Guatemala 29 de agosto de 2024

Jose Santisteban- 21153

Data Science

Laboratorio #3

Para la realización de este laboratorio se utilizaron los conjuntos de entrenamiento para cada serie utilizada en el laboratorio pasado, en este caso las series seleccionadas fueron las siguientes:

- Consumo de gasolina superior
- Importaciones de gasolina superior
- Precios de gasolina superior

El objetivo de este laboratorio es el uso de series de tiempo LSTM y creación de modelos para cada serie, lo cual consiste en la realización de un modelo con hiper parámetros preestablecidos y otro modelo de la misma serie con hiper parámetros 'mejorados'

El modelo LSTM que se utilizó cuenta con los siguientes hiperparámetros:

- Número de unidades LSTM: El modelo tiene dos capas LSTM, cada una con 50 unidades.
- Return sequences: En la primera capa LSTM, se estableció return_sequences=True, lo que permite que la capa devuelva la secuencia completa de salidas. En la segunda capa LSTM, return_sequences=False, por lo que solo se devuelve la última salida de la secuencia.
- Dropout: Se aplicó un dropout de 0.2 en ambas capas de LSTM, lo que implica que el 20% de las neuronas se desactivan aleatoriamente durante el entrenamiento para evitar el sobreajuste.
- Tamaño de la capa de salida: La capa densa (Dense) final del modelo tiene una única unidad, lo que indica que el modelo está configurado para hacer una predicción de un solo valor.
- Optimizador: Se utilizó el optimizador adam, un algoritmo de optimización que ajusta los parámetros del modelo basándose en el gradiente y la tasa de aprendizaje.

- Función de pérdida: Se utilizó la función de pérdida mean_squared_error, que calcula el promedio de los cuadrados de los errores, siendo adecuada para problemas de regresión.

Los hiper parámetros utilizados en los modelos sin mejora son los siguientes:

- Epochs: 50
- Batch size: 32

Los hiper parámetros utilizados en los modelos mejorados son los siguientes:

- Epochs: 100
- Batch size: 40

Consumo de gasolina superior

Durante la realización de este modelo se utilizaron los datos de consumo de gasolina en Guatemala y se hicieron 2 conjuntos de entrenamiento (para el modelo mejorado y sin mejorar). De esta manera podemos asegurar que ambos modelos se hayan entrenado en igualdad de condiciones y así obtener resultados contundentes.

Como se mencionó anteriormente, la función de pérdida que se utilizó fue MSE (Mean Squared Error), y en el momento de entrenar el modelo se obtuvo una pérdida de 0.0975, lo que se puede traducir a que el entrenamiento del modelo fue exitoso.

En el caso del modelo mejorado, tuvo una pérdida de 0.1035 lo que demuestra que fue un poco peor que el modelo anterior, sin embargo no es una diferencia muy significativa.

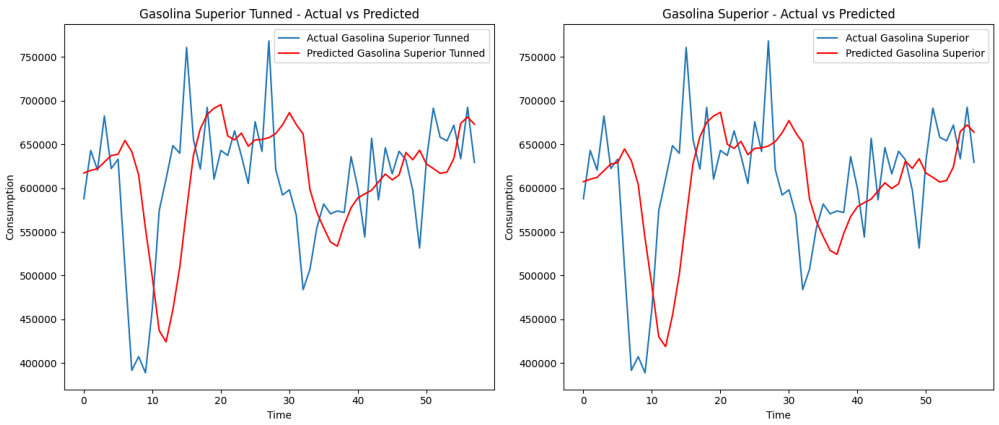

Figura 1: predicción modelos de consumo de gasolina superior

Como podemos observar en la figura 1 los modelos son muy similares, notando una muy pequeña mejoría en el modelo que no está mejorado, lo que puede sugerir un sobreajuste del modelo. Sin embargo a pesar de que ambas predicciones sean similares ambas muestran resultados diferentes a los datos reales, siguiendo solamente la tendencia de valores mínimos y máximos locales y globales.

Importaciones de gasolina superior

Durante la realización de este modelo se utilizaron los datos de importación de gasolina en Guatemala y se hicieron 2 conjuntos de entrenamiento (para el modelo mejorado y sin mejorar). De esta manera podemos asegurar que ambos modelos se hayan entrenado en igualdad de condiciones y así obtener resultados contundentes.

