Taller 2: Predicting Poverty
Big data and Machine learning for Economics

Alison Gissell Ruiz Ruiz - Código 202116230
John Daniel Delgado Vargas - Código 202225721
José Julián Parra Montoya - Código 202213144

1. Introducción

Según el Banco Mundial, la pobreza monetaria, en el ámbito individual y del hogar, tiene por lo menos tres dimensiones: una económica, una social y una demográfica (Mundial (2005)). La dimensión económica se refiere al acceso al mercado de trabajo, tanto en el margen intensivo como en el margen extensivo, y la capacidad de acumular capital (tanto físico como humano). La dimensión demográfica se relaciona a la estructura del hogar: el género del jefe del hogar, el perfil de edad de los integrantes, y en general, condiciones que afecten la capacidad de los integrantes del hogar de recibir un ingreso. Finalmente, la dimensión social se refiere al acceso a bienes y servicios básicos para el bienestar humano.

Aunque estas dimensiones son intuitivas y útiles, al pensar la pobreza teóricamente, la capacidad de verificar estar información en un hogar particular representa un reto práctico. En 2016, el Banco Mundial estimó en 945 millones de dólares el costo de la realización de su programa de encuestas de hogares entre 2016 y 2030 (Mundial (2005)). Dos terceras partes de este monto se dedican a la implementación de las encuestas y el restante a la asistencia técnica necesaria para procesarlas.

Es evidente que ganancias en eficiencia tanto en el proceso de recolección de datos como en procesamiento pueden ayudar a disminuir el costo total del proceso. Ambas cosas están relacionadas con el volumen de información que debe recolectar cada encuestador de los hogares. Si se pudiera disminuir esta información sin perder precisión en la capacidad de identificar a un hogar como pobre, los costos de la implementación de estos programas disminuirían permitiendo una adopción más amplia y por tanto mejores instrumentos de focalización de políticas para combatir la pobreza.

En este documento se identifica que el uso del algoritmo *Extreme Gradient Boosting* (XGBoost) tanto para seleccionar las variables importantes como para realizar la predicción permite obtener las mejores predicciones en la condición de pobreza del ingreso y en el ingreso de los individuos, con 5 y 4 variables, respectivamente. A continuación se describirá el proceso que se realizó para obtener esta conclusión. Inicialmente se realiza una exploración de las variables utilizadas, seguidamente se muestran los resultados de aplicar diferentes métodos de selección de variables (*Recursive Feature Elimination* y Regularización) a diferentes algoritmos, y finalmente se concluye con el mejor algoritmo.

2. Análisis descriptivo

2.1. Características individuales

Los datos que se emplearán consisten en datos provenientes de la Gran Encuesta Integrada de Hogares (GEIH) a nivel individual y a nivel de hogar. En los datos a nivel de individuo se cuenta con 62 variables y a nivel de hogar con 23. A continuación se realiza un análisis de estadísticas descriptivas básicas de estas variables. En el cuadro **??** se presentan las estadísticas descriptivas para las variables continuas. Allí se observa, en la primera fila, que la variable de ingreso total es una variable con una distribución sesgada hacia la derecha pues su media ($775,591) está a la derecha de su mediana ($430,000); este es un rasgo común de las distribuciones de ingreso. Su desviación estándar es de aproximadamente dos veces la media ($1'380,447). También se pueden caracterizar al 10 % de los individuos más ricos de la distribución: son aquellos que reciben más de $1'800,000. En la segunda fila se registra el comportamiento de la edad. La distribución de la edad, al igual que la del ingreso, está sesgada hacia la derecha pues su media (33 años) está a la derecha de su mediana (31 años). La desviación estándar es aproximadamente 0.75 veces su media (21 años). También se puede identificar el último decil de la distribución: el 10 % de los individuos de mayor edad son aquellos mayores a 64 años.

Cuadro 1: Estadísticas descriptivas variables continuas

	Media	Mediana	D.E.	Percentil 90
Individuos				
Ingreso total	775591.56	430000.00	1380447.03	1800000.00
Edad	33.55	31.00	21.64	64.00
Horas Trabajo Semana	44.79	48.00	15.72	60.00
Hogares				
Estim. Arrendamiento	513069.70	350000.00	5046531.07	800000.00
Núm. Cuartos	1.99	2.00	0.90	3.00
Núm. Personas	3.29	3.00	1.77	5.00

Finalmente, la distribución de las horas trabajadas por semana también muestra un sesgo pero en la dirección opuesta a las distribuciones anteriores pues su media (44 horas) está a la izquierda de su mediana (48 horas). La desviación estándar es aproximadamente una tercera parte de su media (15 horas). Así mismo, se puede identificar que el 10 % de los individuos que más trabajan son aquellos que lo hacen más de 60 horas a la semana.

