Hacia una predicción costo-eficiente de la pobreza en Colombia

Daniel Lasso 201813962
Gabriela Mejía 201814214
Matteo Rozo 201815160

El repositorio correspondiente a este reporte se encuentra en el siguiente link

1. Introducción

En los últimos años, los gobiernos han enfocado sus esfuerzos a resolver unos de los grandes problemáticas que afecta el bienestar social de sus ciudadanos, la pobreza. En razón de ello, se han pactado diferentes acuerdos, como la Declaración del Milenio de las Naciones Unidas, los cuales establecen su compromiso con la lucha contra la pobreza. Por lo consiguiente, se han planteado políticas que logren incentivar y permitir la movilidad social, lo cual requiere un proceso de focalización de los recursos hacia los hogares más vulnerables. En Colombia, el Departamento Administrativo Nacional de Estadística (DANE) y el Departamento Nacional Planeación (DNP) ha implementado indicadores que permiten clasificar a los hogares en niveles de pobreza. En particular, encuestas como la Gran Encuesta Integrada de Hogares (GEIH) recolectan información completa sobre las condiciones socioeconómicas de los colombianos. No obstante, uno de los mayores problemas de la predicción de la pobreza es lo costoso y poco eficiente que es la toma de datos. En colombia, aunque no hay cifras sobre el costo por encuestado, el costo total del censo del 2018 fue de 350 mil millones de pesos, monto equivalente al 0.4 % del PIB del 2018 (La República). En general, esto se da porque el costo de las encuestas crece de forma exponencial con la cantidad de preguntas que se hacen. En este sentido, resulta fundamental explorar alternativas que con poca información sobre los hogares logren predecir adecuadamente la pobreza.

La literatura ha identificado que la predicción de la pobreza, puede abarcarse desde, al menos dos aproximaciones:

- Predicción de una variable continua: En esta aproximación, se busca predecir el ingreso de individuos y la categorización de la pobreza/no pobreza se da, desde la agregación a nivel hogar y su respectiva comparación con la línea de pobreza del hogar.
- Predicción de una variable dicótoma: En esta aproximación, se busca predecir la probabilidad de pobreza/no pobreza de los hogares a partir de sus características y distintas agregaciones de los individuos que lo conforman y su respectiva comparación con un umbral de probabilidad a partir del cual el hogar será clasificado como pobre/no pobre.

Estas aproximaciones han sido ampliamente testeadas desde la econometría y el Machine Learning, y como todo, no tiene una única respuesta correcta. En (Verme,) se comparan precisamente estas dos aproximaciones a

través del Modelo Clásico Lineal, LASSO, Logit, y Random Forest. Se provee evidencia de que ex-ante no existe un modelo superior a los demás, antes de ser testeados en un contexto específico de la distribución del ingreso. En últimas, establece que la escogencia de un modelo óptimo depende de la distribución particular del ingreso y de la ubicación de las líneas de pobreza.

Dicho esto, el objetivo de este texto va a ser. Tratar de identificar la posibilidad de predecir la pobreza en los hogares a través de características observables de los individuos. Con esta información, la primera sección presentará los datos que utilizamos. La segunda sección va a utilizar modelos de selección de variables para identificar las variables más asociadas a la pobreza y así, poder establecer modelos más simples y eficientes que requieran muestreos menos costosos. La tercera sección se encargará de utilizar aquellos factores socioeconómicos más importantes encontrados previamente para predecir la pobreza. Esto incluirá tanto modelos de regresión como de clasificación. Finalmente, se realizará una conclusión y reflexión sobre la predicción de la pobreza

