

Challenge 3 Avanzado

Spark ML para construir modelos predictivos, relacionado a mi tema de tesis “Desarrollo de un Asistente Financiero Inteligente para la Optimización de Decisiones de Inversión en el Mercado de Acciones mediante Análisis de Sentimientos y Datos Históricos utilizando la Inteligencia Artificial o1 de OpenAI”

Alumno: Fernando Arturo Arevalo Perez

Código de Alumno: 323018942

MCD

24 de noviembre del 2024

2. Índice

- **Contenido:**

1. Introducción
 2. Objetivos del Proyecto
 3. Descripción de la Base de Datos
 4. Preparación y Limpieza de Datos
 5. Selección y Transformación de Características
 6. División del Conjunto de Datos
 7. Construcción de Modelos Predictivos con Spark ML
 8. Resultados del Modelo de Regresión Lineal
 9. Resultados del Modelo Random Forest Regressor
 10. Comparación de Modelos
 11. Conclusiones
 12. Recomendaciones y Trabajos Futuros
 13. Preguntas y Respuestas
-



3. Introducción

- Contexto:** Importancia del análisis predictivo en el mercado financiero.
 - Relevancia:** Cómo la predicción precisa del precio de cierre puede influir en decisiones de inversión.
 - Tecnología Utilizada:** Introducción a **Spark ML** y sus ventajas en el manejo de grandes volúmenes de datos.
-



4. Objetivos del Proyecto

- **Objetivo General:** Desarrollar modelos predictivos para estimar el precio de cierre de acciones de empresas líderes utilizando Spark ML.

- **Objetivos Específicos:**

- Cargar y preparar los datos financieros.
 - Seleccionar y transformar características relevantes.
 - Construir y evaluar modelos de Regresión Lineal y Random Forest.
 - Comparar el desempeño de los modelos y extraer conclusiones.
-



5. Descripción de la Base de Datos

•**Nombre del Archivo:** datos_acciones.csv




•**Empresas Incluidas:**

- Apple Inc. (AAPL)
- Microsoft Corporation (MSFT)
- Amazon.com, Inc. (AMZN)
- Alphabet Inc. (GOOGL)
- Meta Platforms, Inc. (META)

•**Estructura del Archivo:**

- **Date:** Fecha de la transacción (Formato: YYYY-MM-DD)
- **Ticker:** Símbolo bursátil de la empresa
- **Attribute:** Tipo de atributo financiero (Open, Close, Volume, etc.)
- **Value:** Valor correspondiente al atributo



| |  Date |  Ticker |  Attribute | 1.2 Value |
|-----|--|--|---|---------------------|
| 503 | 2021-12-29 | AAPL | Adj Close | 176.683166503906... |
| 504 | 2021-12-30 | AAPL | Adj Close | 175.5208740234375 |
| 505 | 2020-01-02 | AMZN | Adj Close | 94.90049743652344 |
| 506 | 2020-01-03 | AMZN | Adj Close | 93.74849700927734 |
| 507 | 2020-01-05 | AMZN | Adj Close | 95.41886740830884 |

6. Preparación y Limpieza de Datos

- **Conversión de Tipos de Datos:** Asegurar que las columnas tengan los tipos de datos correctos (ej. Date como fecha, Value como float).
- **Manejo de Valores Nulos:** Eliminación de filas con valores faltantes para garantizar la integridad de los datos.
- **Pivot de Datos:** Transformar el DataFrame para tener atributos como columnas, facilitando el análisis.

7. Selección y Transformación de Características

- **Características Seleccionadas:** Open, High, Low, Volume
- **Etiqueta:** Close (Precio de cierre)
- **Técnicas Utilizadas:**
 - **VectorAssembler:** Para ensamblar las características en un solo vector.
 - **StandardScaler:** Para escalar las características y mejorar el rendimiento del modelo.

8. División del Conjunto de Datos

- **Método de División:** Random Split (80% entrenamiento, 20% prueba)
- **Justificación:** Balancear la cantidad de datos para entrenamiento y evaluación, asegurando representatividad.

```
Número de filas en entrenamiento: 2068  
Número de filas en prueba: 452
```


