

# Predicción de Retornos Financieros

Proyecto end-to-end en Data Science con series temporales y Machine Learning

Los datos me enseñan; yo aprendo construyendo, midiendo y mejorando

Andrés Vallejo | Científico de Datos Jr

Kaggle – Dataset "S&P500 All Assets Daily Update"

GitHub: [github.com/AVALLEJOTORRES/financial-returns-prediction](https://github.com/AVALLEJOTORRES/financial-returns-prediction)

- Mejor modelo: RandomForestRegressor / CatBoostRegressor

- MAE: 0.0537 – 0.0550
- RMSE: 0.0663 – 0.0671
- R<sup>2</sup>: -39.51 / -41.20

- Los resultados reflejan la dificultad del problema, pero muestran un flujo de ML aplicable al ámbito financiero

# CONTEXTO

Este proyecto aborda el reto de predecir retornos financieros a 20 días utilizando Machine Learning aplicado a series temporales de precios históricos del mercado S&P500, obtenidos desde Kaggle.

## Datos analizados:

- 28 variables originales (indicadores técnicos, rezagos, medias móviles, volatilidad, etc.).
- 3.709 registros históricos diarios de la acción de Abbott (ABT).
- Variable objetivo: target\_retorno\_20d

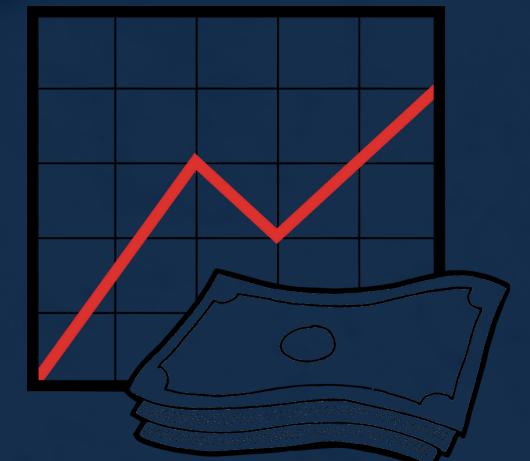
## Variables más influyentes:

- De acuerdo con el análisis de correlación y la interpretación de modelos:
- Volatilidad\_20 (volatilidad de 20 días).
- Signal (indicador técnico de señal).
- SMA\_30 (media móvil simple de 30 días).
- BB\_sup / BB\_inf (bandas de Bollinger).
- Rezagos (lag\_1, lag\_15, lag\_25).

## Modelos destacados:

Los algoritmos probados fueron:

- RandomForestRegressor, CatBoostRegressor, XGBoost, LGBM, ElasticNet y HuberRegressor.
- Ningún modelo logró un  $R^2$  positivo (señal de ruido en el problema).
- Mejores resultados (test):
  - MAE: 0.0537 – 0.0550
  - RMSE: 0.0663 – 0.0671
  - $R^2$ : -39.51 / -41.20



# HERRAMIENTAS

## Python (pandas | scikit-learn | statsmodels)

- Limpieza y preparación de series temporales
- Creación de variables derivadas (rezagos, medias móviles, volatilidad, indicadores técnicos).
- División 80/20 (entrenamiento / test) manteniendo la secuencia temporal.



## Modelos (RandomForest | CatBoost | XGBoost | LGBM | ElasticNet | Huber)

- Entrenamiento y validación de modelos lineales y no lineales.
- Comparación de métricas: MAE, RMSE, R<sup>2</sup>.
- Resultados: errores altos y R<sup>2</sup> negativos, reflejando la dificultad de predecir retornos a 20 días.



## Visualización (Matplotlib | Seaborn | SHAP)

- Mapas de calor de correlación entre predictores y target.
- Gráficos de dispersión y residuos para evaluar desempeño.
- Interpretabilidad con SHAP (beeswarm y waterfall plots).



## Salida a BI (Power BI / Excel)

- Exportación del dataset procesado para análisis adicional.
- Creación de dashboards interactivos en Power BI, conectando métricas financieras y visualizaciones para toma de decisiones.



# HALLAZGOS

## Variables con mayor impacto en los retornos

- Volatilidad\_20: variable más influyente en todos los modelos.
- Signal (indicador técnico derivado del MACD).
- SMA\_30 (media móvil de 30 días).
- Bandas de Bollinger (BB\_sup, BB\_inf).
- Variables de rezagos: lag\_25, lag\_15, lag\_1 también aportan, aunque en menor medida.

