Universitat Oberta de Catalunya (UOC)

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

TRABAJO FINAL DE MÁSTER

Área: Data analysis y big data

Uso de algoritmos data driven para superar al S&P500

Autor: Sergi Garcia Marsol

Tutor: David García Agudiez

Profesor: Albert Solé Ribalta

Barcelona, 12 de julio de 2024

Ficha del trabajo final

Título del trabajo:	Uso de algoritmos data driven para superar el S&P500
Nombre del autor:	Sergi Garcia Marsol
Nombre del colaborador docente:	David García Agudiez
Nombre del PRA:	Albert Solé Ribalta
Fecha de entrega (mm/aaaa):	07/2024
Titulación o programa:	Máster en ciencia de datos
Área del Trabajo Final:	Data analysis y big data
Idioma del trabajo:	Español
Palabras clave	Machine Learning, Big Data, Investment Strategy

Dedicatoria

A Marta que tiene la paciencia que a mí me falta.

A Marc i Alèxia que llenan cada día el espacio que no existía.

Resumen

Ante la automatización de gran parte de las operaciones bursátiles actuales, el presente trabajo final de máster propone la fusión de datos de significancia para los mercados financieros con el objetivo de desarrollar algoritmos que propongan rendimientos superiores al índice de referencia americano S&P500. En este sentido, la meta será superar el benchmarking mediante una estrategia de inversión rotacional que utilizará vehículos de inversión sectoriales.

Este proyecto va más allá de una mera exploración de datos financieros, siendo una inmersión profunda en la sinergia entre diversas fuentes de información. La integración de datos macroeconómicos procura entender las complejidades del entorno global, mientras que la consideración de factores técnicos busca identificar patrones y tendencias cruciales en el mercado. Además, la inclusión de factores Fama-French agrega un nivel de sofisticación al modelado, capturando características específicas de las acciones.

Palabras clave: Big Data, Investment Strategy, Data Driven.

Abstract

Given the automation of a large part of the current stock market operations, this final master's thesis proposes the fusión of data of significance for the financial markets with the objective of developing algorithms that offer superior returns to the American reference index S&P500. In this sense, the goal will be to overcome the benchmarking through a rotational investment strategy that will use sectoral investment vehicles.

This project goes beyond a mere exploration of financial data, being a deep dive into the synergy between various sources of information. The integration of macroeconomic data seeks to understand the complexities of the global environment, while the consideration of technical factors seeks to identify crucial patterns and trends in the market. In addition, the inclusion of Fama-French factors adds a level of sophistication to the model, capturing specific characteristics of the actions.

KeyWords: Big Data, Investment Strategy, Data Driven.

Índice general

Índice de figuras

Índice de cuadros

1. Introducción

1.1. Contexto y motivación

La gestión de carteras consiste en tomar decisiones para llevar a cabo estrategias de inversión eficientes mediante un conjunto de activos financieros con el objetivo de maximizar sus rendimientos. Estos activos pueden ser muy variados: derivados, acciones, bonos, efectivo u otros instrumentos financieros. Mientras que las megatendencias se basan en tomar decisiones a partir de las grandes "trends" económicas y sociales a nivel global, como, por ejemplo, el cambio climático, la longevidad, la tecnología, la digitalización o la robótica, el "factor investing"se basa en identificar y utilizar factores específicos que históricamente han demostrado tener un impacto significativo en el rendimiento de los activos financieros. Esta estrategia permite fundamentar las inversiones en parámetros objetivos donde la opinión del gestor queda minimizada. Aquí es donde el Big Data, gracias a la captura sustancial de un mayor número de datos e identificación de nuevas variables, puede tener un rol fundamental a la hora mejorar la toma de decisiones.

En la actualidad, el crecimiento exponencial de la generación de datos ha alcanzado cifras extraordinarias, con estimaciones que sugieren la creación de 329 millones de terabytes diarios [1]. Este fenómeno se intensifica con el tiempo, con aproximadamente el 90 % de todos los datos almacenados por la humanidad generados en los últimos dos años. Este volumen masivo de información no solo ha transformado la forma en que entendemos y manejamos los datos, sino que también ha influido de manera significativa en diversas áreas, incluida las finanzas y específicamente el campo que nos ocupa en este estudio, la inversión financiera. En este contexto, se observa un cambio importante hacia la gestión pasiva, que actualmente representa aproximadamente el 60 % del mercado de valores. A su vez, los fondos cuantitativos, basados en estrategias algorítmicas, han ganado una cuota de mercado del 20 %. En conjunto, estos datos sugieren que alrededor del 80 % de las transacciones bursátiles son el resultado de ejecuciones automatizadas y decisiones alejadas de intervenciones humanas directas [2]. Este panorama destaca la creciente influencia de los enfoques data-driven en la toma de decisiones financieras.

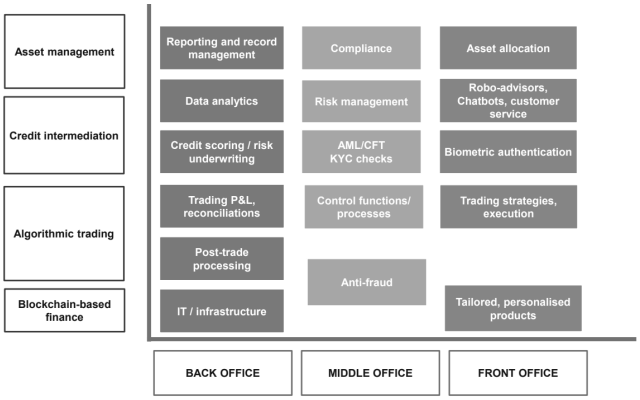

Figura 1: Ejemplos de aplicación de la inteligencia artificial en mercados financieros [3]

En conclusión, los cambios derivados de la automatización y la evolución tecnología propicia estudios acordes a las herramientas que estas nos proporcionan.

Por lo que al autor respecta, la motivación del presente trabajo fue desde el inicio del máster en ciencia de datos una temática que quería abordar. La experiencia profesional y laboral en mercados financieros en varias firmas del sector bancario y por extensión, las ganas de establecer y mezclar dos ramas de conocimiento como son la ciencia de datos y la inversión se materializan en este proyecto cuyos objetivos mostramos en el siguiente apartado.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general del trabajo es el estudio y implantación de modelos de machine learning para obtener resultados de inversión superiores al índice de referencia bursátil americano Standard&Poors 500 (S&P500)

1.2.2. Objetivos específicos

Los objetivos específicos del proyecto son:

- Estudiar la influencia de factores macroeconómicos sobre en el rendimiento del mercado de valores.
- Aplicar el modelado de Fama French como factor predictivo en la estrategia de inversión.
- Elaborar un código de programación ágil con lenguaje Python para la aplicación de algoritmos con el fin de generar una estrategia de inversión rotacional sectorial.

- Usar librerías API Open Access para la importación de los datos necesarios para el estudio.
- Valorar el mejor modelo de inversión en base a cálculos de ratios complementarios.

1.3. Sostenibilidad, diversidad y desafíos ético/sociales

La competencia de compromiso ético y global (CCEG) se define de manera genérica en actuar de manera honesta, ética, sostenible, socialmente responsable y con respeto hacia los derechos humanos y la diversidad, tanto en la práctica académica como profesional. Esta, presenta tres dimensiones destacadas: la dimensión de sostenibilidad, la dimensión del comportamiento ético y responsabilidad social y la dimensión de diversidad, género y derechos humanos.

En lo que respecta a los objetivos de desarrollo sostenible (ODS) comprenden 17 metas a nivel mundial establecidas por las Naciones Unidas con el propósito de enfrentar desafíos como la pobreza, el hambre, la salud, entre otras, persiguiendo la meta de un desarrollo sostenible antes de 2030.

Descritos ambos bloques se debe abordar la aplicación de los mismos en el presente trabajo:

Dimensión de sostenibilidad

En la planificación de este trabajo final de máster se ha considerado su impacto en términos de sostenibilidad, tomando en cuenta los principios delineados en los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas. El enfoque del proyecto se concreta específicamente con el ODS 9 (Industria, innovación e infraestructura), ya que se centra en la exploración de modelos de inversión innovadores y utiliza datos para optimizar la eficiencia de gestión de carteras en el sector financiero.

La gestión de carteras basada en una estrategia rotacional sectorial de ETF que deriva de la predicción de modelos de Machine Learning incluye una serie de medidas arraigadas a la dimensión de sostenibilidad. Estas técnicas permiten analizar grandes cantidades de datos que de manera manual sería mucho más tediosa, es decir, requeriría muchas más unidades de tiempo y trabajo. De esta manera, se contribuye a la reducción del consumo de recursos y gastos asociados. Otra diferencia derivada de estos modelos es la reducción de errores humanos tanto en la ejecución como en la toma de decisiones cosa que propone menores desperdicios de recursos o impactos negativos.

Dimensión de comportamiento ético y responsabilidad social

La presente investigación considera la dimensión ético-social al analizar enfoques de inversión desde una perspectiva ética y responsable. El objetivo es asegurar que las propuestas de solución no promuevan conductas poco éticas ni generen impactos negativos en aspectos como el tratamiento de datos y/o privacidad. En este caso, la investigación respeta los principios éticos de la profesión financiera contribuyendo al logro del Objetivo de Desarrollo Sostenible 8 (Trabajo Decente y Crecimiento Económico).

Cabe destacar la creciente incorporación de los criterios ESG (Enviromental, Social, Governance) en los procesos de inversión. Según CFA Institute [4] entidades como el Sustainability Accounting Standards Board (SASB), la Global Reporting Initiative (GRI) y el Task Force on Climate-related Financial Disclosures (TCFD) están trabajando para establecer estándares y definir la materia para facilitar la integración de estos factores en el proceso de inversión.

En otro sentido, este trabajo utiliza datos sonsacados de librerías públicas. Asimismo, se utilizarán datos financieros sin información personal o privada.

Dimensión de diversidad, género y derechos humanos

En lo que respecta a la diversidad, género y derechos humanos, se ha observado que este proyecto, al enfocarse en modelos de inversión y el análisis de datos financieros, impacta de manera más técnica pero menos inmediata. No obstante, se asegurará de que el diseño de la investigación no refuerce prejuicios ni estereotipos, promoviendo una perspectiva inclusiva y evitando cualquier forma de discriminación. Se presta atención a la perspectiva de género en el lenguaje utilizado y se mantiene una sensibilidad hacia la diversidad en todas las fases del proyecto, a pesar de que la naturaleza técnica del mismo no influya directamente en estas áreas.

1.4. Enfoque y método seguido

1.4.1. Productos obtenidos

El trabajo final de máster pretende generar un producto adaptado a estudios anteriores que aporte un modelo de cartera rotacional sectorialmente mediante el uso de algoritmos vehiculados a través de Exchanged Traded Funds con el fin de superar al índice de referencia S&P500.

En resumen, los productos obtenidos en este proyecto serán:

1. Repositorio Github con el código de implementación del modelo (desde la extracción de factores hasta la comparativa final de la cartera contra benchmark)
2. La presente memoria

1.4.2. Metodología

El estudio del caso seguirá las siguientes etapas:

Recolección de datos

Debemos identificar i recopilar los indicadores macroeconómicos y datos técnicos relevantes. Esto puede incluir indicadores económicos cómo crecimiento del PIB, inflación, tasas de interés entre otros indicadores técnicos cómo medias móviles, RSI, etc.

Preprocesamiento de datos

La etapa de preprocesamiento de los datos implica la limpieza y procesamiento de los datos recopilados. Algunos ejemplos son la gestión de datos nulos, valores atípicos para así asegurar la consistencia e integridad de los datos.

