Imputación de datos perdidos mediante técnicas de *Machine Learning*

Un experimento usando la Encuesta Permanente de Hogares

Germán Rosati
german.rosati@gmail.com

IDAES-UNSAM / PIMSA / UNTREF

07 de Agosto de 2019

Hoja de ruta

1. ¿Qué es y como se genera un dato perdido?
2. ¿Cómo lidiar con los datos perdidos?
   - Técnicas tradicionales (imputación simpe)
   - Técnicas basadas en *Machine Learning*
3. Metodología de imputación utilizada
4. Resultados y discusión

¿Qué es un valor perdido?

- Valor del que se carece una dato válido en la variable observada
- Problema generalizado en investigaciones por encuestas
- Problema cada vez más frecuente en investigaciones que usan registros administrativos o datos de redes sociales, aplicaciones, etc.
- ¿Cómo se generan esos datos perdidos?

Procesos de generación de valores perdidos

Ejemplos

MCAR	
X1	Y
0	NA
0	1
0	1
1	1
1	NA
2	NA
2	1
2	1
3	1
3	1
3	NA
3	1
4	1
4	NA
4	1
4	NA
4	1
4	1

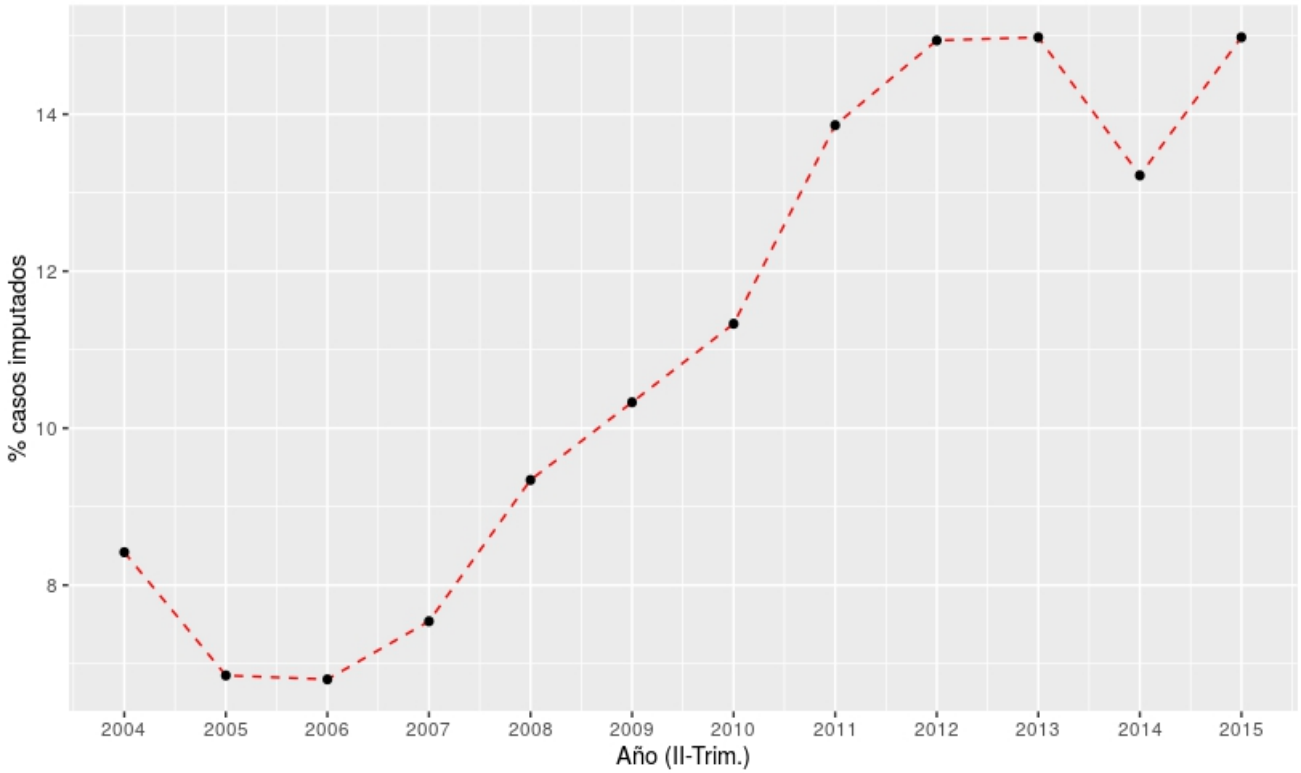
MAR	
X1	Y
0	1
0	1
0	1
1	1
1	1
2	1
2	1
2	1
3	1
3	NA
3	NA
3	1
4	1
4	NA
4	1
4	NA
4	1
4	NA

