Problem Set 2

Autores: Juan Sebastián Vásquez Acevedo y Walter Leonardo Sanchez Salazar
Fecha inicio: 28/06/2022
Enlace a repositorio: https://github.com/wsanch92/Problem_set_2_WS_JSV

Introducción

La pobreza no es un concepto que tenga solo una definición, esto ha generado el interés de diferentes sectores por estudiarla, conceptualizarla y calcularla. Los diferentes estudios entorno a la pobreza han permitido identificar que esta va más allá de la escasez o la falta de recursos que pueden tener los hogares que se encuentran en esta condición. La pobreza genera diferentes entornos que causan que los hogares no cuenten con las herramientas para poder superar su condición de pobreza (en la literatura esto se conoce como las trampas de pobreza). En Colombia, existen dos metodologías para el calculo de la pobreza, una de ellas es la pobreza monetaria, en la cual, bajo unas líneas de pobreza establecidas, se logra identificar a la población pobre y pobre extrema. Por otro lado, esta el Índice de Pobreza Multidimensional, cuya metodología busca identificar a los hogares que se encuentran privados en al menos cinco de las quince carencias que se analizan.

Colombia también cuenta con el Sisbén, que es el sistema de focalización de potenciales beneficiarios de programas sociales, el cual intenta aproximar las condiciones socioeconómicas de la población por medio de una encuesta, clasificándolas en 4 grupos: pobre extremo, pobre, vulnerables y no pobres. Este deseo de identificación de la población pobre ha sido la motivación del presente documento, en el cual se han corrido diferentes modelos para clasificación y para la estimación de ingresos. Para entrenar los modelos se utilizó la Gran Encuesta Integrada de Hogares (GEIH) del 2018. El mejor modelo de clasificación fue un Random Forest que arrojó una sensibilidad de 0.92. Predijo al 33% de los hogares como pobres y al 66% como no pobres. Por otro lado, el modelo de predicción de ingresos predijo que el 9% de los hogares son pobres y el 90% no lo son.

Datos

A continuación, se presentará una tabla que contiene las variables seleccionadas para la elaboración de los modelos de clasificación.

Tabla 1 Variables modelo de clasificación

Variables		
Hogares	164,959	
Pobreza		
Pobre	41,973	(0,25)
No Pobre	122,986	(0,75)
Actividad del jefe de hogar		
Trabajando	104,095	(0,63)
Buscando trabajo	3,934	(2.4%)
Estudiando	2,422	(1.5%)
Oficios del hogar	33,025	(0,2)
Incapacitado permanente para trabajar	2,376	(1.4%)
Otra actividad	19,107	-0,12

Nivel educativo del jefe de Hogar		
Ninguno	8,603	(5.2%)
Preescolar	13	(<0.1%)
Básica primaria	46,619	(0,28)
Básica secundaria	21,615	(0,13)
Media	43,028	(0,26)
Superior o universitaria	45,061	(0,27)
No sabe, no informa	20	(<0.1%)
Ocupación de la vivienda		
Propia, totalmente pagada	62,120	(0,38)
Propia, la están pagando	5,616	(3.4%)
En arriendo o subarriendo	64,343	(0,39)
En usufructo	25,000	(0,15)
Posesión sin titulo	7,717	(4.7%)
Otra	163	(<0.1%)
Edad promedio del hogar	37	(17)
Personas por Unidad de gasto	3.28	(1.77)
Número de cuartos exclusivos	1.99	(0.90)
Proporción de hombres	0.48	(0.28)
Proporción de mujeres	0.52	(0.28)
Proporción de personas en régimen contributivo y especial	0.45	(0.41)
Proporción de personas en régimen Subsidiado	0.36	(0.39)
Proporción de personas que cotizan pensión	0.19	(0.28)
Proporción de personas que No cotizan pensión	0.29	(0.32)
Proporción de personas pensionadas	0.0115	(0.0807)
[1] n (%); Mean (SD)		

