Ronny Cruz
Catalina Esmeral Flórez
Federico Ramirez
Juan Camilo Sánchez
Link al Github: https://github.com/Ronnyjohan3121/Problem_Set_2_Group_7
Big Data y Machine Learning

Problem Set II: Predicting Poverty

I. Introducción

La pobreza ha sido una de las principales motivaciones a la hora de diseñar e implementar políticas públicas. Según datos del Banco Mundial (2017) los niveles de pobreza a menudo se estiman a nivel país extrapolando los resultados de las encuestas realizadas en un subconjunto de la población a nivel de hogar o individuo. Esta serie de encuestas tienen como objetivo i) identificar los hogares pobres, ii) describir y analizar la pobreza y iii) predecir los hogares clasificados como pobres cuando no se tiene suficiente información disponible. Por esta razón, es relevante utilizar modelos de predicción adecuados a la hora de diseñar e implementar políticas públicas que mitiguen este problema.

Este documento muestra los resultados de una serie de modelos que buscan clasificar la pobreza y la predicción de los ingresos de los hogares en Colombia. Para ello, se utilizan datos de la GEIH 2018 y se clasifica como pobre a las personas cuyos ingresos totales sean menores a los establecidos por la línea de pobreza (DANE, 2019). La literatura evidencia que existen tres factores que permiten medir de manera adecuada la pobreza: los ingresos per cápita, la línea de pobreza establecida por cada país y los índices de precios de una canasta de bienes básicos con paridad de poder adquisitivo (Banco Mundial, 2017). Entre los datos disponibles, se utilizaron 8 variables que permiten darle consistencia a los modelos. Como principales resultados se obtuvo que el modelo de GBM dio los mejores niveles de predicción de pobreza y la estimación de ingresos tuvo la mejor especificación a partir de un modelo lasso.

II. Data

a. Construcción de la base de datos

Se extrajeron los datos de la GEIH 2018 divididos en dos grupos: hogares e individuos. Como se busca predecir la pobreza a nivel del hogar, se fusionaron las dos bases de datos a través de un único identificador de id. Posteriormente, se creó la variable de ingresos por hogar que, a su vez, actuará como variable dependiente en algunos de los modelos a estimar. Basados en el diccionario de variables escogimos de manera preliminar las adecuadas para las predicciones, estas indican i) si el hogar es pobre o no, ii) cuántos cuartos hay en el hogar, iii) si la vivienda es propia o no, iv) el número de personas que conforman el hogar, v) el número de personas por unidad de gasto, vi) los ingresos totales del hogar, vii) la línea de pobreza establecida por el DANE, viii) sexo, ix) edad, x) si la persona es jefe de hogar, xi) el tipo de régimen de salud, xii) el nivel educativo de los individuos, xiii) la posición del individuo en su trabajo, xiv) si es cotizante a pensión o no, xv) si el individuo está ocupado y xvi) la actividad a la que se dedica el individuo.

Con las variables anteriores se conformó la nueva base de datos, se verificaron *missing values* y se estimaron estadísticas descriptivas relevantes para el análisis de los datos.

Dentro de las limitaciones que se tuvieron con los datos, estuvo no poder utilizar la variable de género para las estimaciones de los modelos de predicción y regresión de ingreso, esto debido a que la variable solo presentaba datos de un tipo de observación. Se cree que el error pudo haber estado en la creación de la variable en la base *train* de hogares desde la base *train* de personas. Esta variable podía aportar datos relevantes sobre la brecha de género en la composición de ingresos de los hogares.

b. Análisis descriptivo

Al revisar la estructura de los datos se decidió analizar la proporción de personas pobres para cada unidad de territorio. Esto es, diferenciar las ratios de pobreza por ubicación geográfica a partir de la variable *ubic* que le asigna valor de cero (0) a las cabeceras municipales y uno (1) al rural disperso y la variable *pobre* que asigna valor de cero (0) a los no pobres y de uno (1) a los pobres. En la tabla a continuación se puede observar que el 86% de la población pobre vive en las cabeceras municipales y el restante 14% en rural disperso. Esta relación tiene sentido, ya que en Colombia más personas viven en los centros urbanos que en zonas rurales.

