Profesor: Ignasio Sarmiento-Babieri

Problem Set 2: Predicting Poverty

Big Data y Machine Learning para Economía Aplicada, 2024-I
Fecha de Entrega: **14 de abril de 2024**.

Integrantes:
Paula Osorio [1]; Sandra Gamarra Palencia [2]; Erika Macías [3] e Ingrith Sierra [4]

1. Introducción

En los últimos años, el panorama económico y social de América Latina ha sido complejo y poco alentador, dada las diversas situaciones coyunturales presentadas, como por ejemplo los estragos ocasionados con la pandemia del COVID-19, en la que muchos países para frenar las consecuencias sobre la economía y el tejido social. implementaron medias estrictas de confinamiento que afectó significaticamente a la población de los sectores más vulnerables de la sociedad, especialmente los que viven en la pobreza y la pobreza extrema (Lustig & Tommasi (2020). Sin embargo, gracias a las diferentes acciones gubernamentales implementadas, han permitido la recuperación paulatina de los diferentes sectores económicos, que en la actualidad siguen en la construcción y ejecución de políticas públicas que les ayude a recuperar sus indicadores económicos y sociales a los niveles prepandemia.

Según Datos de la Cepal (2023) , el porcentaje de personas en sel porcentaje de personas en situación de pobreza en el año 2022 bajó a 29 % de la población de América Latina (181 millones de personas), 1,2 puntos porcentuales menos que antes del inicio de la pandemia de COVID-19, mientras que la pobreza extrema disminuyó a 11,2 % de la población de la región (70 millones de personas), manteniéndose en niveles similares a 2019.

En ese sentido, la pobreza se define como la carencia frente a un umbral establecido, en el que una persona u hogar será considerado pobre si se encuentra por debajo del umbral respecto de un atributo especifico (Casas & Barichello, 2015). De acuerdo a la Comisión Económica para América Latina y el Caribe (CEPAL) la pobreza es la carencia de ingresos respecto de un umbral de ingreso absoluto correspondiente al costo de una canasta de alimentos básicos.

En 2022, el Banco Mundial estableció la línea de pobreza mundial en 2,15 dólares según la Paridad de Poder de Adquisitivo (PPA) de 2017; en este sentido, una persona en pobreza extrema es aquella que vive con menos de 2,15 dólares al día. En 2017, alrededor de 700 millones de personas se encontraban en esta condición (Banco Mundial, 2022).

La Comisión Económica para América Latina y el Caribe (CEPAL) determina el valor de la línea de pobreza para la región con base en el valor de una canasta de alimentos que cubre los requerimientos mínimos calóricos, la disponibilidad efectiva de los alimentos y sus precios relativos; luego el valor de esa canasta se multiplica por el coeficiente de Orshansky, de esta forma, la línea de pobreza corresponde a los recursos requeridos de los hogares para satisfacer sus necesidades alimentarias (Muñoz, 2009). En Colombia, bajo

[1]Código: 201327186
[2]Código: 202225782
[3]Estudiante de Educación continua
[4]Código: 201720654

esta metodología, la línea de pobreza monetaria extrema per cápita es de $161.099 y la línea de pobreza per cápita es de $354.031 en 2021 (DANE, 2022).

Lora et al. (2024) señalan que los indicadores promedio de 12 meses o 4 trimestres, como las tasas de línea de pobreza, se calculan a partir de secciones transversales repetidas de encuestas de hogares y se interpretan como anuales. Identifican que esta forma de medición implica que los individuos no cambian dentro de un año, lo cual afecta la medición del calculo de medidas del mercado laboral, la desigualdad y la pobreza monetaria.

Para ello, proponen varios métodos para anualizar los datos subanuales. Algunos de ellos se basan en encuestas que requieren preguntas auxiliares mientras que otros requieren técnicas econométricas como el emparejamiento predictivo de medias. Lora et, al. (2024) utilizan la metodología propuesta con datos de Colombia, de esta forma, obtienen medidas de participación laboral, ingreso laboral per cápita, ingreso familiar promedio per cápita, coeficientes de Gini de ingreso laboral e ingreso familiar per cápita, e índices de pobreza monetaria moderada y extrema. Finalmente, muestran la comparación con las medidas tradicionales.

Los datos usados en este trabajo fueron obtenidos del Departamento Administrativo Nacional de Estadística (DANE) a través de la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP), que se revela como un instrumento fundamental para medir la pobreza en Colombia mediante una nueva metodología, en la que se atribuye a la innovadora aproximación adoptada para la evaluación de la pobreza monetaria en Colombia, que implica cambio tanto en la línea de pobreza como en la manera de calcular el ingreso total del hogar. La metodología aplicada no solo se actualiza para reflejar mejor los patrones de consumo actuales en los hogares del país, sino que también integra avances metodológicos reconocidos por la comunidad experta internacional y se basa en las estimaciones más exactas del ingreso agregado de los hogares. Esta estrategia metodológica no solo mejora la medición interna, sino que también facilita la comparación con otros países de la región, alineándose con los estándares y prácticas comunes a nivel regional.