En cuanto a las variables continuas a nivel de hogar, el cuadro 1 muestra que la estimación del arrendamiento que hacen los individuos tiene, al igual que el ingreso a nivel individual, una distribución asimétrica con un sesgo hacia la derecha, pues la media ($513,069) está a la izquierda de la mediana ($3500,000). Así mismo, es una variable con mayor dispersión que esta (aproximadamente 10 veces su media), y la masa de la distribución está ligeramente hacia la izquierda pues el percentil 90 corresponde a tan solo $800,000. También se puede observarque la distribución del número de cuartos es simétrica pues su media (1.99) coincide con su mediana (2). Su desviación estándar es aproximadamente la mitad de la media, y esta poca desviación se refleja también en su rango de valores: el 10 % de familias con más cuartos tan solo tiene más de tres.

Finalmente, se puede observar que el número de miembros del hogar tiene una distribución ligeramente asimétrica. Así mismo se observa que su dispersión es aproximadamente

la mitad de su media, y por tanto, el percentil 90 no está muy lejos: el 10 % de hogares con mayor número de individuos tiene más de 5..

En el cuadro 2se presentan las estadísticas descriptivas de algunas de las variables categóricas. En particular, se limita al promedio o la frecuencia de cada categoría. En la primera fila puede observarse que los existe un porcentaje ligeramente mayor de hombres (1) en los datos. Así mismo, en la fila dos, se observa que estos individuos son en su mayoría jefes de hogar (1) o hijos de los jefes de hogar (3) o sus cónyuges (2).

En cuanto a su seguridad social, se registra que los tipos de regímenes de salud más comunes son el contributivo (1) y el subsidiado (2.). También se tiene que la mayoría de ellos no es formal (2), es decir, no cotiza a un fondo de pensiones. Los individuos, en su mayoría, tienen por lo menos primaria (3). Seguidos de individuos con estudios universitarios (6) e individuos bachilleres (5). También se caracterizan por trabajar solos (1) o en compañías de más de 100 empleados (9). Y en cuanto a sus ingresos, la mayoría no recibe arrendamientos (2) ni pensión alimenticia (2). Finalmente, la mayoría no están en hogares pobres (0).

Cuadro 2: Estadísticas descriptivas variables categóricas

	Media		Media
Individuos		***Hogares***	
Género		*Pobre*	
1	0.48	1	0.20
2	0.53	0	0.80
Reg. Seg. Soc.			
1	0.47		
2	0.05		
3	0.47		
9	0.00		
Parent. Jefe de Hog.			
1	0.30		
2	0.16		
3	0.35		
4	0.08		
5	0.08		
6	0.00		
7	0.00		
8	0.00		
9	0.02		
Nivel Educ.			
1	0.05		
2	0.03		
3	0.26		
4	0.18		
5	0.23		
6	0.25		
9	0.00		
Núm. Emple. Emp.			
1	0.39		
2	0.16		
3	0.06		
4	0.05		
5	0.03		
6	0.03		
7	0.02		
8	0.02		
9	0.24		
Formal			
1	0.38		
2	0.60		
3	0.02		
Recibe Arrend.			
1	0.09		
2	0.91		
Recibe Cuot. Aliment			
1	0.06		
2	0.94		
9	0.00		

3. Selección de variables usando *Random Forest*

Random Forest (RF) es un algoritmo que busca reducir la variación del algoritmo *Decision trees* mediante el empleo de técnicas de remuestreo. Este algoritmo constituye no solamente un algoritmo competente para problemas de regresión y clasificación, si no que también puede utilizarse como un método automático de selección de variables (Genuer et al. (2010)). Una metodología que permite realizar dicha selección es denominada *Recursive Feature Elimination* (RFE). Esta consiste en lo siguiente: en la primera iteración se ajusta un RF con todas las variables predictoras disponibles para obtener un ordenamiento según su importancia. Posteriormente se ajusta un RF para un subconjunto de las primeras S_i variables obtenidas en el ordenamiento y se ajusta el ordenamiento. Finalmente se realiza un ordenamiento de los subconjuntos obtenidos en el procedimiento anterior según los resultados generales.

