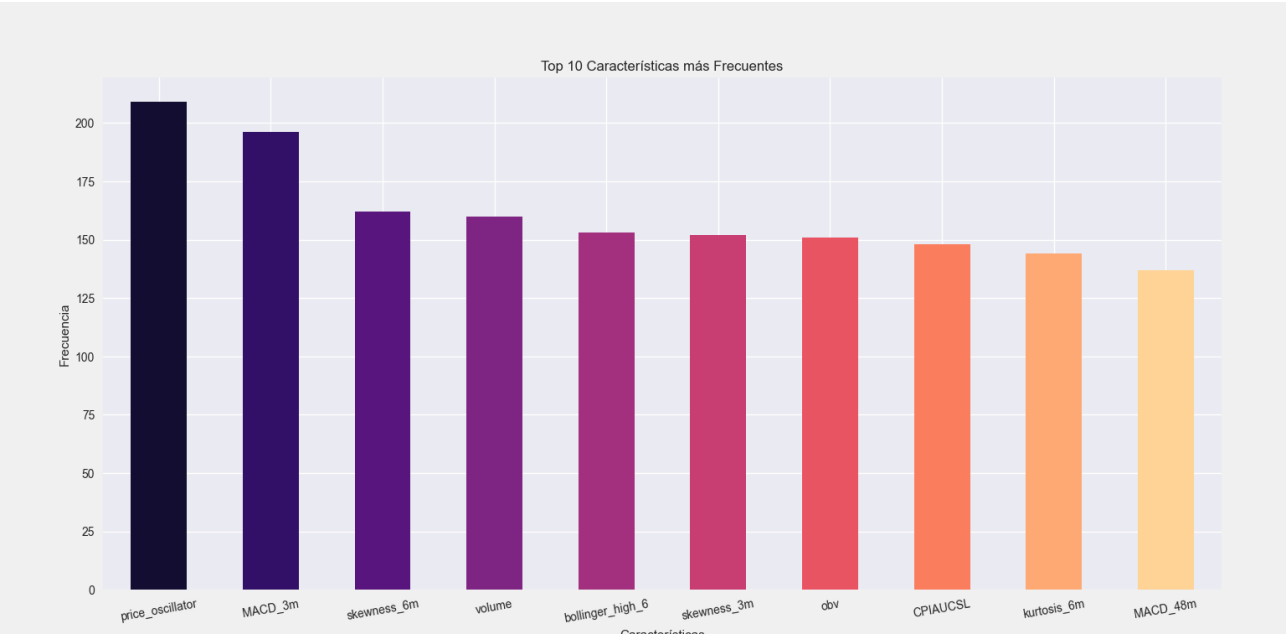

Figura 2: Mejores Atributos para Modelo Random Forest

Cuadro 1: Evaluación de Métodos de Selección de Atributos

ETF	causal	estatico	selectkbest	shap
IYR	0.05567645503982318	0.05243160112410935	0.05691392335257762	0.055347154300501605
SPY	0.05356803606741751	0.05130421668609276	0.05588508245042317	0.05156774472551313
VOX	0.06320073438923558	0.06041522659979954	0.05404554964397212	0.06106401537057868
XLB	0.0662007515455188	0.06432191892348238	0.0631054966904364	0.0662824412751322
XLE	0.1365449695560298	0.14359225864035116	0.1370615932804699	0.1391381152978652
XLF	0.07848694211271928	0.0786272670969303	0.08361590013530608	0.0760394757136519
XLI	0.07753850417265935	0.07097674091125207	0.07235951214274827	0.07349825268301656
XLK	0.06558489267919199	0.06610106894034061	0.06396578718736436	0.06416956735679322
XLP	0.03916848754969415	0.04251798458994643	0.037682008366703976	0.03878574463151345
XLU	0.04577074233866658	0.04945693885582369	0.046939046356177266	0.047173681691264194
XLV	0.04235552111421622	0.04599106089079986	0.04392251595345513	0.04205783877384148
XLY	0.06897207998213048	0.07101740223858023	0.06459569045713807	0.06755361117780674

Modelos Predictivos

Una vez que he creado mi dataset y he desarrollado un proceso para seleccionar los mejores atributos en cada ciclo y para cada ETF, me dispongo ha predecir el rendimiento de cada ETF para el ciclo siguiente, con el objetivo de utilizar este conocimiento para decidir que ETF mantengo en mi cartera y cuales vendo.

Para obtener el mejor rendimiento, podemos seleccionar los mejores hiperparametros con tres modelos diferentes, bien dinámicamente en cada ciclo mensual y para cada ETF o bien podemos estimar cuales son los mas apropiados y mantenerlos durante todo el proceso. son los siguientes:

GridSearchCV

El **GridSearchCV** es un método de búsqueda exhaustiva que prueba todas las combinaciones posibles de los hiperparámetros especificados en una cuadrícula. Este método garantiza encontrar la configuración de hiperparámetros que produce el mejor rendimiento del modelo, según un criterio de evaluación predefinido, generalmente a través de la validación cruzada. Aunque es muy preciso, puede ser computacionalmente costoso, especialmente cuando el espacio de hiperparámetros es grande.