Los resultados muestran que el mejor modelo resulta ser elastic net con todas las variables, que incluye proporciones tomadas de la base de train individuos, y permite darle mayor información sin generar overfit, que parece que solo favorece a elastic net y no tanto a los modelos de random forest y boosting. Por su parte, para el modelo de regresión, el que demostró mayor eficiencia fue el de regresión lineal, dada su sencillez y capacidad para adaptarse a los datos, y predecir de forma indirecta la condición de pobreza de los hogares en Colombia.

2. Descripción de los Datos

Los datos utilizados para el desarrollo de este documento fueron tomados del DANE y de la misión para el “Empalme de las Series de Empleo, Pobreza y Desigualdad - MESEP”, que toma como insumo los datos de la encuesta “Medición de Pobreza Monetaria y Desigualdad 2018” creado a partir de la Gran Encuesta Integrada de Hogares - GEIH - de 2018, que usa el DANE para realizar las estimaciones del Indice de Pobreza Multidimencional (IPM) y clasificar a los hogares como pobres y no pobres basado en la línea de pobreza para Colombia descrita en en el Boletín Técnico Pobreza Monetaria en Colombia Año 2018 y diferentes investigaciones en la línea de Desarrollo Económico y Mercado Laboral. Los datos usados presentan la información de hogares e información de las personas, en la que se realiza un procesamiento de estas para crear una base unificada y lograr realizar un emparejamiento a nivel hogar, en donde a cada individuo se le asigna la información correspondiente a la del hogar. Para lograr el objetivo de este trabajo, la unidad observación es el hogar, en la que se seleccionaron diferentes variables de interés a nivel de individuos y se realizaron cálculos para lograr obtenerlas a nivel de hogar que permitan conocer las características dentro del hogar como lo son, su composición género, nivel educativo, afiliación al sistema de seguridad social y pensional, entre otros.

La base de datos usada en este estudio cuenta con 231.128 hogares, de estos, 164.960 hogares componen la muestra objeto de entrenamiento de los modelos y 66.168 hogares componen muestra de testeo o prueba del modelo que se presentara en la siguiente sección. Para consolidar las bases de datos fue necesario realiza un proceso de imputación de datos para algunas variables de interés que contenían valores missing, esto permite que no se pierdan observaciones en ninguna de las muestras dentro del universo de estudio. Se realizan además el renombramiento de las variables seleccionadas tanto en la base de entrenamiento como en la base de prueba en las bases de hogares, variables relacionadas con las características físicas de la vivienda (total cuartos en la vivienda, y total cuarto usados para dormir); tipo de vivienda (propia pagada, propia pagando, arriendo, usufructo, posesión y otra); cuota de arriendo, número de personas en el hogar, entre otras. Para la base de personas tanto en la base de entrenamiento como en la de testeo, se renombraron y seleccionaron variables como género, nivel educativo, afiliación al sistema de salud y al sistema de pensión, régimen de salud, población en edad de trabaja, entre otras.

La tabla 1 muestra el resumen de las estadísticas descriptivas del universo de estudio, discriminando en la muestra las observaciones de entrenamiento y las de testeo, su vez, dentro de la muestra de entrenamiento, se dividen la muestra entre pobres y no pobres. Se observa que el total de observaciones para la muestra de entrenamiento está compuesta por 164.960 hogares, de estos, 33.024 son pobres que representa el 20 % del total de hogares y 131.936 se clafican como no pobres, mostrando un indicio del problema del desbalance de clases que se tiene en cuenta para trabajar los modelos. Por su parte, la muestra de testeo la componen 66.168 hogares.

El análisis de estás estadísticas muestra que tanto en la muestra de entrenamiento como de testeo no existen diferencias significativas entre las diferentes características de las personas en el hogar; evidenciando que las mujeres son más del 52 % de las personas que componen el hogar; la proporción de personas dentro del hogar que cursan educación superior es dentro de ambas muestras es aproximadamente 56 %. Por otro lado, la proporción de personas en el hogar afiliados al sistema de seguridad social en salud está al rededor del 80 % y la proporción de afiliados al sistema de pensión es del 23 %. En cuando al mercado laboral, se evidencia que en ambas muestras la proporción de Población en Edad de Trabajar (PET) es superior al 86 % y la proporción de desocupados es del 6 %.

