EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Modelos por ensamble.
Objetivo: Que veremos.
Lectura recomendada: Sección 8.2 de [1].

1. INTRODUCCIÓN

Anteriormente, vimos los modelos predictivos basados en `árboles de decisión` (regresión y clasificación). En esta sección del curso estudiaremos `distintas estrategias` para combinarlos para obtener un mejor modelo predictivo al costo de `interpretabilidad`. Algunos de estos modelos continúan representando el estado del arte en competencias como Kaggle para `datos tabulares`.

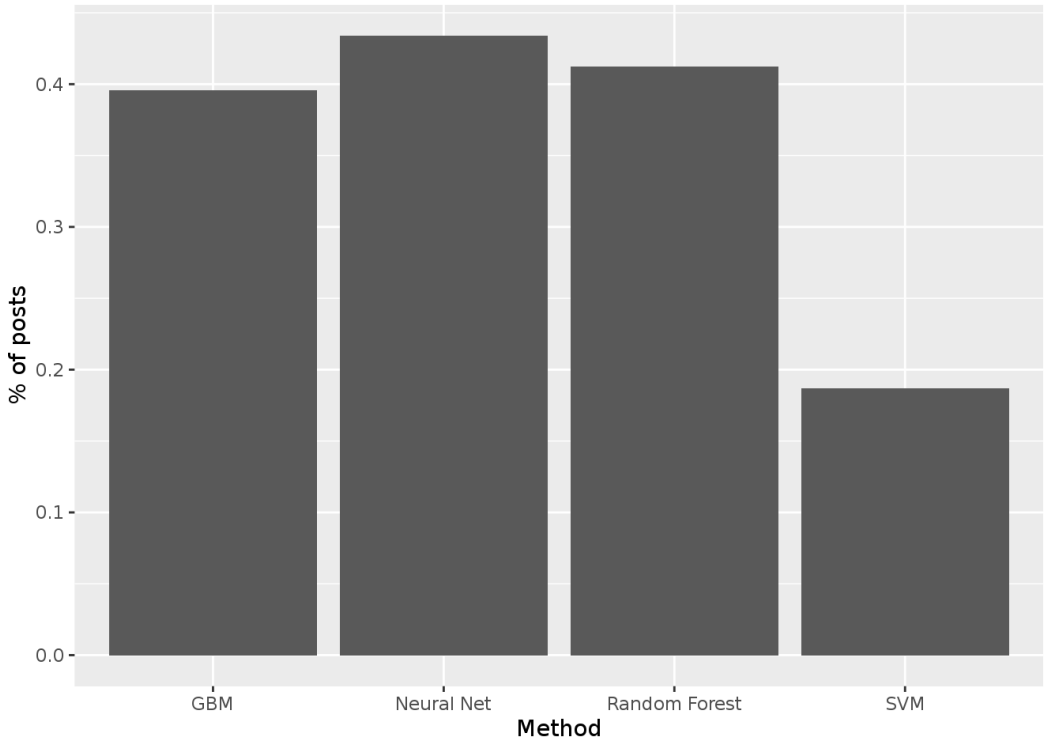

FIGURA 1. *Algoritmos que tienden a quedar en los primeros lugares en las competencias de Kaggle. Tomado de aquí.*

2. REMUESTREO O BOOTSTRAP

Utilizar técnicas de remuestreo nos permite cuantificar la variabilidad de un estimador estadístico sin necesidad de invocar un régimen asintótico para el procedimiento. Asimismo, nos permite controlar, hasta cierto punto, la variabilidad de nuestros estimadores.

Por ejemplo, consideremos la situación en donde tenemos una muestra de n observaciones $Z_1, \ldots, Z_n$ las cuales tienen una varianza σ^2. Es fácil demostrar que la varianza de la media $\bar{Z}_n$ tiene una varianza σ^2/n.

Esto quiere decir, que podemos promediar para reducir la varianza estimada.

2.0.1. Para pensar: Usualmente no tenemos acceso al proceso generador de datos (ya sea $\pi_{X,Y}$ ó π_X). ¿Qué estrategia podemos utilizar?

2.1. Bootstrap

Podemos utilizar la muestra $z_1, \ldots, z_n \overset{\text{iid}}{\sim} \pi$ como un *proxy* de la población de la cual queremos generar observaciones. En este sentido, consideramos que la función de acumulación empírica (`ECDF`, por sus siglas en inglés) sea un *buen* estimador de la función de probabilidad (ó `CDF` por sus siglas en inglés)

$$\pi[X \leq x] \approx \hat{\pi}_n[X \leq x] = \frac{1}{n}\sum_{i=1}^{n} I_{[z_i \leq x]}\,. \tag{1}$$

Con este procedimiento podemos generar B conjuntos de datos

$$z_1^{(b)}, \ldots, z_n^{(b)} \sim \hat{\pi}_n\,, \qquad b = 1, \ldots, B\,, \tag{2}$$

para obtener estimadores $\hat{\theta}_n^{(b)} = t(z_1^{(b)}, \ldots, z_n^{(b)})$ y, a través de un promedio, obtener un estimador

$$\bar{\theta}_{B,n}^{(\text{bag})} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_n^{(b)}\,, \tag{3}$$

con varianza que se reduce a una tasa $1/B$.

