Universidad de los Andes
Maestría en Economía Aplicada

Universidad de los Andes
Facultad de Economía

Irina Andrea Vélez López Código: 201119114
Daniel Casas Bautista Código: 202120803
Miguel Ángel Victoria Simbaqueva Código: 202224043
Lucia Fillippo Aguillón Código: 202213187

Problem Set 3. Predicting Poverty

Big Data & Machine Learning

El presente informe presenta la solución al Problem Set 3, donde se aplicaron diversas herramientas para limpieza de bases de datos y el desarrollo de un modelo predictivo de la pobreza en Colombia. El siguiente repositorio GitHub contiene los resultados del desarrollo de esta taller y el presente informe: https://github.com/irivelez/PS3_Predicting_Poverty.git

1. Introducción

El desarrollo de políticas públicas orientado a mejorar el bienestar de la población pobre y vulnerable es fundamental para fomentar una mejor redistribución de la riqueza, que permita crear condiciones favorables para brindar igualdad de oportunidades a sus habitantes. En este orden de ideas, el despliegue de las políticas públicas será óptimo en la medida en que se focalice adecuadamente la población objetivo, con el fin de que la distribución de recursos disponibles llegue a aquellas personas que más lo necesitan.

A partir de lo anterior, el objetivo es predecir la situación de pobreza de los hogares colombianos, con el fin de que las políticas públicas orientadas a esta población sean correctamente dirigidas, evitando errores de inclusión y exclusión. Para lograr esto, se utilizará un modelo de predicción de la pobreza de los hogares, a partir de datos obtenidos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad en el año 2018 por parte del Departamento Administrativo Nacional de Estadística – DANE. Se utilizará un modelo predictivo que se basará en el siguiente modelo:

$$Poor = I(Inc < Pl)$$

Donde la condición de pobreza existirá cuando el indicador *I* señale que el ingreso *Inc* sea menor a la línea de pobreza *Pl*. El modelo determinará la pobreza por dos vías; primero, con una estrategia de clasificación para predecir hogares pobres y no pobres; segundo, a partir de regresiones para determinar el ingreso de los hogares y así determinar si se encuentra por debajo o por encima de la línea de pobreza. El ejercicio identificará aspectos como el ROC[1], falsos positivos, falsos negativos y demás elementos para predecir la condición de pobreza de los hogares que serían objeto de análisis de las políticas relacionadas con este problema.

2. Datos

2.1 Transformación de los datos y construcción de las bases de datos

[1] La curva ROC (Receiver Operating Characteristic) y el área AUC (Area Under the Curve) son dos métricas comúnmente utilizadas para evaluar y comparar la calidad del rendimiento de un modelo de clasificación

El desarrollo del Problem Set 3 utilizó datos obtenidos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad del año 2018 por parte del Departamento Administrativo Nacional de Estadística – DANE, la cual contiene información que permite realizar un análisis de la pobreza en Colombia, al contener datos de ingreso, de mercado laboral, sociodemográfica, entre otros.

Los datos se encontraban disponibles en 2 bases de datos que contenían información de los hogares y personas, las cuales a su vez ya estaban divididas en un dataset de entrenamiento (train) y pruebas (test).

- **train_hogares.csv** – set de entrenamiento con 23 variables a nivel de hogares
- **test_hogares.csv** - set de prueba con 16 variables a nivel de hogares
- **train_personas.csv** – set de entrenamiento con 135 variables a nivel de individuo
- **test_personas.csv** - set de prueba con 63 variables a nivel de individuo

Como el análisis de pobreza debía realizarse a nivel de hogares, las bases de datos de interés son train_hogares y test_hogares, por lo tanto el objetivo consistió en enriquecer los datos de la base de datos de hogares, tomando información la base de datos de personas, considerando que varios individuos pueden pertenecer a un mismo hogar. Para hacer la unión de los datos se usó como llave el id que identifica al hogar, ya que en la base de datos de personas cada individuo tiene asociado un id que relaciona el hogar al que pertenece.

