Maximizando el ROI de los datos mediante Imputación de valores perdidos con *Ensamble Learning*

Germán Rosati
german.rosati@gmail.com

UNTREF / MTEySS / Digital House

26 de Abril de 2017

Hoja de ruta

1. ¿Qué es y como se genera un dato perdido?
2. ¿Cómo lidiar con los datos perdidos?
   - Técnicas tradicionales (imputación simpe)
3. Ejercicio de aplicación

¿Qué es un dato perdido?

¿Qué es un valor perdido?

- Los datos son caros
  - Encuestas
  - SRMs
  - Logs, Analytics
  - Sistemas de Web Scrapping
- Los datos son muchos
- Los datos están sucios
  - Inconsistencias
  - Errores en la carga, escritura
  - Valores faltantes
  - No respuestas totales o parciales
- Cualquier error impacta en el valor que podemos extrear de un dataset

¿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

Missing Completely at Random -MCAR-

- La probabilidad de que registro tenga un valor perdido en la variable Y no está relacionada ni con los valores de Y, ni con otros valores de la matriz de datos (X)
- Los valores perdidos son una submuestra al azar de los valores totales
- ¿Cuándo no hay MCAR?
  1. Si algún grupo tiene mayor probabilidad de presentar datos perdidos en la variable Y y/o
  2. si alguno de los valores de Y tiene mayor probabilidad de presentar datos perdidos

Procesos de generación de valores perdidos

Missing at Random -MAR- y Missing Not at Random -MNAR-

- **MAR:** La probabilidad de no respuesta en Y es independiente de los valores de Y, luego de condicionar sobre otras variables
- **MNAR:** La probabilidad de no respuesta depende tanto de variables X externas, como de los valores de la variable con datos perdidos (Y)

Procesos de generación de valores perdidos

Resumen

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

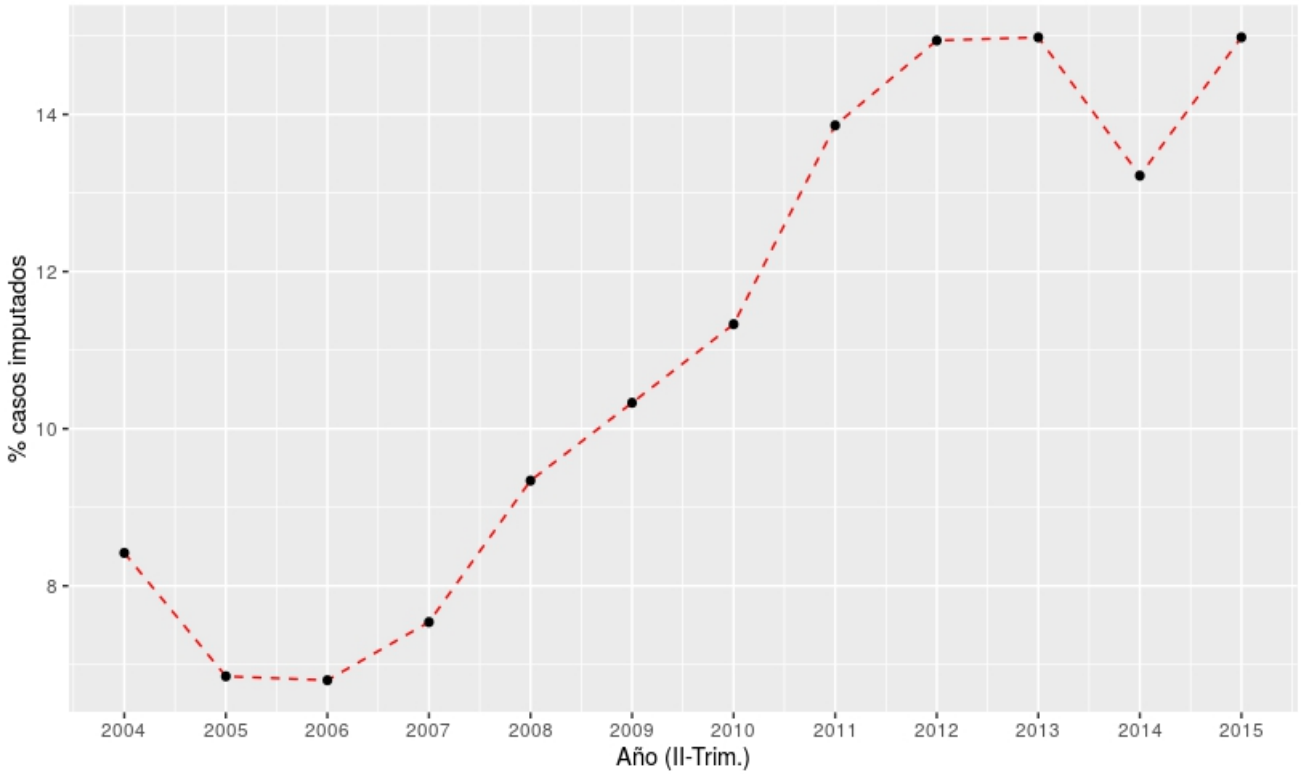
MAR	
X1	Y
0	1
0	1
0	1
1	1
1	1
2	1
2	1
2	1
3	1
3	NA
3	NA
3	1
4	1
4	NA
4	1
4	NA
4	1
4	NA

