PROBLEM SET 2: PREDICTING POVERTY

Big data y machine learning for applied economics

Miguel Fernando Contreras Rivera. Código: 202116189

1. Introducción

Este documento muestra los resultados de una serie de modelos para la clasificación de 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 son menores a un monto establecido (línea de pobreza) (DANE, 2019). Se considera que los datos de la encuesta son idóneos para los fines del estudio dado que contienen gran cantidad de información que permite caracterizar adecuadamente a los individuos y sirven de insumos para predicción. La literatura muestra que hay tres factores influyentes para medir la pobreza: los ingresos per cápita, la línea de pobreza propia de cada país y los índices de precios de una canasta de bienes básicos con paridad de poder adquisitivo (World Bank, 2017). Entre los datos disponibles, se utilizaron 8 variables consideradas relevantes para la construcción de los modelos. Como principales resultados se obtuvo que, de los modelos probados (6 de clasificación y 6 de regresión de ingresos), los que mostraron mejor ajuste y nivel óptimo de complejidad fueron el modelo logit1 (clasificación) y el modelo lasso (regresión de ingresos). Esto sugiere que las variables pueden ser determinantes a la hora de clasificar y predecir. En general, el modelo que mejor logró identificar correctamente a los pobres (82,8 %).

2. Datos

Se utilizaron las bases de datos test_personas, test_hogaers, train_personas y train_hogares. En total se eligieron las siguientes variables para las "X" de los modelos: Clase, P5090, Nper, Npersug, Lp, P6020, P6040, P6050, P6100, P6210, P6430, y Oc. Tras examinar las variables se observaron missing values en P6210 (nivel educativo) y en P6240 (actividad en que se ocupó) tanto en test_personas como en training_personas, sin embargo, dado que estos correspondían a menores de edad casi en su totalidad, se los asignó a "ningún nivel educativo" y a "otra actividad" respectivamente. Los datos se describen a continuación.

Gráfico 1: Resumen de variables – base train_hogares unificada

```
      id              Clase           P5010        P5090           Nper            Npersug          Ingtotug       
 Length:131968      1:119612   Min.   : 1.000   1:49626   Min.   : 1.000   Min.   : 1.000   Min.   :      0  
 Class :character   2: 12356   1st Qu.: 1.000   2: 4520   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:  800000  
 Mode  :character              Median : 2.000   3:51614   Median : 3.000   Median : 3.000   Median : 1400000  
                               Mean   : 1.994   4:19914   Mean   : 3.295   Mean   : 3.282   Mean   : 2096564  
                               3rd Qu.: 3.000   5: 6170   3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 2506926  
                               Max.   :15.000   6:  124   Max.   :22.000   Max.   :22.000   Max.   :85833333  

 Ingtotugarr          Ingpcug             Lp          Pobre       Ingtot_hogar         personas_h    
 Min.   :       0   Min.   :      0   Min.   :167222   0:105576   Min.   :       0   Min.   : 0.200  
 1st Qu.:  900000   1st Qu.:  300000   1st Qu.:275594   1: 26392   1st Qu.:  800000   1st Qu.: 1.000  
 Median : 1586667   Median :  545672   Median :280029              Median : 1400000   Median : 1.500  
 Mean   : 2313750   Mean   :  872398   Mean   :271557              Mean   : 2108550   Mean   : 1.726  
 3rd Qu.: 2798119   3rd Qu.:  988233   3rd Qu.:285650              3rd Qu.: 2524727   3rd Qu.: 2.000  
 Max.   :88833333   Max.   :88833333   Max.   :303817              Max.   :85833333   Max.   :14.000  

 mujer_jh     edad_jh         edu_jh      salud_jh  trabajo_ocu_jh  pension_jh ocu_jh       log_ingh      
 0:54861   Min.   : 11.0   1: 6853   0: 7837   4      :46638   0:38020    0:38018   Min.   : 0.00  
 1:77107   1st Qu.: 37.0   2:    9   1:63761   0      :38018   1:36923    1:93950   1st Qu.:13.59  
           Median : 49.0   3:37220   2: 7454   1      :31345   2:54255              Median :14.15  
           Mean   : 49.6   4:17324   3:52876   2      : 6228   3: 2770              Mean   :14.02  
           3rd Qu.: 61.0   5:34270   9:   40   5      : 5144                        3rd Qu.:14.73  
           Max.   :108.0   6:36272             3      : 2307                        Max.   :18.27  
                           9:   20             (Other): 2288                                       
   pred_log1           pred_log2           pred_log3           pred_log4          pred_pro1        
 Min.   :0.0003971   Min.   :0.0009549   Min.   :0.003748   Min.   :0.003302   Min.   :0.0008257  
 1st Qu.:0.0436741   1st Qu.:0.0415584   1st Qu.:0.044515   1st Qu.:0.042071   1st Qu.:0.0405583  
 Median :0.1154900   Median :0.1155336   Median :0.115591   Median :0.113802   Median :0.1205949  
 Mean   :0.1995766   Mean   :0.1996787   Mean   :0.199679   Mean   :0.199662   Mean   :0.1989330  
 3rd Qu.:0.3040950   3rd Qu.:0.3038145   3rd Qu.:0.304989   3rd Qu.:0.303057   3rd Qu.:0.3111059  
 Max.   :0.9997017   Max.   :0.9997313   Max.   :0.999646   Max.   :0.999745   Max.   :0.9999978  

   pred_pro2         holdout       
 Min.   :0.000641   Mode :logical  
 1st Qu.:0.037767   FALSE:131968   
 Median :0.118673                  
 Mean   :0.198697                  
 3rd Qu.:0.309277                  
 Max.   :0.999999
```

