Predicting Poverty

Camilo Bonilla Rafael Santofimio Nicolás Velásquez

September 30, 2022

1 Introduction

Actualmente se pueden encontrar multiples estudios sobre medición y seguimiento de pobreza en los hogares colombianos. Dentro de estos se encuentra el análisis del comportamiento de pobreza entre 1997 y 2010 de Angulo et al. (2013), concluyendo que la pobreza multidimensional disminuyó significativamente durante esta ventana de tiempo, pero sigue siendo alta para los estándares internacionales si se compara con la media de los paises miembros de la OCDE.

Por otra parte, el desgaste de capital humano y la demanda de recursos para la recolección de información a través de encuestas destinadas a medir estos indicadores pueden ser el mayor desafio al momento de categorizar hogares vulnerables o pobres para destinarlos a programas estatales. Es por esto que la tendencia mundial se ha encaminado a aprovechar el mejoramiento del poder computacional y las técnicas estadísticas para estimar y predecir pobreza. Multiples ejemplos de técnicas de Big Data (BD) y Machine Learnig (ML) se pueden encontrar en la implementación de modelos de clasificación como en paises africanos (Jean et al., 2016), (Kshirsagar et al., 2017) o regiones en Jordania (Alsharkawi et al., 2021). Tras una rápida revisión en bases indexadas no se encontraron publicaciones que empleen estas técnicas (BD y ML) en la predicción de la pobreza monetaria para Colombia.

Teniendo en cuenta lo anterior, el presente trabajo propone dos (2) modelos de predicción de la pobreza monetaria de los hogares colombianos, uno de clasificación y otro de regresión del ingreso, a partir de información reportada en la GEIH de 2018. Ambos muestran el número mínimo de variables empleadas para obtener la menor tasa de falsos negativos, es decir, clasificación de hogares pobres como rico, y sus respectivos hiperparámetros.Se determina que la implementación de ramdon forest para ambos casos resulta siendo los mejores modelos de predicción, en lugar de LM, LOGIT, Lasso, Ridge, elastic net y XgBoost.

En la sección de datos se describe las variables elegidas y sus respesctivas transformaciones, a partir de los registros a nivel de hogar y personas reportados en la GEIH. En la sección modelos y resultados se presenta la especificación de ambos modelos, la forma en cómo se entrenaron y la selección de hiperparámetros. Finamente, en conclusiones y recomendaciones se indican la sintesis del tabajo con los mejores modelos ajustados, y sus respectivas limitaciones a la hora de ser implementados por los diseñadores de políticas públicas. Los archivos, tablas, bases de datos y códigos totalmente replicables se pueden obtener en el siguiente link: ProblemSet2 el cual redirige al repositorio en Github.

2 Data

De la encuesta GEIH se tomaron dos bases de datos que tenían información a nivel de hogar (13 variables) y personas (130 variables) cada una con 164960 y 542191 observaciones respectivamente. Esta última se colapso para obtener de manera agregada información de los hogares colombianos en el 2018, la cual está relacionada con condiciones de empleo, y características de los individuos tales como generó, edad, estado civil, nivel educativo y fuentes de ingresos. Para elegir las variables a emplear en los distintos modelos, primero se genero un filtro que eliminó todas las variables con una proporción de datos faltantes mayores al 30%, dado que se considera riesgoso en términos de poder explicativo y predicción extrapolar valores más allá de esta frontera. Así entonces, quedaron 17 variables a nivel de hogar entre las que se encuentra grado de escolaridad, ocupación, régimen de afiliación a salud, entre otras, y un número final de 164960 unidades de gasto percápita (hogares).

Segundo, aquellas variables que tenían valores faltantes por debajo del umbral, se aplicaron técnicas de imputación específicas para cada una, por ejemplo, en el caso de grado de escolaridad se asignó al hogar el máximo alcanzado por alguno de sus miembros. En cuanto la edad se estableció como el promedio de sus integrantes, y la variable sexo se transformó como la proporción de mujeres, ya que se presumen que estas perciben menores ingresos comparadas con los hombres, y consecuentemente hogares con mayor número de mujeres podrían estar por debajo de la línea de pobreza. Tercero, se convierten a variables dicótomas todas las categóricas. Cuarto, se estandarizan las variables numéricas. Quinto, se realizan transformaciones polinómicas e interacciones a las variables de edad, número de cuartos, dormitorios, personas en el hogar, proporción de afiliados a salud, mujeres y ocupados. Sexto, teniendo en cuenta que la propuesta implica diseñar dos grupos de modelos (clasificación y regresión), se generó una base para cada grupo con sus respectivas muestras de entrenamiento y testeo (proporción 70/30).

