Reproducción de la Competencia ACIC 2016

Federico N. Llanes

Miguel I. Rodríguez Puertas

Universidad Nacional de San Martín

Inferencia Bayesiana Causal

Gustavo Landfried

12 de mayo 2025

Introducción

El presente informe tiene como objetivo comparar el desempeño de la versión original del *CalCause* de la competencia Acic 2016 con los resultados obtenidos a partir de la reimplementación del modelo.

Desarrollo

El **CalCause** es un modelo de inferencia causal presentado en la ACIC 2016 que combina dos técnicas poderosas: **Random Forests** y **Gaussian Process Regressor**. El objetivo principal de este modelo es estimar el *efecto causal individual* (ITE) en escenarios con variables de tratamiento y control. En esencia, **CalCause** realiza dos estimaciones principales: una para los resultados bajo el tratamiento $Y(1)$ y otra para los resultados bajo el control $Y(0)$. La diferencia entre estas predicciones, $\hat{\tau}(x) = \hat{Y}(1|x) - \hat{Y}(0|x)$, permite obtener el efecto causal individual para cada observación. La inclusión de **Random Forests** permite al modelo manejar relaciones no lineales complejas entre las variables, mientras que el uso de **Gaussian Processes** mejora la estimación al ajustar suavemente la predicción y manejar la incertidumbre de manera más eficiente, especialmente en presencia de ruido o poca información.

Métricas

El estimando oficial de la competencia es el efecto promedio del tratamiento sobre los tratados ($SATT$), definido como

$$\text{SATT} = \frac{1}{n_T} \sum_{i:\, z_i=1} \Big(y_i(1) - y_i(0) \Big),$$

donde n_T es el número de unidades tratadas. Todas las métricas que se reportan a continuación se calculan sobre este parámetro.

- *Sesgo promedio (Bias)*:

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)$$

- *Error cuadrático medio (RMSE)*:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2}$$

- *Cobertura del intervalo de confianza al 95 %*:

$$\text{Cobertura}_{95\,\%} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{\theta_i \in [\hat{\theta}_i^L, \hat{\theta}_i^U]\right\}$$

  donde $[\hat{\theta}_i^L, \hat{\theta}_i^U]$ es el intervalo bootstrap del 95 % para el dataset i.

- *Largo promedio del IC*:

$$L_{\text{IC}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i^U - \hat{\theta}_i^L)$$

Metodología

Tuning inicial (10 datasets)

Se realizó una búsqueda en grilla de hiperparámetros sobre 10 datasets simulados distintos a los usados en evaluación final. Los valores fijados tras el tuning fueron:

- Random Forest: `n_estimators`=100, `max_depth`=8, `min_samples_split`=5
- Gaussian Process: kernel = CK $\times$ RBF con $\ell = 1$, $\alpha = 10^{-2}$

Evaluación congelada (77 carpetas $\times$ 20 réplicas)

El modelo se congeló tras el tuning y se evaluó en 1 540 archivos (77 escenarios con 20 réplicas cada uno - 20 % del total de cada carpeta). Se decidió el modelo (RF o GP) usando 3-fold CV sobre el grupo control, y se calculó el IC del 95 % por bootstrap (500 muestras).

Librerías utilizadas

Se utilizaron las siguientes librerías para el desarrollo y evaluación del modelo:

- `scikit-learn`: para implementar los modelos ensamblados de *CalCause* y obtener métricas básicas de desempeño.
- `numpy`: para el cálculo personalizado de métricas como el $RMSE$ y el $SATT$.
- `pandas`: para la carga, filtrado y manipulación de los conjuntos de datos.
- `random`: para la selección aleatoria de conjuntos de datos por escenario.
- `joblib`: para guardar y cargar modelos entrenados y resultados intermedios, optimizando el tiempo de cómputo.

Resultados

A partir de la recreación de la competencia ACIC 2016 usando el modelo CalCause se obtuvieron los siguientes resultados

Métrica	**CalCause original**	**Versión implementada**
Cobertura IC 95 %	82 %	74.5 %
Largo promedio del IC	0.08	0.38
Sesgo promedio (Bias)	$\sim$0.03	0.0567
Error cuadrático medio (RMSE)	$\sim$0.05	0.1779
Modelo elegido (CV interna)	No reportado	RF en 95 % de los casos

Cuadro 1

Comparación entre el desempeño del CalCause original (ACIC 2016) y la versión reimplementada con hiperparámetros fijos.

La versión implementada pierde adaptabilidad en comparación con el CalCause original. A pesar de contar con intervalos de confianza más anchos ($\sim 5\times$), la cobertura cae al 74.5 % vs. 82 % reportado en ACIC 2016. Asimismo, el RMSE también es significativamente mayor (0.18 vs. 0.05), y el sesgo medio se duplica. Además, en el 95 % de los casos el CV eligió RF por sobre GP, lo que evidencia una falta de flexibilidad en el ajuste de hiperparámetros ante escenarios más complejos.

Conclusiones

Luego de realizar la recreación y compararla con el modelo de la competencia, podemos concluir que es necesario realizar una revisión exhaustiva de la optimización de hiperparámetros, tanto para el modelo de **Random Forest** como para el de **Gaussian Process Regressor**. En particular, se recomienda explorar variantes del kernel utilizado en el GPR, así como considerar entrenamientos bajo inferencia bayesiana completa —por ejemplo, mediante muestreo MCMC o métodos variacionales—, ya que se observó una mayor predominancia del desempeño del modelo basado en Random Forest.