2. Descripción de los Datos

Para esto, es importante caracterizar el problema al que nos enfrentamos desde una descripción de los datos que corresponden a la GEIH de 2018 a nivel de hogares y de individuos. Tras realizar una limpieza de la base de datos que se explica con detalle en el anexo, se consolidó una base con 97 variables de interés y 542 941 individuos. Adicionalmente, consolidamos otra base de hogares colapsando la información promedio de estas variables dentro del hogar. Por ejemplo, al colapsar la información de educación de las personas que corresponden a un mismo hogar, obtenemos variables como la proporción de individuos con grado universitario en el hogar. Posteriomente, esta base se pegó con la información de hogar presente en la GEIH como el valor del arriendo, el número de personas en el hogar, etc. Al final la base contiene información de 164 960 hogares y un total de 106 variables. Con esta información podemos realizar teorías a priori, e identificar retos en materia de entrenamiento de los modelos que pueden en últimas generar predicciones erróneas sobre aquella población que queremos clasificar. Primero observamos el balanceo de la muestra de entre hogares catalogados en condición de pobreza y los que no: la Figura 1 demuestra que debemos predecir la pobreza en una muestra donde solo el 20 % de los hogares se encuentran en condición de pobreza.

Figura 1: Hogares en condición de pobreza para la muestra

Esta falta de balance en la muestra genera que deban ajustarse parámetros y llevar a cabo estrategias (Descritas en las Sección 3) que garanticen que siempre que algún hogar se encuentre en condición de pobreza, se catalogue como tal. Si hablamos de realizar esfuerzos para atender a la población vulnerable, llegar a toda la que existe (Aunque proporcionalmente sea poca) es prioridad.

Ahora, ¿Que variables poseemos para realizar el análisis? Para la predicción del ingreso, tenemos características observables a nivel de individuos encuestados que permiten comprender dinámicas del mercado laboral, la

formalidad y sociodemográficos que son en últimas las que se agregan a nivel del hogar para realizar predicciones:

Cuadro 1: Estadísticas descriptivas de variables sociodemográficas para la muestra de entrenamiento

Variable	N	Share
Sexo	542941	
... Hombre	256306	47.2 %
... Mujer	286635	52.8 %
PET	542941	
... Persona en edad de trabajar	447512	82.4 %
... No Pet	95429	17.6 %
Nivel educativo	542941	
... Ninguno	28200	5.2 %
... Preescolar	15247	2.8 %
... Básica primaria (1o - 5o)	134797	24.8 %
... Básica secundaria (6o - 9o)	94204	17.4 %
... Media (10o - 13o)	119629	22 %
... Superior o universitaria	128108	23.6 %
... No sabe,no informa	22689	4.2 %
Ocupados	248269	45.7 %
Desocupados	29965	5.5 %
Inactivos	169278	31.2 %
No info	95429	17.6 %
En este trabajo es	542941	
... Obrero o empleado de empresa particular	91977	16.9 %
... Obrero o empleado del gobierno	12824	2.4 %
... Empleado doméstico	8039	1.5 %
... Trabajador por cuenta propia	114589	21.1 %
... Patrón o empleador	9120	1.7 %
... Trabajador familiar sin remuneración	7619	1.4 %
Afiliado salud	542941	
... Sí	417153	76.8 %
... No	30135	5.6 %
... No sabe, no informa	95653	17.6 %
Cotizante pensiones	542941	
... Sí	95118	17.5 %
... No	443295	81.6 %

Dentro del entrenamiento el ingreso promedio es de $639,271 pesos y la edad promedio es de 33 años, pero tenemos individuos desde los 0 hasta los 110 años, aunque más del 80 % se encuentra en edad de trabajar. En términos de sexo la muestra se encuentra balanceada y el nivel educativo de la población se distribuye mayormente entre Básica primaria, Superior, Media y Secundaria. Este mismo análisis dentro de la muestra de testeo revela que la edad promedio se mantiene cerca a los 33 años, el sexo tiene un balance similar y la educación una distribución porcentual en la muestra equivalente con el entrenamiento, así como la "función"de los individuos dentro del hogar. La distribución departamental es similar para todos los departamentos, menos Bogotá. En la muestra de testeo se excluyen todos los individuos de Bogotá, lo cual representa un reto si la residencia en Bogotá termina siendo un predictor relevante para la pobreza en el entrenamiento de los modelos.

Los individuos dentro de la muestra de entrenamiento se encuentran en mayor medida, ocupados (46 %) o inactivos (31 %); consistente con que 38 % de individuos de la muestra se encuentran trabajando; en oficios del hogar (19.4 %) o estudiando (13.4 %). Podemos comenzar a comprender las dinámicas de la informalidad al observar que un 21 % de los individuos es cuentapropista y el 81 % no cotiza a pensiones, aunque el 76 % si se encuentra afiliado a una EPS, esto no necesariamente nos permite conocer que porcentaje se encuentra afiliada mediante régimen contributivo o subsidiado. Dentro de la muestra de testeo, la distribución de estas variables es similar, por lo que podemos pensar que la caracterización de los individuos en el mercado laboral y su informalidad estará bien capturada por los modelos entrenados.

3. Seleccionando los determinantes fundamentales de la pobreza

Para hacer predicciones más completas y sencillas es importante determinar las variables que están más asociadas a la pobreza. Para permitir que los datos 'hablen por si mismos' y no caer en sesgos como investigadores utilizaremos modelos de selección de variables. En particular, se probarán los modelos Lasso, Ridge y Random Forest, pues son los que identificamos en la literatura con mejor potencial para la selección de variables. Análogamente, es posible realizar metodologías como best subset selection que se basan en adicionar variables y verificar el error de predicción. Este método en particular calcula un modelo para cada combinación posible de variables explicativas. El problema es que para estimarlo en nuestro modelo con 104 dummys requeriría estimar $2{,}02 * 10^{31}$ modelos de regresión lo cual es fisicamente imposible.

Cómo se puede observar en las estadisticas descriptivas, la pobreza es una cateogría desbalanceada, pues hay pocos hogares pobres. Por esto tuvimos que realizar una metodología llamada Oversampling en la cual rebalanceabamos las proporciones de la pobreza. Esta metodología utiliza las observaciones de individuos pobres y genera unas observaciones ligeramente parecidas a estas, hasta que genera una muestra balanceada. En nuestro caso, queríamos que estuviera balanceada 1 a 1. También probamos con undersampling, que a diferencia del anterior método este elimina aleatoriamente observaciones hasta que queden balanceadas. No obstante, en el segundo caso estaríamos perdiendo el poder estadístico resultado de tener una muestra grande, por lo cual, preferímos el primero.
Con esta base, se calibraron tres modelos de predicción de pobreza. En primera instancia, se realizó un modelo lasso de selección de variables, este modelo intuitivamente penaliza el error de predicción -que en este caso va a ser la desviación binomial por tener una variable dependiente dicótoma- con un término de la sumatoria del valor absoluto de los coeficientes. Esto implica que basicamente está revisando qué tan diferente de 0 son los estimadores. Así mismo, qué tanto los penaliza está mediado por un parámetro lambda. Tal que si este parámetro es 0 la estimación es equivalente a la de MCO y si el parámetro es infinito es porque todos los coeficientes son 0. Este lambda óptimo se eligió a través de 5-fold-cross-validation separando los datos en una muestra de entrenamiento (80) y testeo (20). Es importante mencionar que los modelos como Lasso y Ridge son sensibles a la magnitud del estimador. Por eso, para ambos se tuvo que estandarizar las covariables con su media y varianza. Como resultado, el lambda óptimo fue 0.000133547 para el modelo con oversampling, esto tiene sentido porque hay muchas variables, por lo cual, el lambda es pequeño para lograr estabilizar el error de predicción. En segundo lugar, se utilizó un modelo Ridge, cuya intención es similar a Lasso, pero en la cual no se penaliza por el valor absoluto sino por el coeficiente al cuadrado. Este modelo no necesariamente reduce los coeficientes a 0, pero igual nos dará una intuición sobre la importancia de los mismos.
Finalmente, ajustamos un modelo de Random Forest, el cual busca explorar no linealidades de las variables de interés. Este método funciona a través de árboles de decisión, en el cual se hace una segmentación óptima de la muestra en decisiones sobre los datos para ejecutrar una predicción. Por ejemplo, si la familia tiene vivienda propia y el valor del arriendo estimado es superior a 1.000.000 de pesos, entonces lo clasifica como que no es pobre. Sin embargo, los árboles de decisión son sensibles al cambio en observaciones, por lo cual, los árboles aleatorios controlan este efecto a través de generar un árbol para cada remuestreo que se haga de la base de datos. Así mismo, para encontrar los hiperparámetros óptimos se realizó validación cruzada y se tenía como métrica para la la sensibilidad. Esto porque queriamos identificar a los pobres con un mayor detalle. Ya que el proposito es la selección de variables no se discutirá sobre cuáles fueron estos parámetros.

Los resultados fueron robustos para los 3 modelos, por esa razón, solo presentaremos la importancia de las variables elegidas por Random Forest, pero los 3 ofrecen resultados similares. La figura 2 presenta las variables más importantes para la predicción de la pobreza, en las que resalta la linea de pobreza de un hogar pues entre

más baja sea menor es la probabilidad de ser pobre, las horas promedio trabajadas en el hogar que es un indicador del mercado laboral, la tenenencia de la vivienda pues claramente aproxima el estrato y los ingresos, entre otras. Como resultado se tiene que el mejor modelo utiliza principalmente 25 variables de las 106 presentes en la base mencionada anteriormente.

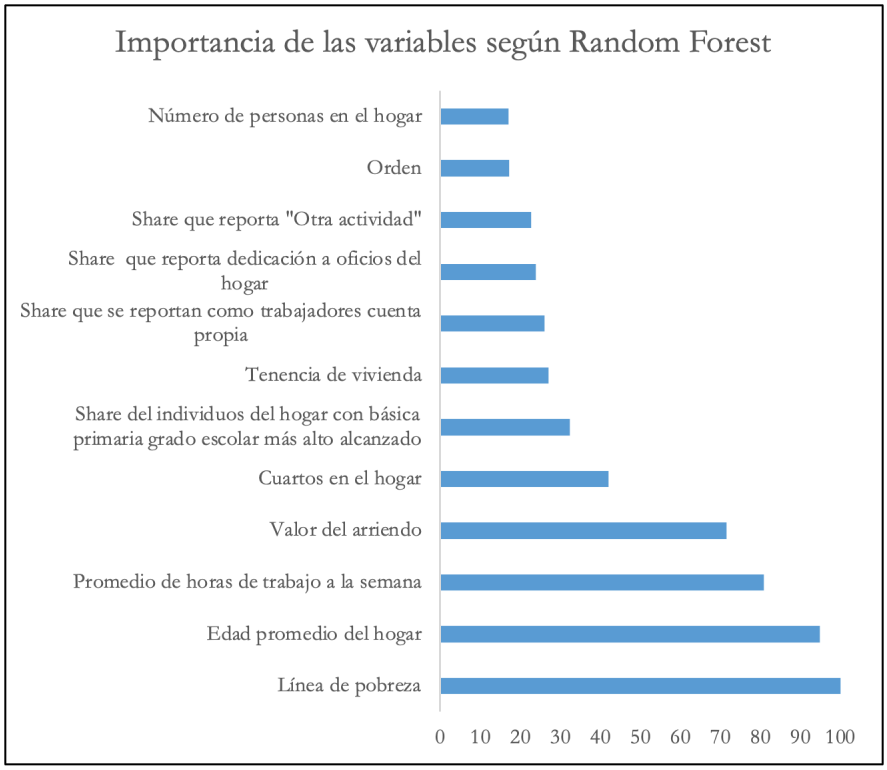

Figura 2: Importancia de las variables según RF

4. Metodología de la predicción de la pobreza

Antes de continuar es crucial comprender que en últimas se va a predecir sobre una muestra de validación dada por el DANE (Ignacio), que será evaluada dándole mayor peso a la tasa de Falsos Negativos (es decir, aquellos hogares que hemos clasificado como no pobres, cuando verdaderamente se encuentran en condición de pobreza). Sin embargo, al evaluar los modelos no se podrá determinar las malas predicciones ante el desconocimiento del status de pobreza de los hogares de este set de validación. Por esta razón para hacer validación fuera de muestra y poder evidenciar la bondad de nuestros modelos, construimos nuestra propia base de testeo, con el objetivo de que se parezca en las observables lo más posible al set de validación. Para esto, realizamos una metodología de propensity score matching y soporte común a partir del modelo que, como fue mencionado anteriormente, nos permite predecir de mejor manera la pobreza. Para el emparejamiento de la muestra, utilizamos un soporte común basado en las observables más importantes que se muestran en la Figura 2. En adelante, esta sección va a predecir la pobreza tanto a través de clasificación, como para regresión. En general, se van a utilizar los determinantes identificados previamente como covariables *

