Financial Regression: Predecir Costo del Oro

Rodrigo Antonio Benítez De La Portilla

A01771433@tec.mx

Abstract: Este estudio presenta la implementación y evaluación de dos modelos de Inteligencia Artificial para predecir el costo de cierre del oro, utilizando el dataset "m15". El modelo estima los coeficientes asociados a variables clave como precio de apertura, máximo diario y precios históricos de cierre, con el objetivo de generar proyecciones precisas que faciliten decisiones de inversión más informadas, identificando momentos óptimos de compra y venta. El modelo corresponde a un algoritmo implementado con el framework **PySpark**. Los resultados muestran que el modelo alcanza un coeficiente de determinación (R^2) cercano a 1.00 en el conjunto de prueba, lo que evidencia su capacidad para explicar alrededor del 100% de la variabilidad en los precios del oro. En cuanto al error cuadrático medio (MSE) el modelo obtuvo un valor de 0.572. Aunque presenta un desempeño sobresaliente, el resultado del error cuadrático medio (MSE) sugiere que el modelo todavía puede acercarse más a la realidad, esto pudiera mejorarse con mejores datos.

Keywords: Oro, Inteligencia Artificial, predicción, Regresión Lineal Multivariable, Gradiente Descendiente, PySpark

INTRODUCCIÓN

En los últimos años las inversiones, trading y criptomonedas han evolucionado el mercado haciendo que día con día más personas inicien en este tipo de mercado a hacer inversiones, este término de inversión se remonta a la Antigüedad. Desde los contratos primitivos en Mesopotamia y Egipto hasta la consolidación de conceptos modernos en Europa entre los siglos XVI y XVII con el auge del comercio marítimo y la creación de fideicomisos y sociedades por comerciantes holandeses, la inversión ha evolucionado hacia enfoques cada vez más cuantitativos. En ese espíritu, este trabajo explora técnicas de aprendizaje automático para estimar el precio de cierre del oro a partir de variables de mercado.

El oro, a menudo considerado activo refugio, concentra ciclos de volatilidad asociados a expectativas macroeconómicas, tipos de interés y fortaleza del dólar [1]. Anticipar su comportamiento facilita decisiones tácticas y de cobertura en portafolios. Para ello, se emplea un conjunto de datos estructurado "m15.csv" con variables de apertura, máximo y mínimo, orientadas a predecir el precio de cierre.

El flujo de trabajo incluye un ETL básico con eliminación de nulos y normalización basada en estadísticas del conjunto de entrenamiento y una partición 80/20 en entrenamiento y prueba. El desempeño se mide con **MSE y R^2**, y se visualizan tanto las predicciones frente a valores reales.

El objetivo es construir una línea base clara, reproducible y explicable para pronóstico de precios del oro, que sirva como punto de partida para iteraciones futuras. Entre las extensiones recomendadas se incluyen validación temporal (para evitar fuga de información), incorporación de rezagos y ventanas móviles, y variables exógenas (índice del dólar, tasas, inflación, volatilidad implícita). Asimismo, se sugiere contrastar regularización, modelos de árboles (Random Forest, Gradient Boosting) y enfoques específicos de series temporales.

En conjunto, este proyecto conecta la tradición histórica de la inversión con herramientas modernas de análisis, mostrando cómo métodos lineales y ensambles pueden apoyar la toma de decisiones en mercados financieros sensibles a la información y al tiempo.

1. DESCRIPCIÓN DEL DATASET

El data set *"m15"* proporciona un registro detallado de los costos de diferentes metales preciosos, combustibles, aceites y otros tipos de químicos.
La parte que se analiza es la del oro, con los precios de apertura, máximos,

cierre y volumen. Los datos abarcan desde el 2002 hasta el 2024 y están organizados en registros de horarios, sumando un total de 419,805 instancias y un peso total del archivo de **1.05 GB**.

2a.- Features.

El data set incluye 5 columnas indispensables para la predicción del costo de cierre del oro. A continuación, se presentan las características disponibles:

- **Date**: La fecha en formato YYYY-MM-DD, representando el día específico de registro
- **xauusd open**: El costo de apertura en USD y dos decimales.
- **xauusd high**: El costo más alto en USD y dos decimales.
- **xauusd low**: El costo más bajo en USD y dos decimales.
- **xauusd close**: El costo de cierre en USD y dos decimales.