La elección de estas 17 variables (9 numericas y 8 categóricas) y sus respectivas transformaciones se debe a que están directamente relacionadas con la capacidad de los hogares colombianos de: generar ingresos (edad, grado escolaridad, horas laboradas y ocupación extra), identificar las fuentes donde provienen estos, perfilar el tipo de seguridad social (salud y pensión) y clasificar según la región. Este último permite incorporar a los modelos la heterogeneidad de las diferentes líneas de pobreza establecidas en el país. Las estadísticas descriptivas de estas variables se reportan en la tabla 1 en apendices.

3 Models and results

Para determinar la mejor especificación se diseñaron 12 modelos, los cuales comprenden regresión lineal (LM), logit, lasso, ridge, elasctic net, random forest y XGBoost. En el problema de clasificación se tomó como medida de desempeño el F1 Score, y en el de regresión el RMSE. El criterio de selección en ambos casos correspondió al que tuviera mayor poder de predicción fuera de muestra y que reportara la menor tasa de falsos negativos. En las tablas 2, 3 y 4 se muestra el desempeño de cada modelo según estas métricas.

El logaritmo del ingreso percápita por unidad de gasto (Ingpcup) fue la variable dependiente para los modelos de regresión, y "Pobre" para los de clasificación. Es importante aclarar que en ambos casos se sustrajeron estas dos variables más la relacionada con la línea de pobreza, así quedando con un total de 14 variables predictoras sin contar sus transformaciones. Esta última se empleó para clasificar los hogares pobres con base en los resultados del mejor modelo de regresión.

i Modelos de Clasificación

Para el modelo de clasificación se especificaron 6 modelos distintos, los cuales se fueron complejizando y mejorando a medida que se iban agregando restricciones, técnicas de regularización y modelación. Los modelos corresponden a: logit , lasso, ridge, elastic net, random forest y xgboost.

Debido a que se presenta un desbalance en la muestra entre hogares pobres y no pobres, siendo los primeros el 20% del universo muestral. Esto puede generar que las predicciones arrojen una falsa precisión (accuracy). Por lo tato, se planteó una grilla que combina sobremuestreo y submuestreo de tal manera que se distribuye cada categoria en un 50%. Al finalizar el ajuste se encuentra que fijar un sobre un sobremuestro del 75% y el 25% restante en submuestreo, representa un mejor desempeño a la hora de predicir bajo f1 score.

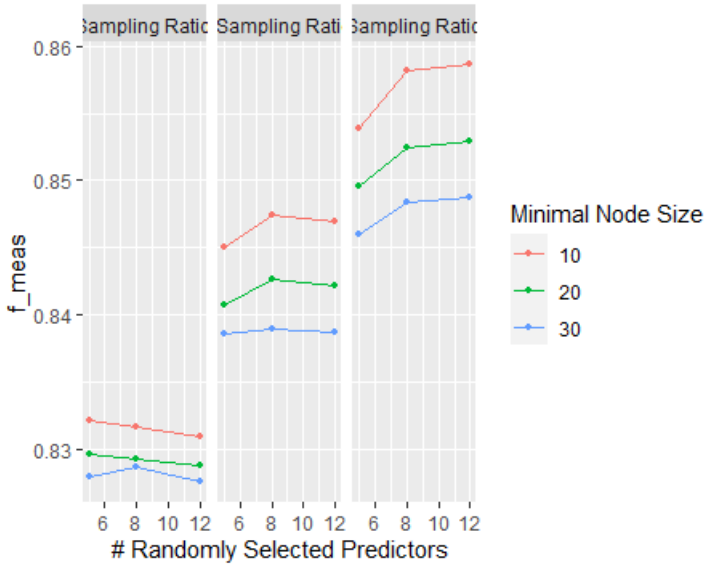
Al entrenar el bosque aleatorio y el xgboost, se escogieron las varibles más importantes con base a los métodos de regularización lasso y elastic net. De tal manera que los coeficientes que tendian a cero en ambos, fueron excluidos de los arboles de desición. De tal forma que se utilizaron tan solo 11 variables para entrenamiento.

