**Introducción**

El propósito de este trabajo es construir un modelo predictivo de la pobreza en los hogares colombianos en base a la GEIH de 2018, para ello, se emplearon dos formas de predecir la pobreza, el primero a nivel de clasificación, pobre o no pobre mediante un modelo logit, y el segundo utilizando la predicción de los ingresos de los hogares y el umbral de pobreza, se determinó la clasificación de pobre y no pobre.

Los datos de este taller fueron suministrados en base a la GEIH de 2018, específicamente de la sección “Medición de Pobreza Monetaria y Desigualdad 2018”, en donde se tienen datos a nivel hogar y a nivel individual. Dado que el enfoque es a nivel hogar, se realizaron varias transformaciones de los datos individuales para obtener la mayor cantidad de variables posibles a nivel hogar, con el fin de utilizar metodologías de machine learning y obtener un buen modelo de predicción y clasificación. El resultado final muestra que a pesar de que los datos ayudan a medir la pobreza, hay un alto riesgo según nuestras metodologías de caer en errores de sobre ajuste y de falsos negativos.

**Datos**

Inicialmente fueron suministradas 4 bases de datos, una de entrenamiento a nivel hogar y una a nivel individual, y una de pruebas a nivel hogar y una a nivel individual. Luego de limpiar las bases de datos (ver apéndice), en donde se obtuvieron la mayor cantidad de variables posibles sin missing values, se unió la información individual a las bases de datos hogares. Es importante mencionar que, en la base de datos de individuos, se tomó como referencia las observaciones del jefe de hogar como representación del hogar, y así utilizar esta información para unirla con la base de datos de hogar.

A continuación, se presentan tablas con las estadísticas descriptivas de las bases de datos después de su limpieza.

Respecto a las tablas anteriores, se resalta que todas las variables categóricas conservan prácticamente las mismas proporciones, de igual manera las líneas de indigencia y pobreza son casi idénticas.

Adicionalmente estas tablas nos permiten observar que en Colombia la mayoría de personas viven en una zona urbana, solo el 38% tienen casa propia pagada en su totalidad, el 94% de los hogares cuentan con afiliación al servicio de salud, cerca del 18% de los hogares reciben ingresos por pensión y/o arriendos, cerca del 29% de los hogares reciben ayudas económicas de otros hogares, personas, instituciones no gubernamentales, intereses por cesantías y/o dividendos-utilizades por cesantías, También llama notoriamente la atención los ingresos por hogares, donde hay un máximo de aproximadamente 85 millones de pesos, mientras se presentan mínimos de cero pesos, para la variable de ingreso total antes de arriendos, la mediana es de $1,400,000, siendo este inferior a la media, así mismo, la mediana del ingreso con imputación de arriendo es de $1,582,735, siendo inferior a la media.

Tabla 3. Estadísticas descriptivas - Base de datos prueba

















**Modelos y resultados**

**Modelos de Clasificación**

***El modelo***

Donde:

* ***Pobre****: hogar debajo de la línea de pobreza. Es la variable que se busca predecir.*
* ***P5090****: tipo de ocupación de la vivienda; propia, arriendo, usufructo, sin título, otro. De acuerdo con el modelo logit corrido esta variable es estadísticamente significativa con un p-valor de 8.18e-11. Los coeficientes de la variable indican un efecto positivo sobre la probabilidad de ser clasificado como pobre, viendo que el estado de “arriendo” y “sin título” tienen los coeficientes más altos.*
* ***nper****: número de personas por hogar. Esta variable presenta un p-valor < 2e-16. Su coeficiente indica un impacto positivo sobre la probabilidad de ser clasificado pobre.*
* ***Oc****: persona ocupada. Presenta un p-valor < 2e-16. Su coeficiente indica un impacto negativo sobre la probabilidad de ser clasificado pobre.*
* ***P6210****: nivel educativo. Presenta un p-valor < 2e-16. Su coeficiente indica un impacto negativo sobre la probabilidad de ser clasificado pobre.*
* ***Dominio****: ciudad. El p-valor para este factor varía depende de su componente, donde la única ciudad que no es estadísticamente significativa en Tunja. Se observa que el coeficiente de las ciudades capitales como Bogotá, Medellín, Manizales, Ibagué, Pereira, Villavicencio, Cali, Barranquilla y Bucaramanga indican un impacto negativo sobre la probabilidad de clasificado pobre. Pertenecer a otras ciudades indica un impacto positivo sobre la probabilidad de ser pobre.*

