Trabajo Práctico 1 - LDD2024

Maximiliano Gandini - 39977503

Septiembre 2024

1. Introducción

Al igual que en el primer trabajo práctico, se eligió el segundo conjunto de datos: Registro anual de nacimientos"; Nacimientos en Argentina por departamento entre 2012 y 2022, publicado por el gobierno de la Nación. El enlace al Collab es: Trabajo Práctico - 2.

En el primer trabajo práctico se eligió la pregunta: ¿Es la tendencia de la baja en la tasa de natalidad a nivel nacional observable en la provincia de Buenos Aires?, con el fin de comprobar que esta tendencia provenía de variables que afectan al país en general.

Este segundo trabajo se basa en una temática similar. Considerando que los datos de todas las provincias contienen implícitamente la información sobre el efecto que está afectando a Buenos Aires, uno debería ser capaz de entrenar un modelo capaz de levantar esos datos y filtrar el ruido para modelar la caída en la natalidad. Se observaron varias posibilidades, pero finalmente se optó por un modelo de `RandomForestRegressor`, una instancia de aprendizaje supervisado con 'target=cantidad de nacimientos' y características (features) a evaluar.

2. Resultados y comentarios

En la figura 1 se pueden ver la totalidad de los datos de validación y la predicción del modelo. Cada línea con diferente color es un municipio en el lapso de una década con su natalidad cayendo. Para los municipios con menor población, esta se vio de manera acelerada y otros municipios tuvieron una caída en forma curva, con cierta resistencia a la tendencia.

Utilizando XGBoost, se ajustó el modelo teniendo en cuenta la complejidad de los árboles de decisión y la velocidad de aprendizaje del método. Posteriormente a cada ajuste, se introdujo una etapa de actualización donde se podaba el árbol de decisiones y se optimizaba con otro criterio. Se separaron los datos en test y train splits por grupo, excluyendo a Buenos Aires del grupo de entrenamiento. Para control adicional, se corroboró con validación cruzada que los resultados son reproducibles tomando otros grupos de provincias.

Se encontró que cualitativamente el mejor criterio para la optimización del gradiente fue el de RMSE (Root Mean Squared Error), el cual, con validación cruzada, otorgaba valores para los estadísticos de MSE $= 0.080$ y $R^2 = 0.96$.

Estos estadísticos no fueron consistentes en sus aparentes mejoras, ya que podrían mejorar a pesar de tener una forma que no capta la caída de manera consistente. Esto es especialmente visible en las poblaciones intermedias, donde se encontraba la mayor cantidad de datos, lo que provocaba sobreajustes en forma de rápidas oscilaciones.

Con el fin de encontrar una guía que permita afinar los hiperparámetros del modelo, se construyó la medida mostrada en la figura 2 y se observaron los residuos. En general, cuando los ajustes otorgaban estadísticos razonables, se encontró que el modelo siempre captaba la caída general en la natalidad. Además, los residuos permitieron encontrar mejores hiperparámetros al revelar que faltaban grados de libertad para las hojas de los árboles de decisión, así como sobreajustes a patrones recurrentes para la caída en la natalidad. (Notar que algunas caídas tienen una curvatura y otras directamente caen en picada).

Existen ciertos casos patológicos donde las predicciones no logran captar incluso a grandes rasgos la caída en natalidad de algunos municipios en particular. Finalmente, se remarca que el modelo sigue fallando en casos donde se mezclan datos de varios municipios. Teniendo en cuenta la etiqueta de cada uno y agregando grados de libertad, esto se puede aligerar, pero eventualmente el modelo comienza a sobreajustar y pierde precisión en otras secciones.

