Problem Set 2: Predicting Poverty

Paula Ramos, Karen Uribe-Chaves y Juan D. Urquijo

July 12, 2022
Repository Link:*Github*

1 *Introduction*

Atacar la pobreza se ha convertido en un objetivo fundamental a la hora de implementar políticas públicas. Según el Banco Mundial (2017) "los niveles de pobreza a menudo se estiman a nivel de país extrapolando los resultados de las encuestas realizadas en un subconjunto de la población a nivel de hogar o individual", con tres objetivos principalmente: i) Identificar los hogares pobres, ii) Mapear la pobreza y iii) Predecir los hogares pobres cuando no se tiene suficiente información recolectada. Por esto, se hace relevante que las decisiones de política pública se realicen basadas en modelos de predicción que permitan identificar los hogares adecuadamente.

La predicción de la pobreza se puede hacer con modelos de variables dependientes continuas o categóricas, como los modelos de regresión (*OLS*) o de clasificación (*Logit*). Este trabajo explora estos dos tipos de modelos predictivos, por medio de la Gran Encuesta Integrada de Hogares (GEIH) del año 2018 en Colombia. Esta encuesta tiene cobertura nacional que permite obtener resultados por zona urbana y rural, grandes regiones y total por departamento. Además, contiene información a nivel de hogares y de personas, lo que permite construir modelos más robustos respecto a la situación de pobreza de los hogares colombianos.

Estamos comparando un total de 16 modelos, 5 con una variable dependiente continua (ingresos) y 11 con una variable dependiente dicotómica (pobre/no pobre), centrándonos en la predicción fuera de la muestra. El objetivo es minimizar la diferencia entre la verdadera tasa de pobreza de los hogares estimada y la tasa de pobreza pronosticada, dando más relevancia a identificar los hogares que sí son pobres. La solución es el uso de dos modelos, después de comparación, que aplican la metodología *Lasso* como método de regularización. Los resultados arrojan que el 34.9% son pobres bajo el modelo de clasificación de *Logit Down-sampling*, mientras que únicamente se logran identificar el 0.13% como pobre bajo la metodología de un modelo de regresión *Lasso* a partir del ingreso del hogar. Se concluye que, los modelos de clasificación lograron un mejor comportamiento en la predicción de pobreza, intuitivamente por el uso de mayor cantidad de determinantes comparado con el modelo de regresión.

2 *Data*

El uso de la GEIH de 2018 nos permite identificar los hogares pobres por medio del ingreso o a través de la probabilidad de ser pobre. Dado que las predicciones son a nivel de hogar, construimos variables para capturar información a nivel de hogar utilizando datos de individuos. Las principales variables utilizadas y su descripción pueden encontrarse en el Anexo 1 de este trabajo. Se analizan las características a nivel de hogar (i.e Vivienda Propia, Tamaño del Hogar, etc) y a nivel del Jefe del hogar (i.e Edad, Educación, uso de servicios financieros, etc.), que resumen los determinantes de la pobreza en Colombia para 2018.

En particular, el empleo, la educación, la salud, el género, etc.; diferenciados por zona serán fundamentales para analizar las predicciones de pobreza realizadas en este trabajo. Finalmente, para los modelos de regresión, el ingreso del hogar se tomó desde la variable "Ingtotugarr" que hace referencia al ingreso total del hogar con imputación de arriendo a propietarios y usufructuarios, lo que permite capturar la totalidad de los ingresos de un hogar - sin importar si vive en vivienda propia o arriendo - con el fin de homogenizar la muestra, tal y como lo plantea el DANE (2018).

Por otro lado, realizamos un análisis descriptivo de la base de entrenamiento, con el fin de entender su distribución y la cantidad de pobres en el país medido por línea de pobreza. Las características del hogar se pueden ver en la Table 1. Se observa que en promedio los hogares tienen un tamaño de 3.3 personas, con un máximo de 28 personas en un mismo hogar y un mínimo de 1. Por su parte,

en la muestra se encuentra que el ingreso promedio de un hogar es de $2,305,654 con un máximo de $88.833.333 y un mínimo de $0, lo que podría indicar tentativamente la presencia de valores atípicos en la distribución. Relacionado con los miembros del hogar, se observa que en promedio el jefe del hogar tiene 50 años, el número de mujeres en el hogar es 2 y el número de ocupados es menor que 2 personas, mientras que el número de afiliados a salud por hogar es de 2.5, en promedio. Finalmente, la educación del jefe del hogar alcanza en promedio 6 años.