Tabla 1. Distribución de pobreza por ubicación geográfica.

Characteristic	**N**	**Overall**, N = 131,968[1]	**0**, N = 105,545[1]	**1**, N = 26,423[1]
Ubic	131,968			
1		119,598 (91%)	96,924 (92%)	22,674 (86%)
2		12,370 (9.4%)	8,621 (8.2%)	3,749 (14%)

[1] n (%)

En un análisis similar, se analizó la distribución de pobreza por tipo de vivienda de los hogares, utilizando la variable *pobre* y *tipo_vivienda*, respectivamente. *Tipo_vivienda* toma valor de uno (1) si la vivienda es propia totalmente pagada, dos (2) si es propia, la están pagando, tres (3) si está en arriendo o subarriendo, cuatro (4) si es de usufructo, cinco (5) si es de posesión sin título (ocupante) y seis (6) otra. De acuerdo a esta clasificación y con los datos de la tabla a continuación, se puede concluir que el 44% de la población pobre vive en arriendo y que tan solo el 1,7% de esta vive en casa propia que la estén pagando. La segunda proporción más grande de pobres se encuentra en los hogares de vivienda propia con un 27%. Sin embargo, dentro de la población no pobre, el 40% vive en vivienda propia pagada y el 39% en arriendo. En general, la mayoría de la población vive en arriendo (39%), pero no es muy distante esta proporción de la población que vive en vivienda propia pagada (38%). En Colombia el sector de vivienda es fuertemente apoyado por el gobierno vía subsidios a la construcción y adquisición de viviendas de interés social (VIS), lo que permite explicar la gran proporción de colombianos que viven en vivienda propia.

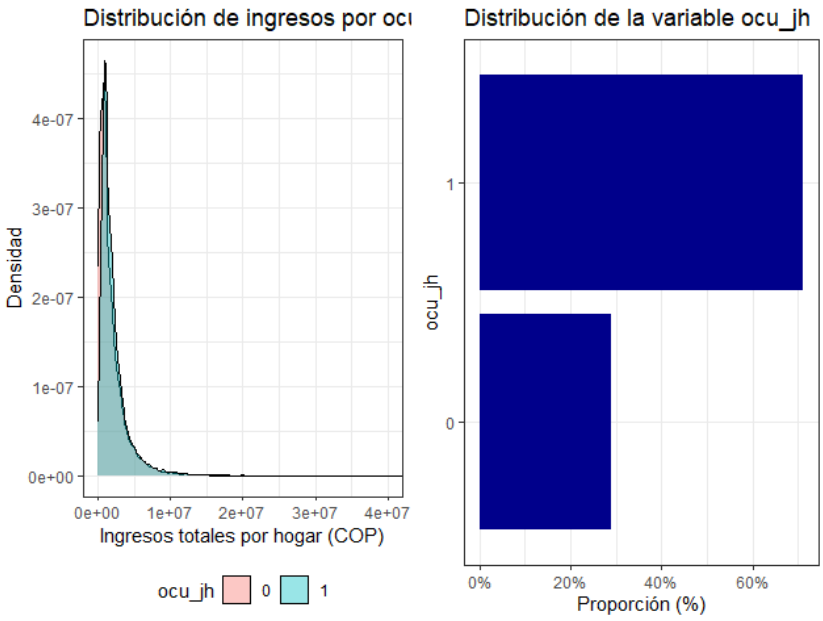
Tabla 2. Distribución de pobreza por tipo de vivienda.