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 los datos perdidos?

¿Cómo lidiar con valores perdidos?

Imputación simple

- **Excluir los casos:** se trabaja solamente con los casos completos en toda la base o solamente en las variables de estudio. Problema: se achica el dataset.
- **Reemplazar por la media o alguna otra medida:** Problema: reducción de la variabilidad de la información y se generan intervalos de confianza más estrechos de forma artificial.
- **Reponderación:** se recalculan los ponderadores de la muestra (a partir de algoritmos de reweighting) para compensar el efecto de los casos con información faltante. Problema: es incómodo trabajar con varios sets de pesos.

¿Cómo lidiar con valores perdidos?

Métodos de imputación simple - Hot Deck

- Método ampliamente usado. INDEC -hasta 2015- y Dirección de Estadística de la Ciudad para realizar imputaciones en EPH y EAH
- Reemplazar valores perdidos de un no respondente ("receptor") con los valores observados de un respondente ("donante") que es similar al receptor.

¿Cómo lidiar con valores perdidos?

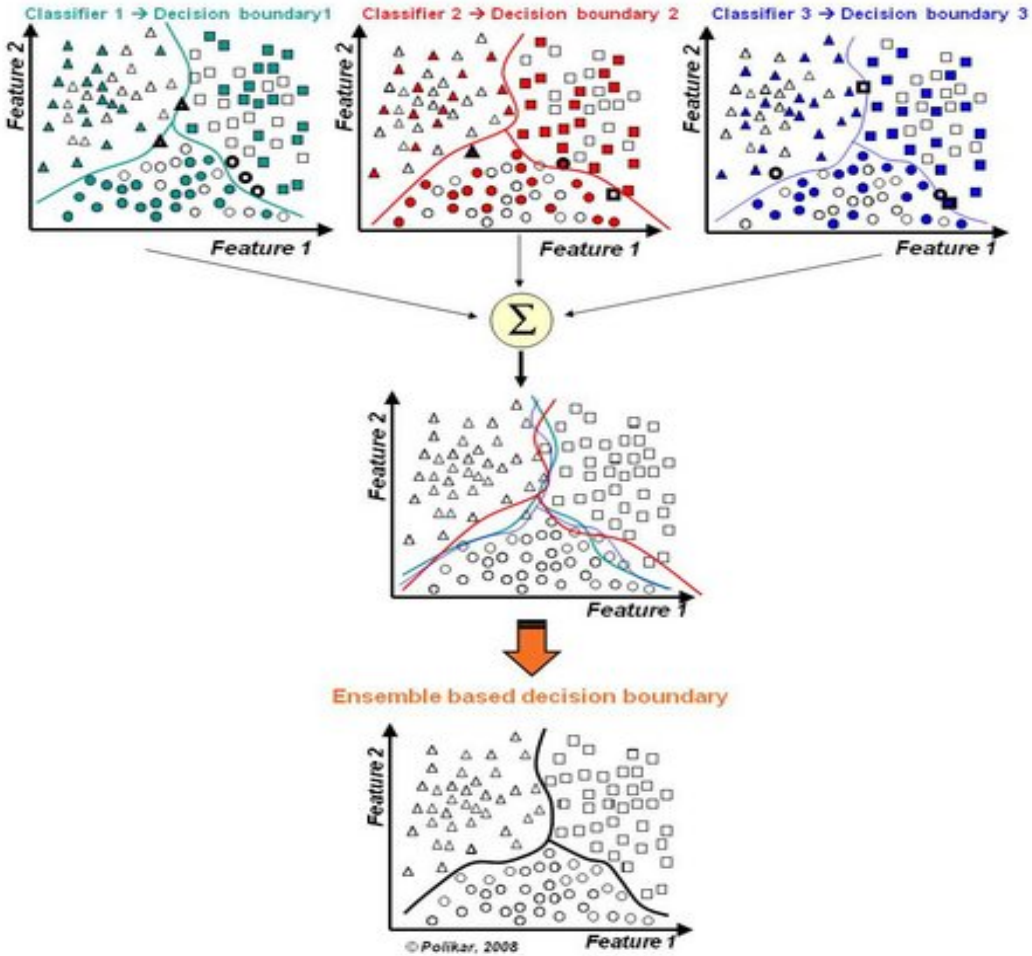
Métodos de imputación simple - Hot Deck

- **Problema 1:** selección de la "métrica" de similitud entre los casos
- **Problema 2:** selección de los donantes. El donante es seleccionado aleatoriamente de un set de potenciales donantes –hot-deck aleatorio- o bien se selecciona un solo caso donante, generalmente a partir de un algoritmo de "vecinos cercanos" usando alguna métrica -hot-deck determinístico-.

¿Cómo lidiar con valores perdidos?

Ensamble Learning

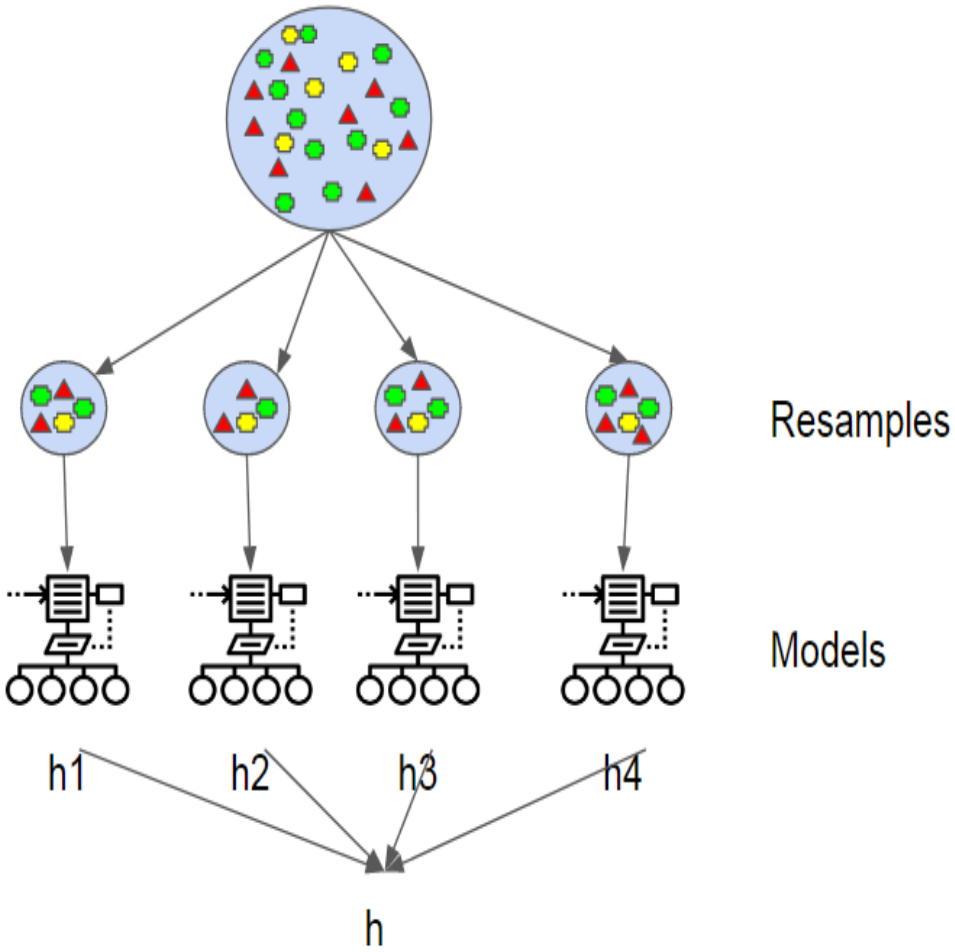
- Técnicas de aprendizaje supervisado donde se combinan varios modelos base.
- Ampliar el espacio de hipótesis posibles para mejorar la precisión predictiva del modelo combinado resultante.
- Los ensambles suelen ser mucho más precisos que los modelos base que los componen.



