Problem Set 2: Machine Learning

Maestría en Economía - Universidad Nacional de La Plata

Carolina Basile

1 Introducción

Este trabajo pretende construir un modelo predictivo que permita conjeturar la situación de pobreza de los hogares de Colombia teniendo como base una encuesta rápida y no muy costosa, basada en pocas preguntas concisas. Con este objetivo se evaluaron 5 tipos de modelos: regresión lineal, regresión logística, árboles de decisión, random forest y adaptative boostings. Luego de evaluar el poder predictivo en base a medidas como el accuracy y los scores ROC-AUC y F1, se detectó como debilidad principal de los modelos a la generación de falsos negativos. Teniendo en cuenta esto, se perfeccionaron los umbrales de clasificación de la pobreza y se seleccionó al modelo de adaptative boostings como el mejor predictor de la situación de pobreza de los hogares en Colombia.

2 Datos

Se utilizan datos del Empalme de las Series de Empleo, Pobreza y Desigualdad correspondientes a la Medición de Pobreza Monetaria y Desigualdad (2018) del Departamento Administrativo Nacional de Estadística de Colombia. Los datos surgen de encuestas asociadas a individuos y hogares divididos en cuatro bases de datos: dos training sets y dos test sets para cada unidad de observación.

Table 1: Bases de datos

Base	Observaciones	Cantidad de variables
Train personas	543.109	135
Test personas	219.644	63
Train hogares	164.960	23
Test hogares	66.168	16

El objetivo es construir un modelo predictivo que permita conjeturar la situación de pobreza de los hogares de Colombia teniendo como base una encuesta rápida y no muy costosa, basada en pocas preguntas concisas. Teniendo en cuenta el objetivo, se utilizaron datos correspodientes a hogares y a los individuos asociados a cada hogar con el objetivo de complementar al primer conjunto de datos. Para lograrlo, se unieron las bases de hogares y las de personas en base al identificador único id. De esta manera, las bases finales utilizadas fueron dos: un train set y un test set que asocian a individuos con los datos correspondientes al hogar que pertenecen. De

cualquier manera, como el test set final no presenta variables que nos permitan evaluar las estimaciones, se crearon un train y un test set dentro del training set para evaluar las predicciones de los modelos. Como todo el análisis fue desarrollado en base al training set y sus subconjuntos y sobre el test set principal sólo se aplicaron los resultados obtenidos en base al trabajo previo, de aquí en adelante al referirnos a la base y los datos nos estaremos refiriendo exclusivamente al training set y sus subconjuntos.

El objetivo es complejo ya que se cuenta con un número relativamente reducido de variables no monetarias. Por lo tanto, se precisa analizar la relación de estas variables con la variable que buscamos aproximar: el ingreso per cápita familiar. En los modelos, se pretende utilizar variables correlacionadas con el ingreso per cápita familiar que nos permitan predecir la situación de pobreza de los hogares.

Se consideró el ingreso per cápita familiar como la variable objetivo luego de analizar el resto de las variables monetarias y considerar a ésta como la mas adecuada dadas las metas. En comparación con el ingreso total del hogar y con las variables de ingreso por individuo, el ingreso per cápita familiar permite reescalar el ingreso teniendo en cuenta la cantidad de miembros del hogar que deben compartir ese ingreso. Toda variable que no considere la cantidad de miembros de un hogar corre el riesgo de sobre-estimar el potencial de ese ingreso. De cualquier manera, resulta útil estudiar al resto de las variables porque pueden funcionar como proxies del ingreso per cápita familiar objetivo.

En principio, se consideraron para el análisis sólo las variables de ingreso con imputaciones por faltantes, extremos y arriendo para evitar subestimaciones o sobreestimaciones basadas en las imperfecciones a las cuales toda encuesta está expuesta como la declaración de ingresos sin considerar costos de un potencial arriendo, la no respuesta y el faltante de ingresos extremos.

El ingreso per cápita familiar posee asimetría positiva, con una media (711.049) significativamente alejada de la mediana (460.000). El ingreso máximo observado para el total del hogar coincide con el ingreso per cápita familiar máximo, lo cual indica que el hogar con mayor ingreso de la muestra esá compuesto por un único miembro. Estas distribuciones pueden observarse en los histogramas presentes en el Anexo 1.

