Asignación de Regresión de Series Temporales

Ryoji Takahashi

December 1, 2024

1 Descripción de la Asignación y Objetivos

La asignación de regresión de series temporales consiste en:

1. Elegir un conjunto de datos multivariado de series temporales que no sea financiero y en el que la variable a predecir no tenga estacionalidad (fuente de referencia).
2. Verificar que la serie temporal no sea estacional.
3. Seleccionar, entrenar y probar los modelos de machine learning/series temporales que permitan: Predecir 100 períodos futuros con la particularidad de que, a partir del momento en que comienza la predicción, no se dispone de los valores de las variables regresoras.
4. Representar gráficamente los resultados y evaluar la bondad del modelo utilizando los KPIs adecuados.
5. Extraer conclusiones del ejercicio realizado.
6. Preparar un entregablé con todos los archivos necesarios para reproducir el análisis y poner en producción el(los) modelo(s) entrenado(s) (integración DevOps).

1.1 Conjunto de Datos

1. Conjunto de datos. household power consumption fue seleccionado para este estudio. Contiene 2,075,259 filas y 8 columnas de registros de potencia activa desde el 16/12/2006 al 26/11/2010 en minutos. He seleccionado "Global active power" medido por "date time" como el objetivo de la serie temporal. Como se muestra en el código, los datos fueron re-muestreados a datos por hora para reducir el ruido, y los valores perdidos fueron interpolados.

1.2 Identificación de Series Temporales No Estacionales

Para determinar si los datos de la serie temporal son estacionales o no, he examinado los siguientes criterios:

- El componente estacional en el gráfico de descomposición es plano o está ausente, lo que confirma que el conjunto de datos no es estacional.
- Las gráficas de la Función de Autocorrelación (ACF) y Función de Autocorrelación Parcial (PACF) no muestran patrones repetitivos a intervalos regulares.
- La prueba de Dickey-Fuller Aumentada (ADF) verifica si la serie tiene una raíz unitaria. Un valor $p > 0.05$ indica que la serie no es estacionaria, pero esto no implica directamente estacionalidad.

Dado que hay casos que son estacionarios pero no estacionales, la última prueba ADF no aborda directamente la cuestión de la no estacionalidad. Sin embargo, es importante considerarla.
Primero, como se muestra en la Figura 1, el gráfico de descomposición es plano.

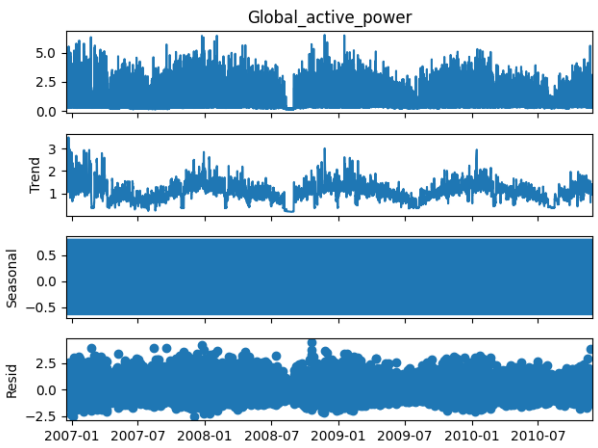

Figure 1: Descomposición Estacional

La Figura 2 muestra que los gráficos no tienen patrones repetitivos perfectos en intervalos de 240.

Finalmente, la prueba ADF muestra que el Estadístico ADF: -14.382566338396966 y el valor p: $9.101197279691604e^{-27}$. Estos resultados confirman que el conjunto de datos es estacionario y no estacional.

1.3 Modelos de Machine Learning

Entrené con el primer 80 % de los datos, probé y validé con el último 20 %, y desplegué los modelos Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), XGBoost con optuna y Long Short Term Memory (LSTM). El

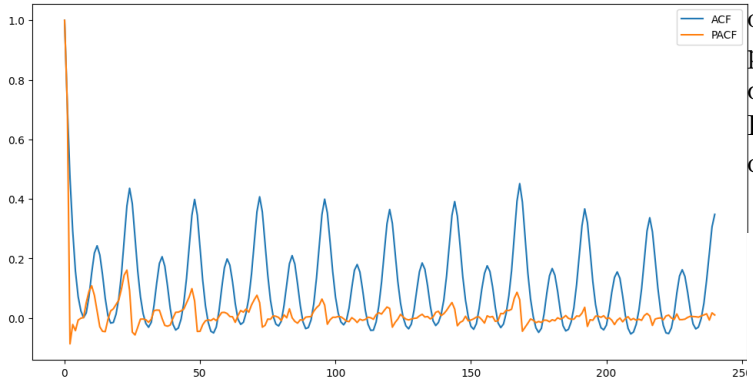

Figure 2: ACF, PACF

modelo SARIMAX es una técnica poderosa de predicción de series temporales que amplía el modelo ARIMA tradicional para tener en cuenta la estacionalidad y factores externos. Dado que observé un mejor rendimiento en XGBoost con optuna y además es muy rápido de calcular, también lo apliqué aquí. LSTM es conocido por proporcionar mayor precisión en predicciones exigentes.
La Figura 3 muestra los resultados de la predicción, y utilicé el Error Cuadrático Medio (MSE), Error Absoluto Medio (MAE) y R-Cuadrado (R^2) como KPIs.

