Asignación de Regresión de Series Temporales

Ryoji Takahashi

December 4, 2024

1 Descripción de la Asignación y Objetivos

Los objetivos de esta asignación de regresión de series temporales son:

1. Elegir un conjunto de datos de series temporales multivariadas (que no sean financieras) donde la variable objetivo no muestre estacionalidad (fuente de referencia).

2. Confirmar la no estacionalidad del conjunto de datos.

3. Seleccionar, entrenar y probar modelos de aprendizaje automático/series temporales que puedan predecir 100 períodos futuros, asumiendo que no se conocen los valores de las variables regresoras más allá del punto de predicción.

4. Representar gráficamente las predicciones y evaluar el rendimiento del modelo utilizando métricas relevantes (KPIs).

5. Extraer conclusiones del análisis realizado.

6. Entregar todos los archivos necesarios para reproducir el análisis y desplegar los modelos entrenados (integración DevOps).

1.1 Selección y Descripción del Conjunto de Datos

El Population Time Series Data fue seleccionado para este estudio. El conjunto de datos contiene 816 filas y 3 columnas de registros de población desde 01/01/1952 hasta 01/12/2019. La variable objetivo es el "value," que representa el tamaño de la población a lo largo del tiempo.

Dado el tamaño reducido del conjunto de datos, se muestreó a intervalos diarios y se interpolaron los valores faltantes. Este preprocesamiento incrementó el tamaño de muestra a 24,807, proporcionando suficientes datos para construir modelos predictivos.

1.2 Confirmación de No Estacionalidad

Para confirmar la no estacionalidad, se llevaron a cabo los siguientes pasos:

- Descomposición estacional: Se analizó un gráfico de descomposición para garantizar que el componente estacional sea plano o insignificante.
- Prueba ADF (Augmented Dickey-Fuller): Esta prueba se aplicó para verificar la estacionariedad. Un p-valor > 0.05 indica que la serie no es estacionaria, aunque esto no confirma directamente la no estacionalidad.

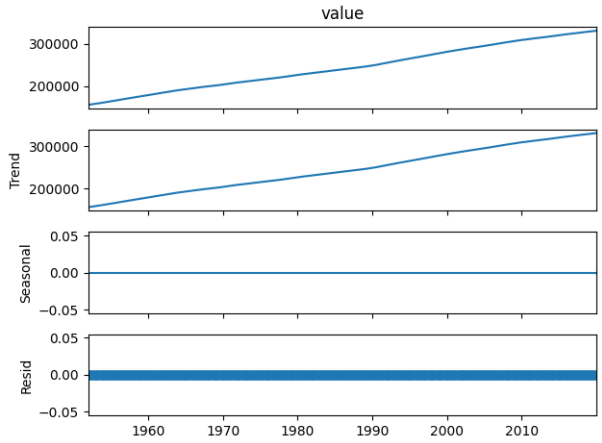

Figure 1: Descomposición Estacional

La figura 1 muestra el gráfico de descomposición. El componente estacional es plano, y los resultados de la prueba ADF (ADF Statistic $= -0.742$, P-valor $= 0.835$) confirman que el conjunto de datos es no estacionario y no estacional. Cabe destacar que se utilizó información interpolada anualmente para generar el gráfico de descomposición.

1.3 Modelos Entrenados y Metodología

El conjunto de datos se dividió en subconjuntos de entrenamiento (80%) y prueba (20%). Los siguientes modelos fueron entrenados y evaluados:

- **SARIMAX (Seasonal ARIMA with Exogenous Variables):** Una técnica de predicción de series tem-

porales que extiende ARIMA para incluir efectos estacionales y regresores externos. Es adecuado para conjuntos de datos pequeños.

- **XGBoost con Optuna:** Un marco de boosting de gradiente optimizado utilizando la biblioteca Optuna para el ajuste de hiperparámetros. XGBoost es efectivo para datos estructurados y maneja eficientemente los valores faltantes. Aunque XGBoost ha mostrado resultados sólidos en conjuntos de datos estacionales, tuvo dificultades en este conjunto de datos no estacional. Esto podría deberse a que XGBoost no está diseñado para capturar dependencias a largo plazo, que a menudo están presentes en series temporales no estacionales.

- **Long Short-Term Memory (LSTM):** Una arquitectura de redes neuronales recurrentes diseñada para predicciones de secuencias. Su capacidad para aprender dependencias a largo plazo lo hace poderoso para la predicción de series temporales.

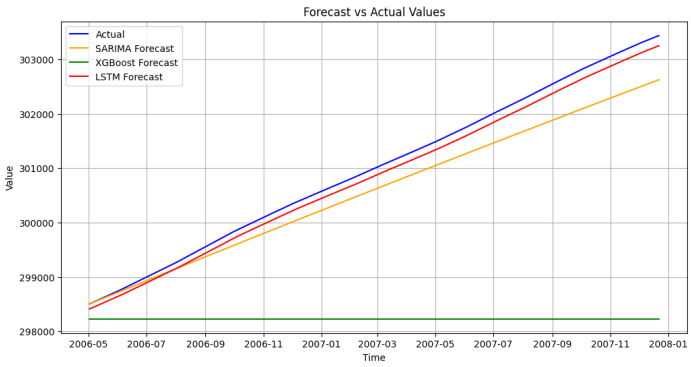

Figure 2: Resultados de Predicción para 100 Períodos Futuros (6 Días Cada Uno)

La Figura 2 ilustra las predicciones para 100 períodos futuros (6 días cada uno).

Modelo	MSE(NMSE)	MAE(NMAS)	R^2
SARIMAX	218267.59 (0.107)	404.41 (0.001)	0.892
XGBoost	9682372.40 (4.797)	2766.88 (0.0092)	-3.777
LSTM	20921.49 (0.005)	141.99 (0.0003)	0.990

Table 1: KPIs (Métricas) Rendimiento de los Modelos Entrenados

La Tabla 1 muestra los KPIs (métricas): Error Cuadrático Medio (MSE), Error Absoluto Medio (MAE) y R^2. NMSE y NMAS representan normalizado del MSE y del MAS, respectivamente.

Como se observa en la Tabla 1, el modelo LSTM superó a SARIMAX y XGBoost en precisión, aunque podría haber sobreajustado debido a su arquitectura compleja. SARIMAX mostró un rendimiento consistente, mientras que se evidencian las limitaciones de XGBoost para conjuntos de datos no estacionales. Cabe destacar que el bajo NMAE de XGBoost indica que los errores son pequeños en relación con la media de la variable objetivo. Sin embargo, NMSE y R^2 muestran que las predicciones del modelo se desvían significativamente de los valores reales, como se muestra en la Figura 2.

2 Conclusión y Trabajo Futuro

El análisis destacó que SARIMAX y LSTM son más adecuados para conjuntos de datos no estacionales, siendo SARIMAX eficiente computacionalmente y LSTM más preciso para sistemas de mayor escala. Los modelos basados en árboles pueden no captar dependencias a largo plazo, lo que los hace menos efectivos para el pronóstico de series temporales no estacionales."

Las mejoras futuras podrían incluir:

- Abordar el sobreajuste en modelos LSTM mediante la incorporación de capas de dropout y early stopping.

- Explorar métodos de predicción en tiempo real y Generative AI para el análisis de series temporales.

El pronóstico de series temporales sigue siendo un campo en evolución, con oportunidades emocionantes en la selección automatizada de modelos y predicciones en tiempo real para aplicaciones a gran escala.