Universidad de
SanAndrés

Big Data

Trabajo Practico N°2

Pilar Ruiz Orrico
Sebastián Einstoss Mastracchio
Sofía Marinkovic Dal Poggetto

Septiembre 2021

1. Parte I: Analizando la base

1.1. Ejercicio 1

En Argentina, para clasificar si una persona es pobre o no el INDEC utiliza la metodología de línea de pobreza. En términos simples, se genera una canasta una canasta básica total (CBT) que funciona como umbral y luego se comparan los ingresos de los individuos con el valor de la CBT. Si el ingreso es mayor se consiera a la persona como "no pobre", caso contrario, será considerada "pobre".

Ahora bien, la CBT no se arma definiendo qué y cuánto consume cada individuo sino que se construye a partir de la canasta básica alimentaria (CBA). Esta última está armada según los requerimientos calóricos de un *adulto equivalente*. Para obtener la CBT de un adulto equivalente alcanza con multiplicar la CBA por la inversa del coeficiente de engel (ICE) que es la proporción de gastos totales sobre los gastos en alimentos. El coeficiente es calculado a partir de los datos de la encuesta de nacional de gastos de los hogares (ENGHo). Tanto la CBA como la CBT dependen de los requerimientos calóricos de los individuos, por lo tanto deben ajustarse para cada persona.

Vale destacar que si bien se tiene en cuenta los requerimientos calóricos, el armado de la canasta tiene en consideración los consumos "típicos"de las personas. Es decir, la misma no surge de minimizar el gasto sujeto a la restricción calórica, sino que se tienen en cuenta cuestiones como los hábitos de consumo prevalecientes.

1.2. Ejercicio 2

1.2.1. Ver en script

1.2.2. Ver en script

1.2.3.

Al analizar la composición de la muestra, se observa una mayor cantidad de mujeres que hombres. Esto refleja tanto algo que puede ser una caracteristica de la sociedad como de la muestra en sí. Dado que no estamos trabajado con la extrapolación de la misma (*pondera*) a la población relevante, esto se debe analizar en torno a la muestra.

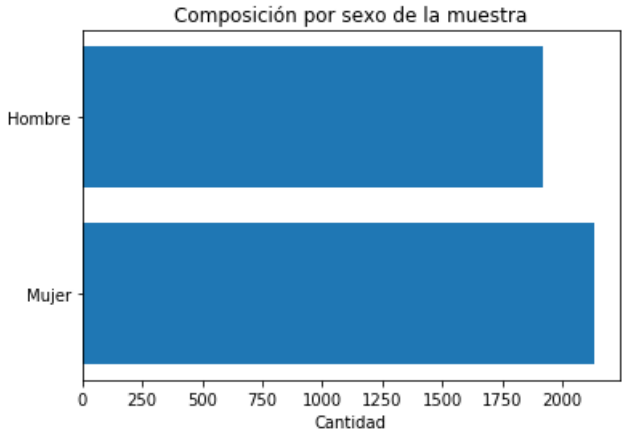

Figura 1: Gráfico de barras de la composición de la muestra

1.2.4.

La matriz de autocorrelación nos permite vislumbrar que variables presentan una correlación entre sí y que signo tiene la misma. Debido a que cada variable tiene una correlación de 1 consigo misma, es esperable que la diagonal principal tenga el color más oscuro.

En cuanto a las correlaciones positivas se destaca la relación entre el estado (*ESTADO*)y la categoría de inactividad (*CAT_INAC*). Esto tiene sentido ya que la categoría de inactividad es únicamente relevante cuando la persona en cuestión se encuentre inactiva. Por otro lado, se observa una correlación positiva entre el estado (*ESTADO*) y la categoría de inactividad (*CAT_INAC*) con el sexo (*CH04*) y la variable relacionada al estado civil (*CH07*). En cuanto al sexo esto podría ocurrir debido a cierta discriminación en el mercado laboral, o por cuestiones externas como la distribución de las tareas de cuidado. Finalmente, en cuanto a la relación con el estado civil, podría estar ocurriendo que personas con determinada condición presenten mayor proporción de gente ocupada (sospechamos que son los casados).

En cuanto a las correlaciones negativas se observa una relación entre el nivel educativo (*NIVEL_ED*) y el estado ocupacional (*ESTADO*), entre el nivel educativo y el estado civil (*CH07*) y entre el ingreso per cápita familiar (*IPCF*) y múltiples variables.

La relación entre el nivel educativo y el estado ocupacional podría explicarse fácilmente porque las personas con alto nivel educativo presentan mayor tasa de empleo que las de menor nivel educativo. El signo proviene que las personas empleadas presentan el menor numero dentro de la variable *ESTADO*. En cuanto a la relación entre el nivel educativo y el estado civil resulta díficil encontrar una posible explicación razonable.

Tal vez la parte más rica proviene de la relación positiva del ingreso per cápita familiar con el nivel educativo y la relación negativa con la categoría de inactividad (donde posiblemente los jubilados sean quienes ganen más dentro de los inactivos).