Como las bases de datos de hogares no tenían la misma cantidad de variables entre el set de train y test, se realizaron los siguientes ajustes en las bases, con el objetivo de dejar ambos sets de hogares con las mismas variables:

Variables de interés: poor e income

- La variable de interés a predecir es *pobre*, por lo tanto fue necesario completar en la base de datos test_hogares la información de pobreza de los hogares disponible en una base adicional llamada sample_submission.csv. Para esto se realizó un left_join.
- Se construyó la variable *tot_income_h* para la base de datos de train_hogares, a partir de la suma de los ingresos de las personas del hogar, disponible en la variable Ingtot de la base train_personas. Esta variable, se agregaría posteriormente a una nueva base de datos llamada p_train_hogares, que es una de las bases de datos finales que se usarán para entrenar los modelos.

Dejando las bases de train y test con las mismas variables

- Se identificaron las diferencias entre las variables disponibles en test y train, tanto para personas como hogares. En el proceso se depuraron de las bases de entrenamiento las diferencias, exceptuando las variables de interés o las que se usarán como dependientes. Los resultados se guardaron en unas nuevas bases de datos, para tener disponibilidad de los datos que permita la construcción de nuevas variables a partir de la información contenida en personas.

Etiquetando algunas variables y creando nuevas variables

- Algunas de las variables independientes etiquetadas son:
    - P5000: “Num_cuartos"
    - P5140: "Arriendo"
    - P5090: "Tipo_vivienda"

    - P6020: "mujer"
    - P6040: "Edad"
    - P6210: "Nivel_educ"
- Creación de nuevas variables en la base de datos de hogares
    - *Num_personas_cuarto* = Nper/P5010. Número de personas por cuarto
    - *edad_prom_h*: edad promedio del hogar. Construida a partir de la edad de cada individuo perteneciente al mismo hogar
    - *horastrab_prom_h*: promedio de horas trabajadas por el hogar. Construida a partir de las horas trabajadas de cada individuo perteneciente al mismo hogar.
    - *max_educ_h*: máximo nivel educativo del hogar. Se asignó a cada hogar el nivel educativo máximo alcanzado por algún miembro del hogar.
    - *max_health_h*: es una variable que toma el valor de 1, si algún miembro del hogar está afiliado al sistema de salud, 0 en caso contrario.

Bases de datos listas para usar en los modelos

- Finalmente, las bases de datos resultado fueron las siguientes:
    - test_hogares: con 16 variables y 66.168 observaciones. Incluye la variable *pobre* tomada de submission template
    - p_train_hogares: con 17 variables y 164.960 observaciones. Tiene la variable adicional *tot_income_h,* construidad a partir de los ingresos de las personas.
- Las bases de datos usadas en los scripts para correr los modelos se encuentran disponibles en la carpeta *stores* del repositorio indicado al principio.
- Para que la base de datos test_hogares tenga la misma cantidad de variables que el set de entrenamiento, es necesario realizar la predicción del ingreso (*tot_income_h*) a partir de los datos de entrenamiento disponibles en p_train_hogares. Esto se desarrolla en el script 3. Predicting Income, y será usado en el punto 2 de modelos y resultados.

