Trabajo Práctico 4 - Regularización aplicada a la EPH

Gil Deza, Hüppi Lo Prete, Walker

Show code

Parte I

Para esta parte recuperaremos lo realizado en los TPs anteriores.

1)

2)

Out[5]:

	CODUSU	ANO4_x	TRIMESTRE_x	NRO_HOGAR	REALIZADA
0	TQRMNOQSYHMOTOCDEIJAH00698520	2022	1	1	1
1	TQRMNOPSSHKKMMCDEIIAD00780111	2022	1	1	1
2	TQRMNOPSSHKKMMCDEIIAD00780111	2022	1	1	1
3	TQRMNORSUHLMNPCDEIIAD00718267	2022	1	1	1
4	TQRMNORSUHLMNPCDEIIAD00718267	2022	1	1	1
...	...	...	...	...	..
6701	TQRMNOPQSHMMKPCDEIJAH00780780	2022	1	1	1
6702	TQRMNOPWPHMLLLCDEIJAH00780781	2022	1	1	1
6703	TQRMNOPWPHMLLLCDEIJAH00780781	2022	1	1	1
6704	TQRMNORUVHLLKQCDEIJAH00718720	2022	1	1	1
6705	TQRMNORUVHLLKQCDEIJAH00718720	2022	1	1	1

6706 rows × 263 columns

Eliminamos todas las columnas duplicadas luego del merge, dado que se encontraban en ambas bases.

3)

Para la limpieza a conciencia de la base de datos eliminaremos, en primer lugar, aquellas variables que contengan más del 50% de sus observaciones con valores faltantes. Esto se debe a que consideramos que serán variables que no aportarán valor informativo suficiente a nuestros modelos predictivos, sino que traerán aparejados problemas para su estimación. Como nuestra base contiene en total 6706 observaciones, realizamos el corte de 50% en 3372 observaciones.

En segundo lugar, eliminamos los outliers de nuestras variables, dado que también pueden generar problemas en nuestras predicciones. Consideramos outliers todas las observaciones que están en el cuantil 1% superior e inferior.

En tercer lugar, chequeamos cuál es el contenido de las variables que nos generarían problemas al no ser numéricas:

```
MAS_500_x es una variable object
CH05 es una variable object
```

Observamos que CH05 contiene la fecha de nacimiento de los individuos (con lo cual podríamos borrarla, dado que tenemos su edad en la base también)

MAS_500_x contienen una letra que indica el tamaño del aglomerado. Al habernos quedado con las observaciones de BsAs y Gran BsAs, todas tienen una "S".

Debido a esto, consideramos que podemos deshacernos de tales variables sin mayores problemas.

A continuación, chequearemos y eliminaremos todas las observaciones que tengan valores negativos sin sentido, por ejemplo, para el ingreso o la edad:

```
0
5205          1          1            33       5409    1    3  ...     0
0
5287          1          1            33       1215    2    4  ...     0
0
5301          1          1            33       2573    1    2  ...     0
0
5576          1          1            32       2676    1    5  ...     0
0
5590          1          1            32       2824    2    2  ...     0
0
5621          1          1            33       5469    2    1  ...     0
0
5728          1          1            33       1427    2    2  ...     0
0
5930          1          1            33       1568    1    3  ...     0
0
6046          1          1            33       3719    1    2  ...     0
0
6357          1          1            33       2397    2    3          0
```

4)

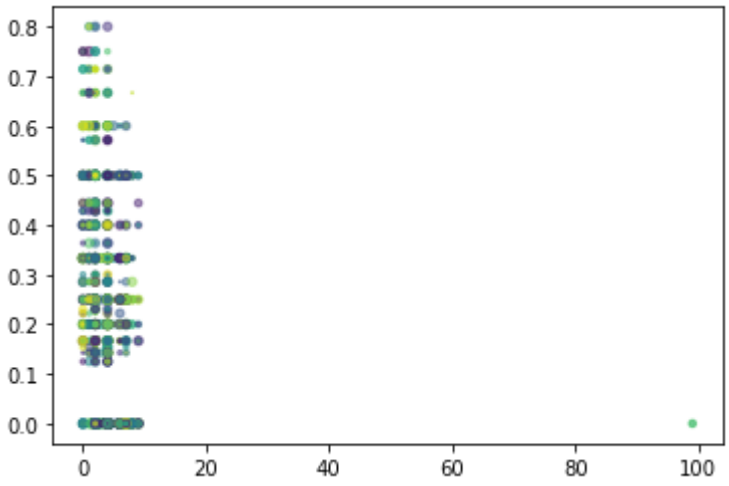
Construimos una variable que indique la proporción de niños menores de 10 años en el hogar:

Construimos una variable que identifique si el cónyuge del jefe de hogar trabaja:

Construimos una variable que indique si quien contesta la encuesta es migrante. Entendemos como migrantes a aquellos individuos que no son oriundos de la localidad en la que residen y además se desplazaron en los últimos 5 años.

