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 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 presentarán 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

* Docente en Universidad Nacional de Tres de Febrero (UNTREF), Argentina; Consultor Analista Experto de Datos en Ministerio de Trabajo, Empleo y Seguridad Social, Argentina; Investigador en Programa de Investigación sobre el Movimiento de la Sociedad Argentina (PIMSA).
Contacto: german.rosati@gmail.com

DEVELOPMENT OF AN IMPUTATION MODEL FOR INCOME VARIABLES WITH LOST VALUES USING ENSAMBLE LEARNING METHODS. APPLICATION TO THE PERMANENT HOUSEHOLD SURVEY (EPH)

Abstract. This paper aims to present some advances made in the development of a missing values and non-response imputation model for income variables in household surveys. The general methodological propose is exposed and the results of some tests. Two imputation methods are evaluated: 1) hot deck (widely used in mayor surveys such as Encuesta Permanente de Hogares and Encuesta Anual de Hogares of the Buenos Aires City) and 2) a LASSO regression model ensamble. The ensamble is generated using the bagging algorithm. The first and second part of the document reviews the main missing data generation mechanisms and its implications for the use of imputation methods. In the third section, several imputation methods are reviewed, emphasizing its assumptions, advantages and limitations. The fourth part analyzes the theoretical and methodological foundations of LASSO and ensamble learning. Finally, the fifth section presents some results of the application of this method to Encuesta Permanente de Hogares data.

Keywords: Regularization; LASSO; Non response.

Original recibido el 22-11-2016
Aceptado para su publicación el 02-03-2017

1.Introducción

Parece obvio remarcarlo pero la presencia de no respuestas y valores perdidos constituyen un problema recurrente e histórico en el análisis estadístico. Se trata de un obstáculo que afecta no solamente a las estadísticas oficiales (encuestas a hogares, datos censales, etc.) sino también a los registros administrativos (de empresas u organismos) y, en términos generales, a cualquier conjunto de datos sobre el cual se busque realizar algún análisis estadístico. La posibilidad de utilizar la información proveniente, por ejemplo, del tráfico de internet, de las aplicaciones en dispositivos móviles (celulares, tablets, etc.) o del *scrapping* de sitios web, replantean en una escala nueva el problema de la generación herramientas que permitan lidiar con la existencia de datos perdidos (y otras inconsistencias que no serán objeto del presente trabajo) de forma eficiente.

Es por ello que contar con métodos que puedan lidiar con el problema de los valores perdidos ha sido (y continúa siendo) una necesidad en el análisis estadístico moderno. Este punto tiene particular relevancia dado que:

> las rutinas de los paquetes estadísticos asumen que se trabaja con datos completos incorporan opciones –no siempre las más adecuadas– para imputar observaciones sin que el usuario se dé cuenta de ello. […] la aplicación de procedimientos inapropiados […] introduce sesgos y reduce el poder explicativo de los métodos estadísticos (Medina y Galvan, 2007, p.10).

El presente documento se propone presentar un primer modelo de imputación de valores perdidos y sin respuesta para las variables de ingreso en encuestas a hogares. El mismo se basa en dos técnicas no demasiado utilizadas aún en ciencias sociales o económicas en el país: modelos de regresión LASSO (*Least Absolute Shrinkage and Selection Operator*) y ensamble *learning*, particularmente a partir del algoritmo *bagging*[1] .

El método desarrollado debería resultar aplicable para la imputación de cualquier tipo de valores perdidos (en variables cualitativas o cuantitativas) y para diversas fuentes de datos. Sin embargo, dada la relevancia particular que presenta el problema de la no respuesta de ingresos en encuestas a hogares (tanto en la Argentina como en la región) se presenta aquí una aplicación para la imputación de variables de ingresos en una onda de la Encuesta Permanente de Hogares (EPH) relevamiento elaborado por el Instituto Nacional de Estadísticas y Censos de la Argentina (INDEC).

El problema no es nuevo y, de hecho, la EPH ha visto incrementarse la proporción de valores de no respuesta totales particularmente en variables de ingreso. Los diversos estudios (Salvia y Donza,1999; Felcman, Kidyba y Ruffo,2004; Pacífico, Jaccoud, Monteforte y Arakaki, 2011) parecen mostrar que la proporción de perceptores de ingresos con ingresos no declarados varió del 8% en 1995 al 24% en 2010 (luego de un descenso entre 1990 y 1994). En ese sentido, el INDEC ha encarado el problema de diferentes formas en la EPH. Durante la EPH en su modalidad virtual, se optó por el método *pairwise;* luego, durante la primera etapa de la EPH-Continua se

1 Se agradecen los comentarios de Javier Burroni a versiones previas de este trabajo. A su vez, Tomás Olego planteó en una charla informal la posibilidad de aplicar un bagging de regresiones LASSO como método general, más allá de la imputación de valores perdidos.

empleó la reponderación de ingresos combinado con la técnica de *hot deck*; y, por último, a finales de 2015 se ha retomado el método de la reponderación (Camelo, 1999; Hoszowski, Messere y Tombolini, 2004; INDEC, 2009).

Para poder repensar estos problemas y lograr el objetivo principal se ha dividido al presente trabajo en cuatro apartados. En la primera parte del documento 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 la segunda se reseñan los métodos de imputación más habitualmente utilizados, sus ventajas y limitaciones. En el tercer apartado se desarrollan los fundamentos teóricos y metodológicos de la técnica de imputación propuesta, junto con una descripción general del algoritmo utilizado. Finalmente, se exponen algunas pruebas realizadas para evaluar su capacidad predictiva comparándolo con el modelo *hot deck*[2].

2.Mecanismos de generación de valores perdidos

Un primer problema a tener en cuenta se vincula con los modelos teóricos que se asumen como generadores de los valores perdidos. En términos generales existen tres procesos generadores de datos perdidos:

1) *Missing Completly at Random* (MCAR): en este caso, la probabilidad de que un registro tenga un valor perdido en la variable Y no está relacionada con los valores de Y ni con otros valores dela matriz de datos (X´s). Es decir, los datos perdidos son una submuestra aleatoria de la muestra general. Este supuesto se viola si: a) algún grupo o subgrupo tiene mayor probabilidad de presentar NR (no respuesta) en la variable Y; y/o b) si alguno de los valores de Y tiene mayor probabilidad de NR.

2) *Missing at Random* (MAR): si la probabilidad de NR en Y es independiente de los valores de Y, por lo tanto de condicionar sobre otras variables.

