Big Data for Applied Economics
Juan Diego López (201425104), María Paula Nieto (201612379), Daniela Jaime (201531520)

Introducción

Medir la pobreza adecuadamente no es una tarea fácil. Existen múltiples variables, que van desde las características personales del agente hasta aspectos contextuales, que determinan si una persona es pobre o no. Al respecto, la Comisión Económica de América Latina y el Caribe (CEPAL) (2018) destaca que la pobreza es una manifestación de las diferentes facetas de la desigualdad, por lo que su medición cobra gran relevancia para entender la realidad regional y definir aquellos factores y políticas que pueden mejorar las condiciones de vida de la población. En ese sentido, dado que los países son afectados diferencialmente por la pobreza, que existe heterogeneidad en su medición.

Así, existen dos enfoques principales para medir la pobreza: i) como carencia de ingresos y ii) los métodos basados en la combinación de múltiples indicadores de carencias (CEPAL, 2018). Con base en lo anterior, la medición de la pobreza definida como un concepto multidimensional en el que diferentes variables y atributos tienen injerencia representa uno de los mayores retos para la academia y el sector público, siendo una importante herramienta de prospectiva y de formulación de política pública en la que los métodos de *machine learning* han adquirido alta relevancia (Namita et al., 2022; Chagalj, 2019). Con esto en mente, a lo largo de este trabajo se presentarán diferentes aproximaciones y modelos para predecir la pobreza en hogares colombianos a partir de la base de datos personal y a nivel hogar de la GEIH en el año 2018 a nivel nacional. El script, figuras y bases utilizadas se podrán encontrar en el siguiente repositorio: https://github.com/marianieto198/Problem_Set_2.git.

Manejo de datos

El proceso de limpieza de datos consistió en i) un análisis de la composición de las bases de train y test, ii) identificación y tratamiento de *missing values*, iii) análisis de distribuciones y compilación de variables en las bases individuales para la estimación de la pobreza en los hogares. En un primer momento, al comparar las bases de *train* y *test* se identificó que se utilizarían para las predicciones aquellas variables que están en las dos bases. Luego, se revisaron las clases de cada variable y se agregaron etiquetas y nombres. Asimismo, al tener bases individuales y por hogar se hace necesario un análisis de la coherencia entre observaciones a partir de su id para comprobar si la suma de los ingresos totales a nivel individual es igual al del hogar. Así, es posible identificar que solamente el 2% de los hogares contaban con ingresos diferentes por pequeñas diferencias, por lo que los ingresos son consistentes entre individuos y hogares.

Ahora bien, para predecir la pobreza de los hogares se cuenta con un amplio repertorio en las bases individuales. De esta manera, se encontró que los ingresos del jefe del hogar son, en promedio, el 65% de los ingresos totales de un hogar, por lo que se decidió usar esta información en los modelos.
El mismo procedimiento se realizó con la base de testeo. En cuanto a los *missing values*, se identificó que el 33.2% de los valores en la base corresponden a estos valores (Apéndice 1). Específicamente, las variables con mayor cantidad de *missing values* se relacionan con el pago de arriendo, amortización o el hecho de que los individuos reciban ingresos aparte del ingreso laboral. Por lo tanto, se definió una variable de arriendo como la combinación de dos variables y se actualizaron los datos de ocupación para el jefe del hogar.

Finalmente, se eliminaron las variables que no se utilizaron para el análisis como la población económicamente activa, el departamento, entre otros. Asimismo, se eliminaron las variables con

missing values mayores al 30% y se les asignó la moda y/o media para imputar algunas variables relevantes. Por otro lado, teniendo en cuenta que la proporción de personas pobres es menor (20%), fue necesario balancear la muestra de pobres mediante el resamsampling teniendo en cuenta una proporción de 50-50. En consecuencia, de acuerdo con el Apéndice 2 y 3 se observa que la mayoría de las observaciones se encuentran en los centros urbanos, cuentan con el jefe del hogar ocupado y mantienen coherencia si se analizan de manera diferenciada por pobre o no.

Modelos de clasificación

Una primera aproximación para predecir la pobreza radica en la clasificación de los hogares como pobres (1) o no pobres (0). En este caso, a partir de una serie de variables se usan modelos de clasificación para llevar a cabo la predicción. En primer lugar, se identificó que la variable "Pobre" de la base será la variable dependiente en estos modelos. Posteriormente, se plantea el método de *random forest*. Lo anterior con un doble propósito: identificar qué variables tienen mayor importancia dentro del modelo y estimar un primer modelo de clasificación con mayor precisión.