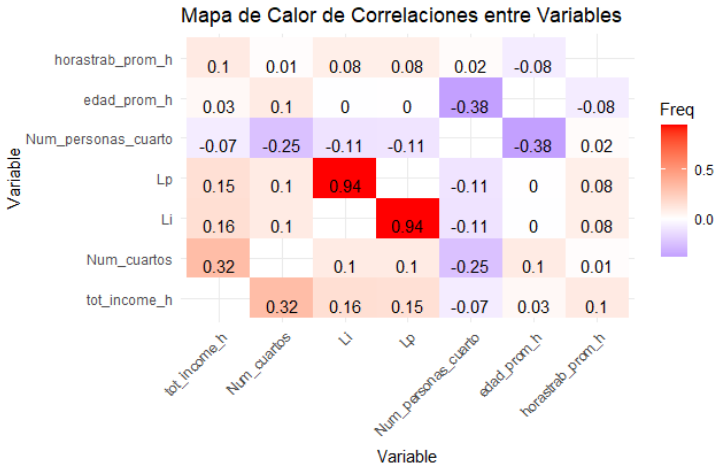
2.2 Análisis descriptivo de los datos

Una de las variables de interés es el ingreso, porque permite identificar un nivel de pobreza al compararla contra la línea de pobreza establecida por el DANE. En este sentido, una descripción general de la variable ingreso del hogar es que su primer cuartil es de $800.000 y el valor medio es de $2.102.586, el cual es 1.6 veces el salario mínimo[2] aplicable en 2017. La línea de pobreza (Lp) refleja el límite de ingresos por debajo del cual un hogar es considerado pobre, señalando que el valor mínimo es de $167.222, el máximo de $303.8107 y la media de $271.605. El nivel de pobreza, teniendo en cuenta los anteriores datos de ingreso, reflejan un total de 33.024 (20%) personas en condición de pobreza monetaria en la muestra, mientras que hay 131.936 (80%) que no lo están. Para 2018, de acuerdo con el DANE, la línea nacional de pobreza monetaria fue de $257.433[3].

General: En una revisión inicial de las correlaciones entre las variables independientes, se puede identificar que existe una relación positiva entre el número de cuartos del hogar y el ingreso total,

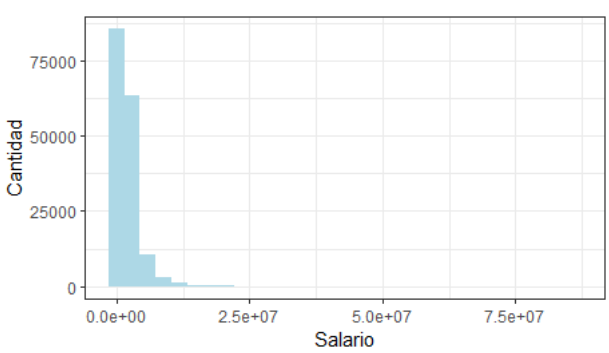
[2] De acuerdo con los valores de salario mínimo y subsidio de transporte aplicable a la fecha. Fuente: https://www.portafolio.co/economia/empleo/salario-minimo-colombia-2017-109538
[3] Fuente: Boletín técnico Pobreza Monetaria en Colombia 2018 (dane.gov.co)

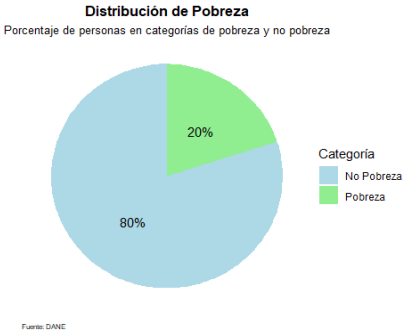
ya que a mayor ingreso existe la posibilidad de tener hogares más amplios; además, otra relación importante a destacar es la relación negativa que hay entre la edad promedio del hogar y el número de cuartos, lo que quiere decir que a mayor edad promedio, la cantidad de cuartos en el hogar se reduce, por lo que los hogares con mayor edad suelen vivir en viviendas más reducidas, esto puede estar también influenciado por varios factores, como el ingreso y los gastos del hogar.



Ingreso: Los ingresos por hogar están concentrados en valores bajos, esto se puede evidenciar en la gráfica de distribución de ingresos, lo cual concuerda con la distribución de ingresos en Colombia, en donde la mayoría de la población es de ingreso bajo y medio. En promedio, los ingresos por hogar están en 2.000.000 de pesos, el mínimo ingreso en un hogar en de 0, siendo estos los casos de pobreza, y el nivel máximo de ingresos por hogar es de 85.000.000 de pesos, los cuales son los valores que sesgan la distribución de ingresos.