Table 1: Estadísticas Descriptivas Variables Numéricas - Por Hogar

Statistic	N	Mean	St. Dev.	Min	Max
Número Personas	164,959	3.295	1.777	1	28
Ingreso Total	164,959	2,305,654.000	2,621,124.000	0.000	88,833,333.000
Edad Jefe Hogar	164,959	49.605	16.385	11	108
Núm. Mujeres	164,959	1.739	1.179	0	14
Núm. Ocupados	164,959	1.508	1.030	0	12
Núm. de menores de edad	164,959	0.920	1.122	0	15
Ingreso Total Jefe Hogar	164,959	1,222,168.000	1,778,210.000	0.000	85,833,333.000
Máx.Educación Jefe Hogar	164,959	6.102	3.718	0	99
Núm. de afiliados a salud	164,959	2.531	1.421	0	17

En el análisis de la información, construimos una gráfica (Figure A1 del Anexo) que nos permita identificar los hogares pobres previo al tratamiento de datos. Lo anterior, debido a que que en los modelos de clasificación es crucial conocer el equilibrio de las etiquetas de clase, para evitar sesgos en las predicciones y la correcta interpretación de la precisión de las predicciones de los modelos. Se encuentra que en la base de entrenamiento, el 20% de la población es pobre, lo que representa 33,023 hogares e indica una muestra claramente desbalanceada. Por su parte en la Figure A2 del Anexo, demuestra la dispersión de los datos. En la muestra, se observa que los hogares clasificados como No Pobres (Pobre = 0) tienen un mayor ingreso, y su media del ingreso se encuentra por encima que la de Pobres (Pobre = 1). Realizando la transformación logarítmica del ingreso, se observa que existe una persona pobre que no recibe ingreso ($0), lo que coincide con la Table 1. Por otro lado se observa una gran dispersión de los datos, especialmente en los hogares No Pobres.

3 *Models and Results*

3.1 Classiffication Models

Se elaboraron 5 modelos logit (Tabla A.1), en los cuales la variable a predecir es la probabilidad de cada hogar de ser pobre. Posterior a ello, se hallaron las predicciones, se definió en principio la regla de 0,5 con el fin de determinar el punto de corte para las matrices de confusión y se calcularon los hiperparámetros para elegir el mejor modelo.

Table 2: Comparación Modelos Logit

	Sensitivity	Specificity	PosPredValue	NegPredValue	Precision	Recall	F1
Modelo 1	0.2740	0.9648	0.6606	0.8415	0.6606	0.2742	0.3873
Modelo 2	0.2735	0.9648	0.6604	0.8414	0.6604	0.2735	0.3868
Modelo 3	0.2592	0.9679	0.6693	0.8392	0.6693	0.2592	0.3737
Modelo 4	0.0030	0.9983	0.3043	0.8000	0.3043	0.0030	0.0059
Modelo 5	0.2737	0.9648	0.6608	0.8415	0.6608	0.2737	0.3871

De acuerdo con la Tabla 2, el mejor modelo es el 1, bajo la siguiente especificación:

$$\begin{aligned} Pobre = ViviendaPropia + TotalMujeresHogar + JefeHogarMujer \\ + Núm.OcupadosHogar + EdadJefeHogar + \\ NúmerodeMenores + Máx.EducaciónJefeHogar + \\ JefeHogarOcupado + Núm.AfiliadosSalud + UsoProductosFinancierosJefeHogar \end{aligned} \quad (1)$$

En este modelo, la métrica F1 (Media armónica de Precision y Recall) es mayor para el modelo 1. Para complementar la interpretación de este parámetro, cabe resaltar que la métrica Precision permite concluir que, cuando el modelo predice que un hogar es pobre acierta el 66,06% de las veces, tomando

como referencia los verdaderos y los falsos positivos. En relación con Recall, el modelo identifica correctamente el 27,42% de los hogares, en relación con los verdaderos positivos y falsos negativos. Adicionalmente, el modelo 1 tiene el mayor valor de área bajo la curva: 0.7915 (Figure A.3).

