Tarea 8 Aprendizaje Automático

Universidad Autonoma de Nuevo León
Alanis Mares, Victor Hugo

18 de Julio, 2024

1 Introducción

Se pretende analizar un conjunto de series de tiempo con diferentes indicadores macroecónomicos para implementar un sistema de prediccion supervisada.

2 Descripción de los datos

Los datos a trabajar cuentan con 16 variables de las cuales once son referentes a la serie de tiempo multivariada y cinco son categoricas. Para propositos de este estudio se utilizaran las once variables numericas, las cuales son:

- Year: Año de observacion.
- Under_five_deaths: Cantidad de muertes en infantes menores de 5 años por cada 1000 habitantes.
- Adult_mortality: Cantidad de muertes en adultos por cada 1000 habitantes.
- Alcohol_consumption: Litros anualizados de consumo de alcohol por capita.
- Hepatitis_B: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Hepatitis B.
- BMI: Índice de grasa corporal promedio.
- Polio: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Polio.
- Diptheria: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para tétanos.
- Incidents_HIV: Casos de VIH por cada 1000 habitantes.
- GDP_per_capita: Producto interno bruto per capita.

- Schooling: Cantidad de años promedio que un adulto (25+) ha pasado estudiando.
- Life_expectancy: expectativa de vida al nacimiento.

2.1 Origen de los datos

Los datos fueron obtenidos de un registro público de conjuntos de datos, los cuales provienen de estudios de múltiples hospitales, a continuación se encuentra el enlace al conjunto de datos.

2.2 Preprocesamiento

Basado en un análisis de características se decidió omitir ciertas variables del conjunto, para más informacion al respecto referirse a este repositorio

3 Metodología

Para este análisis se tomará como variable objetivo la expectativa de vida para cada país, se utilizara un regresor de potenciación de gradiente apoyándonos de la librería `skforecast`. La razón principal de elección de la librería previamente mencionada es su método `ForecasterAutoregMultiSeries` el cual implementa un regresor de potenciación de gradiente multivariado, este método recibe dos argumentos:

- Un conjunto de series de tiempo referentes a nuestra variable objetivo.
- Un conjunto de series de tiempo referentes a variables exógenas a la variable objetivo las cuales se presuponen afectan a nuestra variable objetivo.

Debido a la baja cantidad de observaciones por serie de tiempo se utilizara un regresor que solo toma en cuenta las ultimas 5 observaciones de la serie de tiempo.

4 Metricas de resultados

Para evaluar el modelo y proceder con la optimización de hiper parámetros se procedió a analizar los siguientes tipos de métricas:

- Error promedio al cuadrado.
- Error promedio absoluto.
- Error promedio absoluto porcentual.
- Error promedio cuadrado logarítmico.

Despues de optimizar los hiperparametros se consiguieron las siguientes métricas de desempeño:

Métrica	Promedio	Máximo	Mínimo
Promedio al cuadrado	0.810385	19.811893	0.000090
Promedio absoluto	0.560838	4.408434	0.007003
Promedio absoluto porcentual	0.008592	0.078813	0.000097
Promedio cuadrado logarítmico	2.219828e-04	6.929335e-03	1.693964e-08

Al evaluar el grafico de caja de la métrica "Promedio absoluto" podemos apreciar como se tienen muchos valores atípicos, sin embargo, la gran mayoría de los datos están contenidos en un error entre 0 y 1, lo cual considerando las bajas observaciones y la variable respuesta medida considero que es un error aceptable, siendo esto referente a máximo un año de la esperanza de vida dado los estimadores macroeconómicos usados.

References

[1] Bernd Fritzke (1995) A Growing Neural Gas Network Learns Topologies.

[2] Halil Ertan (2022) Multivariate Time Series Clustering Using Growing Neural Gas and Spectral Clustering

[3] Joaquín Amat Rodrigo, Javier Escobar Ortiz (2022) Modelos de forecasting globales: modelado de múltiples series temporales con machine learning

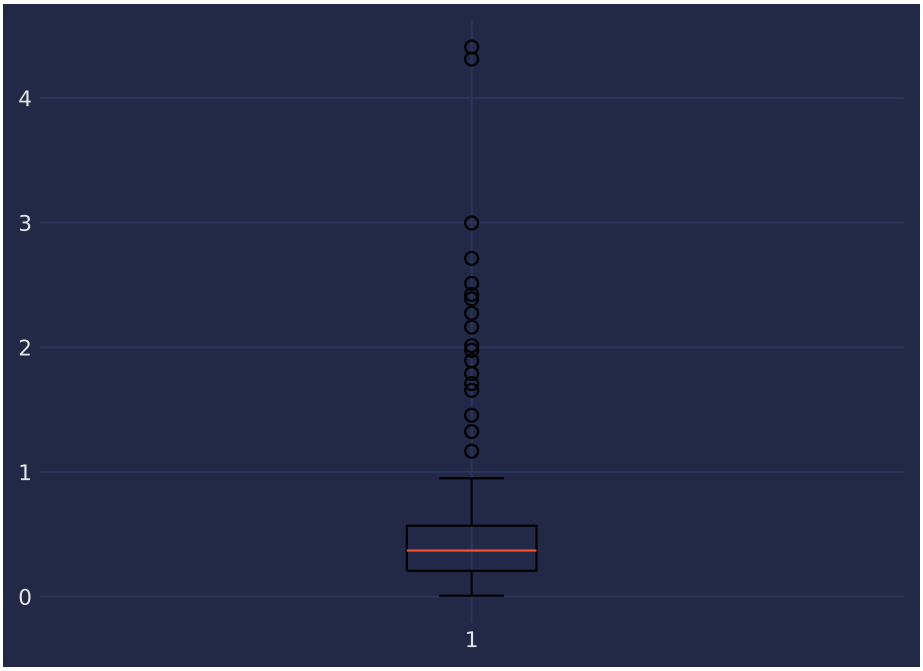

Figure 1: Grafico de caja del error promedio absoluto del modelo