Characteristic	N	Overall, N = 131,968[1]	0, N = 105,545[1]	1, N = 26,423[1]
tipo_vivienda	131,968			
1		49,862 (38%)	42,636 (40%)	7,226 (27%)
2		4,545 (3.4%)	4,103 (3.9%)	442 (1.7%)
3		51,415 (39%)	39,843 (38%)	11,572 (44%)
4		19,940 (15%)	15,756 (15%)	4,184 (16%)
5		6,078 (4.6%)	3,123 (3.0%)	2,955 (11%)
6		128 (<0.1%)	84 (<0.1%)	44 (0.2%)

[1] n (%)

Continuando con el análisis de las variables a utilizar en los modelos de predicción más adelante, se calculó la distribución de ingresos de los hogares por tipo de ocupación de los individuos a partir de la variable *ocu_jh* que toma valores de uno (1) si el individuo tiene trabajo (está ocupado) y de cero (0) si es desocupado. Se utilizó la variable de distribución de ingreso, ya que es un buen indicador de pobreza monetaria. Según el gráfico presentado a continuación, la población ocupada tiene ingresos que van desde valores cercanos a cero hasta 80 millones de pesos colombianos al mes. La población desocupada tiene ingresos más bajos cercanos a cero y es menor a la cantidad de personas ocupadas en el país.

Ilustración 1. Distribución de los ingresos por tipo de ocupación.



Entre otros datos relevantes se encontró que aquellos individuos que tienen niveles de posgrado tienen un nivel de ingreso mayor al del resto de la población y son a su vez la menor proporción de colombianos en la distribución.

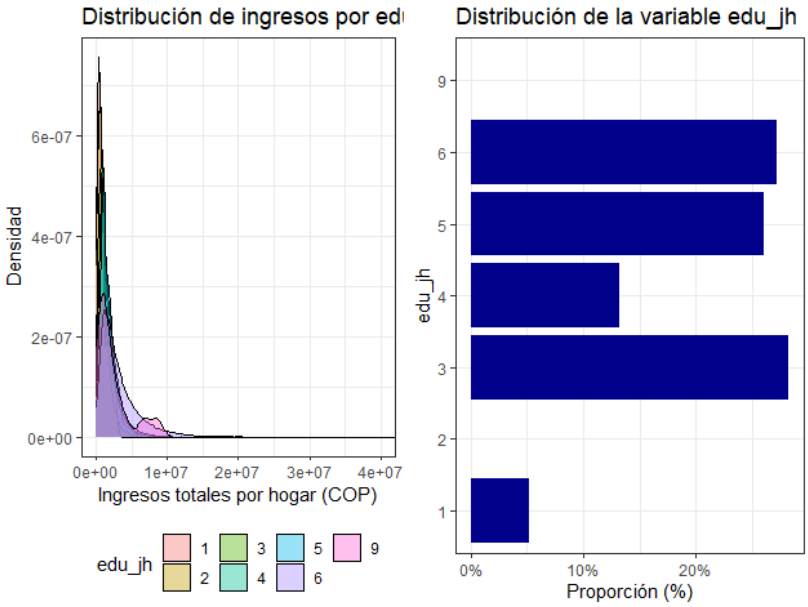


Ilustración 2. Distribución del ingreso por nivel educativo.

III. Modelos y resultados

a. Modelos de clasificación

Ahora bien, para clasificar a la población en pobre y no pobre se utilizó la línea de pobreza establecida por el DANE que establece que se considera pobre a aquellas personas que tienen ingresos inferiores o iguales a 354 mil pesos mensuales. Esta categorización tiene una limitante y es que la mayoría de la población en Colombia se ubica alrededor de esa medida, por lo que cualquier corrección en este parámetro de clasificación afectaría de manera importante la asignación para cada individuo y hogar.

Para predecir la pobreza, se testearon los datos con 4 modelos logit, 2 probit, un árbol de decisión, un bosque aleatorio y un modelo *grading boosting machine.* Los resultados de estas predicciones se muestran a continuación:

Tabla 3. Comparación de las predicciones de los modelos.

