Universidad de los Andes
Maestría en Economía Aplicada
Andres Felipe Martinez, Angela Paola Morales Guio y Oscar Cortes
Repositorio: https://github.com/paolamguio/Problem_Set_2-G16

Predicting Poverty

1. Introduction

A pesar de los avances del país en la reducción de la pobreza, en el año 2018 Colombia se ubicó en términos de pobreza moderada en el puesto 11 de 14 países de Latinoamérica (CEPAL, 2019). Lo anterior resalta la importancia de evaluar el impacto de los programas sociales del Gobierno Nacional enfocados a la reducción de la pobreza, que requiere principalmente de la identificación de los hogares y personas pobres en las diferentes regiones del territorio colombiano para la entrega de ayudas monetarias y en especie a quien más lo necesite y les permita superar la línea de pobreza monetaria.

El principal instrumento de focalización individual en Colombia ha sido el Sistema de Identificación de Potenciales Beneficiarios (Sisbén), cuya cuarta versión se comenzó a implementar desde el año pasado. Adicionalmente, el sistema más rápido y fácil de clasificación de la pobreza es la estratificación socioeconómica, donde los estratos 1 y 2 se podrían considerar como pobres. A pesar de la importancia de identificar los pobres, un error de cálculo de la pobreza en el año 2010 conllevó a la exclusión de 4 millones de pobres en Colombia y durante la crisis del Covid 19, el Departamento Nacional de Planeación (DNP) identificó a 3 millones de hogares pobres que nunca habían recibido una transferencia directa del Estado (Fedesarrollo, 2021).

En concordancia con lo anterior, para la definición de los modelos de predicción en la identificación de un hogar o persona pobre, se estudiará la metodología del instrumento de focalización individual del Sisbén IV implementado por DNP recientemente y los resultados presentados por el DANE sobre pobreza monetaria 2021.

En este documento, la primera sección corresponde a la introducción. En la segunda sección se describe la metodología de la limpieza de los datos provenientes de la GEIH del año 2018. En la tercera sección se presentan los modelos y resultados empleados en la clasificación de los hogares pobres. En la cuarta sección se desarrolla la predicción de la pobreza a partir de la estimación del ingreso de los hogares. Finalmente, la última sección tiene las conclusiones.

2. Datos

El objetivo del trabajo se enfoca en construir modelos predictivos de pobreza a nivel hogar haciendo uso de la Gran Encuesta Integrada de Hogares - GEIH año 2018. Estos datos se encuentran segmentados en dos bases de datos, por un lado, se tiene la base que cuenta con la información de la variable pobreza e ingreso, con la cual se entrenarán los modelos y por el otro lado, se tiene la base de datos que no cuenta con estas variables y por lo tanto se realiza la predicción de los hogares pobres sobre esta última considerando el mejor modelo entrenado.

Partiendo de la base de entrenamiento, se realizaron los siguientes ajustes: i. Se colapsa la base de datos de personas para obtener información agrupada por hogar de las variables relevantes y se consolida con la base de datos de hogares, ii. Se procede a realizar ajuste de los missing values de cada variable, determinando dejar cero en estos valores para las variables: cuota de amortización, arriendo, horas trabajadas promedio, porcentaje de trabajo formal y tasa de desempleo. Lo anterior, debido a que los hogares que no reportaron cuota de amortización y arriendo se asumen que es porque no pagan estos conceptos, las personas que no tienen horas de trabajo, se asume que no se encuentran ocupadas, lo que también sucede para la tasa de trabajo formal, pata la tasa de desempleo, corresponde a hogares en que las PET están inactivas, y por lo tanto se asumió una tasa de desempleo cero ya que no se encuentran desempleadas. Para la variable años de educación promedio por hogar, se encontraron cuatro valores en missing value, se procedió a eliminarlos al no ser un número de hogares significativos.

Para la base de testing, se replicaron los ajustes mencionados anteriormente, sin embargo, se encontró un missing value para la educación promedio por hogar y se procedió a imputar el valor considerando la educación promedio de los hogares que tenían individuos con la misma edad promedio. Lo anterior se realizó considerando el objetivo del trabajo el cual es la proyección de hogares pobres y no pobres, más allá de la inferencia estadística que se pueda determinar por los datos.

Se eligen alrededor de 26 variables a analizar considerando que pueden servir para predecir pobreza, son variables que corresponden a la ciudad, tipo de vivienda, años de educación promedio, número de personas por hogar, número de cuartos, edad promedio, número de hijos promedio, porcentaje de mujeres, horas trabajadas promedio, tipo de trabajo (formal o informal), valor arriendo,

tasa de desempleo del hogar, tapa de ocupación y tasa de participación e ingreso promedio por hogar. Estas variables pueden servir para identificar si un hogar es pobre o no[1].

Tabla 1. Estadísticas descriptivas.