Por otro lado, se puede observar que el ingreso de la actividad principal presenta una media (1.265.281) y una mediana (850.000) mucho mayores que las del ingreso de la actividad secundaria (608.732 y 250.000) lo cual puede estar reflejando la complementariedad de las actividades o una tendencia a menores ingresos promedios por actividad ante la presencia de una actividad secundaria. De la misma manera, los ingresos en especie, por ayuda de hogares y por ayuda de instituciones presentan medias y medianas relativamente bajas, indicando que estos ingresos colaboran poco en el ingreso total familiar y son potenciales indicativos de situaciones de pobreza. Por el contrario, ingresos por jubilaciones y pensiones, de desocupados e inactivos, intereses o dividendos y arriendos tienen mayor potencial para generar un efecto positivo significativo sobre el ingreso total disponible en un hogar y, por lo tanto, evitar situaciones de pobreza.

Table 2: Bases de datos

Estadisticos descriptivos	Media	Desvío Estándar	Mínimo	Máximo	Mediana
Ingreso total del hogar	2.503.725	2.640.962	4.167	88.833.333	1.800.000
Ingreso per cápita familiar	711.049	960.146	2.083	88.833.333	460.000
Ingreso de la actividad principal individual	1.265.281	1.711.962	6.667	52.000.000	850.000
Ingreso de la actividad secundaria individual	608.732	1.656.914	6.000	48.000.000	250.000
Ingreso en especie individual	246.371	284.032	2.000	8.000.000	180.000
Ingreso de desocupados e inactivos individual	717.752	746.657	7.000	6.000.000	600.000
Ingreso por intereses o dividendos individual	643.318	1.795.105	83	25.000.000	166.667
Ingreso por jubilaciones y pensiones individual	1.882.954	1.556.259	704.000	18.000.000	1.375.000
Ingreso por ayudas de hogares individual	357.441	649.108	1.667	30.000.000	200.000
Ingreso por ayudas de instituciones individual	70.841	63.244	2.500	800.000	65.000
Ingreso por arriendos individual	902.848	1.392.386	30.000	30.000.000	550.000
Ingreso total individual	1.145.262	1.693.550	83	55.000.000	780.000

Fuente: Elaboración propia en base a datos del Empalme de las Series de Empleo, Pobreza y Desigualdad (2018).

Teniendo en cuenta esto, se calcularon un conjunto de variables consideradas relevantes para aproximar la situación de pobreza de los hogares que podrían ser preguntadas directamente al jefe de hogar de forma sencilla y rápida, reduciendo los costos de las encuestas y manteniendo la confianza en los resultados.

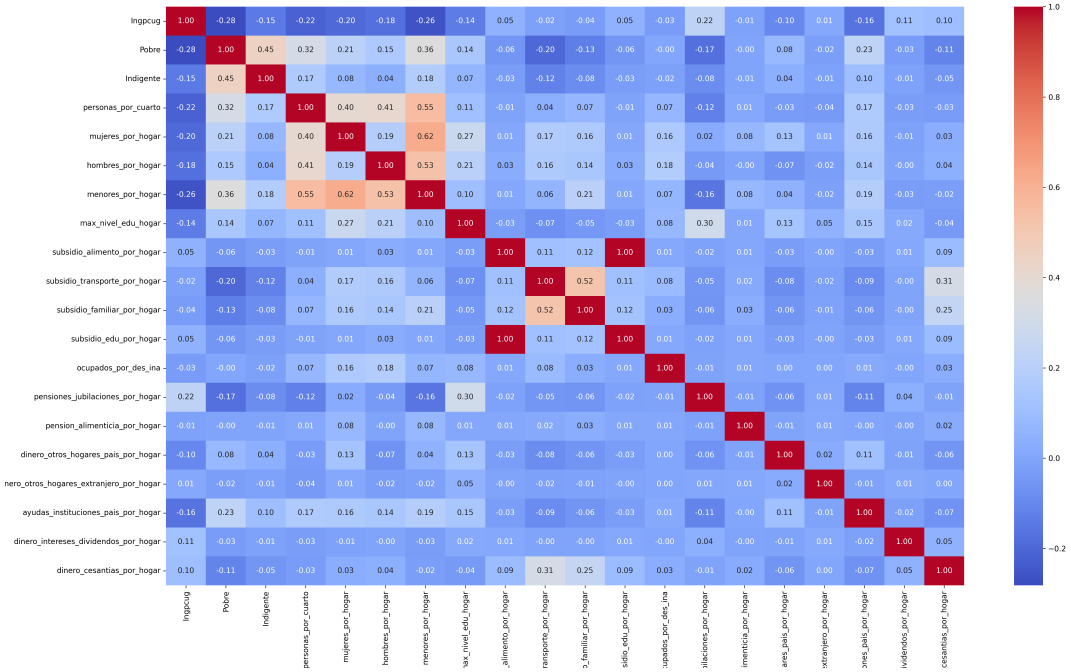
El conjunto de variables calculadas y utilizadas en los modelos es el siguiente:

- Personas por cuarto: Cantidad de miembros del hogar por cuartos para dormir
- Mujeres: Cantidad de mujeres en el hogar
- Hombres: Cantidad de hombres en el hogar
- Menores: Cantidad de menores de 18 años en el hogar
- Máximo nivel educativo: Máximo nivel educativo alcanzado por algún miembro del hogar
- Subsidio de alimento: Dummy que vale 1 si en el hogar se percibe al menos un subsidio de alimento
- Subsidio de transporte: Dummy que vale 1 si en el hogar se percibe al menos un subsidio al transporte
- Subsidio familiar: Dummy que vale 1 si en el hogar se percibió en el último mes al menos un subsidio familiar
- Subsidio educativo: Dummy que vale 1 si en el hogar se percibió en el último mes al menos un subsidio educativo
- Ocupados por desocupados e inactivos: Ratio de ocupados sobre desocupados e inactivos
- Jubilaciones y pensiones: Dummy que vale 1 si en el hogar se percibe al menos un ingreso por jubilaciones y/o pensiones
- Pensión alimenticia: Dummy que vale 1 si en el hogar percibió en los últimos 12 meses al menos un ingreso por pensión alimenticia

- Dinero de otros hogares o instituciones en el país: Dummy que vale 1 si en el hogar percibió en los últimos 12 meses dinero de otos hogares o personas residentes en el país
- Dinero de otros hogares o instituciones fuera del país: Dummy que vale 1 si el hogar percibió en los últimos 12 meses dinero de otros hogares o personas residentes fuera del país
- Ayuda instituciones: Dummy que vale 1 si en el hogar se percibió en los últimos 12 meses al menos un ingreso por ayuda de instituciones del país
- Intereses o dividendos: Dummy que vale 1 si en el hogar se percibió en los últimos 12 meses al menos un ingreso por intereses de préstamos, depósitos de ahorros, utilidades, ganancias o dividendos
- Cesantías: Dummy que vale 1 si en el hogar se percibió en los últimos 12 meses al menos un ingreso en concepto de cesantías y/o intereses de cesantias

Finalmente, se evaluó la correlación entre estas variables y la variable de interés a través de una matriz de correlación. En esta, podemos observar que la cantidad de menores en el hogar se correlaciona positivamente con la situación de pobreza de los hogares y que la cantidad de hombres, mujeres y menores en el hogar se correlacionan negativamente con el ingreso per cápita familiar (lógicamente ya que forman parte del denominador). La cantidad de personas por cuarto y las ayudas de instituciones también se correlacionan negativamente con el ingreso per cápita.

Figure 1: Matriz de correlación



3 Modelos y resultados

Se han utilizado cinco métodos distintos para predecir la situación de pobreza de los hogares de Colombia en el 2018:

- Regresión lineal
- Modelo logístico
- Árbol de decisión
- Random Forest
- Adaptative Boosting

Con el objetivo de comparar el poder predictivo de cada modelo se utilizaron las mismas variables predictivas con todos los métodos. Las variables utilizadas son aquellas que fueron creadas con éste propósito y que fueron descritas en la sección anterior.

Para todos los modelos excepto la regresión lineal se consideró un umbral de 0.5 para determinar la condición de pobreza de los hogares. Es decir, si un modelo arroja un probabilidad de ser pobre de 0.51, entonces se lo clasifica como pobre. Si, por el contrario, el modelo arroja una probabilidad de 0.49, se lo considera no-pobre.

