Predicción y Análisis del S&P 500 mediante Modelos Estadísticos y de Machine Learning

Daniel David Florez Bocanegra

7 de noviembre de 2025

Resumen

Este informe presenta el desarrollo y evaluación de diversos modelos para la predicción del retorno mensual del índice S&P 500. El objetivo fue doble: (i) estimar los retornos futuros del S&P 500 con distintos enfoques estadísticos y de aprendizaje automático, y (ii) analizar la influencia de los principales stocks sobre el comportamiento del índice.

Se compararon cinco modelos: **Regresión Ridge**, **ARIMA**, **SARIMAX-PCA**, **SARIMAX con todas las features** y **Random Forest Regressor**. Se evaluaron las métricas MSE, MAE y R^2. Los resultados muestran diferencias significativas en desempeño y estabilidad, destacando el **Random Forest** como el modelo más consistente para la predicción del retorno.

1. Introducción

El índice S&P 500 resume el comportamiento de las 500 empresas más grandes de Estados Unidos y es un referente clave de la economía global. Predecir su retorno constituye un desafío complejo debido a la naturaleza no estacionaria de los mercados y la interacción entre múltiples variables.

El propósito de este trabajo es comparar modelos estadísticos tradicionales (ARIMA, SARIMAX, haciendolos variar en estrategias) con enfoques de regresión regularizada (Ridge) y de aprendizaje automático (Random Forest), para determinar cuál ofrece mejor capacidad predictiva sobre los retornos mensuales del S&P 500 (Tomados en logaritmo). Adicionalmente, se busca interpretar los coeficientes o importancias para identificar qué acciones contribuyen más a la variación del índice.

2. Metodología

Los datos utilizados consisten en los retornos logarítmicos mensuales del S&P 500 y de sus acciones individuales. Los modelos se entrenaron con un subconjunto histórico y se validaron con un conjunto de testeo posterior.

Los enfoques comparados fueron:

- **Regresión Ridge**: modelo lineal regularizado, con penalización L2 para reducir sobreajuste.
- **ARIMA**: modelo univariado clásico, empleado como línea base.
- **ARIMA Rolling**: versión actualizada secuencialmente en cada paso, simulando una predicción realista.
- **SARIMAX-PCA**: modelo con variables exógenas reducidas mediante Análisis de Componentes Principales. (También se prueba usando rolling)
- **SARIMAX (todas las features)**: versión que incorpora la totalidad de las variables exógenas.
- **Random Forest Regressor**: modelo no lineal basado en ensamble de árboles.

Las métricas de evaluación fueron el **error cuadrático medio (MSE)**, el **error absoluto medio (MAE)** y el **coeficiente de determinación** (R^2).

3. Resultados por Modelo

3.1. Regresión Ridge

El modelo Ridge presentó un error medio en test de MAE ≈ 0.00675, MSE $= 0.000109$ y un R^2 casi nulo (0.0005), indicando una capacidad predictiva que aunque mejor que predecir la media, el margen en que lo hace es extremadamente bajo. El R^2 nos permite concluir que el modelo falla en generar buenas predicciones de los retornos, pero sigue siendo mejor que nada. El MSE y el MAE, aunque bastante pequeños hay que recordar que en realidad se están calculando los log-retornos lo que implica que el orden de los valores se reduce drasticamente.

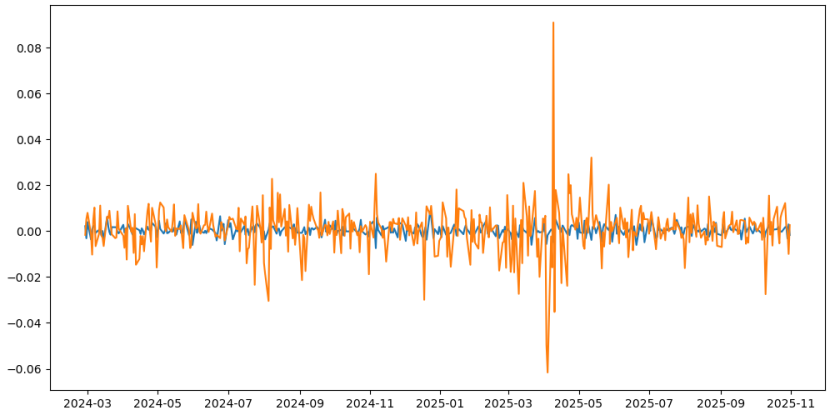

Figura 1: Gráfica de la serie de tiempo predicha por Ridge vs la serie de tiempo real del S&P500