Cuando observamos las características en la muestra de entrenamiento para los pobres y los no pobres, se observa diferencias importantes entre estos dos grupos principalmente en la proporción de individuos dentro del hogar que tienen educación superior, en la que se observa que para los pobres es del 38,9 % y para los no pobres del 56 %; además se muestra que la proporción del afliados al sistema de seguridad social en salud para los pobres es del 69 % y para los no pobres es del 83,6 %. Consecuentemente, la proporción de personas en el hogar que hacer parte o están afiliados al sistema de pensión en el país es para los pobres de apenas el 4 %, mientra que para los hogares no pobres esta es del 29 %; la porporción de PET en los hogares pobres es del 76 % y los desocupados es del 11 %, en contraste con la proporción de hogares no pobres que componen la PET es del 89 % y la proporción de desocupados es del 5 %.

Estos resultados entre estos grupos reflejan las brechas y diferencias existentes entre pobres y no pobres en el país, que pueden ser explicadas por múltiples factores socioeconómicos que influyen en las oportunidades y acceso a recursos. Se evidencian disparidades en el acceso a la educación de calidad y oportunidades de formación avanzada y las barreras económicas y sociales que enfrentan los hogares más pobres, limitando su capacidad para invertir en educación y, por lo tanto, reduciendo sus oportunidades de empleo calificado y de mejor remuneración. Consecuentemente, la menor afiliación al sistema de seguridad social y al sistema de pensiones en los hogares pobres refleja una precariedad laboral más marcada. Los trabajos informales, que son más comunes entre los pobres, raramente ofrecen beneficios sociales como salud y pensiones. Esta falta de protección y seguridad social perpetúa la vulnerabilidad de estos hogares a enfermedades y a la pobreza en la vejez. Asimismo, la mayor tasa de desocupación en los hogares pobres también es indicativa de las

dificultades en el acceso al mercado laboral que estos enfrentan, donde la falta de habilidades educativas y técnicas reduce su empleabilidad.

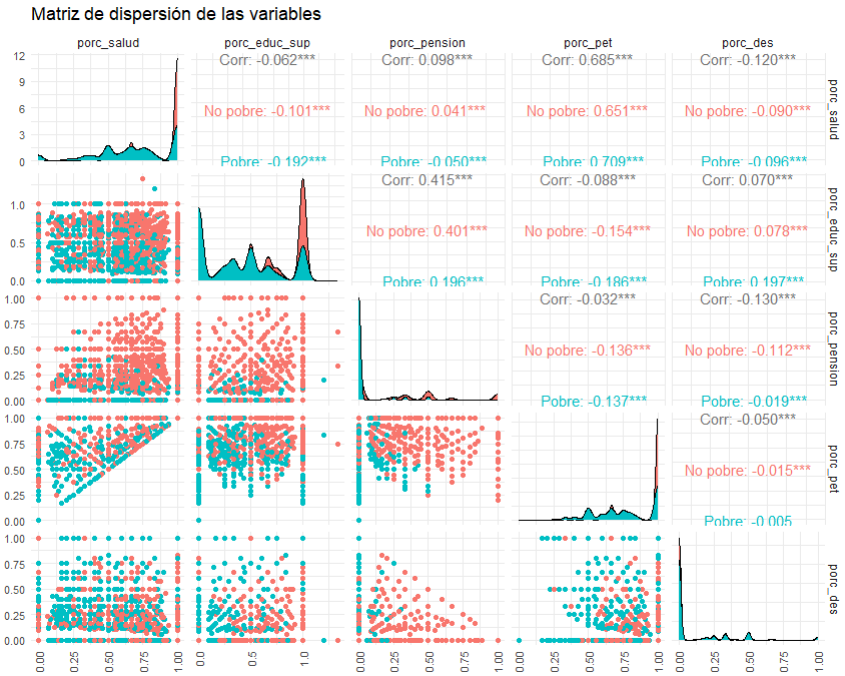
Tabla 1: Estadísticas Descriptivas de Hogares en Colombia clasificado en pobre-no pobre

	Muestra de entrenamiento						Muestra de testeo	
	Todos		Pobres		No Pobres		Todos	
	Media	Desv. Estándar	Media	Desv. Estándar	Media	Desv. Estándar	Media	Desv. Estándar
A. Composición								
Prop. Mujeres	0,526	0,277	0,546	0,241	0,521	0,285	0,523	0,276
B. Educación								
Edu Superior	0,569	0,388	0,389	0,364	0,613	0,381	0,561	0,389
C. Salud y Pensión								
Salud	0,807	0,255	0,692	0,280	0,836	0,240	0,808	0,254
Pensión	0,241	0,320	0,041	0,122	0,290	0,335	0,232	0,316
D. Mercado Laboral								
PET	0,866	0,187	0,765	0,215	0,891	0,171	0,865	0,187
Desocupado	0,064	0,169	0,113	0,224	0,052	0,150	0,063	0,169
Observaciones	164.960		33.024		131.936		66.168	