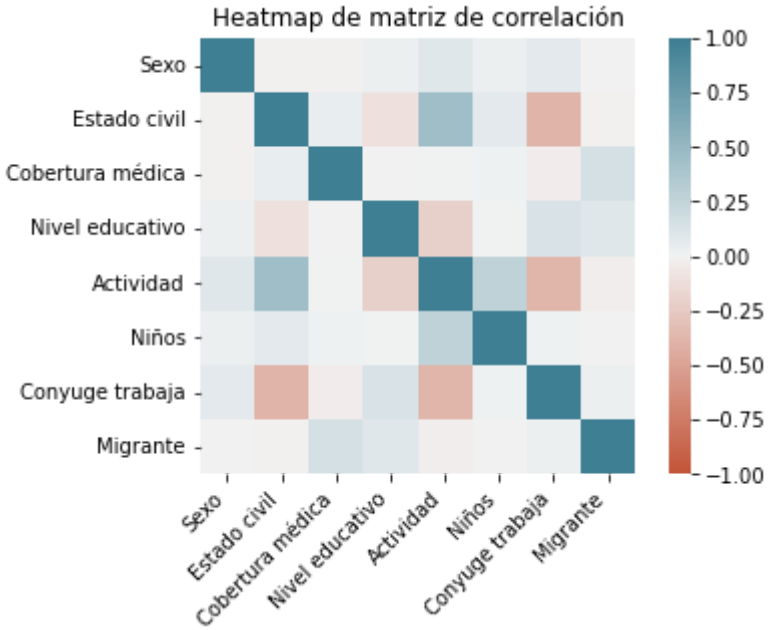
5)

Creamos un grafico de dispersión entre la proporción de niños en el hogar y el máximo nivel educativo de quien contesta



Out[16]:

```
Text(0.5, 1.0, 'Heatmap de matriz de correlación')
```



6)

```
La cantidad de personas que no respondieron cuál es su ingreso total familia
r es: 1542
1       435000
2       120000
3        15000
6       210000
8        63000
         ...
5224     22000
5225     16000
5226    160000
5228     53500
5230    282000
Name: ITF_x, Length: 3689, dtype: int64
0       0
4       0
5       0
7       0
14      0
       ..
5218    0
5219    0
5221    0
5227    0
5229    0
Name: ITF_x, Length: 1542, dtype: int64
1        69353.9820
2        96551.6220
3        97639.5276
6        89208.2592
8        96551.6220
           ...
5224    132180.5304
5225     60922.7136
5226     60650.7372
5228     61194.6900
5230     88664.3064
Name: ingreso_necesario, Length: 3689, dtype: float64
```

7)

8)

```
                            CODUSU  ANO4_x  TRIMESTRE_x  NRO_HOGAR  \
1     TQRMNORSUHLMNPCDEIIAD00718267    2022            1          1
2     TQRMNOQUUHLMTSCDEIJAH00719592    2022            1          1
3     TQRMNORRUHLLKNCDEIJAH00718712    2022            1          1
6     TQRMNOQXUHJMMUCDEIJAH00693031    2022            1          1
8     TQRMNOPQPHMMQLCDEIJAH00701137    2022            1          1
...                             ...     ...          ...        ...
5224  TQRMNORQYHMMMPCDEIJAH00698194    2022            1          1
5225  TQSMNOTXQHKMLQCDEIJAH00780680    2022            1          1
5226  TQRMNOQWPHKOTMCDEIJAH00780657    2022            1          1
5228  TQRMNOSRWHLOLSCDEIJAH00719039    2022            1          1
5230  TQRMNOQRXHMMPPCDEIJAH00780782    2022            1          1

      REALIZADA  REGION_x  AGLOMERADO_x  PONDERA_x  IV1  IV2  ...  niños_p
rop  \
1             1         1            32       2785    2    3  ...    0.000
000
2             1         1            33       2461    1    4  ...    0.000
000
3             1         1            33       2964    2    3  ...    0.000
000
6             1         1            33       1876    1    9  ...    0.000
000
8             1         1            33       1983    1    4  ...    0.000
000
...         ...       ...           ...        ...  ...  ...  ...
...
5224          1         1            33       2913    1    2  ...    0.428
571
5225          1         1            33       1864    1    1  ...    0.333
333
5226          1         1            33       4796    1    3  ...    0.333
333
5228          1         1            33       2783    1    3  ...    0.333
333
5230          1         1            33       2622    1    4  ...    0.250
000

      conyuge_trabaja  migrante   id          Edad  adulto_equiv  Varon  \
1                   0         0  V20  46 a 60 años          1.00      1
2                   0         0  V20  46 a 60 años          1.00      1
3                   0         0  V20  46 a 60 años          1.00      1
6                   0         0  V20  46 a 60 años          1.00      1
8                   0         0  V20  46 a 60 años          1.00      1
...               ...       ...  ...           ...           ...    ...
5224                0         0   V2        2 años          0.46      1
5225                0         0   V2        2 años          0.46      1
5226                0         0   V2        2 años          0.46      1
5228                0         0   V2        2 años          0.46      1
5230                0         0   V2        2 años          0.46      1

      ad_equiv_hogar  ingreso_necesario  pobre
1               2.55         69353.9820    0.0
2               3.55         96551.6220    0.0
3               3.59         97639.5276    1.0
6               3.28         89208.2592    0.0
8               3.55         96551.6220    1.0
...              ...                ...    ...
5224            4.86        132180.5304    1.0
5225            2.24         60922.7136    1.0
```

```
5226            2.23         60650.7372    0.0
5228            2.25         61194.6900    1.0
5230            3.26         88664.3064    0.0

[3689 rows x 143 columns]
```

```
El pocentaje de hogares bajo la linea de pobreza es 0.34064281543033087
```