Con el fin de mejorar la capacidad predictiva del modelo 1, se estimó nuevamente utilizando las siguientes metodologías: Logit-Cross validation, Logit-Lasso Accuracy, Logit-Lasso ROC, Logit-Lasso Sensitivity, Up-sampling y Down-sampling, como se evidencia en la Tabla A.2. En este proceso se halló que el mejor modelo fue Downsampling con un parámetro de ROC (0.7913).

Una vez elegido el modelo a través del cual se mejoró el equilibrio entre clases a través de Down-sampling, se procede a determinar el cut-off con la base de evaluación, hallando el punto de la curva ROC más cercano a la esquina superior izquierda del gráfico - Threshold de 0.526 (Table A.3). Sin embargo, se prefiere el criterio indicado incialmente de 0.5, porque la predicción le atina a más hogares que son verdaderamente pobres.

Se estiman los 6 modelos en la base de testing, seleccionando el modelo Downsampling, de acuerdo con la predicción que acierta una mayor cantidad de pobres (25% del total de hogares pobres) y una mayor cantidad de hogares que no son pobres(29% del total de hogares no pobres). Aunque es posible que nos haga falta una variable que nos permita mejorar la predicción, por último, se predice con este modelo en la base test, hallando un 34.9% de hogares pobres con respecto al total de hogares.

3.2 Income regression Models

Se elaboraron 5 modelos de regresiones lineales para predecir el ingreso con la muestra ya expuesta de entrenamiento. Se escogió el modelo con menor MAE (*Mean Absolut Error*) para posteriormente compararlo con la linea de pobreza (Lp) dada para cada hogar. Así, se le asignó como pobre a los ingresos predichos por debajo de Lp y no pobre en el caso contrario, de acuerdo a la ecuación (2).

$$Pobre = I(Ing < Lp) \tag{2}$$

Para la definición de la forma funcional de cada modelo se usaron las variables especificadas en el apéndice A1. En ese orden de ideas:

- El Modelo 1 se estimó por *OLS*
- el Modelo 2 se estimó usando el método de regularización de *Ridge*
- El Modelo 3 se estimó usando el método de regularización de *Lasso*
- El Modelo 4 se estimó usando la técnica de *k-fold Cross validation* con un k=5 sin el efecto marginal de la edad al cuadrado
- El Modelo 5 se estimó de la misma manera que el Modelo 4 incluyendo el efecto marginal de la edad al cuadrado.

La forma base de los modelos estimados para el ingreso de los hogares fue:

$$\begin{aligned} IngresoDelHogar = Dominio + ViviendaPropia + PersonasEnElHogar \\ + TotalMujeresHogar + JefeHogarMujer + Núm.OcupadosHogar \\ + EdadJefeHogar + EdadJefeHogar^2 + NúmerodeMenores + Máx.EducaciónJefeHogar \\ + JefeHogarOcupado + UsoProductosFinancierosJefeHogar + Núm.AfiliadosSalud \end{aligned} \tag{3}$$

Los MAE para cada modelo fue: Modelo 1 ($1,317,009); Modelo 2 ($1,299,164); Modelo 3 ($1,294,976); Modelo 4 ($1,317,472) y Modelo 5 ($1,317,388).

El de mejor comportamiento fue el modelo que usó la regularización de *Lasso* (Modelo 3), donde cabe mencionar que en un principio se incluyeron todas las variables posibles para estimar, que finalmente arrojó un modelo con siete (7) variables (Número de cuartos en el hogar, Vivienda Propia, Jefe Hogar Mujer, Núm. Ocupados Hogar, Número de Menores, Máx. Educación Jefe Hogar y Uso Productos Financieros Jefe Hogar). En el Apéndice, Figura A.4, se puede ver el comportamiento de los coeficientes de todas las variables en función del lambda de este proceso de regularización.

Como resultado final, podemos ver en la Tabla 3 el calculo de pobres según Línea de Pobreza con los datos originales en la base de entrenamiento; las predicciones del mejor modelo en la base de entrenamiento y las predicciones del mejor modelo en la base de prueba.