> El muestreo $z_1^{(b)}, \ldots, z_n^{(b)} \sim \hat{\pi}_n$ implica tomar muestras **con** reemplazo del conjunto de datos observado. Nota que las remuestras son del mismo tamaño que la muestra original. Es decir, cada remuestra b tiene n observaciones. Como el procedimiento es con reemplazo, esto puede ocasionar que pueda haber algunas observaciones que repitan en la remuestra.

La estrategia de remuestreo nos puede ayudar a cuantificar la estabilidad de ciertos estimadores. Por ejemplo, consideremos los datos que teníamos sobre el ingreso para un conjunto de 140 observaciones. El interés es construir un suavizador que relacione `Edad` con `Ingreso`. Utilizaremos un suavizador de *splines* con 15 grados de libertad, ver Fig. 2.

A través de remuestreo podemos cuantificar la estabilidad de dicha estimación.

REFERENCIAS

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. 1

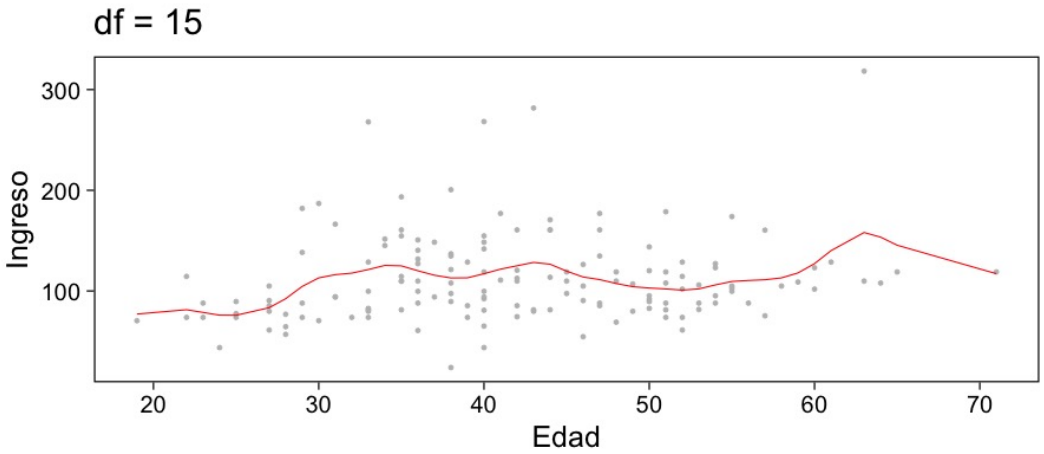

FIGURA 2. *Suavizador por* splines *con 15 grados de libertad.*

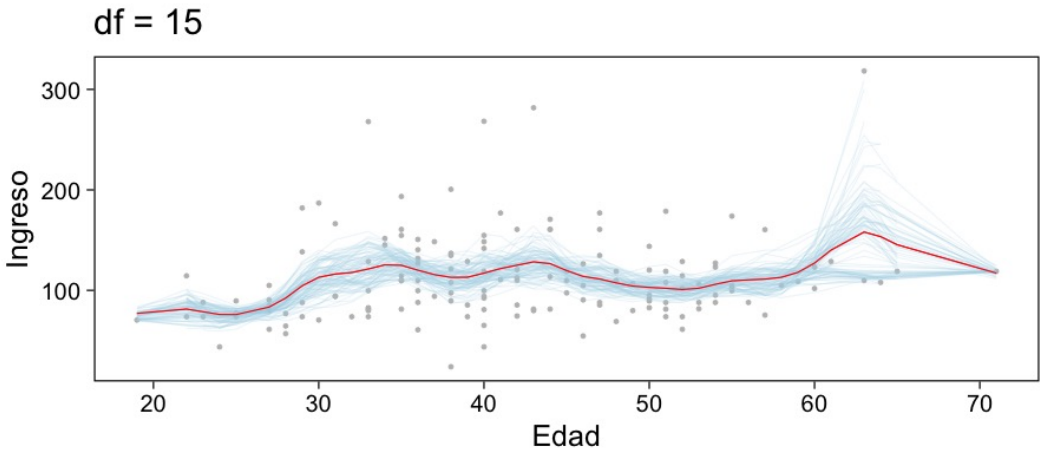

FIGURA 3. *Suavizador por* splines *con 15 grados de libertad, réplicas con remuestreo.*