3) *Non Missing at Random* (NMR): en este caso, la probabilidad de NR depende tanto de variables X´s externas, como de los valores de la variable con datos perdidos (Y).

El supuesto de MAR sería satisfecho si la probabilidad de no respuesta en ingresos dependiera de, por ejemplo, el nivel educativo: es decir, si hubiera una mayor probabilidad en los niveles educativos altos de no haber respondido la variable ingresos. Bajo el proceso MAR, si bien existe esa probabilidad diferencial en cada grupo, al interior de cada uno de ellos (en el ejemplo anterior el nivel educativo) la probabilidad de no respuesta en ingresos no está relacionada con los valores del ingreso: dentro de los niveles educativos altos, todos los individuos tienen la misma probabilidad de presentar NR en la variable ingresos. En términos generales, los datos perdidos no son generados por un proceso MAR si los casos con NR en una variable particular tienden a tener mayores o menores valores en esa variable que los casos con datos no perdidos y se trataría de un proceso NMR.

Ahora bien, se le llama “ignorable” al proceso generador de datos cuando el mismo

2 Este modelo fue propuesto como otro mecanismo de imputación de datos perdidos y se usó ampliamente en 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.

es MAR o MCAR. Esto implica que no es necesario "modelar" de forma explícita el mecanismo generador de los datos perdidos en el proceso de estimación de los parámetros. Si los datos no son MAR o MCAR, entonces, se dice que el proceso generador de datos perdidos es "no ignorable". La gran mayoría de los mecanismos de imputación parten del supuesto de que los datos perdidos han sido generados por un proceso MAR.

3.Métodos de imputación más habituales

¿Cómo lidiar con datos perdidos? Existen dos grandes estrategias generales para afrontar el problema. La primera supone la exclusión de los casos con valores perdidos. Es decir, se eliminan de la estimación de los parámetros aquellos registros que contengan valores perdidos. Se trata de la forma más simple. Existen dos variantes de esta estrategia:

1) Exclusión *listwise:* se trabaja solamente con los casos completos en toda la base. De esta forma, se reduce el tamaño de muestra y se asume que los valores perdidos tienen igual distribución que los no perdidos.
2) Exclusión *pairwise:* emplea solamente los datos completos de cada variable. Se generan diferentes tamaños de muestra en cada una de las estimaciones.

La segunda estrategia consiste en el reemplazo de los valores perdidos por algún valor estimado. Es decir, se busca generar mecanismos de imputación. Es posible identificar dos grandes tipos de mecanismos de imputación: los basados en "imputación simple" y los basados en "imputación múltiple". Entre los primeros podemos reseñar:

1) Medias no condicionadas: se imputa la media (de los casos completos). Se asume un patrón MCAR. El impacto de este mecanismo es la reducción de la varianza y la generación de intervalos de confianza más estrechos de forma artificial.
2) Reponderación: se recalculan los ponderadores de la muestra (a partir de algoritmos de *reweighting)* para compensar el efecto de los casos con información faltante.
3) Medias condicionadas: se forman categorías a partir de covariables correlacionadas con la variable de interés, y se imputan los datos omitidos con observaciones provenientes de la submuestra que comparte características. Se asume un patrón MCAR y existirán tantos promedios como categorías se formen, lo cual contribuye a atenuar los sesgos en cada celda pero de ninguna manera los elimina. Este método resulta equivalente a correr una regresión.
4) *Hot deck*: se busca reemplazar los valores perdidos de una o más variables de un no respondente ("receptor") con los valores observados de un respondente ("donante") que es similar al receptor. En algunas versiones el donante es seleccionado aleatoriamente de un set de potenciales donantes (*random hot deck)*; en otros casos se selecciona un solo caso donante, generalmente a partir de un algoritmo de "vecinos cercanos" usando alguna métrica (*deterministic hot deck)*.

En todos estos casos (y muchos otros métodos que no se mencionan aquí) la imputación del valor perdido se realiza a partir de un solo valor estimado. Los métodos basados en las llamadas "imputaciones múltiples" generan un conjunto de posibles

valores como estimación de los valores a imputar, los cuales son agregados de alguna manera. En general, se utilizan métodos de simulación de Monte Carlo y se sustituyen los datos faltantes a partir de un número (mayor a 1) de simulaciones. “La metodología consta de varias etapas, y en cada simulación se analiza la matriz de datos completos a partir de métodos estadísticos convencionales y posteriormente se combinan los resultados para generar estimadores robustos” (Medina y Galván, 2007, p. 31).

El método propuesto en este documento se basa en una lógica similar a la de imputación múltiple, es decir, generará varias estimaciones para los valores perdidos a imputar y las agregará para generar el valor de imputación final. A su vez, se propone realizar una combinación de algunas técnicas modernas de regresión con algoritmos utilizados por los métodos de aprendizaje automático (*machine learning*), particularmente, con los llamados ensambles de modelos o ensamble *learning.* Dichos métodos (uno de los cuales se desarrolla en el siguiente apartado) buscan lograr una mejora en la capacidad predictiva de los clasificadores/modelos utilizados a partir de la generación de sucesivas submuestras y reestimación de los modelos. Particularmente, se aplicará una versión del algoritmo *bagging (Bootstrap Aggregating*) (Breiman, 1996).

4.Descripción de las técnicas de imputación propuestas

4.1.Regresión LASSO

Los modelos de regresión LASSO (Hastie, Tibshiriani y Wainwrigh, 2015; Tibshiriani, 1996) se basan en el modelo lineal múltiple y buscan lograr la “regularización” del mismo. Supongamos que partimos de la siguiente expresión correspondiente a un modelo lineal:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\ \beta_j + e_i$$

El objetivo, entonces, es predecir el resultado (y_i) a partir de la expresión anterior dónde β_0 y $\beta=(\beta_1,...,\beta_p)$ son los parámetros a estimar y e_i es un término de error aleatorio. El método de Mínimos Cuadrados Ordinarios (MCO) estima los parámetros mencionados a partir de la minimización de la siguiente función objetivo (también llamada en la jerga del aprendizaje automático “función de pérdida”, *loss function*):

$$min\ \beta_0, \beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j\ x_{ij} \right)^2$$

Es decir, busca minimizar la suma de los cuadrados de los residuos (RSS, por sus siglas en inglés). Típicamente, todos los parámetros $\beta=(\beta_1,\ldots,\beta_p)$ serán diferentes a cero. Esto hace que el modelo sea difícil de interpretar si p (la cantidad de parámetros) es muy grande. Es más: si *p>n* los estimadores mínimo-cuadráticos no son únicos. En este caso, habrá un conjunto infinito de modelos que igualan a cero la

función objetivo. Este conjunto de modelos muy probablemente sobreajusten[3] a los datos. Esto se debe a que, en líneas generales, los estimadores de mínimos cuadrados tienen bajo sesgo (*bias*) pero alta variancia.