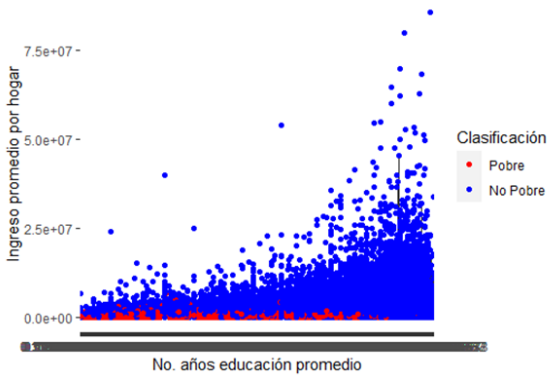
Descriptivas principales por hogar	Base training	Pobre		Base Testing
		Si	No	
No. Hogares	164,955	33,022 (20%)	131,933 (80%)	66,168
No. Personas	3.30 (1.78)	4.14 (2.03)	3.08 (1.64)	3.31 (1.79)
Edad promedio	37 (17)	31 (16)	39 (17)	37 (17)
Mujer como Jefe de Hogar	68,684 (42%)	15,501 (47%)	53,183 (40%)	27,838 (42%)
No. De hijos	1.17 (1.12)	1.74 (1.32)	1.02 (1.01)	1.17 (1.12)
Años educación	10.3 (4.1)	8.0 (3.5)	10.9 (4.1)	10.2 (4.1)
No. Personas con trabajo formal				
0	93,346 (57%)	29,060 (88%)	64,286 (49%)	38,282 (58%)
1	51,434 (31%)	3,822 (12%)	47,612 (36%)	20,181 (30%)
2	17,320 (10%)	134 (0.4%)	17,186 13%)	6,611 (10.0%)
Tipo de vivienda				
Propia, totalmente pagada	62,118 (38%)	9,051 (27%)	53,067 (40%)	25,235 (38%)
Propia, la están pagando	5,616 (3.4%)	537 (1.6%)	5,079 (3.8%)	2,148 (3.2%)
En arriendo o subarriendo	64,341 (39%)	14,435 (44%)	49,906 (38%)	25,310 (38%)
En usufructo	25,000 (15%)	5,169 (16%)	19,831 (15%)	10,259 (16%)
Posesión sin titulo	7,717 (4.7%)	3,786 (11%)	3,931 (3.0%)	3,178 (4.8%)
No. de personas por cuarto	1.73 (0.83)	2.25 (1.15)	1.60 (0.67)	1.73 (0.83)
Ingreso Total	1,801,441 (2,339,945)	$554,163 ($467,797)	$2,113,626 ($2,510,813)	
Horas trabajadas	39 (20)	34 (22)	40 (19)	39 (20)
Tasa ocupación	0.58 (0.34)	0.43 (0.31)	0.62 (0.33)	0.58 (0.34)
Tasa desempleo	0.09 (0.23)	0.17 (0.31)	0.07 (0.20)	0.09 (0.23)
Tasa participación	0.65 (0.33)	0.54 (0.31)	0.67 (0.32)	0.64 (0.33)

1 Mean (SD); n (%)

Fuente: Elaboración propia

De la tabla 1 se infiere que del total de 164.955 observaciones (hogares) de la base training, el 20% corresponde a hogares clasificados como pobres y el 80% restante a hogares no pobres, el 57% de los hogares en promedio sus individuos no cuentan con trabajo formal, el 42% de los hogares tienen como jefe de hogar a una mujer, se tiene una edad promedio de 34 años y 10,5 años de estudio en promedio, sin embargo, respecto a este último dato se evidencia brecha entre los hogares pobres y no pobres, teniendo como años promedio de educación 8 y 11 respectivamente, situación que esta enlazada con los ingresos promedio de los hogares como se observa en la siguiente gráfica:

Gráfico 1. Dispersión entre años de educación e ingreso total entre hogares pobres y no pobres



Fuente: Elaboración propia

Adicionalmente, se observa que a mayores años de educación disminuye la participación de los hogares pobres, siendo más notoria la brecha de ingresos, la cual también es evidente entre los hogares pobres y no pobres con igual número de años de educación.

Aunado a esto, se considera que existe una relación entre ser pobre, tener más hijos, trabajar menos horas a la semana, no contar con trabajo formal y no contar con vivienda propia, esto a su vez se estaría viendo reflejado en menos ingresos promedio de los hogares. Esto, se evidencia en los gráficos relacionados en el Anexo 1.

Por otro lado, comparando estas estadísticas generales de la base training respecto a la base testing, se concluye que que no existen diferencias significativas en la media de las variables presentadas, lo que quiere decir que ambas bases son muy similares.

3. Modelos y resultados

Modelos de clasificación.

[1] El proceso de generación de bases de datos se encuentra en los scripts 1. data_cleaning_training y 2. data_cleaning_testing incluidos en el repositorio.

Para el modelo de clasificación, se evaluaron un total de 54 modelos entre Logit, lasso, ridge, elastic net y remuestreo de estos, cutoffs alternativas para cada modelo, árbol, ramdom y forest, estos modelos se ejecutaron considerando dos selecciones de variables (modelo 1 con 47 variables y modelo con 77 variables, en el caso del modelo 1 son 26 variables que considerando los factores se convierten en 47 y para el caso del modelo 2 se considera un modelo más complejo que incluye interaciones de las variables como mujer-número de hijos, mujer-educación promedio y ciudad-educación).

Los modelos se evaluaron considerando las siguientes particiones de la base de datos de training: i. Training. Compuesta por el 70% de observaciones de la base training principal, se utilizó para la estimación de los modelos, ii. Evaluación. Corresponde a la tercera parte del 30% restante de la base de datos training principal, se utilizó para evaluar el cut-off óptimo para los modelos, iii. Testing. Compuesta por las dos terceras partes restantes de la base de datos training principal, se utilizó para predecir pobreza con los modelos estimados y obtener las matrices de clasificación con el fin de escoger el mejor modelo.

Se tiene como objetivo elegir el modelo con sensibilidad más alta ya que esta representa el número de verdaderos positivos respecto a la suma de verdaderos y falsos negativos, nos interesa predecir el menor número de falsos negativos debido a que hogares pobres y que serán soporte de políticas públicas, no deberían quedar clasificados de forma errónea.