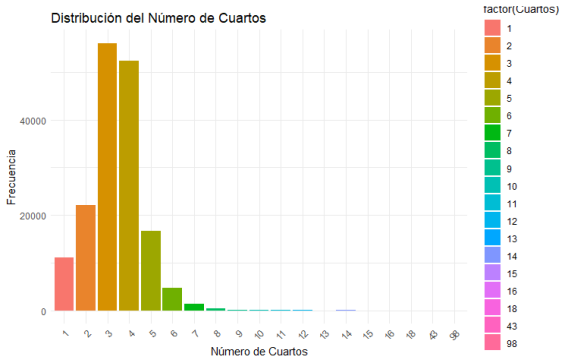
Pobreza: Se puede identificar en el siguiente gráfico de torta, que en los datos obtenidos, el 20% de los hogares está por debajo de la línea de pobreza, mientras que el resto se encuentra sobre este valor. Cabe resaltar que esta distribución proviene de la información obtenida de la base de datos submission template, la cual es una aproximación a los datos reales.

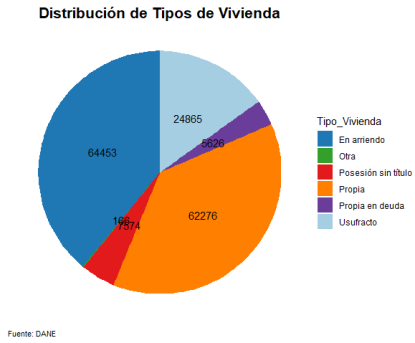




Número de cuartos: La distribución del número de cuartos en el hogar es similar a la distribución de salario vista anteriormente, ya que, a mayor salario en el hogar existe la posibilidad de tener casas con espacios más amplios.

Tipo de vivienda: Se puede observar que la mayor parte de la muestra vive en casa en arriendo, seguido de casa propia, lo cual se puede relacionar con que un porcentaje más bajo de la población cuenta con los recursos suficientes para tener casa propia. Llama la atención que una gran cantidad de hogares vive en usufructo.

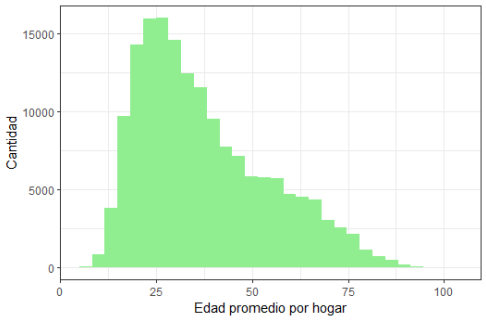


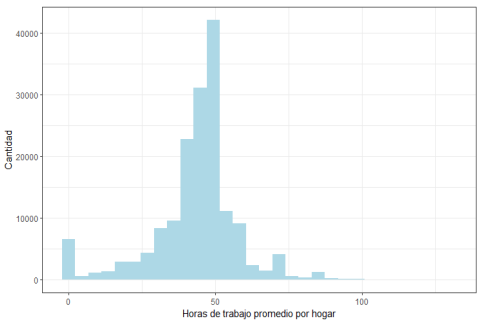


Número de personas por cuarto: se identifica que una gran cantidad de hogares tiene en promedio 2 y 3 personas por cuarto. Esta variable es de interés ya que, mientras más personas vivan por cuarto la probabilidad de que sean hogares pobres es más alta, mientras que si solo vive una persona en el cuarto, la probabilidad de que sea un hogar pobre es menor.

Edad promedio del hogar: los datos se concentran en valores bajos, a saber, la mayor cantidad de hogares tiene una edad promedio de 25 años, por lo que son hogares jóvenes. Mientras que existen pocos hogares en donde la edad promedio sea de más de 50 años. La edad puede ser un factor relevante a la hora de predecir la pobreza ya que, a mayor edad existe la posibilidad de tener más experiencia y por lo tanto más salario.

