Juan Sebastián Durán Durán

Pablo Cerezo Lesmes

https://github.com/juansduran/Taller-2

Problem Set 2. Problem Set 2: Predicting Poverty "Wars of nations are fought to change maps. But wars of poverty are fought to map change" M. Ali MECA 4107

¿Cómo podemos elaborar herramientas que efectivamente clasifiquen a las personas de acuerdo con parámetros establecidos? ¿Qué utilidad podemos obtener de los modelos de predicción?

En política pública, por ejemplo, tener un buen mecanismo de clasificación puede hacer la diferencia entre proveer efectivamente los recursos de un Estado o de manera análoga -e indeseable- privar de un programa a personas en condición de vulnerabilidad. En Colombia, por ejemplo, un mecanismo de clasificación de personas es el SISBEN. Por medio de esta herramienta, las personas son incluidas en categorías para focalizar la inversión. En ese sentido, el problema de una mala clasificación radica en que, las personas ricas podrían acceder a programas sociales, mientras que, aquellas personas de menores ingresos quedarían excluidas de estos programas.

En concreto, este ejercicio busca realizar una buena predicción a partir de la clasificación de personas pobres y no pobres. Los datos que se utilizan surgen a partir de la encuesta de "Medición de la Pobreza Monetaria y Desigualdad" para el año 2018 elaborada por el DANE. El reto específico de este ejercicio consiste en que, a partir de la información suministrada, se desarrollen modelos de predicción tales qué se minimice la probabilidad de hacer una clasificación errada – falsos positivos y falsos negativos. En principio se suministraron 4 bases de datos:

Base de Datos:	Número de observaciones	Número de variables
test_hogares	66168	16
test_personas	219169	63
train_hogares	164960	23
train_personas	543584	135

En referencia a la tabla anterior, las bases de datos de entrenamiento y testeo están a nivel de personas y de hogares. Así mismo, vale la pena resaltar que, estos datos no tienen un mismo tamaño de observaciones ni tampoco contienen las mismas variables. Lo anterior implica que, en primer lugar, se debe explorar la información contenida en estas bases. En segundo lugar, se debe definir intuitivamente que variables son las necesarias para un buen modelo los modelos que se diseñaran y, finalmente, unir en una misma base, la información de hogares y personas. Una vez finalizada la exploración, y limpieza de base de datos, resultamos con información suficiente para entrenar, evaluar y testear los modelos predictivos con respecto a la pobreza y el ingreso.

Particularmente, las variables que utilizamos para realizar las predicciones fueron las siguientes:

- Pobreza: se utilizó como variable dependiente, precisamente esta variable es contra la que se van a correr los diferentes modelos. Esta variable se construyó a partir de los ingresos del hogar,

si el hogar tiene un ingreso inferior a línea de pobreza toma el valor de 1, mientras que, si el ingreso del hogar está por encima de la línea de pobreza toma el valor de 0.