La precisión predictiva de un modelo de regresión puede ser incrementada a través del encogimiento de los valores de los coeficientes o, incluso, haciéndolos cero. Haciendo esto, se introduce algún sesgo pero se reduce la variancia de los valores predichos y, por lo tanto, se incrementa la precisión predictiva total. En muchos casos, cuando existen muchos predictores puede ser necesario identificar un subconjunto más pequeño de estos predictores que muestren los efectos más grandes.

Es por ello que resulta útil imponer restricciones en el proceso de estimación. A esta operación se la define como "regularización". Existen varios métodos de regularización de modelos lineales: *non negative garrotte* (Breiman,1995); *ridge regression* (Hoerl y Kennard, 1970). Este trabajo se centrará en el LASSO. Este método utiliza la norma ℓ_1 como me dida de penalización para definir las restricciones al modelo lineal. LASSO busca minimizar la siguiente expresión:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j\,x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

Se parte de la minimización de la RSS y se agrega una restricción: el segundo término $\lambda\sum_{j=1}^{p}|\beta_j|$ se hace pequeño cuando los coeficientes son pequeños y, por lo tanto, tiene el efecto de reducir los coeficientes β estimados. λ constituye un parámetro de *tunning* y su función es controlar el impacto relativo de ambos términos. Cuando el coeficiente es igual a cero ($\lambda=0$), LASSO es equivalente a un modelo lineal estimado por MCO.

Por el contrario, a medida que el parámetro se hace más grande (λ), el término de restricción lo hace en igual proporción y todos los coeficientes se reducen para poder satisfacer dicha restricción. En el límite, cuando λ es lo suficientemente grande todos los coeficientes se hacen igual a cero y solamente queda como parámetro el intercepto (β_0), es decir, algo casi equivalente a predecir y_i solamente con la media de la distribución.

A su vez, la utilización de la norma ℓ_1 como medida de las distancias (absolutas) entre los coeficientes provoca que algunos de los coeficientes β se hagan cero para determinados valores de λ. Esta propiedad es definida como *sparsity* y tiene va-

3 El sobreajuste (u *overfitting*) se produce como consecuencia de sobreentrenar un algoritmo de aprendizaje automático o un modelo de predicción sobre un conjunto de datos sobre el que se conoce el valor de la variable a predecir. En general, se busca un modelo o algoritmo que logra una buena performance predictiva en datos "nuevos", es decir, que permita generalizar la predicción a datos no observados previamente. Cuando se produce el sobreentrenamiento del modelo, existe la posibilidad de que el mismo ajuste "demasiado bien" a los datos de entrenamiento y, por ende, no capte la verdadera señal de los datos, sino que la confunda con ruidos y errores aleatorios de los datos. Como consecuencia, el modelo presenta un elevado ajuste en los datos de entrenamiento pero una mala performance en datos "nuevos". Existe amplia bibliografía al respecto, por ejemplo, ver Hastie, Tibshirani y Friedman, 2009.

rias propiedades teóricas y computacionales útiles (Hastie, Tibshiriani y Wainwright, 2015).

Este punto marca una diferencia con otro método de regularización ampliamente utilizado: la regresión *Ridge*. Esta regresión hace uso de la norma ℓ_2, en la cual el término que define la penalización es $\lambda\sum_{j=1}^{p}\beta_j^2$, por lo que todos los coeficientes se reducen pero ninguno llega a presentar valores iguales a cero. De esta forma, puede verse que en la regresión LASSO para cada valor de λ existe un set diferente de coeficientes β estimados, por ende, un modelo diferente. Esto hace que sea crucial la selección de un correcto valor de λ. Generalmente, la determinación de su valor se realiza a través de procesos de validación cruzada.

4.2.Ensamble Learning

La propuesta de imputación construida combina los modelos de regresión LASSO con una técnica del llamado ensamble *learning* (o ensamble de modelos, "clasificadores basados en comités" o "sistemas de clasificadores múltiples"). El objetivo general de los ensambles de modelos es incrementar la capacidad predictiva de clasificadores/modelos (*base learners*) a partir de la generación de submuestras de los datos originales y la estimación para cada una de esas submuestras de un modelo. Luego, las estimaciones provenientes de cada uno de esos modelos generados se agregan de alguna manera y se obtiene la estimación final. De esta manera, se obtiene una capacidad predictiva que puede ser superior a la capacidad que presenta la aplicación de un solo clasificador base.

Este *base learner* puede ser de cualquier tipo (regresiones, árboles de clasificación, redes neuronales, etc.) e, incluso, puede plantearse la construcción de un ensamble con diferentes modelos base. Existen numerosos algoritmos para la construcción de ensambles de modelos y muchas aplicaciones a diversos problemas (Polikar, Zhang y Ma, 2012; Okun, Valentini y Re 2011; Zhou, 2012). Este trabajo se centrará en uno de los más utilizados: el algoritmo *bagging*[4]. Originalmente, este método fue diseñado para ser utilizado sobre árboles de decisión (Breiman, 1996). Sin embargo, la lógica puede ser aplicada (y de hecho, lo ha sido) a diversos problemas de regresión o clasificación con distintos modelos base (regresiones, redes neuronales, clasificadores bayesianos, etc.).

El algoritmo simplemente entrena un determinado conjunto de clasificadores independientes, cada uno construido a través del remuestreo con reposición de n registros del *training set*. La diversidad del ensamble está asegurada por la variación de las muestras *bootstrap* y por la utilización de clasificadores débiles (sensibles a perturbaciones menores en los datos de entrenamiento). En ese sentido, clasificadores lineales son buenos candidatos para este propósito. Los clasificadores son combinados por alguna forma de simple *majority voting* (medias, medianas, modas, etc.).

