PROBLEM SET 2- PREDICTING POVERTY

Camila Alejandra Velasco Contreras

Jairo Alexander Torres Preciado

Enlace repositorio: camivel/ProblemSet2 (github.com)

1. Introducción

En décadas recientes, las metodologías de aprendizaje de máquinas se han usado ampliamente en diversas aplicaciones empíricas. En asuntos sociales y, en particular, en temas de pobreza, también se han implementado modelos que permiten predecirla con bastante precisión. En el contexto latinoamericano se destacan trabajos sobre pobreza en Ecuador (Ochoa M. et al, 2021), Argentina (Chagalj C., 2019; Dabús A., 2020) y Costa Rica (Kim J., 2021). En el caso colombiano, Sabogal H. et al. (2021) incorporan métodos de aprendizaje de máquinas al análisis de la pobreza multidimensional entre 2016 y 2019, con el fin de entender mejor cada una de las dimensiones que componen esta medición.

En el presente ejercicio se utilizaron datasets de la Gran Encuesta Integrada de Hogares (2018) para predecir la pobreza monetaria en Colombia a partir de dos enfoques. El primero consiste en implementar modelos de clasificación de los hogares y, el segundo, se basa en la predicción de ingreso para luego clasificar los hogares usando la línea de pobreza. En ambos casos se estimaron diversos modelos y se encontró que el modelo que mejor predice en el ejercicio de clasificación es el árbol con variables seleccionadas, hiperparámetros ajustados y muestra balanceada (upsample), mientras que en la predicción de ingreso el mejor fue la regresión lineal por mínimos cuadrados ordinarios.

2. Datos

El objetivo de este ejercicio es predecir si un hogar es pobre o no de acuerdo a las métricas que usa el DANE, para ello contamos con dos bases de datos (*train_hogares* y *train_personas*), una a nivel individuo y otra a nivel hogar. Estas bases contienen las variables que queremos predecir, una variable categórica que toma el valor de 1 si el hogar es pobre y cero de lo contrario y el ingreso pércapita de la unidad de gasto con imputación de arriendo (*Ingpcug*); estas bases las usamos para entrenas los distintos modelos de predicción. Dado que las bases sobre las que queremos hacer nuestra predicción (test) no tienen las mismas variables que las bases de training el primer paso fue identificar aquellas variables independientes que tienen en común las bases *traing_hogares* y *test_hogares*, así como las comunes entre *train_personas* y *test_personas*. Nuevamente, esto con el objetivo de no entrenar nuestros modelos con características con las que no vamos a contar a la hora de hacer la predicción. Para esto creamos dos listas, una a nivel hogar y otra a nivel personas, con las variables en común entre las bases ya mencionadas. Borramos todas aquellas que no se encuentran en las listas.

Una vez hecho esto nos concentramos en la limpieza de la base a nivel individuo con el objetivo de extraer información valiosa a nivel hogar que complemente la que ya tenemos en la otra base, dado que la predicción se hará a nivel hogar y no individuo. Juntamos las bases de *test* y *training* para que todas las transformaciones que se realicen queden en ambas bases, generamos un identificador para posteriormente volverlas a separar. En esta nueva base nos quedamos con las variables que consideramos relevantes y con ellas creamos ocho variables nuevas a nivel hogar: sexo del jefe de hogar (*sexojefe*), edad del jefe de hogar (*edad jefe)*, nivel educativo del jefe de hogar (*niveleduc_jefe)*,

número de menores de edad en el hogar (*No_menores*), número de adultos mayores de 60 en el hogar (*No_mayores)*, una dummy que indica si el jefe de hogar es informal o formal (*informal)*, una dummy de si alguno de las personas en el hogar recibe ingresos por dividendos (*otro_dividendo)*, una dummy indicativa de si alguien en el hogar recibe ingresos por pensiones o arriendos (*otro_pens)*. Nos quedamos con estas variables y nos quedamos solo con una unidad por hogar.

Antes de separar las bases, designamos como factor a todas las variables que los son, con sus respectivas etiquetas. Dividimos las bases de nuevo en *train* y *test* y hacemos un merge con las respectivas bases a nivel hogar. Quedamos sólo con una base a nivel hogar para training y otra en la que haremos la predicción. Cambiamos nombres de variables y nos aseguramos nuevamente de que todas las variables categóricas estén designadas como factor. Creamos una nueva variable llamada *arriedo* que indica el valor que paga de arriendo la familia, y en el caso de no vivir en arriendo el estimado de cuanto pagarían.