A continuación, se desarrolla el planteamiento y los resultados de cada modelo:

3.1 Regresión lineal

Se parte de un clásico modelo lineal

$$IPCF = X\beta + u$$

donde la variable independiente es el ingreso per cápita familiar.

Este modelo será considerado el bechmark. Se espera un poder predictivo bajo debido a las fuertes limitaciones lineales que aplica sobre el efecto de las variables independientes sobre la dependiente.

Table 3: Regresión lineal

Variables	**Coeficientes**	**P-Value**
Personas por cuarto	277.873.16	0.000
Mujeres	80.660,74	0.000
Hombres	80.582,45	0.000
Menores	-326.696,32	0.000
Máximo nivel educativo	39.422,61	0.000
Subsidio de alimento	249.853,40	0.000
Subsidio de transporte	-9.299,36	0.217
Subsidio familiar	48.134,53	0.000
Subsidio educativo	249.853,40	0.000
Ocupados por desocupados e inactivos	-310.234,68	0.000
Jubilaciones y pensiones	674.387,93	0.000
Pensión alimenticia	177.523,29	0.000
Dinero de otros hogares o instituciones en el país	7.855,21	0.266
Dinero de otros hogares o instituciones fuera del país	155.826,48	0.000
Ayuda instituciones	233.286,03	0.000
Intereses o dividendos	1.165.366,35	0.000
Cesantías	396.390,53	0.000

Podemos observar que todas las variables excepto las dummies de subsidio al transporte y de dinero de otros hogares dentro del país resultan estadísticamente significativas para explicar el ingreso per cápita familiar. El R-Cuadrado es de 0.33, por lo que se espera que el poder predictivo sea bajo. En base a estas estimaciones, se aplican los coeficientes estimados sobre el test set. Se comparan los ingresos estimados con la línea de pobreza correspondiente y se determina la condición de pobreza de cada hogar en base a su diferencia. Finalmente, se comparan las predicciones con las situaciones observadas obteniendo un accuracy de 77,78% y la siguiente matriz de confusión:

Table 4: Matriz de confusión: Regresión lineal

		Predicción	
		0	1
Real	0	35728	3924
	1	7072	2764

3.2 Modelo logístico

Se estima un clásico modelo logístico

$$\hat{p}_i = \frac{e^{X_i\hat{\beta}}}{1 + e^{X_i\hat{\beta}}}$$

donde se busca predecir la probabilidad de que un hogar sea considerado pobre. La estimación se itera hasta 500 veces con el objetivo de optimizar el modelo y las variables son previamente normalizadas para facilitar este proceso.

El modelo final es utilizado para predecir las probabilidades asociadas al test set, arrojando un accuracy de 83.89% y la siguiente matriz de confusión:

Table 5: Matriz de confusión: Regresión logística

		Predicción	
		0	1
Real	0	35728	1791
	1	6181	3655

3.3 Árbol de decisión

Se crea el modelo de árbol de decisión con una profundidad máxima de 5 niveles, para evitar un proceso de overfitting dada la baja complejidad de los datos y se define al coeficiente de gini como el criterio de agrupamiento del árbol.

Esta metodología nos arroja 32 nodos terminales que recursivamente forman parte de un arbol de 5 niveles de profundidad.

En base a éste resultado se lleva a cabo un proceso de prunning poco exitoso. El modelo original no sufre de sobreajuste, motivo por el cual se obtiene un alpha óptimo de 0, generándose un nuevo árbol de decisión con ahora 9877 nodos terminales y una profundidad de 32 niveles (serio overfitting) que genera efectos negativos sobre el poder predictivo.

El accuracy del primer árbol de decisión sobre el test set es de 83.47% y el del segundo de 82.76%. Sus correspondientes matrices de confusión son:

Table 6: Matriz de confusión
Árbol de decisión con 5 niveles y 32 nodos terminales

		Predicción	
		0	1
Real	0	35728	1791
	1	6181	3655

Table 7: Matriz de confusión
Árbol de decisión con 32 niveles y 9877 nodos terminales

		Predicción	
		0	1
Real	0	35728	1791
	1	6181	3655

3.4 Random Forest

A continuación, se llevó a cabo una predicción basada en un Random Forest, donde se evalúan 100, 150 y 200 árboles, con un máximo de 5, 7 y 9 características para la división de los nodos y una profundidad de 10 y 20 niveles. El criterio para definir las divisiones es el Coeficiente de Gini. Estos hiperparámetros son seleccionados en base a una busqueda aleatoria, tomando la mejor combinación basada en una evaluación de los resultados en base al accuracy. Se decidió realizar la selección en base a una búsqueda aleatoria en lugar de una prueba de todas las combinaciones posibles (grid search) para reducir el costo computacional del procedimiento. El proceso de evaluación de combinaciones aleatorias se itera 10 veces.