3. Modelos y resultados

3.1. Modelos de clasificación

Con el objetivo de predecir Pobres (1) y No Pobres (0) se elaboraron 6 modelos de clasificación: logit1. logit2. logit3. logit4. probit1 y probit2. Cada uno utilizó distintas combinaciones de las variables seleccionadas.

- **Modelo de clasificación elegido (modelo logit1)**

El modelo de clasificación elegido tiene la siguiente estructura:

$$Pobre = \hat{\beta}_0 + \hat{\beta}_1 P5090 + \hat{\beta}_2 Personas_h + \hat{\beta}_3 jh_{mujer} + \hat{\beta}_5 jh_{edu} + \hat{\beta}_6 jh_{ocu} + \hat{\beta}_7 ocu_{id} + \hat{\beta}_8 jh_{edad} + \hat{\beta}_9 jh_{salud} + \hat{\beta}_{10} jh_{trabajo_{ocu}} + \hat{\beta}_{11} jh_{pension}$$

Dado que en este punto se tenía en cuenta principalmente la capacidad que tenía el modelo a la hora de predecir adecuadamente los hogares pobres, el criterio de sensibilidad es entonces el que cobra mayor relevancia. El modelo de clasificación con mejor poder predictivo logró clasificar correctamente al 82,8 % de los pobres. Esto implica que, cuando el modelo predice que un hogar es pobre, acierta en este porcentaje.

La justificación de la elección de las variables es que, al tratar de clasificar a la población pobre es necesario conocer las condiciones de su entorno que pueden determinar dicho estado como, por ejemplo: que deban pagar arriendo frente a tener vivienda propia (ello limitaría el presupuesto de la unidad de gasto, además en la revisión inicial de datos se evidenció que la mayoría de pobres viven en arriendo. También se tiene en cuenta que, si una unidad de gasto logró comprar una vivienda, es presumible que sus ingresos sean mayores para haberla podido pagar, por lo que no es más probable que se encuentre por encima de la línea de pobreza); que el hogar tenga muchas personas (requerirían mayores ingresos para suplir sus necesidades y superar la línea de pobreza); que el jefe de hogar sea mujer (históricamente se ha observado que tienden a haber más familias pobres con madres cabeza de familia); que el jefe de hogar tenga un bajo nivel de educación (probablemente no pueda estudiar por su condición de pobreza); y que el jefe de hogar se encuentre desocupado (no tendría un ingreso fijo); que hayan muchas personas por habitación (esto indicaría hacinamiento, el cual es característicos de los hogares pobres); que estén en un determinado régimen de salud (pues los hogares pobres tienden a estar en subsidiado); que se encuentren ocupados y cuál sea el tipo de ocupación (se observó en la revisión previa que las personas en cargos de obrero tienen menores ingresos, lo que los pone por debajo de la línea de pobreza).