Table 3: Pobres según Línea de Pobreza

	No	Si
Pobres Lp: train	159,592	5,367
Pobres Lasso: train	164,750	209
Pobres Lasso: test	66,082	85

En conclusión, los modelos lineales estudiados no son una buena alternativa a la hora de predecir variables categóricas (Pobre o No Pobre) que dependen de la correcta predicción del ingreso del hogar. Esto se puede deber a los valores atípicos observados en la distribución del ingreso y al desbalance que existe entre las clases de la variable que se quiere predecir. Al predecir el ingreso de los hogares para luego compararlo con la linea de pobreza utilizando el modelo con menor MAE en la base de testing, se obtiene que solo el 0.13% de los hogares son pobres; esto podría sugerir que el modelo está sobrestimando el ingreso de los hogares.

4 *Conclusions and recommendations*

4.1 Conclusiones

Las predicciones de pobreza encontradas dan cuenta de la dificultad de medir y predecir los índices de pobreza a nivel mundial. La heterogeneidad de los resultados encontrados, brinda evidencia frente a la cual la inclusión o exclusión de variables, afectan las métricas de manera relevante; y la definición de la forma funcional respecto a los determinantes de la pobreza aún tiene espacio para su estudio.

En particular, se encontró que con los modelos de clasificación es posible identificar más hogares pobres dentro y fuera de muestra, comparados con los modelos de regresión lineal. Para el modelo de *Logit Down-Sampling* las predicciones finales arrojaron un 35% de hogares pobres; por lo cual son preferibles respecto al objetivo de reducir el error tipo 2 (falsos negativos) o, en otras palabras, encontrar la mayor cantidad pobres. En contraste, cabe resaltar que a través de la clasificación de los hogares pobres a través del ingreso por debajo de la línea de pobreza, se identifica que el 3,25% son hogares pobres. Por esta razón, el modelo de regresión lineal planteado no tuvo un porcentaje de predicción adecuado, clasificando como pobres únicamente al 0.13% de los hogares de la muestra de prueba.

4.2 Recomendaciones

A partir de lo encontrado, consideramos que para los modelos de predicción de pobreza es relevante definir la función de optimización, en función de los objetivos específicos del Estado. En este caso, se sobre ponderó la importancia de identificar los pobres (reducir error tipo 2), que pueden implicar la preferencia por maximizar coberturas de programas para la reducción de la pobreza, pero menor optimización de los gastos asociados (No pobres clasificados como pobres). Ante un cambio en las preferencias u objetivos, la comparación de los modelos y la selección de los mismos cambiaría, afectando las predicciones finales.

Por otro lado, para futuras mediciones se propone calcular el ingreso a partir del uso de variables individuales con el fin de validar su impacto en las predicciones realizadas; y comparar las métricas del DANE y su idoneidad. Otra opción que ayudaría a mejorar las predicciones de pobreza, sería contar con más características físicas del hogar, el acceso a servicios de salud, la ubicación geográfica, así como los hábitos de consumo e inversión de sus integrantes.

5 Referencias

Gibson, John. (1999). "A Poverty Profile of Cambodia, 1999." Report to the World Bank and the Ministry of Planning, Phnom Penh.

Goulden, Chris (2010). "Cycles of poverty, unemployment and low pay". Joseph Rowntree Foundation. United Kingdom 2010.

Guevara, Diego (2005). "Dinámica de la pobreza en Colombia: análisis de los ingresos de jefes de hogar urbano 1984-2003". Revista Economía y Desarrollo, Volumen 4 Número 2. Universidad Autónoma de Colombia.

Haughton, Jonathan Khandker, Shahidur (2009). "Handbook on Poverty and Inequality". The World Bank. Washington, DC.

Núñez, Jairo y Espinosa, Silvia (2005). "No siempre pobres, no siempre ricos: vulnerabilidad en Colombia". Documento CEDE 2005-15.

A *Appendix*

1. Variables utilizadas y relevancia

- *Tamaño del Hogar*: Gibson (1999) encuentra que en Camboya los pobres tienden a vivir en hogares más grandes (con un mayor número de personas), lo que podría indicar que aquellos hogares con mayores densidades, sufren de una condición de pobreza crónica que los lleva a una trampa de pobreza. La pobreza en Colombia, según Núñez y Espinosa (2005) está relacionada con altas tasas de dependencia, hogares grandes.