3.2. ARIMA

El modelo ARIMA prediciendo de forma estática (todo el testeo de golpe) obtuvo MAE ≈ 0.006742, MSE $= 0.000109$ y $R^2 = -0{,}0003$. El hecho de que el R^2 sea negativo nos indica que el modelo es peor que predecir la media y en la grafica podemos observar de hecho que ARIMA tras los primeros valores termina prediciendo una recta constante que simula la media.

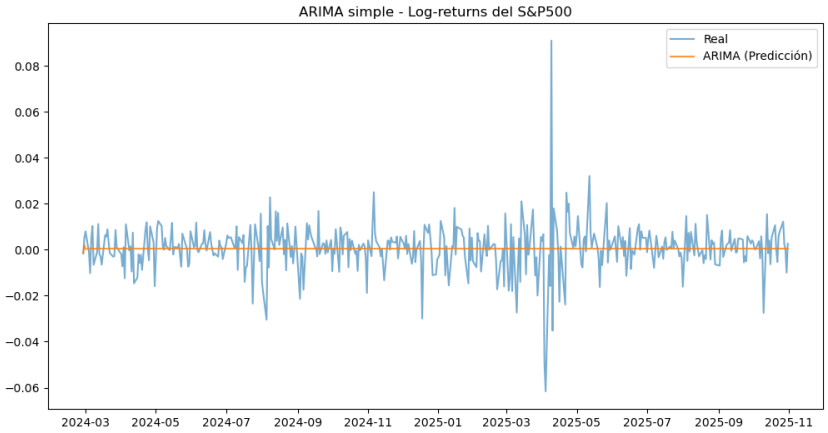

Figura 2: Gráfica de la serie de tiempo predicha por ARIMA vs la serie de tiempo real del S&P500

Cuando se actualizó dinámicamente en cada paso ("rolling"), el rendimiento mejoró ligeramente con MAE ≈ 0.00674, MSE $= 0.000108$ y $R^2 = 0{,}002$. Aquí podemos observar que el modelo mejoro y ahora es mejor que predecir la media, aunque de nuevo, R^2 es muy pequeño y no resulta en una gran predicción, puesto que no se llega a capturar prácticamente nada de la varianza del problema.

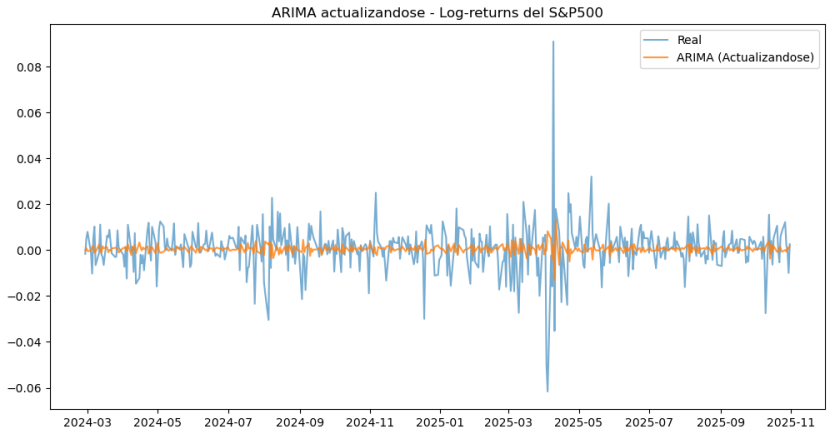

Figura 3: Gráfica de la serie de tiempo predicha por ARIMA Rolling vs la serie de tiempo real del S&P500

3.3. SARIMAX-PCA

SARIMAX con variables exógenas reducidas por PCA mostró un MAE ≈ 0.00674, MSE $= 0.000110$ y $R^2 = -0{,}0082$. Parece que la información adicional comprimida por PCA en SARIMAX no dio mejores resultados que ARIMA Rolling, pero si mejor (aunque descartable) que ARIMA simple, nuevamente el modelo no captura nada de la varianza y es peor que predecir la media.