## Claves del preprocesamiento de datos

- Eliminación de la columna ABT para evitar fuga de datos (precio absoluto).
- Definición de target\_retorno\_20d como variable objetivo.
- División 80/20 temporal: entrenamiento y test sin mezclar el orden cronológico.
- Escalado con RobustScaler aplicado solo a modelos lineales.
- Para modelos no lineales (árboles, boosting): se mantuvieron valores originales sin escalar.

## Resultados de los modelos

- RandomForest, CatBoost, XGBoost, LGBM, ElasticNet, Huber fueron probados.
- Ningún modelo logró un  $R^2$  positivo, confirmando la dificultad de predecir retornos financieros a 20 días.
- Los modelos lineales con regularización (ElasticNet y Huber) ofrecieron los resultados menos malos (errores relativos menores que los demás).

# RECOMENDACIONES

## ⚡ Reconoce la complejidad de los retornos financieros

Los retornos a 20 días son altamente ruidosos y difíciles de predecir. Los modelos no lograron obtener  $R^2$  positivo, lo que refleja que la señal es débil frente al ruido del mercado.

👉 Beneficio: este aprendizaje muestra que la predicción de retornos financieros requiere enfoques más avanzados (ej. redes neuronales recurrentes o modelos híbridos con variables macroeconómicas).

## 📊 Complementa modelos con interpretación

Se utilizaron técnicas como SHAP para evaluar el impacto de las variables. Esto permitió confirmar que indicadores como Volatilidad\_20, Signal, SMA\_30 y Bandas de Bollinger tienen mayor relevancia en las predicciones.

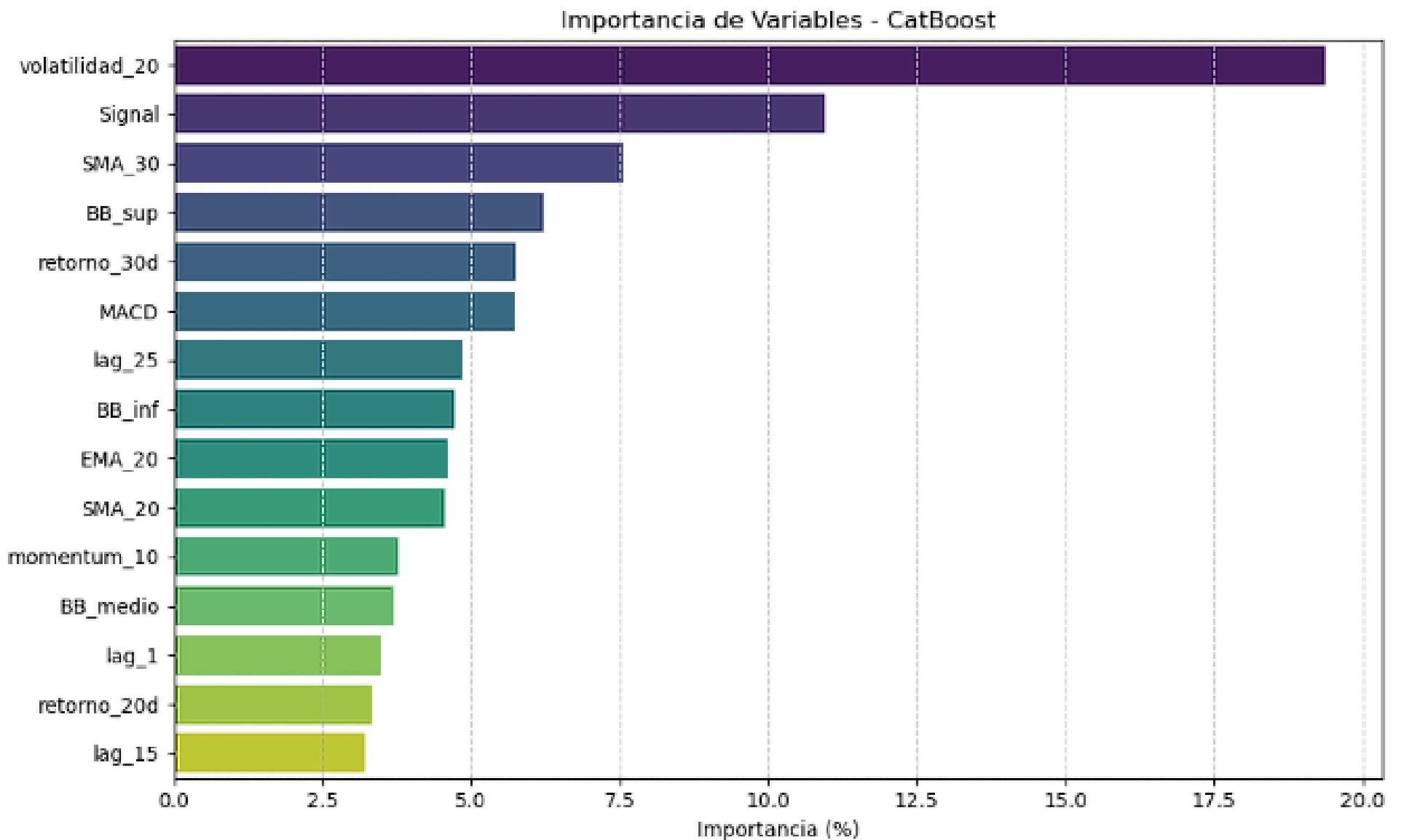
👉 Beneficio: aunque el ajuste fue limitado, se logró interpretar qué factores influyen más en los resultados, lo cual aporta valor al análisis financiero.