- *Vivienda Propia*: Toma el valor de 1 si el individo tiene vivienda propia y 0 de lo contrario (i.e. en arriendo, usufructo, etc). Un gasto asociado a un arriendo, o la inestabilidad de no tener vivienda propia incide en la pobreza del hogar.

- *Total mujeres en el hogar y Jefe de Hogar Mujer*: El género del jefe del hogar influye sobre la condición de pobreza que atraviesa el hogar, específicamente si el jefe es mujer los hogares tienden a ser más pobres. (Guevera, 2005).

- *Número de Ocupados en el Hogar y si el Jefe del Hogar es ocupado*: El empleo o la posesión de activos parecen tener gran importancia dentro de los determinantes de la pobreza. Goulden (2010), encuentra que un individuo tiene más riesgo de caer en la pobreza si es desempleado, inactivo o trabajador por cuenta propia.

- *Edad del Jefe del Hogar y número de menores de edad*: La edad del jefe del hogar, que usualmente es el encargado económico, influye en su nivel de ingreso y por tanto en si el hogar es pobre o no. Por otro lado, entre más dependientes, medido por numero de menores de edad, puede generar menor ingreso total.

- *Educación Jefe del Hogar*: La educación, medida como el máximo nivel de educación alcanzado por el individuo o su asistencia a una institución educativa, está asociado con la capacidad de generar ingresos para mantener una calidad de vida adecuada.

- *Número afiliados a salud*: Intentamos capturar con esta variable el hecho que la mala salud de las personas pobres suele tener relación con su condición de pobreza. Las bajas capacidades de las personas pobres (mal estado nutricional, condiciones de vida peligrosas, incapacidad para acceder a tratamiento médico) implican que los choques de salud en los individuos pobres se repiten con una mayor frecuencia y del mismo modo, tardan más en recuperarse.

- *Uso Productos Financieros Jefe del Hogar*: Dado que un hogar pobre es vulnerable a un choque financiero, consideramos que esta variable permite capturar la capacidad de suavizar el consumo del hogar e incidir en su condición de pobreza.

- *Dominio*: Identificar la ubicación del individuo puede dar información sobre su condición de pobreza. Según Haughton & Khandker (2009), la pobreza es mayor en áreas caracterizadas por el aislamiento geográfico, una escasa base de recursos, baja probabilidad de lluvia y condiciones climáticas inhóspitas.

2. Gráficas Destacadas Estadísticas Descriptivas

Figure A.1: Clasificación Pobres - No Pobres - Colombia (2018)