Una vez analizada la sensibilidad de los 54 modelos, se elige como mejor modelo un Elastic-Net utilizando remuestreo con enfoque Up-sampling para las variables del modelo 2, resultado de comparar la sensibilidad de todos los modelos fuera de muestra. A continuación, se presentan los mejores cinco modelos:

Tabla 2. Resultados modelos clasificación

		Validación cruzada - training						Matriz de confusión fuera de muestra			Medidas fuera de muestra		
Modelo	Descripción	Cut-off	ROC	Sens	Espec	AUC	Kappa		No	Si	Sens	Espec	Accuracy
Logit - ridge down sample	Modelo 1	0,5	0,730	0,861	0,573	0,717	0,433	Si	885	5.719	0,8660	0,5644	0,6248
								No	14.892	11.494			
logit - ridge up sample	Modelo 1	0,5	0,818	0,858	0,599	0,729	0,458	Si	921	5.683	0,8605	0,5954	0,6485
								No	15.710	10.676			
Logit EN down sample	Modelo 2	0,5	0,888	0,837	0,773	0,805	0,610	Si	1.079	5.525	0,8366	0,7770	0,7889
								No	20.502	5.884			
Logit - EN up sample	Modelo 2	0,5	0,889	0,839	0,774	0,806	0,612	Si	1.083	5.521	0,8360	0,7764	0,7884
								No	20.487	5.899			
Forest	Modelo 1	N/A	0,906	0,575	0,942	0,869	0,559	Si	1.510	3.783	0,7147	0,8981	0,8687
								No	24.876	2.821			

Fuente: Elaboración propia

El modelo seleccionado presenta una sensibilidad de 0.8360 fuera de muestra encontrándose dentro de los más altos, sin embargo, no se selecciona el más alto ya que presenta un accuracy menor. Este modelo se entrenó con la partición training mencionada anteriormente realizando un ajuste a la submuestra por medio de up-sampling, al ser un elastic-net los hiperparámetros que se tuvieron en cuenta fueron Alpha y lambda, seleccionando los parámetros que presentaban mayor sensibilidad dentro de muestra con validación cruzada 5 fold[2].

Las variables más relevantes del modelo seleccionado que presentan coeficientes más altos fueron: % personas que reciben subsidio familiar, años educación promedio, valor arriendo mensual, tasa de ocupación, edad promedio, % personas con trabajo formal, reciben otros ingresos, horas trabajadas promedio semana, interacción edad promedio y % mujeres, ciudad de residencia Pereira, interacción educación promedio y % mujeres y ciudad de residencia Bogotá. se evidencia que las dos variables más relevantes subsidio familiar y años de educación, tienen sentido, pues se espera que a mayores años de educación la probabilidad de ser pobre disminuya[3].

Modelos de regresión con ingreso

Para el modelo de regresión, se evaluaron un total de 10 modelos entre OLS, lasso, ridge, elastic net y remuestreo (up sample y down sample), para cada modelo se utilizó 2 formulas, la primera de ellas con un total de 19 variables, la segunda formula se limitó solo a aquellas variables que se consideraban más significativas como lo son el tipo de vivienda, el dominio (ciudades), número de personas por cuarto, si el hogar vive en arriendo, edad promedio del hogar, edad promedio al cuadrado, si la cabeza del hogar es mujer, número de hijos, educación promedio del hogar, porcentaje de mujeres en el hogar, porcentaje de miembros del hogar que tienen un trabajo formal, y la tasa de ocupación de los miembros del hogar, las cuales correspondían con las variables identificadas por el DANE y el DNP que tenían mayor incidencia de pobreza en los hogares colombianos.

Los modelos se evaluaron considerando las siguientes particiones de la base de datos de training: i. Training. Compuesta por el 70% de observaciones de la base training principal, se utilizó para la estimación de los modelos con métrica RMSE ii. Evaluación. Corresponde al otro 30% restante de la base de datos training principal iii. Clasificación. Se clasificaron que hogares eran pobres, es decir, cuyos ingresos estimados estaban por debajo de la línea de progreso y se procedió realizar la matriz de confusión. Finalmente,

[2] Adicionalmente, en las estimaciones se consideró otro hiperparámetro correspondiente a seleccionar el cut-off optimo, sin embargo, el modelo seleccionado con cut-off 0.5 presentó mejor desempeño fuera de muestra que los modelos con cut-off optimo. Ver Anexo 2.

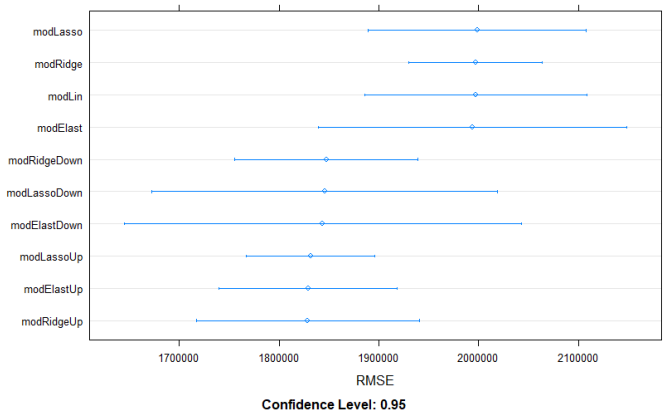
[3] Ver anexo 3.

se encontró que los modelos con la fórmula con 19 variables tenían mejor ajuste que la simplificada con las variables más importantes, los resultados se presentan a continuación:

Tabla 3. Resultados modelos regresión con ingreso

Modelo	RMSE	Sens	Spec
OLS	2,581,912	0.6326	0.8456
Lasso	2,498,772	0.6196	0.8526
Ridge	2,579,860	0.6301	0.8460
Elasticnet	2,579,908	0.6300	0.8459
Lasso Upsample	2,319,757	0.6676	0.8329
Ridge Upsample	2,363,387	0.6752	0.8279
Elasticnet Upsample	2,363,741	0.6752	0.8278
Lasso Downsample	2,425,635	0.6690	0.8316
Ridge Downsample	2,490,962	0.6768	0.8257
Elasticnet Downsample	2,490,691	0.6766	0.8256

Gráfico 2. Intervalos de confianza de RMSE



Se tiene como objetivo elegir el modelo con sensibilidad más alta ya que esta representa el número de verdaderos positivos respecto a la suma de verdaderos y falsos negativos, nos interesa predecir el menor número de falsos negativos debido al efecto que tiene clasificar a hogares como no pobres que si lo son y que no reciben ayudas de los Gobiernos nacionales y subnacionales al igual que entidades no gubernamentales.