Modelo	Accuracy	Sensitivity	Specificity
logit2	0.8277	0.63242	0.85049
logit3	0.8245	0.62577	0.84609
logit4	0.8274	0.63092	0.85040
probit1	0.8243	0.62988	0.84448
probit2	0.828	0.63934	0.84923
arbol1	0.8199	0.67423	0.82880
forest1	0.8204	0.72139	0.82526
gbm1	0.8297	0.6372	0.8532

De las predicciones anteriores, se obtuvo que el mejor modelo para predecir la probabilidad de que un individuo sea pobre, con las especificaciones escogidas y los datos disponibles, fue el GBM con un *accuracy* de 0,8297. Esta medida de ajuste es apropiada, ya que según la literatura se debe procurar que sea mayor a 0,5 pues este es el ajuste para un evento completamente aleatorio. El modelo de GBM presentó la mejor predicción después de 150 árboles o iteraciones y por lo general, estos modelos suelen predecir mejor los resultados, ya que utilizan árboles de decisión débiles para crear un modelo más fuerte que sea capaz de hacer predicciones más precisas.

b. Regresiones de ingreso

Con el objetivo de predecir los ingresos de los hogares, se elaboraron 5 modelos de regresión uno por MCO, otro por lasso, uno de ridge, uno de forward y uno de backward. A continuación, se presentan los resultados de los modelos estimados:

Tabla 4. Comparación de los errores de los modelos

Modelo	RMSE	Rsquared	MAE
ols	2171063	0.2617702	1165625
lasso	2169580	0.2622595	1165001
ridge	2173772	0.2602621	1160080
forward	2232456	0.2193596	1218035
backward	2236249	0.2164671	1192307

El modelo de regresión elegido fue el lasso y tiene la siguiente estructura:

$$Ingtotug = \beta 0 + \beta 1\_\text{Ubic} + \beta 2\_Personas\_h + \beta 3\_\text{tipo\_vivienda} + \beta 4\_Npersug + \beta 5\_jh\_\_edad + \beta 6\_jh\_\_salud + \beta 7\_edu\_\text{jh} + \beta 8\_ocu\_\_id\ \_$$

En este modelo lasso de ingresos totales por unidad de gasto se utilizaron las siguientes variables: la ubicación de los hogares, ya que está relacionada con su nivel de pobreza y se perciben menores ingresos en las zonas rurales y dispersas; el número de personas en el hogar (los hogares pobres tienen, en promedio, más personas y es presumible que no todas las personas tengan ingreso); el nivel de educación del jefe del hogar, ya que los trabajos no calificados con pocos estudios a los que generalmente acceden quienes tienen nivel bajo de educación, tienden a tener menores salarios; que el jefe de hogar se encuentre desocupado, ya que dentro de la unidad de gasto se cuenta como un ingreso potencial menos, así hayan otras personas trabajando; el régimen de salud, ya que las personas de menores ingresos tienden a estar en subsidiado). El modelo se eligió tras revisar sus estadísticos y ver que sus errores eran menores que los otros modelos estimados y su R-cuadrado era ligeramente mayor.

IV. Conclusiones y recomendaciones

Debido a que la distribución de los ingresos en Colombia está agrupada hacia el lado izquierdo de la curva, establecer una línea de pobreza como indicador para la clasificación puede llevar a una pérdida de precisión en el objetivo de identificar correctamente todos los hogares pobres que existen en el país. En este sentido, un cambio marginal en la línea de pobreza hacia la izquierda o hacia la derecha puede recaer en una cantidad significativa de hogares determinados (o no) como pobres. Al respecto, se deberían diseñar otras formas de clasificación que mitiguen ese sesgo y que permitan el diseño de mejores políticas públicas que propendan la reducción de los niveles de pobreza en el país.

Por otro lado, en el análisis hecho se puede concluir que los modelos son bastante sensibles a las variables que se incluyan o excluyan, así como la forma funcional y el método de estimación que se utilice. Es por ello que previo al diseño de modelos, las revisiones de literatura pueden ser útiles para tener noción teórica de las variables que se vayan a considerar, en especial en situaciones donde hay limitaciones en acceso a información.