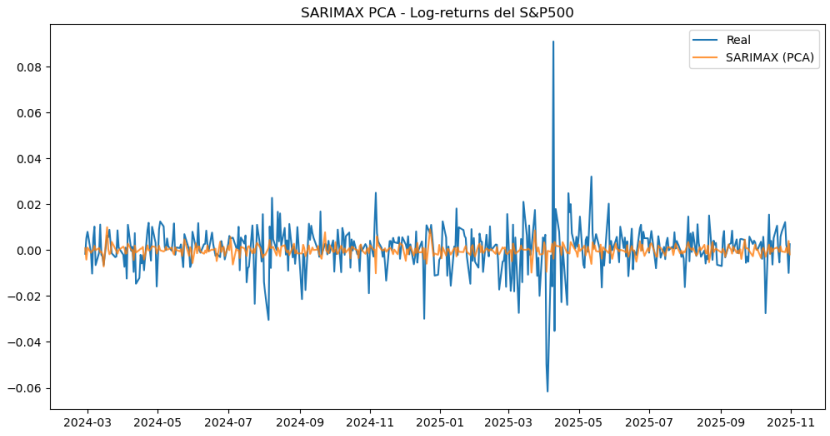

Figura 4: Gráfica de la serie de tiempo predicha por SARIMAX PCA vs la serie de tiempo real del S&P500

Al hacer rolling en SARIMAX-PCA obtuvimos los siguientes resultados, MAE ≈ 0.006764, MSE $= 0.000110$ y $R^2 = -0{,}0127$

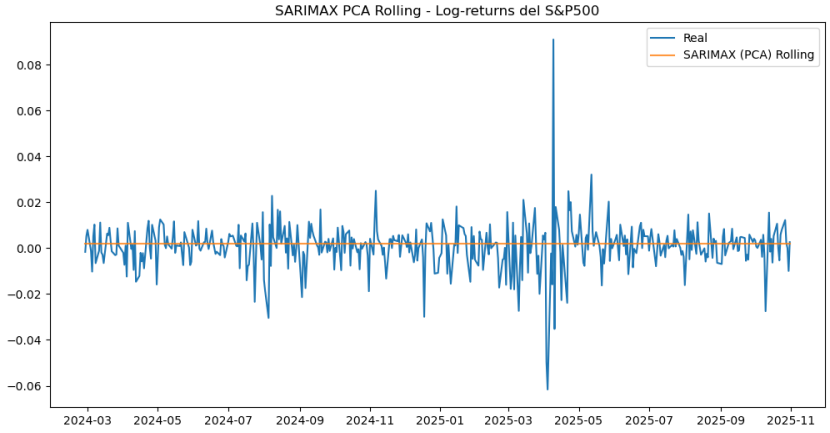

Figura 5: Gráfica de la serie de tiempo predicha por SARIMAX PCA Rolling vs la serie de tiempo real del S&P500

A diferencia del caso de ARIMA, parece que hacer rolling afecto negativamente a SARIMAX-PCA y decidio por predecir una recta constante.

3.4. SARIMAX con todas las features

Este modelo obtuvo el peor desempeño general con MAE ≈ 0.01886, MSE $= 0.000574$ y un $R^2 = -4{,}2797$, lo que podria indicarnos un sobreajusto, sin embargo, va más allá, si nos fijamos en la grafica SARIMAX esta prediciendo valores mucho más altos (o bajos) que los reales lo que no había pasado con ningún otro modelo que solía mayormente predecir valores de menor magnitud.