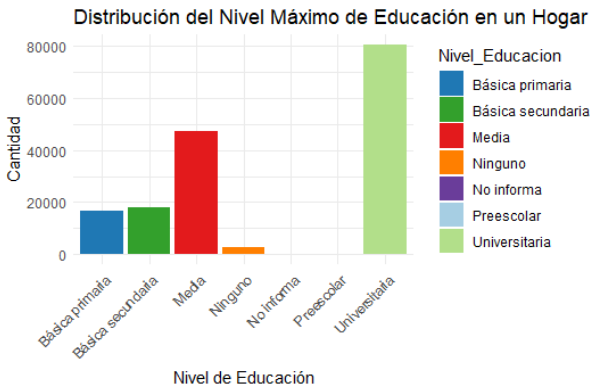
Promedio de Horas trabajadas en el hogar: esta variable muestra que la mayoría de los hogares trabaja entre 47 y 48 horas, lo que concuerda con lo establecido por ley. La distribución de esta variable parece ser normal; sin embargo, se ve una cantidad alta de hogares en 0, estos pueden ser aquellos hogares en línea de pobreza, ya que no reciben ingresos por su trabajo.

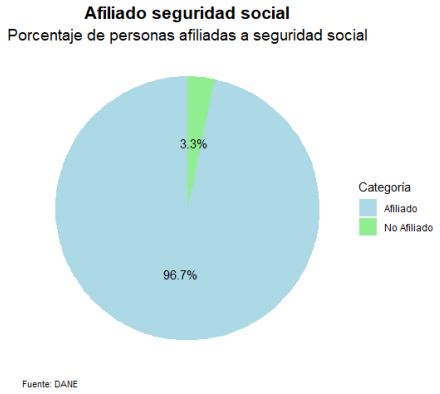




Máximo nivel educativo alcanzado en un hogar: se identifica que la mayoría de los hogares tiene algún individuo que terminó la universidad. Seguido de educación media y básica secundaria. Es de resaltar que casi 40.000 hogares en la muestra solo alcanzaron un nivel de educación de básica primaria y básica secundaria, lo cual puede influir directamente en si el hogar está por debajo de la línea de pobreza o no.

Afiliación seguridad social: se utilizó esta variable para identificar si alguna persona en el hogar está afiliada a seguridad social o no, esta variable nos puede brindar información que permita establecer si el hogar es pobre o no, en la media en que, si al menos una persona está afiliada puede significar que tiene ingresos suficientes para no estar debajo de la línea de pobreza. A saber. Se observa que el 3% de los hogares no cuenta con ningún individuo afiliado a seguridad social, esta es una distribución similar a la vista anteriormente de línea de pobreza.





3. Modelo y resultados

3.1 Clasificación

Se utilizarán modelos de clasificación binarios para realizar la predicción de aquellos hogares que son pobres y aquellos que no lo son. Para ello se utilizarán al menos tres modelos con diferentes variables predictivas hasta alcanzar el mejor resultado; además, se utilizarán distintos métodos de predicción como Logit, Lasso (tomando como métrica la sensibilidad o el ROC y haciéndolo upsample o downsample) y Elastic Net.

Para hacer este ejercicio se ha dividido la muestra de entrenamiento en tres partes; la primera, es una mini muestra de training, la cual contiene el 70% de la base de datos principal (training de hogares) y ha sido utilizada para la estimación de los modelos; la segunda, que es de evaluación, la cual ha sido útil para desarrollar técnicas de post procesamiento, evaluando el punto de quiebre óptimo de los modelos; finalmente, el tercer modelo es de testeo, cuyo objetivo es el de predecir la pobreza a partir de los modelos estimados. Para cada modelo y especificación se presentarán los resultados, de acuerdo con las métricas de ROC, sensibilidad[4], la especificidad[5], precisión[6] y el coeficiente kappa. Dado que no se trata de una muestra balanceada, la precisión no será la única variable importante por considerar.

- **Modelo 1**

[4] La cual hace referencia a la capacidad de detectar verdaderos positivos.
[5] La cual hace referencia a la capacidad de detectar verdaderos negativos.
[6] La cual hace referencia a la proporción de predicciones correctas.