MAR	
X1	Y
0	1
0	1
0	1
1	1
1	1
2	1
2	1
2	1
3	1
3	1
3	1
3	1
4	1
4	NA
4	1
4	NA
4	NA
4	NA

¿Por qué es importante imputar datos?

Un ejemplo: EPH

Proporción de casos imputados (sin datos en alguna variable de ingresos) en EPH. Total de aglomerados urbanos, 2003-2015 (II-Trimestre de cada año)



¿Cómo lidiar con valores perdidos?

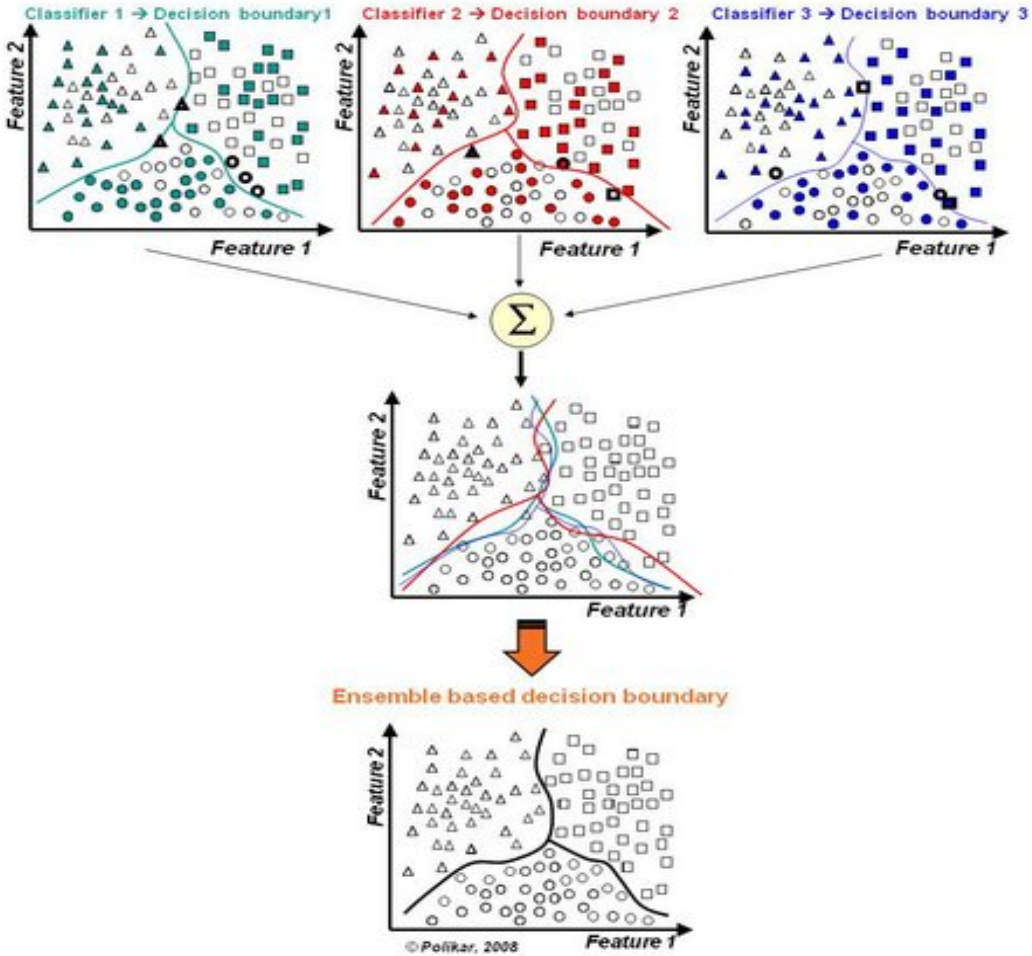
Imputación simple

- Exclusión de casos $\rightarrow$ se achica el dataset
- Reemplazo por la media o alguna otra medida $\rightarrow$ intervalos de confianza más estrechos de forma artificial