Como se mencionó anteriormente, la función de pérdida que se utilizó fue MSE (Mean Squared Error), y en el momento de entrenar el modelo se obtuvo una pérdida de 0.0742, lo que se puede traducir a que el entrenamiento del modelo fue exitoso.

En el caso del modelo mejorado, tuvo una pérdida de 0.0701 lo que demuestra que tuvo un mejor rendimiento que el modelo anterior, sin embargo no es una diferencia muy significativa.

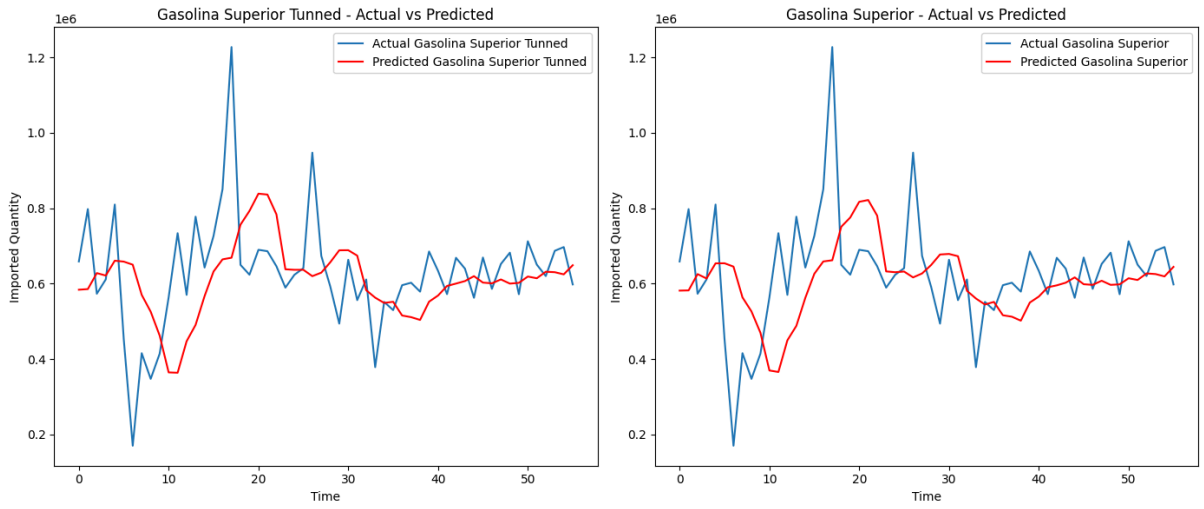

Figura 2: predicción de modelos de importación de gasolina superior

De la misma manera que la serie anterior, ambas predicciones de ambos modelos resultaron muy similares, de tal manera que es casi indistinguible ver las diferencias directamente, por lo que el uso de cualquiera de los dos modelos tendría el mismo resultado. Además, igualmente que en la serie anterior, los modelos muestran resultados diferentes a los datos reales, siguiendo solamente la tendencia de valores mínimos y máximos locales y globales.

Precios de gasolina superior

Durante la realización de este modelo se utilizaron los datos de precios de gasolina en Guatemala y se hicieron 2 conjuntos de entrenamiento (para el modelo mejorado y sin mejorar). De esta manera podemos asegurar que ambos modelos se hayan entrenado en igualdad de condiciones y así obtener resultados contundentes.

Como se mencionó anteriormente, la función de pérdida que se utilizó fue MSE (Mean Squared Error), y en el momento de entrenar el modelo se obtuvo una pérdida de 0.1703, lo que se puede traducir a que el entrenamiento del modelo fue exitoso.

En el caso del modelo mejorado, tuvo una pérdida de 0.1708 lo que demuestra que tuvo un rendimiento muy similar al modelo anterior ya que es una diferencia de 0.0005, lo que supone que ambos modelos se entrenaron de una manera exitosa.

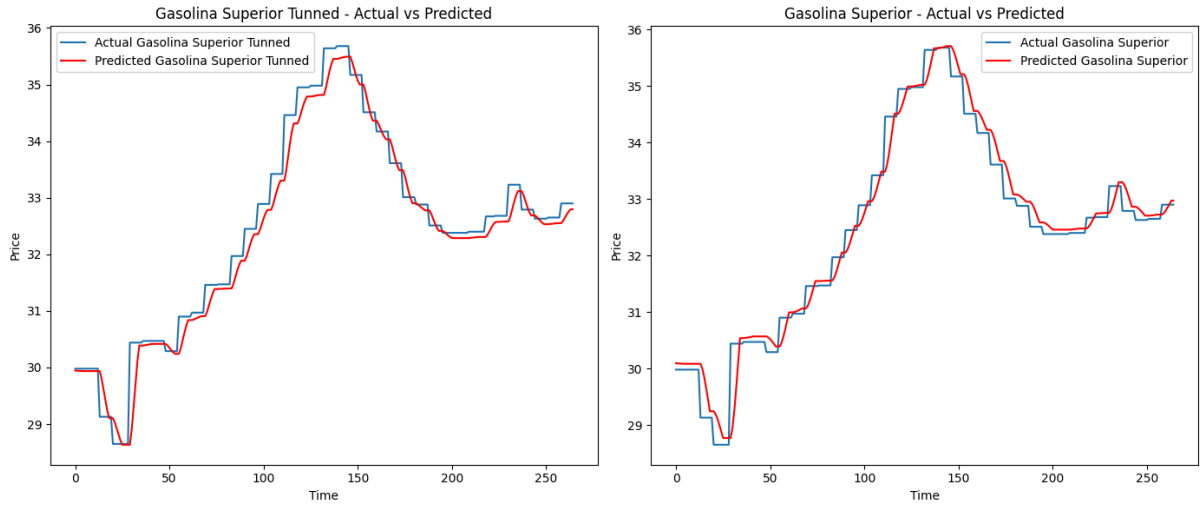

Figura 3: predicción de modelos de precio de gasolina superior

A diferencia de las dos series anteriores, en estas predicciones si se nota una diferencia, aunque no significativa, notable en donde sobresalta el modelo con los hiperparametros mejorados 'Gasolina Superior Tunned'. En el modelo mejorado se puede observar que las predicciones realizadas se apegan un poco más a los valores reales en comparación con el otro modelo.

A pesar de que el modelo mejorado tenga un mejor apego a los valores reales, es importante resaltar que el modelo sin mejorar tuvo resultados muy apegados a la realidad, por lo que el uso de este modelo no debería de suponer algún problema ya que predice resultados de manera correcta.

ARIMA vs LSTM

En la comparación entre ARIMA y LSTM, es importante destacar que, aunque los modelos LSTM son más complejos en términos de arquitectura, resultan más sencillos de utilizar en la práctica. Esto se debe a que los LSTM no requieren una preparación tan exhaustiva de los datos como ARIMA. Mientras que ARIMA necesita que los datos estén formateados y limpios de manera específica, incluyendo la necesidad de que la serie sea estacionaria y sin tendencias o estacionalidades no modeladas, los LSTM pueden manejar series temporales con patrones complejos y no lineales sin tanta preprocesamiento. En el laboratorio anterior, los malos resultados se obtuvieron con el modelo ARIMA, lo que subraya la importancia de un enfoque cuidadoso en la preparación de los datos y la configuración del modelo en este tipo de técnicas.

Esto se puede ejemplificar con los resultados de predicción obtenidos para los precios de gasolina superior utilizando el modelo ARIMA:

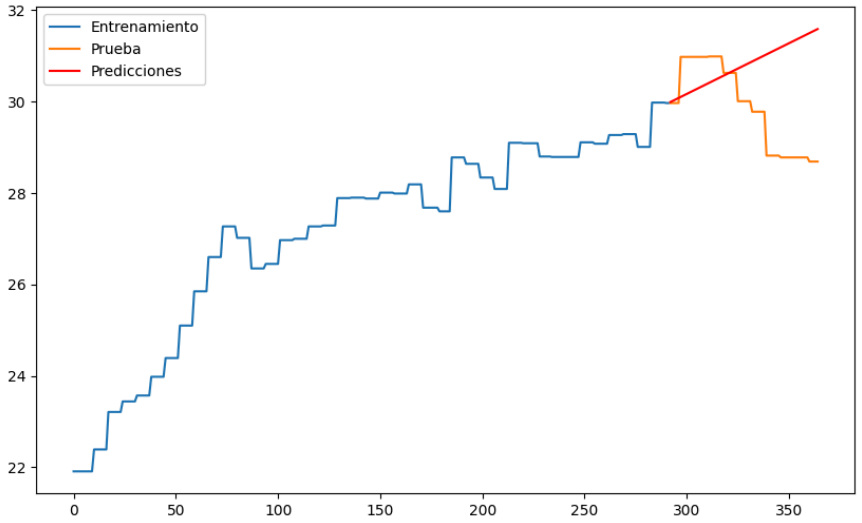

Figura 4: predicciones de modelo ARIMA con precios de gasolina superior

En la figura 4 se puede observar claramente el rendimiento superior de LSTM a comparación con ARIMA ya que este solo muestra una tendencia positiva pero los valores de la predicción no coinciden de ninguna manera con los valores reales de la serie de precios de gasolina superior, en cambio en la figura 3 los valores de predicción son precisos en comparación de la serie utilizada.

Cabe resaltar que es probable que el preprocesamiento de los datos en el modelo ARIMA no hayan sido realizados de manera correcta, pero esto solamente resalta la sencillez de utilizar el modelo LSTM.

Conclusiones

En conclusión, el mejor modelo para la predicción de los precios de gasolina superior resultó ser el LSTM. Su sencillez de uso y capacidad para manejar patrones complejos en las series temporales lo hicieron superior en comparación con otros enfoques. En particular, el modelo ARIMA no pudo competir eficazmente, ya que los valores de predicción generados por ARIMA no se apegaron a los datos de prueba, lo que resultó en un rendimiento significativamente inferior. Esto demuestra que, para este tipo de datos, LSTM es una herramienta más robusta y confiable para realizar predicciones precisas.

Link de repositorio

https://github.com/Jskenpo/DS_LAB3.git