- Ingresos: esta variable utilizó como variable dependiente. Es una variable continua que registra el ingreso total por persona, como los resultados se deben mostrar a nivel de hogar fue necesario agrupar los ingresos por hogar y luego dividirlos entre el número de personas que componen el hogar.
- Clase: esta variable indica si el hogar se encuentra en una cabecera municipal o resto. Al estar definida de esta forma, esta variable se utilizó como un proxy de si el hogar se encuentra ubicada en zona rural o urbana. Esta variable se incluyó entendiendo que, se espera que las condiciones de vida y el nivel de ingresos de las personas que se encuentran en la zona urbana sean mayores a aquellos que se encuentran en las zonas rurales.
- T_hab: esta es una variable discreta y hace referencia al total de habitaciones con las que cuenta la vivienda donde habita el hogar. En ese sentido, se esperaría que una vivienda con un mayor número de cuartos sea habitada por hogares de mayor ingreso. Vale la pena aclarar que dentro de estos cuartos no solo cuentan los dormitorios, sino también la sala, comedor y el estudio. Esta variable puede tener algún número de outliers si es que el hogar habita en una vivienda con más hogares.
- Dormitorios y Dormitorios2: esta es una variable discreta que hace referencia específicamente a las habitaciones donde duermen las personas del hogar. La importancia de esta variable es que logra indicar si el hogar se encuentra en hacinamiento o no. Dormitorios2 hace referencia al número de dormitorios elevado al cuadrado, se espera que, a mayor número de habitaciones ocupadas por integrantes del hogar haya un rendimiento creciente en los ingresos del hogar.
- Num_mujeresh: esta variable es una variable discreta que muestra el número de mujeres que componen el hogar. En principio, dado que injustamente a las mujeres se les ha asignado una carga de trabajo no remunerado en el hogar, se esperaría que un mayor número de por hogar generen un menor nivel de ingresos.
- Mun_adulth: es una variable discreta y describe el número de adultos que componen el hogar. La intuición económica detrás de esta variable indica que, si hay una mayor cantidad de personas en edad de trabajar y trabajan entonces, se esperaría que el hogar reporte mayor nivel de ingresos.
- Subsidio: esta variable de carácter dicótomo toma el valor de 1 si el hogar recibió algún tipo de subsidio y 0 de lo contrario. Para elaborar esta variable, fue necesario identificar si algún miembro del hogar recibió un subsidio de tipo de alimentación, transporte, educación o subsidio familiar, de recibir algún subsidio se marcaba para el hogar la identificación de si recibió o no subsidio.
- Fam_rural: es la interacción entre las variables de clase (urbano y rural) y el número de personas por hogar. Esta variable discreta busca evidenciar si un mayor número de personas en la zona rural genera impacto en el ingreso del hogar.
- Variables de ciudades Medellín (Mdll), Cali, Barranquilla (Bqa), Quibdó (Qbd), Riohacha (Rioh): este conjunto de variables es de carácter dicotómico. Específicamente se creó una variable para cada una de esas ciudades con el propósito de identificar si estar en uno de estos lugares está relacionado con ser o no pobre o con tener mayores o menores ingresos. Precisamente, para hacer este tipo de comparación se incluyen Medellín, Cali y Barranquilla que son unas de las

ciudades principales en Colombia junto con Riohacha y Quibdó que no presentan las mismas condiciones de crecimiento y condiciones de vida para la población.

Para modelos de clasificación se optó por un modelo Logit Ridge con *upsampling.* Este fue el modelo que arrojó los mejores resultados sobre sensibilidad. Teniendo en cuenta que el objetivo principal es evitar los falsos negativos y que queden personas pobres sin el subsidio. Los primeros tres modelos evaluados dan una sensibilidad de 1 que es "perfecta" y no presentan falsos negativos, pero estos son modelos *naïve* que clasifican a todos como "pobres" por lo que se evita el problema del falso negativos. Por lo tanto, no presentan un verdadero valor predictivo. Una explicación del buen desempeño de este modelo es el *up sampling* que se realiza. La base de train tiene una proporción de 5.8 % de pobres en la muestra. Esta metodología simula nuevas observaciones de la minoría (pobres en este caso) en la muestra para poder tener más información para realizar la predicción. Por último, este modelo tiene en cuenta todas las variables creadas y tuvo el mejor resultado a pesar de la penalización por número de variables que presenta un modelo Ridge.

Los hiperparámetros que maximizan la sensibilidad para este modelo son $\alpha = 1$ (característico de un modelo Ridge) y $\lambda = 0.1925$.