Procedemos a analizar los *missing values* de nuestra base train final. Las variable *informal* tiene 6% de missings, imputamos 0 (Formal). El nivel educativo del jefe de hogar tiene 1.2% de missings, imputamos con la moda. Finalmente, pegamos nuestras variables dependientes (*Pobre* y *Ingpcug).*

El Anexo 1 presenta un resumen de las estadísticas descriptivas de la base que quedó después del procesamiento. En la muestra hay un mayor número de jefes de hogar hombres con respecto a las mujeres. Así mismo, la mayor parte de los hogares no perciben ingresos por arriendos y pensiones o por concepto de dividendos e intereses. También se observa que, en promedio, hay 3 habitaciones por hogar, de las cuales 2 de ellas se usan para dormir y que cada hogar está compuesto por 3 personas, una de las cuales en un menor de edad. El ingreso total promedio de la unidad de gasto o jefe de hogar es de $2305640.

3. Modelos y Resultados

3.1. Modelos de clasificación

Se escogieron 6 modelos distintos tratando de maximizar la capacidad predictiva del modelo, para ello se usaron técnicas de regularización, remuestreo y model tuning. El primer modelo es un logit con todas las variables independientes de la base final, en total 14. Además, se calcularon modelos Lasso, Ridge y Elastic net escogiendo el lambda óptimo en cada caso y el punto de corte (threshold) que máximiza. Finalmente se calculó otro logit y un elastic net con las mismas variables dependientes luego de balacear la base por medio del método upsampling. Los métodos de regularización nos permiten ver en el Anexo 3 que entre las variables con mayor importancia se encuentran el número de menores de edad en el hogar, si el jefe de familia es informal, si alguien de la familia recibe ingresos de arriendos o pensiones, el número de habitaciones y si el jefe de hogar cuenta con educación superior.

Adicionalmente se estimaron cuatro árboles de decisión, cuyos resultados son, en general, mejores que los de los demás modelos estimados (Ver Anexo 2). El primer árbol no tiene ajuste alguno e incluye todas las variables de la base final, al igual que los árboles 2 y 3. El segundo árbol está ajustado con algunos hiperparámetros como el cost complexity de 0.0001, un mínimo de data points de 2, 14 y 17 y una profundidad de 4, 8 y 16. El tercer árbol contempla una grilla con cost complexities de 0.01, 0.001 y 0.0001, además de profundidades de 10, 15 20, 25 y 30. Estos nuevos hiperparámetros se incluyen debido a que aumentaron las observaciones de la muestra luego de hacer un upsamble para balancer las clases. Por último, el cuarto árbol contiene las mismas especificaciones

del tercer árbol, pero solo incluyendo las seis variables más importantes (Ver árbol 4 del Anexo 4). Este último modelo resulta siendo el modelo seleccionado dado que tiene un buen comportamiento en las métricas de Recall y F1, las cuales son de especial interés pues queremos evitar predecir falsos negativos. Además, el hecho de que contenga menos variables reduce el riesgo de sobreajuste.

3.2. Modelos de regresión de ingreso

Para este modelo se usaron las seis variables independientes que el modelo de lasso indicaba como relevantes (*No_menores, informal,i ng_arrie_pen, Nhabitaciones, ing_dividendo, sexojefe, niveleduc_jefe*). El mejor modelo fue la estimación por mínimos cuadrados ordinarios, el cuál se usó para predecir la pobreza con base en la línea de pobreza.

4. Conclusiones

Tras estimar diez modelos diferentes de clasificación de la pobreza, los árboles de decisión se destacaron por tener mayor precisión respecto a los modelos de regresión por regularización. En términos de Accuracy, los árboles arrojaron cifras por encima de 80, lo que significa que predicen correctamente las clases por encima de 80%. Por su parte, los modelos de regularización tuvieron puntajes de entre 21 y 25. En la predicción de ingreso, el mejor modelo fue la regresión por mínimos cuadrados ordinarios.