Nota: DANE - Medición de Pobreza Monetaria y Desigualdad 2018. Cálculo de autoras. Los valores representados en la tabla corresponden a la proporción de personas que se encuentra dentro de cada categoría

La figura 1 representa la matriz de dispersión para las variables presentadas en la tabla 1 para la población pobres y no pobres. Cada cuadro en la diagonal principal muestra la distribución de una variable única, con histogramas para los no pobres superpuestos sobre los pobres, proporcionando una visión inmediata de cómo cada grupo se distribuye en cada variable. Así, los coeficientes de correlación sugieren la fuerza y la dirección de la relación entre estas variables analizadas, mostrando que la correlación entre la proporción de individuos con educación superior y la proporción de individuos con pensiones es positiva (Corr: 0.098***) para toda la muestra, lo que indica una tendencia general a que mayores niveles de educación coinciden con mayores tasas de afiliación a pensiones. También se evidencia la relación negativa entre la proporción de la PET y la proporción de desocupados para los pobres y una correlación ligeramente positiva para los no pobres, que sugiere que para los hogares pobres, un mayor nivel de actividad económica no necesariamente conlleva a una menor desocupación, lo cual podría reflejar empleo precario o subempleo.

Figura 1: Gráfico de dispersión de las variables



3. Análisis de modelos y resultados

3.1. Modelo de regresión: Ingresos

La pobreza monetaria es un fenómeno complejo multifactorial que ha afectado a millones de hogares en América Latina, privándolas de recursos básicos para una vida digna . Determinar los ingresos de un hogar es fundamental para comprender y abordar este problema de manera efectiva. Los ingresos de un hogar se refieren a los recursos financieros que reciben sus miembros, ya sea a través de salarios, beneficios sociales, pensiones u otras fuentes (MacEwan. Artur, 2010).

Sin embargo, evaluar los ingresos de un hogar no se limita simplemente a sumar cifras; también implica considerar la estabilidad de estos ingresos a lo largo del tiempo, así como los gastos necesarios para cubrir las necesidades básicas, como vivienda, alimentación, salud y educación. Además, es crucial tener en cuenta las diferencias regionales y contextuales que pueden influir en la capacidad de un hogar para satisfacer sus necesidades básicas con los ingresos disponibles. Por lo tanto, determinar con precisión los ingresos de un hogar es un primer paso crucial para diseñar políticas efectivas de lucha contra la pobreza y promover la igualdad de oportunidades (Bassetto. Giovanni).

Para ello, identificar variables clave que puedan predecir los ingresos de un hogar es esencial para establecer un umbral efectivo de pobreza. Estas variables corresponden al nivel educativo, experiencia laboral, el tipo de ocupación, compocisión familiar y la accesibilidad a servicios básicos como la salud y la vivienda. Además, es importante considerar el tipo de ingresos que recibe el hogar, como subsidios y arrendamientos, dado que estos también influyen significativamente en su situación financiera, y por consiguiente su condición de pobreza.

Dado lo anterior, se implementaron tres metodologías para predecir el ingreso de un hogar mediante regresiones que permitieran capturar las relaciones entre las variables independientes y el ingreso por hogar. Posteriormente, con el ingreso medio de los hogares determinados pobres, se estable el umbral de ingresos que permite clasificar si un hogar es considerado pobre o no.

3.1.1. Regresión lineal

El modelo de regresión lineal, al emplear variables socioeconómicas como el número de personas en el hogar, la clase social, la salud y la educación, tiene como objetivo predecir el ingreso de un hogar. Una vez realizadas estas predicciones, se establece un umbral de ingreso para discernir si un hogar puede ser clasificado como pobre o no.

Para predecir el ingreso se tomo la base de datos de hogares tomando como variables dependiente ingreso y como explicativas el numero de personas, mujeres, afiliación a salud, subsidio de transporte, otros ingresos, recibe ayudas, total de cuartos y tipo de vivienda. Se toman estas variables pues se considera que son aquellas que explican el ingreso y nos pueden dar una aproximación a predecir indirectamente pobres (1) y no pobres (0). Los algoritmos usados fueron regresión lineal, elastic net, boosting y random forest.

$$\begin{aligned}\text{Ingreso} = \beta_0 &+ \beta_1 \text{nper} + \beta_2 \text{nsalud} + \beta_3 \text{neducsup} + \beta_4 \text{sub\_transporte} \\ &+ \beta_5 \text{otros\_ing} + \beta_6 \text{ayud} + \beta_7 \text{ndes} + \beta_8 \text{tot\_cuartos} \\ &+ \beta_9 \text{tipo\_vivienda} + \mu\end{aligned}$$

Donde β_0 es el intercepto, $nper$ el número de personas en el hogar, $nsalud$ si están afiliados a salud, $neducsup$ nivel educativo, $sub\_transporte$ representa la variable subsidio de transporte, $otros\_ing$ si recibe otros ingresos, $ayud$ si recibe ayuda, $ndes$ que representa el número de personas desempleadas en el

hogar, $tot\_cuartos$ el número de cuartos y $tipo\_vivienda$ que representa el tipo de vivienda. Se toman estas variables, ya que se considera que son las que explican el ingreso.