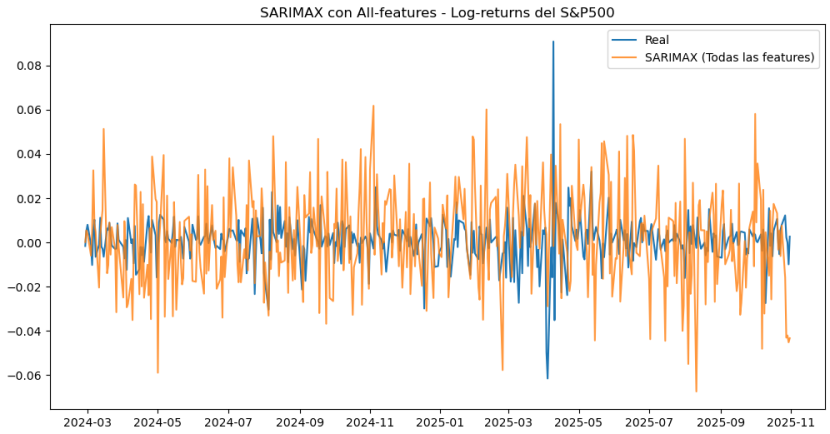

Figura 6: Gráfica de la serie de tiempo predicha por SARIMAX all features vs la serie de tiempo real del S&P500

3.5. Random Forest Regressor

El modelo Random Forest alcanzó el mejor balance entre error y estabilidad, con MAE ≈ 0.00650, MSE $= 0.000108$ y $R^2 = 0{,}0067$ aunque la mejora es marginal.

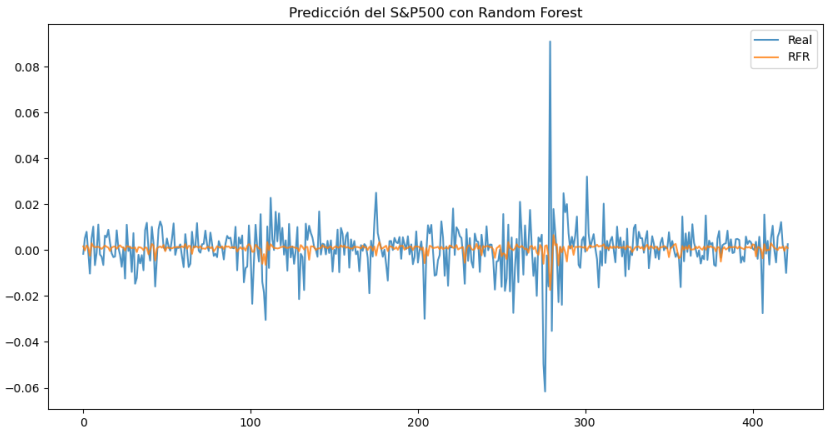

Figura 7: Gráfica de la serie de tiempo predicha por RFR vs la serie de tiempo real del S&P500

Cuadro 1: Comparación de métricas de desempeño en el conjunto de test

Modelo	**MAE**	**MSE**	**R^2**
Ridge Regression	0.0067530678210317285	0.000109	0.0005
ARIMA (batch)	0.006742186082260763	0.000109	-0.0003
ARIMA (rolling)	0.006745749477160255	0.000108	0.0020
SARIMAX-PCA	0.00674187797802022	0.000110	-0.0082
SARIMAX-PCA-Rolling	0.00676479726880968	0.000110	-0.0127
SARIMAX (full)	0.018862522699635478	0.000574	-4.2797
Random Forest	0.006503350160976997	0.000108	0.0067

4. Análisis de Importancia de Stocks

Para calcular la importancia de cada stocks se sumaron los pesos de cada una de sus features. Ridge y Random Forest Regression (RFR) no tuvieron ningun stock en común, mientras que RFR y SARIMAX (usando todas las features) tuvieron a Microsoft (MFST) en común, por otro lado, Ridge y SARIMAX tuvieron en común dos stocks; Lamb Weston Holdings (LW) y The Sherwin-Williams Company (SHW).

5. Conclusiones

Ningún modelo logró un poder predictivo alto (R^2 muy pequeños o negativos), lo que refleja la dificultad inherente de predecir retornos financieros. RFR, ARIMA Rolling y Ridge fueron los mejores tres modelos en orden. ARIMA Rolling logró vencer todas las otras variantes de ARIMA, incluidas las más avanzadas como SARIMAX, en este experimento. Los tres modelos terminaron generando los 20 stocks más relevantes muy diferentes entre ellos lo que nos imposibilita de construir un portfolio unificado usando los resultados de los tres modelos.