Modelo Predictivo para las Acciones del Índice S&P 500

Gabriela Zumarraga

Business Understanding

Contexto

S&P 500 - Standard & Poor’s 500

Índice del rendimiento de las 500 mayores empresas que cotizan en Estados Unidos

Indicador clave de la salud económica de USA. Abarca distintos tipos de industrias, como tecnología, atención médica, servicios financieros, bienes de consumo y energía, seleccionadas por su tamaño, liquidez y representatividad económica.

OBJETIVO

Identificar patrones relevantes en los precios que puedan ayudar a entender mejor el comportamiento de los precios de las acciones en el mercado para anticipar movimientos futuros.

VISION

1. **Tomar decisiones** informadas sobre la compra o venta de acciones para mejorar sus rendimientos.

2. **Gestión de riesgos:** identificar patrones que podrían reducir el riesgo asociado con las inversiones.

3. **Optimizar estrategias:** Facilita la creación de estrategias de trading automatizadas basadas en predicciones de movimientos del mercado.

ALCANCE

- EDA
- Data Wrangling
- Probar diversos modelos predictivos para identificar cuál es el más adecuado para predecir si el precio de una acción del índice S&P 500 subirá o bajará en un día. Esto se logrará mediante el análisis y evaluación de las métricas de rendimiento obtenidas de cada modelo, seleccionando finalmente el modelo con el mejor desempeño para resolver el problema planteado.

Data Understanding

Fuente de datos

kaggle

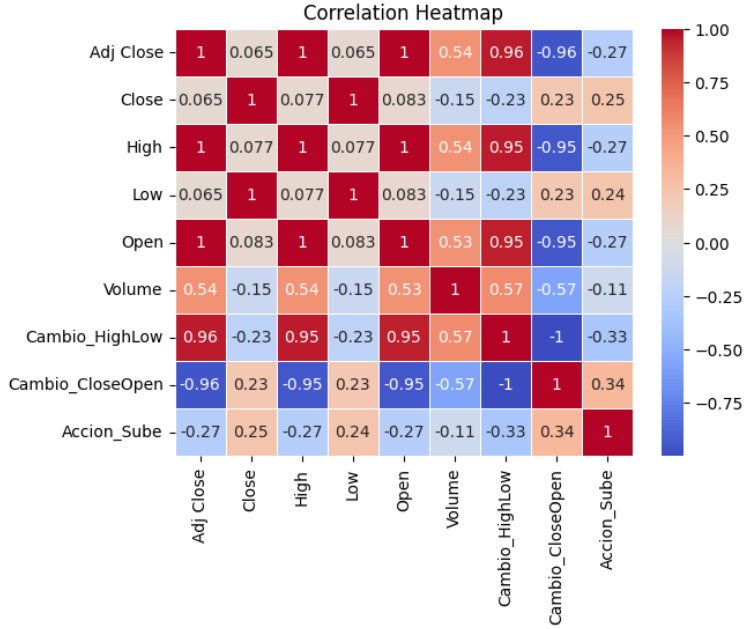
S&P 500 - Standard & Poor’s 500 (2010-2024)

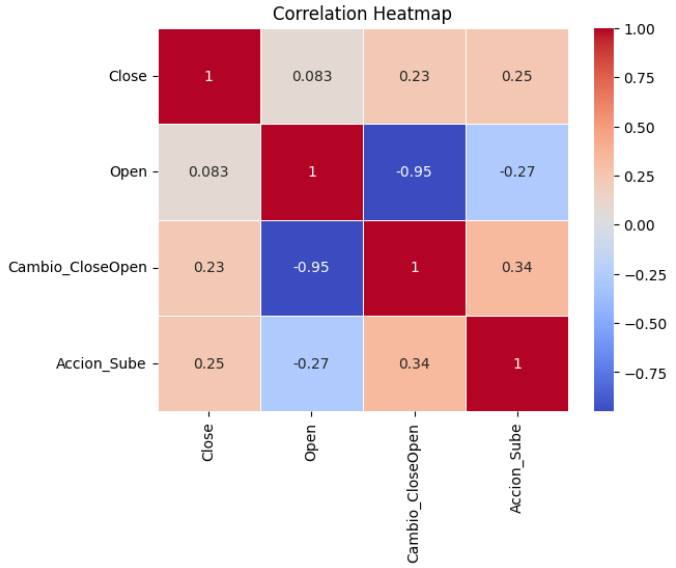
Precios históricos de acciones para las empresas que se encuentran en el índice S&P 500.

Data Preparation

Normalización de datos

- **Duplicados:** No se encontraron.
- **Valores Nulos:** 67.33% imputados con días cercanos.
- **Transformación:** Ajuste de tipos de datos.
- **Columnas Eliminadas:** Alta correlación y strings irrelevantes.
- **Variables Derivadas:** Columna CambioOpenClose y Columna binaria que determina si sube o baja la accion.
- **Variables Lag:** 5, 10 y 20 días.





Limitaciones

- Variables lag: incremento en dimensionalidad puede limitar el probar varios modelos ya que requiere de mayor poder de procesamiento.

Modeling

Modelos

Problema de clasificación

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- VotingClassifier

- Iteración entre los modelos y diferentes parámetros.
- Se considero el mejor modelo el que tiene mejor precisión

Métricas

- Precision
- Recall
- F1

- Minimizar falsos positivos
- Evaluar cuántos de los casos predichos como positivos son realmente positivos

Resultados

Mejor modelo: Regresion Logistica y Random Forest

Lag 20 dias

```
Mejor modelo: RandomForestClassifier
Hiperparámetros: {'n_estimators': 100}
Exactitud: 0.8280535447669538
Recall: 0.5783936993759873
Precision: 0.5463569501031801
F1: 0.5619190665883019
```

Lag 10 dias

```
Mejor modelo: LogisticRegression
Hiperparámetros: {'max_iter': 500}
Exactitud: 0.8218496620900011
Recall: 0.5067385142780323
Precision: 0.5346184757942511
F1: 0.5203052838023592
```

Precision esperada – 70%

- Se obtuvo una precisión de 54%
- El modelo logra predecir un 54% de acciones que suben. De ese porcentaje solo el 57% es acertado.

Costo del error

- ¿Qué costo tiene actuar pensando que una acción subirá cuando en realidad baja? (Compra equivocada, pérdida de capital).
- ¿Qué implica perder la oportunidad de invertir cuando la acción realmente sube? (Costos de oportunidad).

Beneficios

- Incremento en rendimiento del portafolio.
- Reducción de pérdidas al evitar decisiones basadas en predicciones incorrectas.
- Maximización de oportunidades al identificar con mayor certeza las acciones que suben.

Proximos pasos

- Considerar tunear mejor los modelos para encontrar una mejor precisión o considerar otros modelos
- Considerar tendencias de mas largo plazo (no solo un día) para tomar decisiones

Dspliegue

- Contenerizar el modelo
- Arquitectura de microservicios y diponibilizar una API
- Entorno de producción en la nube (AWS, Azure o Google Cloud)
- Considerar costos de infraestructura en los diferentes proveedores y tiempo de desarrollo

Limitaciones y riesgos

- Infraestructura para procesamiento para manejar variables lag y ejecutar evaluaciones o entrenamiento de modelos mas complejos.
- Considerar un dataset mas fiable para evitar imputar valores cercanos si el porcentaje de valores nulos es grande
- Considerar la precisión del modelo

Gracias