La principal desventaja del ensamble LASSO usado en este trabajo se vincula al

4 Otro algoritmo ampliamente utilizado es *boosting* (en sus diversas variantes). Tiene la particularidad de que, a diferencia de *bagging* que realiza un muestreo *bootstrap* de todos los datos indistintamente, en cada una de las iteraciones se centra en aquellos registros en los que el clasificador funciona "peor", es decir, en aquellos registros peor clasificados (Schapire y Freund, 2012).

tiempo de cómputo. Los ensambles de modelos tienden a mostrar requerimientos computacionales notablemente más grandes que las imputaciones simples. Así, en el modelo estimado en el trabajo, se utilizó una rutina programada en lenguaje R[5]. El tiempo en una PC con procesador i7 y 16*GB* de memoria *RAM* para realizar la imputación correspondiente a la primera prueba (reseñada en el apartado 5.1) fue alrededor de 15 minutos. La estimación de errores de test usando validación cruzada (desarrollada en el apartado 5.2) tomó aproximadamente una hora.

No obstante, existe una ventaja que el método propuesto tiene en relación a las técnicas de imputación habitualmente utilizadas (específicamente, aquellas basadas en imputación simple o en la eliminación –*listwise o pairwise*– de casos). La construcción de ensambles de modelos introduce variabilidad en la estimación al remuestrar una determinada cantidad de veces los datos con valores a imputar. En efecto, esto permite potenciar la capacidad predictiva del modelo y generar clasificadores más eficientes. Por otro lado, el operador LASSO tiene dos particularidades que resultan sumamente útiles a los efectos de la construcción de un modelo de imputación. Dado que se trata de un método de regularización que utiliza la norma ℓ_1, permite realizar selección de variables (*feature selection*): debido a que algunos de los coeficientes β_j se hacen cero para satisfacer la restricción impuesta a la minimización de la suma de residuos al cuadrado, LASSO realiza una selección de variables. Este punto resulta de utilidad en tanto que permite comenzar con un conjunto de variables predictoras construido con un criterio amplio. En cambio, en métodos como *hot deck* suele ser necesario realizar un análisis previo para poder identificar las variables más correlacionadas con la dependiente para limitar el tamaño de la matriz a utilizar. De esta forma, el modelo *bagging*-LASSO permite realizar la selección de variables de forma automática.

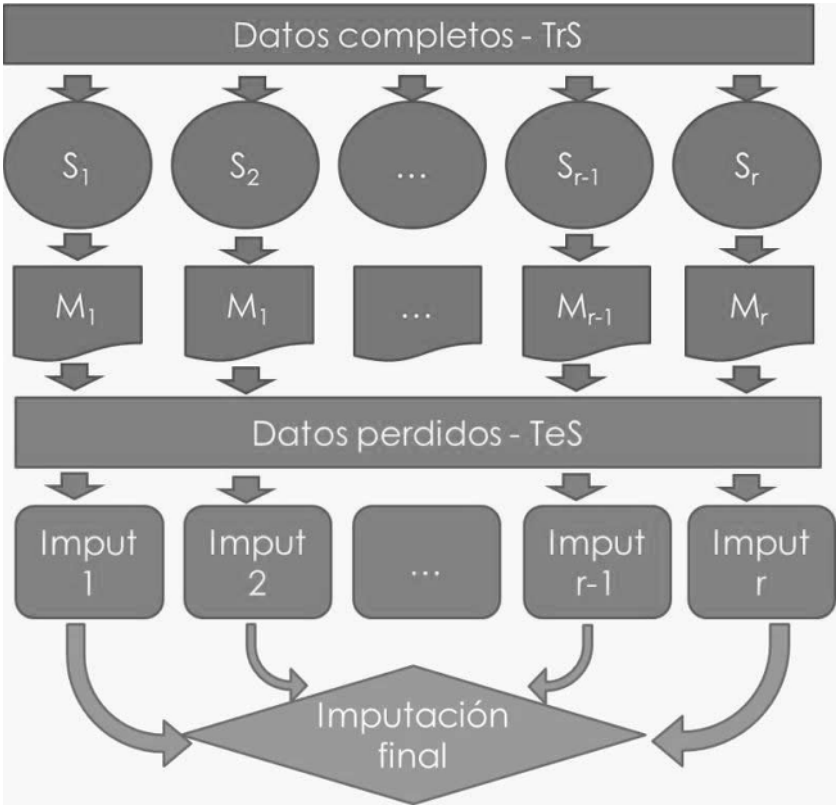
4.3.Algoritmo general empleado para la imputación

Si bien para cada una de las pruebas se varían algunos de los parámetros, la descripción general del algoritmo es la siguiente:

1. Se genera una base de datos con los datos completos TrS (sin NR)
2. Se genera otra base (TeS) con los casos con NR.
3. Para r = 1: rep
   3.1. En TrS se genera una muestra *boostrap* (por muestreo aleatorio simple o estratificado) de tamaño n=n (con reposición)
   3.2. En la muestra generada se estima una regresión LASSO
   3.3. A partir del modelo generado se estiman los datos perdidos en el TeS
4. Luego de las rep repeticiones se generan la misma cantidad imputaciones de cada uno de los valores perdidos en TeS.
5. Se agregan (a partir de la mediana) y ese es el valor final de imputación.

5 El autor se encuentra trabajando en la implementación de un paquete de imputación de valores perdidos para lenguaje R que utilice el método *bagging*-LASSO

Esquema 1. Secuencias en el modelo de imputación basado en un ensamble de regresiones LASSO



Fuente: elaboración propia.

Se utilizó como soporte empírico la base de datos usuaria de la Encuesta Permanente de Hogares (EPH) correspondiente al segundo trimestre de 2015. El objetivo fue lograr la imputación de los valores perdidos de la variable correspondiente a los ingresos de la ocupación principal -p21- de la población ocupada.

Las variables predictoras utilizadas fueron las siguientes:

Tabla 1. Variables incluidas en el modelo de imputación[6]

Variable	Dimensión
Región	Contexto
Aglomerado	
Tamaño	
Relación de parentesco	Sociodemográficas
Sexo	
Edad	
Situación conyugal	
Tipo de cobertura médica	
Lecto-escritura	
Nivel educativo	
Lugar de nacimiento	
Lugar de residencia (5 años atrás)	
Cantidad de ocupaciones	Laboral
Total de horas trabajadas (semana de referencia)	
Intensidad de trabajo	
Búsqueda de mayor cantidad de horas de trabajo	
Categoría ocupacional	
Carácter de la ocupación	
Calificación de la ocupación	
Rama de actividad del establecimiento	
Tamaño del establecimiento	
Tiempo que trabaja en forma continua en el empleo	
Cobertura previsional	
Percepción de ingresos por programas sociales	Otros ingresos
Monto total de ingreso no laboral	

Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

6 En general, se mantuvieron las formas de clasificación/codificación de las variables originales del INDEC. La única excepción es la referida a aportes jubilatorios –que se registra solamente para los asalariados– en la cual se agregó una categoría específica indicando si el caso correspondía a un trabajador independiente.