Al considerar los resultados del modelo de *random forest,* resulta pertinente destacar que, en este caso, el interés radica en predecir correctamente a los hogares que son pobres. En ese sentido, el interés se concentrará en la sensibilidad: el ejercicio busca predecir correctamente los hogares pobres con un costo relativamente bajo de falsos positivos. De esta manera, se podrán focalizar correctamente las intervenciones de política pública, por ejemplo. Este modelo tiene una sensibilidad de solo 36%, lo que sugiere que el modelo no clasifica correctamente a los hogares pobres. Sin embargo, esta metodología sirve para identificar las variables más relevantes en el modelo. Así destacan las siguientes: Arriendo, la educación del jefe del hogar, la pensión del jefe del hogar, el número de personas en el hogar, entre otros.

A partir de lo anterior, se tomó la base train de hogares y se dividió en tres submuestras: i) base de entrenamiento para el preprocesamiento (115.471 observaciones), ii) base de evaluación para el post-procesamiento (32.990 observaciones) y iii) base de testeo (16.496 observaciones). En este caso, las submuestras se utilizaron para probar diferentes modelos y encontrar el que tenga una mayor sensibilidad. Después de la definición de las submuestras mediante *cross-validation* se corta en cinco partes la base para balancear la muestra entre pobres y no pobres. Luego, se empiezan a correr modelos de predicción construidos a partir de los resultados de la importancia de las variables en el random forest:

(1) $\text{Pobre} = \beta 0\ \text{Arriendo} + \beta 1\text{PensionJefe\_2} + \beta 2\text{ArriendosPenJefe\_2} + \beta 3\text{TipoTrabajoJefe\_1} + \beta 4\text{Npersug} + \beta 5\text{EducacionJefe\_6} + \beta 6\text{P5090} + \beta 7\ \text{P5000} + \beta 8\text{EdadJefe} + \beta 9\text{OcupadoJefe} + \beta 10\text{HorasTrabJefe} + \upsilon$

(2) $\text{Pobre} = \beta 0 + \beta 1\text{Arriendo} + \beta 2\text{PensionJefe\_2} + \beta 3\text{TipoTrabajoJefe\_1} + \beta 4\ \text{Npersug} + \beta 5\text{EducacionJefe\_6} + \beta 6\text{P5000} + \beta 7\text{EdadJefe} + \beta 8\text{OcupadoJefe} + \upsilon$

Cinco modelos de predicción se definen para cada modelo de regresión y se corren a partir de la base de entrenamiento. El primero consiste en un modelo *logit* con función binomial a partir del cual se obtiene un *accuracy* de 0.80 para el modelo (1) y 0.81 para el (2). Es decir, que aproximadamente un 80% de las observaciones totales se clasificaron de manera correcta. Adicionalmente, el área bajo la

curva (ROC) para (1) es del 0.76 y para (2) es de 0.83, para los dos modelos la especificidad es de 0.16 y la sensibilidad es de un 98%. Es decir, que los modelos detectan correctamente el 98% de los pobres que en verdad son pobres y un 16% de los verdaderos no pobres son clasificados como tal. Ahora bien, para verificar la pertinencia de incluir las variables especificadas con anterioridad en el modelo se utiliza lasso con una grilla de 200. A partir de esto se observa las diferentes penalizaciones y escogemos el *lambda* que maximiza la sensibilidad del modelo. Asimismo, se prueba Ridge para identificar cuál funciona mejor, en este caso para (1) y (2) ofrecen un *lambda* similar de 1.03. Teniendo esto en cuenta, se corre el modelo de regresión definiendo el *lambda* obtenido a partir de Lasso.