¿Cómo lidiar con valores perdidos?

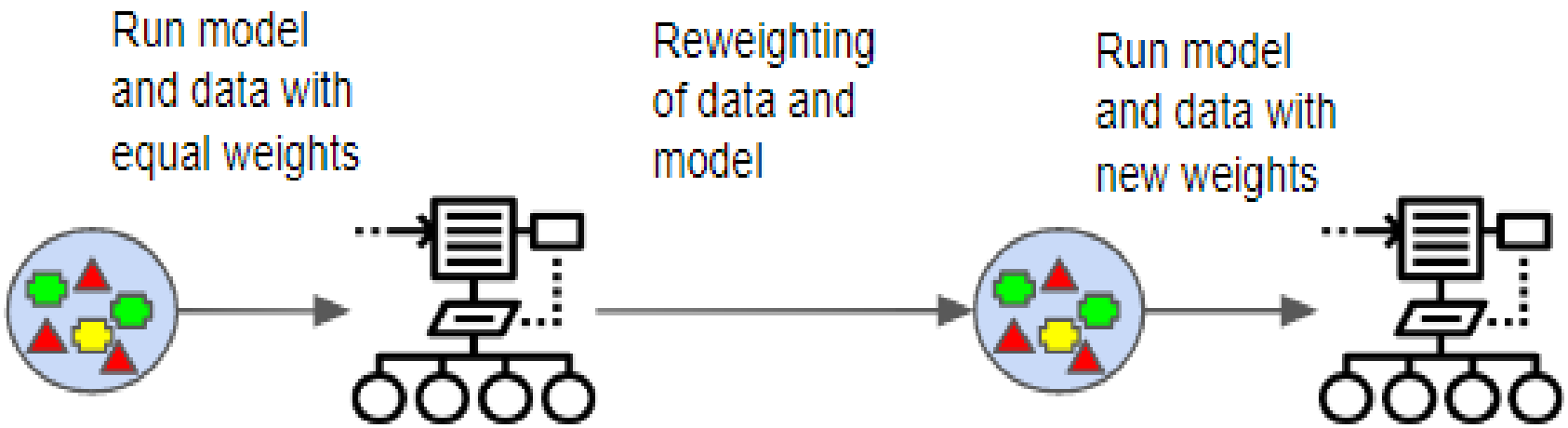
Ensamble Learning - Bagging

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



¿Cómo lidiar con valores perdidos?