En la tabla 1 y tabla 2 se encuentran las metricas y parámetros con los cuales se seleccionaron los mejores modelos de clasificación para cada tipo de especificación.La metrica que se utilizó para maximizar el poder predictivo en esta sección fué F SCORE el cual posee BETA igual a 1.75 lo cual le da un peso mayor en 75% al recall que a presission, es decir, que le da mayor importancia a no cometer falsos negativos que a falsos positivos.

Como se puede ver en las tablas mencionadas anteriormente el mejor modelo para la clasificación de pobres segun la métrica F SCORE es Random Forest, este modelo fué entrenado bajo los siguientes hiperparámetros: 200 arboles, un mínimo de 10 datos en cada nodo y 12 predictores mínimos escoginos al azar. Este modelo alcanzó un F Score por fuera de muestra de 0.857 superior a los demás.

De todos los modelos empleadosa en esta sección tres mejores modelos fueron Random Fores, xgboost y regresión lineal. Random Forest mejoró las predicciones por fuera de muestra en 3 puntos y 4 puntos respectivamente.

Figure 3.1: Ajustes de hiperparámetros para modelo de clasificación



Note: Esta figura muestra el ajuste de los hiperparametros del modelo de ramdon forest para clasificacción
Fuente: Elaborado a partir (Dane, 2008)

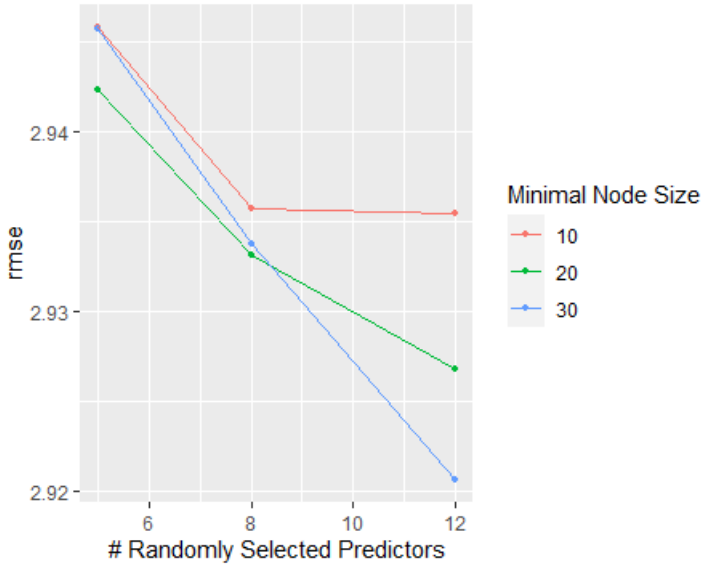
Como se observa en la gráfica 3.1 a medida que se aumentaron los numeros de predictores el modelo se ajustó mejor a la muestra de testeo. Los numeros de observaciones de nodos fueron mejor con 10 y el reatios de sobre muestreo tuvo mejores resultados con el paámetro 0.75.

ii Modelos de regresión de ingreso

Para el modelo de regresión y predicción del ingreso se especificaron 6 modelos distintos, los cuales se fueron complejizando complementando y potenciando a medida que se iban agregando restricciones, nuevas técnicas de regularización y de modelación. Dentro de estos modelos se encuentran: regreción lineal , lasso, ridge, elastic net, random forest y xgboost.

En esta sección de igualemanera se redujo la dimensionalidad de las variables bajo la combinación de la regularización elastic net y lasso, de 14 a 11 variables para entrenar los modelos y arboles de desición. Al igual que en el problema de clasificación se fijafron 200 áboles y profundidad 6.

Figure 3.2: Ajustes de hiperparámetros para modelo de regresión



Note: Esta figura muestra el ajuste de los hiperparametros del modelo de ramdon forest para regresión
Fuente: Elaborado a partir (Dane, 2008)

A la luz de los resultados, se observa que para el problema de regresión del ingreso, el mejor modelo que minimiza el RMSE es ramdon forest con los parámetros ajustados correspondientes al tamaño de muestra de 30 obervaciones minimas por nodo y 12 predictores. Se pueblo establecer que medida que los predictores aumentaron los parámetros se ajustaron mejor y se desempeñaron con mayor precición respecto a los demás modelos como lasso, ridge o modelo lineal. Adcionamente, se encontró que a pesar que el Rgboost es más extensivo en la parametrización este no produjo mejores resoultados.