- Reponderación $\rightarrow$ es incómodo trabajar con varios sets de pesos.
- Hot Deck $\rightarrow$ problema en la selección de métrica de similitud y en la selección de los donantes

¿Cómo lidiar con valores perdidos?

Ensamble Learning

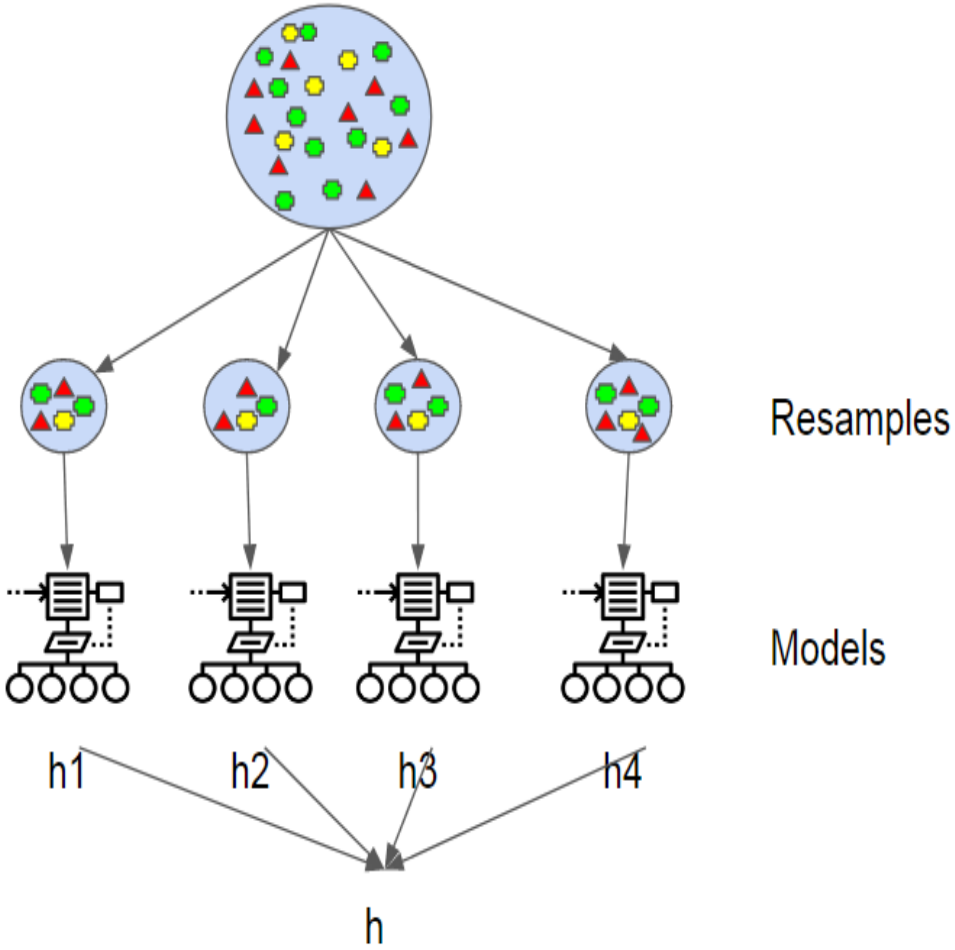
- Técnicas de aprendizaje supervisado donde se combinan varios modelos base.
- Ampliar el espacio de hipótesis posibles para mejorar la precisión predictiva del modelo combinado resultante.
- Los ensambles suelen ser mucho más precisos que los modelos base que los componen.