Para hacer la predicción indirecta de pobreza se calcula la media de los hogares pobres de la base de entrenamiento, omitiendo las colas, de esta forma, se clasifica si un hogar es pobre o no. El ingreso se utiliza como medida indirecta de pobreza, ya que permite determinar si un hogar es pobre o no según sus ingresos. Es una forma ampliamente utilizada en los métodos de medición de pobreza que optan por utilizar un umbral de ingreso para definir quiénes son pobres o no. Además, el ingreso es una de las variables que define si una persona u hogar puede acceder a satisfacer sus necesidades básicas, de ahí que entidades como la CEPAL calculen la línea de pobreza con base en el acceso a una canasta básica de alimentos.

3.1.2. Random forest

Para hacer este modelo se usaron 500 bootstrap con 5 variaciones aleatorias seleccionadas en cada partición y 100 como el número mínimo de observaciones. Se utiliza una regresión similar a las anteriores añadiendo como variable de tiempo relacionada al tiempo de antiguedad en el trabajo actual.

$$\begin{aligned}\text{Ingreso} = \beta_0 &+ \beta_1 \text{nper} + \beta_2 \text{clase} + \beta_3 \text{nmujeres} + \beta_4 \text{nsalud} \\ &+ \beta_5 \text{neducsup} + \beta_6 \text{subtransporte} + \beta_7 \text{otrosing} + \beta_8 \text{ayuinst} \\ &+ \beta_9 \text{ndes} + \beta_{10} \text{totcuartos} + \beta_{11} \text{Tipovivienda} + \beta_{12} \text{tiempocon} + u\end{aligned}$$

3.1.3. Elastic net

El modelo Elastic Net es un método de regresión que combina las propiedades de la regresión de Ridge (L2) y la regresión Lasso (L1). Este método utiliza una penalización combinada de las normas L1 y L2 para regularizar el modelo, lo que ayuda a evitar el sobreajuste y a mejorar la generalización del modelo. Para este caso, el modelo Elastic Net se entrena utilizando el método de regularización glmnet, permitiendo optimizar una función de pérdida que combina dos términos: el término de ajuste del modelo y el término de regularización.

Para realizar la regresión elastic net se utilizo la siguiente función de regresión

$$\begin{aligned}\text{Ingreso} = \beta_0 &+ \beta_1 \text{nper} + \beta_2 \text{clase} + \beta_3 \text{nmujeres} \\ &+ \beta_4 \text{nsalud} + \beta_5 \text{neducsup} + \beta_6 \text{subtransporte} \\ &+ \beta_7 \text{totcuartos} + \beta_8 \text{tipovivienda} + \beta_9 \text{cuotaamort} + \mu\end{aligned}$$

Donde β_0 es el intercepto, $nper$ el número de personas en el hogar, clase, mujer, $nsalud$ si este afiliado a salud, $neducsup$ nivel educativo, $subtransporte$ representa a la variable subsidio de transporte, $totcuartos$ que son el total de cuartos que tiene una vivienda, el tipo de vivienda y , $ndes$ que representa el número de personas desempleadas en el hogar, $tot\_cuartos$ el número de cuartos, $tipo\_vivienda$ que representa el tipo de vivienda y $cuotaamort$ que representa el valor de una cuota de amortización (créditos). Esta ultima variable se añade ya que se considera que pagar créditos puede afectar el ingreso.

3.2. Modelo de clasificación: Pobreza

Para predecir si un hogar se clasifica como pobre o no pobre, se tomó la base de datos con las variables organizadas y las creadas a partir de la base de datos de individuos. En total se tenía la disponibilidad de 25 variables independientes además de la variable pobre para los datos de entrenamiento, siendo el modelo

posible con mayor cantidad de variables el siguiente:

$$\begin{aligned} Pobre = \beta_0 &+ \beta_1 \text{Clase} + \beta_2 \text{tot\_cuartos} + \beta_3 \text{tot\_(cuartos\_dorm)} \\ &+ \beta_4 \text{tipo\_vivienda} + \beta_5 \text{cuota\_amort} + \beta_6 \text{cuota\_arriendo} \\ &+ \beta_7 \text{nper} + \beta_8 \text{lin\_indig} + \beta_9 \text{lin\_pobreza} \\ &+ \beta_{10} \text{Ingtotug} + \beta_{11} \text{nmujeres} + \beta_{12} \text{nsalud} \\ &+ \beta_{13} \text{neducsup} + \beta_{14} \text{npension} + \beta_{15} \text{npet} \\ &+ \beta_{16} \text{ndes} + \beta_{17} \text{jefe\_mujer} \\ &+ \beta_{18} \text{jefe\_(salud\_cont)} + \beta_{19} \text{jefe\_(educ\_sup)} \\ &+ \beta_{20} \text{porc\_mujer} + \beta_{21} \text{porc\_salud} \\ &+ \beta_{22} \text{porc\_(educ\_sup)} + \beta_{23} \text{porc\_pension} \\ &+ \beta_{24} \text{porc\_pet} + \beta_{25} \text{porc\_des} + \mu \end{aligned}$$

Para poder tener una noción del desempeño del modelo comparable entre todos, se dividió la base de train en un 70/30, para tener una submuestra de train, de modo que con el restante 30 % de los datos de train se pudiera calcular un F1.

3.2.1. Elastic Net

Se probaron 2 modelos de Elastic Net, el primero con una versión reducida de las variables mencionadas anteriormente, eligiendo las que según la literatura económica pueden dar señales de pobreza más claras, como: la ubicación de la vivienda (cabecera), el tamaño de la vivienda, que en este caso se mide en total de cuartos, la cantidad de personas que viven en el hogar, además de los que están en edad de trabajar que son los que pueden atraer ingresos, por lo mismo se incluyó la cantidad de desocupados y la cantidad que cotizan pensión. Finalmente se incluye si la jefe del hogar es mujer, que tendría acceso a opciones laborales de menor ingreso y menos tiempo para trabajar, y si el jefe tiene educación superior, que le brinda acceso a un mercado laboral diferente, siguiendo la siguiente ecuación:

$$\begin{aligned} Pobre = \beta_0 &+ \beta_1 \text{Clase} + \beta_2 \text{tot\_cuartos} + \beta_3 \text{nper} \\ &+ \beta_4 \text{npet} + \beta_5 \text{ndes} + \beta_6 \text{npension} \\ &+ \beta_7 \text{jefe\_mujer} + \beta_8 \text{jefe\_(educ\_sup)} + \mu \end{aligned}$$

Con este modelo se prueban los parámetros de $\alpha = 0, 0.2$ y 0.4 además de los parámetros λ = 100, 0, 0.01, 0.0001, 0.000001 con un fold de 5 para hacer cross-validation. El resultado del modelo no arroja ni un Ridge ni un Lasso, sino una combinación de ambos, ya que el parámetro α queda en el valor de 0.4. Adicionalmente, el mejor parámetro λ corresponde a 0.01. Los resultados de F1 de estos modelos fueron 0.41 con la submuestra de variables y 0.56 con la totalidad de las variables.

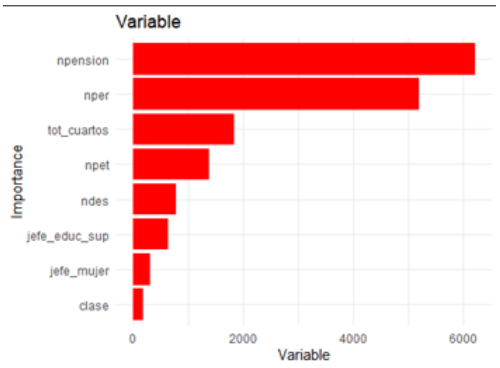
3.2.2. Random Forest

Se probaron 2 modelos de Random Forest. Inicialmente, uno con una submuestra de las variables mencionadas al inicio, y otro con todas las variables, de la misma forma que se probó para Elastic Net, que incluía las variables como se muestra a continuación:

$$\begin{aligned} Pobre = \beta_0 &+ \beta_1 \text{Clase} + \beta_2 \text{tot\_cuartos} + \beta_3 \text{nper} \\ &+ \beta_4 \text{npet} + \beta_5 \text{ndes} + \beta_6 \text{npension} \\ &+ \beta_7 \text{jefe\_mujer} + \beta_8 \text{jefe\_(educ\_sup)} + \mu \end{aligned}$$

En este caso, se establecieron los hiperparámetros a través de un 5-fold cross-validation con los valores de: i) variables seleccionadas para cada submuestra (2, 4, 6, 8), siendo la última equivalente a bagging, ii) número mínimo de observaciones (10, 50, 100, 500, 1000) buscando optimizar el criterio Gini. El mejor resultado se encontró bajo una submuestra de 4 variables y un número mínimo de 100 observaciones por nodo. Adicionalmente, se tomó 500 como el número de Bootstrap simples y árboles a estimar. El resultado de F1 de los modelos de Random Forest fue de 0.49 con la submuestra de variables y de 0.36 con todas las variables.