Ingeniería de Características

Una vez superadas las etapas anteriores aplica la creación de características adicionales o transformaciones que puedan mejorar el poder predictivo del modelo.

Selección de modelos

Cobra especial importancia la elección del modelo de aprendizaje. Los modelos comunes para la predicción de series temporales incluyen modelos de regresión, árboles de decisión, clasificación u otros.

Entrenamiento del Modelo

Seleccionado el modelo realizaremos el entrenamiento mediante las librerías correspondientes utilizando el conjunto de datos destinado a ello. Se deben ajustar los hiperparámetros según

sea necesario y considera el uso de técnicas como la validación cruzada para garantizar la robustez.

Evaluación y backtesting

La evaluación del rendimiento vendrá de la mano de métricas comunes que incluyen precisión, precisión, recuperación y puntuación F1. Además, la evalúa se combinará con el estudio de backtesting versus el benchmark (S&P500).

Optimización

En caso que fuera necesario, se optimiza el modelo y la estrategia según la etapa anterior.

1.4.3. Planificación

Debemos organizar el trabajo de manera acorde a la presentación del trabajo bajo la temporización de la materia por parte de la Universidad. Las pruebas de evaluación continua proponen entregas parciales que serán revisadas por el tutor. Por otro lado, el desarrollo e implementación del modelo deberá ejecutarse de manera paralela. El gráfico de gantt muestra visualmente la generalidad metodológica y de planificación.

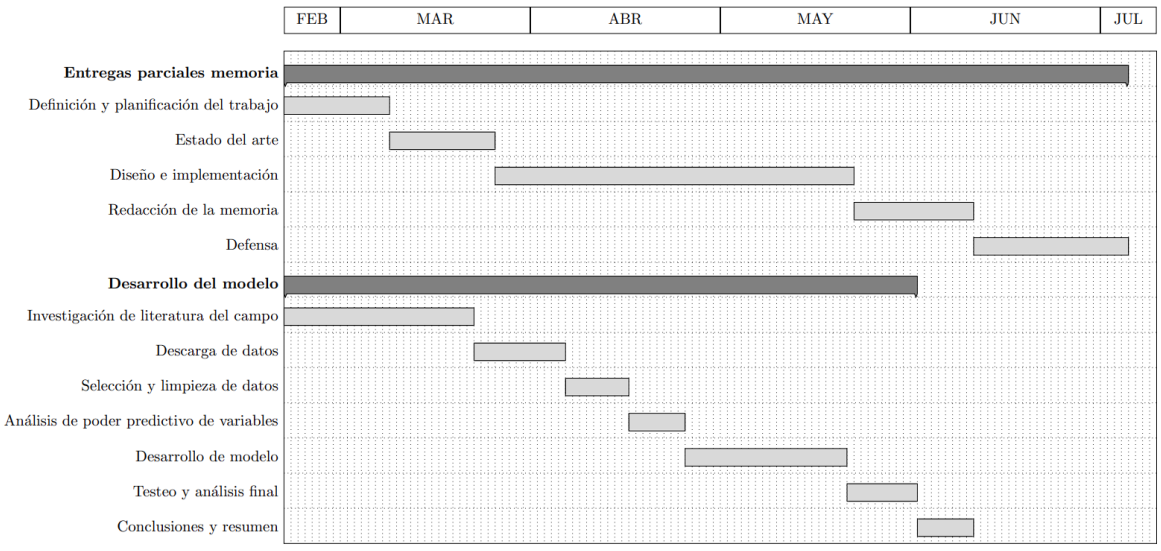

Figura 2: Planificación del trabajo final

La distribución de las entregas parciales del trabajo responden a la siguiente temporización:

PEC1	**PEC2**	**PEC3**	**PEC4**
Resumen	Estado del arte	Métodos	Resultados
Justificación		Recursos	Conclusiones
Objetivos		Desarrollo	Memoria
Metodología			
Planificación			

Cuadro 1: Distribución de las tareas por entregas

La tabla anterior proporciona una visión general de las tareas asociadas con cada entrega, clasificadas por categorías. Cada entrega se asocia con un conjunto específico de actividades o secciones que se espera abordar empezando por la primera actividad de introducción y planteamiento del proyecto hasta la entrega final de cierre anterior a la defensa del mismo.

2. Estado del arte

La investigación del trabajo académico desempeña una función esencial en cualquier estudio. El objetivo del apartado radica en proporcionar una perspectiva general sobre cómo diversos investigadores han abordado la problemática que el presente trabajo de investigación pretende tratar. Podríamos definirlo como una actualización contextualizada del área de conocimiento que aplica.

2.1. Machine Learning

La realidad del sector financiero pasa por una evolución drástica producida en las últimas décadas basadas fundamentalmente en el avance tecnológico [5]. Desde las primeras aportaciones de Markowitz con su teoría moderna de carteras [6] dónde proponía una visión más innovadora basada en su comprensión de que el rendimiento de una acción individual no era tan importante como el rendimiento y la composición de toda la cartera del inversor hasta la actualidad, han aparecido elementos de cambio que nos sitúan en un presente mucho más complejo. En este sentido, algunos de los factores clave que han propiciado la incursión del Machine Learning al estado actual [5] son:

- Cambios en la estructura de mercados que generan oportunidades de arbitraje entre precios por la integración de más tipologías de activos y mayor número de geografías implicadas.

- Mejora de los métodos estadísticos, gestión de datos y potencia de las computadoras para el tratamiento de estos.

- Desarrollo de nuevas estrategias de inversión basadas en factores de riesgo y no en clase de activos.

- Mayores rendimientos del algorithmic trading versus rendimiento de gestor humano.

La gestión activa de fondos de inversión ha estado vinculada a dos vertientes: por un lado, la inversión sistemática o quant y por otro, la inversión discrecional. Ambas metodologías han ido convergiendo en el tiempo pues el aumento de la información ha permitido eficiencias de data driven, es decir, el uso de datos para la toma de decisiones.

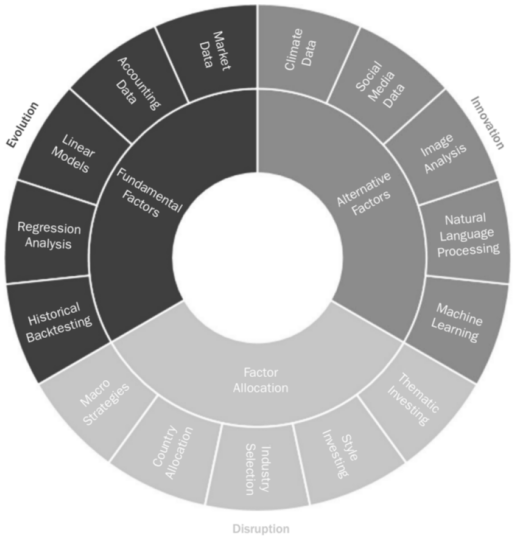

Figura 3: Evolución, innovación y disrupción, presente y futuro de la inversión factorial [7]

En resumen, la figura anterior muestra una posible clasificación de las aplicaciones de la data driven en base al momento temporal. Iniciada por un proceso de disrupción que se basaba en el uso de factor allocation, posteriormente apareció el movimiento de evolución que aporto las herramientas tecnológicas a la gestión de carteras para finalmente establecerse en un contexto donde los factores alternativos como el Machine Learning, el proceso de lenguaje natural, el análisis de sentimiento en redes entre otros hacen que la especialización permita mayores eficiencias y tomas de decisión de los gestores.

Cabe señalar a autores de presente destacados como Marcos Lopez de Prado quién desempeña un papel muy relevante a la vez que reconocido por la comunidad. Sus aportaciones más recientes de su última obra [8] comentan como el futuro del Machine Learning deberá dominar las decisiones financieras, reduciendo la especulación y aumentando la precisión en las inversiones. Por otra parte, propone varios desafíos importantes como poner solución a uno de los errores críticos en la gestión de inversiones, especialmente en el ámbito cuantitativo, dónde identifica la falta de un enfoque estructurado y colaborativo en la toma de decisiones. Asimismo, su obra hace referencia a una serie de investigadores como Krauss et al. [9] o Raffinot [10] que permitirán mediante sus estudios y trabajos un enfoque más significativo al proyecto.

Finalmente, los estudios recientes sobre Machine Learning en el campo de la inversión demuestran un creciente interés en la aplicación de algoritmos avanzados para la mejora de precisión en la toma de decisiones. En este sentido también podemos señalar el "metalabeling" como proceso de etiquetar las predicciones de un modelo de Machine Learning con información

adicional sobre la confianza o la calidad de esas predicciones. Su aplicación permite una gestión más sofisticada de las estrategias de inversión, ya que no solo se tienen en cuenta las predicciones del modelo, sino también la fiabilidad percibida de esas predicciones. En conclusión, proporciona un valor de probabilidad sobre la certeza de la predicción.

2.2. Fama-French: Five Factor Model

Es indudable el impacto reciente del modelo Fama French en la gestión de carteras y la valoración de activos. En este sentido y puesto que lo utilizaremos como variable de predicción parece interesante dar una visión más extensa sobre su aplicación.

El modelo Fama-French [11], nació de la mano de los economistas Eugene Fama y Kenneth French cómo extensión del modelo de valoración de activos financieros [12] que pretendía explicar rendimientos de activos financieros mediante el estudio de factores de riesgo adicionales. Concretamente el modelo Fama-French Three Factor calcula la tasa de rentabilidad de una inversión en función del riesgo global del mercado, el grado en que las pequeñas empresas superan a las grandes y el grado en que las empresas de alto valor superan a las de bajo valor. Sobre esta base los mismos autores incluyeron posteriormente dos nuevos factores en el modelo Fama-French Five Factor [13]: la comparativa de rendimientos de empresas con rentabilidad operativa alta y aquellas con rentabilidad operativa baja y por último, el factor conservador menos agresivo, que mide la diferencia entre empresas que invierten de forma agresiva y aquellas que lo hacen de forma más conservadora.

$$r = \beta_m(Rmkt - Rf) + \beta_{smb}(smb) + \beta_{hml}(hml) + \beta_{rmw}(Rmw) + \beta_{cma}(cma) + \varepsilon$$

- r = retorno de mercado esperado
- Rf = Risk-free rate
- ß = Coeficientes de sensibilidad
- (Rmkt – Rf) = Prima de riesgo de mercado
- smb = Exceso de retorno de compañías de baja capitalización menos compañías de larga alta capitalización
- hml = Exceso de retorno de empresas de valor (alto price to book ratio) menos empresas de crecimiento (bajo price to book ratio)

- rmw = Exceso de retorno entre compañías de alta rentabilidad operativa menos compañías de baja rentabilidad operativa

- cma = Exceso de retorno entre compañías con inversión conservadora menos compañías con inversión agresiva

- ε = Riesgo

2.3. Estudios relacionados

Referente al presente trabajo, basará su estrategia en una inversión rotacional sectorial determinada por los factores predictivos que mediante ETFs propondrán la compra de uno u otro con el fin de superar el benchmark de referencia S&P500.

Para ello, debemos definir los Exchanged Traded Funds (ETF) como fondos de inversión que se negocian en bolsa y combinan características de los fondos mutuos y las acciones individuales. Los ETFs tienen como objetivo replicar el rendimiento de un índice subyacente, sector del mercado, materia prima u otro activo financiero [14]. Las características más relevantes de los ETF son los bajos costes de gestión, la liquidez, diversificación y la casi nula intervención del factor humano.

Con todo ello, varios autores profundizan en la importancia de los ETFs como vehículos de inversión eficientes [15].