Estos tres modelos iniciales están utilizando un *cutoff* de ½ para definir los hogares por encima y por debajo de la línea de pobreza, por lo que también se buscó el *cutoff* que se encontrara ubicado más cerca al ideal de la predicción del modelo. Para esto se utiliza la muestra de evaluación de post-procesamiento y se obtuvo para (2) un ROC de 0.76 y un umbral de 0.79. Se compara la especificidad y la sensibilidad con los dos puntos de corte y se evidencia que con un punto de corte de 0.79 hay una mayor sensibilidad. No obstante, los modelos con la muestra de entrenamiento son bastante optimistas por lo que en respuesta al desbalanceo de los datos se definieron también modelos de up-sampling y down-sampling con el propósito de que la cantidad de hogares pobres y no pobres estén balanceadas a la hora de realizar la predicción. Con up-sampling para (1) se obtiene un lambda de 0.013 y uno muy similar con el down-sampling[1]. Ya con estos cinco modelos definidos se evalúan a partir de la muestra de testeo y se elige el modelo con una mayor sensibilidad. Las métricas de cada modelo se encuentran en la tabla que se encuentra en el anexo 5, se identifica entonces con el modelo de regresión (1) que el modelo de predicción logit es el que tiene una mayor sensibilidad de 72.5%, seguido por el modelo lasso. Sin embargo, el modelo (2) también bajo el modelo de predicción logit cuenta con una mayor sensibilidad con 73.2%, por lo que este será el elegido.

Modelos de predicción de ingreso

La medición de la pobreza entendida como un problema de predicción del ingreso, requiere que, además, se determine una línea de pobreza y clasificar a los individuos. En este caso, se decidió usar la variable Ingreso Total de la Unidad de Gasto (Ingtotug) como la variable dependiente debido a que tiene en cuenta los ingresos derivados de arriendos y es usada para calcular la línea de pobreza. Adicionalmente, con el propósito de identificar las variables con mayor poder de predicción se estimaron los modelos de Ridge y Lasso, aunque solo éste lleva a cabo la selección de variables. Como destacan James et al. (2021), los modelos Lasso dado que seleccionan variables pueden generar modelos menos complejos. Sin embargo, solo tendrán buen desempeño en los escenarios donde pocos predictores tengan coeficientes significativos, por lo que también se estima el modelo ridge.

Igualmente, se estima un modelo de *XGBoost* con los siguientes hiperparamétros: 250 a 500 rondas; de 4 6 hojas por árbol para evitar un sobreajuste; lambda 0.05 para tener mayor pureza pero con mayor costo computacional; entre 10 y 50 observaciones en cada hoja lo que evita un sobreajuste del modelo; gamma 0.01 lo que generó una penalización alta y en *subsample* se determinó en 0.6. Finalmente, se estimaron modelos OLS con todas las variables y modelo de regresión lineal con las variables con mayor poder predictivo. Además, se estimó un modelo de *elastic net*.

[1] Cabe destacar que estos modelos se corrieron con diferentes semillas para mayor robustez en las predicciones.

Posteriormente, la capacidad predictiva de cada uno de estos modelos se midió a través del MSE (Anexo 7). Así, se encontró que el modelo con el menor error cuadrático medio fue el estimado a través de la metodología de XGBoost. En este caso, observamos que los modelos estimados usando los métodos de regularización de Lasso y Ridge no generan valores del MSE muy distintos del error obtenido con el modelo de regresión lineal estimado con todas las variables de la muestra. Lo anterior, puede estar relacionado con el hecho de que XGBoost construye un modelo robusto a partir de modelos más simples y débiles usando una penalización de ridge. Como se mencionaba anteriormente, la penalización de ridge funciona adecuadamente cuando muchas de las variables tienen importancia dentro del modelo. Los resultados obtenidos en la regularización dan cuenta que este puede ser el caso. Al respecto del entrenamiento del modelo, este fue llevado a cabo con una división aleatoria de la muestra entre training (115472 observaciones) y test (49488 observaciones).

Ahora bien, usando la línea de pobreza estimada por el DANE para el año 2018, se convierte la variable dependiente del modelo elegido como un indicador binario que nos permite clasificar a los individuos a partir de la predicción del ingreso realizada. Si el ingreso del hogar predicho es mayor a la línea de pobreza multiplicada por el número de personas de la unidad de gasto, este hogar será clasificado como no pobre; si es menor, será clasificado como pobre. En el caso del modelo XGBoost, al evaluar con una parte de la muestra de entrenamiento, se obtuvieron las siguientes estadísticas: especificidad de alrededor de 64%. Lo anterior, quiere decir que el modelo tuvo la capacidad de predecir correctamente a los hogares pobres en un 64%.