Ensamble Learning - Boosting

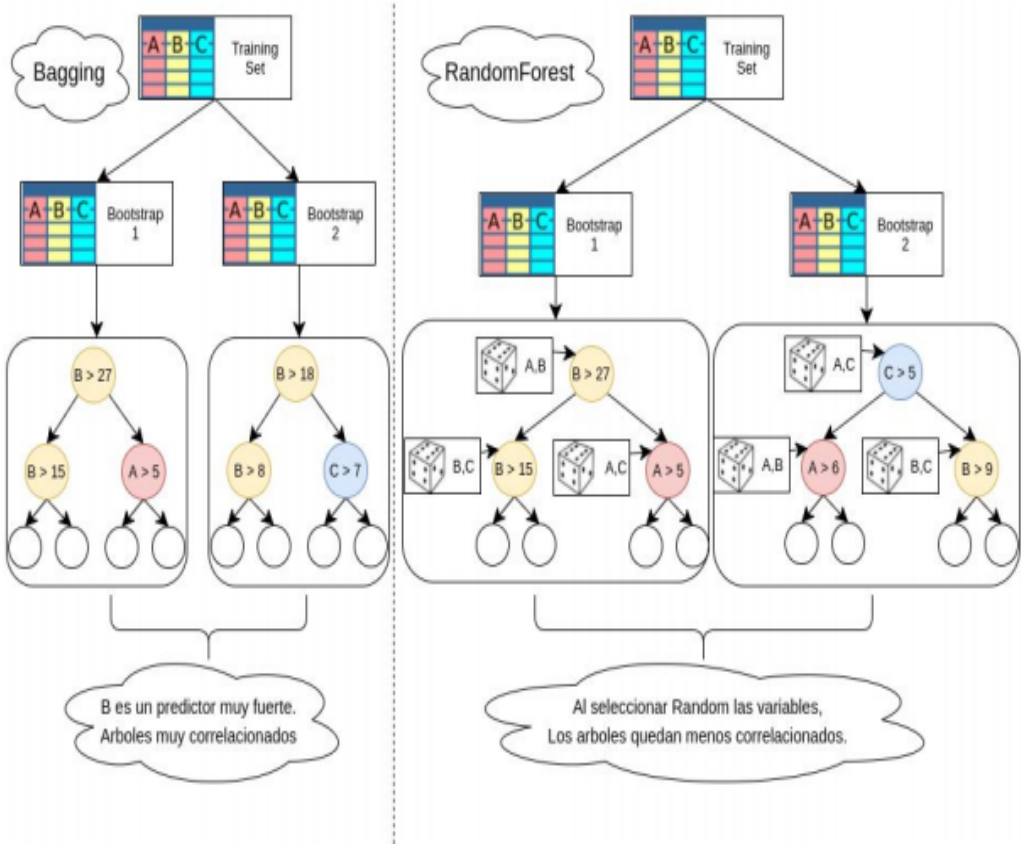


- Construcción secuencial de los estimadores
- Mayor peso en aquellos casos en los que se observa una peor performance.
- Ejemplos: AdaBoost y Gradient Tree Boosting, XGBoost.

¿Cómo lidiar con valores perdidos? - Método posible 1

Ensamble Learning - Random Forest

- Variación del algoritmo Bagging
- Modelo base: árboles de decisión
- Feature Bagging, en cada iteración y en cada split del árbol de decisión, el algoritmo selecciona aleatoriamente un subconjunto de variables predictoras



Un ejercicio de aplicación

LAB: Imputando datos perdidos con Random Forest

- Dataset: EPH 2do. trimestre de 2015
- Población: Ocupados en la semana de referencia
- Objetivo: Generar un imputador de la variable ingresos
- Variables predictoras sociodemográficas, laborales y otros ingresos
- Pipeline
  1. Partición Train-Test
  2. Sobre Train: entrenamos un clasificador Random Forest
  3. Sobre Train: entrenamos dos imputadores Hot-Deck
  4. Sobre Test: evaluamos los resultados