4 Conclusions and recommendations

Durante el ajuste de los diferentes modelos pudimos observar que el mejor ajuste segun las métrics de F Scroe y rmse tamnto para el modelo de regresión como de clasificación fué Bosques aleatorios.

En las gráficas de la especificación de parametros para cada uno se observa una optimización de las métricas. Como el objetivo del parámetro es darle más peso a los falsos negativos que a los falsos positivos observamos que estas predicciones pueden estan más centradas al problema como tal.

En sistesis se obtuvieron modelos ajustados que predicen de forma acertada por fuera de muestra los hogares pobres de la muestra. Sirve aclarar que las variables más fuertes encontradas en cada modelo fueron genero, cotiza a pensión y numero de personas en el hogar.

5 Appendix

Table 1: Estadísticas descriptivas variables continuas estandarizadas

	variable	mean	sd	min	max
1	Edad promedio personas hogar	0	0.0025	-0.0046	0.0094
2	Linea de probreza	271502.6478	33695.8782	167222.4776	303816.6902
3	log del ingreso	13.0128	3.2807	-34.5388	18.3023
4	Numero cuartos hogar	0	0.0025	-0.0047	0.1876
5	Numero de dormitorios	0	0.0025	-0.0027	0.0357
6	Proporcion afiliados a salud	0	0.0025	-0.0078	0.0019
7	Proporcion mujeres	0	0.0025	-0.0047	0.0042
8	Proporcion ocupados	0	0.0025	-0.0039	0.0039
9	Proporcion personas hogar	0	0.0025	-0.0032	0.026

Fuente: Elaborado a partir de (Dane,2018)

Table 2: Reporte modelos LM, LOGIT, Lasso, Ridge y Elastic

	Problema	Modelo	Penalidad	Mixtura	Metrica	Valor	Error	over_ratio
1	Reg.	Lineal	N/A	N/A	rmse	3.347	0.060	N/A
2	Clas.	Lineal	N/A	N/A	f_meas	0.816	0.0003	0.5
3	Reg.	Ridge	N/A	0	rmse	3.335	0.072	N/A
4	Clas.	Ridge	0.0001	0	f_meas	0.814	0.001	0.75
5	Reg.	Lasso	0.00055	1	rmse	3.333	0.072	N/A
6	Clas.	Lasso	0.0001	1	f_meas	0.815	0.001	0.25
7	Reg.	Elastic Net	0.00055	0.9	rmse	3.334	0.072	N/A
8	Clas.	Elastic Net	0.000325	0.3666667	f_meas	0.815	0.0005	0.25

Fuente: Elaborado a partir de (Dane,2018)

Table 3: Reporte modelos ramdon forest

	Problema	Modelo	mtry	min_n	Metrica	Valor	Error
1	Reg.	Random Forest	12	30	rmse	3.188	0.055
2	Clas.	Random Forest	12	10	f_meas	0.857	0.001