Una vez analizada la sensibilidad de los 40 modelos, se elige como mejor modelo un ridge utilizando remuestreo con enfoque Down-sampling para las variables del modelo 1, resultado de comparar la sensibilidad de todos los modelos fuera de muestra. También se llevó a cabo la identificación de las principales variables en la regresión: edad promedio, educación promedio, tipo de, otros ingresos, edad promedio al cuadrado, si vive en arriendo y el número del hogar

4. Conclusiones y recomendaciones:

Como se evidenció a lo largo del documento, existen diferentes métodos de estimación para realizar predicciones de variables discretas y continuas, no existen modelos malos y buenos, el resultado final depende de la base de datos utilizada y las variables seleccionadas, así como del criterio de la importancia de las medidas de clasificación para evaluar.

Para este trabajo, es representativo seleccionar si un hogar es pobre o no y por lo tanto, los hogares que son predichos como no pobres y en realidad lo son pueden estar quedando por fuera de políticas que se pretendan implementar, por esta razón, la medida que nos interesó fue la sensibilidad. Acorde con los datos arrojados, el modelo de clasificación óptimo para el ejercicio es el elastic-net por remuestreo upsample que da como resultado una sensibilidad de 0.8360 fuera de muestra y utilizando este modelo en la base de datos de testing, arroja que el 35,9% de los hogares de la muestra son pobres. En cuanto al modelo de ingreso ridge por remuestreo downsample, se seleccionó considerando un RMSE de 2,490,962, arrojando para la muestra de testing un total de 67.68% de hogares pobres.

Los modelos seleccionados pueden ser utilizados para clasificar hogares de los cuales no se cuente con información del ingreso promedio y servirán para orientar al gobierno en la focalización correcta de política.

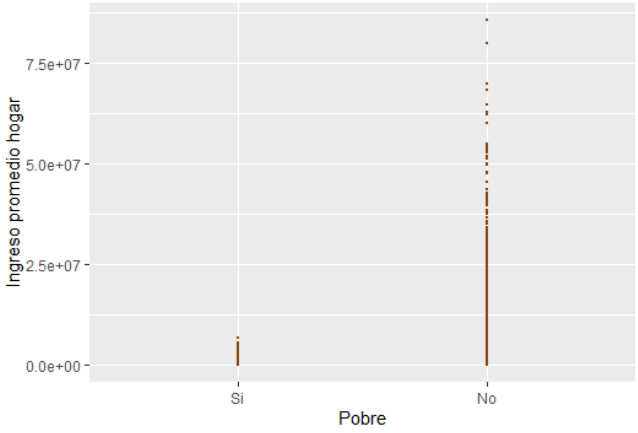
Anexo 1. Análisis de Datos – Modelos de clasificación Base de datos training

Tabla 1. Estadísticas descriptivas generales

Estadísticas por hogar	nbr.null	nbr.na	min	max	range	median	mean	std.dev
No. Personas	-	-	1	28	27	3	3,30	1,78
Li	-	-	99544,84	131125,57	31580,72	121603,86	120415,70	7179,03
Lp	-	-	167222,48	303816,6902	136594,21	280028,7104	271604,40	33543,76
No. cuartos por hogar	-	-	1	18	17	3	3,39	1,21
No. de personas por cuarto	-	-	0,2	16	15,8	1,5	1,73	0,83
Edad promedio	-	-	5,67	102	96,33	33,5	37,45	16,88
No. hijos	55974	-	-	12	12	1	1,17	1,12
No. de personas con trabajo formal	93346	-	-	8	8	-	0,58	0,76
Años educación promedio	2605	-	-	27	27	10,5	10,30	4,10
No. horas trabajadas promedio	22138	-	-	130	130	44	39,18	19,65
Ingreso total por hogar	13220	-	-	85833333,33	85833333,33	1187717	1801440,60	2339945,38
Participación en trabajo formal	93346	-	-	1	1	-	0,34	0,43
% personas con subsidio familiar	80203	-	-	9	9	0,2	0,42	0,56
Si alguna persona cuenta con segundo trabajo	153588	-	-	1	1	-	0,07	0,25
Si alguna persona cuenta con otros ingresos	71127	-	-	1	1	1	0,57	0,50
Si alguna persona cuenta con otros ingresos instituciones	140035	-	-	1	1	-	0,15	0,36
Tasa ocupación	22138	-	-	1	1	0,5	0,58	0,34
Tasa desempleo	138670	-	-	1	1	-	0,09	0,23
Tasa inactivas	59582	-	-	1	1	0,33	0,35	0,33
Tasa participación	16498	-	-	1	1	0,67	0,65	0,33