5. Referencias

Chagalj, C. (2019) Predicción de la pobreza en Argentina usando Random Forest. [Tesis de maestría]. Universidad de San Andrés. Buenos Aires.

Dabús, A. (2020). Pobreza en Argentina: un análisis predictivo utilizando herramientas de machine learning. [Tesis de maestría]. Universidad de San Andrés. Buenos Aires

Kim, J. (2021). Using Machine Learning to Predict Poverty Status in Costa Rican Households. Arxiv, Working Paper.

Ochoa, M., Castro-García, R., Arias, A., Machado, A. y Sucozhañay, D. (2021). Machine Learning Approach for Multidimensional Poverty Estimation. Revista Tecnológica - Espol, 33(2), 205-225.

Sabogal, H., García-Bedoya, O., y Granados, O. (2021). Un análisis de la pobreza en Colombia basado en aprendizaje automático. [Tesis de maestría]. Universidad Jorge Tadeo Lozano. Bogotá.

Anexos

Anexo 1. Estadísticas descriptivas

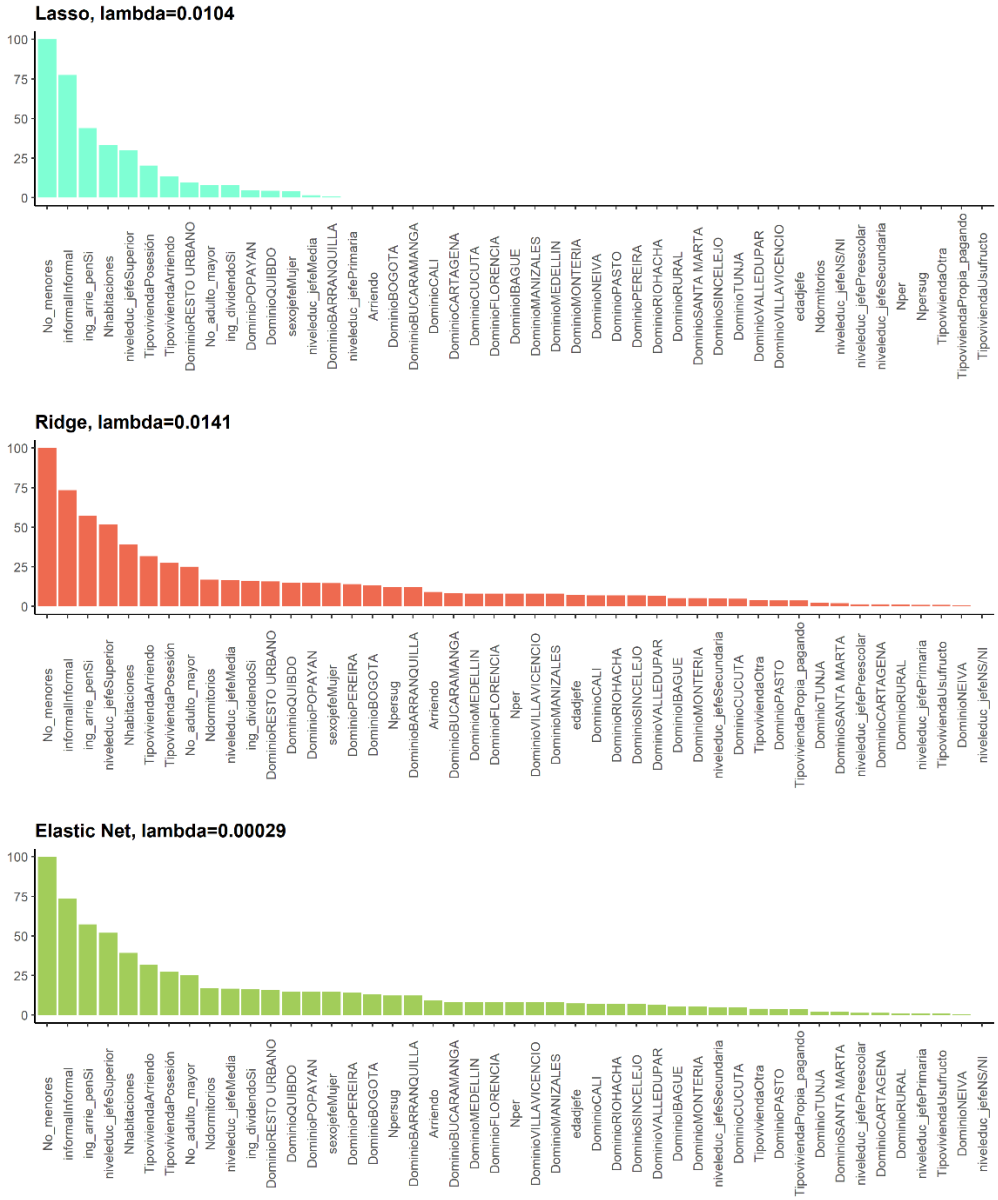
Variable	Valor mínimo	Mediana	Media	Valor máximo
Ingreso total unidad de gasto	0	1400000	2089017	85833333
Ingreso total unidad de gasto y arriendo	0	1582735	2305640	88833333