A continuación, se observa un cuadro comparativo de 6 de los modelos evaluados:

Tabla 1. Otros modelos de clasificación

Modelo	Sensibilidad	Especificidad	Precisión	TN	TP	FN	FP
logit1	0.63741	0.85054	0.828	25096	2222	1264	4410
logit2	0.63104	0.85231	0.8279	25017	2297	1343	4335
logit3	0.62511	0.84849	0.8251	25067	2156	1293	4476
logit4	0.62781	0.85201	0.8272	25003	2289	1357	4343
probit1	0.63363	0.85017	0.8273	25081	2212	1279	4420
probit2	0.63363	0.85017	0.8273	25081	2212	1279	4420

3.2. Modelos de regresión de ingresos

Con el objetivo de predecir los ingresos, se elaboraron 6 modelos de regresión: ols. ols1. lasso. lasso1. ridge. ridge1.

- **Modelo de regresión de ingresos elegido (modelo lasso)**

El modelo de regresión elegido tiene la siguiente estructura:

$$Ingtotug = \hat{\beta}_0 + \hat{\beta}_1 Clase + \hat{\beta}_2 Personas\_h + \hat{\beta}_3 Npersug + \hat{\beta}_4 jh\_mujer + \hat{\beta}_5 jh\_edad + \hat{\beta}_6 jh\_salud + \hat{\beta}_7 jh\_edu + \hat{\beta}_8 ocu\_id$$

En este modelo lasso de ingresos totales por unidad de gasto el MAE fue de 1166245. Entre las justificaciones de la elección de las variables se encuentra que: estar ubicado en zonas rurales puede ser indicativo de menores ingresos (suelen tener salarios más bajos, además, la revisión de datos inicial mostró que el porcentaje de pobres rurales era casi el doble frente al dato de cabeceras); 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); que el jefe de hogar sea mujer (suele haber una brecha de ingresos negativa para las mujeres); bajo nivel de educación del jefe de hogar (los trabajos no calificados con pocos estudios a los que generalmente acceden quienes tienen nivel bajo de educación, tienden a tener menores salarios); el jefe de hogar se encuentre desocupado (dentro de la unidad de gasto se cuenta como un ingreso potencial menos, así hayan otros trabajando); el régimen de salud (revisión previa mostró que los de menores ingresos tienden a estar en subsidiado).

A continuación, se observa un cuadro comparativo de 6 de los modelos evaluados:

Tabla 2. Otros modelos de regresión

Modelo	RMSE	RSQUARED	MAE
ols	2168636	0.2575818	1166884
ols1	2278202	0.1803901	1297821
lasso	2167996	0.2578326	1166245
lasso1	2276542	0.1810331	1297322
ridge	2168812	0.2571984	1161187
ridge1	2279079	0.180104	1292623

4. Conclusiones y Recomendaciones.

Tras el ejercicio llevado a cabo, inicialmente se pudo observar que, en los modelos de clasificación, no necesariamente la elección de mayor cantidad de variables genera resultados más acertados. Previo a las regresiones aquí presentadas se corrieron modelos con 21 variables que aparecían en ambas bases. También se destaca la importancia de hacer primero un lasso para determinar las variables que se deberían utilizar en el modelo, pues con ello se reduce la búsqueda.