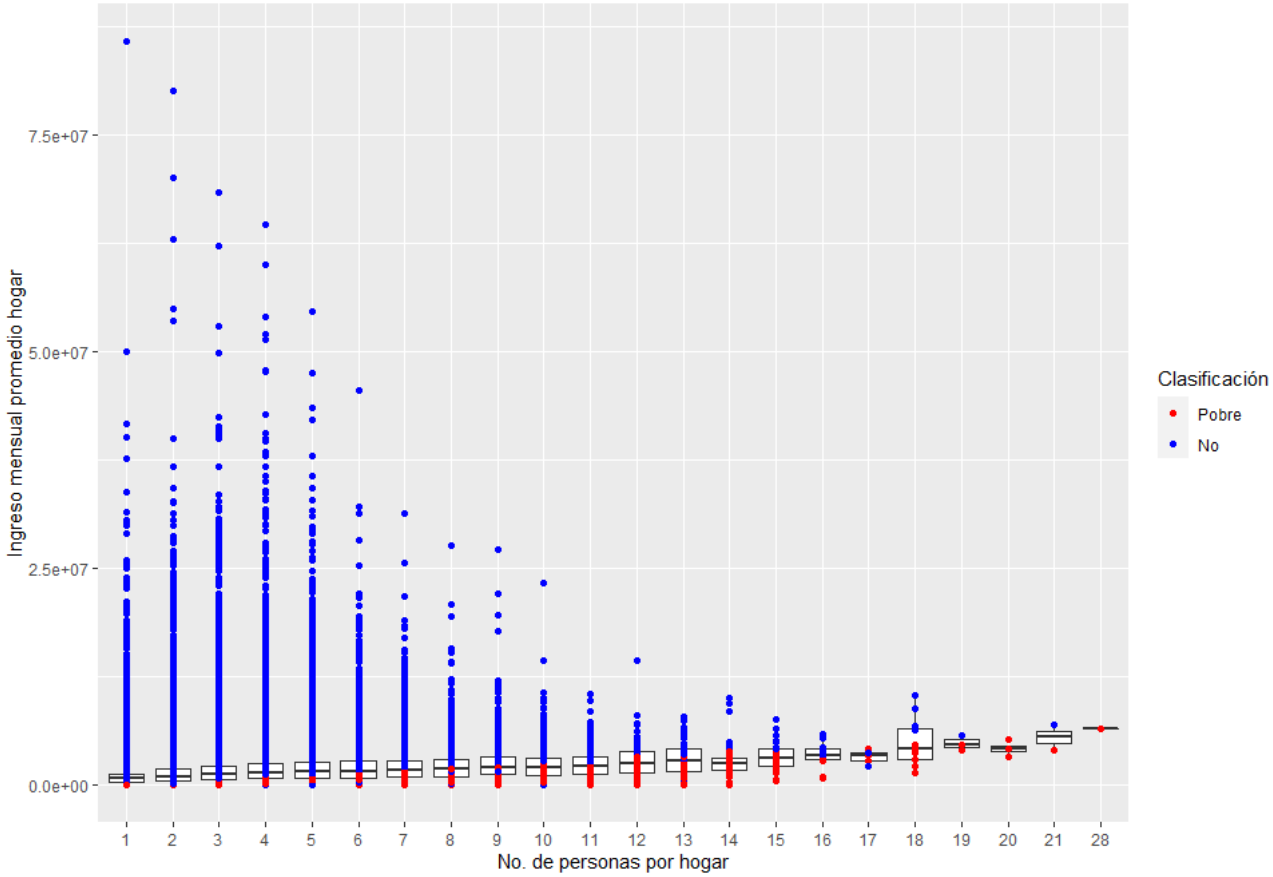
Fuente: Elaboración propia

Gráfico 1. Ingresos promedio mes entre hogares pobres y no pobres



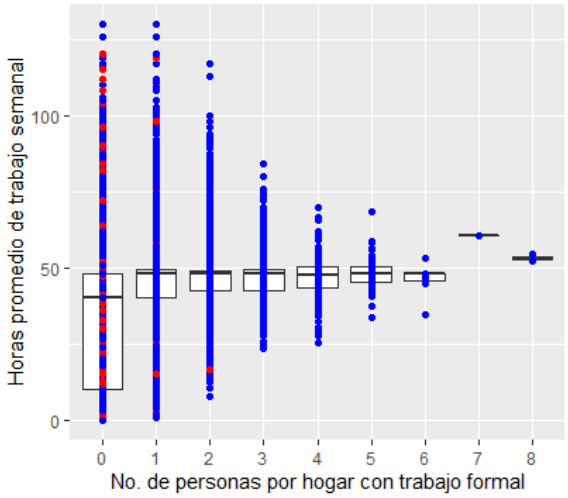
Fuente: Elaboración propia

Gráfico 2. Box-Plot por número de personas e ingresos en los hogares pobres y no pobres



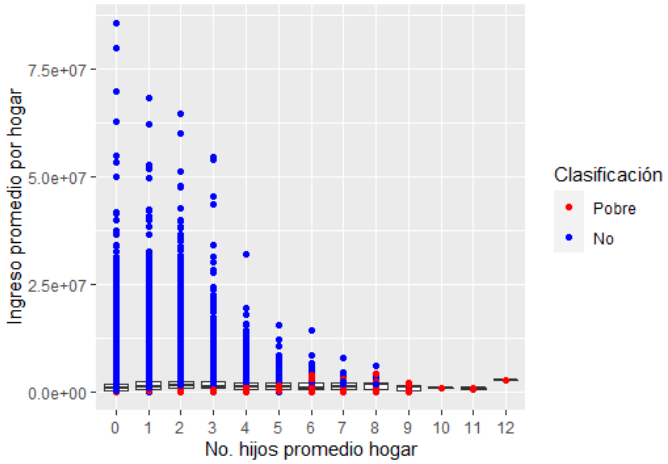
Fuente: Elaboración propia

Gráfico 3. Box-Plot por número de personas con trabajo formal y número de horas trabajadas en los hogares pobres y no pobres



Fuente: Elaboración propia

Gráfico 4. Box-Plot por número de hijos e ingresos promedio en los hogares pobres y no pobres



Fuente: Elaboración propia

Las gráficas de caja muestran que el ingreso promedio por hogar depende del número de personas por hogar, se puede evidenciar que a mayor número de personas en el hogar, se presentan menores ingresos promedio hogar y aumenta el número de hogares pobres. Los hogares con menor número de personas reportan mayores ingresos, esta relación también se observa con el número de hijos promedio por hogar.

Tabla 2. Estadísticas descriptivas generales base de datos testing.