## 🔍 Usa los resultados como punto de partida, no como fin

Más que predecir retornos exactos, el modelo sirve para explorar la dinámica de los mercados y entender relaciones entre indicadores técnicos.

👉 Beneficio: convierte este pipeline en una base sólida para futuros proyectos que integren más datos (fundamentales, noticias, sentimiento, etc.) y técnicas más potentes.

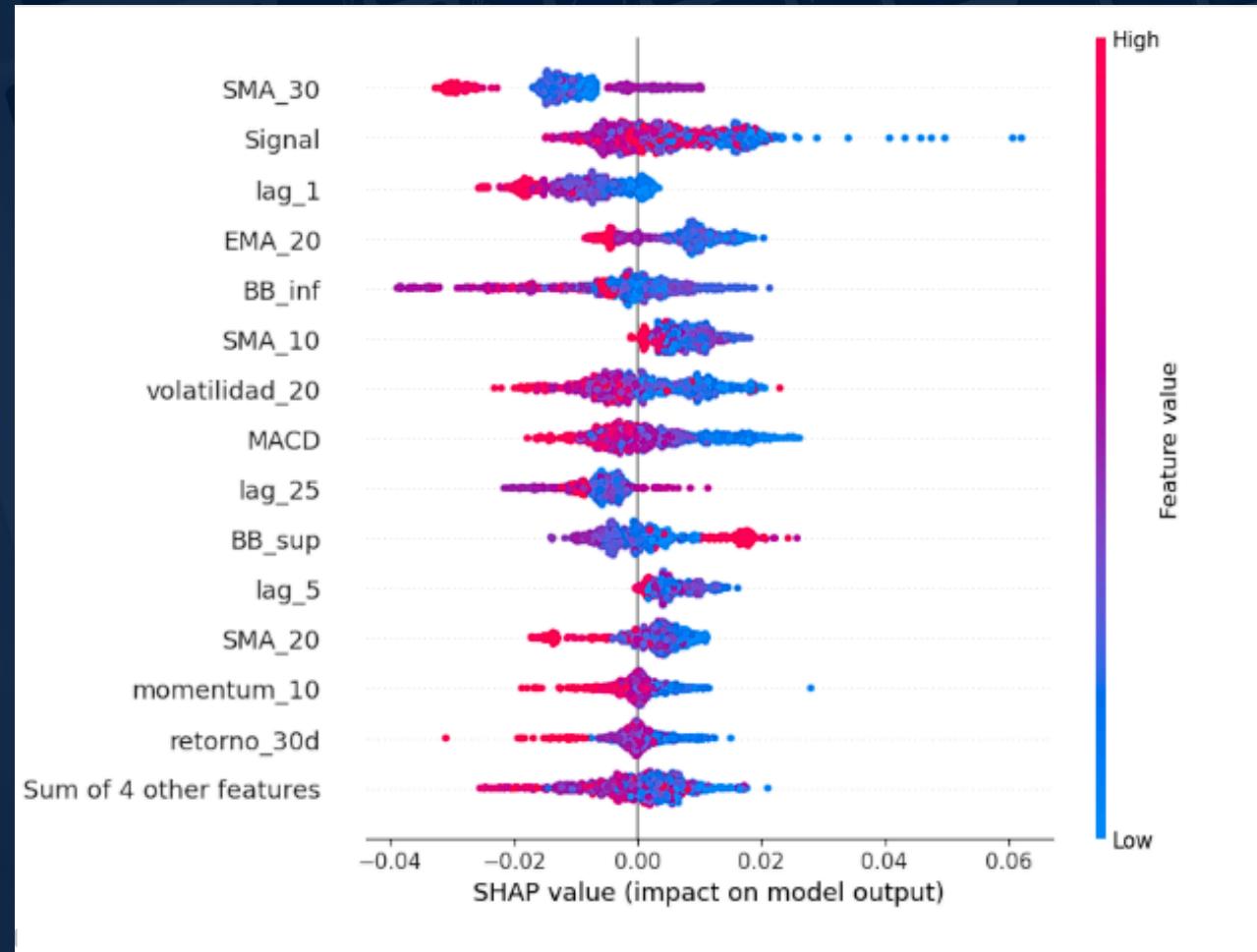




## 🔍 Interpretación de predicciones

- SMA\_30, Signal, EMA\_20 y BB\_inf fueron las variables con mayor influencia en las predicciones.
- Volatilidad\_20 y MACD también mostraron un aporte importante, confirmando su papel como indicadores técnicos clave.
- Variables de rezagos (lag\_25, lag\_15, lag\_1) contribuyen en menor medida, pero refuerzan la dinámica temporal de los retornos.
- El gráfico tipo beeswarm de SHAP evidencia cómo cada variable puede aumentar o reducir la predicción del retorno a 20 días, aunque el efecto global es bajo por el alto nivel de ruido.
- Esta explicación aporta transparencia y justificación en el proceso de modelado, mostrando qué factores financieros tienen mayor peso en las decisiones del modelo.

# ¿TODAS LAS VARIABLES IMPORTAN IGUAL?



No. El modelo muestra que algunas variables tienen un peso mucho mayor que otras en la predicción de los retornos financieros a 20 días.

- SMA\_30, Signal y EMA\_20 fueron las variables más influyentes, reflejando la importancia de las medias móviles y señales técnicas.
- Volatilidad\_20 y MACD también aportaron relevancia, capturando la dinámica del riesgo y el momentum.
- Variables de rezagos (lag\_1, lag\_15, lag\_25) ofrecen valor adicional, aunque su impacto es menor frente a los indicadores técnicos principales.
- El uso de SHAP permite visualizar qué factores suman o restan al retorno estimado, aportando transparencia y confianza en la interpretación del modelo.

# RESULTADOS ESPERADOS

Gracias al desarrollo del modelo de predicción de retornos financieros y al análisis exploratorio realizado, se espera lograr impactos positivos en las siguientes áreas clave:

## 🔍 Comprensión de la dinámica de retornos

El modelo ayuda a identificar qué indicadores técnicos influyen más en la predicción de retornos a 20 días, permitiendo un análisis menos sesgado y más basado en datos (ejemplo: medias móviles, volatilidad, MACD).



## Comunicación clara y visual

Los resultados se han traducido a insights visuales fáciles de interpretar mediante gráficas como heatmaps de correlación, gráficos de residuos, scatterplots y SHAP. Esto facilita la comprensión incluso para audiencias no técnicas.



## Apoyo a la toma de decisiones

Aunque los modelos no lograron un ajuste predictivo perfecto ( $R^2$  negativo y errores altos), la metodología aplicada es útil como flujo de trabajo replicable, aportando transparencia y destacando la dificultad de predecir retornos financieros en mercados ruidosos.

## Idea bonita

Uso de herramientas visuales avanzadas  
Implementé visualizaciones con heatmaps de correlación, gráficos de residuos y SHAP para interpretar relaciones entre indicadores financieros y retornos.

## Idea sencilla

Imputación de valores nulos  
Se reemplazaron valores faltantes en indicadores técnicos (ejemplo: medias móviles o RSI) con métodos simples para mantener la consistencia de la serie temporal.

## Idea creativa

Modelos no lineales de alto rendimiento  
Entrené modelos como RandomForest, CatBoost, XGBoost y LGBM, validando que aun con errores altos, sirven para explorar la dinámica de series financieras ruidosas.

## Idea simple

Codificación de variables categóricas  
Se crearon indicadores técnicos (SMA, EMA, volatilidad, momentum, bandas de Bollinger) para enriquecer el dataset y mejorar la capacidad predictiva.

## Idea brillante

Escalado de variables  
Se aplicó RobustScaler únicamente en los modelos lineales, manteniendo sin escalar los datos para los modelos no lineales (árboles y boosting).

## Idea original

Tratamiento de outliers  
Se detectaron y ajustaron retornos extremos que distorsionaban los modelos, lo cual redujo ruido y mejoró la estabilidad del entrenamiento.

# APRENDIZAJES