Conclusiones y recomendaciones

A partir del análisis anterior, es posible identificar que no es necesario formular modelos complejos para predecir la pobreza de los hogares con un alto nivel de precisión, sensibilidad y ajuste. Como recomendación para futuros ejercicios de predicción del ingreso o la probabilidad de pobreza de los hogares se identifica la necesidad de organizar y analizar la estructura de los datos, revisar el *overfitting* de los modelos y experimentar con diferentes niveles de complejidad. Se sugiere, además, identificar teórica y empíricamente aquellas variables que cuentan con una mayor relevancia para la pobreza de los individuos y que puede que afecten de manera heterogénea las estimaciones realizadas. En este caso, el modelo de clasificación con una mayor sensibilidad para la predicción de la pobreza de los hogares consistió en un modelo logit construido a partir de variables que contaban con una mayor importancia sobre la dummy de "Pobre" encontradas mediante *random forest;* este modelo contó con un 73.2% de sensibilidad. Para la predicción del ingreso, el modelo con menor MSE fue estimado con XGBoost. A partir de lo anterior, se logró clasificar correctamente al 64% de la muestra como pobres.

Referencias

Comisión Económica para América Latina y el Caribe (CEPAL), Medición de la pobreza por ingresos: actualización metodológica y resultados, Metodologías de la CEPAL, N° 2 (LC/PUB.2018/22-P), Santiago, 2018.

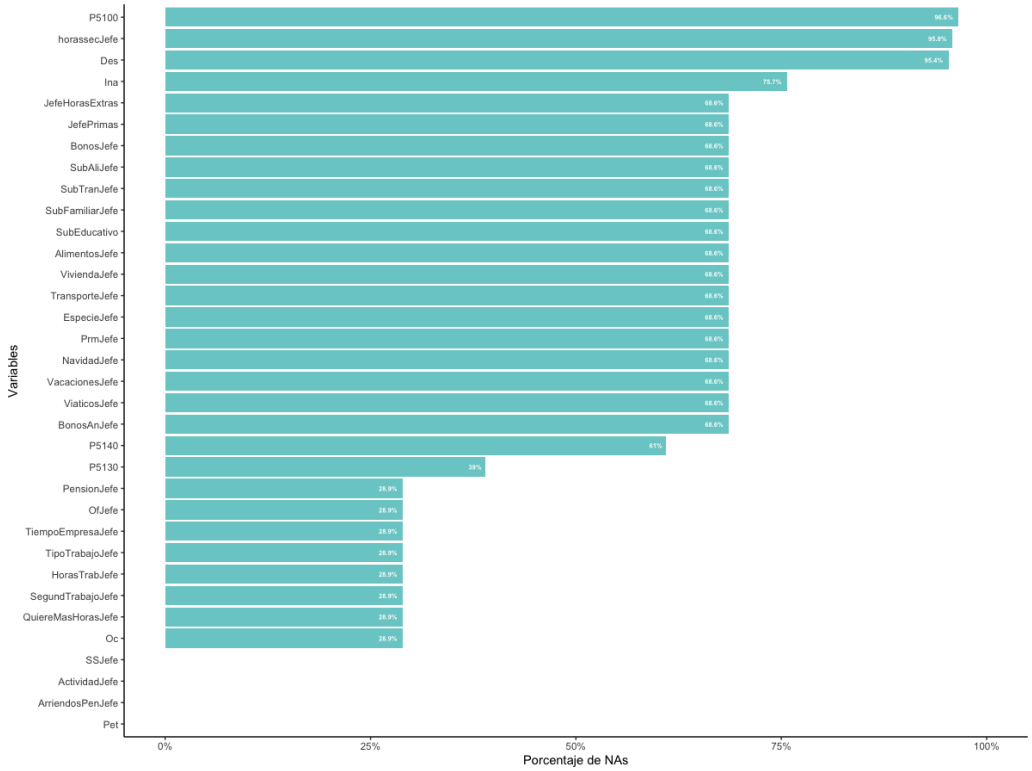
Dabús, A. (2019). Pobreza en Argentina : un análisis predictivo utilizando herramientas de machine learning.

Srivastava, Namita & , Saras & mohan, Dr. (2022). Approaches on Poverty and its measurement problem: Public policy outlook.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R.

