Proyecto 2 Data Science

Brandon Reyes Morales - 22992
Nancy Mazariegos 22513
Santiago Pereira Alvarado - 22318
Andre Jo 22199

Conclusiones

Principales Hallazgos

Estructura y calidad de los datos:

- El dataset presenta una alta dimensionalidad con 558 variables numéricas y 1,961 observaciones, lo que constituye un problema de big data que requerirá técnicas de reducción dimensional.
- El proceso de limpieza fue efectivo, eliminando 134 filas con valores faltantes del conjunto de prueba sin pérdida significativa de información.

Patrones de correlación identificados:

- Existe alta multicolinealidad entre variables relacionadas del mismo instrumento financiero (correlaciones de 0.81 a 1.00), especialmente en futuros japoneses de oro, platino y caucho, indicando redundancia significativa en los datos.
- Se detectaron correlaciones negativas moderadas a fuertes (-0.45 a -0.77) entre commodities y bonos estadounidenses (SPTL, VGLT), sugiriendo un efecto de diversificación natural entre estos activos.
- Las acciones del sector energético/minero muestran alta sincronía (0.93-0.97), confirmando que responden a factores macroeconómicos similares.

Comportamiento temporal de los commodities:

- La mayoría de las series analizadas (oro estándar, oro mini, platino) muestran tendencias alcistas sostenidas durante el período observado, con crecimiento especialmente acelerado en la segunda mitad del período.
- El caucho presenta un comportamiento atípico con mayor ciclicidad y volatilidad, alternando entre niveles de 150 y máximos superiores a 400, lo que indica mayor sensibilidad a factores externos.

Implicaciones para Etapas Posteriores

Para el preprocesamiento:

- Se implementarán técnicas de reducción dimensional debido a la alta correlación entre variables similares.
- Creación de variables sintéticas que capturen la información común de grupos de variables altamente correlacionadas.
- El tratamiento diferenciado del caucho necesitará modelos específicos dada su mayor volatilidad.

Para el modelado predictivo:

- La presencia de tendencias claras sugiere que modelos de series de tiempo como ARIMA o LSTM podrían ser efectivos para capturar patrones temporales.
- Las correlaciones negativas identificadas pueden aprovecharse para estrategias de cobertura y diversificación en las predicciones.
- La alta dimensionalidad favorece el uso de algoritmos robustos como Random Forest o Gradient Boosting que pueden manejar muchas variables.

Para la validación:

- Es crucial mantener el orden temporal en la división train/test para evitar data leakage.
- Se necesitará evaluación por bloques temporales dada la naturaleza de series de tiempo del problema.
- Las diferentes volatilidades entre commodities necesitarán métricas de evaluación específicas por activo.