Teniendo en cuenta lo anterior, el primero de los modelos elegidos es el siguiente:

$$Poor = \beta_0 + \beta_1 Npersug + \beta_2 Lp + \beta_3 Tipo_{vivienda} + \beta_4 Dominio + \beta_5 Num\_cuartos$$

Donde:

- $Poor$: Es una variable dummy que es 1 si la persona es pobre y 0 en caso contrario
- $Npersug$: Es una variable categórica que señala el número de personas por unidad de gasto
- Lp: Es una variable continua que refleja la línea de pobreza que aplica al hogar, de acuerdo con su sitio de residencia
- $Tipo_{vivienda}$: Es una variable categórica que refleja el tipo de vivienda y toma los siguientes valores. (a: Propia, totalmente pagada; b: Propia, la están pagando; c: En arriendo o subarriendo; d: En usufructo; e: Posesión sin título; f: Otra)
- $Dominio$: Es una variable que refleja la ciudad donde reside el hogar
- $Num\_cuartos$: Es una variable categórica que refleja el número de cuartos del hogar

Los resultados obtenidos con este modelo se mostrarán más adelante en la Tabla 1.

- **Modelo 2**

Posteriormente se utilizó un segundo modelo, donde se añadieron tres variables adicionales a las contenidas en el Modelo 1, las cuales se describen a continuación:

- $Máx_{Educ}$: Es una variable categórica que refleja el máximo nivel educativo alcanzado por el hogar, donde se encuentran las opciones de ninguno, preescolar, primaria, secundaria, media, superior, o no sabe.
- $Num_{hab\_ocupadas}$: Es una variable numérica que refleja el número de habitaciones en las que duermen las personas del hogar, es decir las habitaciones ocupadas.
- $Num_{personas-cuarto}$: Es una variable continua que refleja el número de personas por habitación.

Los resultados obtenidos con este modelo se mostrarán más adelante en la Tabla 1.

- **Modelo 3**

Posteriormente se utilizó un tercer modelo donde se añadieron tres variables adicionales al Modelo 2, las cuales se describen a continuación:

- Acc_{salud}: Es una variable dummy que toma el valor de 1 si el hogar está afiliado a salud
- $Edad_{prom,h}$: Es una variable continua que muestra la edad promedio del hogar
- $H_{trabajadas\ por\ hora}$: Es una variable continua que muestra la edad promedio del hogar

Teniendo presente lo anterior, a continuación se reflejan los resultados para los tres modelos elegidos para cada especificación utilizada. Nótese que se ha subrayado en azul oscuro aquellos resultados donde la precisión es mayor, esto es, al utilizar la especificación logit y elastic net para el tercer modelo.

Table 1. Resultados para los 3 modelos predictivos

MODELO 1							
Especificación	**Alpha**	**Lambda**	**ROC**	**Sens**	**Spec**	**Accuracy**	**Kappa**
Logit	N.A.	N.A.	0.7733385	0.2024085	0.9686044	0.8148902	0.2285727
Lasso (Sensibilidad)	0	0.0094356945	0.7734394	0.1762496	0.9740862	0.8140242	0.2058859
Lasso (ROC)	0	0.0098838153	0.7734398	0.1753862	0.9742054	0.8139462	0.2050074
Lasso Upsample	0	0.0136766552	0.7735692	0.7016445	0.7067255	0.7041850	0.4083700
Lasso Downsample	0	0.0130565713	0.7718717	0.6991289	0.7059481	0.7025385	0.4050769
Elastic net	0.1	0.0192128642	0.7735979	0.13519821	0.9822547	0.8123181	0.1671635
MODELO 2							
Especificación	**Alpha**	**Lambda**	**ROC**	**Sens**	**Spec**	**Accuracy**	**Kappa**
Logit	N.A.	N.A.	0.8080366	0.2842526	0.9599701	0.8244076	0.3084832
Lasso (Sensibilidad)	0	0.0124646013	0.8078311	0.254985565	0.9673044	0.8243990	0.288747289
Lasso (ROC)	0	0.0118994705	0.8078318	0.254985565	0.9673044	0.8243990	0.288747289
Lasso Upsample	0	0.0157192623	0.8082528	0.7252616	0.7381536	0.7317076	0.4634152
Lasso Downsample	0	0.0157192623	0.8068945	0.7258057	0.7349560	0.7303808	0.4607616
Elastic net	0.1	0.0025040422	0.8080862	0.2754898	0.9622343	0.8244596	0.3028093
MODELO 3							
Especificación	**Alpha**	**Lambda**	**ROC**	**Sens**	**Spec**	**Accuracy**	**Kappa**
Logit	N.A.	N.A.	0.8632603	0.4629197	0.95516	**0.8564067**	0.4828833
Lasso (Sensibilidad)	0	0.0118994705	0.8634538	0.424544538	0.9631227	0.8550731	0.461762106
Lasso (ROC)	0	0.0124646013	0.8634541	0.424415033	0.9631443	0.8550644	0.461678243
Lasso Upsample	0	0.0157192623	0.8648494	0.7789851	0.7903712	0.7846781	0.5693563
Lasso Downsample	0	0.0157192623	0.8645168	0.7785116	0.7879647	0.7832381	0.5664763
Elastic net	0.1	0.0002504042	0.8632949	0.4599844	0.9558750	**0.8563894**	0.4815270