2. EXTRACCIÓN Y LIMPIEZA DE DATOS.

La extracción de datos implica recuperar información de fuentes para su análisis [2]. En este estudio se utilizó el archivo “m15.csv” y se aplicaron pasos de preprocesamiento para entrenar y evaluar modelos de predicción del precio de cierre del oro.

La transformación de datos prácticamente es cambiar los valores de las columnas numéricas del conjunto de datos para usar una escala común, sin distorsionar las diferencias en los intervalos de valores ni perder información, en este caso todas las columnas excepto la de la fecha, se normalizaron a tipo de dato double.

3a.- Carga y filtrado de datos.

Se cargó el dataset y se eliminaron las filas con valores faltantes mediante dropna, conservando únicamente observaciones completas.
La columna “date” no se empleó como característica predictora; solo conserva el orden del archivo. Las columnas para features, se normalizaron para evitar algún error, se hicieron tipo double.

3. CONSTRUCCIÓN Y EVALUACIÓN DEL MODELO

3a.- Selección de variables predictoras y objetivo.

Las variables predictoras para entrenar al modelo son: “xauusd open”, “xauusd high”, “xauusd low”. Mientras que el valor que queremos predecir o la variable objetivo es: “xauusd close”. El resto de las columnas no se utilizaron para esta línea base, la unidad exacta depende de la fuente original, pero en el CSV

aparece como conteo negociaciones a gran escala.

3b.- Cálculo del error

El error del modelo cuantifica qué tan lejos están las predicciones de la realidad y guía todo el proceso de aprendizaje [3].
La métrica elegida es el Mean Squared Error (MSE), calculado sobre los datos normalizados: para cada muestra se obtiene el residuo como la diferencia entre la predicción y el valor real, se eleva al cuadrado, se suman todos los residuos y se promedian dividiendo entre el número de ejemplos la fórmula es la siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

Ecuación 1. *MSE*

En el código, esto se implementa en la función mse.
Minimizar este MSE es el objetivo del entrenamiento.

3c.- Modelado y evaluación.

En la evaluación del rendimiento para el modelo, se ocupó la métrica de coeficiente de determinación ($R2$), este coeficiente nos explica cuanto porcentaje llega a poder explicar nuestro modelo, si tenemos un coeficiente de 1.0, nuestro modelo es capaz de explicar el 100% de las predicciones con alta seguridad.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Ecuación 2. *Coeficiente de determinación*

Dónde:

- R^2: Coeficiente de determinación
- RSS: Suma de cuadrados de los residuos
- TSS: Suma total de cuadrados

3d.- Entrenamiento y Prueba.

Para evaluar y potencialmente mejorar la capacidad predictiva de nuestro modelo de regresión en datos financieros, dividimos el conjunto de datos en tres segmentos: entrenamiento (80%) y prueba (20%). Esta división es fundamental en el aprendizaje automático, ya que utilizar los mismos datos para entrenar y evaluar el modelo puede provocar un sesgo significativo, donde el algoritmo memoriza los ejemplos en lugar de aprender patrones generalizables, resultando en predicciones pobres para datos nuevos. El conjunto de entrenamiento se utiliza para optimizar los parámetros.
El conjunto de prueba permanece completamente aislado hasta que el

entrenamiento finaliza. Solo entonces lo utilizamos para obtener una evaluación imparcial y definitiva del rendimiento del modelo, calculando tanto el MSE como el coeficiente de determinación (R^2) para determinar la precisión y exactitud de nuestras predicciones del precio de cierre del oro. Nuestro modelo calcula sistemáticamente dos métricas fundamentales MSE (error cuadrático medio) y R^2 (coeficiente de determinación) para los conjuntos de entrenamiento y prueba. El MSE cuantifica la diferencia promedio al cuadrado entre los precios de cierre del oro predichos y los valores reales, mientras que el R^2 indica la proporción de varianza en la variable dependiente que el modelo explica. Ambas métricas se implementan como funciones independientes en nuestro código y se visualizan mediante gráficos de barras para comparar el rendimiento entre los dos conjuntos. Nuestro enfoque de división del dataset, busca mantener un equilibrio óptimo entre los conjuntos para detectar y prevenir el sobreajuste, permitiéndonos monitorizar visualmente la evolución de ambos errores en una gráfica temporal. Si el modelo comenzara a memorizar los datos de entrenamiento en lugar de aprender patrones generalizables, observaríamos una divergencia entre estas curvas, el error de entrenamiento continuaría disminuyendo mientras el error de prueba aumentaría o se estancaría.