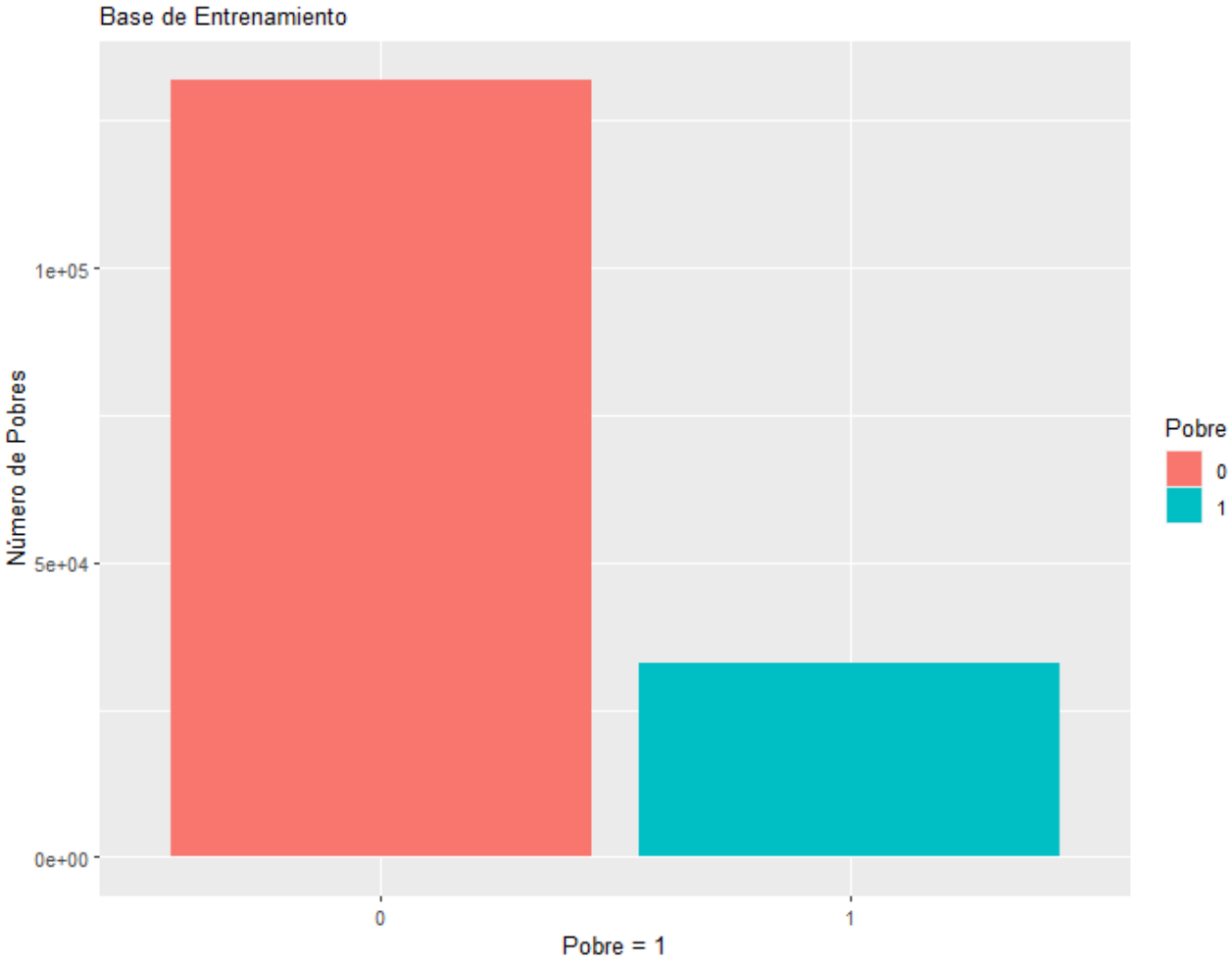
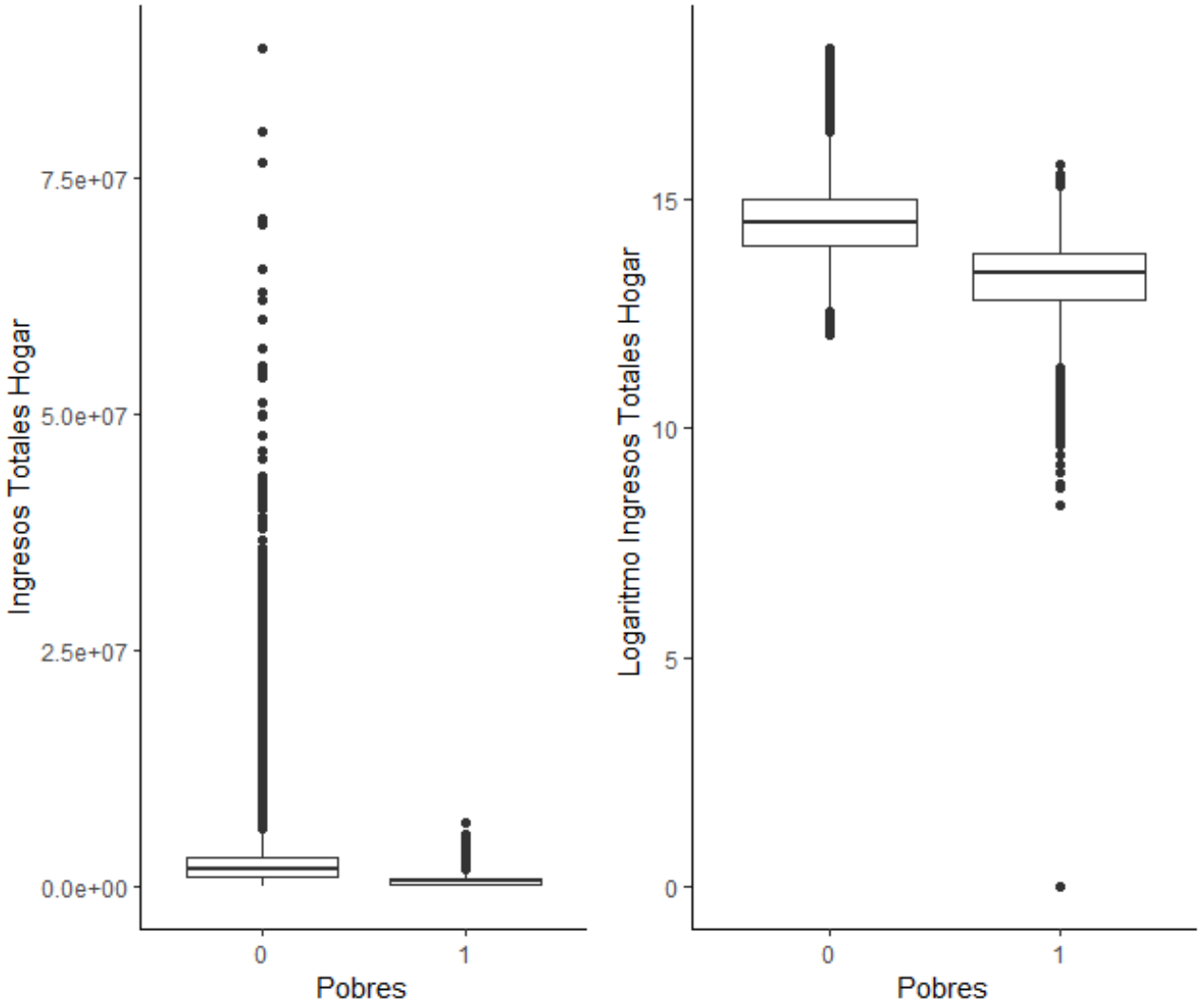