Anexos

Anexo 1. Porcentaje de NAs en cada variable



Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 2. Estadísticas descriptivas variables relevantes

Characteristic	N = 164,957[1]
Pobre	
No_Pobre	131,935 (80%)
Pobre	33,022 (20%)
Clase	
1	149,605 (91%)
2	15,352 (9.3%)
Cabeza de hogar femenina	68,686 (42%)
Jefe de hogar ocupado	117,232 (71%)
Horas trabajadas por Jefe de hogar	40 (0, 48)
Personas en el hogar	3.00 (2.00, 4.00)
Preescolar	13 (<0.1%)
Básica Primaria	46,618 (28%)
Básica Secundaria	21,614 (13%)
Media	43,028 (26%)
Superior o universitaria	45,061 (27%)
TiempoEmpresaJefe	24 (0, 120)
Edad Jefe de Hogar	49 (37, 61)

[1] n (%); Median (IQR)

Fuente: Elaboración propia con base a GEIH 2018 (DANE)

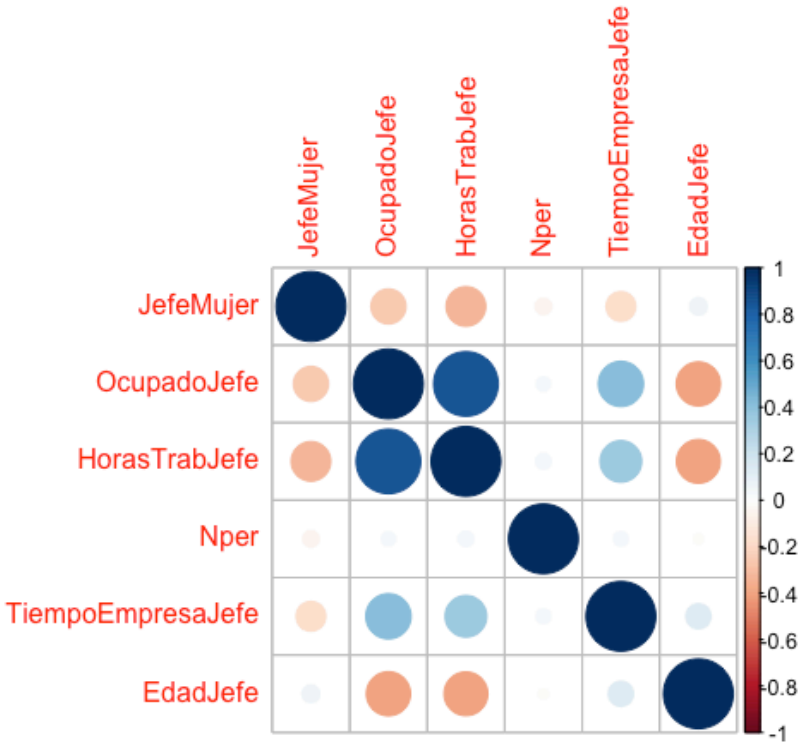
Anexo 3. Estadísticas descriptivas variables relevantes

Characteristic	**No_Pobre**, N = 131,935[1]	**Pobre**, N = 33,022[1]
Clase		
1	121,261 (92%)	28,344 (86%)
2	10,674 (8.1%)	4,678 (14%)
Cabeza de hogar femenina	53,185 (40%)	15,501 (47%)
Jefe de hogar ocupado	96,093 (73%)	21,139 (64%)
Horas trabajadas por Jefe de hogar	35 (25)	29 (25)
Personas en el hogar	3.08 (1.64)	4.14 (2.03)
Preescolar	7 (<0.1%)	6 (<0.1%)
Básica Primaria	34,137 (26%)	12,481 (38%)
Básica Secundaria	16,267 (12%)	5,347 (16%)
Media	34,747 (26%)	8,281 (25%)
Superior o universitaria	41,591 (32%)	3,470 (11%)
TiempoEmpresaJefe	81 (120)	68 (118)
Edad Jefe de Hogar	50 (16)	47 (16)

[1] n (%); Mean (SD)

Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 4. Gráfico de Correlaciones



Fuente: Elaboración propia con base a GEIH 2018 (DANE)

