Asignación de Regresión de Series Temporales

Ryoji Takahashi

December 2, 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. El 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:

- Si el conjunto de datos no es estacional, el componente estacional en el gráfico de descomposición es plano o ausente.
- 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.

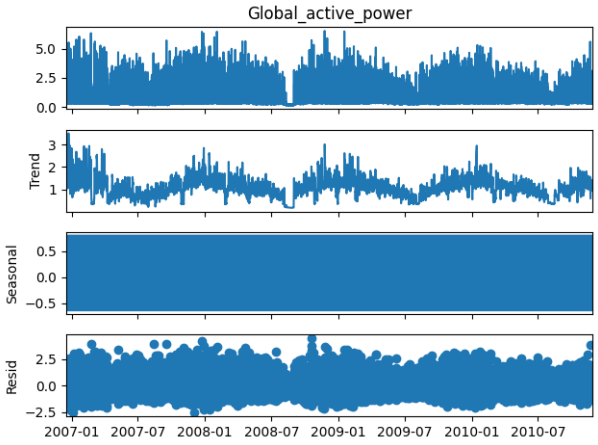

Figure 1: Descomposición Estacional

Primero, como se muestra en la Figura 1, el gráfico de descomposición es plano.

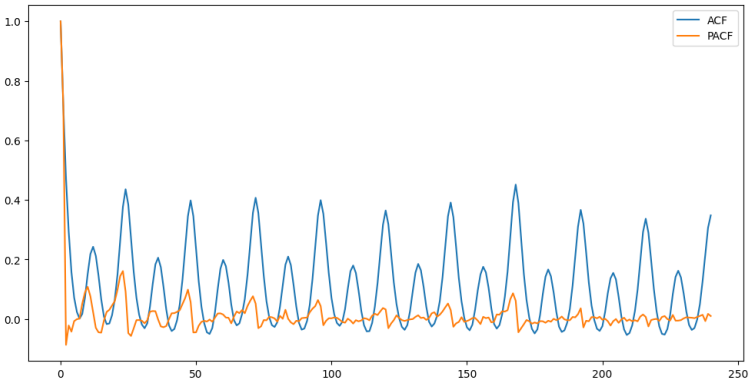

Figure 2: ACF, PACF

La Figura 2 muestra que los gráficos no tienen patrones repetitivos perfectos en intervalos de 240. La prueba ADF muestra que el Estadístico ADF: -14.382 y el valor p: $9.101 \times e^{-27}$. Por lo tanto, 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 %. Desplegué los modelos Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), XGBoost con optuna y Long Short Term Memory (LSTM). El 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. Como he visto un mejor rendimiento al usar XGBoost con optuna en los casos de clasificación, lo he aplicado aquí también. LSTM es conocido por proporcionar mayor precisión en predicciones exigentes.

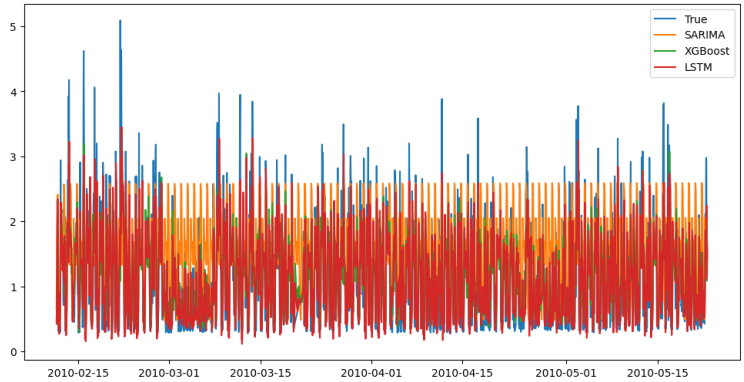

Figure 3: Resultados de Predicción de 100 períodos (24h) futuros

Según la asignación, realicé gráficos de predicciones para 100 períodos de tiempo (24 horas cada uno) en La Figura 3. (Es difícil ver los detalles). He utilizado

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).

el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R2) como indicadores clave de rendimiento (KPIs), como se muestra en la Tabla 1. Aunque la Figura 3 no muestra claramente los resultados, la Tabla 1 indica que XGBoost tiene un rendimiento ligeramente mejor que LSTM. Esto puede deberse a que no pude realizar una optimización de hiperparámetros (Optuna, Keras Tuner o GridSearchCV) para el LSTM debido a limitaciones de hardware.

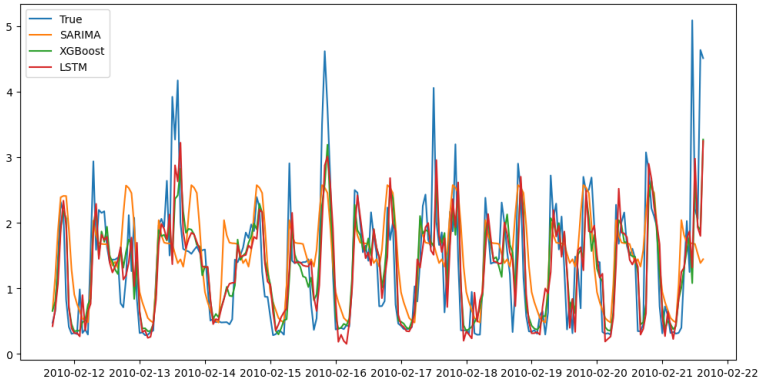

Figure 4: Resultados de Predicción de 10 períodos (24h) futuros.

Además, representé gráficamente los resultados de las predicciones para 10 períodos de tiempo (24 horas) en el futuro en la Figura 4. Esto ofrece una visión más clara en comparación con la Figura 3. Como se muestra, las predicciones de XGBoost y LSTM tienen un mejor desempeño 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).

La Tabla 2 muestra los valores de los indicadores KPI. Los resultados de XGBoost y LSTM son ligeramente peores que los anteriores. Esto sugiere que la precisión de cada modelo varía dentro de un rango determinado. Para demostrar el rango de precisión de cada modelo, experimenté con predicciones de 1, 5, 10, 20, 50, 100 y 200 períodos de tiempo (24 horas) en el futuro. El MSE en SARIMA está aproximadamente en el rango de $0.371 \leq \text{MSE} \leq 0.827$, en XGBoost entre $0.211 \leq \text{MSE} \leq 0.4287$, y en LSTM entre $0.239 \leq \text{MSE} \leq 0.465$. (Otros indicadores KPI no se muestran aquí). Por lo tanto, XGBoost y LSTM son los modelos más precisos para este estudio.

Finalmente, para la entrega y la reproducibilidad, guardé los modelos entrenados como archivos **pkl** y **keras**, lo que permite a otros cargar estos archivos y ejecutar pruebas fácilmente.

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.