Problem Set 2 – Prediciendo Pobreza

Julio 12,2022

Andrea Margarita Beleño, María Valeria Gaona Guevara

*Abstract*—En el presente documento, se realizará la predicción de cuáles hogares son pobres por medio de los datos adquiridos por la Gran Encuesta de Hogares 2018. Esto se realizará por medio de dos metodologías: Clasificación y Regresión. Además, se identificarán parámetros como ROC, falsos positivos, falsos negativos y demás elementos para obtener dos modelos en donde se pueda predecir de la manera más acertada dichos hogares que son objetivo de ser implementados en las políticas relacionadas con el enfrentamiento de este problema socioeconómico. El link al Github del presente taller, se encuentra en el siguiente enlace: <https://github.com/mvgaona/Problem-Set-N-mero-2>

# Introducción

Generar una política para la población adecuada es fundamental para la construcción de una sociedad justa y en la que todos tengan oportunidades. Es por eso que es fundamental la ejecución acertada de modelos que puedan predecir correctamente la población objetivo y dicha política pueda ser aplicada a las familias adecuadas. Por lo tanto, En el siguiente documento se presentan dos modelos de predicción de pobreza en los hogares Colombianos, ya que es esencial conocer adecuadamente cuales hogares son pobres, para que la política pueda ser aplicada para quienes se encuentran en condición de pobreza y no existan casos en donde algunos hogares no sean identificados como pobres y con ello, no puedan contar con las ayudas que se plantean dentro de dicha política. El primer modelo se ejecutará por medio de clasificación de hogares pobres y finalmente, el segundo se realizará por medio de una regresión en donde se toman los ingresos y se compara con la línea de pobreza para posteriormente, definir si son pobres o no.

# Datos

***Modelo de clasificación***

La pobreza puede estar dada por diferentes variables. Sin embargo, es fundamental contar con las variables relevantes para que este modelo sea robusto, pero no se incurran en gastos que entorpezcan la investigación.

La variable Npersug (No. Personas en la unidad de gasto) Evidencia aquellas personas que dentro del hogar están dentro de la unidad de gasto. De acuerdo con el análisis, la moda de esta variable es 3, es decir, 3 personas por unidad de gasto es el valor más común entre unidades de gasto por familia. Además, se evidencia que el rango va de 1 a 28 personas por UG.

La línea de pobreza (Lp) establece el límite de ingresos por debajo del cual un hogar es considerado pobre. El valor mínimo es COP 167,222; el máximo COP 303,8107; la media COP 271,605; la moda COP 281,549.3. Además, de acuerdo con DANE(2019) evidencia que la línea de pobreza monetaria nacional fue de $257.433 pesos.

La variable Dominio es una variable categórica que indica en donde vive el hogar. Existen 25 niveles, entre ellos Bogotá, Villavicencio, rural, entre otros.

La variable categórica P5090 (OcViv) hace referencia al tipo de ocupación que tiene hogar en la vivienda, es decir, arriendo, propia, entre otros. Por otra parte, la variable numérica P5000 hace referencia a la cantidad de habitaciones que cuenta la vivienda que tiene el hogar, evidenciando que el mínimo es 1 habitación, máximo 98 y la cantidad de habitación más común es 3.

# Modelo y resultados

***Modelo de clasificación***

La pobreza puede estar dada por diferentes factores. Sin embargo, es importante conocer cuales variables son las esenciales dentro del modelo para que no se presente sobre ajuste y se pueda generar una predicción acertada. Por lo tanto, se tomaron diversos modelos para comparar cuál es el mejor modelo que puede clasificar los hogares pobres con la menor cantidad de variables posibles.

Para cada modelo se realizó dividió la muestra de entrenamiento entre 3 (Entrenamiento, test y evaluación), en donde se confrontan diferentes metodologías para poder observar cuál puede predecir mejor en la base test oficial y así, generar resultados óptimos sin sobreajustar. Se utilizaron diferentes modelos: Lasso, Ridge, Logit, Up Sample, Down Sample.

Como primer modelo, se seleccionó la variable OcViv explicada anteriormente, para poder conocer si es suficiente realizar la predicción solo con esta variable, es decir, Pobre = *b*0 + *b1*OcViv. Dentro de todas las metodologías analizadas, se encuentra que en la partición realizada, el modelo implementado con Logit\_Lasso\_Upsample genera el siguiente resultado:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** |
| 1,0232930 | 0.60395 | 0.5561771 | 0.5924888 | 0.5743329 |

## Por otra parte, se realizó la Confussion Matrix con la matriz de entrenamiento completa y generando la clasificación con el Threshold, por medio del método Logit, para así poder tener elementos más completos acerca del modelo:

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix and Statistics** | | |
| **Predicción** |  |  |
|  | 0 | 1 |
| 0 | 58148 | 9588 |
| 1 | 73788 | 23436 |
| **Acurracy** | 0,4946 |  |
| **Sensitivity** | 0,7097 |  |
| **Sensitivity** | 0,4407 |  |
| **Pos Pred Value** | 0,2411 |  |
| **Neg Pred Value** | 0,8585 |  |