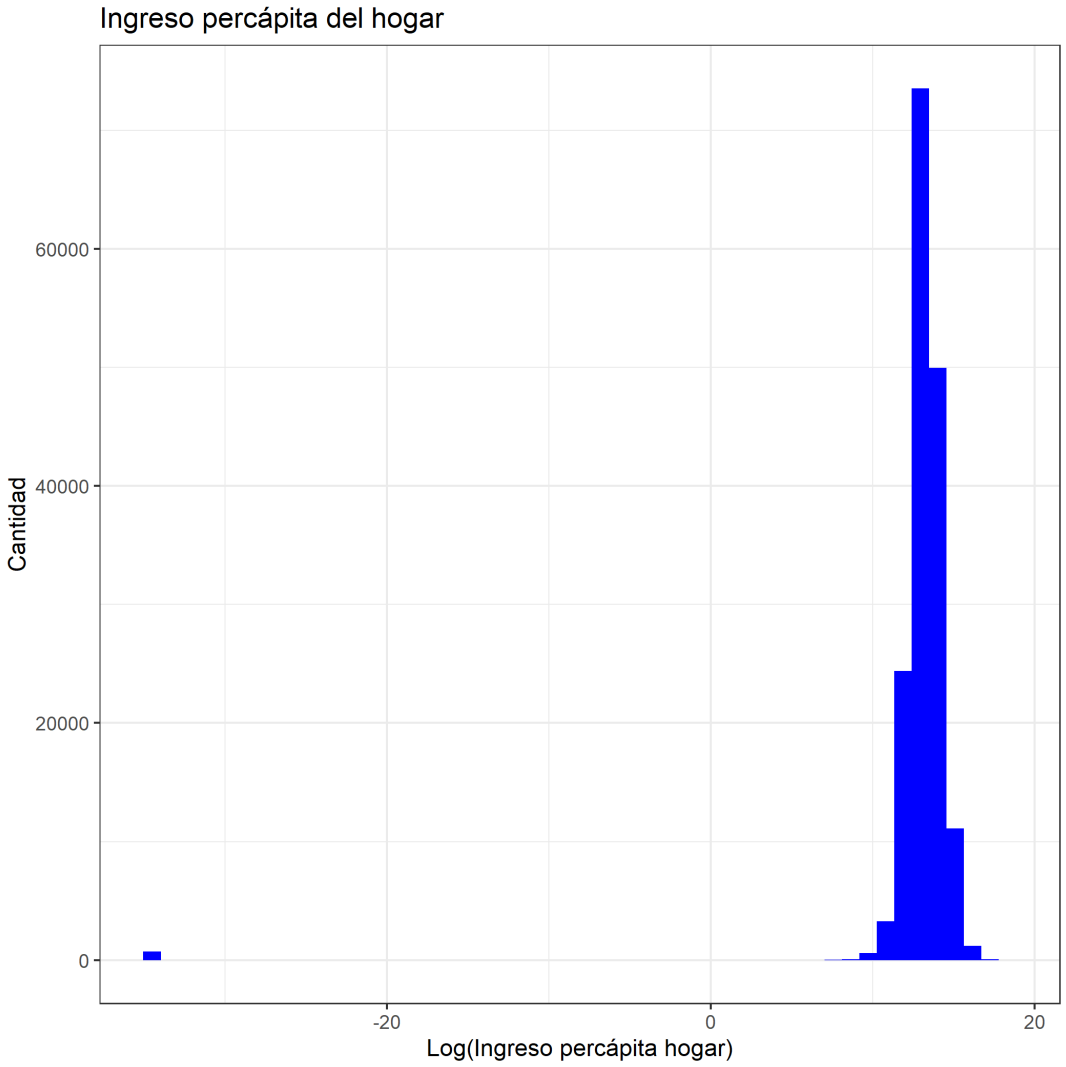
Fuente: Elaborado a partir de (Dane,2018)

Table 4: Reporte modelos XgBoost

	Problema	Modelo	mtry	min_n	sample_size	Metrica	Valor	Error
1	Reg.	XGBoost	5	20	1	rmse	3.192	0.052
2	Clas.	XGBoost	12	10	1	f_meas	0.830	0.001

Fuente: Elaborado a partir de (Dane,2018)

Figure 5.1: Distribuciión del ingreso de los hogares



Note: Esta figura muestra la distribución del ingreso de los hogares colombianos reportados en la GEIH 2018

Fuente: Elaborado a partir (Dane, 2008)

References

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., and Alyaman, M. (2021). Poverty classification using machine learning: The case of jordan. *Sustainability.*, pages Vol. 13, Page 1412, 13(3). Disponible en: https://doi.org/10.3390/SU13031412.

Angulo, R., Diaz, Y., and Pardo, R. (2013). Multidimensional poverty in colombia, 1997-2010. *ISER Working Paper Series.*, page np. Disponible en: https://ideas.repec.org/p/ese/iserwp/2013-03.html.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science.*, pages 353(6301), 790–794. Disponible en: https://doi.org/10.1126/SCIENCE.AAF7894/SUPPL_FILE/JEAN.SM.PDF.

Kshirsagar, V., Wieczorek, J., Ramanathan, S., and Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. page n.p. Disponible en: https://doi.org/10.48550/arxiv.1711.06813.