*Si una de las categorías es importante, entonces en los modelos se tomo todos los niveles de esa variable factor

5. Resultados de la pobreza

5.1. Clasificación

La figura 3 resume los resultados de los 8 modelos planteados utilizando oversampling para resolver los problemas de imbalance muestral. Dado que nos interesa la generalización y capacidad predictiva, es importante fijarse en las métricas fuera de muestra. Las estimaciones se presentan para un modelo de probabilidad lineal, lasso, ridge y random forest (RF). En particular, se puede evidenciar que el Recall, es decir, la capacidad que tienen los modelos para no dejar personas en condición de pobreza sin clasificarlos como pobres, es muy superior para modelos como RF, que identifica el 91 % de las personas en condición de pobreza. En segundo lugar se encuentra sorprendentemente el modelo de regresión lineal (0.88), que le gana en esta métrica a los modelos penalizados de Ridge y lasso. Por otra parte, RF presenta las mejores métricas de Accuraccy -proporción de aquello que fueron bien clasificados-, logrando clasificar correctamente el 92 % de los hogares. Los otros modelos presentan un resultado similar. Finalmente, F1 es una métrica que armoniza los resultados de accuracy y recall, ofreciendo una medida combinada que resulta importante en problemas de muestras imbalanceadas. En este caso la mejor corresponde a RF y los demás modelos tienen valores similares. El mejor modelo de RF se calibró a través de validación cruzada y un análisis de grilla, en particular, se obtuvo que el mejor número de variables para hacer el Bagging fue de 49, por lo que utilizando arboles con 49 variables resultó más útil. Segundo, que la mejor profundidad máxima de los árboles fue de 6 y que el número de estimadores óptimos era de 100. Los demás parámetros son los que están por default en el software de R.

Modelo	Muestreo	Evaluación	Sensibility	Specificity	FPR	Accuracy	Precision	Recall	F1
Linear Model	SMOTE - Oversampling	Dentro de muestra	0.75	0.88	0.12	0.81	0.78	0.87	0.82
Linear Model	SMOTE - Oversampling	Fuera de muestra	0.75	0.88	0.12	0.81	0.78	0.88	0.82
Lasso	SMOTE - Oversampling	Dentro de muestra	0.86	0.77	0.23	0.81	0.85	0.77	0.80
Lasso	SMOTE - Oversampling	Fuera de muestra	0.86	0.77	0.23	0.81	0.84	0.77	0.81
Ridge	SMOTE - Oversampling	Dentro de muestra	0.86	0.74	0.26	0.80	0.84	0.74	0.79
Ridge	SMOTE - Oversampling	Fuera de muestra	0.86	0.77	0.23	0.81	0.84	0.77	0.81
RF	SMOTE - Oversampling	Dentro de muestra	1.00	1.00	0.00	1.00	1.00	1.00	1.00
RF	SMOTE - Oversampling	Fuera de muestra	0.92	0.91	0.09	0.92	0.92	0.91	0.92

Figura 3: Resultado de comparación entre las métricas de los modelos de regresión

La curva ROC de la figura 4 nos muestra precisamente el trade off entre la regla de clasificación: Podemos observar como en un punto, la pendiente de la curva cambia drásticamente hacia el aumento del False Positive Rate (Dado que estamos haciendo un rebalanceo, entonces tenderemos cada vez más a categorizar como pobres, hogares que no se encuentran en esa condición) Sin embargo, esto es consistente con nuestra prioridad de que ningún hogar pobre se quede sin clasificarm aunque pogresivamente se convertirá en un problema mayor porque estaré sencillamente prediciendo mal; y por ejemplo destinando recursos a atender hogares que no se encuentran en condición de vulnerabilidad