Fuente: elaboración propia datos GEIH 2018

Para los modelos de clasificación se seleccionaron 14 variables, entre ellas la variable categórica que define los pobre y no pobres. Esta variable nos permite evidenciar que del total de hogares (164,959) el 25% son pobres y el 75% no lo son. Los 13 restantes fueron seleccionados por la relación que guardan con la pobreza, como es el caso del número de cuartos exclusivos en el hogar. Esta variable esta relacionada con la vivienda digna, así como con el hacinamiento que en ocasiones en los hogares pobres puede existir. En la base de entrenamiento, el número de cuartos exclusivos de los hogares en promedio es de 1.99, con una desviación estándar de 0.9. El nivel educativo del jefe de hogar es otra de las variables que incluimos en el modelo, ya que como lo ha mostrado la literatura a mayor nivel educativo se esperan mayores niveles de ingreso, los cuales guardan una relación con la pobreza monetaria. El 27% de los jefes de hogar manifiesta tener educación superior o universitaria y el 28% educación básica primaria. Así mismo, incluimos la actividad del jefe de hogar, en donde podemos evidenciar que en su mayoría (63%) manifiestan estar trabajando, seguido del 20% que manifiestan como su actividad principal los oficios del hogar, el 2.4% se encuentran buscando trabajo y el 1.5% estudiando.

Otra de las situaciones que pueden generar menores ingresos y de alguna forma incidencia en pobreza es la dependencia de algunos integrantes del hogar en el jefe. Por ello calculamos la edad media del hogar, que en la base analizada en promedio es de 37 años, así como el número de personas por unidad de gasto, en donde podemos ver un promedio de 3.28 personas con una desviación estándar de 1.77. Así mismo calculamos la proporción tanto de hombres como de mujeres en el hogar la cual corresponde en promedio a 0.48 y 0.52 respectivamente. Con esto buscamos capturar las diferencias que existen en materia de género, así como la carga de cuidado que recae en las mujeres y que en la mayoría de las ocasiones no es remunerada.

Para el modelo de predicción de ingresos se utilizaron las siguientes variables:

Tabla 2 Variables modelo de predicción de ingresos

Variables		
Personas	543,583	
Género		
Hombre	256,783	(47%)
Mujer	286,800	(53%)
Actividad		
Trabajando	208,956	(47%)
Buscando trabajo	13,491	(3.0%)
Estudiando	72,874	(16%)
Oficios del hogar	105,103	(23%)
Incapacitado permanente para trabajar	7,338	(1.6%)
Otra actividad	40,384	(9.0%)
Missing	95,437	
Posición ocupacional		
Obrero o empleado de empresa particular	92,078	(37%)
Obrero o empleado del gobierno	12,763	(5.1%)
Empleado doméstico	7,975	(3.2%)
Trabajador por cuenta propia	114,909	(46%)
Patrón o empleador	9,169	(3.7%)
Trabajador familiar sin remuneración	7,658	(3.1%)
Trabajador sin remuneración en empresa o negocios de otros hogares	712	(0.3%)
Jornalero o peón	3,274	(1.3%)
Otro	145	(<0.1%)
Missing	294,900	
Cotizante a pensión		
Sí	95,121	(38%)
No	147,941	(60%)
Ya es pensionado	4,568	(1.8%)
Missing	295,953	
Ocupación de la vivienda		
Propia, totalmente pagada	213,123	(39%)
Propia, la están pagando	18,952	(3.5%)
En arriendo o subarriendo	200,479	(37%)
En usufructo	81,050	(15%)
Posesión sin titulo	29,471	(5.4%)
Otra	508	(<0.1%)
Recibe ayudas institucionales		
Sí	28,918	(31%)
No	64,759	(69%)
No sabe, no informa	17	(<0.1%)
Missing	449,889	
Ingreso total sin imputar	773,258	(1,372,249)
Missing	95,437	
Ingreso total imputado	637,497	(1,280,236)
Edad promedio del hogar	34	(22)
Horas trabajadas al mes	86	(113)
Missing	294,900	
n (%); Mean (SD)		