Estadísticas por hogar	nbr.null	nbr.na	min	max	range	median	mean	std.dev
No. Personas	0	0	1	22	21	3	3,312311087	1,79076566
No. De personas por cuarto	0	0	0,2	12	11,8	1,5	1,733690391	0,833106227
Edad promedio	0	0	7	101	94	33,4	37,43497851	16,87501555
No. hijos	22397	0	0	11	11	1	1,166651554	1,118659572
No. de personas con trabajo formal	38282	0	0	7	7	0	0,556945956	0,749621974
Años educación promedio	1075	0	0	27	27	10,4	10,20779081	4,097525066
No. Horas trabajadas promedio	8905	0	0	130	130	44	39,02967523	19,65453794
% personas subsidio familiar	32897	0	0	9	9	0,142857143	0,409732954	0,557371703
Si alguna persona cuenta con segundo trabajo	61678	0	0	1	1	0	0,067857575	0,251503241
Si alguna persona cuenta con otros ingresos	29200	0	0	1	1	1	0,558699069	0,496546217
Si alguna persona cuenta con otros ingresos instituciones	55770	0	0	1	1	0	0,157145448	0,363940597
Tasa ocupación	8905	0	0	1	1	0,5	0,579627289	0,338375389
Tasa desempleo	55445	0	0	1	1	0	0,089606273	0,227929196
Tasa participación	6651	0	0	1	1	0,666666667	0,644219633	0,326715732

Fuente: Elaboración propia

Anexo 2.

Tabla 3. Resultados modelos de clasificación

		Validación cruzada - training						Matriz de confusión fuera de muestra			Medidas fuera de muestra		
Modelo	**Descripción**	**Cut-off**	**ROC**	**Sens**	**Espec**	**AUC**	**Kappa**		**No**	**Si**	**Sens**	**Espec**	**Accuracy**
Logit	Modelo 1	0,50	0,898	0,526	0,946	0,862	0,523	Si	3.154	3.450	0,5224	0,9475	0,8624
								No	25.000	1.386			
	Modelo 2		0,899	0,529	0,946	0,863	0,526	Si	3.126	3.478	0,5267	0,9461	0,8622
								No	24.965	1.421			
Logit rfThresh	Modelo 1	0,21		0,825	0,803	0,897		Si	1.166	5.438	0,8234	0,8069	0,8102
								No	21.290	5.096			
	Modelo 2	0,21		0,825	0,805	0,897		Si	1.162	5.442	0,8240	0,8083	0,8114
								No	21.327	5.059			
Logit - Lasso	Modelo 1	0,50	0,897	0,477	0,957	0,861	0,499	Si	3.479	3.125	0,4732	0,9562	0,8595
								No	25.230	1.156			
	Modelo 2		0,898	0,482	0,956	0,861	0,503	Si	3.435	3.169	0,4799	0,9556	0,8604
								No	25.215	1.171			
Logit - Lasso rfThresh	Modelo 1	0,23		0,81	0,811	0,895		Si	1.244	5.360	0,8116	0,8153	0,8146
								No	21.513	4.873			
	Modelo 2	0,22		0,815	0,804	0,895		Si	1.204	5.400	0,8177	0,8110	0,8123
								No	21.399	4.987			
Logit - Ridge	Modelo 1	0,50	0,898	0,525	0,947	0,863	0,524	Si	3.161	3.443	0,5214	0,9479	0,8625
								No	25.010	1.376			
	Modelo 2		0,899	0,528	0,947	0,863	0,526	Si	3.135	3.469	0,5253	0,9475	0,8630
								No	25.000	1.386			
Logit - Ridge rfThresh	Modelo 1	0,21		0,827	0,801	0,896		Si	1.163	5.441	0,8239	0,8064	0,8099
								No	21.278	5.108			
	Modelo 2	0,21		0,826	0,803	0,897		Si	1.151	5.453	0,8257	0,8058	0,8098
								No	21.263	5.123			
Logit - ElasticNet	Modelo 1	0,50	0,899	0,523	0,947	0,862	0,523	Si	3.168	3.436	0,5203	0,9478	0,8622
								No	25.009	1.377			
	Modelo 2		0,899	0,528	0,947	0,863	0,525	Si	3.146	3.458	0,5236	0,9469	0,8622
								No	24.985	1.401			
Logit - ElasticNet rfThresh	Modelo 1	0,21		0,825	0,803	0,896		Si	1.175	5.429	0,8221	0,8077	0,8105
								No	21.311	5.075			
	Modelo 2	0,21		0,828	0,801	0,897		Si	1.141	5.463	0,8272	0,8041	0,8087
								No	21.217				

		Validación cruzada - training						Matriz de confusión fuera de muestra			Medidas fuera de muestra		
Modelo	**Descripción**	**Cut-off**	**ROC**	**Sens**	**Espec**	**AUC**	**Kappa**		**No**	**Si**	**Sens**	**Espec**	**Accuracy**
										5.169			
Logit - lasso up sample	Modelo 1	0,50	0,896	0,084	0,790	0,814	0,627	Si	1.092	5.512	0,8346	0,7928	0,8012
								No	20.920	5.466			
	Modelo 2		0,898	0,836	0,794	0,815	0,631	Si	1.099	5.505	0,8336	0,7969	0,8043
								No	21.028	5.358			
Logit - lasso up sample rfThresh	Modelo 1	0,52		0,818	0,801	0,894		Si	1.197	5.407	0,8187	0,8063	0,8088
								No	21.275	5.111			
	Modelo 2	0,51		0,826	0,795	0,895		Si	1.126	5.478	0,8295	0,8011	0,8068
								No	21.137	5.249			
logit - ridge up sample	Modelo 1	0,50	0,818	0,858	0,599	0,729	0,458	Si	921	5.683	0,8605	0,5954	0,6485
								No	15.710	10.676			
	Modelo 2		0,818	0,858	0,598	0,728	0,457	Si	921	5.683	0,8605	0,5954	0,6485
								No	15.710	10.676			
logit - ridge up sample rfThresh	Modelo 1	0,53		0,759	0,728	0,825		Si	1.683	4.921	0,7452	0,7226	0,7271
								No	19.067	7.319			
	Modelo 2	0,53		0,759	0,728	0,825		Si	1.683	4.921	0,7452	0,7226	0,7271
								No	19.067	7.319			
Logit - EN up sample	Modelo 1	0,50	0,897	0,839	0,789	0,814	0,627	Si	1.100	5.504	0,8334	0,7908	0,7993
								No	20.865	5.521			
	Modelo 2		0,889	0,839	0,774	0,806	0,612	Si	1.083	5.521	0,8360	0,7764	0,7884
								No	20.487	5.899			
Logit - EN up sample rfThresh	Modelo 1	0,53		0,813	0,808	0,895		Si	1.240	5.364	0,8122	0,8129	0,8128
								No	21.450	4.936			
	Modelo 2	0,52		0,818	0,793	0,889		Si	1.235	5.369	0,8130	0,7984	0,8013
								No	21.066	5.320			
Logit - lasso down sample	Modelo 1	0,50	0,894	0,836	0,787	0,811	0,623	Si	1.085	5.519	0,8357	0,7906	0,7996
								No	20.861	5.525			
	Modelo 2		0,895	0,835	0,789	0,812	0,625	Si	1.096	5.508	0,8340	0,7936	0,8017
								No	20.940	5.446			
Logit - lasso down sample rfThresh	Modelo 1	0,50		0,834	0,787	0,893		Si	1.092	5.512	0,8346	0,7922	0,8007
								No	20.902	5.484			
	Modelo 2	0,51		0,822	0,795	0,894		Si	1.141	5.463	0,8272	0,8004	0,8058
								No	21.119	5.267			
Logit - ridge down sample	Modelo 1	0,50	0,730	0,861	0,573	0,717	0,433	Si	885	5.719	0,8660	0,5644	0,6248
								No	14.892	11.494			
	Modelo 2		0,809	0,860	0,586	0,723	0,446	Si				0,5813	0,6375