RandomizedSearchCV

A diferencia de GridSearchCV, **RandomizedSearchCV** no prueba todas las combinaciones posibles, sino que selecciona al azar un número fijo de combinaciones de hiperparámetros para probar. Esto reduce significativamente el tiempo de cálculo necesario para la búsqueda, a costa de una posible pérdida de exhaustividad. Sin embargo, en muchos casos, RandomizedSearchCV puede alcanzar resultados muy cercanos a los óptimos con una fracción del tiempo de procesamiento.

Optuna

Optuna es un marco de optimización de hiperparámetros diseñado específicamente para automatizar la búsqueda de los mejores hiperparámetros. Optuna utiliza un enfoque basado en la historia de las pruebas anteriores para proponer candidatos potencialmente prometedores, optimizando así el proceso de búsqueda. Optuna es particularmente útil para espacios de búsqueda complejos y de alta dimensión.

Además, podemos elegir entre tres modelos, XGBRegressor, LGBMRegressor y RandomForestRegressor

XGBRegressor

El **XGBRegressor** es una implementación del algoritmo de *Gradient Boosting* desarrollado bajo el marco de XGBoost. Este modelo es conocido por su alta eficiencia y capacidad de manejar grandes conjuntos de datos con alta dimensionalidad. Utiliza árboles de decisión como base learners y optimiza tanto funciones de pérdida diferenciables como no diferenciables, lo

que lo hace extremadamente flexible. El modelo incorpora regularización ($L1$ y $L2$), lo que ayuda a prevenir el sobreajuste.

LGBMRegressor

El **LGBMRegressor** pertenece a la familia de LightGBM, una implementación de gradient boosting que se diferencia por su eficiencia en el uso de la memoria y velocidad de ejecución, gracias a su técnica de crecimiento de árbol basada en gradientes basados en histogramas. LightGBM divide los árboles por hojas en lugar de por niveles, lo cual permite obtener mejores resultados con menos tiempo de cálculo y menos recursos.

RandomForestRegressor

El **RandomForestRegressor** es un algoritmo de ensemble que opera construyendo una multitud de árboles de decisión durante el entrenamiento y produciendo la media de las predicciones de los árboles individuales para obtener un resultado más estable y robusto. El modelo es particularmente bueno para evitar el sobreajuste, gracias a su naturaleza de ensemble y el método de bagging utilizado para entrenar los árboles de forma independiente con diferentes subconjuntos de datos.

Similitudes y Diferencias

Los tres modelos son algoritmos de aprendizaje supervisado basados en árboles, lo que les permite capturar relaciones no lineales y complejas en los datos. Sin embargo, difieren en su enfoque de construcción y optimización de árboles. XGBoost y LightGBM utilizan técnicas de boosting, donde los modelos se construyen secuencialmente corrigiendo errores de los modelos anteriores, mientras que RandomForest utiliza bagging, donde cada modelo se construye de forma independiente y el resultado es un promedio de todos los modelos.

Cada uno de estos modelos, seran entrenados para cada ciclo y para cada ETF, añadiendo cada mes, los datos del mes concluido y que serviran para el entrenamiento de estos modelos en la siguiente predicción.

Optimización de Carteras

La optimización de nuestra cartera, la basamos en el ratio de Sortino que ajusta el retorno por el riesgo, lo cual es ideal para inversores que se preocupan más por las caídas que por la volatilidad general. A diferencia del ratio de Sharpe, que considera la volatilidad total, el ratio de Sortino solo tiene en cuenta la volatilidad negativa, es decir, las fluctuaciones que resultan en pérdidas.

La función criterio para la optimización basada en el ratio de Sortino se define como:

$$\text{Criterio} = -\left(\frac{\mu - r_{\text{TB3MS}}}{\sigma_{\text{neg}}}\right)$$

donde:

- μ es la media de los retornos de la cartera, representando el retorno esperado.
- r_{TB3MS} es la tasa mensual de los bonos del Tesoro de EE.UU. a 3 meses, utilizada como aproximación de la tasa libre de riesgo.
- σ_{neg} es la desviación estándar de los retornos negativos de la cartera, enfocándose solo en las caídas.

Utilizando Scipy, definimos la función criterio con este ratio y tratamos de maximizarlo.

4. Glosario

S&P 500 Índice que representa las 500 empresas públicas más grandes de Estados Unidos, considerado como uno de los mejores reflejos del mercado de acciones estadounidense.

Machine Learning Rama de la inteligencia artificial que se centra en el desarrollo de sistemas capaces de aprender y mejorar a partir de la experiencia, sin estar explícitamente programados para ello.

Deep Learning Subcampo del machine learning que utiliza redes neuronales profundas para modelar abstracciones complejas en datos, permitiendo el reconocimiento de patrones, la clasificación y la predicción con alta precisión.

