Reporte Téorico: Comparación de Modelos Predictivos para el S&P500

Análisis Experimental de Series Temporales

1. Datos y Configuración Experimental

Se utilizó la serie temporal diaria del S&P500, acompañada de una matriz de variables exógenas derivadas de los precios de las acciones individuales del índice. En algunos modelos, estas variables se redujeron mediante **PCA** (10 componentes principales), mientras que otros trabajaron con la matriz completa.

El conjunto de datos se dividió de forma cronológica, aplicando esquemas de validación temporal (*expanding window*) para evitar fugas de información y asegurar que cada predicción se base exclusivamente en datos pasados.

2. Modelos Evaluados

2.1. Ridge Regression (con búsqueda de hiperparámetros)

Se empleó **Ridge Regression** de `sklearn.linear_model`, aplicando **GridSearchCV** con validación *expanding window* (5 splits). El parámetro de regularización α se buscó en un espacio logarítmico de 10^{-3} a 10^{3} con 20 valores. El mejor valor encontrado fue $\alpha = 1000$

Métricas utilizadas: MAE, MSE, R^2. **Evaluación:** *expanding window*. **Gráfica:** evolución temporal de $\hat{y}_t$ vs y_t.

2.2. ARIMA (5,0,0)

El modelo **ARIMA** se ajustó con la función `ARIMA` de `statsmodels`, usando orden (5,0,0). Se omitió el parámetro b (diferenciación) al asumir estacionareidad con base en pruebas previas. Se realizaron dos variantes:

- **ARIMA estático:** entrenamiento único en todo el set de entrenamiento.
- **ARIMA rolling:** el modelo se re entrena cada día con todos los datos hasta ese punto antes de predecir el siguiente valor.

Métricas utilizadas: MAE, MSE, R^2. **Evaluación:** rolling y no rolling. **Gráfica:** y_{pred} vs y_{real} sobre el tiempo.

2.3. SARIMAX (2,0,2) con PCA (10 componentes)

Se utilizó **SARIMAX** de `statsmodels` con orden (2,0,2) y las variables exógenas definidas como las 10 componentes principales obtenidas por **PCA**. Esto reduce el ruido y la dimensionalidad, preservando las tendencias estructurales.

Se aplicó nuevamente la estrategia *rolling*, reentrenando el modelo diariamente. Esto permite incorporar información más reciente, a costa de un incremento significativo en tiempo de cómputo.

Métricas utilizadas: MAE, MSE, R^2. **Evaluación:** rolling y no rolling. **Gráfica:** y_{pred} vs y_{real} sobre el tiempo.

2.4. SARIMAX (2,0,2) con todas las features

En esta versión se preservaron todas las variables exógenas originales sin aplicar PCA. Debido al alto costo computacional, no se aplicó *rolling*, sino un solo entrenamiento y predicción completa.

Métricas utilizadas: MAE, MSE, R^2. **Evaluación:** no rolling debido al alto costo computacional. **Gráfica:** y_{pred} vs y_{real} sobre el tiempo.

2.5. Random Forest Regressor

Se usó un **Random Forest Regressor** con la siguiente configuración:

```
rfr_model = RandomForestRegressor(
    n_estimators=300,
    max_depth=8,
    random_state=42,
    n_jobs=-1,
    oob_score=True
)
```

El modelo captura relaciones no lineales entre las variables exógenas y el S&P500

Métricas utilizadas: MAE, MSE, R^2. **Evaluación:** conjunto de validación temporal. **Gráfica:** y_{pred} vs y_{real}.

3. Análisis de Importancia de Variables

Para los modelos que utilizaron todas las features (Ridge, SARIMAX completo y RFR), se calculó la agregada por stock:

1. Se sumaron los pesos o importancias de todas las features asociadas a cada acción.
2. Se seleccionaron los 20 valores más altos.

3. Se compararon los resultados entre los tres modelos para hallar intersecciones, identificando qué acciones son consistentes como predictores significativos en distintos enfoques.

Los resultados mostraron no encontrar muchas acciones compartidas debido a los diferentes enfoques que toman Ridge, RFR y SARIMAX al aplicar los pesos de las variables. Cuando existen variables altamente correlacionadas Ridge distribuye su peso entre todas las relacionadas con ella. RFR capta relaciones no-lineales por lo que termina otorgando pesos diferentes. SARIMAX no tiene regularización como Ridge e intenta capturar el efecto temporal en las variables. Por estas diferencias es que se terminan encontrando resultados muy distintos en la asignación de los pesos.

4. Conclusiones Generales

- **Ridge Regression** ofrece una referencia lineal robusta y estable, pero distribuye el peso entre variables correlacionadas, reduciendo la interpretabilidad individual. Finalmente el modelo no resulta muy efectivo para predecir los retornos, pero resulta más seguro que predecir la media
- **ARIMA** y **SARIMAX** El modelo ARIMA se desempeño mejor que el resto de sus variantes incluso a su variante mejorada por Rolling, similarmente, SARIMAX PCA Rolling gozó de la misma mejoría. SARIMAX con todas las features resulto no solo mucho más costoso, sino que peligroso, al predecir retornos muy superiores a los reales y caídas de magnitudes no reales.
- **Random Forest Regressor** Mostró ser mejor que la media, como se esperaba al tratar de captar relaciones no lineales y ser un modelo más "potente"

Los modelos no resultaron muy eficaces y aunque era esperado de los modelos lineales, al final el desempeño de RFR tampoco atacó de mejor manera el problema, quizás con un fine-tuning de sus parámetros tras hacer una pre-selección de los stocks a usar pueda mejorar el desempeño de RFR. El resultado de las métricas puede variar por ejecucción sobre todo en RFR y Ridge si se busca un alpha diferente al modificar la definición del logspace, hay casos donde Ridge podría funcionar peor que RFR reduciendo la severidad de la regularización pero conservando más información sobre los coeficientes.