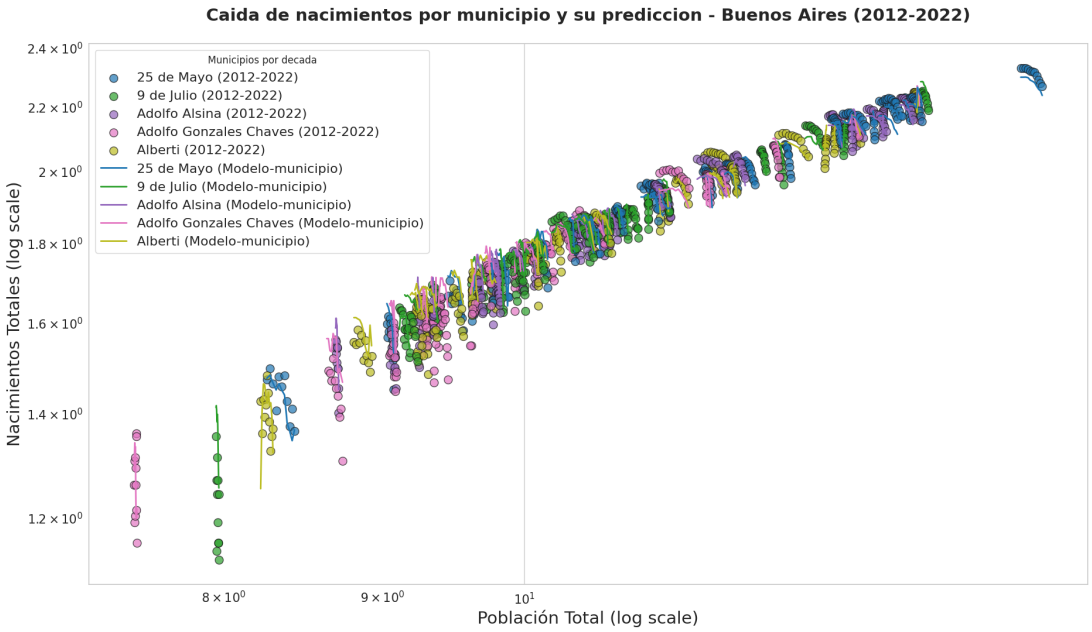

Figura 1: Cada línea con diferente color es un municipio en el lapso de una década con su natalidad cayendo. En línea continua pueden verse las predicciones del RFR para cada uno de los municipios entre 2012 y 2022. Se tomaron los logaritmos de la población total y la cantidad de nacimientos para poder visualizar todos los municipios.

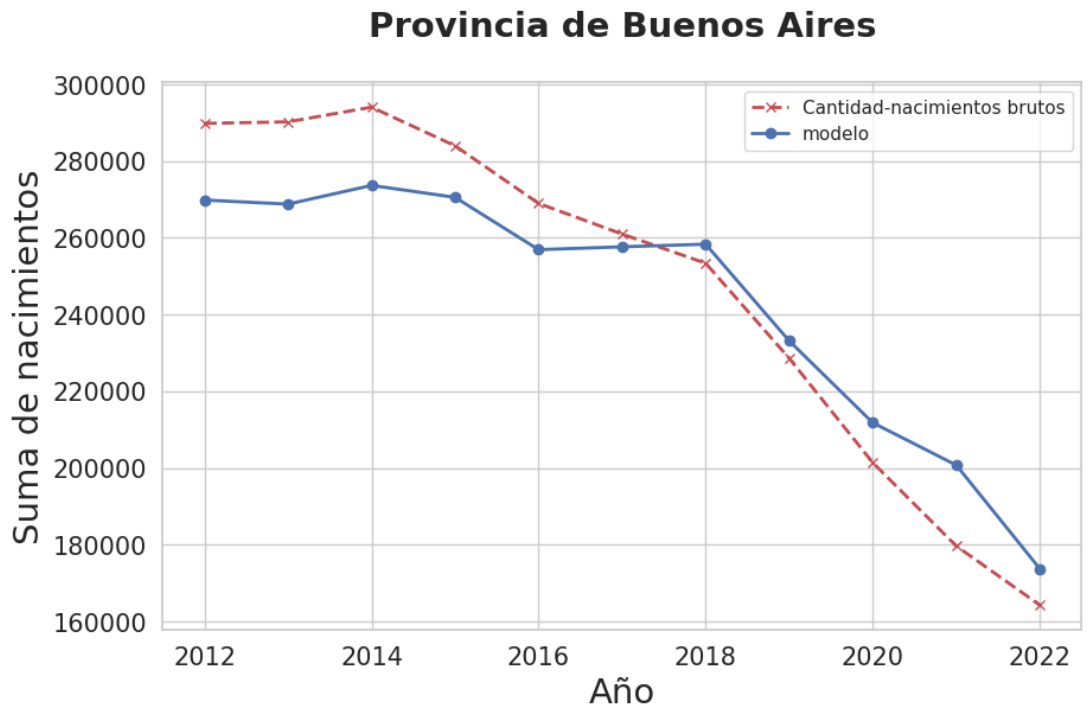

Figura 2: Suma de nacimientos brutos en la provincia de Buenos Aires anualizada. El modelo, en líneas generales, capta la baja en la natalidad.