El monitoreo constante de estas métricas en los dos conjuntos es crucial para nuestro modelo financiero. Un rendimiento desequilibrado (por ejemplo, R^2 alto en entrenamiento, pero bajo en prueba) indicaría que nuestro modelo podría fallar al predecir precios futuros del oro en escenarios reales de mercado. Las visualizaciones que implementamos, especialmente la comparación de errores finales y scores R^2 entre los tres conjuntos, nos proporcionan una herramienta efectiva para evaluar si estamos logrando un equilibrio adecuado entre aprender de los datos históricos y mantener la capacidad de generalizar a nuevos datos del mercado de oro.

4. VISUALIZACIÓN DE RESULTADOS DEL CASO BASE

Después de pasar por el entrenamiento y prueba y realizar las pruebas pertinentes, los resultados muestran que el modelo alcanzó una exactitud para explicar las predicciones (R^2) con el siguiente valor:

- R^2: 1.00
- MSE: 0.5726

Este resultado es muy bueno, pues sugiere que el modelo es capaz de

capturar aproximadamente el 100% de la variabilidad de los datos.

Los valores del coeficiente R^2 en los conjuntos de entrenamiento, validación y prueba son extremadamente similares, lo que indica que el modelo no está experimentando overfitting. Esta consistencia en los datos entre los tres conjuntos de datos sugiere que el modelo es estable y no está reaccionando de manera excesiva a las particularidades del conjunto entrenamiento, lo que indica una varianza baja. Por otro lado, el modelo, tampoco está presentando un caso de underfitting, pues el modelo está no está haciendo suposiciones simplificadas sobre las relaciones entre los datos, lo que lleva a un ajuste eficiente.

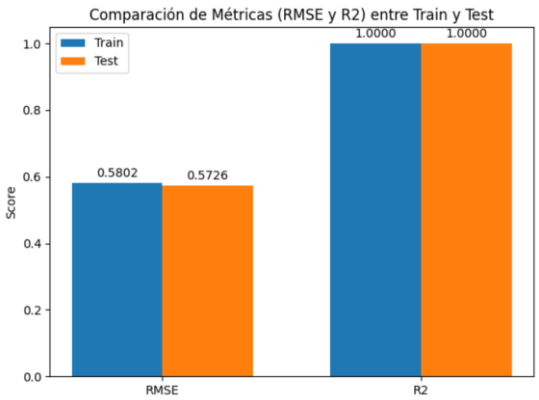

Figura 1. *Métricas entre train y test*.

La gráfica nos indica como el error empieza alto y conforme pasa el entrenamiento y las épocas este disminuye considerablemente, hasta que llega un momento en dónde el modelo a encontrado un buen balance para los datos. variables, calculando primero la media y desviación estándar de cada característica.

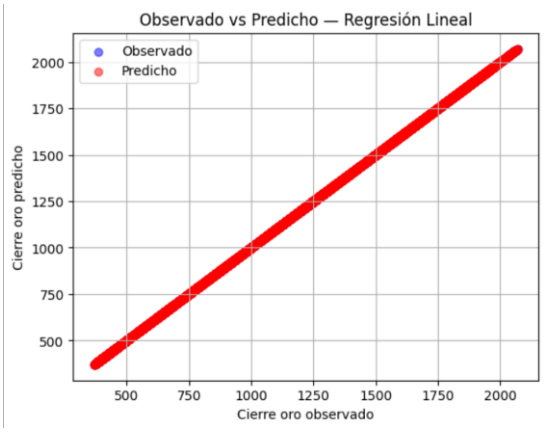

Figura 2. *Comparación de resultados*.

Cómo podemos observar la gráfica nos muestra que los valores que el modelo predijo van muy similares a los valores reales, y es por eso que no hay ninguna línea azul pues los datos predichos coinciden con los observado

REFERENCIAS

1. A Brief History of Investments | Indian River Financial Group, Inc. (2016). Paulmilleradvisor.com. https://www.paulmilleradvisor.com/blog/brief-history investments#:~:text=La%20inversi%C3%B3n%20comenz%C3%B3%20en%20el,largo%20plazo%20en%20el%20extranjero

2. Mohamad. (2017, March 9). MSE – Error Cuadrático Medio. Centro de Ayuda. https://support.numxl.com/hc/es/articles/115001223423-MSE-Error Cuadr%C3%A1tico-Medio

3. Team, A. A. (2023, April 27). ¿Qué es la extracción de datos? Tipos, usos y beneficios | Astera. Astera. https://www.astera.com/es/type/blog/ what-is-data-extraction-a-brief-guide/