Reto HidroPredictiva

Equipo Overflow (4)

Objetivo

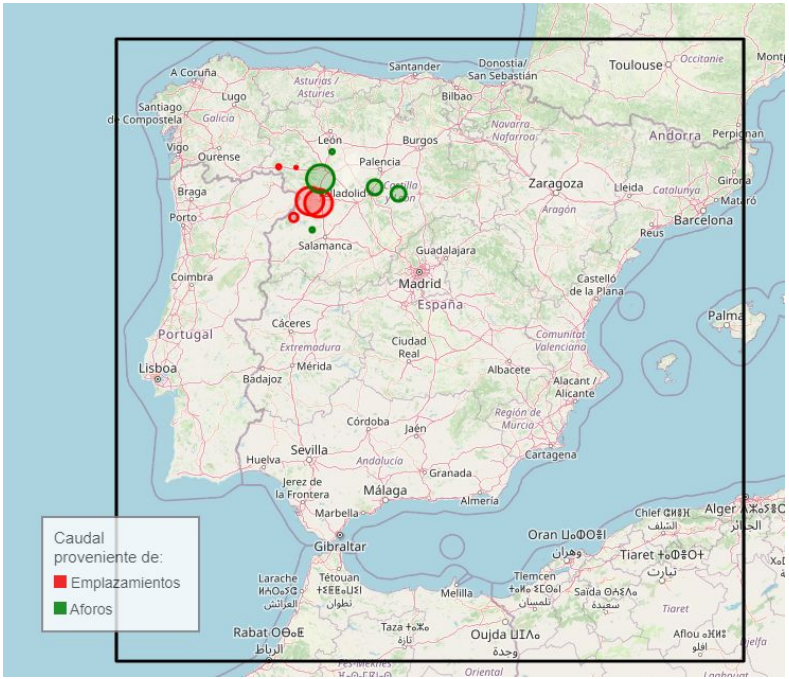
Programa Nacional de Algoritmos Verdes (PNAV) presenta el desafío de desarrollar modelos eficientes de predicción hidráulica.

Pronósticos para 5 puntos de la cuenta del Duero a 24 y a 48 horas.

Problema

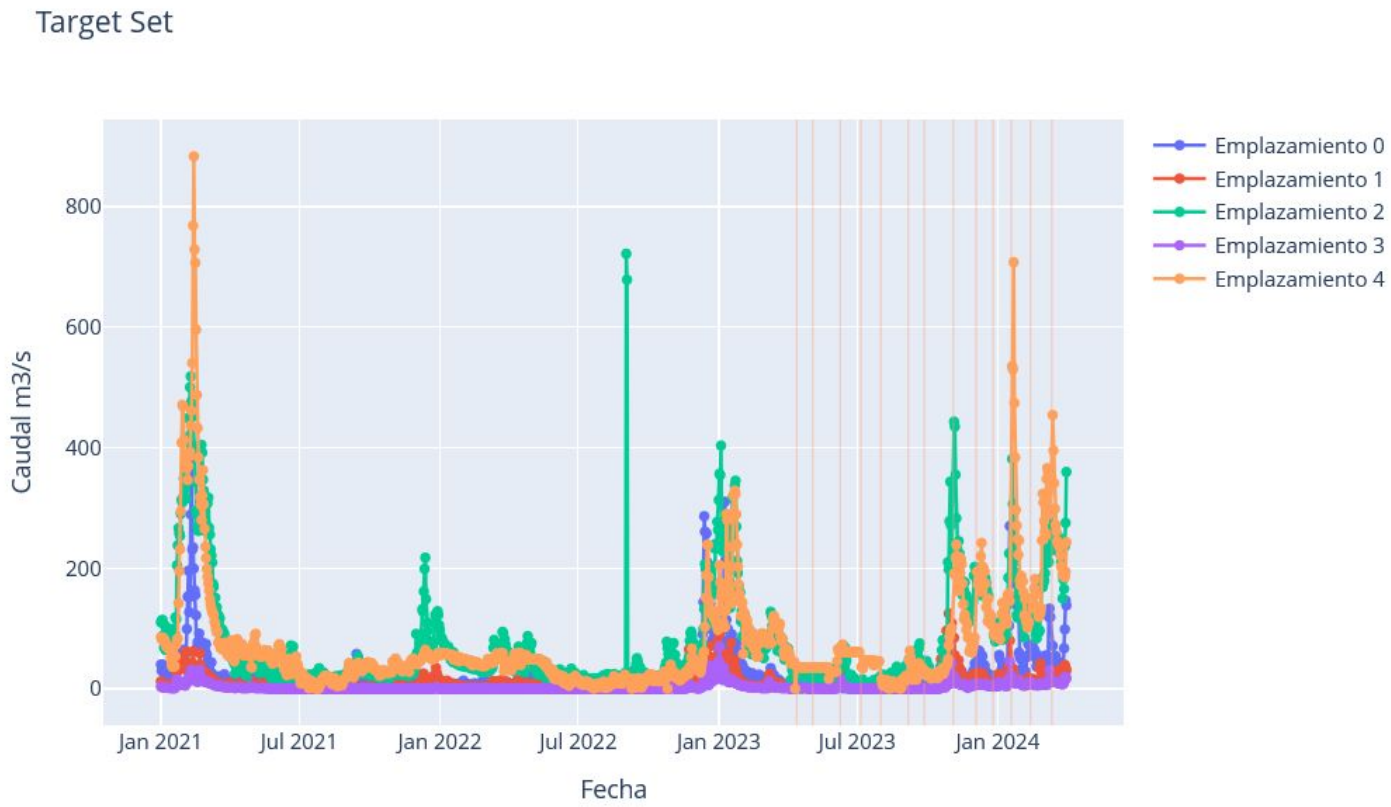
- Volumen de datos elevados
- Eficiencia recursos computacionales.
- Precisión en las predicciones para planificación estratégica y operación de la cuenca hidrográfica.