¿Cómo lidiar con valores perdidos?

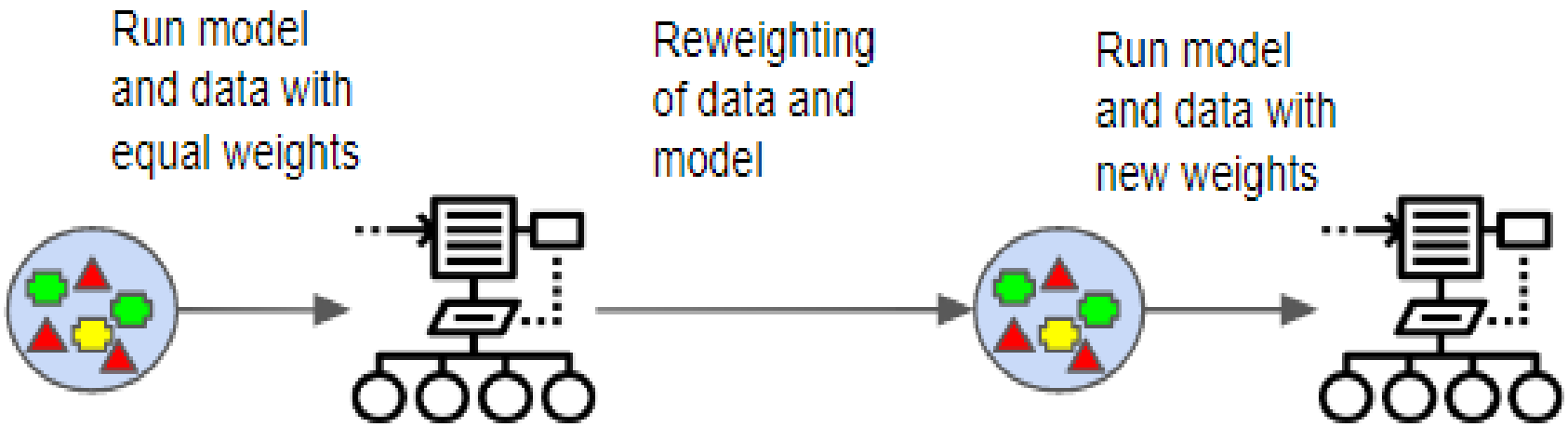
Ensamble Learning - Bagging

- Construcción de estimadores independientes -Boostrap-
- Combinación las predicciones mediante función agregación.
- Ejemplos: Random Forest, ExtraTrees, etc.



¿Cómo lidiar con valores perdidos?