El algoritmo RFE es entonces una aplicación del algoritmo *Backward Stepwise Selection* (BSS) utilizando bosques aleatorios. A continuación se aplica este algoritmo al problema de regresión. Inicialmente se toma un Subconjunto de 15 variables de las 62 posibles. Estas variables se escogen de manera tal que exista un número razonable de observaciones que no tengan datos faltantes (de tomarse las 62 no logra obtenerse un subconjunto de filas tal que cada variable tenga un valor no faltante) y que intuitivamente tengan sentido. Adicionalmente la variable "Oficio" se descarta debido a la dificultad computacional que trae su alto número de categorías.

Usando este conjunto de 15 variables se emplea el algoritmo RFE en una muestra aleatoria de una quinta parte de las observaciones (para hacerlo tratable computacionalmente). Este procedimiento arroja un subconjunto de 5 variables que permiten minimizar la Raíz Cuadrada del Error Cuadrático Medio (RMSE) las cuales son: el género ("P6020"), la edad ("P6040"), el nivel educativo ("P6210"), las horas de trabajo en la ocupación principal durante la última semana ("P6800") y si el individuo recibe pago por arrendamiento o pensión ("P7495"). Estas variables se emplean como argumentos de dos algoritmos de regresión cuyo resultado consiste en la generación de una predicción para el ingreso individual. Una vez esta predicción se agrega a nivel de hogar, se puede obtener el ingreso per capita y al ser comparado con la línea de pobreza puede obtenerse la predicción respecto a la condición de pobreza del hogar. Los resultados de la aplicación de este procedimiento se presentan en el cuadro 3.

Cuadro 3: Resultados RFE Regresión

Modelo	(0.75)FNR+(0.25)FPR	Número de variables
Regresión Lineal	0.4746745	5
XGBoost	0.4221759	5

Como puede observarse, el algoritmo *Extreme Gradient Boosting* (XGBoost) presenta los mejores resultados utilizando una función de pérdida personalizada que consiste en una suma ponderada entre la tasa de falsos negativos (FNR) y la tasa de falsos positivos (FPR).

Seguidamente, se considera el mismo problema de predicción directamente sobre la condición de pobreza del hogar, lo que configura un problema de clasificación. Tras aplicar la metodología RFE se encuentra que el subconjunto óptimo de variables para explicar la condición de pobreza son: la edad promedio, la edad del jefe del hogar, el ingreso promedio y el número de habitantes del hogar. Los resultados del uso de estas variables en diferentes algoritmos se reportan en el cuadro 4. En primer lugar se emplean los algoritmos Probit y

Logit sin la aplicación de ninguna técnica de remuestreo. Seguidamente estos algoritmos se entrenan en muestras de entrenamiento sometidas a procesos de remuestreo (SMOTE y *Undersampling*) para tratar los problemas de balancero de clases. Finalmente se utiliza el algoritmo XGBoost. Se encuentra que el modelo Probit con una muestra de entrenamiento sometida a SMOTE produce los mejores resultados.

Cuadro 4: Resultados RFE Clasificación

Modelo	(0.75)FNR+(0.25)FPR	Número de variables
Probit	0.7262859	4
Logit	0.7261416	4
Probit (SMOTE)	0.6828354	4
Logit (SMOTE)	0.6839173	4
Probit (Undersampling)	0.6823886	4
Logit (Undersampling)	0.6835628	4
XGBoost	0.7256521	4

4. Selección de variables mediante regularización

Se realizan los métodos de regularización Ridge y Lasso para lograr determinar cuales son las variables más importantes para la predicción. Para Lasso, se destacan 4 variables como las más importantes, estas son: edad de trabajar (Pet) que no tiene variación en sus datos, por lo tanto no se tiene en cuenta para entrenamiento; si el individuo recibe pago por arrendamiento o pensión ("P7495"); Número de miembros del hogar ("Nper"); y el número de compañeros de trabajo ("P6870"). Se realizaron diversas pruebas agregando variables de menos importancia y se encontró que no existía variación significativa.

Para el caso de Ridge, se tiene adionalmente la variable de número de personas en la unidad de gasto ("Npersug"). Esta se tiene en cuenta para entrenar y se repiten las pruebas realizadas con Lasso, pero no se encuentra mejores al agregar otras variables al modelo. Los resultados obtenidos se presentan en el cuadro 5.