Con los dos modelos subrayados, al subir las predicciones a Kaggle, se ha obtenido una probabilidad de 0.81 se predecir la pobreza.

3.2 Modelo de regresión de ingreso y de predicción indirecta de pobreza

Otro enfoque de abordar la clasificación de la pobreza de los hogares es a través de la predicción del ingreso. Para esto, se utilizaron los datos de ingreso disponibles en la base de entrenamiento y se realizó la predicción para la base de datos test_hogares.

Se realizó la predicción usando el siguiente modelo, ajustados con diferentes modelos de regularización, y midiendo el RMSE para cada uno los modelos regularizados.

$$\begin{aligned} log_{income} = {} & \beta_0 + \beta_1 Dominio + \beta_2 Num_{cuartos} + \beta_3 Num_{hab\_ocupadas} + \beta_4 Tipo_{vivienda} \\ & + \beta_5 Nper + \beta_6 Npersug + \beta_7 Li + \beta_8 Lp + \beta_9 pobre \\ & + \beta_{10} Num_{personas-cuarto} + \beta_{11} Edad_{prom,h} + \beta_{12} H_{trabajadas\ por\ hora} \\ & + \beta_{13} Máx_{Educ} + \beta_{14} Acc_{salud} \end{aligned}$$

La medición del RMSE para cada modelo de regularización, indica que el mejor RMSE fue con Lasso, seguidamente de Elastic Net sin usar variables estandarizadas.

Table 2. Cálculo de RMSE para los 3 modelos de regularización

Modelo	Muestra	RMSE
Elastic Net con var estandarizadas	Dentro	204593540
Elastic Net	Dentro	204547123
Lasso	Dentro	203380810
Ridge	Dentro	216888177

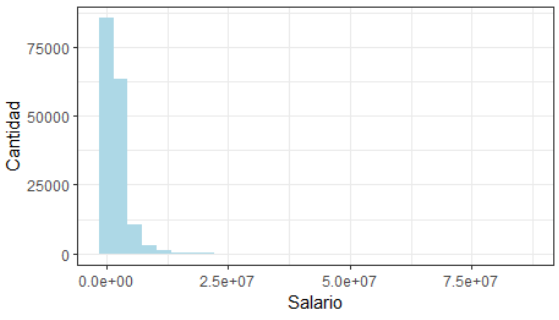
Considera lo anterior, se presentan las estadísticas descriptivas de la variable ingreso predicha con lasso en la base *test*, comparada con la base de *train*

Table 3. Estadísticas descriptivas de la variable ingreso

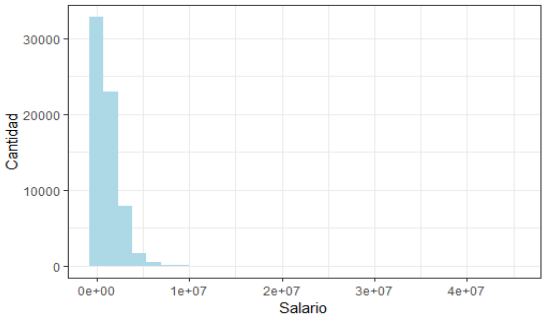
Medida	Train	Test
Min.	0	17,718
1st Qu.	800,000	357,297
Median	1,400,000	760,547
Mean	2,102,586	1,239,859
3rd Qu.	2,518,242	1,760,526
Max.	85,833,333	52,730,261