Fuente: elaboración propia datos GEIH 2018

En total, la base cuenta con 543,583 personas, de la cual el 53% son mujeres y el 47% hombres. Otra variable que incluimos en el modelo fue actividad, en esta variable fue necesario imputar los datos faltantes. Al momento de la imputación logramos identificar que los datos correspondían a niños, niñas y adolecentes (NNA) entre los 0 y 11 años, por lo cual tomamos la decisión de que los niños y niñas entre 0 y 2 años entrarían en el grupo de otra actividad y NNA entre 3 y 11 años se imputarían con la actividad de estudiando. Posición ocupacional fue otra variable en la cual fue necesario realizar imputaciones, en este caso, se imputaron los valores faltantes como cero, ya que se logro identificar que los datos que no contaban con una ocupación se debían a que en la variable de actividad

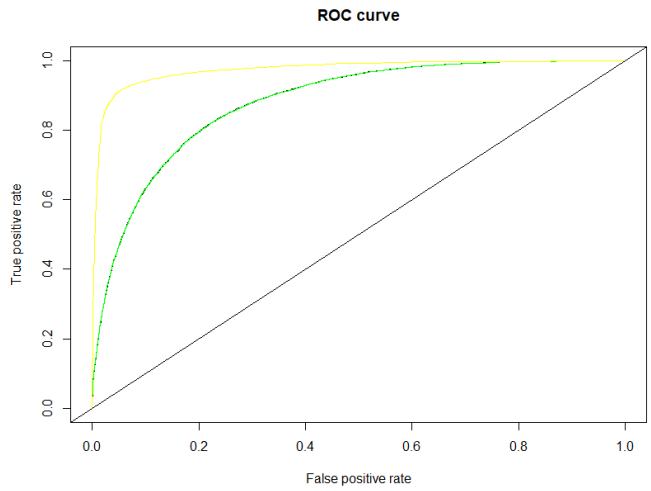
habían reportado no estar trabajando. Esto mismo se comprobó en la variable de horas trabajando por lo cual se aplicó la metodología mencionada anteriormente.

Recibe ayudas institucionales fue otra variable que se incluyo en el modelo, en nuestra base de entrenamiento el 31% de las personas manifestaron recibir alguna ayuda mientras que el 69% contestaron que no. Esta variable contaba con un total de 449,889 datos faltantes y 17 que correspondían a la respuesta "no sabe o no informa", los cuales fueron imputados como 0, ya que supondremos que las personas prefieren contestar que no reciben una ayuda del estado para así recibir algún beneficio futuro. Para los ingresos, decidimos filtrar la base para las personas mayores de 15 y que sus ingresos fueran positivos, esto nos llevo a que el ingreso promedio de las personas fuera de $637,497 pesos, los cuales tienen una desviación estándar de $ 1,280,236 pesos.

Modelos y resultados

Dentro de los modelos de clasificación se entrenaron 7 modelos: k-vecinos cercanos (1,5,7,10,13 vecinos), Logit, Logit Cross Validation k-fold (CV-K), Lasso CV-K, Ridge CV-K, Elasticnet y Random Forest. Para los modelos que utilizan un punto de corte para la clasificación de sus predicciones, utilizamos un punto de corte bayesiano (0.5) y un punto óptimo obtenido a partir de la curva de ROC. Para la selección del mejor modelo utilizamos el área bajo la curva de ROC (AUC), así como la curva de ROC y la sensibilidad. A continuación, se observa que la mejor curva de ROC se obtiene a partir del entrenamiento de un modelo Random Forest (curva amarilla):

Imagen 1 Curva de ROC



Fuente: elaboración propia datos GEIH 2018

Tabla 3 Sensibilidad y AUC de modelos de clasificación

Modelos	Sensibilidad	AUC
Logit	0.613056945437217	-
Logit-cv	0.816059089826066	0.88028216
Lasso_logit	0.612103883726471	0.88028216
Lasso_logit_Thres	0.806051941863236	-

Ridge_logit	0.612103883726471	0.88028216
Ridge_logit_Thres	0.806051941863236	-
Elastic_Net	0.612103883726471	0.88028216
Elastic_Net_thresh	0.806051941863236	-
Random_Forest	0.868954014772456	0.97462638
Random_Forest_Thresh	0.919704550869669	-
Knn_1	0.55928838058931	-
Knn_5	0.582479548884124	-
Knn_7	0.582320705265666	-
Knn_10	0.580335160034946	-
Knn_13	0.584782781351759	-

Fuente: elaboración propia datos GEIH 2018.