Otro punto para resaltar en este tipo de ejercicios es la necesidad de hacer un buen análisis de la literatura relevante, previo a la elaboración de los modelos, pues con ello se pueden alivianar las cargas en materia de cantidad de datos, al dar la posibilidad de identificar las variables más significativas, y gracias a ello, poder trabajar con bases más pequeñas que demanden menor capacidad en cuanto a las especificaciones técnicas de los equipos utilizados para el procesamiento de datos.

Referencias

DANE. (2019, julio). *Archivo Nacional de Datos.* Retrieved from http://microdatos.dane.gov.co/index.php/catalog/608/study-description

World Bank. (2017). *Monitoring Global Poverty: Report of the Commission on Global Poverty.* Retrieved from worldbank.org: https://documents1.worldbank.org/curated/en/353781479304286720/pdf/110040-REVISED-PUBLIC.pdf

Anexos

Gráfico 1. Npersug Pobres y No Pobres – Train Hogares

Characteristic	N	Overall, N = 542,941[1]	0, N = 406,473[1]	1, N = 136,468[1]
Npersug	542,941	4.00 (3.00, 5.00)	4.00 (3.00, 5.00)	5.00 (4.00, 6.00)

[1] Median (IQR)

Gráfico 2. Vivienda Propia Pobres y No Pobres – Train Hogares

Characteristic	N	Overall, N = 542,941[1]	0, N = 406,473[1]	1, N = 136,468[1]
viviendapropia	542,941			
0		311,667 (57%)	217,456 (53%)	94,211 (69%)
1		231,274 (43%)	189,017 (47%)	42,257 (31%)

[1] n (%)

Gráfico 3. Pobres y No Pobres en Cabecera y Otras áreas – Train Hogares

Characteristic	N	Overall, N = 542,941[1]	0, N = 406,473[1]	1, N = 136,468[1]
Clase	542,941			
1		491,173 (90%)	373,714 (92%)	117,459 (86%)
2		51,768 (9.5%)	32,759 (8.1%)	19,009 (14%)

[1] n (%)

Gráfico 4. Ingresos por Unidad de Gasto – Train Hogares

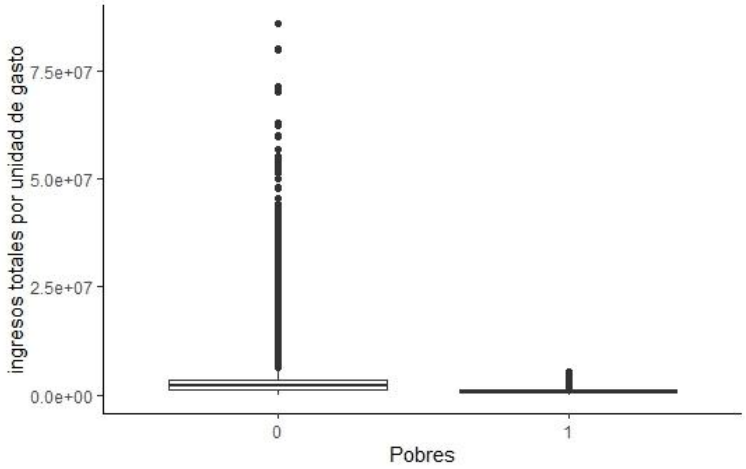


Gráfico 5. Distribución de Pobres – Train Hogares

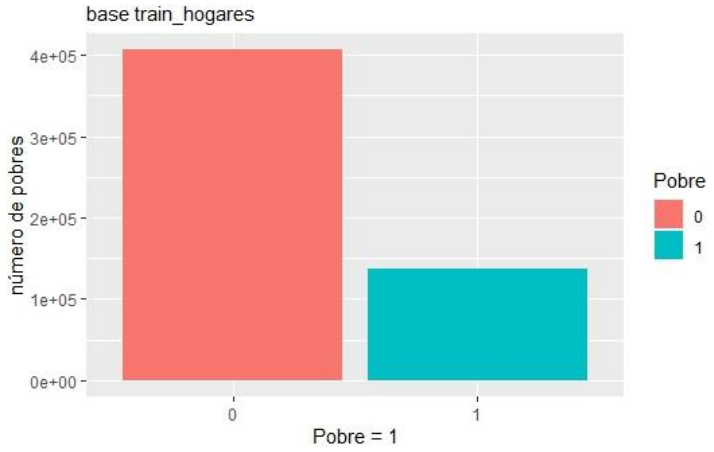


Gráfico 6. Gráfico de cajas del modelo seleccionado (logit1)