El estudio de nuestra estrategia rotacional de inversión en ETFs sectoriales implica el libro de referencia Machine Learning for Algorithmic Trading [5] de Stefan Jansen que será utilizado como manual de aplicación sobre las técnicas de validación y procesos. El autor mediante su obra acerca el uso de algoritmos de Machine Learning en su aplicación al campo de inversión. La obra cubre una amplia gama de técnicas de Machine Learning, desde la regresión lineal hasta el aprendizaje profundo por refuerzo, y ofrece código base para implementación de estrategias de predicción con modelos. De la misma forma propone un ese mismo código en espacios virtuales.

Liew y Mayster (2018) [16] publican su artículo "Forecasting ETFs with Machine Learning Algorithms"dónde aplican técnicas de Machine Learning centrándose en la dirección del movimiento de varios ETFs. Como conclusión encuentran que los algoritmos de deep neural networks, random forests y support vector machines presentan baja previsibilidad a corto plazo. También destacan la importancia del volumen alto de factores predictivos como conjunto de información por su poca contribución individual.

Otro ejemplo como el publicado por Bartram, S. M. et al (2021) [17] quienes examinan el uso del Machine Learning destacando el potencial del aprendizaje por refuerzo y señalando que el desempeño de los fondos cotizados en bolsa que utilizan Machine Learning tiende a ser mixto.

Finalmente, algunos trabajos que pueden orientar la base del presente trabajo se sitúan en el repositorio Github de Ken Chiang [18] o el artículo de Pinsky y Yang [19].

3. Metodologia

En esta sección se presenta la metodología utilizada para desarrollar el proyecto descrito. Por limitaciones de tiempo el trabajo no presenta tantos modelos como inicialmente se planteó. En un primer estadio, con un código sencillo se implementaban varios modelos de regresión lineal, clasificación y árboles de decisión. A medida que se profundizaba en ellos se lucia la necesidad de limitarse en algunos en concreto. Así pues, finalmente se decidió presentar los modelos de regresión lineal, random forest y light gradient boosting.

En referencia al modelo de regresión lineal, es uno de los métodos más simples y utilizados en aprendizaje automático. Basa su aprendizaje en la suposición que existe relación lineal entre las variables independientes y dependientes. Trata de encontrar la mejor respuesta minimizando la suma de los cuadrados de las diferencias entre los valores observados y los predichos por el modelo. Como principal inconveniente, el modelo puede limitar su rendimiento ante la multicolinealidad de las variables independientes.

Respecto al segundo modelo, se trata de un algoritmo de aprendizaje automático que utiliza múltiples árboles de decisión para obtener predicciones. Su principal ventaja se centra en la capacidad de gestión de grandes conjuntos de datos evitando el sobreajuste en base a la aleatoriedad que presenta seleccionando características que impiden la correlación entre sus múltiples árboles. Por otra parte, la aparición de más árboles se traduce en una mayor complejidad del modelo perdiendo interpretabilidad.

Finalmente, un modelo de aprendizaje supervisado ofrece algunas ventajas, principalmente que el conjunto de entrenamiento se modifica cada vez en función de los errores acumulativos cometidos por el modelo hasta el momento. En otras palabras, procede secuencialmente utilizando versiones ponderadas de los datos. Tal como se confirma en el estudio, estos modelos tienden a tener un rendimiento ligeramente superior al de la regresión lineal y bosques aleatorios. Por último, comentar que LightGBM, el modelo utilizado, fue un desarrollo de código abierto de la compañia Microsoft.

3.1. Obtención de datos

Para realizar el presente trabajo era necesario obtener información respecto a los distintos activos con los que se iba a desarrollar la estrategia de rotación sectorial. En ese sentido, era importante la descarga de todos los precios históricos de los vehículos de inversión señalados. Aunque el código base de Stephan Jansen presente una posible descarga de los mismos a través del portal stooq, nuestra descarga se ha realizado mediante la API de Yahoo Finance. Esta se descarga mediante una función que almacena los tickers solicitados gracias a la capacidad de almacenamiento integrada en el formato .ipynb. Los datos proporcionados por la API yfinance permiten la descarga para uso personal de múltiples factores como valor de las acciones, dividendos, información corporativa entre otros. Así pues, la posibilidad de extracción de información se presenta suficiente para nuestro cometido.

Se realizo la extracción de histórico de precios a cierre de los tickers a continuación mencionados entre los años 2000 y 2024 con los siguientes campos:

- Date: Fecha diaria de referencia al precio del ETF
- Close: Precio de cierre al que cotiza el ETF
- Ticker: Dato alfabético correspondiente a la identificación del ETF

Por otra parte los tickers correspondientes a los vehículos de inversión de gestión pasiva seleccionados son los siguientes:

Ticker	Sector
XLE	Energía
XLB	Materiales
XLI	Industriales
XLK	Tecnología
XLF	Financiero
XLP	Consumo básico
XLY	Consumo discrecional
XLV	Salud
XLU	Servicios públicos
^GSPC	Índice S&P500

Una vez incorporado al documento de trabajo los precios históricos de los ETF se manipulo el dataset para adaptarlo a precios mensuales, es decir, eliminando los precios intermedios del mes dejando así sólo los precios de inicio de mes para su posterior tratamiento.

Realizados los pasos anteriores, calculamos el rendimiento intermensual de los activos en base a la diferencia de precios del mes anterior con el posterior.

$$\text{Monthly performance} = \left(\frac{Price_{n+1} - Price_n}{Price_n} \right) \times 100$$

3.2. Feature engineering

En lo que respeta a las variables independientes, es decir, aquellas que servirán para facilitar las predicciones de los precios, podemos dividirlas en dos secciones distintas. Por una banda aparece los datos macroeconómicos y por otra los datos técnicos de los activos.

3.2.1. Datos macroeconómicos

Para iniciar con los datos financiero, se ha desarrollado un conjunto de herramientas en código Python para descargar, limpiar y visualizar datos macroeconómicos utilizando principalmente la biblioteca de pandas_datareader y la API de código abierto de la Reserva Federal en Estados Unidos (FRED).

Primero, definimos una función diseñada para descargar los datos financieros de la API. Esta función es esencial para nuestra investigación, ya que nos permite acceder a una amplia gama de datos macroeconómicos y financieros relevantes. La función toma dos parámetros: una lista de símbolos (tickers) y una clave de API, necesaria para autenticarse con la API de FRED. En este sentido, el método se asemeja mucho a la extracción de los precios de los ETF. En este caso, los tickers utilizados han sido:

- **GS2**
  Este ticker representa la tasa de interés de los bonos del Tesoro de EE.UU. a 2 años. Es un indicador importante de las expectativas del mercado sobre la política monetaria a corto plazo.

- **GS10**
  Este ticker representa la tasa de interés de los bonos del Tesoro de EE.UU. a 10 años. Es ampliamente utilizado como una medida del costo de endeudamiento a largo plazo del gobierno de EE.UU. y como un indicador de las expectativas económicas futuras.

- **CPIAUCSL**
  Este ticker representa el Índice de Precios al Consumidor (CPI, por sus siglas en inglés) para todos los consumidores en EE.UU. Es una medida clave de la inflación que mide los cambios en los precios de una cesta de bienes y servicios representativa del consumo de los hogares.

- **PPIACO**
  Este ticker representa el Índice de Precios al Productor (PPI, por sus siglas en inglés) para todas las mercancías para consumo final. Es una medida de la inflación que mide los cambios en los precios recibidos por los productores de bienes y servicios.

- **EXUSEU**
  Este ticker representa la tasa de cambio entre el dólar estadounidense (USD) y el euro (EUR). Es un indicador importante de la fortaleza relativa entre dos de las economías más importantes.

- **WPS0571**
  Este ticker representa el Índice de Sentimiento del Consumidor de la Universidad de Michigan. Es una medida de la confianza del consumidor en la economía de EE.UU., que puede influir en las decisiones de gasto y ahorro de los consumidores.

- **UNRATE**
  Este ticker representa la tasa de desempleo de EE.UU. Es una medida clave del mercado laboral que indica el porcentaje de la fuerza laboral que está desempleada y en busca de empleo.

Una vez almacenados los datos anteriores en un DataFrame de pandas, manipulamos los índices para compatibilizarlos con todo nuestro trabajo y finalmente llevamos a cabo un proceso de limpieza para eliminar cualquier valor faltante (NaN) que pueda afectar la integridad de nuestro análisis.

Por otra parte, una parte importante de nuestro trabajo versa sobre la incorporación de factores fama french en nuestro modelado. En este sentido, se descarga un conjunto de datos de factores Fama-French utilizando la biblioteca pandas_datareader. Este conjunto de datos

proporciona información sobre los factores de riesgo de mercado, tamaño y valor, que son ampliamente utilizados en la investigación financiera tal y como se expone en apartados anteriores. Al igual que con los datos financieros, llevamos a cabo un proceso de preparación de los valores para posteriormente fusionarse todos los registros descargados.

Preparado y limpiado el data, realizamos un análisis exploratorio para comprender mejor las relaciones y tendencias subyacentes en los datos. Utilizamos la biblioteca seaborn para visualizar la correlación entre diferentes variables financieras utilizando un mapa de calor. Esta visualización nos permite identificar patrones y relaciones importantes entre las variables financieras, lo que nos ayudará en nuestro análisis posterior. En este caso, vemos sobre todo que el impacto más relevante se materializa en el modelo de regresión lineal pues es afectado de manera muy negativa ante la multicolinealidad.

En resumen, nuestro proceso de código nos permite descargar, limpiar y analizar los datos macroeconómicos lo que permite analizar la coherencia y obtener la información necesaria para pasar al siguiente paso de modelado y aplicación de técnicas de predicción mediante aprendizaje automático.

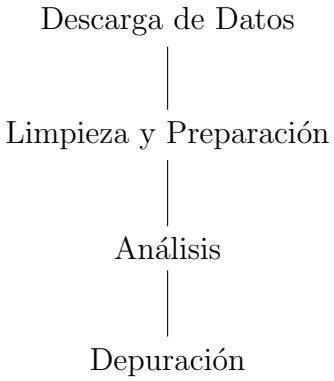

Figura 4: Proceso de Análisis de Datos

Por último, cabe destacar que en el proceso presentado se ha desestimado la incorporación de otras características por alta correlación (9X %) así como el cálculo de una nueva que proviene de la diferencia entre el GS10 - GS2 por aportar mayor significancia que las descritas por separado.

3.2.2. Datos de análisis técnico

Para abordar esta sección es importante destacar que el estudio de las características de análisis técnico se ha determinado en base al benchmark de referencia, es decir, el S&P500. Para este caso, nuevamente se ha descargado el histórico mediante la biblioteca de yfinance con el ticker correspondiente al índice, aunque esta vez se ha realizado la petición de mayores variables pues en este estudio no se limita a obtener el precio de cierre. Gracias a esta información podemos procesar los siguientes campos técnicos:

Simple Moving Average (SMA)
La simple moving average se trata de una métrica estadística que se calcula sumando los precios de cierre de un activo durante un período de tiempo determinado y dividiendo el total por el número de períodos.

$$\text{SMA}(n) = \frac{\sum_{i=1}^{n} \text{Close}(i)}{n}$$

Donde

- $\text{Close}(i)$ es el precio de cierre en el período i.
- n es el número de períodos.

Relative Strenght Index (RSI):
El índice de fuerza relativa (RSI) es un indicador de momentum que mide la velocidad y el cambio de los movimientos de precios.

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

Donde:

- RS es la relación entre el promedio de los cierres alcistas y bajistas en un período de tiempo determinado.

$$\text{RS} = \frac{\text{SMA}(n_{\text{up}})}{\text{SMA}(n_{\text{down}})}$$

Donde:

- n_{up} es el número de períodos en los que el precio de cierre subió.
- n_{down} es el número de períodos en los que el precio de cierre bajó.

Diferencia entre Precio de Apertura y Cierre:
Esta métrica calcula la diferencia entre el precio de apertura y el precio de cierre de un activo en un período de tiempo específico. Se calcula utilizando la fórmula:

$$\text{DifOpenClose} = \text{Open} - \text{Close}$$

Donde:

- Open es el precio de apertura.
- Close es el precio de cierre.

Diferencia entre Precio Máximo y Mínimo:
Esta métrica calcula la diferencia entre el precio máximo y el precio mínimo de un activo en un período de tiempo específico.

$$\text{DifHighLow} = \text{High} - \text{Low}$$

Donde:

- High es el precio máximo.
- Low es el precio mínimo.

Finalmente, fusionamos los datos técnicos con los datos históricos de volatilidad y los datos macroeconómicos para crear un conjunto de datos completo que contenga una amplia gama de factores que puedan influir en los rendimientos del mercado. Este conjunto de datos fusionado se utiliza posteriormente para realizar un análisis de regresión para predecir los rendimientos de los activos financieros.

3.3. Aplicación de modelos

En este subapartado se pretende dar una visión más amplia sobre la aplicación de los tres modelos seleccionados. Para ello, se describen la selección de hiperparámetros, la selección de parámetros la aplicación del modelo para generar predicciones y la convergencia con los tickers con mayor resultado mensual para nuestra cartera.

El primer paso vino dado de la segregación de los datos del Dataset dónde existían todas las variables fusionadas, tanto las dependientes como las independientes.

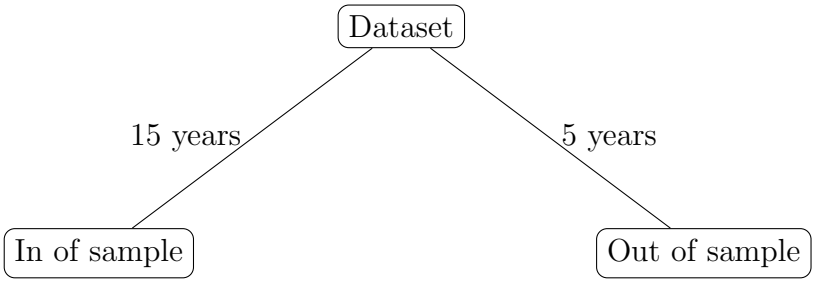

Figura 5: Segregación de datos de Dataset en dos subconjuntos

Algunas de las razones para segregar en este formato el conjunto de datos de entrenamiento y testeo se enmarcan en:

- **Generalización Mejorada**: Entrenar los modelos en un conjunto de datos amplio de 15 años permite capturar mejor las tendencias a largo plazo y los patrones históricos en los datos. Al aplicar el modelo entrenado a un conjunto de datos más reciente de 5 años (out of sample), se puede evaluar de manera efectiva la capacidad del modelo para realizar previsiones con una nueva incorporación de datos.
- **Eficiencia Computacional**: La división de datos (75-25) permite una optimización más eficiente del proceso de búsqueda de parámetros e hiperparámetros. Al tener un conjunto de entrenamiento considerablemente mayor, se pueden realizar ajustes más precisos en el modelo. El conjunto de prueba más pequeño, pero representativo, permite validar estos ajustes sin requerir recursos computacionales excesivos.
- **Estabilidad de Modelos**: Entrenar el modelo con un conjunto de datos de 15 años proporciona una base sólida y estable para estimar los parámetros e hiperparámetros. Esta segmentación ayuda a evitar el sobreajuste, ya que el modelo es probado en un conjunto distinto que no ha sido expuesto durante el entrenamiento. Esto garantiza que las evaluaciones de desempeño sean más robustas y fiables.

3.3.1. Búsqueda de hiperparámetros

Los hiperparámetros de los modelos son configuraciones externas al modelo mismo que afectan su comportamiento y rendimiento. De hecho, aunque su desarrollo vendrá en los apartados finales, su modificación resulta en cambios importantes en el producto final.

Los hiperparámetros son propios de cada modelo, no obstante, la librería utilizada en cada uno de ellos fue GridSearchCV. Esta, mediante el listado de los hiperparámetros a buscar busca en base a el error cuadrático medio el rendimiento para evaluar los mejores para cada uno. Este

proceso proporciona una metodología sistemática para optimizar los modelos lo que permite mejorar su capacidad predictiva y su rendimiento en la predicción del valor de cada ETF.

Modelo	**Hiperparámetros**
Linear Regression	fit_intercept positive
LightGBM	learning_rate num_leaves feature_fraction min_data_in_leaf num_boost_round
Random Forest	n_estimators max_depth min_samples_split min_samples_leaf bootstrap

Cuadro 2: Hiperparámetros de los modelos de Regresión Lineal, LightGBM y Random Forest

Aunque es importante describir la función de cada uno de ellos, se ha considerado resumir al máximo los conceptos.

- **fit_intercept**: Determina si se ajusta o no la intercepción en el modelo de regresión lineal.
- **positive**: Controla si los coeficientes de la regresión deben ser positivos o no.
- **n_jobs**: Número de trabajos paralelos a ejecutar durante el ajuste del modelo.
- **learning_rate**: Tasa de aprendizaje del modelo.
- **num_leaves**: Número máximo de hojas que puede tener un árbol (límite de expansión).
- **feature_fraction**: Fracciones de características a considerar en cada árbol.
- **min_data_in_leaf**: Número mínimo de puntos de datos que deben aparecer en cada árbol.
- **num_boost_round**: Número de iteraciones que puede realizar el modelo.
- **n_estimators**: Número de árboles en el bosque modelo.
- **max_depth**: Profundidad de cada árbol.

- **min_samples_split**: Número mínimo de muestras por nodo.
- **min_samples_leaf**: Número mínimo de muestras por hoja.
- **bootstrap**: Controla si se utiliza el reemplazo al construir árboles.

3.3.2. Búsqueda de parámetros

El trabajo realizado implementa una estrategia de validación "walk-forward"para la evaluación de cada uno de los modelos en el conjunto de datos de series temporales. Esta estrategia implica entrenar y validar el modelo en ventanas temporales deslizantes a lo largo del tiempo.

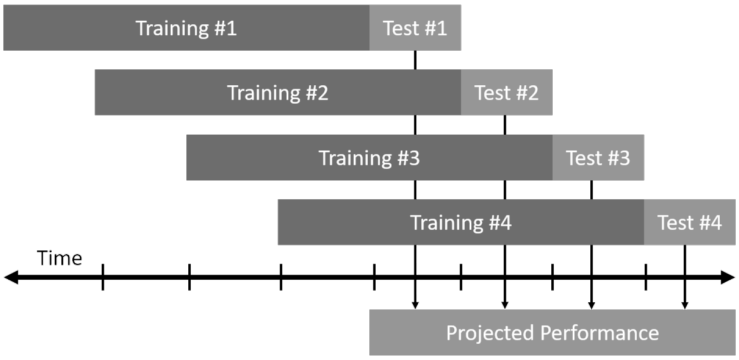

Figura 6: Proceso walk forward en series temporales [20]

La función propone como entrada las variables dependientes e independientes, así como la especificación del modelo a evaluar. Por otra parte, se estipulan un rango de parámetros de entrenamiento (train length), de longitud de prueba (test length) y de pasos adelante (lookahead). Luego, entrenado el modelo, se evalúa el rendimiento en el conjunto de prueba utilizando la correlación de Spearman como métrica. La función devuelve el promedio de las correlaciones de Spearman obtenidas a lo largo de todas las ventanas de validación.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Donde:

- ρ es la correlación de Spearman.
- d_i es la diferencia entre los rangos de los pares de datos.

- n es el número de pares de datos.

En resumen, se realiza una búsqueda exhaustiva de los mejores parámetros para el modelo. Se exploran diferentes combinaciones de valores para los parámetros presentados mediante bucles. En cada iteración, se utiliza la función y se registra el conjunto de parámetros que maximiza la correlación de Spearman permitiendo encontrar la combinación óptima de parámetros.

3.3.3. Estrategia de predicciones

La estrategia de prediccion desarrollada fue modificada a medida que se trabajaba en ella. Inicialmente el autor codifico sobre predicciones alpha o lo que es lo mismo, predicciones sobre el exceso de retorno que producían los distintos vehículos de inversión respecto al índice de referencia. No obstante, la aplicación de hiperparámetros destino muchos esfuerzos y finalmente se optó por simplificar la predicción a rendimientos de los fondos. Por otra parte, también es importante señalar la aplicación de estrategias long short en uno de los modelos que hizo descartar la aplicación a los demás por sus resultados. Así pues, las estrategias viables se vinculan a estrategias en largo.

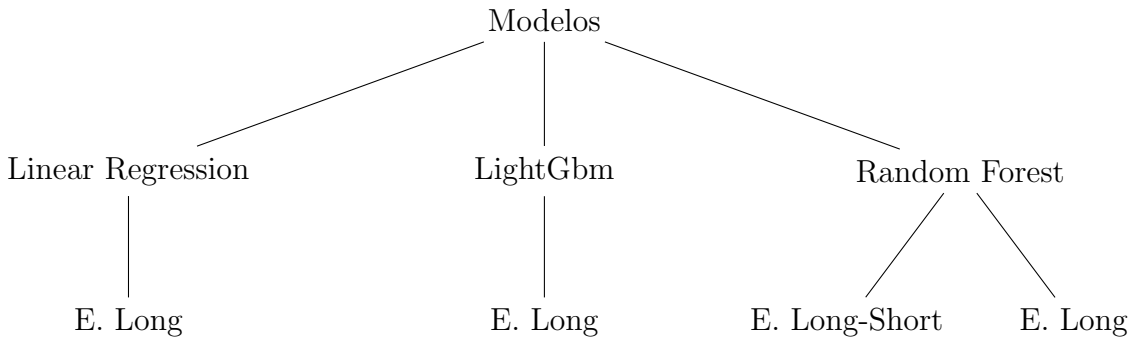

Figura 7: Aplicación de distintas estrategias

Los distintos modelos se entrenaron utilizando la técnica de validación cruzada de series temporales, que es especialmente adecuada para datos financieros debido a su naturaleza secuencial. Las predicciones obtenidas se combinaron para generar un conjunto completo de resultados. Este proceso se realiza para cada uno de los ETF. Asimismo, como cada modelo presenta distintos parámetros, se han unificado los datasets finales para presentar resultados a partir de una fecha concreta y así poder realizar comparativas más precisas.

Con base en las predicciones del modelo, se construye un marco de señales que identifica los activos financieros que generan mayor rendimiento para cada mes. Estas señales se utilizan

posteriormente en un proceso de backtesting, donde se evalúa el rendimiento de una estrategia de inversión basada en estas señales. Este paso es fundamental para determinar la viabilidad y la eficacia de la estrategia propuesta en condiciones históricas. Otro aspecto relevante es el número de ETFs utilizados. En este caso, la selección se ha basado en tres vehículos de inversión en estrategia de compra (long).

Para finalizar el estudio se realiza una evaluación exhaustiva del rendimiento del modelo propuesto. Se calculan métricas clave, como el rendimiento anualizado, el rendimiento acumulado total, la volatilidad, la ratio de Sharpe o el drawdown máximo. Estas métricas proporcionan una comprensión completa de la eficacia y el riesgo asociado con la estrategia de inversión propuesta.

Finalmente, también es importante valorar la importancia de las características en nuestro modelo pudiendo así mejorar la interpretación del modelo.