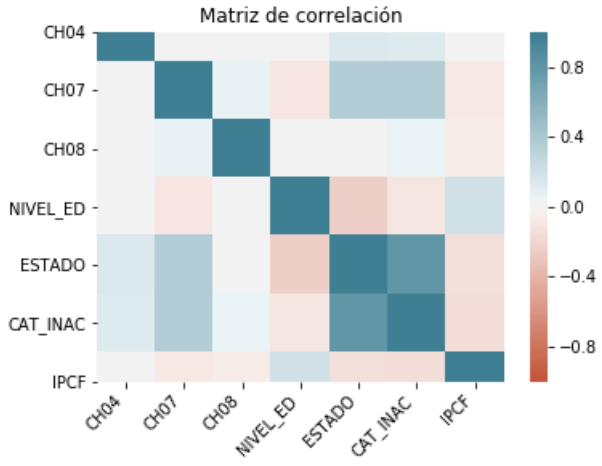

Figura 2: Matriz de correlación

1.2.5.

En la muestra se encontraron 213 desocupados, 1682 inactivos, 1737 ocupados.

1.2.6.

El ingreso per cápita familiar promedio es de 23.384,5 para los ocupados, 11.920,3 para los desocupados, y 16.739,6 para los inactivos.

1.2.7. Ver en script

1.3. Ejercicio 3

1.3.1.

En la muestra encontramos que 1549 personas no contestaron contra las 2500 que sí lo hicieron. Es un número elevado en nuestra opinión ya que representa casi un 40 % del total de la muestra.

1.3.2. Ver en script

1.4. Ejercicio 4

Luego de realizar todas las transformaciones necesarias, encontramos 838 personas pobres entre las que efectivamente respondieron.

2. Parte II: Clasificación

2.1. Ver en script

2.2. Ver en script

2.3. Ver en script

2.4. Algunas aclaraciones sobre el filtrado de las bases

Para poder trabajar con los diferentes modelos se realizó una limpieza más profunda de las bases dado que se contaba con algunas variables que eran string y muchas observaciones con algunos datos faltantes.

Esta limpieza se realizó en todas las bases: tanto en las de respondieronçomo en las de "norespondieron".

En primer lugar, se eliminó la columna CH05 que contenía fecha de nacimiento, y dado que ya se contaba con la edad de los individuos (CH06), se consideró que no era necesario transformala y mantenerla en entre las variables explicativas. Adicionalmente, se eliminó CODUSU, PP09A_ESP, PP09C_ESP que, como solo contenían variables strig, no se podía utilizar en los modelos. Una última variable que se optó por eliminar fue IMPUTA ya que no se pudo reemplazar los valores de no respuesta (NaN).

En segundo lugar, se convirtió a la variable MAS_500 en una dummy, que tomaba 1 cuando el valor original era **S** y cero en el caso contrario. La última transformación que se hizo sobre las bases fue la del manejo de los datos sin respuesta (NaN). En particular, con el objetivo de no perder observaciones, se optó por completar los missing values con la mediana de las observaciones de cada variable en cada base. De esta manera se consiguen dos cosas: por un lado, no se pierden observaciones, y por el otro lado, las variables dummy siguen funcionando de igual manera.

Una vez hecho esto, se realizaron las estimaciones y predicciones con los tres métodos.

2.5. ¿Cuál de los tres métodos predice mejor?

Para analizar cual de los tres métodos predice mejor la cantidad de pobres, en el apartado anterior, además de entrenar a los modelos y darles los datos para la predicción, se calcularon una serie de métricas. En particular, se calculó la tasa de presición de la predicción (accuracy rate, que mide del total de las predicciones cuántas fueron correctas), la matriz de confusión (para mirar los falsos positivos y los falsos negativos) y se graficó la curva de ROC de cada predicción (calculando el área bajo la curva, AUC, que implica que a mayor área mejor predicción).

En primer lugar, dado que se está prediciendo la cantidad de individuos que son pobres, lo que más preocupa son los falsos negativos. Si luego se desea realizar alguna política para reducir la tasa de pobreza es mejor contar con una sobreestimación, donde uno se asegura que los individuos verdaderamente pobres sean tenidos en cuenta a tener menos individuos de los que se buscaba captar.

En este sentido, a priori el mejor modelo pareciera ser el de análisis discriminante lineal ya que no solo tiene la menor cantidad total de falsos, 199, sino que tiene la menor cantidad de falsos negativos (77, contra 99 de vecinos cercanos y 156 del logit). Si embargo es necesario analizar las otras mértricas antes de tomar una decisión.

En segundo lugar, en lo que refiere a la tasa de presición, nuevamente es el método de análisis discriminante lineal el que tiene la mejor métrica: 0,73 contra 0,719 del logit y 0,673 de vecinos cercanos. Si bien las tasas no son muy altas, parecen adecuadas dado que el modelo se entrena solo sobre una parte de la muestra. Una mayor tasa implica una mayor presición en la estimación (en términos generales ya que no identifica sesgos hacia algun falso en particular).

Por último, en lo que refiera a la curva de ROC y el área bajo esta (AUC), la AUC del modelo de análisis discriminante lineal es la mayor (0,72 contra 0,66 de vecinos cercanos y 0,63 del modelo logit). De la misma manera, la curva de ROC que más se aleja de la línea de 45 grados es la de análisis discriminante lineal. A continuación se incluyen los gráficos de las tres predicciones.

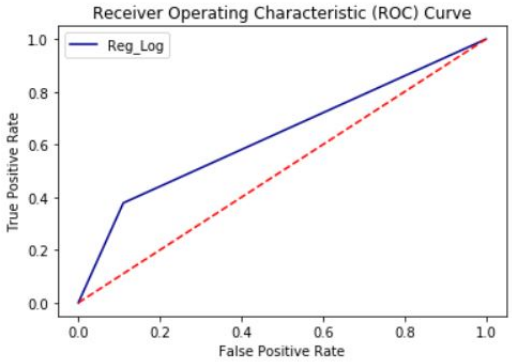

Figura 3: Logit

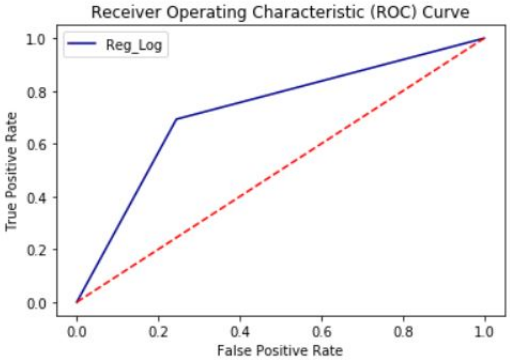

Figura 4: Análisis discriminante lineal

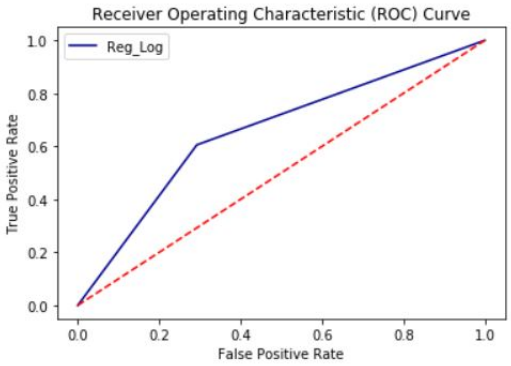

Figura 5: Vecinos cercanos

Dado que el método de análisis discriminante lineal presenta los mejores resultados para las tres métricas analizadas, se puede concluir que es el método que mejor predice la cantidad de pobres en la muestra.

2.6. Proporción de pobres en la base norespondieron

Utilizando el método de análisis discriminante lineal, se encontró que 597 individuos del total de 1549 serían pobres. Es decir, el 38,54 % de la muestra "norespondieron"se identificaría como pobre.

2.7.

La inclusión de una cantidad excesiva de variables podría inducir un probelma de *overfit*. Lo cierto es que esto puede conocerse ex-post y el hecho de quitar ciertas variables no nos garantiza que estemos solucionando el problema.

Dado que vamos a elegir solo un conjunto de variables para observar si la capacidad predictiva mejora, conservamos unicamente aquellas que creemos que son las primeras candidatas desde la literatura.

La primera variable a incluir es el nivel educativo (*NIVEL_ED*), ya que espesperable que personas con mayor formación obtengan mayores ingresos y, por lo tanto, tengan menos probabilidad de ser pobres.

La segunda variable es el estado ocupacional (*ESTADO*), donde las personas que no tienen empleo tienen un menor ingreso, lo que podría explicar porque están por debajo de la linea de pobreza.

La siguiente variable es el tipo de establecimiento en que trabaja la persona (*PP04*), el cual puede ser una variable importante para el nivel de ingresos.

Otra variable incluida es la intensidad con la que trabaja (*INTENSI*), la cual puede estar dando indicios sobre la estabilidad de la fuente de ingresos. Dado que esto mismo está relacionado con la informalidad, también incorporamos la variable *PP07H* sobre los aportes jubilatorios.

Observamos una reducción en la capacidad predictiva del modelo, lo cual en una primera instancia nos pareció sorprendente. Puede estar ocurriendo que nos falten incorporar ciertas variables que aumentarían fuertemente la capacidad predictiva, y es probable que despúes de determinado umbral esta comience a caer. Aún intentando con múltiples especificaciones no logramos obtener un mejor resultado.

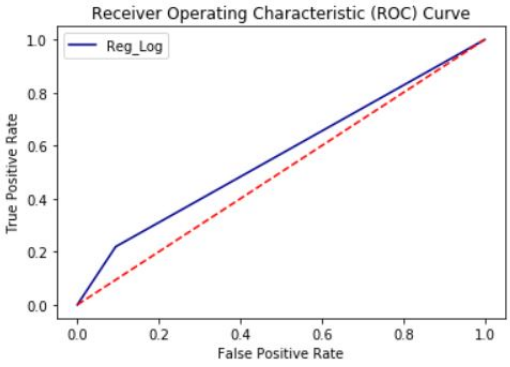

Figura 6: Logit con selección de variables