A continuación se presenta una tabla comparativa entre los modelos estudiados:

	ROC	Sens	Spec	Accuracy	Kappa	Lamnda	Alpha
Logit lasso sin Sampling	0,7359071	1	0	0,9449079	0	1,023293	0
Logit Ridge sin Interacciones ni ciudades	0,5	1	0	0.9449079	0	1,023293	1
Logit Elastic Net sin interacciones	0,792248	0,6501987	0,79811933	0,724194	0,4483907	0,4045618	0
Logit Ridge Up sapmle	0,7253793	0,7980202	0,5580348	0,678275	0,356055	0,192563	1
Logit Lasso Down sapmle sin ciudad	0.7967158	0.6261031	0.8196502	0.7278811	0.4457515	0.5101979	0

Modelos de regresión.

El modelo con mejor MSE es el siguiente:

$$Ingreso = \beta_0 + \beta_1 T_{hab} + \beta_2 Dormitorios + \beta_3 Clase + \beta_4 num_{mujeresh} + \beta_5\, mun_{adulth} + \beta_6 subsidio + \beta_7 Mdll + \beta_8 Cali + \beta_9 Bqa + \beta_{10} Qbd + \beta_{11} Rioh + \beta_{12} Dormitorios2 + \beta_{13}\, fam\_rural + \varepsilon$$

Para este modelo se utilizó un *down sample,* es decir, se eliminan de forma aleatorios valores de "no pobres" en la muestra para tener una base más balanceada. Este modelo utilizó todas las variables que se tenían para el modelo, incluyendo interacciones y variables categóricas por ciudad. Se debe tener en cuenta que, aunque el modelo presenta el mejor resultado de la variable deseada, éste sigue siendo un resultado bastante alto. Los modelos sin tener ajustes al sampleo mostraron

resultados peores de predicción, lo que indica la importancia de rebalancear la muestra para entrenar un modelo.

Modelos	RMSE	MSE	Requerid	MAE
DownSample Completa	1.803.241	3,2517E+12	0,2212784	9394759
Upsample Completa	1.845.479	3,4058E+12	0,207994	918369
Regresión sin resampleo	2.309.201	5,3324E+12	0,1426602	1253953
Upsample y preprocesamiento	1.849.596	3,421E+12	0,2044551	920673
Regresión sin resampleo y con preprocesamiento	2.302.387	5,301E+12	0,1476736	1249887

Después de haber corrido estos modelos para hacer una clasificación y realizar predicciones sobre nuestras fuentes de datos, se puede concluir lo siguiente: en primer lugar, un buen modelo de predicción depende tanto de la información que se le provea como del modelo que se ejecute. Es decir, si se recoge información que recoge adecuadamente variables necesarias y analíticamente se corre un modelo que capte la información, el modelo de predicción obtiene buenos resultados. Este ejercicio, por el contrario, permitió retarnos como investigadores en el sentido que, al no poseer las variables necesarias, buscamos a partir de diferentes modelos el que generara una mejor predicción con los recursos que se tenían. En este caso en particular, el poder de la herramienta permitió entrenar efectivamente con una muestra pequeña al modelo, de tal forma que se obtuvo una buena clasificación. Más allá de los resultados, el entrenar modelos y poder diferenciar la utilidad de cada uno permitió que con un logit ridge up sample y con un DownSample se hiciera un ejercicio autónomo de predicción.

Anexo 1.

Tabla de estadísticas descriptivas

	Overall
n	66168
ClaseNum (mean (SD))	0.10 (0.30)
T_hab (mean (SD))	3.41 (1.21)
Dormitorios (mean (SD))	1.99 (0.90)
Nper (mean (SD))	3.31 (1.79)
num_mujeresh (mean (SD))	1.75 (1.19)
mun_adulth (mean (SD))	2.38 (1.20)
Mdll = 1 (%)	3905 (5.9)
Cali = 1 (%)	2777 (4.2)
Bqa = 1 (%)	2735 (4.1)
Qbd = 1 (%)	1792 (2.7)
Rioh = 1 (%)	2361 (3.6)
Dormitorios2 (mean (SD))	4.79 (4.36)
fam_rural (mean (SD))	0.34 (1.17)