Finalmente, se seleccionan 150 árboles con un máximo de 7 características para la división de los nodos y una profundidad máxima de 10 niveles.

De esta manera, se realizan las predicciones pertinentes sobre el test set, definiendo a las variables Menores por Hogar, Personas por Cuarto y Subsidio de Transporte por Hogar como las tres variables de mayor importancia (30%, 17% y 15% correspondientemente) en la determinación del output.

Esta predicción arroja un accuracy de 84.02% y la siguiente matriz de confusión:

Table 8: Matriz de confusión: Random Forest

		Predicción	
		0	1
Real	0	37980	1672
	1	6236	3600

3.5 Adaptative Boosting

Finalmente, se lleva a cabo una predicción que surge de un algoritmo de Adaptative Boosting que entrena a un modelo tomando secuencialmente 300 árboles de decisión de profundidad 1, asignándoles una ponderación en base a su rendimiento y combinándo sus predicciones para generar un resultado final.

Este proceso aplicado sobre el test set arroja un accuracy de 84.05% y la siguiente matriz de confusión:

Table 9: Matriz de confusión: Random Forest

		Predicción	
		0	1
Real	0	37509	2143
	1	5749	4087

4 Evaluación de los resultados

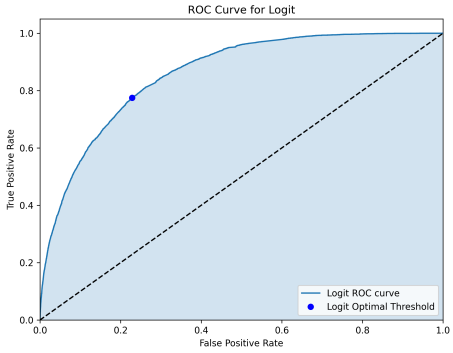
Las matrices de confusión permiten dar cuenta del hecho de que la mayor debilidad de los modelos surge de la generación de falsos negativos.

En esta instancia, nos quedamos con los 4 modelos con mayor accuracy (regresión logística, árbol de decisión, random forest y adaptative boosting) y calculamos medidas alternativas al accuracy para detectar mejor sus problemas y buscar una solución.

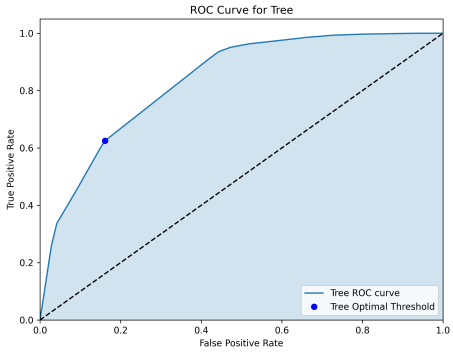
Comenzamos por evaluar el score ROC-AUC de los modelos:

ROC-AUC

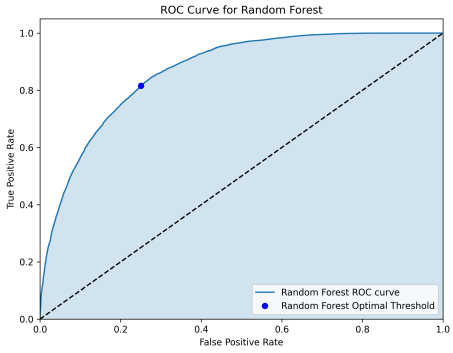
(a) Regresión logística



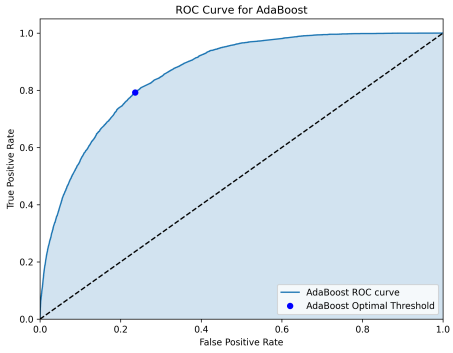
(b) Árbol de decisión



(c) Random Forest



(d) Adaptative Boosting



El área debajo de la curva (AUC) es mayor al 80% en todos los casos, indicando una gran capacidad general de clasificación. Sin embargo, este tipo de score está sesgado ante la presencia de clases desbalaceadas como las que existen en este tipo de análisis. Es necesario, considerar un score que tenga en cuenta el hecho de que las personas

clasificadas como pobres representan una proporción baja de la población.

El F1 score tiene en cuenta no sólo la proporción de positivos/negativos correctamente/incorrectamente identificados sino también la proporción de predicciones positivas (es pobre) correctas respecto a todas las predicciones positivas, permitiendo dar cuenta de la sensibilidad del modelo a la generación de valores positivos.

Table 10: F1 Score

Modelo	**Cantidad de variables**
Regresión logística	0.48
Árbol de decisión	0.45
Random Forest	0.48
Adapatative Boosting	0.51

Indicando serios problemas asociados al desbalance en las clases.

Para solucionarlo, se estiman umbrales óptimos basados en la idea de minimizar la distancia euclidiana de la curva ROC al punto ideal (True Positive Rate = 1 y False Positive Rate = 0).

Esto nos arroja los siguientes umbrales óptimos para cada modelo:

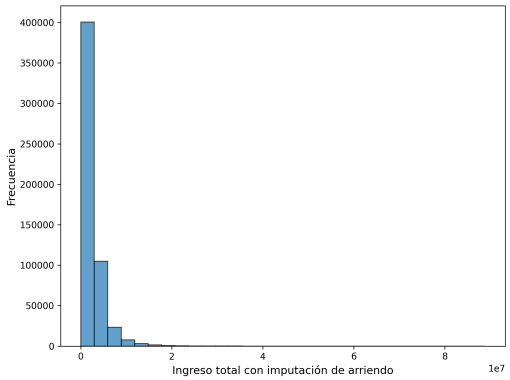
- Regresión logística: 0.22
- Árbol de decisión: 0.24
- Random Forest: 0.20
- Adaptative Boosting: 0.50

Incrementando el poder predictivo de los modelos logístico, random forest y adaptative boosting en términos del F1 score en casi 0.1.

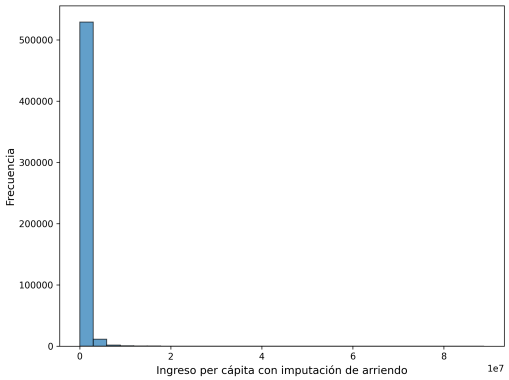
Al evaluar los resultados sobre el test set original, concluimos que el modelo de adaptative boosting es el que mejor predice la situación de pobreza de los hogares en Colombia.

Anexo 1: Histogramas

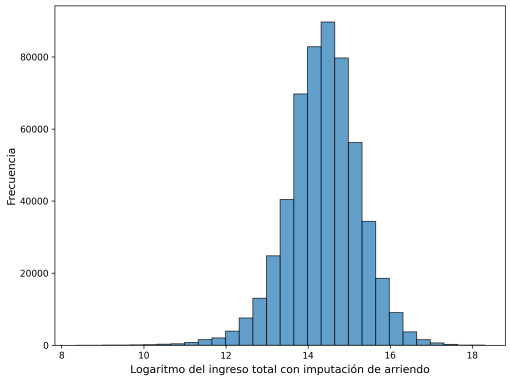
(a) Ingreso total del hogar



(b) Ingreso per cápita.



(c) Logaritmo del ingreso total del hogar



(d) Logaritmo del ingreso per cápita