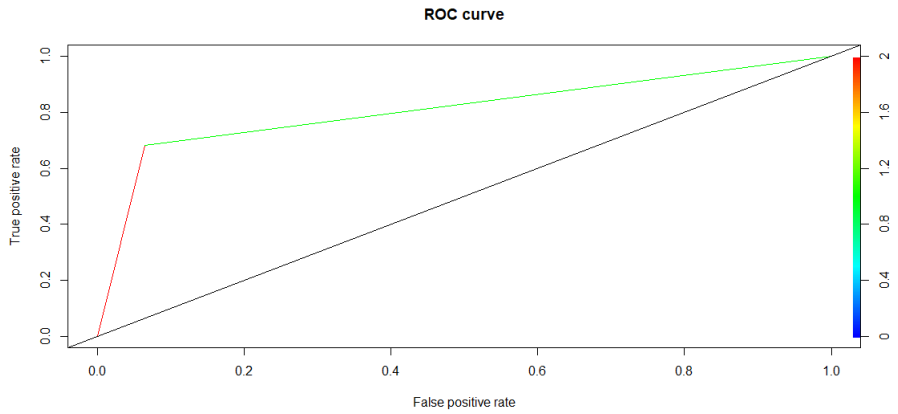

Figura 4: Curva ROC del modelo de RF

En suma, para clasificar la pobreza RF presenta una solida captura del proceso generador de datos, pues puede aprender de los efectos heterogeneos que pueden tener los grupos de individuos. Por ejemplo, es muy diferente un hogar con muchos individuos como independientes en ocupación pero que la edad promedio son menos de 30 años, a otro donde la edad promedio es superior a los 60, pues el trabajo y los ingresos por este medio independiente pueden ser sustancialmente distintos. Por otra parte, en cuanto a Ridge y lasso dependen en gran medida de la habilidad de los investigadores de capturar con interacciones estas no linealidades. Lo cual, pese a que lo intentamos mitigar a través de la consolidación de interacciones y con polinomios de grado 2 en algunas variables, no se pudo resolver satisfactoriamente.

5.2. Regresión

De forma simétrica, se realizá el ejericio para predecir la pobreza a través del ingreso. Cabe recordar que estos resultados son producto de predecir el ingreso y contrastarlo con la línea de pobreza de cada hogar con el fin de obtener la predicción. En este orden de ideas, la figura 5 nos muestra un resultado diferente al ya establecido por clasificación, el modelo de Regresión Lineal es aquel que será elegido para la estimación en prueba. Pero, a diferencia de la otra aproximación, se destaca que este también tiene el menor RMSE con un valor de 946625,3 por fuera de muestra. En este caso Ridge y lasso sugieren una penalización de las variables, luego de aplicar validación cruzada con los subdatos de entrenamiento encontramos que el lambda de lasso es pequeño, aproximadamente 0.01 lo que sugiere poca penalización y es natural ya que estas variables fueron seleccionadas prviamente. Estos presentan resultados superiores al modelo de regresión lineal en sensibilidad y precisión. Pero, como nos interesa mucho el recall del modelo porque queremos predecir bien a las personas pobres, el modelo de regresión lineal ofrece resultados superiores a este, identificando al 55 % de los pobres en la muestra de testeo, mientras que sus hermanos mayores al 30 % aproximadamente. Finalmente, es importante mencionar que un modelo que identifique al 55 % de los pobres es casi similar a asignarlo con una moneda al aire. Por lo cual, en defninitiva elegimos el modelo de RF también para esta metodología, que presenta todas las métricas de interés por encima del 90 %. Para este modelo se encontraron los hiperparámetros óptimos a través de validación cruzada, lo que ofrece los siguientes parámetros óptimos: 23 variables óptimas para hacer el bagging (resampleo), 5 como la profunidad máxima del árbol y 1000 árboles.