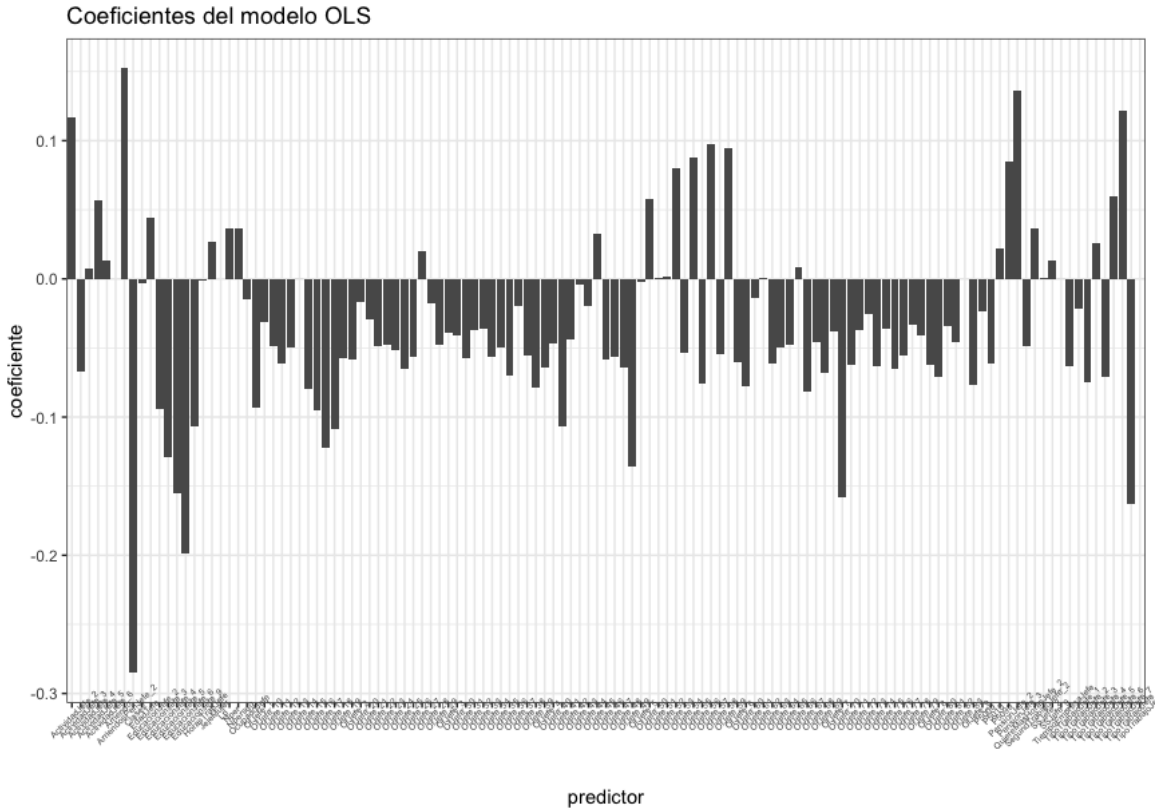
Anexo 5. Métricas modelos de clasificación

Modelo	Métrica	(1)	(2)
Logit	Sensibilidad Especificidad	72.5% 82.8%	73.2% 82.5%
Lasso	Sensibilidad Especificidad	64.3% 84.02%	65.4% 83.5%
Lasso-Threshold	Sensibilidad Especificidad	41.7% 92.6%	39.6% 91.8%
Up-sampling	Sensibilidad Especificidad	40.5% 93.06%	39.4% 91.9%
Down-sampling	Sensibilidad Especificidad	40.4% 93.02%	40.35% 87.7%