***KNN****:* *Confusion Matrix and Statistic*

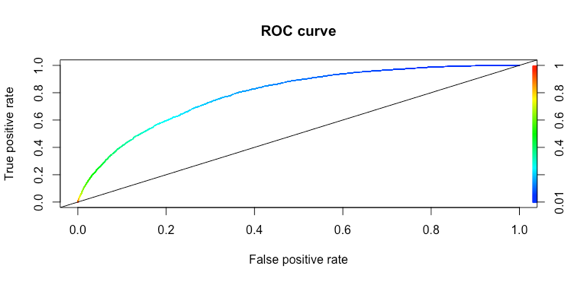
|  |  |  |  |
| --- | --- | --- | --- |
|  | *Reference* | | |
| *prediction* |  | *No* | *Si* |
| *No* | *80853* | *15888* |
| *Si* | *11529* | *7201* |

*Sensitivity : 0.31188 Specificity : 0.87520*

***Logit***

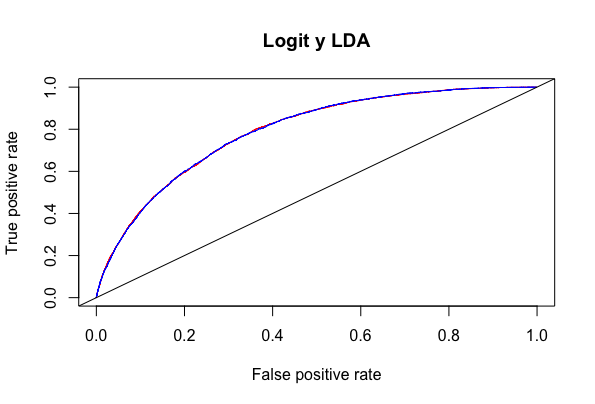
|  |  |  |  |
| --- | --- | --- | --- |
|  | *Reference* | | |
| *prediction* |  | *No* | *Si* |
| *No* | *23634* | *3833* |
| *Si* | *2752* | *2771* |

*Sensitivity : 0.4196 Specificity : 0.8957*

**

*AUC: 0.7909132*

***LDA***

******

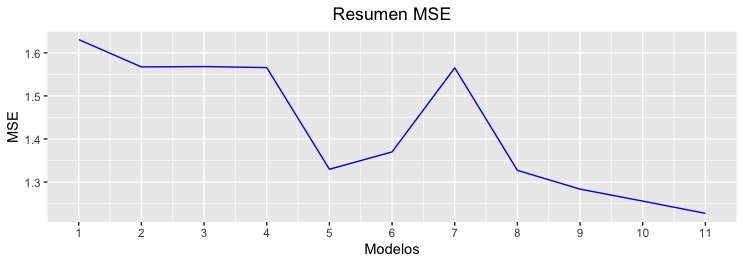
*AUC: 0.7908166*

Se utilizan las distintas metodologías de clasificación siguiendo los pasos que permitan evaluar estas muestras en base a la categorización de las distintas variables para observar efectos de control sobre variables que no son lineales o numéricas. Los resultados nos muestran que el mejor modelo está dado por la metodología de Logit.

Dado el desbalance de la muestra que puede llevar a sesgo de predicción, se utilizan métodos de sub-sampling ya sea aumentando, disminuyendo o una mezcla de ambas para balancear la muestra y mejorar la predicción.Los métodos utilizados fueron upsampling y downsampling. En este caso, el método upsample aumenta el número de observaciones clasificadas como pobres.

**Modelos de regresión de ingresos**

*Basados en la lógica económica se consideran las variables consignadas en modelo (variables descritas en el numeral anterior) presentado junto con variables de línea de pobreza y de indigencia que no resultan significativas para la predicción.*

**

*Se corren 11 modelos diferentes y se analizan los resultados para mostrar que la mejor combinación es la que controla por tipo de ocupación de vivienda, número de personas por hogar, ubicación, educación y personas ocupadas.*

Conclusión

En principio, la limpieza y análisis de las bases de datos, que primero nos hicieron ver la necesidad de evaluar por hogar y no por persona, debiendo tener que encontrar la manera de agregar los datos, muestra que todo proceso predictivo depende de una adecuada limpieza de datos.

Con la modelación se puede concluir que las mediciones del DANE cumplen su propósito de dotar de una base para logar. Una predicción de la situación de pobreza, al analizar por ingreso se puede comprobar que las variables que lo explican en nuestro caso, permiten también una clasificación acorde a lo que indica la proporción inicial de 20% de personas clasificadas como pobres. Sin embargo, también es preciso señalar que se encontró un riesgo de error tipo II, es decir clasificar personas no pobres como pobres. Inferimos que probablemente falta alguna variable más robusta de control que se encuentran dentro de los errores estocásticos.