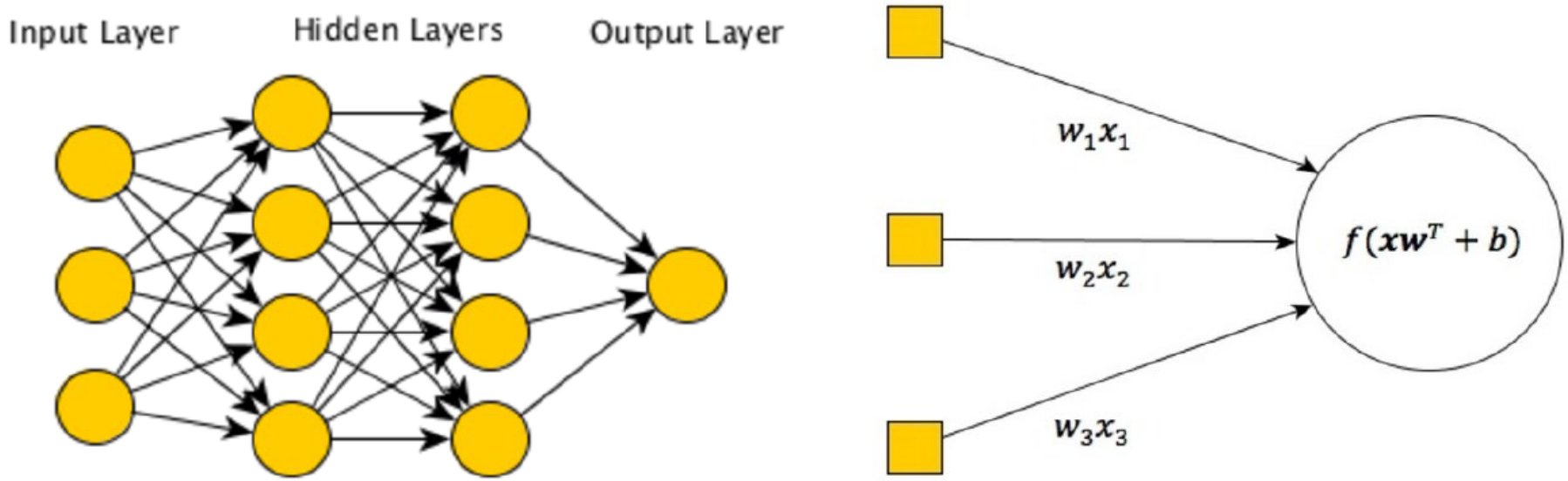
Ensamble Learning - Boosting



- Construcción secuencial de los estimadores
- Mayor peso en aquellos casos en los que se observa una peor performance.
- Ejemplos: AdaBoost y Gradient Tree Boosting, XGBoost.

¿Cómo lidiar con valores perdidos?

Ensamble Learning - Multi Layer Perceptron

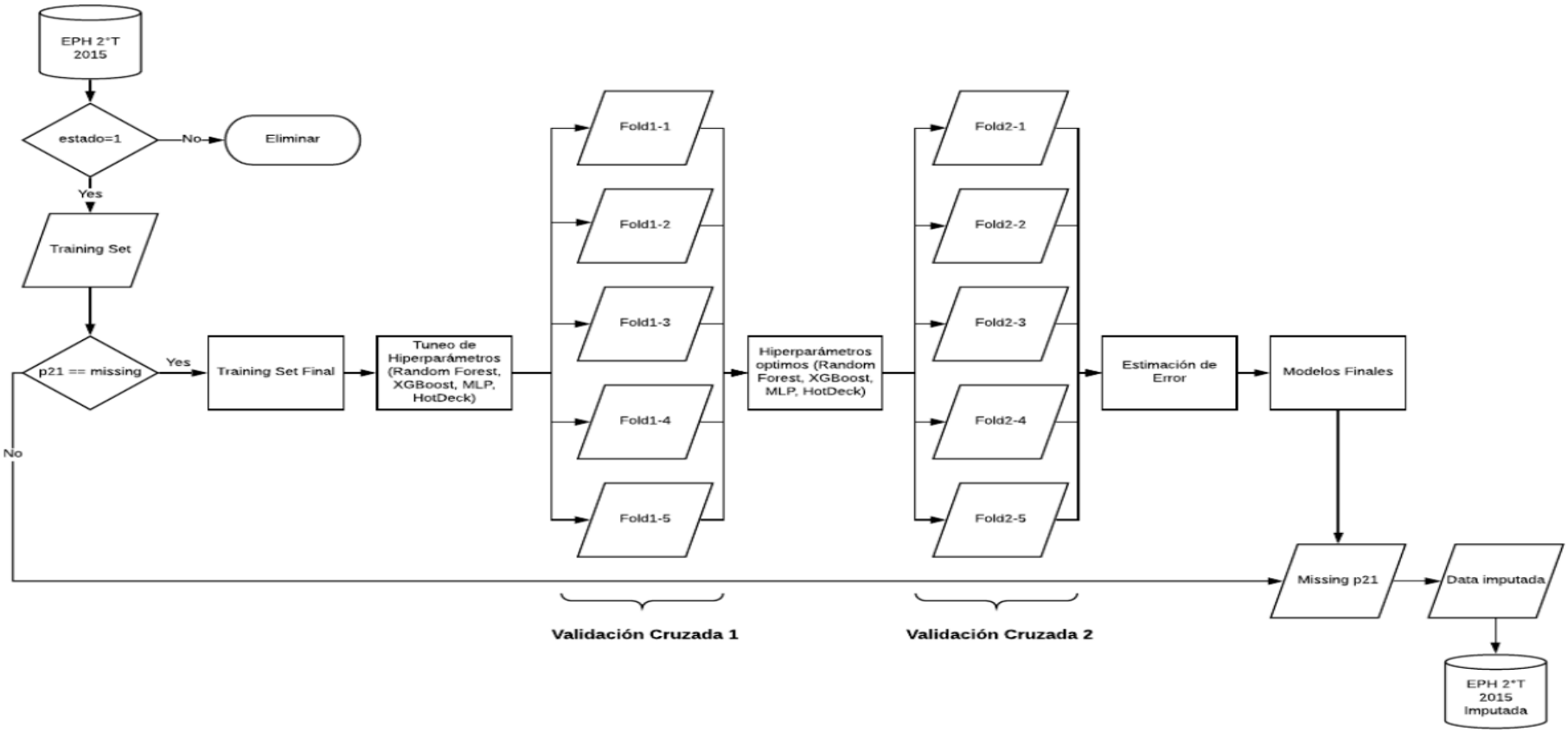


Fuente: https://technology.condenast.com/story/a-neural-network-primer

- Cada neurona aplica una transformación lineal $x_i w_i^T + b$ seguida de una función de activación
- Al apilar capas de neuronas se aplican sucesivas de transformaciones lineales que permiten la construcción de modelos altamente no lineales

Experimento con EPH

Pipeline



- Dataset: EPH 2do. trimestre de 2015
- Población: Ocupados en la semana de referencia
- Variables predictoras sociodemográficas, laborales y otros ingresos

Experimento con EPH

Estrategia de validación 1

- Estimación de métricas de error
- Supuesto: Proceso de generación de datos perdidos MCAR o MAR

Tabla 3. Métricas de performance predictiva de los diferentes algoritmos entrenadas