Figure A.2: Cajas y Bigotes Análisis de Ingreso - Colombia (2018)



3. Anexos Modelos de Clasificación

Table A.1: Modelos iniciales

	Dependent variable:				
	Pobre				
	(1)	(2)	(3)	(4)	(5)
Vivienda Propia =1	−0.712***	−0.714***	−0.638***		−0.712***
	(0.016)	(0.016)	(0.015)		(0.016)
Número Total de Mujeres Hogar	0.126***	0.126***	0.148***		0.126***
	(0.009)	(0.009)	(0.008)		(0.009)
Jefe de Hogar Mujer =1	0.038**	0.040***		0.142***	0.038**
	(0.015)	(0.015)		(0.013)	(0.015)
Número de Ocupados Hogar	−0.715***	−0.715***	−0.782***		−0.715***
	(0.011)	(0.011)	(0.009)		(0.011)
Edad Jefe Hogar	0.002***	0.002***		−0.023***	0.001
	(0.001)	(0.001)		(0.0004)	(0.002)
Edad Jefe Hogar^2					0.00001
					(0.00002)
Núm. Menores Edad Hogar	0.853***	0.853***	0.801***		0.853***
	(0.009)	(0.009)	(0.008)		(0.009)
Máx. Grado Escolar Jefe Hogar	−0.060***	−0.060***		−0.050***	−0.060***
	(0.002)	(0.002)		(0.002)	(0.002)
Jefe Hogar Ocupado = 1	−0.181***	−0.177***		−0.646***	−0.179***
	(0.019)	(0.019)		(0.015)	(0.020)
Núm. Afiliados Salud Hogar	0.061***	0.062***	0.092***		0.062***
	(0.007)	(0.007)	(0.007)		(0.007)
Uso Prod. Financieros Jefe Hogar = 1	−1.036***			−1.254***	−1.035***
	(0.133)			(0.126)	(0.133)
Constante	−1.156***	−1.164***	−1.500***	0.398***	−1.136***
	(0.036)	(0.036)	(0.016)	(0.030)	(0.060)
Observations	164,959	164,959	164,959	164,959	164,959
Log Likelihood	−66,666.120	−66,705.200	−67,328.020	−80,224.950	−66,666.020
Akaike Inf. Crit.	133,354.200	133,430.400	134,668.000	160,461.900	133,356.000
Note:					*p<0.1; **p<0.05; ***p<0.01