Ingreso percápita unidad de gasto y arriendo	0	544500	869748	88833333
Número habitaciones	1	3	3.389	98
Número dormitorios	1	2	1.991	15
Arriendo	20	350000	481105	8000000
Número de personas	1	3	3.295	28
Número de personas en unidad de gasto	1	3	3.282	28
Número de menores	0	1	0.9204	15
Número adultos mayores	0	0	0.46	7
Edad jefe de hogar	11	49	49.6	108
Sexo jefe de hogar			Hombre: 96272	Mujer: 68688
Ingreso arriendos o pensiones			Sí: 36184	No: 128776
Ingreso dividendos o intereses			Sí: 72231	No: 92729

Anexo 2. Métricas de modelos de clasificación

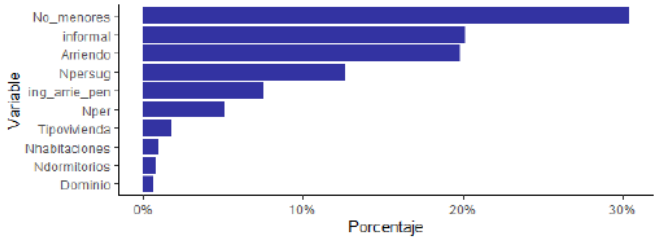
Modelo	Base	Accuracy	Precision	Recall	F1
Árbol 1	Test	83.46	69.2	30.23	42.08
Árbol 2: Grid Search	Test	84.33	64.04	48.21	55.01
Árbol 3:Grid Search+Upsample	Test	84.39	64.39	48.04	55.03
Árbol 4: Grid Search+Upsample+Selected Variables	Test	84.29	65.91	43.43	52.36
Logit Threshold	Test	21.99	21.59	23.53	30.7
LassoThreshold	Test	24.28	22.49	22.43	34.3
Ridge Threshold	Test	22.39	22.06	23.73	31.3
Elastic Net Threshold	Test	22.33	22.07	23.36	31.3
Elastic Net UpSample	Test	23.89	24.62	20.96	34.1
Logit Upsample	Test	23.73	24.413	20.97	33.9

Anexo 3. Importancia variables Ridge, Lasso y Elastic Net

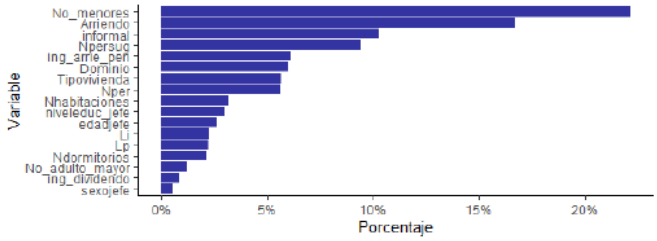


Anexo 4. Importancia variables en árboles de clasificación

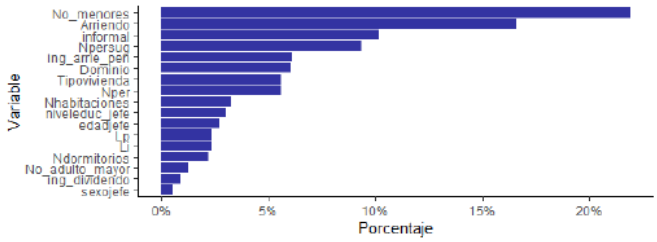
Árbol 1



Árbol 2



Árbol 3



Árbol 4