3.4. Descripción de los productos obtenidos

El proceso de modelado de aprendizaje automático aplicado al análisis financiero ha resultado en la creación de una estrategia de inversión que puede ser rigurosamente comparada con el índice de referencia. Mediante la utilización de técnicas de regresión lineal, árboles de decisión o aprendizaje supervisado se ha desarrollado un marco metodológico sólido que ayuda a predecir el rendimiento futuro de varios Exchange Traded Funds.

La estrategia de inversión generada a partir de este proceso ha sido sometida a un exhaustivo backtesting, donde se ha evaluado su desempeño en condiciones históricas.

Finalmente, este enfoque demuestra la capacidad de la ciencia de datos para ofrecer soluciones prácticas y rentables en el ámbito financiero. La estrategia de inversión desarrollada constituye una herramienta valiosa para los inversores y gestores de activos, proporcionándoles una ventaja competitiva al tomar decisiones de inversión fundamentadas en datos sólidos y análisis predictivos. Además, crea un punto de partida para futuras investigaciones más profundas en la mejora y refinamiento de técnicas de modelado financiero utilizando aprendizaje automático.

3.5. Valoración económica del trabajo

3.5.1. Coste económicos derivados del trabajo

Los principales portales de selección de personal presentan de media que los sueldos de perfiles junior en el campo de la data science se sitúan en los 27.000€ brutos. Esto se refiere a un sueldo entrada, es decir, el sueldo que pagaría el mercado a un perfil con experiencia inferior a los dos años. Con el cálculo de horas realizadas, unas ochenta horas, el computo invertido en la realización del presente trabajo se estima en dos semanas de trabajo a jornada completa. Esto supone un importe total de 1.125€ en mano de obra.

En lo que supone el coste de electricidad de la computadora esta sería de 100 horas. El consumo de vatios por hora de los ordenadores se estima en 50. 5000 vatios entre 1000 para calcular el Kw son 5 que a un precio de 0,2347 (media anual 2023) suma un total de 1,17€

Referente a la amortización del ordenador utilizado podemos realizarla con el método lineal dónde vida útil será igual a 5 y el precio de 1.500€ saldrá a 300€ anuales. De estos, podemos imputar la mitad (es decir, 150€) al trabajo de final de máster que corresponde a un semestre de trabajo.

Por el lado del software, el coste de las licencias ha sido nulo. En todo caso, los aplicativos software utilizados han sido:

- **Anaconda Navigator** para utilización de Jupiter Notebook.
- **Overleaf** para redacción y presentación de memoria.

$$\begin{aligned} \text{Coste de mano de obra} &= 1125\,€ \\ \text{Coste de electricidad} &= 1,17\,€ \\ \text{Coste de amortización del ordenador} &= 150\,€ \\ \text{Coste total} &= 1276,17\,€ \end{aligned}$$

3.5.2. Costes de mantenimiento

Por lo que respecta a los costes de mantenimiento, el trabajo realizado presenta una posibilidad potencial de mantener a largo plazo un spread significativo contra el benchmark. No

obstante, es importante tener en cuenta la necesidad de infraestructura para llevar a cabo la gestión propuesta. En el contexto de una gestora de activos que dedique el empleo al control y gestión de activos puede ser viable aplicar alguno de los modelos como estrategia de inversión. Fuera de una institución, se hace difícil implementar un modelo que requiera la supervisión constante, así como la orden de compra y venda de los activos.

3.5.3. Beneficios económicos del trabajo

En el marco de este trabajo, se ha explorado el potencial de los algoritmos data-driven para superar el rendimiento del índice S&P500 en el mercado financiero. Si bien el objetivo principal radica en alcanzar un rendimiento superior al del benchmark, es importante destacar que este logro no sólo implica un éxito académico, sino también potenciales beneficios económicos.

Por otra parte, es importante destacar que, en el mundo real, los inversores corren con costes adicionales en forma de comisiones de compra venta o comisiones de gestión. Estas comisiones, expresadas como la Total Expense Ratio (TER), representan un porcentaje anual de los activos gestionados por el fondo. Aunque no se presenten explícitamente en el trabajo, pueden mermar significativamente el rendimiento neto de una estrategia de inversión a largo plazo.

Superar consistentemente al S&P500, incluso después de considerar las comisiones de gestión, puede generar un crecimiento de la riqueza significativo para los inversores. Esto se traduce en mayores beneficios económicos, ya sea en forma de ingresos adicionales para los inversores individuales o mayores rendimientos para las instituciones financieras.

4. Resultados

En este apartado se presentan los resultados obtenidos del proceso anteriormente mencionados, por lo tanto, esta sección responde a la descripción metodológica detallada. Para ello plantearemos una introducción común para posteriormente centrarse en cada uno de los modelos estudiados.

En primer lugar debemos fijarnos en las correlaciones. Buscar correlaciones entre las variables independientes en un modelo de machine learning se considera altamente importante dado que existen diversos problemas que podrían dificultar el resultado óptimo de los modelos. En algunos casos, variables independientes demasiado correlacionadas podrían transformarse en problemas de multicolinealidad, lo que afectaría la precisión del modelo. Por otra banda, la información redundante aumentaría la complejidad del modelo sin resultar en una mejora predictiva. Otro aspecto importante es como la incorporación de variables sin valor añadido dificulta el procesamiento computacional sin obtener mejoras en los resultados. En este sentido, se han eliminado variables de inflación (tal y como muestra la imagen siguiente ya existen suficientes) así como se ha eliminado el Treasury a dos años y diez años incorporando el spread entre ellos. El estudio realizado denotó una mejora significativa del modelo con el cambio.

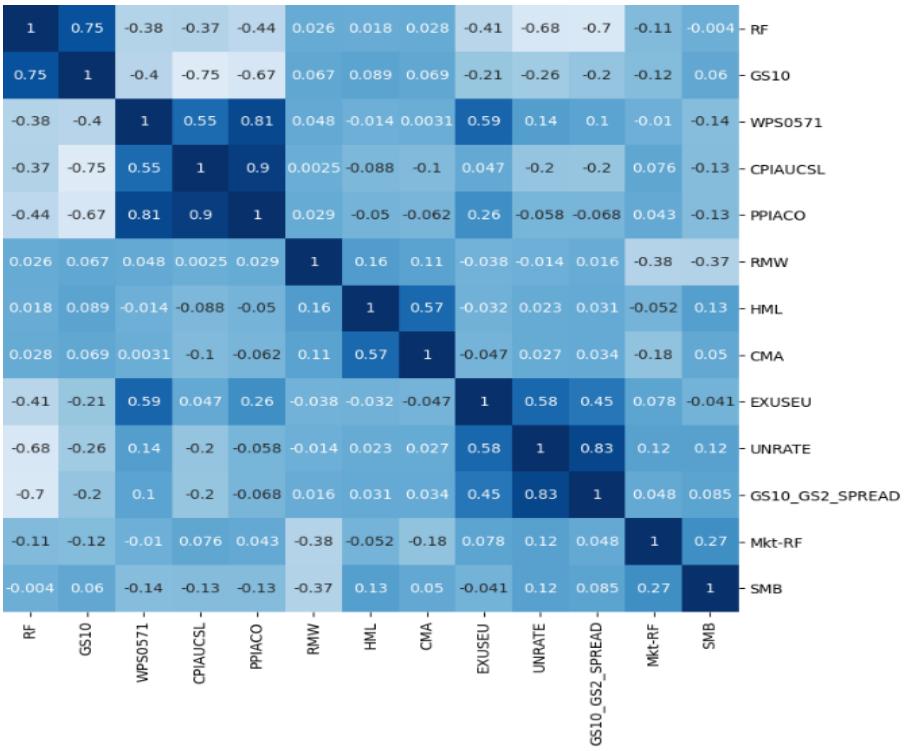

Figura 8: Aspecto final de la matriz de correlación de las variables independientes utilizadas

Por otra parte, tal y como se comenta en apartados anteriores, uno de los objetivos iniciales fue el diseño y aplicación de una estrategia long short, es decir, utilizar modelos de predicción para identificar aquellos vehículos de inversión que dadas las variables independientes aportaban un resultado mejor y un resultado peor en un determinado momento. Una vez obtenida la señal, la estrategia recogería los tres activos con mayor rendimiento y los tres activos con peor rendimiento (a la inversa).
Este estudio se descartó en una etapa temprana del trabajo pues una vez aplicada en uno de los modelos denotó rendimientos planos. Esta estrategia podría ser analizada con mayor detalle pues presenta un estilo de gestión que puede dar cobertura a movimientos bruscos del mercado. En todo caso, el presente trabajo busca mejorar el rendimiento del índice de referencia.

	S&P500	**Linear Regression Long Short**
Annualized Returns (%)	12.85	3.81
Total Cumulative Return (%)	83.05	20.58
Volatility (Standard Deviation) (%)	19.0	6.0
Sharpe Ratio	0.68	0.64
Maximum Drawdown (%)	-24.17	-8.06

Cuadro 3: Comparativa de ratios S&P500 y el modelo de Regresión Lineal Long Short de 01/01/2019 hasta 01/01/2024

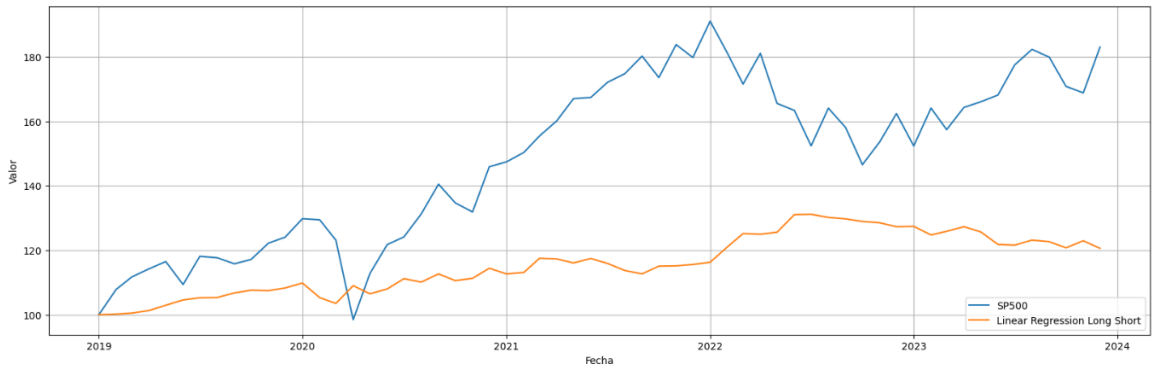

Figura 9: Rendimiento acumulado S&P500 versus el modelo de Regresión Lineal long short de 01/01/2019 hasta 01/01/2024

4.1. Linear Regression

Sabemos que los modelos de regresión lineal son más simples en comparación con los otros modelos presentados en este trabajo. Por otro lado, este tipo de modelos tienen menos hiperparámetros que ajustar. Dadas los pocos elementos manipulables en este caso, presentamos el ajuste de dos hypeparámetros combinados:

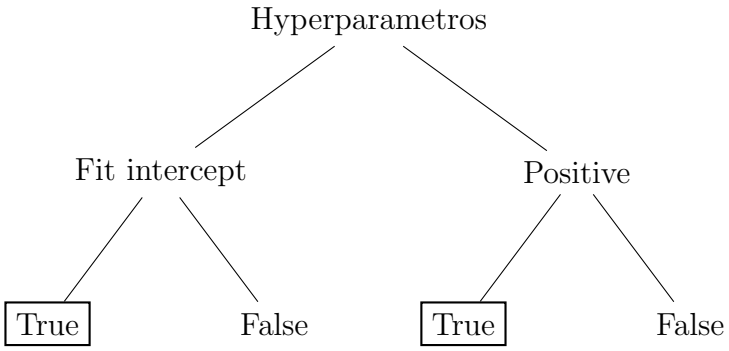

Figura 10: Hyperparametros para el modelo de regresión lineal

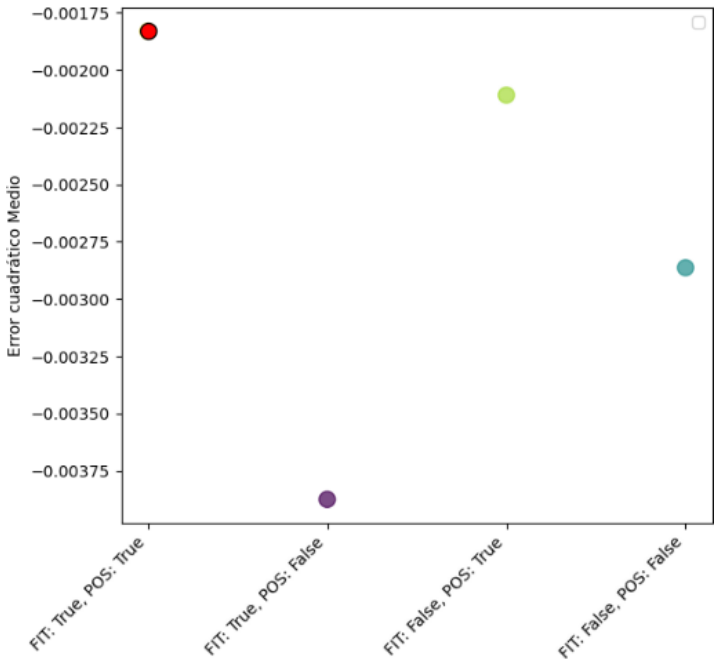

Figura 11: Combinación de hyperparámetros para modelo de Regresión Lineal

La siguiente figura muestra un mapa de calor dónde se combinan distintos valores de train y test para un lookahead dado. En este caso, sabiendo que el parámetro lookahead no modifica en exceso los coeficientes, el valor de test presenta valores favorables a partir de doce hacia dieciocho. En relación al valor de train presenta un aumento significativo a medida que se

incrementa el valor, sobretodo en conjuntos de lookahead más elevados. Finalmente se selecciona el mayor valor en ambos casos.

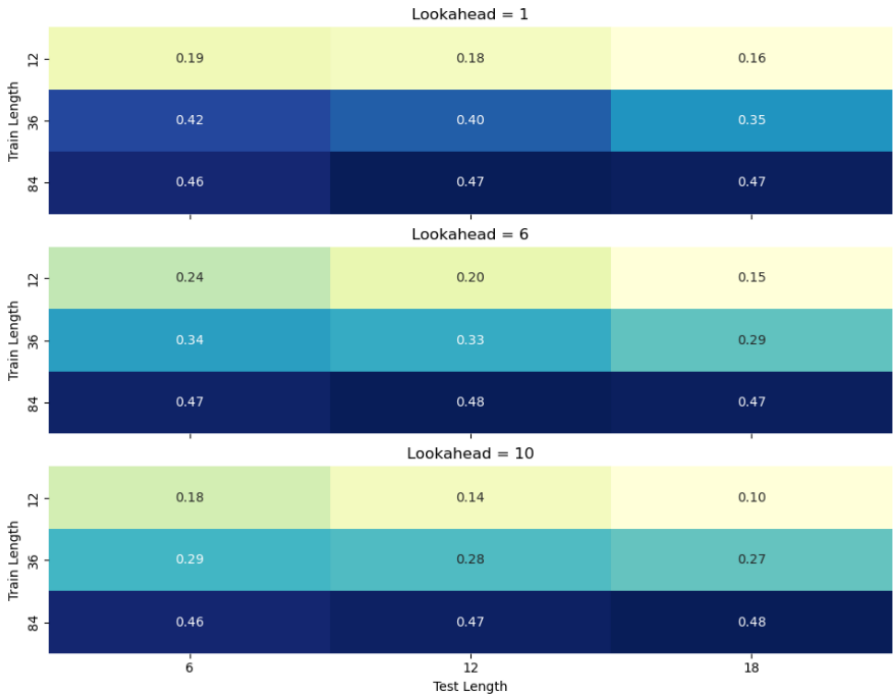

Figura 12: Búsqueda por coeficiente de spearman para Linear Regression

El resultado final puede visualizarse en el cuadro siguiente. Esta muestra como el S&P500 ha superado consistentemente al modelo de regresión lineal en varios aspectos clave. Los rendimientos anualizados y acumulados del S&P500 son significativamente mayores. Por otro lado, la menor pérdida máxima del S&P500 también indica una mejor capacidad para gestionar periodos de mercado desfavorables, ofreciendo una mayor protección del capital durante caídas significativas. Se aprecia que la estrategia implementada ofrece una menor volatilidad que el S&P500 aunque no supone una métrica a considerarse como negativa dado que deriva del perfil inversor.

	S&P500	**Linear Regression**
Annualized Returns (%)	12.85	6.03
Total Cumulative Return (%)	83.05	34.02
Volatility (Standard Deviation) (%)	19.0	18.0
Sharpe Ratio	0.68	0.34
Maximum Drawdown (%)	-24.17	-33.86

Cuadro 4: Comparativa de ratios S&P500 y el modelo de Regresión Lineal de 01/01/2019 hasta 01/01/2024

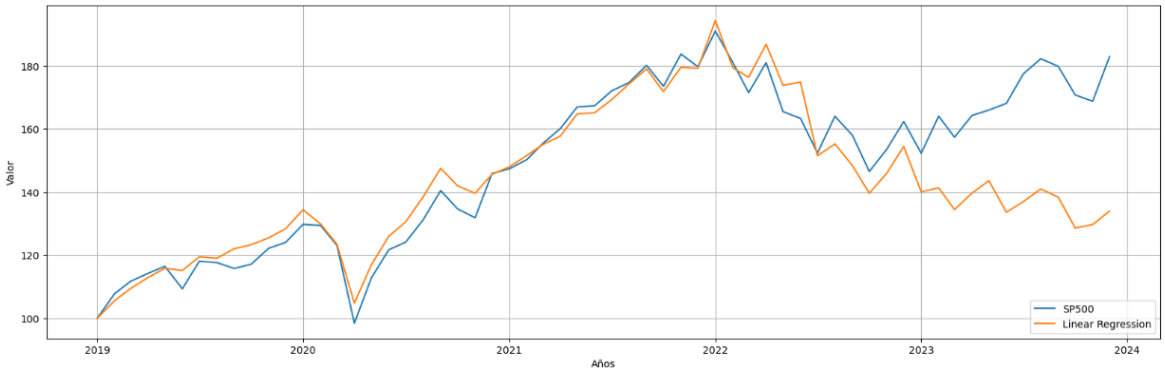

Figura 13: Rendimiento acumulado S&P500 versus el modelo de Regression Lineal de 01/01/2019 hasta 01/01/2024

Finalmente, se incide en las ocho variables independientes más importantes para la predicción del modelo. Las características RF, EXUSEU, Dif_Open_Close, y GS10 son las principales contribuyentes al modelo, mientras que las otras características tienen una influencia mucho menor. Esto implica que en términos de predicción o análisis, concentrarse en estas cuatro características podría ser más efectivo y eficiente. Además, destaca que la principal variable, aquella que el modelo propone como más relevantes es referente al modelo Fama French.

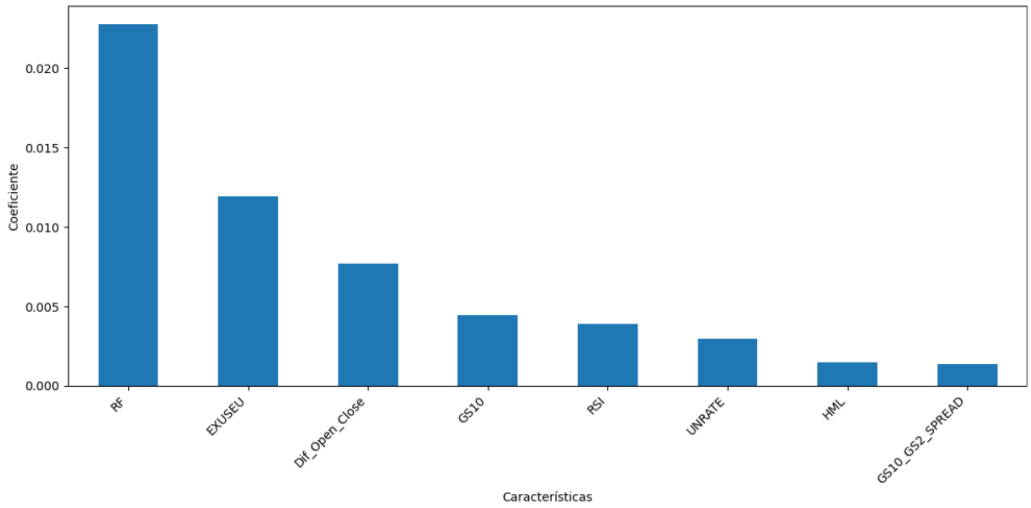

Figura 14: Importancia de las características principales en Linear Regression

4.2. Light Gradient Boosting Machine

En lo que respecta al modelo Lightgbm, podemos decir que los hiperparámetros presentados están configurados para optimizar el rendimiento del modelo enfocándose en un equilibrio entre la capacidad de ajuste del modelo y la prevención del sobreajuste. La elección concreta que vemos a continuación supone varias lecturas:

Propone una tasa de aprendizaje moderada, permitiendo que el modelo se entrene de manera estable y se ajuste gradualmente. Además, los árboles están equilibrando la complejidad del modelo. Se destaca la utilización del 60 % de las características en cada iteración reduciendo así el sobreajuste. Configura un mínimo de 20 datos para cada hoja d'estudio. Finalmente, también indica un entrenamiento durante 100 rondas de boosting, proporcionando suficiente iteración para captar patrones complejos.

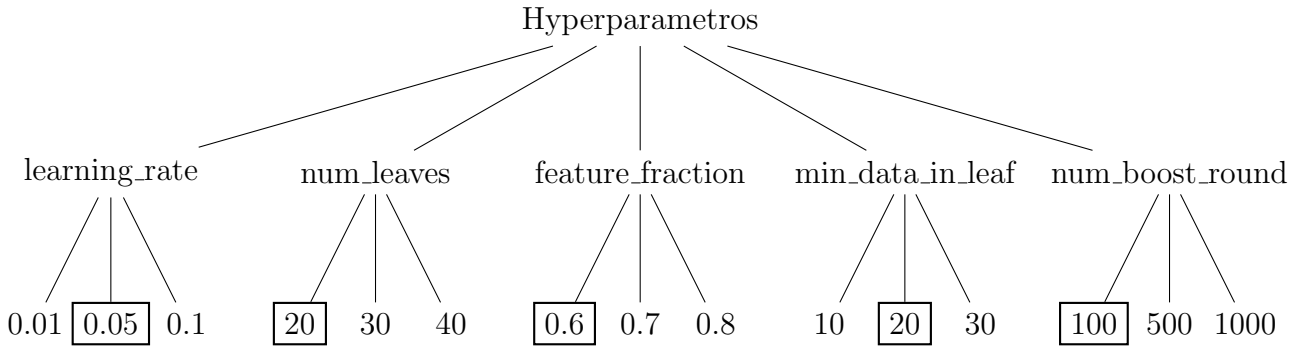

Figura 15: Hiperparámetros para el modelo LightGBM

Tal y como se presentaba en el modelo anterior, nuevamente el gráfico de relación entre el coeficiente de sperman basado en las variables de train y test lenght para un lookahead dado ponen de manifiesto que a medida que la longitud del conjunto de prueba (Test Length) aumenta, el coeficiente de Spearman tiende a mejorar, especialmente cuando el lookahead se situa en uno. Respecto al conjunto de entrenamiento, es interesante destacar que para los valores más bajos del rango propuesto el coeficiente de spearman es irrelevante por lo qual no se puede seleccionar. Otra conclusión es que a mayor valor de lookahead menor coeficiente global. Por todo ello, se observa un área óptima donde el coeficiente de Spearman es más alto (dieciocho de test lenght y en los ochenta y cuatro de train lenght siendo lookahead uno), indicando una mejor correlación en estas combinaciones.