		Validación cruzada - training						Matriz de confusión fuera de muestra			Medidas fuera de muestra		
Modelo	**Descripción**	**Cut-off**	**ROC**	**Sens**	**Espec**	**AUC**	**Kappa**		**No**	**Si**	**Sens**	**Espec**	**Accuracy**
									909	5.695	0,8624		
								No	15.337	11.049			
Logit - ridge down sample rfThresh	Modelo 1	0,50		0,849	0,606	0,739		Si	1.121	5.483	0,8303	0,5997	0,6459
								No	15.824	10.562			
	Modelo 2	0,53		0,737	0,738	0,818		Si	1.872	4.735	0,7167	0,7342	0,7307
								No	19.372	7.014			
Logit EN down sample	Modelo 1	0,50	0,887	0,838	0,772	0,805	0,609	Si	1.090	5.514	0,8349	0,7761	0,7879
								No	20.479	5.907			
	Modelo 2		0,888	0,837	0,773	0,805	0,610	Si	1.079	5.525	0,8366	0,7770	0,7889
								No	20.502	5.884			
Logit EN down sample rfThresh	Modelo 1	0,52		0,82	0,791	0,887		Si	1.231	5.373	0,8136	0,7967	0,8001
								No	21.022	5.364			
	Modelo 2	0,52		0,823	0,789	0,888		Si	1.210	5.394	0,8168	0,7952	0,7995
								No	20.981	5.405			
Logit -lasso smote oversampled	Modelo 1	0,50	0,887	0,921	0,559	0,766	0,501	Si	3.710	2.894	0,4382	0,0789	0,1508
								No	2.082	24.304			
	Modelo 2		0,891	0,912	0,608	0,782	0,539	Si	3.866	2.738	0,4146	0,0880	0,1534
								No	2.322	24.064			
Logit -lasso smote oversampled rfThresh	Modelo 1	0,56		0,819	0,772	0,879		Si	5.401	1.203	0,1822	0,2252	0,2166
								No	5.943	20.443			
	Modelo 2	0,57		0,811	0,769	0,875		Si	5.373	1.231	0,1864	0,2252	0,2175
								No	5.943	20.443			
Logit ridge SMOTE Oversampled	Modelo 1	0,50	0,5	1	0	0,571	0	Si		6604	1,0000	-	0,2002
								No		26386			
	Modelo 2		0,5	1	0	0,571	0	Si		6604	1,0000	-	0,2002
								No		26386			
Logit EN SMOTE	Modelo 1	0,50	0,904	0,843	0,785	0,818	0,629	Si	5062	1542	0,2335	0,1539	0,1698
								No	4060	22326			
	Modelo 2		0,907	0,849	0,795	0,826	0,644	Si	5009	1595	0,2415	0,1490	0,1675
								No	3931	22455			
Logit EN SMOTE rfThresh	Modelo 1	0,54		0,812	0,809	0,894		Si	5362	1242	0,1881	0,1867	0,1870
								No	4927	21459			
	Modelo 2	0,56		0,815	0,802	0,893		Si	5409	1195	0,1810	0,1935	0,1910
								No	5105	21281			
Tree	Modelo 1	N/A	0,872	0,546	0,917	0,842	0,485	Si	1495	3414	0,6955	0,8864	0,8580
								No	24891	3190			
	Modelo 2		0,869	0,534	0,935	0,855	0,509	Si	1548	3433	0,6892	0,8868	0,8570
								No	24838	3171			
forest	Modelo 1	N/A	0,906	0,575	0,942	0,869	0,559	Si	1510	3783	0,7147	0,8981	0,8687
								No	24876	2821			
	Modelo 2		0,898	0,568	0,939	0,864	0,545	Si	1585	3698	0,7000	0,8951	0,8639
								No	24801	2906			

Fuente: Elaboración propia

Anexo 3.

Tabla 4. Coeficientes modelo elastic-net up samplimg.

Predictores por hogar	Coeficiente
% personas que reciben subsidio familiar	0,8399
Años educación promedio	0,7573
Valor arriendo mensual	0,7474
Tasa de ocupación	0,6522
Edad promedio	0,5686
% personas con trabajo formal	0,5542
Reciben otros ingresos	0,5266
Horas trabajadas promedio semana	0,3854
Interacción edad promedio y % mujeres	0,1830
Ciudad de residencia Pereira	0,1378
Interacción educación promedio y % mujeres	0,1293
Ciudad de residencia Bogotá	0,1264

Fuente: Elaboración propia

La tabla anterior presenta las variables más representativas del modelo, en total se estimó el modelo con 76 variables, resultados reportados en el archivo excel denominado coeficientes que reposa en la carpeta 3.Stores.