Contexto

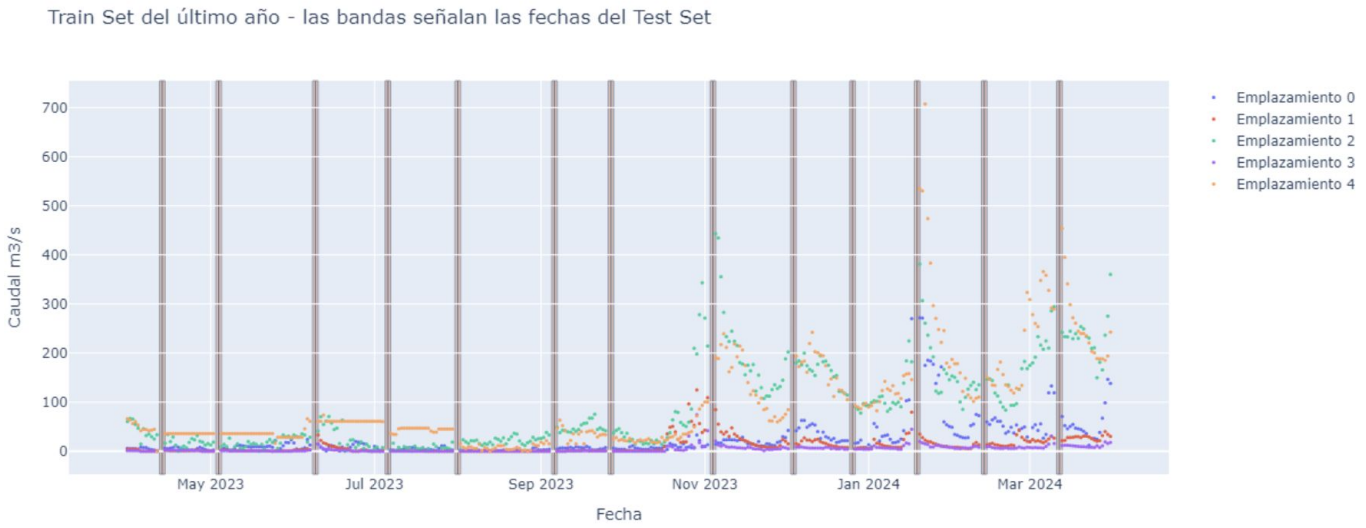


Datos aforo + GFS → Caudal emplazamiento

Contexto


Target Set
Caudal m3/s
800
600
400
200
0
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024
Fecha
Emplazamiento 0
Emplazamiento 1
Emplazamiento 2
Emplazamiento 3
Emplazamiento 4

Contexto


Train Set del último año - las bandas señalan las fechas del Test Set
Caudal m3/s
700
600
500
400
300
200
100
0
May 2023
Jul 2023
Sep 2023
Nov 2023
Jan 2024
Mar 2024
Fecha
Emplazamiento 0
Emplazamiento 1
Emplazamiento 2
Emplazamiento 3
Emplazamiento 4

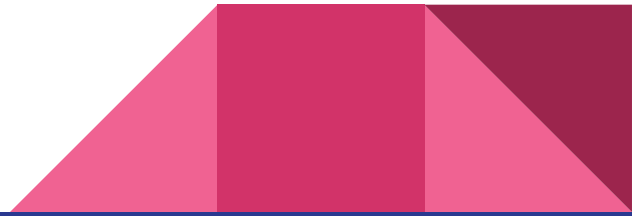
Nuestra solución

2 (24 y 48h) modelo XGBoost por desplazamiento.