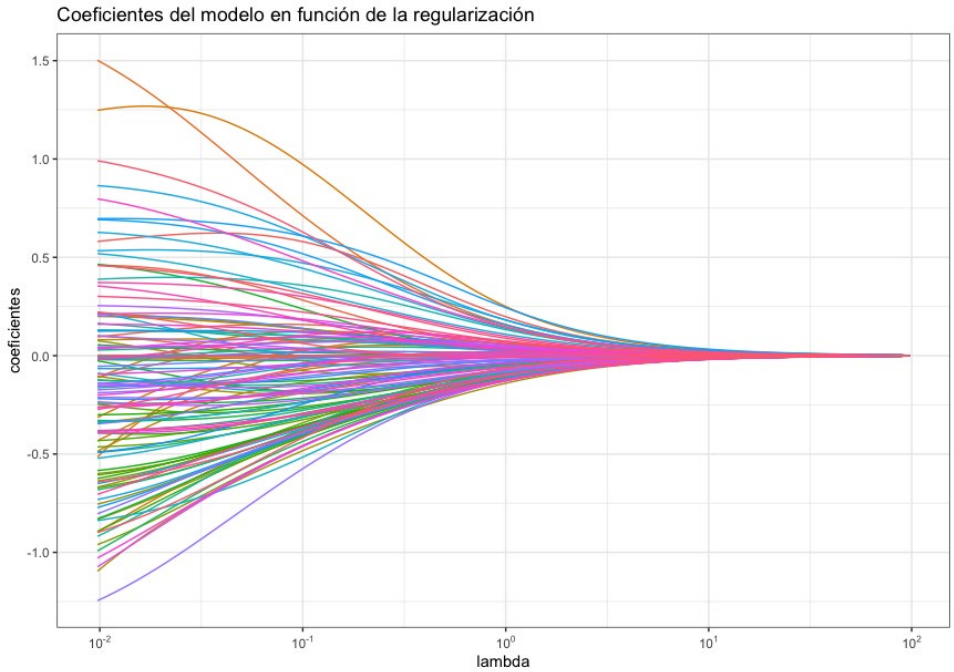
Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 6. Resultados Regularización Modelos Clasificación

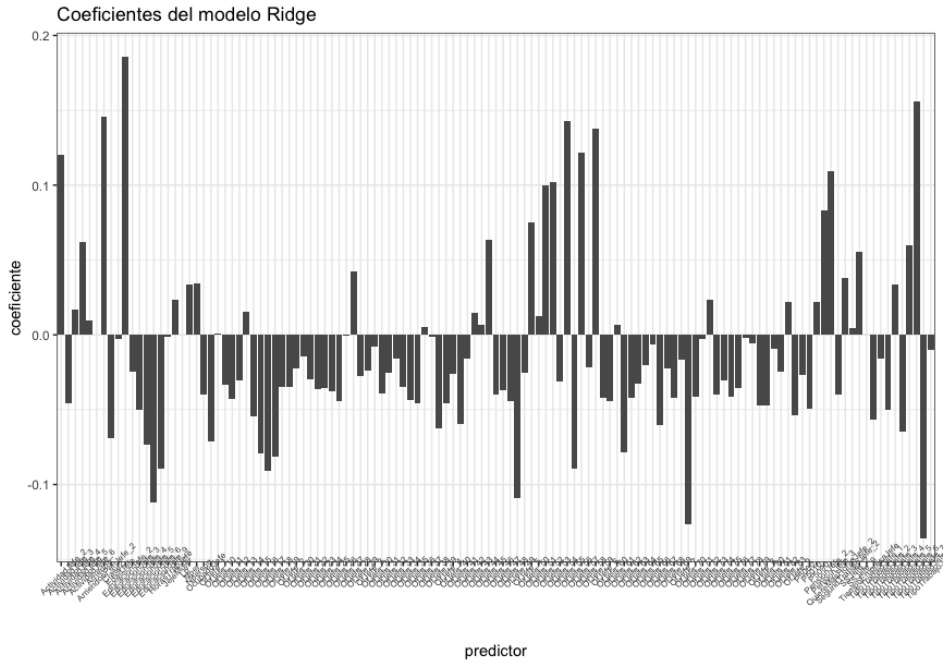
6.1. Coeficientes modelos OLS con todas las variables