Anexo 4.

Gráfico 1. Coeficientes de regresión modelo Ridge Downsampling.

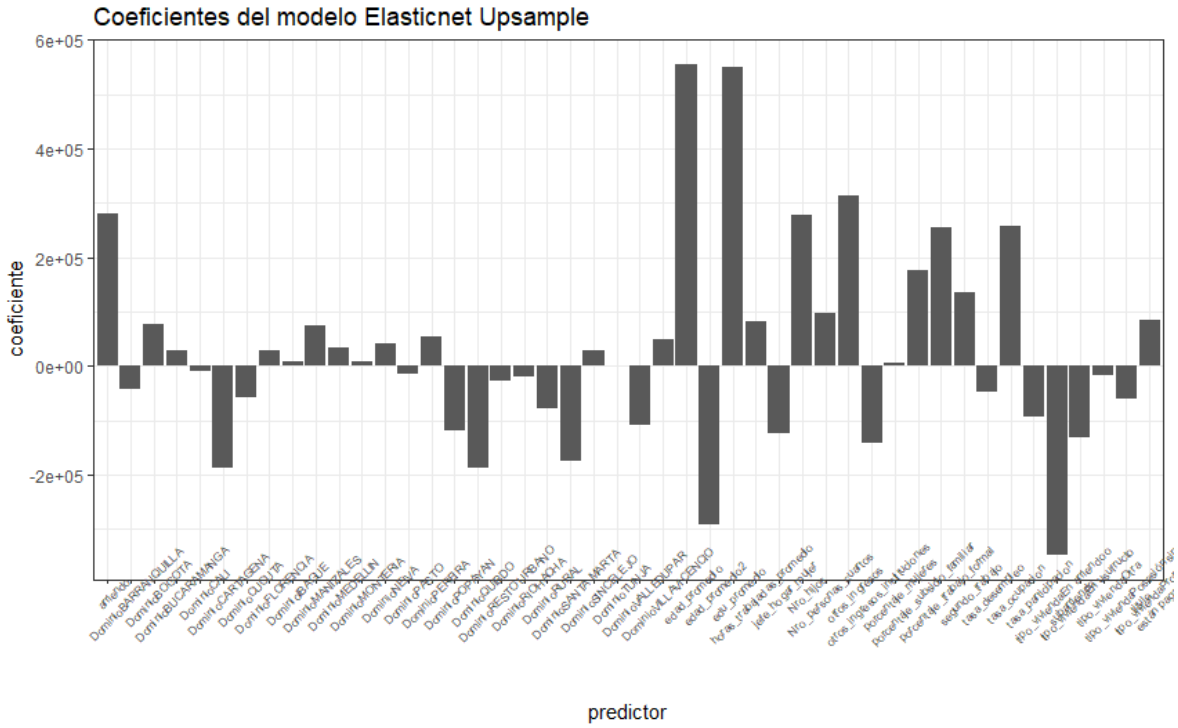


Tabla 5. Coeficientes de regresión modelo Ridge Downsampling.

predictor	coeficiente	Abs (Dev Std)
(Intercept)	1486457.174	1486457.174
edad_promedio	555192.9102	555192.9102
edu_promedio	550847.8654	550847.8654
tipo_viviendaEn arriendo osubarriendo	-347055.4212	347055.4212
otros_ingresos	313627.6142	313627.6142
edad_promedio2	-290718.8685	290718.8685
arriendo	280700.9017	280700.9017
Nro_hijos	277899.066	277899.066
tasa_ocupacion	258388.7373	258388.7373
porcentaje_trabajo_formal	254559.8658	254559.8658
DominioQUIBDO	-188304.2995	188304.2995
DominioCARTAGENA	-187603.7033	187603.7033
porcentaje_subsidio_familiar	176433.7423	176433.7423
DominioSANTA MARTA	-173987.8096	173987.8096
otros_ingresos_instituciones	-140250.0694	140250.0694
segundo_trabajo	135803.7332	135803.7332
tipo_viviendaEn usufructo	-130183.5559	130183.5559
jefe_hogar_mujer	-123773.9733	123773.9733
DominioPOPAYAN	-118956.8554	118956.8554
DominioVALLEDUPAR	-107513.9449	107513.9449
Nro_personas_cuartos	98650.77367	98650.77367
tasa_participacion	-93047.76906	93047.76906
tipo_viviendaPropia, laestán pagando	84172.92661	84172.92661
horas_trabajadas_promedio	81321.52437	81321.52437
DominioBOGOTA	77660.78843	77660.78843
DominioRURAL	-77152.15018	77152.15018
DominioMANIZALES	74980.96459	74980.96459
tipo_viviendaPosesión sintitulo	-59135.26661	59135.26661
DominioCUCUTA	-57181.04514	57181.04514
DominioPEREIRA	53561.21713	53561.21713
DominioVILLAVICENCIO	50357.51618	50357.51618
tasa_desempleo	-46841.82403	46841.82403
DominioBARRANQUILLA	-41394.82208	41394.82208
DominioNEIVA	40436.59567	40436.59567
DominioMEDELLIN	33927.12556	33927.12556
DominioFLORENCIA	28767.66993	28767.66993
DominioBUCARAMANGA	28690.1532	28690.1532
DominioSINCELEJO	27882.46987	27882.46987
DominioRESTO URBANO	-27296.09206	27296.09206
DominioRIOHACHA	-18765.6824	18765.6824
tipo_viviendaOtra	-15934.08487	15934.08487
DominioPASTO	-14112.58417	14112.58417
DominioCALI	-10171.54719	10171.54719
DominioIBAGUE	8047.697537	8047.697537
DominioMONTERIA	7822.066789	7822.066789
porcentaje_mujeres	5023.960402	5023.960402
DominioTUNJA	0	0