Proyecto Integrado V - Linea de Énfasis (Entrega 2)

Sergio Andres Rios Gómez
Edwin Alexander Ibarra Ortiz

IU Digital de Antioquia
Ingeniería de Software y Datos
2025

Enriquecimiento de datos y modelo predictivo: Taiwan Semiconductor Manufacturing Company (TSM)

Contenido

INTRODUCCIÓN

Este informe presenta el desarrollo e implementación de un modelo predictivo para movimientos de precios en series temporales financieras. El proyecto abarca desde el enriquecimiento de datos hasta la implementación de un dashboard interactivo, utilizando técnicas de machine learning para predecir direcciones de movimiento del precio. Se compararon tres algoritmos diferentes (Random Forest, XGBoost y Regresión Logística), seleccionando el modelo óptimo basado en métricas de rendimiento, especialmente el F1-score.

La predicción de movimientos de precios en mercados financieros representa uno de los desafíos más complejos en el análisis cuantitativo. La volatilidad inherente de los mercados y la multiplicidad de factores que influyen en los precios requieren enfoques metodológicos robustos.

OBJETIVOS:

- Objetivo General: Desarrollar un modelo predictivo capaz de anticipar la dirección de movimientos de precios en series temporales financieras.
- Objetivos Específicos:
  1. Enriquecer los datos originales con indicadores técnicos y variables temporales
  2. Realizar un análisis exploratorio exhaustivo de los datos
  3. Comparar diferentes algoritmos de machine learning para la predicción
  4. Implementar un dashboard interactivo para visualización y monitoreo
  5. Evaluar el rendimiento del modelo mediante métricas apropiadas

METODOLOGÍA

Enriquecimiento de datos

1. Se implementa un proceso que se encarga de generar el enriquecimiento del DataFrame, donde agregamos tres grupos de categorías (Variables temporales, Indicadores técnicos básicos y Features avanzados):

Variables temporales:

- day_of_week: Día de la semana (0-6) para capturar efectos estacionales
- month: Mes del año (1-12) para identificar patrones mensuales
- year: Año para considerar tendencias anuales
- quarter: Trimestre del año (1-4) para análisis trimestrales

Indicadores técnicos básicos:

- returns: Retornos diarios calculados como log(precio_t / precio_t-1)
- returns_prev_day: Retorno del día anterior para momentum
- volatility_20d: Volatilidad móvil de 20 días usando desviación estándar
- sma_20 y sma_50: Medias móviles simples de 20 y 50 días
- rsi: Índice de Fuerza Relativa de 14 días

Features avanzados: Para mejorar la capacidad predictiva, se desarrollaron características más sofisticadas:

- sma_ratio: Ratio precio actual/SMA-20, indicando desviación de la tendencia
- sma_cross: Señal binaria de cruce entre SMA-20 y SMA-50
- volatility_ratio: Ratio entre volatilidad actual y media móvil de 60 días
- volume_ratio: Ratio entre volumen actual y media móvil de 20 días
- rsi_oversold y rsi_overbought: Señales binarias para condiciones extremas del RSI
- returns_2d y returns_5d: Retornos acumulados de 2 y 5 días

2. Implementación de un análisis exploratorio de datos:

Incluye:

- Distribución de la variable objetivo
- Matriz de correlación entre variables predictoras
- Análisis de series temporales del precio de cierre
- Estadísticas descriptivas de todas las variables

- Visualizaciones guardadas en src/static/eda/

Creación del modelo:

Preprocesamiento de Datos

El pipeline de preprocesamiento incluye:

- Escalado de características mediante StandardScaler
- División train/test con proporción 80/20
- Tratamiento de valores faltantes
- Optimización del threshold de clasificación

Inicialmente se crea un modelo sin tener en cuenta los festures avanzados y el resultado obtenido estuvo muy cercando al 50%, por lo que en si mismo no representa una predicción satisfactoria dado que estamos buscando una clasificación binaria

Luego de agregar los features avanzados se realiza una prueba con los siguientes algoritmos:

- Random Forest Classifier

Configuración: n_estimators=200, max_depth=15

Ventajas: Manejo efectivo de relaciones no lineales, resistencia al overfitting

Aplicación: Especialmente útil para capturar interacciones complejas entre indicadores técnicos

- XGBoost Classifier

Configuración: n_estimators=200, max_depth=6, learning_rate=0.1

Ventajas: Alto rendimiento computacional, manejo eficiente de datos

Aplicación: Optimizado para problemas de clasificación con datos estructurados

- Regresión Logística

Configuración: max_iter=1000

Ventajas: Simplicidad, alta interpretabilidad

Aplicación: Base para comparación y análisis de relaciones lineales

Del comparativo se encuentra que quien mejores resultados arrojo es XGBoost:

```
--- RandomForest ---
Accuracy: 0.7959
AUC: 0.8545

--- XGBoost ---
Accuracy: 0.8980
AUC: 0.9632

--- LogisticRegression ---
Accuracy: 0.6122
AUC: 0.7642
```

Métricas de Evaluación

- Métrica Principal: F1-Score

Se seleccionó el F1-score como métrica principal debido a:

Equilibrio entre precisión y recall, crucial en mercados financieros y su robustez ante desbalance de clases