En cada iteración se estima un modelo LASSO utilizando como variable dependiente el logaritmo (en base 10) del ingreso y como predictores las variables anteriormente mencionadas,

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

buscando minimizar la siguiente función de pérdida

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

donde

y_i = ingresos totales de la i-ésima unidad

X_{ij} = j-ésima variable predictora de la i-ésima unidad

β_j = coeficiente de regresión correspondiente a la j-ésima variable predictora

5.Pruebas realizadas

El modelo fue evaluado de dos maneras diferentes. En primer lugar, se compararon los resultados obtenidos con la imputación realizada por el INDEC a través del método *hot deck* secuencial. El objetivo era establecer qué tan diferentes eran ambas imputaciones y qué posible impacto podía tener la utilización del modelo LASSO al realizar las estimaciones con la información publicada por dicho organismo. Sin embargo, el problema con esta aproximación es que se desconocen las especificidades del modelo de imputación utilizado por el INDEC. En efecto, no se conocen, por ejemplo, cuáles son las variables predictoras utilizadas. Es por ello que se utilizó, además, una segunda estrategia de validación: se generaron valores perdidos de forma aleatoria sobre los casos completos de la EPH y se evaluó la performance predictiva del modelo LASSO y de un método basado en *hot deck*[7], utilizando las mismas variables.

Tanto en la primera prueba como en la segunda, se utilizó el algoritmo *bagging* LASSO especificado más arriba. A su vez, se realizó una aproximación lo más "grosera" intentando emular el peor escenario posible, interviniendo lo menos posible sobre la información presentada:

1) se aplicó muestreo aleatorio simple (sin estratificar por dominio);

2) para el primer set de pruebas se realizó un primer filtro de *outliers* (se excluyeron aquellos casos a más de 2 desvíos estándar de la media), en cambio, para el segundo no se realizó ningún filtro para evaluar la performance ante la existencia de casos extremos;

3) no se utilizó ningún tipo de peso ni ponderación: probabilidad de inclusión en las muestras *bootstrap* cada caso fue de 1/n;

7 Para definir los casos donantes en el método *hot deck* se usó una función basada en la distancia de Manhattan. Se optó por un esquema restrictivo en el cual se definió que cada caso "donante" solo pudiera funcionar como tal una sola vez, es decir, que solamente "donara" el valor a un solo valor perdido y a su vez, si cada valor perdido encuentra más de un donante "cercano" se selecciona de forma aleatoria cuál de todos los casos donantes se usará para la imputación.

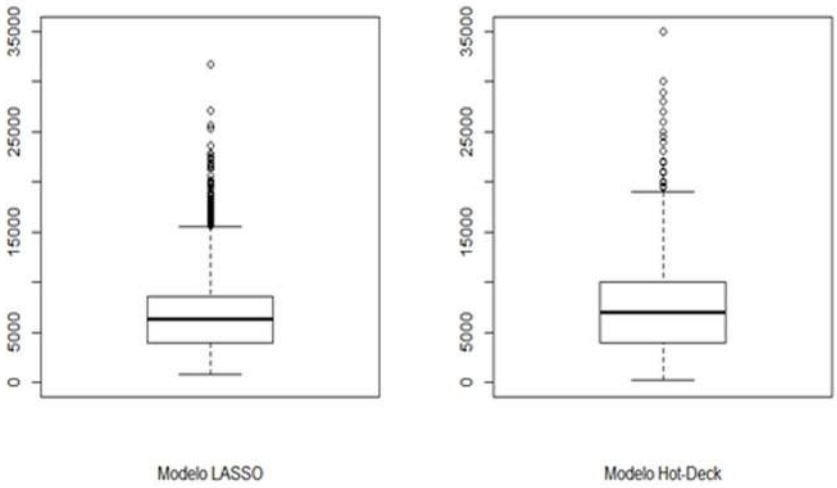
4) no se estimaron modelos para cada aglomerado, sino que se incluyó el aglomerado como variable independiente.

En las siguientes secciones se avanzará sobre ambas formas de validación.

5.1.Comparación del modelo *bagging*-LASSO con la imputación realizada por INDEC mediante *hot-deck*

Para la primera de las pruebas realizadas, se corrió el algoritmo desarrollado previamente con 5.000 repeticiones. Se presentan las comparaciones con los casos de la base de datos imputados por el INDEC con el método *hot deck*.

Gráfico 1. *Boxplots* de casos imputados con LASSO y *hot deck* (casos imputados por INDEC), en escala monetaria y logarítmica.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

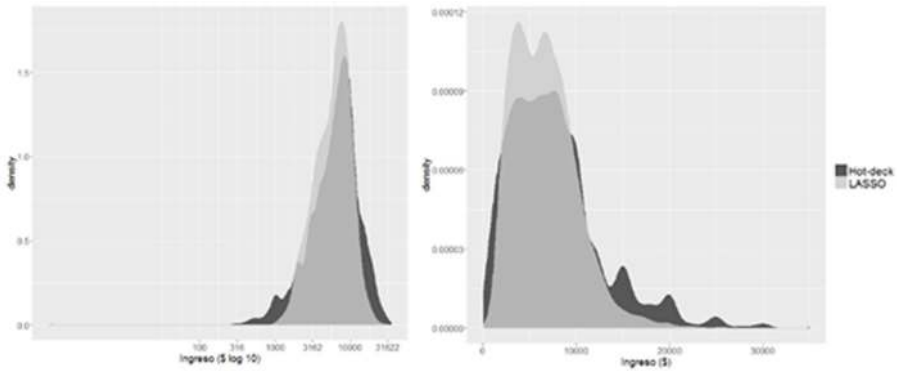
Tabla 2. Descriptivos de casos imputados con LASSO y *hot deck* (casos imputados por INDEC), en escala monetaria y logarítmica.