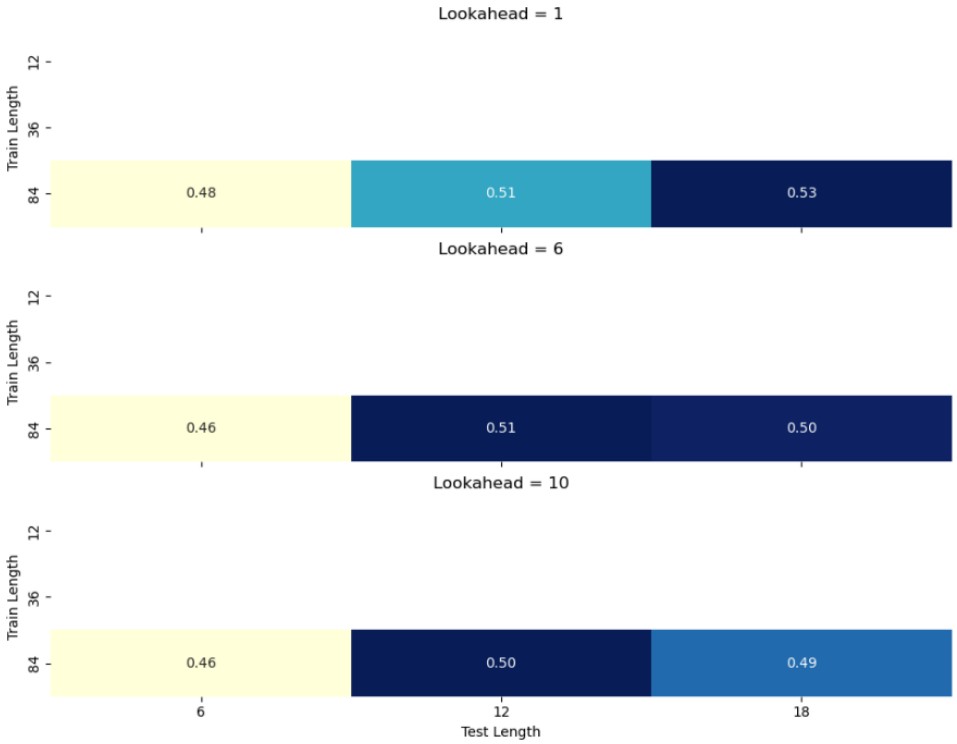

Figura 16: Búsqueda por coeficiente de spearman de train y test lenght para Lightgbm

Gracias al estudio realizado en este modelo podemos interpretar que Lightgbm sabe identificar los movimientos de mercado con cierto delay. Es importante visualizar como los últimos 2 años el modelo ha sabido mantener un rango lateral en un momento dónde el benchmark ha sufrido comportamientos negativos. En todo caso, queda claro como en mercados alcistas, el modelo tiende a quedarse por debajo del benchmark (S&P500).

Las métricas ofrecidas destacan una ligera mejora comparativa en términos de retorno aunque es importante resaltar como las otras métricas proponen la estrategia más volátil. Concretamente como suma de los datos estudiados, se considera el modelo de Lightgbm como una alternativa no determinante para superar el índice de referencia de manera consistente.

	S&P500	**LightGBM**
Annualized Returns (%)	12.85	13.86
Total Cumulative Return (%)	83.05	91.34
Volatility (Standard Deviation) (%)	19.0	22.0
Sharpe Ratio	0.68	0.63
Maximum Drawdown (%)	-24.17	-34.45

Cuadro 5: Comparativa de ratios S&P500 y el modelo de LightGBM de 01/01/2019 hasta 01/01/2024

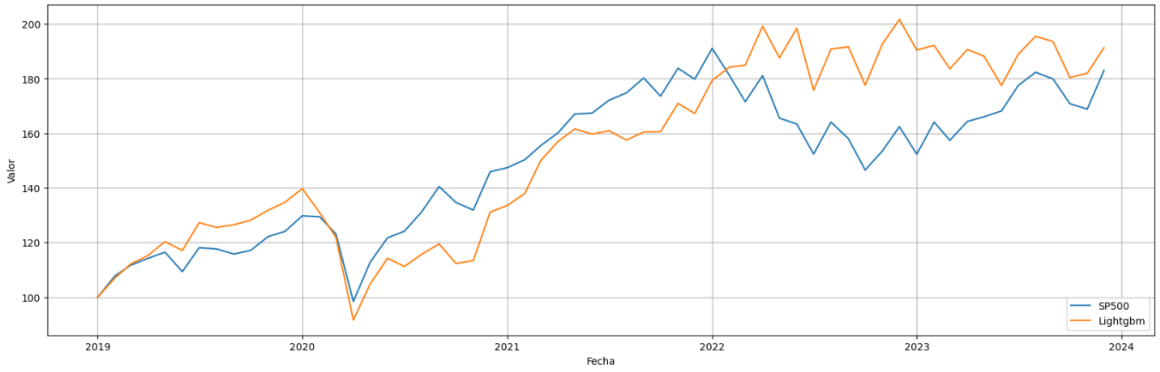

Figura 17: Rendimiento acumulado S&P500 versus el modelo de Lightgbm de 01/01/2019 hasta 01/01/2024

Como punto y final, cabe destacar la importancia de las características en la predicción de rendimiento en Light Gradient Boosting Machine. En este caso, vemos que las variables con mayor aportación al modelo son una combinación de datos técnicos, datos macroeconómicos y datos del modelo Fama French. En este sentido, se visualiza una diversidad de componentes importante que a diferencia de Linear Regression ofrece mayor similitud del modelo contra su índice comparativo.

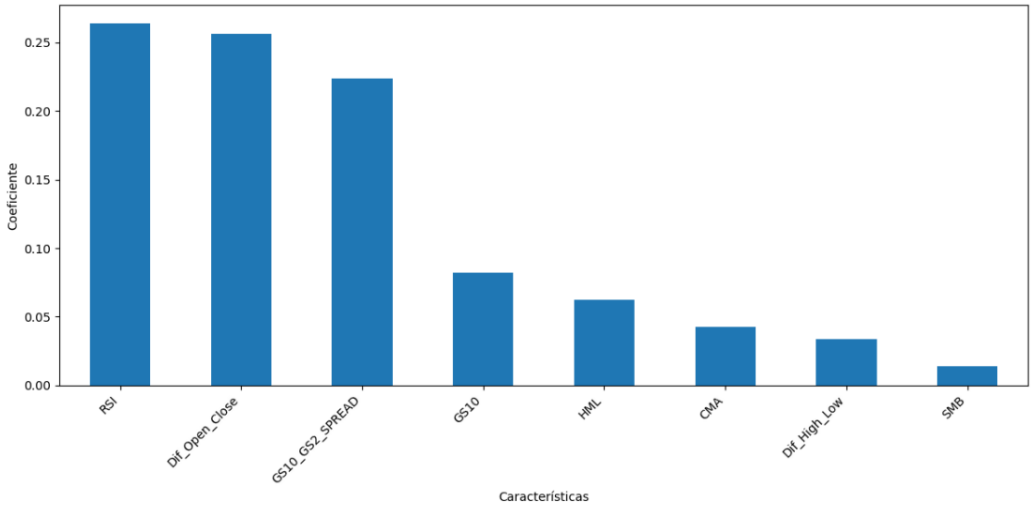

Figura 18: Importancia de las características principales en Lightgbm

4.3. Random Forest

El último modelo presentado se trata de Random Forest. En este caso, se proponen distintos hyperparámetros adaptados al mismo. Este, propone el uso de bootstrap, lo que permite mejorar la estabilidad y la precisión del modelo mediante el uso de múltiples muestras de entrenamiento. Además, establece una profundidad máxima de 30, lo que hace hincapié en un modelo complejo que tiende a buscar un buen nivel de ajuste. Configura un mínimo de 4 datos para cada hoja, asegurando que cada hoja tenga suficiente información para hacer predicciones significativas. La optimización del modelo también establece un mínimo de 2 muestras para dividir un nodo, permitiendo un equilibrio adecuado entre la división de nodos y la complejidad del árbol. Finalmente, indica un entrenamiento utilizando 100 árboles, proporcionando suficiente iteración para captar patrones complejos sin caer en el riesgo de sobreajuste.

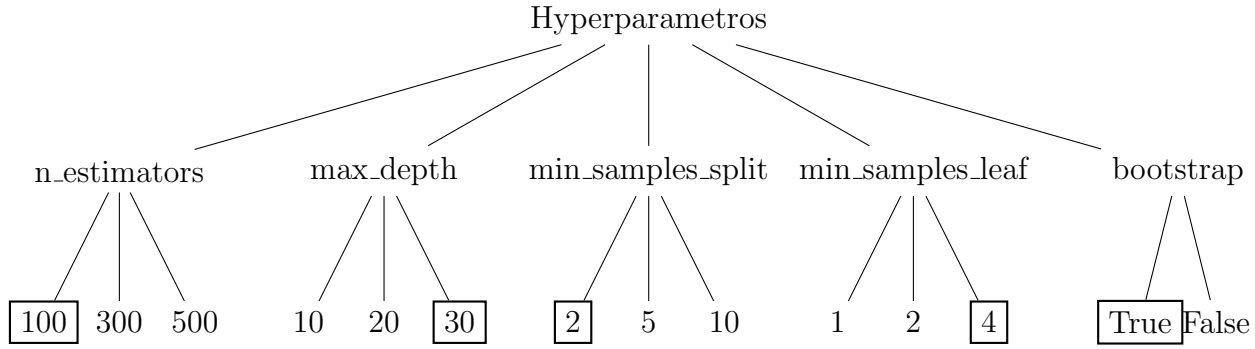

Figura 19: Hiperparámetros para el modelo Random Forest

El siguiente gráfico muestra la correlación de Spearman para el modelo de Random Forest, tal y como se ha hecho anteriormente evaluando diferentes combinaciones de parámetros de longitud de entrenamiento (train) y longitud de prueba (test) para varios horizontes de predicción (lookahead). En general, se observa que una mayor longitud de entrenamiento, especialmente de ochenta y cuatro, tiende a ofrecer las mejores correlaciones. A medida que la longitud de prueba aumenta, también mejora la correlación, lo que sugiere que conjuntos de prueba más largos proporcionan una mejor estimación del rendimiento del modelo.

Para un lookahead de uno, las correlaciones son más altas cuando se utilizan longitudes de entrenamiento y prueba mayores. Con un lookahead de seis, la correlación se mantiene alta pero es ligeramente inferior a la del lookahead de uno.