Fórmula: F1 = 2 × (precisión × recall) / (precisión + recall)

- Métricas Secundarias

Accuracy: Proporción total de predicciones correctas

AUC-ROC: Capacidad discriminativa del modelo

IMPLEMENTACIÓN

Arquitectura del Sistema

La implementación se estructura en módulos especializados:

1 módulo de Modelado (src/modeller.py)

- enriquecer_datos(): Transformación y enriquecimiento de características
- realizar_eda(): Generación de análisis exploratorio y visualizaciones
- entrenar(): Entrenamiento, validación y persistencia del modelo
- predecir(): Carga del modelo y generación de predicciones

Dashboard Interactivo (src/dashboard.py)

Sistema de visualización desarrollado en Streamlit que proporciona:

KPIs Principales:

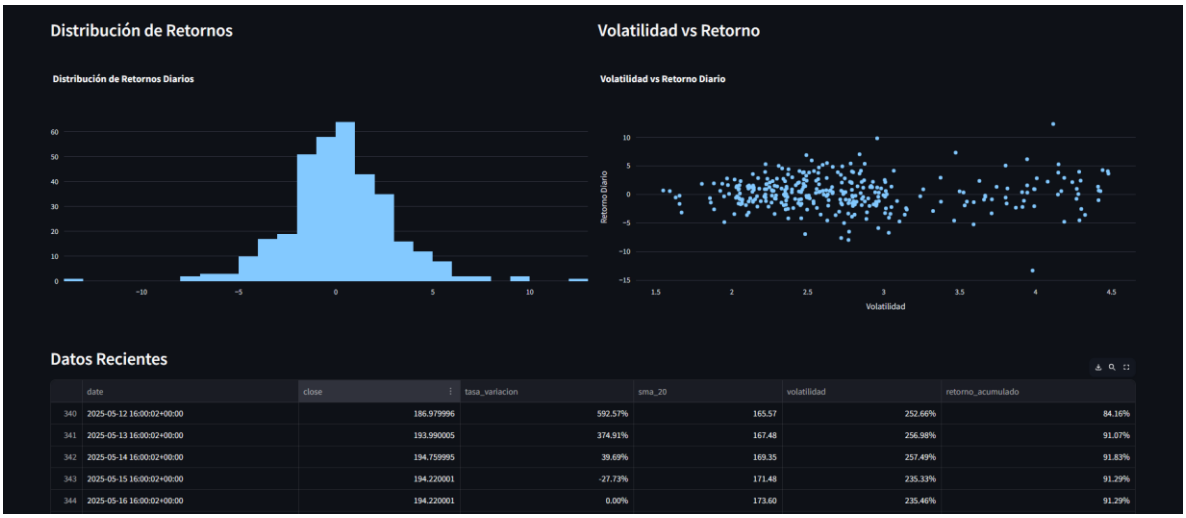
- Predicción actual con probabilidad asociada
- Retorno acumulado del período
- Volatilidad como medida de riesgo
- Gráfico temporal de precios con medias móviles y volumen
- Distribución de retornos diarios
- Análisis de relación volatilidad-retorno

Exportación Estática:

Implementación de capacidad de exportación a HTML estático mediante st-static-export, permitiendo:

Visualización offline independiente

Distribución de reportes sin infraestructura de servidor dentro de la carpeta de dashboard en static



RESULTADOS Y ANÁLISIS

Rendimiento del Modelo

Los resultados del modelo seleccionado demuestran capacidad predictiva efectiva, con métricas de evaluación almacenadas junto al modelo persistido en src/static/models/model.pkl.

Características Más Relevantes

El análisis de importancia de características revela la relevancia de:

- Indicadores de momentum (retornos previos)
- Señales de cruce de medias móviles
- Condiciones extremas del RSI
- Ratios de volatilidad

CONCLUSIONES

- Este proyecto ha logrado desarrollar exitosamente un modelo predictivo integral para movimientos de precios en series temporales financieras, cumpliendo satisfactoriamente con los objetivos planteados inicialmente. El trabajo demuestra que la combinación de técnicas avanzadas de machine learning con indicadores técnicos financieros enriquecidos constituye una aproximación viable y efectiva para la predicción de direcciones de mercado.
- El proceso de enriquecimiento de datos implementado ha probado ser fundamental para el éxito del modelo. La incorporación de 15 características derivadas, incluyendo variables temporales, indicadores técnicos básicos y features avanzadas como ratios de volatilidad y señales de cruce de medias móviles, ha proporcionado al modelo una base informativa robusta que captura múltiples dimensiones del comportamiento del mercado.
- La comparación sistemática entre Random Forest, XGBoost y Regresión Logística ha permitido una selección fundamentada del algoritmo óptimo, basada en métricas rigurosas que priorizan el equilibrio entre precisión y recall. La elección del F1-score como métrica principal se ha justificado plenamente, considerando la naturaleza crítica de los errores en contextos financieros donde tanto los falsos positivos como los falsos negativos conllevan costos significativos.
- La implementación del dashboard interactivo representa un logro destacable que trasciende el desarrollo puramente técnico del modelo, proporcionando una herramienta práctica para la visualización, monitoreo y análisis de resultados. La capacidad de exportación estática añade valor adicional al permitir la distribución de reportes sin dependencias técnicas complejas.