- Selección de features más importantes por desplazamiento.
- Lags para variables importantes.

¿Por qué XGBoost?

- Rápido
- Híbrido entre random forest y gradient boosting
- Redes neuronales (LSTM, transformers):
  - Datos insuficientes
  - Mayor complejidad computacional y espacial



Preprocesamiento y selección de variables

Objetivo: Modelo más sencillo posible pero que mantenga precisión.

Paso 1:

Análisis de correlación entre features e y → 4 variables más importantes

Paso 2:

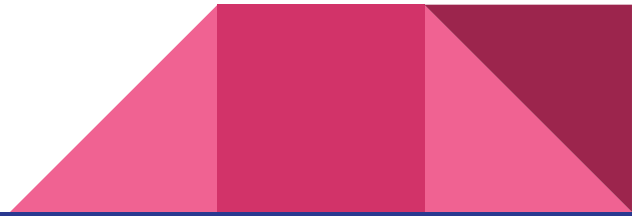
Análisis de correlación tamaño lag óptimo → hasta 4 días

Paso 3:

Añadir variables temporales

¿Por qué hemos elegido este método para reducción de dimensionalidad?

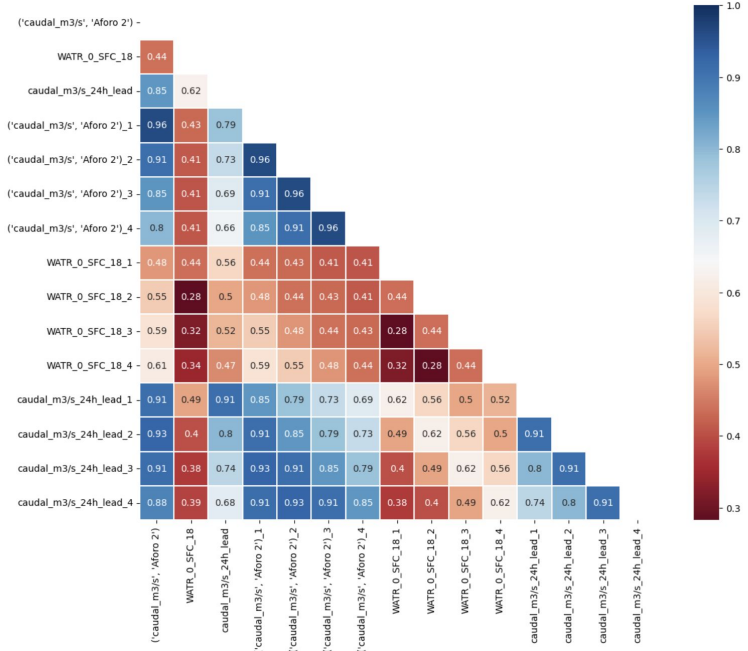
- Explicabilidad y control.



Análisis de correlación

*Hemos eliminado del modelo las referencias a los valores del caudal

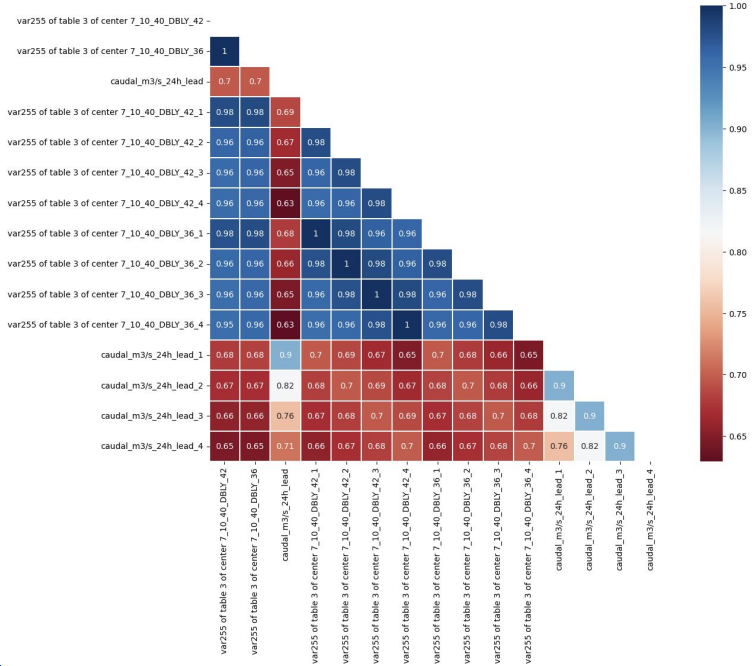
Emplazamiento 0



Análisis de correlación

*Hemos eliminado del modelo las referencias a los valores del caudal

Emplazamiento 1



Análisis de correlación

Emplazamiento i

Testing y validación

Validación cruzada:

- Time Series Split
- 5 folds
- Aprovechamos todo el dataset

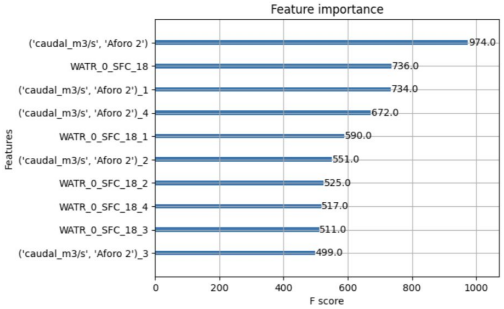
Código eficiente

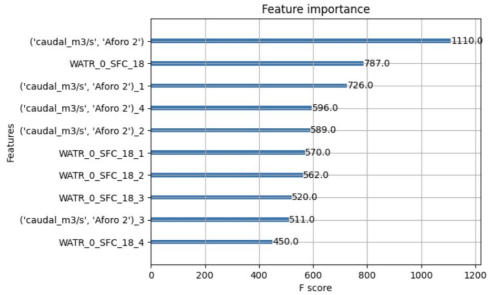
Hemos empleado librerías optimizadas a bajo nivel como:

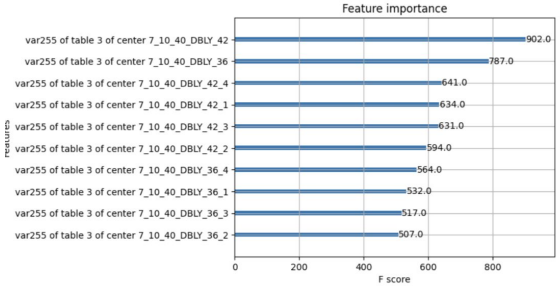
- XGBboost
- pandas
- numpy
- …

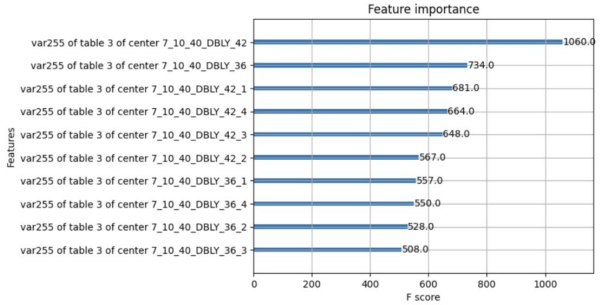
Beneficios

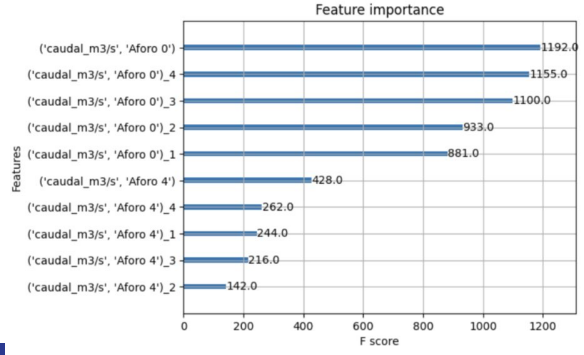
Sacamos máximo provecho a las variables: F score











Beneficios

- **0.003544** kWh de consumo de electricidad (incluyendo codigo que analisis).
- **0.002089** g.CO2eq/s estimación media anual 65.8649 kg.CO2eq/año.

- Sin selección de variables y lags (370%):
- 0.013351 kWh
- 0.014929 g.CO2eq/s anual

Resultados

- MAE: 29 (Sin el uso de la variable del caudal de los emplazamientos)

Metricas del conjunto de validación:

Emplazamientos	MAE caudal_m3/s_24h	MAE caudal_m3/s_48h
Emplazamiento 0	9.7629	9.8054
Emplazamiento 1	3.8009	3.9101
Emplazamiento 2	20.50	23.33
Emplazamiento 3	1.5718	1.0964
Emplazamiento 4	22.700	23.60

Mejoras

- Entrenar con más datos
  - Considerar franjas del testset
- Crear modelos específicos para franjas irregulares.
- Probar juntar modelos.
- Seguir experimentando con hiperparametros.
- Analizar otras métricas de ajuste