Cuadro 5: Resultados Regularización Regresión

Modelo	(0.75)FNR+(0.25)FPR	Número de variables
Regresión Lineal (Especificación Ridge)	0.3520840	4
Regresión Lineal (Especificación Lasso)	0.3550469	3

5. Selección de variables mediante Boosting

Los algoritmos de Boosting tienen como base otros modelos, mejorándolos al hacer iteraciones dependientes donde se penalizan los errores de cada una de manera que en la siguiente se tenga un ajuste diferente. El uso de estos algoritmos está ganado popularidad en las competencias de Machine Learning, como sugiere el portal *Synced Review*. Debido a esto se empleará la variación del Boosting conocida como XGBoost inicialmente para la selección de variables y posteriormente para la predicción de la pobreza.

El primer paso es el pre-procesamiento de los datos. Para esto se relizan dos pasos. El primero es tomar la tabla de hogares conviertiendo en binarias las variables categóricas;

el segundo consiste en el tratamiento de los datos por individuo. El dataset de test de personas cuenta con 63 variables de las cuales únicamente 11 son variables continuas, las otras 52 consisten en variables categóricas. Este ejercicio deja un total de 344 variables en el set de datos a trabajar.

Como los datos se encuentran desbalanceados, es necesario equilibrar las categorías para optimizar el aprendizaje orientado al objetivo del algoritmo que es predecir la pobreza. Se realiza un *Upsampling* para no perder información de los No Pobres e igualar la importancia para el algoritmo con los Pobres.

A continuación se busca determinar las variables más importantes. Para esto se definieron parámetros base con nrounds alto. Al finalizar el proceso se obtiene el valor de importancia de las variables, y por validación cruzada se empiezan a poner cada una de la variables en orden de importancia hasta obtener el número óptimo de variables que minimizan la función de pérdida objetivo.

El número óptimo de variables para clasificación fue 5, y para regresión fue 4. Las variables de mayor importancia para el modelo de clasificación fueron ("P5130") Arriendo potencial, ("P6100") seguridad social, ("P6800") horas trabajadas, ("P6920") pensión y ("P7500s3") otros ingresos. Estas variables hacen mucho sentido, ya que las horas trabajadas, pagando seguridad social y presión en un hogar muestran que hay mayor formalidad en el trabajo y esto reduce la probabilidad de estar por debajo del umbral de la pobreza. El mismo proceso con el modelo de regresión usa 3 de estas variables y agrega una nueva variable ("P5140") Arriendo pagada.

Con las variables identificadas quitando complejidad al algoritmo se realizó la optimización de hiperparametros mediante *grid search*. La profundidad óptima del árbol fue de 4. A medida que se incrementaba la profundidad del árbol aumentó el overfiting. Los resultados de los modelos entrenados se presentan en el cuadro 6.

Cuadro 6: Resultados XGBoost Regresión y Clasificación

Modelo	(0.75)FNR+(0.25)FPR	Número de variables
XGBoost (Clasificación)	0.11	5
XGBoost (Regresión)	0.31	4

Al comparar el resultado de este algoritmo con el resultado de las secciones anteriores, se puede concluir que esta metodología de selección de variables y de entrenamiento del algoritmo presentó los mejores resultados tanto en la predicción de la condición de pobreza como de ingreso de los individuos.

6. Conclusiones

A nivel general, si se emplean modelos sencillos como regresión lineal, Probit y Logit, se puede observar que al abordar el problema como un problema de regresión, la función de pérdida basada en la tasa de falsos positivos y falsos negativos da errores mucho más bajos, por lo cual, si se emplean modelos básicos es más favorable predecir el ingreso de cada persona y compararlo con la linea de pobreza que realizar un modelo de clasificación.

El XGBoost arrojó los mejores resultados dentro de los algoritmos probados, ya que al tener una estructura más compleja, permite capturar no linealidades en los datos, teniendo una ventaja muy considerable al abordar el problema como clasificación frente a Logit y Probit.

Los entrenamientos realizados con variables extraídas por medio de Lasso y Ridge tienen mejores resultados que tomando como base las variables arrojadas por RFE, sin embargo, ninguno de estos métodos supera el uso de XGBoost para seleccionar las variables y realizar la predicción.

Referencias

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236.

Mundial, B. (2005). Poverty manual. *World Bank Institute*.