Modelo	Muestreo	Evaluación	Sensibility	Specificity	FPR	Accuracy	Precision	Recall	F1	RMSE
Linear Model	SMOTE - Oversampling	Dentro de muestra	0.89	0.56	0.44	0.83	0.57	0.56	0.56	1036254.1
Linear Model	SMOTE - Oversampling	Fuera de muestra	0.89	0.55	0.45	0.82	0.57	0.55	0.56	946625.3
Lasso	SMOTE - Oversampling	Dentro de muestra	0.97	0.30	0.70	0.84	0.72	0.30	0.42	1072361.1
Lasso	SMOTE - Oversampling	Fuera de muestra	0.97	0.30	0.70	0.84	0.72	0.30	0.42	984930.9
Ridge	SMOTE - Oversampling	Dentro de muestra	0.97	0.29	0.71	0.84	0.72	0.29	0.42	1076523.8
Ridge	SMOTE - Oversampling	Fuera de muestra	0.97	0.29	0.71	0.83	0.71	0.29	0.41	988470.8
RF	SMOTE - Oversampling	Dentro de muestra	0.00	1.00	0.00	1.00	1.00	1.00	1.00	466294.5
RF	SMOTE - Oversampling	Fuera de muestra	0.92	0.91	0.09	0.92	0.92	0.91	0.92	1574964.4

Figura 5: Resultado de comparación entre las métricas de los modelos de clasificación

Por otra parte, se procedío a predecir y a clasificar con los datos de muestreo que separamos anteriormente. Por un lado, en la clasificación, se encontró que un 37.6 % de la muestra es clasificado como pobre. Por otro lado, en la regresión, se observa que un 20.4 % de la muestra se predice como pobre. De tal forma, podemos apreciar las diferencias que se generan al utilizar cada enfoque, donde el primero tienen sesgo mucho menor a predecir pobres que el segundo.

6. Conclusiones y recomendaciones

En conclusión, podemos decir que es posible lograr predicciones acertadas de la pobreza, incluso cuando los hogares en esta condición son cada vez más dificiles de identificar a medida que se reducen. Lograr estas predicciones sin embargo, se encuentra sujeta a poder obtener información certera, relevante y consistente con la teoría y literatura económica que logren capturar al menos 3 factores: 1) Condiciones de la vivienda; 2) Informalidad y 3) Educación. De ahí, nuestro análisis permitiría a las entidades estatales realizar encuestas rápidas, con menor número de variables y eficientes que le permitan medir constantemente la pobreza sin invertir grandes sumas de recursos que podrían, entre muchas cosas, ser invertidas en programas contra la pobreza. Esto deja la puerta abierta a nuevas metodologías, incluso digitales, donde se realicen preguntas simples y se aprovechen formas de consultar información puntual a escala. Adicionalmente, encontramos que el mejor modelo es Random Forest, pues este logra capturar no linearidades en los datos y que tratar de predecir la pobreza es más certero estadísticamente que predecir los ingresos y posteriormente clasificarlo con las lineas de pobreza.

7. Anexo

7.1. Limpieza de Datos

Para la limpieza de datos, primero, se tomó la información sobre la GEIH a nivel personas que incluye más de 100 variables. Sobre estas se restringió el número de variables a un conjunto de 60 factores socioeconómicos, la única información que presentaba la base de testeo. Esto porque, naturalmente, no podemos elaborar un modelo con variables para las cuales no tenemos información en aquellos individuos que queremos predecir. En segunda instancia, con este subset de la base inicial procedimos a imputar los valores faltantes. Las variables de ingreso de los individuos para los que no tuviesemos información las reemplazabamos por 0, para variables factor que tuvieran una respuesta binaria: Sí o No, reemplazamos los valores faltantes por la moda y aquellas variables factor

que presentaran la opción de respuesta: No sabe o no informa, reemplazamos los valores faltantes por estos. En cuanto a otras variables numéricas como la edad, el número de personas en el hogar, entre otras, reemplazamos los valores faltantes por la moda. Por otra parte, si bien la literatura refiere a la importancia de identificar los valores atípicos, no los manipulamos. Esto no es una limitación ya que modelos como Random Forest van a capturar estas no-linealidades. Finalmente, convertimos las variables factor en dicótomas.

8. Referencias

Referencias

Verme, P. (2020). Which model for poverty predictions? *Discussion Paper, No. 468, Global Labor Organization (GLO)*. Descargado de https://www.econstor.eu/bitstream/10419/213811/1/GLO-DP-0468.pdf