Se puede observar que, si bien existe una diferencia en todas las medidas de ambas bases de datos. Su comportamiento y distribución es similar; en ambos casos la mayor cantidad de observaciones se encuentran en niveles de ingreso bajo, también existen datos altos de ingresos que sesgan la muestra, lo cual se puede observar en las siguientes gráficas.

Distribución de salario (*train*)



Distribución de salario (*test*)



A partir de la predicción del ingreso para la base *test,* se comparó este resultado con la línea de pobreza para realizar la clasificación del hogar como pobre si el ingreso se encontraba por debajo de la línea de pobreza, y como no pobre en caso contrario. De esta manera, se realizó la predicción indirecta del nivel de pobreza de los hogares, a partir de la estimación del ingreso de los hogares.

3.3 Modelo de clasificación final y resultados

El modelo final utilizado fue el 3, utilizando Elastic net, dado que su precisión fue igual de buena que Logit, y tiene la ventaja de manejar eficientemente variables no lineales como, por ejemplo, el ingreso per cápita, la educación promedio, antigüedad en el trabajo entre otras; además, puede ser bueno prediciendo muestras desbalanceadas.

$$Poor = \beta_0 + \beta_1 Npersug + \beta_2 Lp + \beta_3 Tipo_{vivienda} + \beta_4 Dominio + \beta_5 Num_{cuartos} + \beta_6 Máx_{Educ} + \beta_7 Num_{hab\_ocupadas} + \beta_8 Num_{personas-cuarto} + \beta_9 Acc_{salud} + \beta_{10} Edad_{prom,h} + \beta_{11} H_{trabajadas\ por\ hora}$$

Los resultados de la comparación de los modelos se resumen en la Tabla 1 de la sección 3.1

4. Conclusiones y recomendaciones

- Con el objetivo de implementar políticas públicas focalizadas, implementar modelos de predicción de la pobreza es fundamental para los gobiernos. Esto puede permitir la fácil identificación de la población con más necesidades y asignar de manera óptima los recursos disponibles que permitan la mejora en el bienestar social.
- En el presente documento, el primer modelo predictivo tuvo su mejor rendimiento haciendo uso de las especificaciones Logit y Elastic Net. En estos casos, los valores de precisión fueron de 0.814 y 0.812, respectivamente.
- La predicción indirecta de la pobreza a partir de la estimación del ingreso no resultó siendo tan buena como la predicción directa realizada a través de los modelos de clasificación. Esto puede ser resultado de que Machine Learning es mucho mejor realizando predicciones de nuevos valores a partir de la correcta identificación de patrones en los datos, en lugar de resolver estimaciones con problemas de regresión.
- En el segundo modelo predictivo, donde se incluyeron tres nuevas variables al Modelo 1, nuevamente se obtuvieron las mejores precisiones al utilizar Logit y Elastic Net. En estos casos, los valores de precisión fueron de 0.82432 y 0.8242, respectivamente; sin embargo, nótese la leve diferencia en la precisión de las especificaciones, aproximadamente 0.0000520.
- Finalmente, para el tercer modelo predictivo, donde se incluyeron tres variables predictivas adicionales al Modelo 2, nuevamente los mejores resultados se obtuvieron con las especificaciones Logit y Elastic Net, alcanzando valores más altos que en los anteriores modelos, llegando a 0.8564 y 0.8563, respectivamente.
    - El número de variables afecta considerablemente el performance y aprendizaje de cada modelo. Nótese la diferencia entre los resultados evolutivos entre el primer modelo de clasificación y el modelo final: Esto se debe a que en el primer modelo hay seis (6), en el segundo nueve (9), y finalmente, en el tercero doce (12) variables.