	Escala monetaria ($)		Escala logarítmica	
Estadístico	Modelo LASSO	Modelo *Hot-Deck*	Modelo LASSO	Modelo *Hot-Deck*
Mínimo	$760	$200	2,881	2,301
1er. Quartil	$3.923	$4.000	3,594	3,602
Mediana	$6.285	$7.000	3,798	3,845
Media	$6.606	$7.729	3,759	3,781
3er. Quartil	$8.581	$10.000	3,934	4
Máximo	$31.830	$35.000	4,503	4,544
Desv. Est.	$3.411	$5.131	0,261	0,333
Coef. de Var	51,68%	66,38%	6,95%	8,81%
Gini	0,282	0,354	0,036	0,048

Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

Puede notarse en las tablas y en los *boxplots* anteriores que en el centro de la distribución (digamos, entre el tercer y primer cuartil) casi no parece haber diferencias entre ambos modelos. Sin embargo, en las colas ambas estimaciones se alejan bastante. Esto es así, especialmente, si se considera la escala absoluta (en unidades monetarias). Como era de esperarse, las diferencias son notablemente menores en la escala logarítmica. Al mismo tiempo, puede notarse como la variabilidad (tanto medida a través del desvío estándar como del coeficiente de variación) es más pequeña en la estimación por el método LASSO que en el que utiliza *hot deck.*

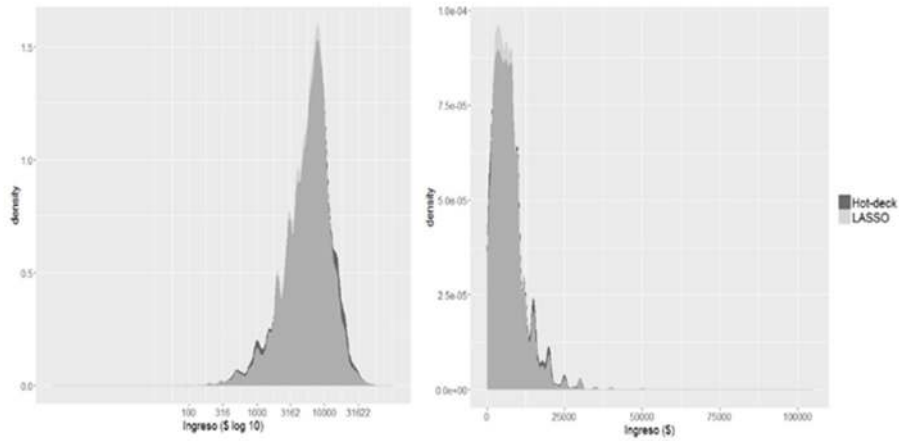
Gráfico 2. *Density plot* de distribución de casos imputados con LASSO y *hot deck* (aplicado por INDEC), escala logarítmica y escala monetaria.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

En los gráficos de densidad (*kernels*) queda claro como las estimaciones a través de *hot deck* son más altas que las realizadas por el método *bagging*-LASSO. Se observan unas "colas gruesas" tanto en el extremo superior como en el inferior de la distribución, observable en las áreas coloreadas con gris oscuro.

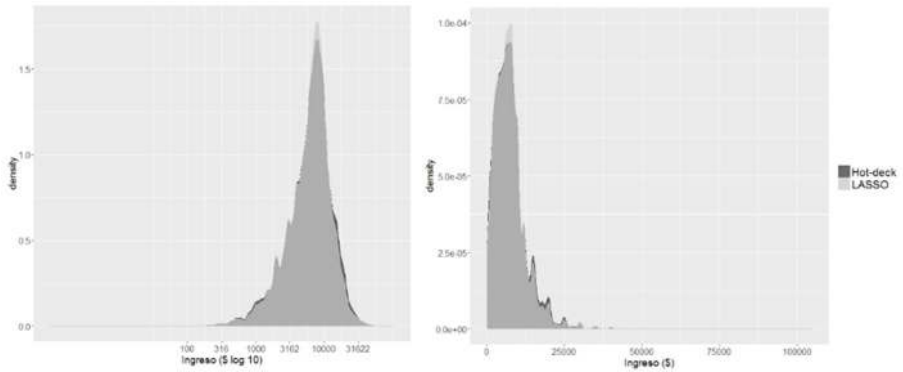
Gráfico 3. *Density plot* de distribución completa (respuesta más imputados) con LASSO y *hot deck* (aplicado por INDEC), escala logarítmica y escala monetaria.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

A su vez, se observa que al analizar la distribución conjunta total (es decir, el total de respuestas con casos completos y con casos imputados) puede verse que la estimación no parece alterarse sustancialmente. Lo mismo sucede, si se analiza la distribución condicionada a las dos grandes categorías ocupacionales: asalariados y trabajadores independientes (patrones y trabajadores por cuenta propia).

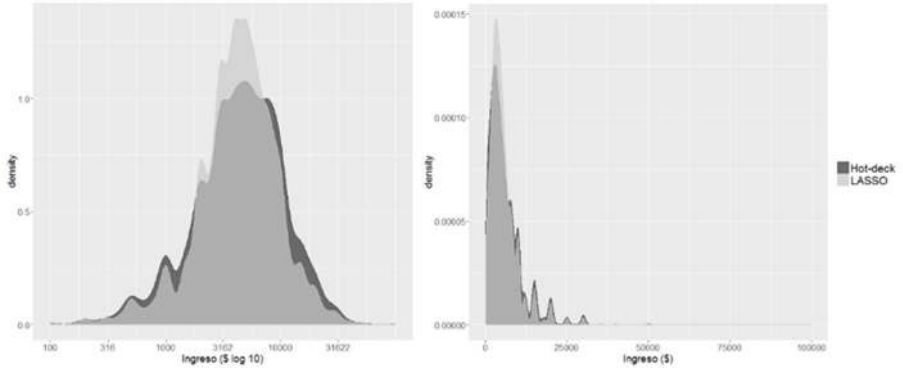
Gráfico 4. *Density plot* de distribución completa (respuesta más imputados) con LASSO y *hot deck* (aplicado por INDEC) de trabajadores asalariados, escala logarítmica y escala monetaria.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

Para el caso de los asalariados casi no se observan diferencias entre ambas distribuciones. Puede verse que tanto en el caso de la distribución en escala logarítmica como absoluta, las distribuciones completas se superponen. Se verifica, igual que en los gráficos anteriores, que el método *hot deck* presenta una mayor densidad en los valores extremos superiores.