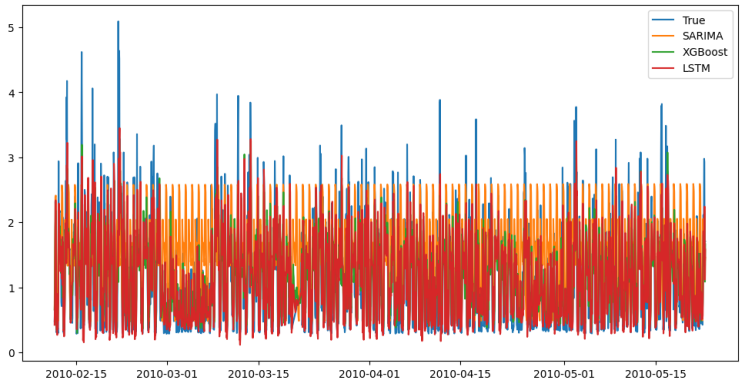

Figure 3: Resultados de Predicción de 100 períodos (24h) futuros

Como se especifica en la asignación, se graficaron 100 períodos de 24 horas. Es difícil ver los detalles. Por lo tanto, mostré todos los resultados en la Tabla 1.

Métricas de Rendimiento del Modelo			
Modelo \ KPI	MSE	MAE	R^2
SARIMAX	0.629	0.598	-0.061
XGBoost	0.266	0.358	0.551
LSTM	0.290	0.376	0.510

Table 1: Comparación de Métricas de Rendimiento de Modelos (100 períodos de 24 horas).

Como se puede observar, XGBoost tiene un rendimiento ligeramente mejor que LSTM. Esto podría deberse a que, por limitaciones de hardware, no realicé ajuste de hiperparámetros en LSTM. Si lo hubiera hecho, probablemente habría obtenido resultados más precisos que con XGBoost.
La Figura 4 proporciona una vista más clara en comparación con la Figura 3.

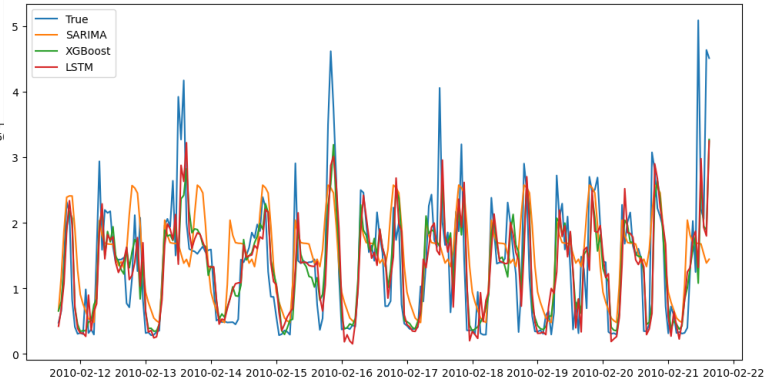

Figure 4: Resultados de Predicción de 10 períodos (24h) futuros.

Como se muestra, las predicciones de XGBoost y LSTM son mejores que las de SARIMAX.

Métricas de Rendimiento del Modelo			
Modelo \ KPI	MSE	MAE	R^2
SARIMAX	0.599	0.543	0.344
XGBoost	0.425	0.417	0.535
LSTM	0.459	0.439	0.497

Table 2: Comparación de Métricas de Rendimiento de Modelos (10 períodos de 24 horas).

Es importante señalar que debido a la menor cantidad de puntos de datos, las precisiones de XGBoost y LSTM son ligeramente peores que en los resultados previos. Sin embargo, SARIMAX es ligeramente mejor. Esto indica que SARIMAX es más adecuado para conjuntos de datos de series temporales más pequeños.
Finalmente, para la entrega y reproducción, como se indica en el código, guardé los modelos entrenados en archivos .pkl y .keras para que puedan cargarse fácilmente y ejecutar pruebas.

2 Conclusión y Discusión Adicional

El análisis de series temporales está en constante evolución, impulsado por los avances tecnológicos y la creciente disponibilidad de datos. Mi objetivo es seguir explorando la selección automatizada de modelos, modelos no lineales y machine learning con grandes conjuntos de datos. En particular, el pronóstico en tiempo real utilizando inteligencia artificial generativa es un área intrigante. Estoy por realizar análisis futuros utilizando todos los avances mencionados.