Big Data: Problem Set 2

Paolo Valcarcel Pineda[1], Camila Ciurlo Aragon[2]
p.valcarcel@uniandes.edu.co[1], c.ciurlo@uniandes.edu.co[2]
Código: 202010427[1], 202214407[2]
Universidad de los Andes
GitHub: https://github.com/CamilaCiurlo/Pobreza

13 de julio de 2022

1. Introducción

La medición de la pobreza es una de las tareas más arduas que hoy tiene investigadores y gobiernos, sobre todo en escenarios en los que se cuenta con pocos recursos y capacidad de levantar información estadística de la población; como puede ser el caso de países en vía de desarrollo. Sin embargo, este tipo de mediciones son esenciales para el diseño de políticas públicas cada vez más ajustadas a la realidad de los individuos y sus necesidades.

Por esa razón, hoy, las ciencias sociales se encaminan progresivamente a la construcción de modelos que puedan explicar este tipo de fenómeno de manera más acertada, sin que eso conlleve grandes inversiones. Este documento es una aproximación a un modelo de predicción de pobreza por dos vías, a partir de la construcción del ingreso de los hogares y a través de un ejercicio de clasificación. Para evaluar los modelos se estimó su capacidad predictiva fuera de muestra.

2. Base de datos

De acuerdo con el objetivo del estudio se realizó una selección y limpieza de la base de datos con la que se trabajó. En primer lugar, se redujo la base a 13 variables fundamentales que, por un lado, son un identificador de los hogares y, por el otro, pueden explicar el nivel de ingresos (modelo de ingresos) al interior de un hogar, o la probabilidad del mismo de ser pobre (modelo de clasificación).

La variable dependiente elegida para aproximar el ingreso del hogar fue Ingpcug, la cual representa el Ingreso per cápita de la unidad de gasto con

imputación de arriendo, mientras que la dependiente del modelo de clasificación fue la variable Pobreza.

Respecto a las independientes de los modelos estimados, se generaron variables como años de escolaridad (Esc), que tiene un efecto negativo sobre la pobreza, número de personas que conforman la unidad de gasto (Npersug) que tiene un efecto directo sobre la manera en que se mide la pobreza. [1]

Otras variables importantes son, la edad y su componente cuadrático que captura la no linealidad del los ingresos por edad, el sexo (diferenciación de género), jefatura del hogar, vivienda propia y el número de habitaciones que tienen los hogares.

3. Modelo de clasificación

El modelo empleado para generar la clasificación, proviene de una adaptación de las variables empleadas en el documento de Nuñez y Ramirez (2002), el cual tiene la siguiente forma:

$$Pobre_i = \beta_0 + \beta_1 Nper_i + \beta_2 Sexo_i + \beta_3 Jefe_i + \beta_4 Vivienda_i + \beta_5 Edad_i + \beta_6 Edad2_i + \beta_7 Esc_i + \beta_8 Habit_i + \epsilon_i$$

Para poder ralizar las verificaciones pertinentes, se realizó una partición de la base Train de hogares y personas en tres submuestras: train, test y base de evaluación.

Una vez realizada esa partición, se realizaron los siguientes métodos: Logit, Ridge, Lasso, Lasso con punto de corte, Ridge con punto de corte, logit upsampling, logit downsampling, ridge upsampling y ridge downsampling. De estos modelos se obtuvieron diversos hyper par´metros, en ese sentido, se obtuvo un lambda de 0.006056507 para el modelo Logit Lasso (alpha=0), un lambda de 0.006056507 para el modelo Logit Lasso con punto de corte (alpha=0) y para el modelo Ridge y Ridge con punto de corte un lambda de 0.0001 (alpha=1).

A continuación se presentan las matrices de confusion de los modelos estimados:

[1] Un hogar es pobre si el ingreso per cápita de los integrantes es menor a un umbral mínimo que representa una canasta básica.

Figura 1: Matrices de confusión

Logit	No	Si
No	4558	76819
Si	7452	19886

Lasso	No	Si
No	3789	77588
Si	6586	20752

Lasso Thresh	No	Si
No	24682	56695
Si	18871	8467

Lasso Up	No	Si
No	25599	55778
Si	17912	9426

Lasso Down	No	Si
No	25604	55773
Si	17917	9421

Ridge	No	Si
No	4503	76874
Si	7409	19929

Ridge Thresh	No	Si
No	26161	55216
Si	19492	7846

Ridge Up	No	Si
No	26349	55028
Si	18217	9121

Ridge Down	No	Si
No	26379	54998
Si	18224	9114

Como podemos observar, los modelos tienen clasificaciones muy diferentes. No obstante, se sabe de la importancia de predecir correctamente a los pobres que son pobres y a los no pobres que son no pobres, sin embargo, los datos no se llegan a ajustar correctamente a la proporción que manejaba la base Train de pobre y no pobres (80/20), siendo que el porcentaje de pobres es mucho menor que el de no pobres y que al momento de hacer la predicción en la base test, el nivel de pobres era muy cercano al 90 %, lo cual no cuadra con el número de pobres hayado en el modelo de ingreso.