Gráfico 5. *Density plot* de distribución completa (respuesta más imputados) con LASSO y *hot deck* (aplicado por INDEC), de trabajadores independientes, escala logarítmica y escala monetaria.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

La diferencia fundamental se observa en el caso de los trabajadores independientes. Allí se observa que existen diferencias entre los extremos de la distribución. Puede verse qué, como resultado de la imputación sobre la distribución completa, en los valores mayores de la escala se observan mayor densidad de casos para la imputación *hot deck* por sobre la realizada con *bagging*-LASSO.

5.2.Comparación del modelo *bagging*-LASSO y método *hot deck* sobre datos perdidos generados al azar

Para la segunda prueba, se generaron valores perdidos aleatoriamente y se comparó la performance predictiva de ambos métodos. Para estimar el *test-error* se utilizó un proceso de validación cruzada de *k-fold*=9. Es decir, se dividió el *training set* (total de registros con valores completos –sin perdidos– en la base de datos correspondiente al segundo trimestre de 2015) en 9 partes de (aproximadamente) igual tamaño conformadas de forma aleatoria. Se corrió el algoritmo mencionado previamente en nueve iteraciones. En cada una de las iteraciones cada k parte de la base de datos como *test-set* y el resto como *training-set.*

Esquema 2. Representación gráfica del procedimiento de validación cruzada para k=9 iteraciones.

Iteraciones (k)	Base de datosTOTAL				Resultado
I1	Test-set 1				MSE_1
I2		Test-set 2			MSE_2
Ii			Test-set...		MSE_i
I9				Test-set 9	MSE_9

Fuente: elaboración propia.

Para cada una de las iteraciones se calculó el *test-error* como la media del cuadrado de la diferencia entre el valor estimado por el ensamble de modelos ($\hat{y}_i$) y los valores reales (y_i).

$$MSE_k = \sum_{(i \in C_k)} \frac{(y_i - \hat{y}_i)^2}{n_k}$$

La estimación final del test-error ($CV_{(k)}$) está constituida por el promedio de los *test-error* (MSE_k) a lo largo de las k=9 iteraciones:

$$CV_{(k)} = \sum_{k=1}^{k} \frac{n_k}{n} MSE_k$$

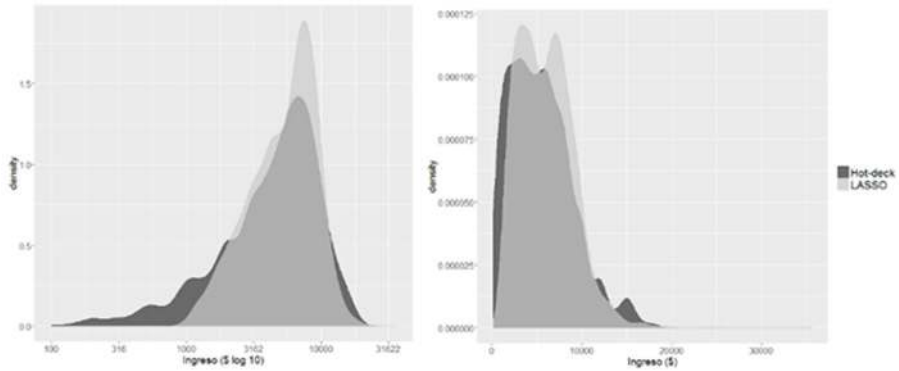
Tabla 3. Error cuadrático medio (MSE) estimado por validación cruzada (k=5) método LASSO y *hot deck*, en escala monetaria.

	Escala absoluta ($)	
	Modelo LASSO	Modelo *Hot-Deck*
$CV_{(k)}$ (MSE)	$\$^2$16.911.175	$\$^2$25.480.031
$CV_{(k)}$ (RMSE)	$3.994	$4.933

Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

Los resultados muestran que existe una reducción de alrededor del 34% en el MSE al utilizar el modelo *bagging*-LASSO en comparación con la imputación realizada por el método *hot deck*. En relación al RMSE (es decir, la raíz cuadrada del MSE) se observa una reducción de alrededor del 20%. Es decir, que *bagging*-LASSO mejora considerablemente el error de predicción de la variable de ingresos laborales. En ese sentido, es esperable que en datos perdidos "originales" (es decir, no generados artificialmente) el método consiga una mejor performance que *hot deck.*

Gráfico 6. *Density plot* de distribución de casos imputados con LASSO y *hot deck* (perdidos generados aleatoriamente), escala logarítmica y escala monetaria.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

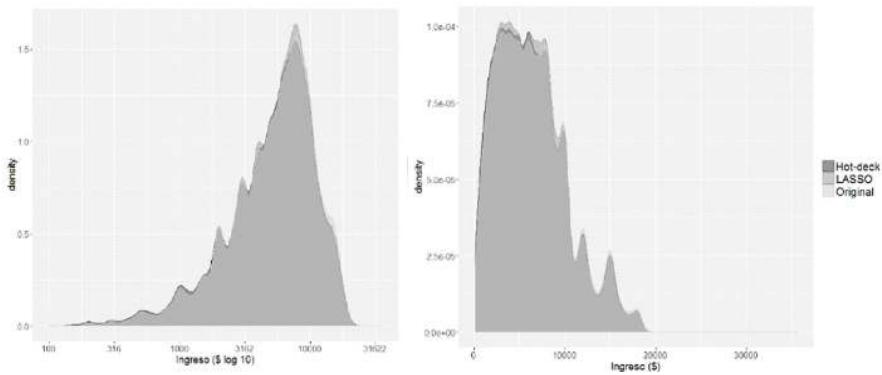
Tabla 4. Descriptivos de casos imputados con LASSO y *hot deck* (imputadas por modelo propio), en escala monetaria y logarítmica.

Estadístico	Escala absoluta ($)		Escala logarítmica	
	Modelo LASSO	Modelo Hot-Deck	Modelo LASSO	Modelo Hot-Deck
Mínimo	$917	$100	2,962	2
1er. Quartil	$3.466	$2.600	3,537	3,415
Mediana	$5.762	$5.000	3,761	3,699
Media	$5.938	$5.410	3,709	3,608
3er. Quartil	$7.919	$7.800	3,899	3,892
Máximo	$35.700	$18.000	4,553	4,255
Desv. Est.	$3.092	$.3508	0,249	0,382
Coef. de Var	52,10%	64,20%	6,74%	10,58%
Gini	0,286	0,361	0,038	0,056

Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

Como puede verse, se observan tendencias similares al ejercicio desarrollado en base a la imputación mediante hot deck realizada por el INDEC: menor variabilidad en las imputaciones generadas por LASSO y diferencias menores en los valores centrales y una mayor diferencia en las colas de las distribuciones.