¿Cómo lidiar con valores perdidos? - Método posible 2

Bagging-LASSO

SaberEs

REVISTA DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

INICIO ACERCA DE INICIAR SESIÓN REGISTRARSE BUSCAR ACTUAL ARCHIVOS AVISOS

Inicio > Vol. 9, Núm. 1 (2017) > **Rosati**

SciELO Scientific Electronic Library Online

latindex CATALOGO

Núcleo Básico de Revistas Científicas Argentinas NB

REDIB Red Iberoamericana de Innovación y Conocimiento Científico

Dialnet

EconBib

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning. Aplicación en la Encuesta Permanente de Hogares (EPH)

Germán Federico Rosati

Resumen

El presente documento se propone exponer los avances realizados en la construcción de un modelo de imputación de valores perdidos y sin respuesta para las variables de ingreso en encuestas a hogares. Se presentará la propuesta metodológica general y los resultados de las pruebas realizadas. Se evalúan dos tipos de modelos de imputación de datos perdidos: 1) el método hot-deck (ampliamente utilizado por relevamientos importantes en el Sistema Estadístico Nacional, tales como la Encuesta Permanente de Hogares y la Encuesta Anual de Hogares de la Ciudad de Buenos Aires) y 2) un ensamble de modelos de regresión LASSO (Least Absolute Shrinkage and Selection Operator). El mismo se basa en la generación de múltiples modelos de regresión LASSO a través del algoritmo bagging y de su agregación para la generación de la imputación final. En la primera y segunda parte del documento plantea el problema de forma más específica y se pasa revista a los principales mecanismos de generación de los valores perdidos y las implicancias que los mismos tienen al momento de generar modelos de imputación. En el tercer apartado se reseñan los métodos de imputación más habitualmente utilizados, enfatizando sus ventajas y limitaciones. En la cuarta parte, se desarrollan los fundamentos teóricos y metodológicos de las dos técnicas de imputación propuestas. Finalmente, en la quinta sección, se presentan algunos resultados de la aplicación de los métodos propuestos a datos de la Encuesta Permanente de Hogares.

Palabras clave

Regularización; LASSO; No respuesta

¿Cómo lidiar con valores perdidos? - Método posible 2

Bagging-LASSO

- Se aplica el algoritmo bagging a la imputación de ingresos laborales en la EPH del II trimestre de 2015
- En cada remuestra se estima la siguiente regresión LASSO

$$log_{10}(y_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + e_i \tag{1}$$

- Buscando minimizar la siguiente función de costo:

$$CF = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

¿Cómo lidiar con valores perdidos? - Método posible 2

Bagging-LASSO

1. Se crea un dataset con todos los casos completos *TrS*
2. Se crea otro dataset con los casos son datos perdidos *TeS*
3. Se fija la cantidad de iteraciones *rep*
4. Para r entre 1 y *rep*
   1. Se extrae una muestra boostrap (MAS con reemplezo y $n = n^*$) de *TrSet*
   2. En la muestra generada se estima una regresión LASSO
   3. Con los parámetros estimados en el paso anterior se realiza la predicción en el *TeS*
5. Luego de *rep* iteraciones se generan *rep* predictiones de los casos perdidos en *TeS* y se agregan usando la mediana

¿Cómo lidiar con valores perdidos? - Método posible 2

Bagging-LASSO

- Generación de datos perdidos de forma aleatoria
- Para evaluar performance de Bagging-LASSO y hot-deck se estima el *RMSE* a través de *k-fold cross validation*, con $k = 9$
  1. $RMSE - LASSO = 3{,}994\$$
  2. $RMSE - hotdeck = 4{,}933\$$
- Es decir, una reducción de alrededor del 20 %
- La imputación Bagging-LASSO mejora considerablemente el error de predicción de los ingresos laborales
- Esperable que en datos perdidos "originales" (es decir, no generados artificialmente) el método consiga una mejor performance que hot-deck

Resumen

- Procesos generadores de datos perdidos
- Técnicas "tradicionales"
- Ensamble Learning como alternativa
- Random Forest logra una reducción considerable en el *RMSE* entre casos perdidos comparado a Hot Deck
- Baggin-LASSO también logra una mejora sustantiva

¿Preguntas?

german@digitalhouse.com
german.rosati@gmail.com