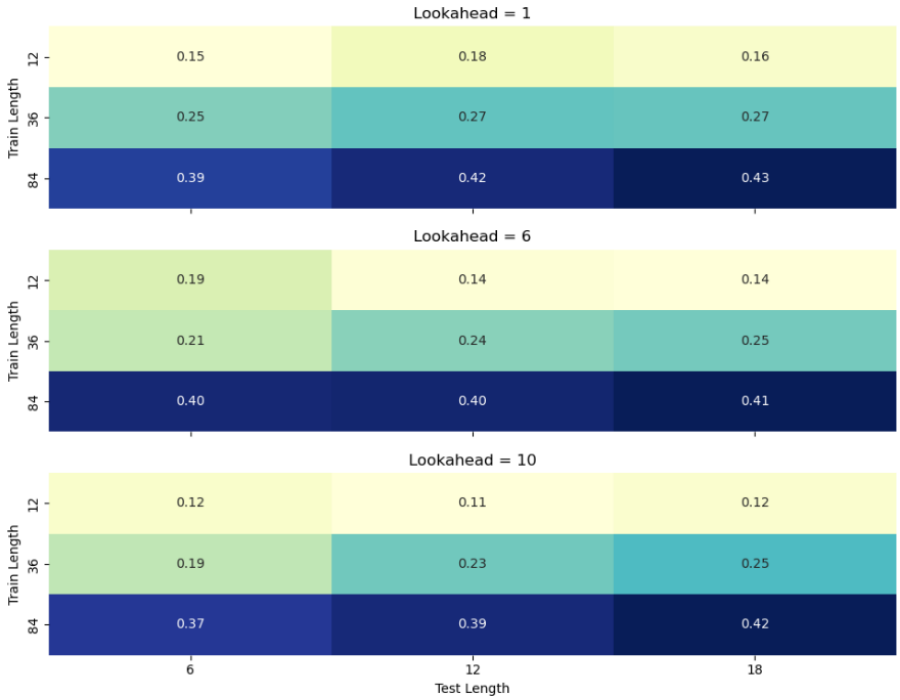

Figura 20: Búsqueda por coeficiente de spearman de train y test lenght para Random Forest

Finalmente se muestran las principales métricas de rendimiento entre el modelo de Random Forest y el índice S&P500. Los rendimientos anualizados y el retorno acumulativo total del modelo de Random Forest son superiores a los del S&P500, lo que sugiere que el modelo ha logrado batir el benchmark durante los ocho años de período analizado. Además, la volatilidad del modelo de Random Forest es ligeramente menor que la del S&P500, lo cual es positivo ya que implica menor riesgo. La ratio de Sharpe, que mide la relación entre rendimiento y riesgo, es también superior en el modelo de Random Forest, indicando una mejor compensación por el riesgo asumido. Como punto negativo de la comparativa, el modelo propone mayor drawdown, es decir, una caída máxima superior al índice de referencia. En lo general, el modelo se podría considerar mejor estrategia que el benchmark.

	S&P500	**Random Forest**
Annualized Returns (%)	12.85	10.23
Total Cumulative Return (%)	83.05	62.76
Volatility (Standard Deviation) (%)	19.0	21.0
Sharpe Ratio	0.68	0.49
Maximum Drawdown (%)	-24.17	-35.23

Cuadro 6: Comparativa de ratios S&P500 y el modelo de Random Forest de 01/01/2019 hasta 01/01/2024

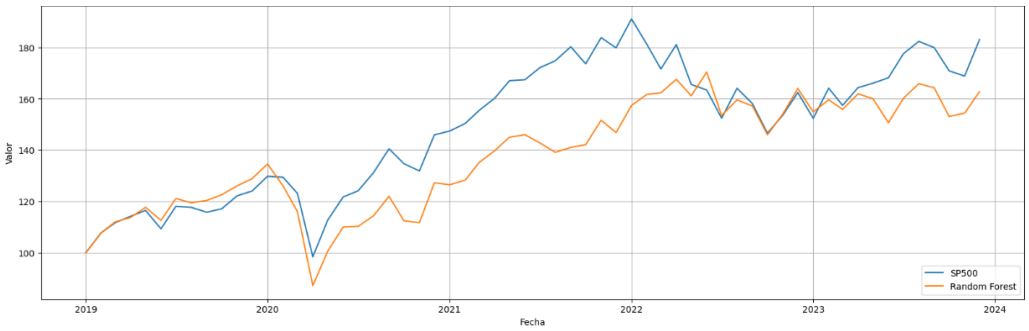

Figura 21: Rendimiento acumulado S&P500 versus el modelo de Random Forest de 01/01/2019 hasta 01/01/2024

Como punto y final, presentamos las características más importantes para Random Forest. Estas divergen de los anteriores modelos. En este caso, se muestra como principal determinante la volatilidad, es decir el valor que representa los movimientos de mercados (a más movimiento mayor valor en la característica). Otros factores interesantes del modelo son datos técnicos así como algunas variables de Fama French. En todo caso, también debemos centrar nuestra atención en las variables que corresponden a cambios en los precios (VIX, Diferencias entre High y Low, RSI o HML correspondiente a Fama French).

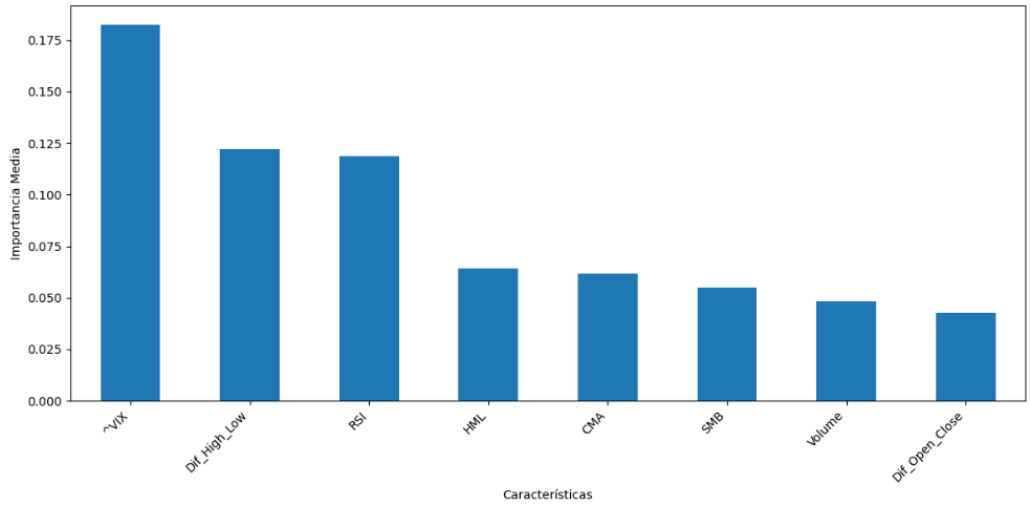

Figura 22: Importancia de las características principales en Random Forest

5. Conclusiones finales y prospectiva

Para terminar, es importante recoger todos aquellos aspectos que se han ido identificando en el transcurso del trabajo. En este sentido, se destacan varios puntos clave.

En primer lugar, se observa que todos los modelos son extremadamente sensibles a los cambios en los hiperparámetros. Esto destaca la importancia de un ajuste cuidadoso para lograr resultados óptimos. Cualquier variación en ellos determina un resultado ampliamente diferenciado y por ello, se debe analizar con profundidad para cada modelo concreto.

Por otra parte, y de la mano de lo mencionado, también se ha destacado el impacto de nuevas variables. En el caso actual, las variables independientes incorporadas no han sido demasiado elevadas, en este sentido, se puede enfatizar más la incorporación de datos técnicos que ayuden a crear métricas relevantes para los modelos evadiendo los aportes que deriven en impactos negativos en el rendimiento.

Es relevante como el estudio de cada modelo requiere una importante inversión de tiempo. Por ello, se considera crucial destacar la necesidad de estudiar profundamente cada modelo individualmente. Aunque pueda parecer obvio, este punto es esencial para maximizar su efectividad. El campo de estudio parece no tener limitaciones al combinarse con aspectos financieros.

En cuanto al número de ETFs utilizados en el análisis, se descubrió que menos puede ser más en términos de resultados. Experimentar con diferentes cantidades de vehículos de inversión, es decir, manipulando el número de ETFS utilizados en la estrategia de compra mensual puede ser beneficioso, pero los datos sugieren que un mayor número de ETFs puede conducir a un rendimiento menor. Es por ello que este estudio también podría extenderse en el tiempo en base a este cambio.

Por último, respecto a los distintos rendimientos de los modelos presentados, Linear Regresión juntamente a Random Forest presentan valores por debajo del benchmark. A diferencia de ellos, Lightgbm si supone una diferencia positiva con respecto al índice en el período estudiado. No obstante, deberíamos valorar la consistencia con en el tiempo. Además, cabe destacar que el período de tiempo estudiado supone un mercado alcista con movimientos estables cosa que podría facilitar el aprendizaje de nuestros modelos. Estas estrategias podrían variar con el tiempo y se debe poner énfasis en su estudio en tiempos más volátiles. En lo que respecta a la importancia de las características, vemos como aquellas estrategias que se centran en datos técnicos y valores de Fama French tienden a presentar mejores resultados.

Como punto y final, mi investigación resalta la complejidad y la importancia del proceso de modelado en el campo de la ciencia de datos. Cada paso, desde la selección de características hasta el ajuste de parámetros, requiere atención meticulosa para obtener resultados óptimos que no aseguran la mejora comparativa de rendimiento de los modelos sobre el índice de referencia.

Bibliografía

[1] Fabio Duarte. Amount of data created daily. exploding topics. Diciembre 2023.

[2] Yun Li. 80 % of the stock market is now on autopilot. Junio 2019.

[3] OECD. Artificial intelligence, machine learning and big data in finance: Opportunities, challenges, and implications for policy makers. Agosto 2021.

[4] CFA Institute. What is esg investing and analysis?, 2024. Consultado el 21 de Marzo de 2024.

[5] S. Jansen. *Machine Learning for Algorithmic Trading: Predictive Models to Extract Signals from Market and Alternative Data for Systematic Trading Strategies with Python*. Expert Insight. Packt Publishing, 2020.

[6] H. Markowitz. Portfolio selection. *The Journal of Fincance, 7*, 1952.

[7] D. Melas. The future of factor investing: The journal of portfolio management. *Quantitative strategies: Factor investing 7th Edition, 48(2)*, 2022.

[8] M Lopez del Prado. *Advances in Financial Machine Learning*. Whiley, 2018.

[9] Christopher Krauss, Xuan Anh Do, and Nicolas Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the sp 500. *European Journal of Operational Research*, 259(2):689–702, 2017.

[10] T. Raffinot. Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management*, 2017.

[11] E.F. Fama and K.R. French. The cross-section of expected stock returns. *Journal of Finance, American Finance Association 47(2)*, 1992.

[12] W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance, 19*, 1964.

[13] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Fama-Miller Working Paper*, 2014.

[14] Z. Bodie, A. Kane, and A. J. Marcus. *Investments (10th ed.)*. McGraw-Hill Education, 2014.

[15] Edwin J. Elton, Martin J. Gruber, Stephen J. Brown, and William N. Goetzmann. *Modern Portfolio Theory and Investment Analysis*. John Wiley Sons Inc, 2014.

[16] Mayster Boris Liew, Jim Kyung-Soo. Forecasting etfs with machine learning algorithms. *The Journal of Alternative Investments*, 2018.

[17] S. M. Bartram, J. Branke, G. D. Rossi, and M. Motahari. Machine learning for active portfolio management. *The Journal of Financial Data Science*, 3(3):9–30, July 2021. © 2021 Portfolio Management Research. All rights reserved.

[18] C. Ken. Etfs sector rotation strategy, 2022.

[19] E. Pinsky and Y.H. Yang. A simple rotation strategy with sector etfs. *Technical Analysis of Stocks and Commodities*, Diciembre 2022.

[20] S. Nawara. Avoiding data leakage in machine learning.