Gráfico 7. *Density plot* de distribución de casos completos (perdidos generados aleatoriamente más completos) con LASSO y hot deck, escala logarítmica y monetaria.



Fuente: elaboración propia en base a INDEC, Microdatos EPH 2do. trimestre de 2015.

Puede notarse, en los gráficos anteriores, un patrón similar al que resulta de las pruebas anteriores: en relación al modelo *bagging*-LASSO, el método *hot deck* tien-

de a sobreestimar los valores extremos de la distribución (tanto superiores como inferiores).

6.Discusión de resultados y nuevos problemas

En el presente documento se intentó elaborar una propuesta de un modelo de imputación de datos con respuesta faltante en variables de ingreso basado en la combinación de algunas técnicas de aprendizaje automático y de modelos de regresión LASSO. Luego de exponer los fundamentos de cada una de las técnicas se realizaron dos tipos de evaluación de los modelos con base en los datos de la EPH (2do trimestre de 2015) buscando lograr la imputación de la variable correspondiente a los ingresos laborales de los individuos.

Se compararon los resultados del método propuesto con el llamado *hot deck* (el más utilizado en el Sistema Estadístico Nacional argentino). Los resultados mostraron:

a) que existe poca diferenciación con respecto a la distribución de casos imputados por el INDEC a partir del método *hot deck*;

b) que al analizar el MSE entre dos formas "comparables" de imputación se observa la caída considerable del método *bagging*-LASSO en alrededor de un 33% y de un 20% al considerar el RMSE.

Un punto importante para remarcar se vincula al hecho de que la imputación *bagging*-LASSO parece funcionar mejor (menor MSE) que el método *hot deck* ante la presencia de casos extremos. A su vez, resulta importante remarcar la potencialidad del modelo propuesto para realizar un proceso de selección automática y regularización de las variables predictoras del ingreso, lo cual permitiría construir *sets* de predictores con un criterio conceptual más amplio que el que permiten otros métodos. Es más, en el límite podría pensarse en un enfoque más empírico de predicción e imputación de no respuestas en variables de ingreso.

Ahora bien, se abren algunas posibilidades interesantes para continuar el trabajo sobre este tipo de modelos de imputación. En primer lugar, dado que el modelo LASSO aparece no tan sensible a las variabilidad producida por las muestras *boostrap,* sería interesante pensar en introducir alguna variabilidad a nivel de los predictores.

Una posibilidad en ese sentido sería avanzar hacia un esquema similar al de *random forests*, realizando un muestreo aleatorio de los predictores en cada modelo. Es decir, no solamente se efectuaría un *bootstrap* de los casos, sino también de las predictoras de la base de datos.

Una segunda posibilidad implica utilizar otro de los algoritmos más comunes de ensamble de modelos: el denominado *boosting*. En efecto, dado que en este algoritmo el clasificador es entrenado con sucesivas muestras en cuya selección se otorga una mayor ponderación a los casos peor clasificados, es esperable una mejora en la capacidad predictiva del modelo.

Referencias bibliográficas

Allison, P. (2002). Missing Data.*Sage University Papers on Quantitative Applications in the Social Sciences.* 07-136. California: Sage.

Breiman, L. (1995). Better subset selection using the nonnegative garrote.*Technometrics,* 37. 738–754.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24. 123-140.

Camelo, H. (1999). Subdeclaración de ingresos medios en las encuestas de hogares, según quintiles de hogares y fuente del ingreso. 2° Taller Programa para el Mejoramiento de las Encuestas y la Medición de las Condiciones de Vida en América Latina y el Caribe (MECOVI), Buenos Aires.

Felcman, D., Kidyba, S. y Ruffo, H. (2004). Medición del ingreso laboral: ajustes a los datos de la encuesta permanente de hogares para el análisis de la distribución del ingreso (1993–2002).14° Taller Programa para el Mejoramiento de las Encuestas y la Medición de las Condiciones de Vida en América Latina y el Caribe (MECOVI), Buenos Aires.

Hastie, T., Tibshiriani, R. y Wainwright, M. (2015).*Statistical Learning with Sparsity. The Lasso and Generalizations.*Florida: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R. y Friedman, J. (2009).*The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Berlin: Springer.

Hoerl, A. E. y Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems.*Technometrics*, 12. 55-67.

Hoszowski, A., Messere, M., y Tombolini, L. (2004). Tratamiento de la no respuesta a las variables de ingreso en la Encuesta Permanente de Hogares de Argentina. 14° Taller Programa para el Mejoramiento de las Encuestas y la Medición de las Condiciones de Vida en América Latina y el Caribe (MECOVI), Buenos Aires.

INDEC (2009). Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la EPH. Metodología N° 15, INDEC, Buenos Aires.

Medina, F. y Galván, M. (2007). Imputación de datos: teoría y práctica. *Serie Estudios Estadísticos y Prospectivos*, 54, Santiago de Chile: CEPAL. Disponible en http://www.cepal.org/es/publicaciones/4755-imputacion-datos-teoria-practica

Okun, O., Valentini, G. y Re, M. (2011). *Ensambles in Machine Learning Applications*. Berlín: Springer.

Pacífico, L., Jaccoud, F., Monteforte, E., y Arakaki, G.A. (2011). La Encuesta Permanente de Hogares, 2003–2010. Un análisis de los efectos de los cambios metodológicos sobre los principales indicadores sociales. X Congreso de Nacional de Estudios del Trabajo, (ASET), Buenos Aires.

Polikar, R., Zhang, C., y Ma, Y. (eds.) (2012), *Ensamble Machine Learning. Methods and Applications*. Berlín: Springer

Salvia, A. y Donza, E. (1999). Problemas de medición y sesgos de estimación derivados de la no respuesta completa a las preguntas de ingresos en la EPH

(1990-1998). *Revista Estudios del Trabajo*, 18. 93-110.

Schapire, R. y Freund, Y. (2012). *Boosting: Foundations and Algorithms.* Massachussets: MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, 58. 267–288.

Zhou, Z. (2012). *Ensamble Methods*. *Foundations and Algorithms*. Florida: Chapman & Hall/CRC.

Fuentes utilizadas

Microdatos de la Encuesta Permanente de Hogares (EPH) correspondientes al segundo semestre de 2015, Instituto Nacional de Estadística y Censos (INDEC), disponibles en www.indec.gob.ar.