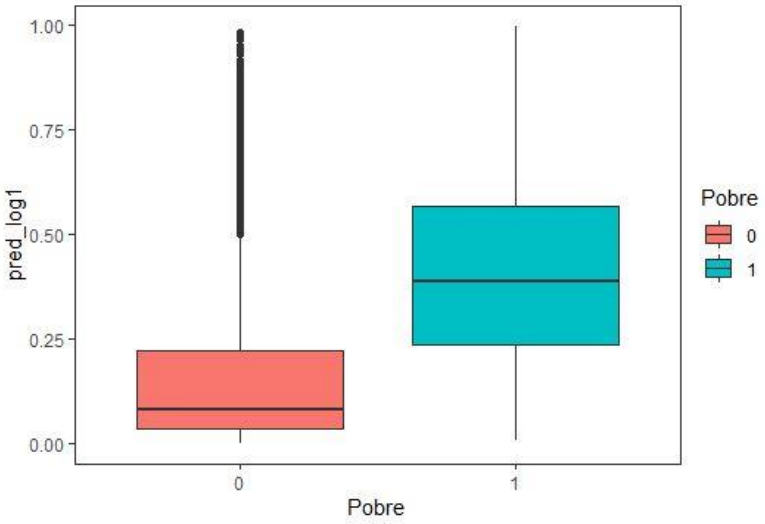
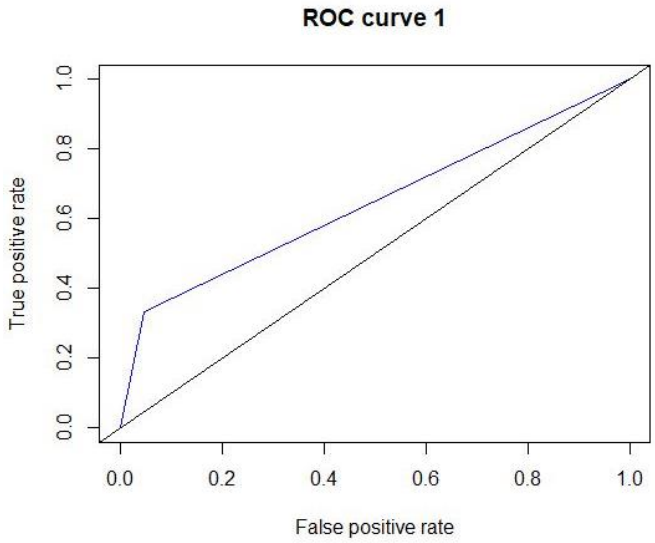


Gráfico 7. Curva ROC del modelo seleccionado (logit1)



