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(2018) 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 |

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Conclusiones y recomendaciones

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

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>
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**