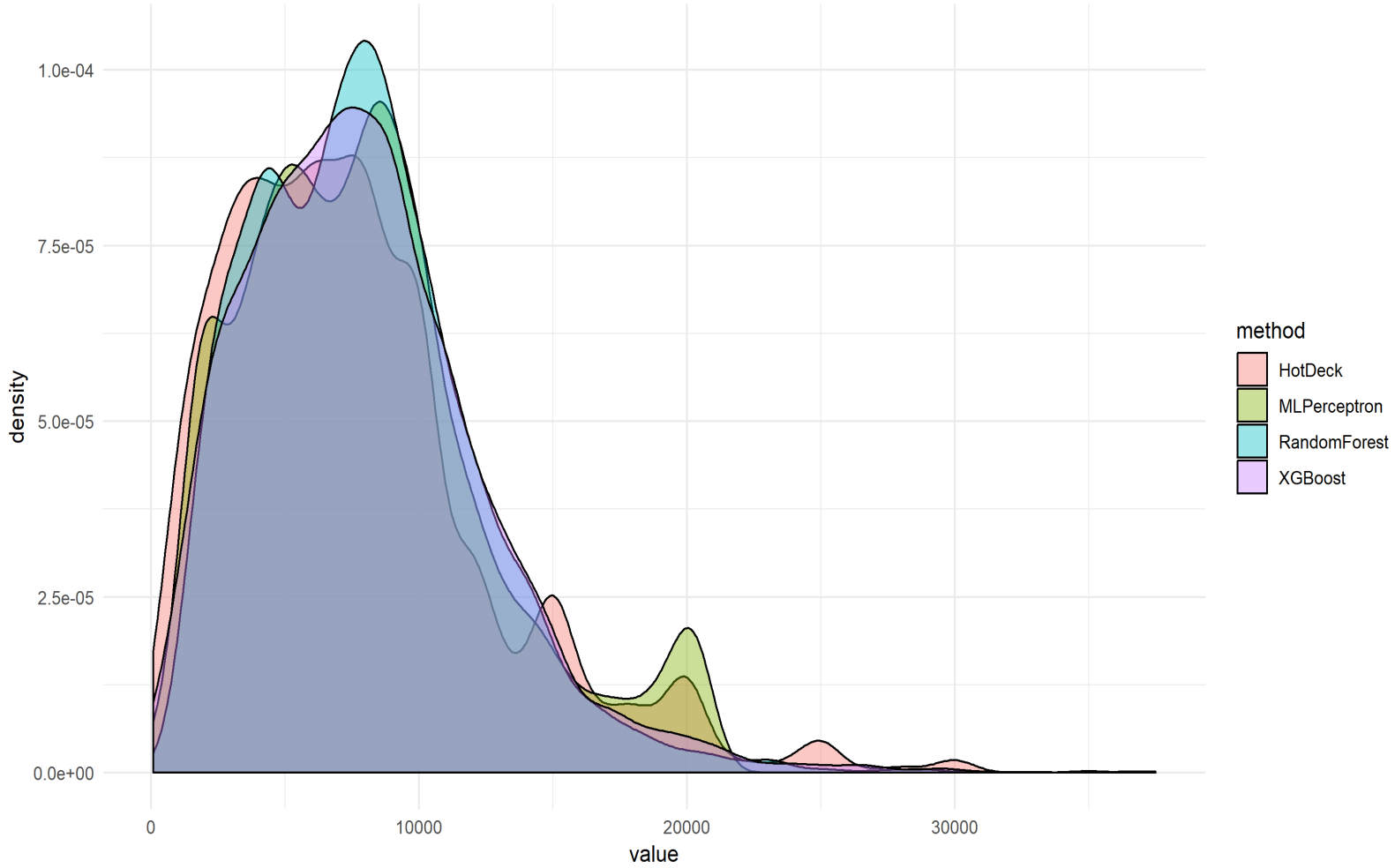
Algoritmo	RMSE	MAE
Hot Deck	$5930.6	$3740.6
Random Forest	$2800.6	$1561.9
XGBoost	$3260.8	$2016.8
MLP	$3974.2	$2293.1

Fuente: elaboración propia en base a microdatos de la EPH - 2do. trimestre de 2015

Experimento con EPH

Estrategia de validación 2

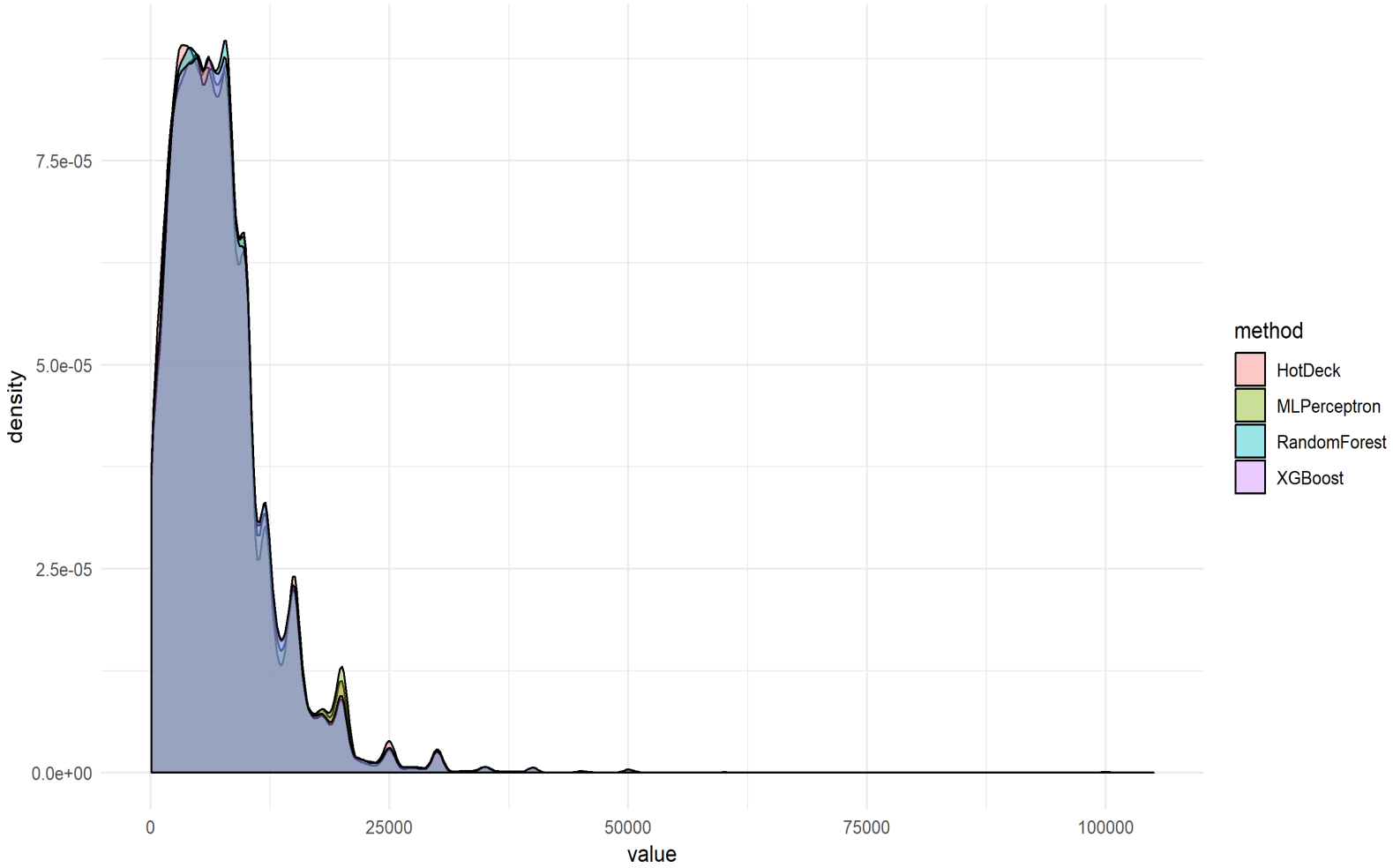
Comparación de distribuciones sobre datos perdidos reales (es decir, imputados por INDEC)



Experimento con EPH

Estrategia de validación 2

Comparación de distribución de datos completos (imputados + respuesta)



Resumen

- Machine Learning como alternativa para la imputación
- Reducción considerable en el *RMSE* entre casos perdidos comparado a Hot Deck -entre 30 % y 50 %-
- Problemas a futuro
  - Extensión del alcance del ejercicio
  - Mejoras en tuneo de hiperparámetros (algoritmos de búsqueda más inteligentes, diferentes funciones de activación, etc.)
  - Propiedades de los estimadores y estimaciones de medidas basadas en ingresos al utilizar estas técnicas
  - Performance relativa a HotDeck en procesos de generación de datos no aleatorios

¿Preguntas?

german.rosati@gmail.com