Es en función a este criterio que se escogió el modelo Ridge con punto de corte, el cual en las evaluaciones nos arrojó un valor de 0.7559402. Una vez enviada la predicción a la base Test, se procedió a clasificar a los pobres. Se tuvo el inconveniente de que existia una predicción por cada individuo por hogar, esto es, podian haber individuos pobres y no pobres dentro de un hogar. Para poder determinar si el hogar era pobre o no, se consideró el estado de pobreza del jefe de hogar y con esto realizamos la categorización por hogar final.

4. Modelo de ingreso

Una vez elegida la variable a explicar se construyeron dos modelos de regresión lineal múltiple.

$$Ingpercapita_i = \beta_0 + \beta_1 Npersug_i + \beta_2 Sexo_i + \beta_3 Jefe_i + \beta_4 Vivienda_i + \beta_5 Edad_i + \beta_6 Edad2_i + \beta_7 Esc_i + \beta_8 Habit_i + \epsilon_i$$

$$Ingpercapita_i = \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Edad2_i + \beta_4 Esc_i + \epsilon_i$$

Todas las variables, en los dos modelos, tuvieron signos intuitivos y fueron significativas al 99 %. Ahora bien, el R2 ajustado del modelo 2 explicó poco más del 14 % de la variabilidad de la variable dependiente y el segundo el 21 %. Pero el interés de este documento es estimar la capacidad predictiva de los modelos, no cuánto se ajustan a la muestra. En ese sentido, se dividió la muestra en una base Train y una Test y, una vez diseñados los modelos en Train, se corrieron predicciones de ingreso fuera de muestra, en el Test. Los resultados muestran, para los dos casos, que el error cuadrático medio (MSE) fuera de muestra es menor que dentro de muestra y eso, sin duda, es una buena señal para el desarrollo del ejercicio. El error cuadrático medio del primer modelo fuera de muestra fue menor que el primero, así que se eligió para el desarrollo de siguientes pruebas (Figura 1).

Figura 2:

	Log MSE In	Log MSE Out
Modelo 1	27.31	27.25
Modelo 2	27.40	27.34

Ahora bien, como ya se mencionó líneas atrás, para evitar el sobreajuste del modelo sobre la muestra de entrenamiento se realizaron tres modelos de regularización. Se hizo una nueva partición de la muestra dada, en una base de entrenamiento, una de testeo y otra más de evaluación. Se diseñaron tres métodos Ridge, Lasso y Elastic Net que, si bien aumentan el sesgo, hacen que los coeficientes tiendan (Ridge) o sean cero (Lasso), con eso se reduce la varianza, se atenúa el efecto de la correlación entre predictores y se reduce la influencia de predictores menos relevantes.

Para el caso del Modelo Ridge, se establecieron los hiperparámetros Alpha=0 y Lamda=100. A través del método de validación cruzada se estableció el rango del Lamda óptimo (Gráfico 1) y, se optó por tomar el Lamda mínimo dentro del intervalo para hacer la predicción. Una vez hechas las predicciones de la variable Y (Ingpecug) fuera de muestra, se calculó el error cuadrático medio del modelo fuera de muestra. Para el caso del Modelo Lasso, se establecieron los hiperparámetros Alpha=1 y Lamda=100. Al correr el modelo, y dado que Lasso penaliza el uso de variables no relevantes al modelo, los Lamdas se redujeron a 77. Se estableció el rango del Lamda óptimo (Gráfico 2) y se eligió el Lamda mínimo para la predicción fuera de

muestra. El cálculo del error cuadrático medio (MSE) fuera de muestra para Ridge y Lasso, mostró que predecía mejor el Lasso (Figura 2).

Figura 3:

	Log MSE In
Ridge	27.26
Lasso	27.25
E Net	27.25

Ahora bien, dado que el Lasso dio mejores resultados, para el caso del Modelo Elastic Net se eligió un Alpha más cercano a 1. Se establecieron los hiperparámetros Alpha=0,7 y Lamda=100. Al igual que en los otros dos casos, se estableció el rango del Lamda óptimo y se eligió el Lamda mínimo dentro del intervalo para la predicción fuera de muestra. Una vez hecho el cálculo de la predicción, el MSE fuera de muestra mostró que el modelo Lasso seguía siento ligeramente mejor que los otros dos.

Una vez elegido el modelo óptimo (Lasso), se pasó a predecir el ingreso per cápita por hogar de la base Test que no contaba con observaciones de ingresos. Una vez más se estableció un modelo Lasso, se estableció el Lamda mínimo dentro del intervalo óptimo y se evaluó la predicción de ingresos por cada persona del hogar. Con eso, se pudieron sumar los ingresos de los individuos del hogar y dividirlos por el número de integrantes. De acuerdo con la predicción de ingresos, de la base Test de 66.168 hogares, 61.879 no son pobres y 4.289 si lo son. Es decir que cerca del 84 % de hogares están por encima de la línea de pobreza y el 6 % restante está por debajo.