Adicionalmente, se observa en la tabla 3 que el modelo con mejor desempeño, determinado mediante la AUC y la sensibilidad, con un punto de corte optimo de 0.339 para la clasificación de sus predicciones, es el Random Forest, afirmando así, que este es el mejor modelo para la identificación de pobres. Este modelo se entrenó en el 70% de la muestra de entrenamiento utilizando Cross Validation K-fold; en el 20% de la muestra se testeo el desempeño del modelo; y en el 10% restante se evalúa el punto de corte óptimo que maximiza los verdaderos positivos en la clasificación en la curva de ROC. Final mente, se obtiene la siguiente matriz de confusión entre la clasificación observada y la predicción:

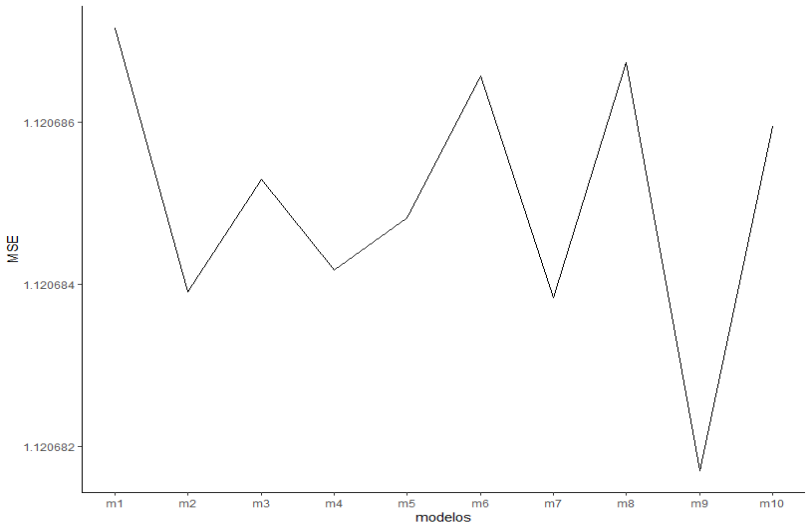
Tabla 4 Matriz de confusión

	Referencia	
Predicción	Pobre	No Pobre
Pobre	7720	1443
No Pobre	674	23153

Fuente: elaboración propia datos GEIH 2018

Al correr el modelo de clasificación en la base de nuevos datos (test) logramos obtener las predicciones, en las cuales el modelo predice que el 33% de los hogares son pobres y el 66% no lo es.

Imagen 2 Errores cuadrático medio



Fuente: Elaboración Propia datos GEIH 2018

Para los modelos de estimación de ingresos se corrieron 10 regresiones en las cuales se iba aumentando su complejidad por medio de la inclusión de más variables y de interacciones en las mismas. Por medio del calculo del MSE haciendo uso de la metodología de Cross Validation K-fold se logra determinar que el mejor modelo es el número 9 en el cual hace uso de la siguiente ecuación:

$$Ingtot_{log} \sim \beta 0 + \beta 1 edad + \beta 2 edad_{sqr} + \beta 3 tiempo_{empresa} + \beta 4 tiempo_{empresa_{sqr}} + \beta 5 mujer + \beta 6 nivel_{educativo} + \beta 7 mujer : nivel\_educativo + \beta 8 actividad + \beta 9 ocupación\_empleo + \beta 10 ReciAyudaInst + \beta 11 ocupacion\_vivienda + \varepsilon$$

Al hacer uso de este modelo en la base de datos nuevos (test) se logra predecir que el 9% de los hogares son pobres y que el 90% no lo son.