En el modelo Pobre = *b0+b1* Lp +*b2*OcVivl , dentro de la partición de la matriz de entrenamiento se evidenció que el modelo implementado por Logit\_Lasso\_Upsample es el que mejor puede predecir fuera de muestra:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** |
| 0.2428436 | 0.636416 | 0.6174397 | 0.5773408 | 0.0.5973902 |

## Por otra parte, se realizó la Confussion Matrix con la matriz de entrenamiento completa y generando la clasificación con el Threshold para así poder tener elementos más completos acerca del modelo:

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix and Statistics** | | |
| **Predicción** |  |  |
|  | 0 | 1 |
| 0 | 71303 | 11691 |
| 1 | 60633 | 21333 |
| **Acurracy** | 0,5616 |  |
| **Sensitivity** | 0,646 |  |
| **Specificity** | 0,5404 |  |
| **Pos Pred Value** | 0,2603 |  |
| **Neg Pred Value** | 0,8591 |  |

En el siguiente modelo se analizan las siguientes variables Pobre = *b0+ b1P5000+ b2 OcViv*, las cuales ya han sido explicadas anteriormente. Por lo tanto, siguiendo el proceso de análisis de los anteriores modelos, se se encuentra que en la partición realizada, el modelo implementado con Logit\_Lasso\_Upsample genera el siguiente resultado:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** |
| 0.01892497 | 0.6397923 | 0.5935276 | 0.6061653 | 0.5998464 |

Por otra parte, realizando la Confussion Matrix, por medio del método Logit, se puede observar los siguientes parámetros, en donde se tomó como punto de clasificación Threshold:

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix and Statistics** | | |
| **Predicción** |  |  |
|  | 0 | 1 |
| 0 | 79995 | 13351 |
| 1 | 51941 | 19673 |
| **Acurracy** | 0,6042 |  |
| **Sensitivity** | 0,5957 |  |
| **Specificity** | 0,2747 |  |
| **Pos Pred Value** | 0,2747 |  |
| **Neg Pred Value** | 0,8570 |  |

Finalmente, en el modelo Pobre = *b0+ b1*P5000+ *b2*OcVivl+ *b3*Dominio, se realiza la metodología de partir en tres la base de datos de entrenamiento para finalmente realizar diversas evaluaciones y conocer cuál modelo es el que predice mejor. Por lo tanto, se encontraron estos dos modelos Logit\_Lasso\_Upsample(1) y Logit\_Lasso\_DownSamplesample (2), los cuales contienen un parámetro de Accurracy mayor respecto a los otros submodelos que se estaban comparando, siendo el modelo 1, el mejor modelo que puede predecir fuera de muestra.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** |
| 1 | 0,009435694 | 0,699791 | 0,6644197 | 0,629493 | 0,6469564 |
| 2 | 0,009883815 | 0,698254 | 0,6559219 | 0,6357918 | 0,6458568 |

De igual manera, se presentan los siguientes parámetros por medio de la Confussion Matrix por medio del método Logit, en donde se realizó la clasificación de los hogares pobres Colombianos por medio de Threshold:

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix and Statistics** | | |
| **Predicción** |  |  |
|  | 0 | 1 |
| 0 | 86009 | 11898 |
| 1 | 45927 | 21126 |
| **Acurracy** | 0,6495 |  |
| **Sensitivity** | 0,64 |  |
| **Specificity** | 0,6519 |  |
| **Pos Pred Value** | 0,3151 |  |
| **Neg Pred Value** | 0,8785 |  |

De acuerdo con lo anterior, el modelo óptimo para poder predecir en la base de datos Test Hogares es Pobre = *b0+ b1*P5000+ *b2*OcVivl+ *b3*Dominio, ya que este por medio de Logit\_Lasso\_Upsample cuenta con un Accuracy mayor, el cual evidencia cuál es el porcentaje de las predicciones correctas respecto a las predicciones totales. A su vez, de acuerdo a las diferentes Confussion Matrix cuenta con una proporción de Sensitivity y specificity favorable para poder realizar la predicción fuera de muestra.Además, este el mejor modelo que predice la cantidad de No pobres, falsos negativos, etc.

El número de habitaciones, la ubicación de la familia y el tipo de dominio de la vivienda son factores que explican si una familia puede ser clasificada como pobre, ya que dependiendo de donde se encuentre el hogar, el costo de vida puede ser más costoso al igual que adquirir un inmueble. Por otra parte, el número de habitaciones implica qué tan cómoda puede llegar a estar la familia en un espacio, asumiendo que entre más habitaciones tiene una casa, la familia tiene más poder adquisitivo para comprar o arrendar un inmueble, es decir, cuenta con mayores oportunidades y así, su subsistencia puede ser cubierta por el mismo hogar y no entraría a clasificar el hogar como en estado de pobreza. Por lo tanto, se procedió a realizar la predicción de los hogares pobres desde la base de datos Test hogares completa, demostrando a continuación el número de hogares pobres y no pobres:

|  |  |
| --- | --- |
| **No Pobres** | **Pobres** |
| 2421 | 63747 |

1. DANE “Boletín Técnico Pobreza Monetaria en Colombia Año 2018” Bogotá D.C Mayo 3,2019 Tomado de: <https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf>