PROBLEM SET 2: PREDICTING POVERTY

MARIA CAMILA CELY MORENO – SARA OSPINA GIRALDO

Repositorio Github: https://github.com/SaraOspinaG/ProblemSet2_Cely_Ospina

INTRODUCCIÓN

El desarrollo del presente taller tiene como objetivo principal construir un modelo que **prediga la pobreza a nivel de hogares en Colombia** con el menor número de variables posible, de manera tal que se aporte a la identificación de hogares pobres de forma rápida y poco costosa. Para ese fin, y a partir de los datos de la GEIH[1] (DANE, 2019), se entrenan distintos modelos, tanto de clasificación como de regresión, con el fin de aproximar la mayor cantidad de verdaderos positivos posible. Cabe resaltar que la GEIH es una encuesta mediante la cual se solicita información directamente a los hogares, y se utiliza para medir y reportar indicadores relacionados con mercado laboral, ingresos y pobreza monetaria, por lo cual se puede afirmar que es una fuente confiable y con suficiente información para lograr el objetivo del ejercicio. Este ejercicio busca predecir la pobreza a través de dos canales principales. El primero tiene que ver con las características de la persona jefe de hogar, siendo especialmente relevante el hecho de que sea una mujer la jefe del hogar: los hogares con jefatura femenina tienen más probabilidad de estar por debajo de la línea de pobreza, y obtienen menores ingresos totales aun cuando, en promedio, dichas jefes de hogar tienen más educación que los hombres. El segundo tiene que ver con las características de la vivienda, y tienen especial relevancia el valor que los hogares destinan a pagar arriendo, y el número de habitaciones que los integrantes del hogar destinan a dormir. Ambos canales dan pistas de posibles caminos de política pública que contribuirían a la disminución de la pobreza, por un lado, encaminando acciones a la disminución de la brecha de género en el mercado laboral; y por el otro, encaminando acciones a la disminución del déficit cualitativo y cuantitativo de vivienda, así como a generar subsidios no solamente para la compra de vivienda sino para el arriendo de la misma.

DATOS

Los datos utilizados provienen de la GEIH, y contienen información tanto a nivel del hogar como a nivel del individuo. Se toma la decisión de no agregar los datos individuales por hogar, sino que solamente se toman los datos individuales de las personas jefes de hogar, ya que se considera que estos datos capturan de forma suficiente las características socioeconómicas del hogar como un todo. Teniendo en cuenta que no todas las variables de la GEIH aparecen en las muestras de train y test, se realiza una selección de las que se consideran más relevantes de acuerdo con la literatura y con lo analizado en el primer Problem Set. Se hace evidente que los hogares con jefatura femenina (valor de 1) tienen mayor probabilidad de estar por debajo de la línea de pobreza que los hogares con jefatura masculina (valor de 0). Adicionalmente, si bien las mujeres jefas de hogar tienen en promedio mayores niveles de educación, reciben en promedio menores ingresos y pagan mayores valores de arriendo. Lo anterior da cuenta de distintas brechas de género a nivel educativo, en el mercado laboral y quizá en el mercado inmobiliario, por lo cual en nuestro trabajo planteamos como una de las variables explicativas principales el hecho de que el hogar tenga jefatura femenina.

[1] Gran Encuesta Integrada de Hogares

Characteristic	0, N = 88,956[1]	1, N = 63,026[1]
Hogar clasificado como pobre (1)	15,519 (17%)	13,700 (22%)
Básica primaria (1o - 5o)	24,406 (27%)	17,996 (29%)
Superior o universitaria	23,589 (27%)	17,700 (28%)
Edad jefe de hogar	47 (36, 59)	50 (37, 62)
Valor arriendo / num personas hogar	123,333 (75,000, 200,000)	133,333 (75,000, 233,333)
Ingreso total unidad de gasto con imputacion de arriendo	1,700,000 (984,822, 2,914,808)	1,488,367 (858,558, 2,553,976)

[1] n (%); Median (IQR)

La variable de si el hogar es clasificado como “Pobre” según el DANE se obtiene directamente de la GEIH, sin embargo, es posible reproducir esta clasificación utilizando la variable de Ingreso total de unidad de gasto con imputación de arriendo (Ingtotugarr) y definiendo si este ingreso es menor a la Línea de pobreza (Lp) definida para el departamento, multiplicada por el Número de personas del hogar (Nper). De esta manera, ya que se utilizará Ingtotugarr como variable dependiente de los modelos de regresión a plantear, se construye la variable de valor de arriendo, por medio de la combinación de las columnas en las que las personas que realmente pagan arriendo reportan dicho valor, y las personas que no pagan arriendo manifiestan el valor estimado que pagarían si lo hicieran. Por último, se construyen las variables de arriendo per-personas por hogar y N°cuartos per-personas por hogar, de esta forma se controla por el tamaño de las unidades de gasto y como el ingreso se distribuye.

Con respecto a la proporción de hogares pobres en las muestras a utilizar (training, evaluation y testing), se comprueba esta corresponde a alrededor del 19% de la muestra en todos los casos. Por lo tanto, se genera una muestra UpSampled para lograr que esta proporción sea de 50%, lo cual se utilizará en las pruebas de mejores modelos de predicción.

	Training	Evaluation	Testing	UpSampled
Characteristic	**N = 121,586[1]**	**N = 10,132[1]**	**N = 20,264[1]**	**N = 196,174[1]**
Hogar clasificado como pobre (1)	23,499 (19%)	1,858 (18%)	3,862 (19%)	98,087 (50%)

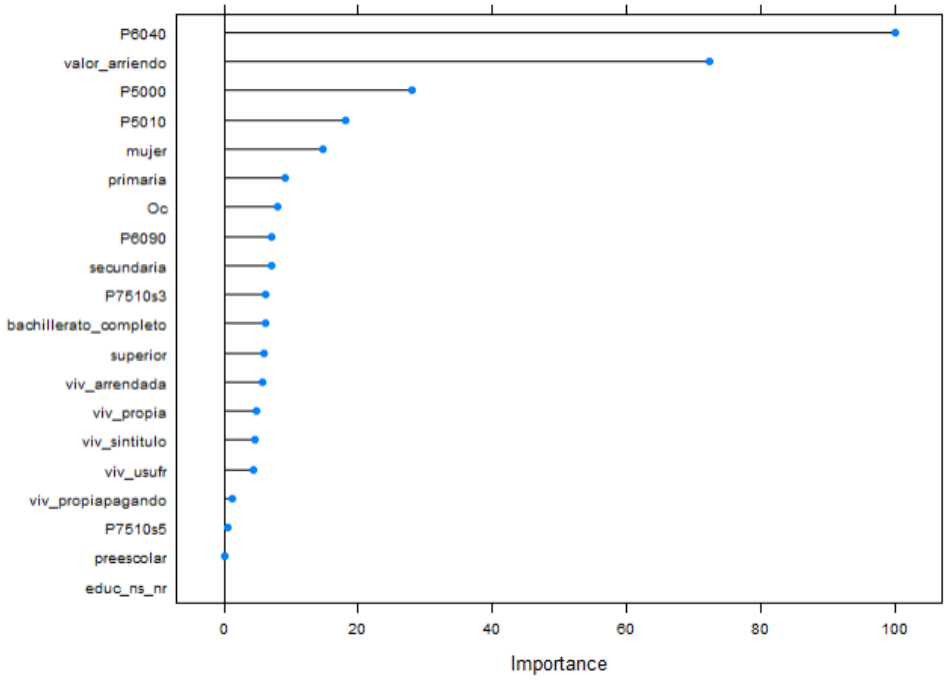
[1] n (%)

MODELOS Y RESULTADOS

MODELOS DE CLASIFICACIÓN

Con el fin de predecir de la mejor manera posible si los hogares se pueden clasificar como pobres o no, se toma como variable dependiente la variable Pobre suministrada por la GEIH, que toma valor de 1 cuando el hogar tiene un ingreso total por debajo de la línea de pobreza definida para el departamento en el que se ubica. Adicionalmente, se separa la muestra de train en tres sub-muestras (training, evaluation y testing) con el fin de reducir desbalanceo en la muestra, se hace una partición inicial de 80%-20%.

En primer lugar, la totalidad de las primeras variables explicativas escogidas se corren en un modelo de **Random Forests** que permita aproximarnos a la Variable Importance, y así, escoger de entre estas las más relevantes para correr los modelos subsiguientes.



A partir de este ejercicio, se observa que las variables identificadas como más importantes son la edad del jefe del hogar, el valor de arriendo que pagan, el número de cuartos en los que duermen, si el jefe de hogar es mujer y su nivel educativo. El modelo de RF, al ser probado por fuera de muestra, arroja una Sensitivity de 0,43.

Posteriormente, se plantean dos versiones de modelos en boosting mediante la metodología **XGBoost**, el primero con seis variables explicativas (número de personas por cuarto, edad del jefe de hogar, valor del arriendo, si el jefe de hogar es mujer, si el jefe de hogar está empleado y si el jefe de hogar cursó solamente la primaria); y el segundo con solamente cuatro variables explicativas (valor del arriendo por persona, número de personas por cuarto, si el jefe de hogar es mujer y la edad del jefe de hogar). Ambos modelos se entrenan con la muestra **UpSampled.** El segundo modelo se plantea con el fin de reducir las variables explicativas, ya que XGBoost castiga los modelos entre más variables incluyan. Sorprendentemente, los dos modelos se comportan muy semejante prediciendo fuera de muestra, obteniendo Sensitivity de 0,39 y 0,38 respectivamente.

Debido a que una de las instrucciones del taller especifica que se debe usar la menor cantidad de variables posibles en la predicción, y que las Sensitivities de ambos modelos están muy semejantes, se escoge el modelo de cuatro variables para analizar su comportamiento al **cambiar el Cutoff.** Mediante el análisis de ROC, se concluye que el mejor Threshold para este caso es de 0,46.

Con base en lo encontrado respecto a la importancia de las variables en los ejercicios anteriores, en primer lugar, se corrieron cuatro variaciones de un modelo Logit, para los primeros tres se complejizaron las variables y para el cuarto de entrenó con **Upsample** (sens dentro de muestra del 1 al 4: 0.09<0.24<0.25<0.75). Posteriormente, se probó correr **Logit con ElasticNet entrenada en Upsample** (sens dentro de muestra: 0.1103) y tres modelos adicionales de **Logit con Lasso**, los primeros dos entrenados con la muestra normal y el tercero con Upsample(sens dentro de muestra del 1 al 3: 0.19< 0.19<0. 0.75). Por último, se intentó **cambiar el Cutoff** de este último modelo para optimizar la sensibilidad y la especificidad, pero se encuentra que la proporción de verdaderos positivos es menor y la capacidad predictiva es muy baja. A pesar de estos resultados dentro de muestra, se corren también fuera de muestra para probar su capacidad predictiva. Con base en esto se encuentra que la Sensitivity de los modelos cambia significativamente:

nombre modelo	sensitivity
logit_enet	0.772695161
logit4	0.770357596
lasso_ups	0.766461654
xgboost min	0.734514929
xgboost	0.713996301
xgboost cutoff	0.372971513
random forests	0.345440189
logit3	0.270378374
logit2	0.259729466
lasso1	0.20882249
lasso2	0.194017911
logit1	0.109865564

nombre modelo	sensitivity	specificity	num var
logit_enet	0.772695161	0.71659	8
logit enet 4var	0.813732416	0.632271	4
logit enet 3var	0.816069982	0.62624	3

De acuerdo con estos resultados, se observó que **Logit con Elastic Net entrenada en Upsample con métrica de Sensitivity,** el cual obtuvo la mejor proporción entre Sensitivity y Verdaderos Positivos en relación con los demás. Con base en esto, se corrieron dos modelos adicionales con las mismas especificaciones, pero con menos variables. Finalmente, se escogió el modelo con 4 variables y la mejor proporción entre Sensibility y Specificity.

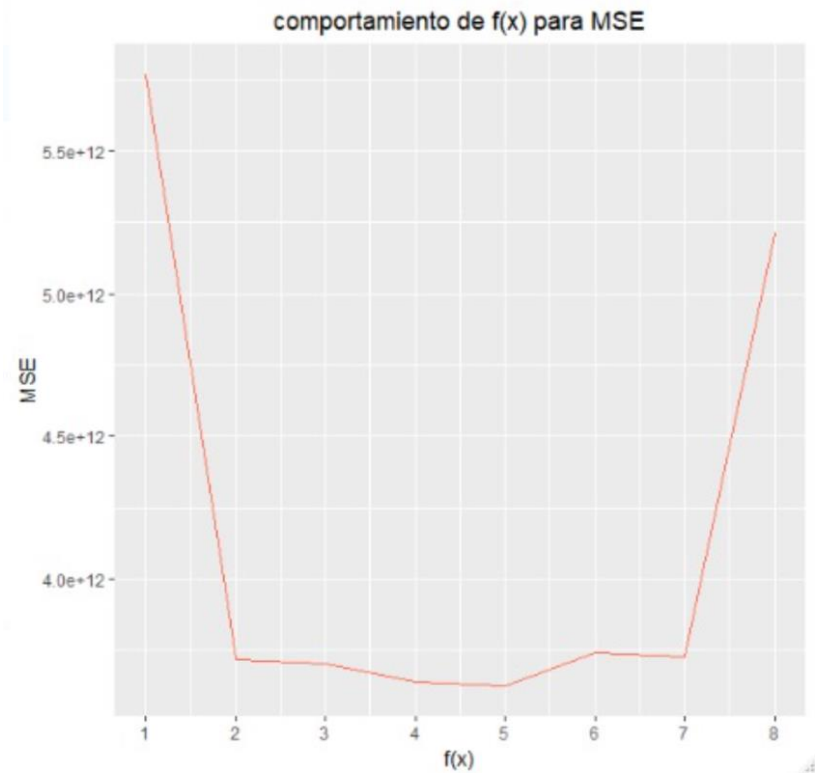
MODELOS DE REGRESIÓN

Para el ejercicio de predicción del ingreso, con base en lo encontrado en el ejercicio anterior, se corrieron siete modelos de mínimos cuadrados ordinarios, modificando la complejidad del modelo en cuanto a las variables incluidas y sus interacciones en cada uno de ellos. Posteriormente, se probaron estos modelos fuera de muestra y se obtuvo los MSE de cada modelo, previo a escoger el modelo final. Se puede observar en la siguiente gráfica el comportamiento de cada uno de los modelos, partiendo del menos complejo al más complejo. De acuerdo con lo encontrado, el modelo 5 es el que tiene un MSE fuera de muestra más bajo, pero, el modelo 4 que cuenta con un MSE bastante similar se corrió con 3 variables menos. De esta forma, se escoge el modelo 4 pues tiene una capacidad predictiva similar pero menos variables. (ver gráfica 1 en anexo). Por último, se prueba este modelo en la muestra de prueba (test_final) donde se obtienen los resultados de la predicción de ingreso.

CONCLUSIONES Y RECOMENDACIONES

- Se encuentra evidencia de que los modelos de clasificación, a pesar de tener un buen ajuste dentro de la muestra, esto no necesariamente se mantiene cuando se prueba fuera de muestra.
- Inicialmente, se tenía como hipótesis que ser mujer sería una de las variables que más importan dentro del modelo. A pesar de que se encontró evidencia de que, si importa, no está cerca de ser la principal.
- Se obtiene un resultado inesperado al ser la variable de edad que más influencia nuestra y. Esto puede sugerir recomendaciones de política en el marco de la seguridad pensional, ya que se encontró que es más probable que un hogar sea pobre a medida que aumenta la edad del jefe de hogar.
- Por último, como se pensaba en un inicio, se encuentra que las condiciones del hogar son determinantes para la condición de pobreza. Esto sugiere que las condiciones de pobreza se ven fuertemente relacionadas con la tenencia o el arriendo de vivienda y las condiciones habitacionales de esta como el número de cuartos en relación con el número de personas que lo habitan. Esta conclusión es clave para el diseño y la implementación de política pública de vivienda, ya que se recomienda concentrar esfuerzos en la reducción del déficit cuantitativo y cualitativo de la vivienda, al ser esta principal herramienta para la disminución de pobreza en el País.

APÉNDICE



Gráfica 1