6.2. Coeficientes modelo Ridge en función de lambda

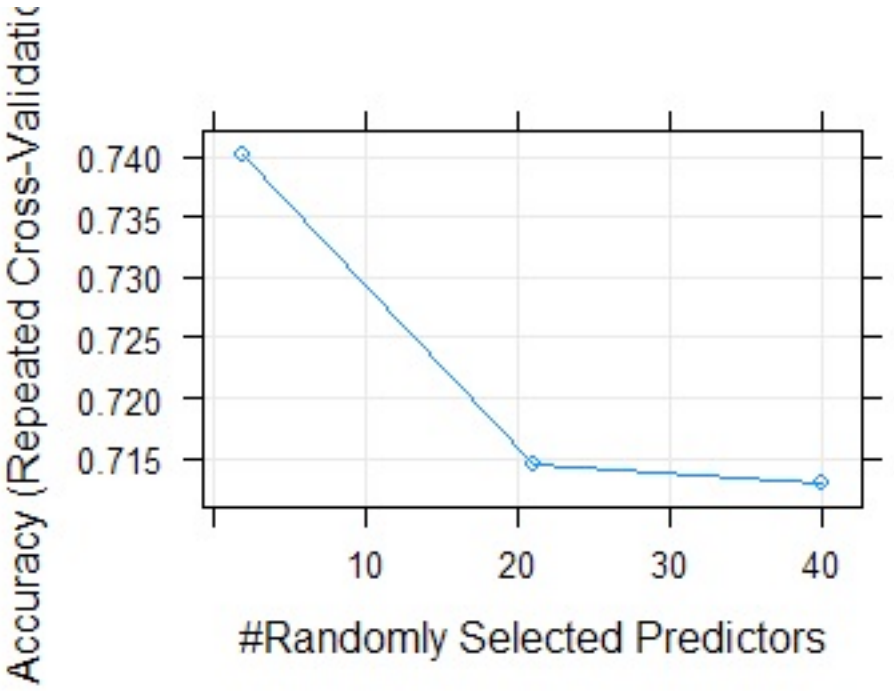


6.3. Coeficientes modelo ridge



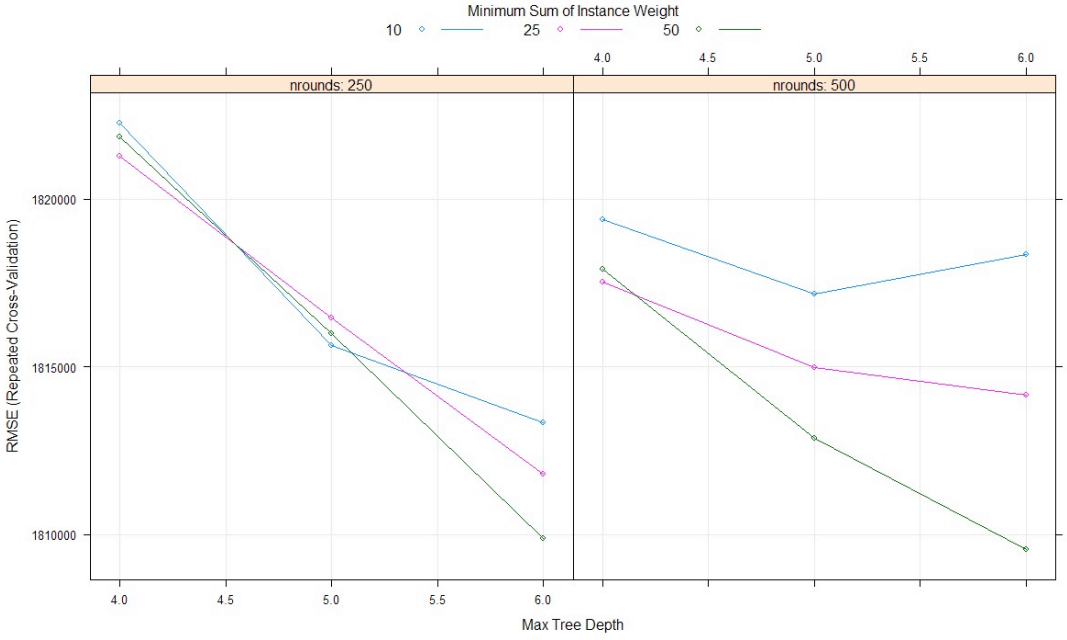
Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 7. Resultados Random Forest Ingreso



Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 8. Resultados modelo XGBoost Ingreso



Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 9. Resultados modelos de predicción de ingreso

Modelo	MSE Valores	RMSEValores
OLS – Todas las variables	4559910213781.03	213.539.462.717.809
Ridge	4568810659868.1	213.747.763.961.827
Lasso	4558798080424.24	213.513.420.665.405
XGBoost	3495298557010.06	186.957.175.765.202
Elastic Net	5251005115946.5	229.150.717.126.229
OLS – Pocas variables	5810889269003.55	241.057.861.705.516
RidgeEducacionYActividad	5458455714585.57	233.633.381.916.745

Fuente: Elaboración propia con base a GEIH 2018 (DANE)

Anexo 10. Matriz de confusión (Pobres a partir de ingreso)

Y Test	0 No Pobre	1 Pobre
0 No Pobre	37220	2254
1 Pobre	5956	4058

Fuente: Elaboración propia con base a GEIH 2018 (DANE)