Figure A.3: Comparación ROC de los modelos inciciales

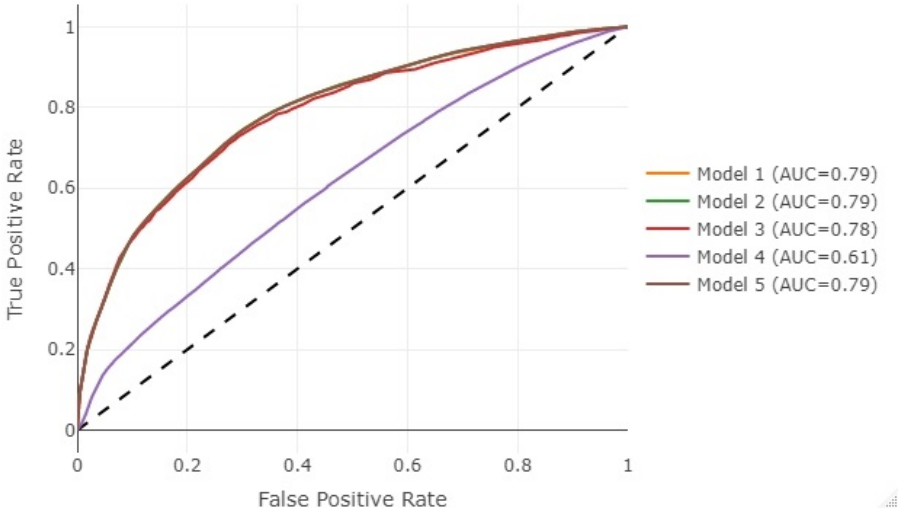


Table A.2: Modelos estimados para mejorar capacidad predictiva

Model		lambda	ROC	Sens	Spec	Acc	Kappa
Logit-Cross validation	1	none	0.79	0.96	0.28	0.83	0.31
Logit-Lasso Accuracy	55	0154	0.79	0.97	0.23	0.83	0.28
Logit-Lasso ROC	54	0.01402	0.79	0.97	0.23	0.83	0.28
Logit-Lasso Sensitivity	100	1.02329	0.77	1.00	0.00	0.80	0.00
Logit-Upsampling	57	0.01855	0.79	0.75	0.69	0.72	0.43
Logit-Downsampling	57	0.01855	0.79	0.74	0.69	0.72	0.43

Table A.3: Alternate cutoff Closest toplefttt ROC

	threshold	specificity	sensitivity
1	0.53	0.73	0.71

Figure A.4: Coeficientes en función de la regularización de Lasso

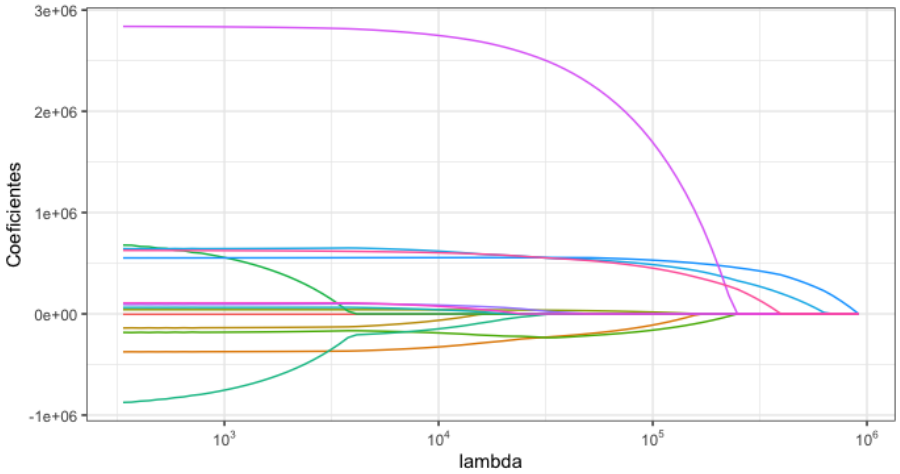


Table A.4: Matriz confusión mejor Modelo de Regresión

	Y pred =0	Y pred =1
Y=0	159, 592	0
Y=1	5, 158	209