Figura 2: Relevancia de las variables para el modelo de clasificación random forest



Buscando identificar las mejores variables, se revisó la importancia de las variables en el modelo de la submuestra, ya que el resultado de estos modelos no fue mejor que el de Elastic Net. Se procedió a elegir algunas de las que tenían mayor importancia para correr los modelos de boosting, como se presenta a continuación.

3.2.3. Boosting

Se probaron 3 modelos de ADABoost. Inicialmente, se hizo un cross-validation de 5 fold para probar los hiperparámetros de: i) cantidad de bifurcaciones (1, 2 y 5), ii) la cantidad de iteraciones o modelos (50, 300, 500), y iii) la tasa de aprendizaje (Breiman y Freud).
Se corrió teniendo en cuenta las variables que tenían mayor importancia en el modelo de Random Forest y una variación de las que ya se habían utilizado, teniendo en cuenta que algunas podían reflejar relaciones más importantes. El modelo fue el siguiente:

$$\begin{aligned} Pobre = \beta_0 &+ \beta_1 \text{tot\_(cuartos\_dorm)} + \beta_2 \text{tot\_cuartos} \\ &+ \beta_3 \text{nper} + \beta_4 \text{npet} + \beta_5 \text{ndes} \\ &+ \beta_6 \text{npension} + \beta_7 \text{nsalud} \\ &+ \beta_8 \text{jefe\_(educ\_sup)} + \mu \end{aligned}$$

Por ejemplo, se removió la variable de si la jefe del hogar es mujer, que bajo teoría económica parece ser relevante, pero que en el modelo de Random Forest tenía baja importancia, al igual que clase. Este primer modelo dio mejor bajo ROC con hiperparámetros de tasa de aprendizaje Breiman, 500 modelos y 5 bifurcaciones.

Con base en esto se probaron los siguientes dos modelos. Se editaron los hiperparámetros para correr dos modelos adicionales, pero bajo las mismas variables predictoras. Específicamente, se corrieron ambos con cross-validation de 5 fold con bifurcaciones de 5 o 10 y tasa de aprendizaje asociada a Breiman, pero con cantidad de modelos de 500 o 750 para el primer modelo, y de 750 y 950 para el segundo.

Siempre resultó mejor 5 bifurcaciones, pero la cantidad de modelos siempre daba mejor con la cantidad mayor (750 el primer modelo y 950 el segundo), por lo que se podría seguir incrementando la cantidad

para mejorar el modelo.

Los resultados de estos modelos bajo el F1 calculado con la submuestra de train nos daban entre 0.47 y 0.5, y subimos el mejor a Kaggle, que resultó en un F1 de 0.49.

3.3. Modelo final

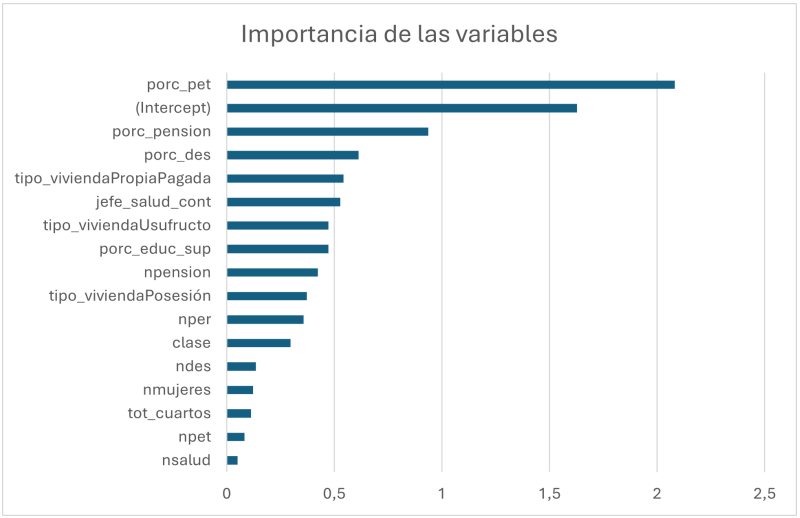
Por el lado de los modelos de clasificación, el que logró un mejor resultado en términos de F1 fue el Elastic Net con todas las variables. La ecuación del modelo Elastic Net es la siguiente:

$$\begin{aligned} Pobre = \beta_0 &+ \beta_1 \text{Clase} + \beta_2 \text{tot\_cuartos} + \beta_3 \text{tot\_(cuartos\_dorm)} \\ &+ \beta_4 \text{tipo\_vivienda} + \beta_5 \text{cuota\_amort} + \beta_6 \text{cuota\_arriendo} \\ &+ \beta_7 \text{nper} + \beta_8 \text{lin\_indig} + \beta_9 \text{lin\_pobreza} \\ &+ \beta_{10} \text{Ingtotug} + \beta_{11} \text{nmujeres} + \beta_{12} \text{nsalud} \\ &+ \beta_{13} \text{neducsup} + \beta_{14} \text{npension} + \beta_{15} \text{npet} \\ &+ \beta_{16} \text{ndes} + \beta_{17} \text{jefe\_mujer} + \beta_{18} \text{jefe\_(salud\_cont)} \\ &+ \beta_{19} \text{jefe\_(educ\_sup)} + \beta_{20} \text{porc\_mujer} + \beta_{21} \text{porc\_salud} \\ &+ \beta_{22} \text{porc\_(educ\_sup)} + \beta_{23} \text{porc\_pension} \\ &+ \beta_{24} \text{porc\_pet} + \beta_{25} \text{porc\_des} + \mu \end{aligned}$$

Se muestra a continuación el peso que le da a cada una de las variables cuyos pesos son mayores a 1. El resto de las variables incluidas en el modelo y que no aparecen en la gráfica, el modelo las arrojó a cero.

Para corregir desbalances en la base, se intentó truncar la base de datos de train, para que se contara con la misma cantidad de observaciones clasificadas como pobre, ya que la base en general contaba con muchos mas datos clasificados como no pobre. El resultado intentando mejorar elastic net con todas las variables fue inferior, logrando un F1 en kaggle de 0.39.

Figura 3: Gráfico de relevancia de las variables



Los parámetros utilizados corresponden a una unión de Ridge y Lasso, con un α de 0.4. y adicionalmente, se tiene un λ de 0.01. Se llegó a este modelo con una submuestra del 70 % de los datos de train, para calcular un F1 preliminar y se intentó mejorarla truncando la base de datos para tener la misma cantidad de hogares no pobres que de hogares pobres (aprox 33000), pero el resultado de F1 subido a kaggle bajo de 0.56 a 0.39, estando aún por debajo del adaboost, por lo que quedó seleccionado el elastic net con todas las variables.

Para el caso de modelo de predicción, el modelo que logró un mejor resultado en relación al F1 fue el de regresión lineal de la forma:

$$\begin{aligned}\text{Ingreso} = \beta_0 &+ \beta_1 \text{nper} + \beta_2 \text{clase} + \beta_3 \text{nmujeres} + \beta_4 \text{nsalud} \\ &+ \beta_5 \text{neducsup} + \beta_6 \text{sub\_transporte} + \beta_7 \text{otros\_ing} + \beta_8 \text{ayu\_inst} \\ &+ \beta_9 \text{ndes} + \beta_{10} \text{tot\_cuartos} + \beta_{11} \text{tipo\_vivienda} + \beta_{12} \text{tiempo\_con} + \mu\end{aligned}$$

Con los resultados de predicción de los ingresos, se clasificó en pobre y no pobre los hogares en relación a la media de ingresos de los hogares pobres (excluyendo percentil y percentil 95). De esta forma, se obtuvo un F1 de 0.32. Aunque se intentó mejorar el resultado con otros ajustes, como lo fue Elastic Net, booting y random forest, el resultado fue inferior a 0.3. Con estos resultados, el mejor modelo entre los de clasificación y regresión, resulto ser el de clasificación, probablemente por los efectos que tiene la base desbalanceada en los de regresión.

4. Conclusiones

En resumen, el mejor modelo de clasificación resulta ser elastic net con todas las variables, que incluye proporciones tomadas de la base de train individuos, y permite darle mayor información sin generar overfit, que parece que solo favorece a elastic net y no tanto a los modelos de random forest y boosting. Se cree que el incremento de variables favoreció al modelo de elastic net y a los otros no, ya que tiene más opciones de donde elegir para la regularización. Adicionalmente, como se incluyen las proporciones que surgen de los datos de individuos, algunas de estas logran darle más fuerza el modelo de elastic-net. Para el modelo de regresión, el que demostró mayor eficiencia fue el de regresión lineal, dada su sencillez y capacidad para adaptarse a los datos, y predecir de forma indirecta la condición de pobreza de los hogares.

El performance de los modelos de regresión resultaron peores que los de clasificación, que puede deberse al desbalance de la base que afecta en mayor medida el poder predictivo de una regresión vs. el de clasificación.

5. Link de GitHub

Para replica el ejercicio, puede acceder al repositorio, donde encontrará las líneas de código necesarias. Recuerde que puede hacer uso de $README$ como instructivo para navegar dentro del git.