Gráfica 8: Distribución-intotug

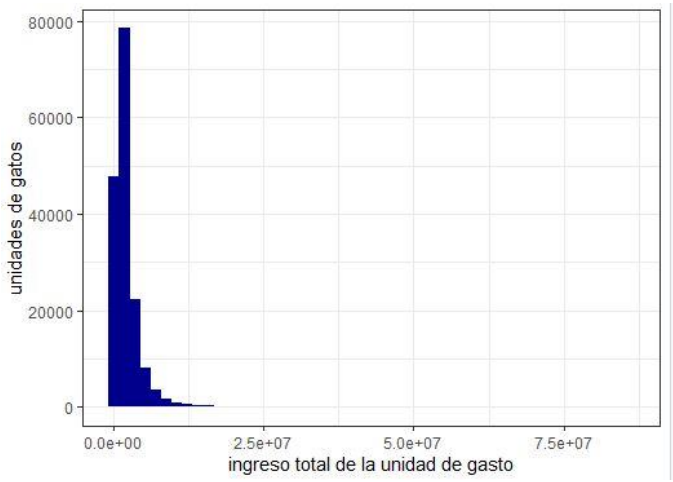


Gráfico 9: Jefe de hogar por sexo

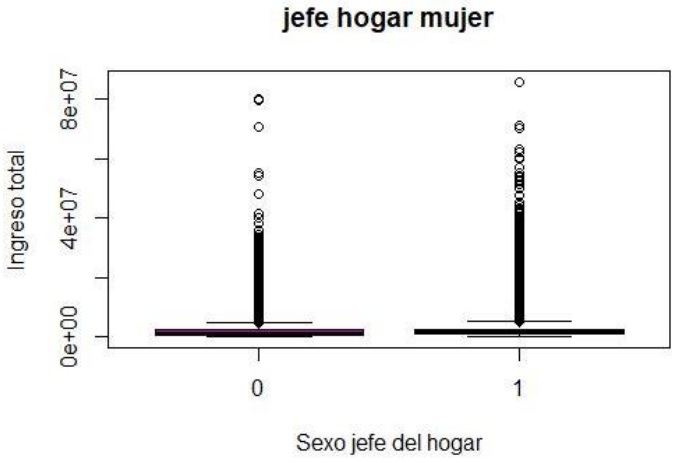


Gráfico 10: Ingresos por clase (Cabecera vs. otros)

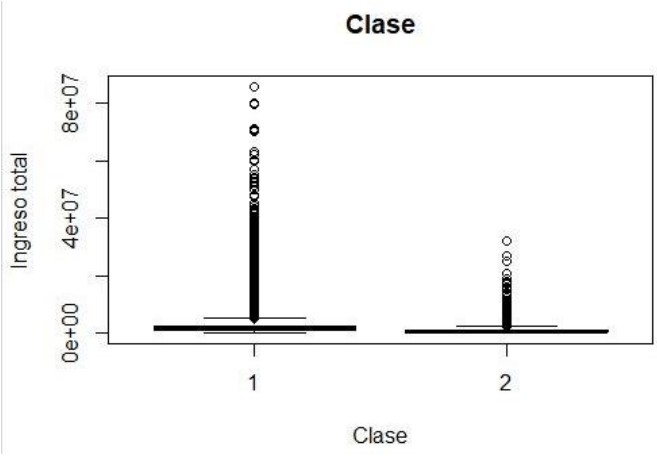


Gráfico 11: Máximo nivel de educación del jefe de hogar

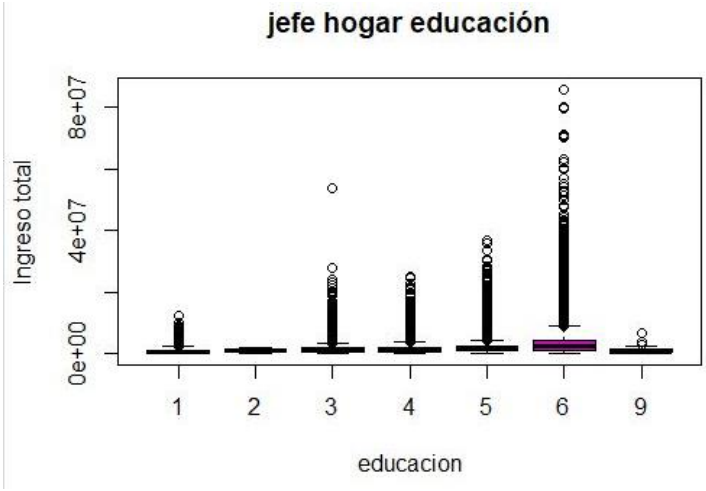


Gráfico 12: Coeficientes de modelo de regresión elegido