9. Construcción de Modelos Predictivos con Spark ML

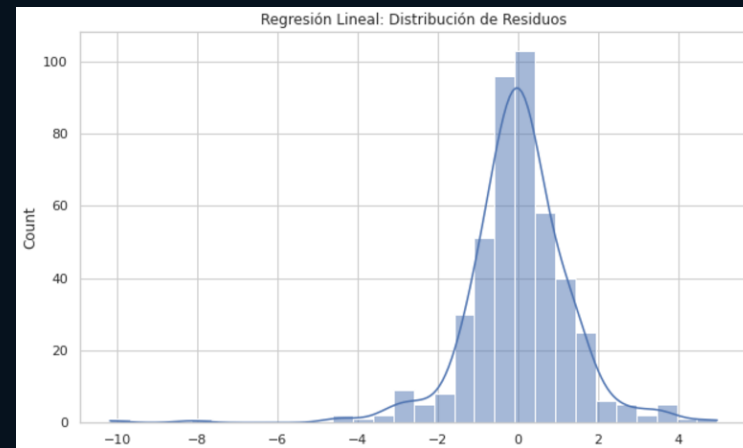
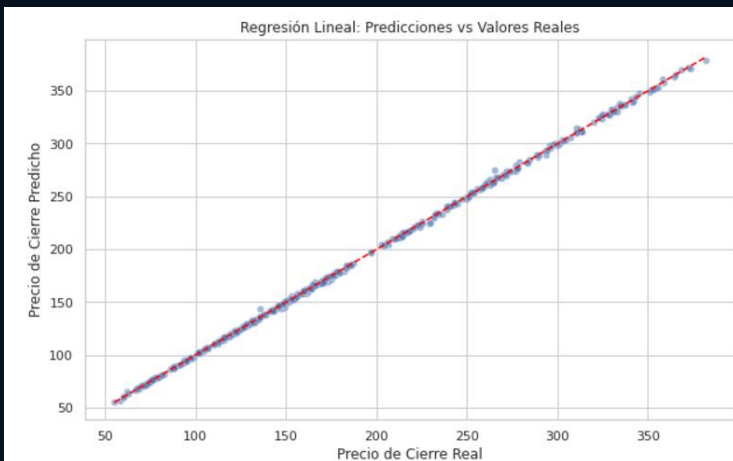
- **Modelos Utilizados:**

- **Regresión Lineal:** Modelo sencillo para relaciones lineales.

- **Random Forest Regressor:** Modelo más complejo capaz de capturar relaciones no lineales.

- **Configuración del Pipeline:** Integración de transformación de características y entrenamiento del modelo en un pipeline cohesivo.

| | Date | Ticker | 1.2 Close | 1.2 prediction |
|----|------------|--------|---------------------|---------------------|
| 1 | 2020-01-02 | GOOGL | 68.43399810791016 | 68.20983063279137 |
| 2 | 2020-01-03 | AMZN | 93.74849700927734 | 94.11354101770385 |
| 3 | 2020-01-03 | META | 208.6699981689453 | 209.563636856137... |
| 4 | 2020-01-06 | META | 212.600006103515... | 211.393781431441... |
| 5 | 2020-01-07 | MSFT | 157.5800018310547 | 158.102744681646... |
| 6 | 2020-01-08 | META | 215.220001220703... | 215.290660173830... |
| 7 | 2020-01-09 | MSFT | 162.089996337890... | 161.559566517294... |
| 8 | 2020-01-13 | AAPL | 79.23999786376953 | 78.90020351974532 |
| 9 | 2020-01-15 | AAPL | 77.83499908447266 | 78.25375850963206 |
| 10 | 2020-01-15 | AMZN | 93.10099792480469 | 93.25530942031982 |
| 11 | 2020-01-15 | GOOGL | 71.95999908447266 | 71.96982397105032 |
| 12 | 2020-01-15 | MSFT | 163.179992675781... | 163.664964923882... |
| 13 | 2020-01-16 | AMZN | 93.89700317382812 | 93.64357975727162 |
| 14 | 2020-01-17 | AAPL | 79.68250274658203 | 79.31525542766418 |
| 15 | 2020-01-21 | GOOGL | 74.11250305175781 | 74.0923162052441 |



10. Resultados del Modelo de Regresión Lineal

•RMSE: 1.3438

• R^2 : 0.9997

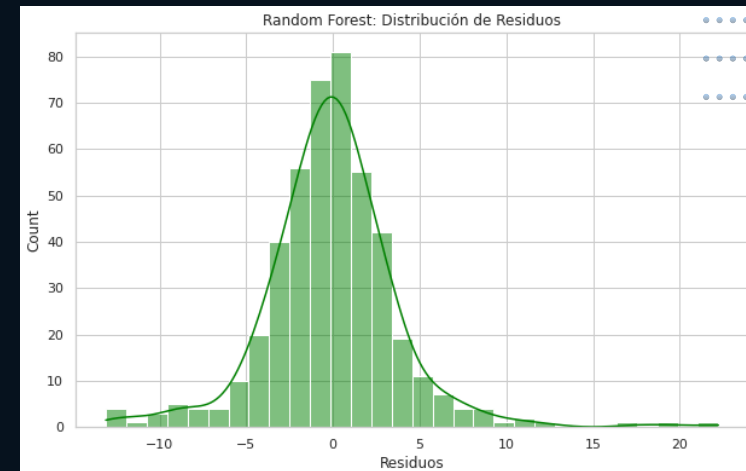
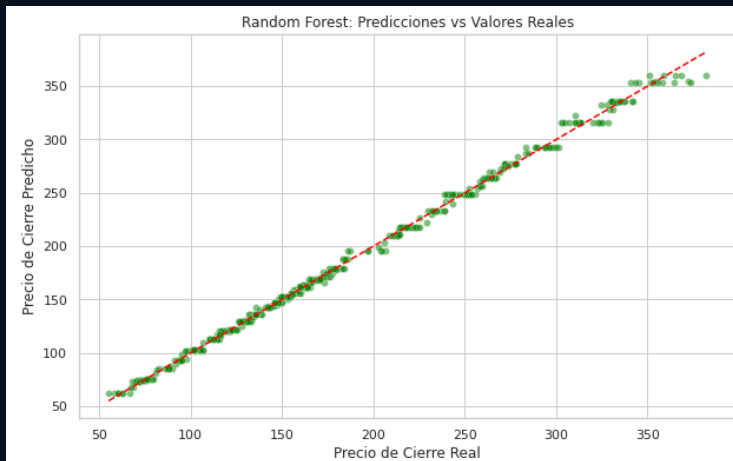
•Interpretación:

- RMSE Bajo: Alta precisión en las predicciones.
- R^2 Cercano a 1: El modelo explica casi toda la variabilidad en el precio de cierre.

Resultado

El modelo de Regresión Lineal ha mostrado un desempeño excepcional, con un RMSE muy bajo y un R^2 cercano a 1. Esto sugiere que el modelo es altamente preciso en la predicción del precio de cierre de las acciones y explica casi toda la variabilidad observada en los datos. La simplicidad y la interpretabilidad de la regresión lineal han sido ventajosas en este contexto, proporcionando resultados claros y confiables.

| | 📅 Date | 📈 Ticker | 1.2 Close | 1.2 prediction |
|----|------------|----------|---------------------|---------------------|
| 1 | 2020-01-02 | GOOGL | 68.43399810791016 | 67.29238108587458 |
| 2 | 2020-01-03 | AMZN | 93.74849700927734 | 92.94757013212997 |
| 3 | 2020-01-03 | META | 208.6699981689453 | 210.155748402058... |
| 4 | 2020-01-06 | META | 212.600006103515... | 210.155748402058... |
| 5 | 2020-01-07 | MSFT | 157.5800018310547 | 158.9027527657248 |
| 6 | 2020-01-08 | META | 215.220001220703... | 217.232283925879... |
| 7 | 2020-01-09 | MSFT | 162.089996337890... | 162.811547803445... |
| 8 | 2020-01-13 | AAPL | 79.23999786376953 | 75.48666506936904 |
| 9 | 2020-01-15 | AAPL | 77.83499908447266 | 75.48666506936904 |
| 10 | 2020-01-15 | AMZN | 93.10099792480469 | 92.94757013212997 |
| 11 | 2020-01-15 | GOOGL | 71.95999908447266 | 73.96333066014977 |
| 12 | 2020-01-15 | MSFT | 163.179992675781... | 163.272526581245... |
| 13 | 2020-01-16 | AMZN | 93.89700317382812 | 92.94757013212997 |
| 14 | 2020-01-17 | AAPL | 79.68250274658203 | 75.48666506936904 |
| 15 | 2020-01-21 | GOOGL | 74.11250305175781 | 74.68079481907854 |



11. Resultados del Modelo Random Forest Regressor

•RMSE: 3.8549

•R²: 0.9977

•Interpretación:

- **RMSE Mayor que Regresión Lineal:** Menor precisión en comparación con la regresión lineal.
- **R² Alto:** El modelo aún explica una gran parte de la variabilidad, pero no tan cerca de 1 como la regresión lineal.

Resultados:

el modelo de Random Forest Regressor también ha demostrado un rendimiento sólido, con un RMSE ligeramente mayor pero aún significativamente bajo, y un R² alto. Aunque este modelo es más complejo y puede capturar relaciones no lineales en los datos.

12. Comparación de Modelos

- **Regresión Lineal vs Random Forest:**

- **RMSE:** Regresión Lineal < Random Forest (1.3438 vs 3.8549)
- **R²:** Regresión Lineal > Random Forest (0.9997 vs 0.9977)

- **Conclusión:**

- **Regresión Lineal Supera a Random Forest:** En este caso específico, la Regresión Lineal ha superado al Random Forest en términos de precisión predictiva. Esto podría deberse a la naturaleza lineal de la relación entre las características seleccionadas y el precio de cierre, haciendo que un modelo lineal sea más adecuado para este conjunto de datos.

13. Conclusiones

- **Eficacia de Spark ML:** Spark ML facilitó la construcción y evaluación de modelos predictivos de manera eficiente y escalable.
- **Desempeño de Modelos:**
 - **Regresión Lineal:** Excelente precisión y capacidad explicativa para la predicción del precio de cierre.
 - **Random Forest Regressor:** Buen desempeño, pero inferior a la regresión lineal en este caso específico.
- **Implicaciones para la Tesis:** Los modelos desarrollados proporcionan herramientas robustas para la predicción financiera, con potencial para integrarse en estrategias de inversión.
- Este challenge ha demostrado que Spark ML es una herramienta poderosa para el modelado predictivo en el ámbito financiero, permitiendo construir modelos precisos y eficientes. Los excelentes resultados obtenidos con la Regresión Lineal destacan la importancia de seleccionar modelos alineados con la naturaleza de los datos. Continuar explorando y refinando estos modelos fortalecerá aún más la capacidad predictiva y aportará valiosos insights para mi investigación de tesis.