Backtesting Técnica utilizada para evaluar la eficacia de una estrategia o modelo de trading mediante la aplicación de reglas de trading a datos históricos para determinar cuán efectiva hubiera sido.

XGBoost Algoritmo de aprendizaje supervisado que se utiliza para la construcción de modelos predictivos, conocido por su eficiencia, precisión y capacidad para manejar datos estructurados de gran tamaño.

LSTM (Long Short-Term Memory) Tipo de red neuronal recurrente especializada en aprender dependencias a largo plazo, ampliamente utilizada en el procesamiento del lenguaje natural y otras secuencias de datos temporales.

Redes Convolucionales Tipo de redes neuronales profundas que son especialmente potentes para tareas de procesamiento de imágenes, reconocimiento de patrones y clasificación, debido a su capacidad para capturar relaciones espaciales y temporales en los datos.

CAPM (Capital Asset Pricing Model) Modelo que describe la relación entre el riesgo sistémico y el rendimiento esperado de los activos, utilizado para evaluar la inversión bajo riesgo.

Teoría Moderna de Portafolios Marco conceptual propuesto por Harry Markowitz que utiliza la diversificación para optimizar la selección de la cartera, equilibrando el riesgo contra el retorno esperado.

Riesgo Sistémico Riesgo de colapso de todo un sistema financiero o mercado, no solo de componentes individuales, a menudo debido a interconexiones y dependencias.

Riesgo Específico Riesgo que afecta a una empresa o sector particular, también conocido como riesgo no sistémico, que puede mitigarse mediante la diversificación.

Algoritmos Genéticos Métodos de optimización y búsqueda basados en los principios de selección natural y genética para resolver problemas complejos mediante la evolución de soluciones.

Factores Alfa Factores utilizados en la gestión de inversiones para medir el rendimiento de una inversión contra un índice de referencia, buscando superar el mercado.

Estrategia Rotacional Estrategia de inversión que implica moverse entre sectores o activos en diferentes momentos para capitalizar sobre las condiciones cambiantes del mercado.

Trading Algorítmico modalidad de operación en mercados financieros que se caracteriza por el uso de algoritmos, reglas y procedimientos automatizados en diferentes grados, para ejecutar operaciones de compra o venta de instrumentos financieros.

Bibliografía

[1] Diego Luengas Domínguez, Esperanza Ardila Romero, and John Freddy Moreno Trujillo. Metodología e interpretación del coeficiente de hurst. *ODEON (Bogotá)*, (5):265–290, 2010.

[2] Francesco Guidi and Mehmet Ugur. An analysis of south-eastern european stock markets: Evidence on cointegration and portfolio diversification benefits. *Journal of International Financial Markets, Institutions and Money*, 30:119–136, 2014.

[3] Yuxuan Wang. Comparative analysis and research of investment portfolio management models. *Advances in Economics, Management and Political Sciences*, 63:95–100, 2023. © 2023 The Author(s). Published by EWA Publishing. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

[4] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, September 1964. First published: September 1964.

[5] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[6] Jonathan A. Brogaard. High frequency trading and its impact on market quality. Northwestern University, Kellogg School of Management, Northwestern University School of Law, JD-PhD Candidate, 2010. First Draft: July 16, 2010. November 22, 2010.

[7] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

[8] Robert Novy-Marx. The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1):1–28, 2013.

[9] Ramon Lawrence. Using neural networks to forecast stock market prices, December 1997.

[10] Lijuan Cao and Francis E.H. Tay. Financial forecasting using support vector machines. *Neural Computing and Applications*, 10:184–192, 2001.

[11] Huina Mao, Scott Counts, and Johan Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. 2011.

[12] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[13] Lawrence Takeuchi and Yu-Ying (Albert) Lee. Applying deep learning to enhance momentum trading strategies in stocks, December 2013.

[14] Chien-Feng Huang. A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, 12(2):807–818, 2012.

[15] R.J. Kuo and Tzu-Hsuan Chiu. Hybrid of jellyfish and particle swarm optimization algorithm-based support vector machine for stock market trend prediction. *Applied Soft Computing*, 154:111394, 2024.

[16] Ruilin Hu and Tianyang Luo. Xgboost-lstm for feature selection and predictions for the s&p 500 financial sector. *Advances in Economics, Management and Political Sciences*, 59:249–257, 01 2024.

[17] Omer Berat Sezer and Ahmet Murat Ozbayoglu. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing*, 70:525–538, 2018.

[18] Ali Shavandi and Majid Khedmati. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications*, 208:118124, 2022.

[19] Thanadon Praphutikul and Yachai Limpiyakorn. Xgboost-based multi-factor stock selection model for rotational trading. In *Proceedings of the 2023 5th International Conference on Information Technology and Computer Communications (ITCC '23)*, pages 107–115, New York, NY, USA, June 2023. ACM.