La tasa de pobreza que obtenemos se asemeja a la del INDEC. Para el primer semestre de 2022 para AMBA obtenemos una tasa de 34.06%, mientras que en el INDEC la tasa de pobreza para esta misma región es de 28.2%.

```
La suma del PONDIH de los hogares pobres es 1262824
La suma del PONDIH de todos los hogares 4978383
La tasa de hogares bajo la linea de pobreza de AMBA es 25.366148004281712 %
```

La tasa de pobreza que obtenemos se asemeja a la del INDEC. Para el primer semestre de 2022 para AMBA obtenemos una tasa de 25.36%, mientras que en el INDEC la tasa de pobreza para esta misma región es de 28.2%.

Parte II

1)

2)

3)

4)

Parte III

1)

2)

Out[39]:

	Modelo	Configuración	Error Cuadratico Medio	Falsos positivos	Falsos negativos	Verdaderos positivos	Verdaderos negativos	Val AU
0	Regresión Logística	{'penalty': 'l1', 'C': 1, 'n_components': 1, '...	0.206765	80	132	220	675	0.75
1	Analisis discriminante lineal	{'penalty': 'l1', 'C': 1, 'n_components': 1, '...	0.207587	80	137	215	675	0.75
2	Vecinos cercanos	{'penalty': 'l1', 'C': 1, 'n_components': 1, '...	0.19598	34	170	182	721	0.73
3	Arbol	84	0.19598	64	60	292	691	0.87
4	Bagging	Predeterminada	0.19598	67	158	194	688	0.73
5	Gradient Boosting	Predeterminada	0.19598	26	105	247	729	0.83
6	Gradient Boosting	Predeterminada	0.19598	26	105	247	729	0.83

3)

A partir de la medida de Accuracy score y en la mayoría de las medidas de precisión, podemos determinar que el mejor modelo es un Arbol de regresión con una profundidad de 79 nodos. Asimismo, en la problemática en cuestión consideramos necesario poner atención a los "Falsos negativos", es decir, aquellas personas que son pobres y se predice como que no lo son dado que esto lo consideramos el error más grave. En este sentido, también observamos que el modelo elegido tiene la menor cantidad de "Falsos negativos".

4)

Con respecto a las predicciones del TP3 observamos una mejoría en la capacidad predictiva. En el caso anterior, elegimos regresión logística con un AUC = 0.7608, Accuracy score = 0.81, ECM = 0.198057 y Falsos Negativos = 132. En el caso actual podemos hacer dos comparaciones. Por un lado, si comparamos con el modelo de regresión logística vemos que las predicciones empeoraron levemente porque aumentó el ECM y las medidas de precisión disminuyeron. Por otro lado, si lo comparamos con el modelo óptimo actual vemos que mejoraron las predicciones: AUC = 0.8870, Accuracy Score = 0.90, ECM = 0.19598 y Falsos negativos = 53.

5)

```
Hogares pobres predichos: 1516.0
Hogares que no reportaron ingreso: 1542
Proporción de hogares pobres: 98.31387808041504 %
```

Vemos que la proporción de hogares pobres obtenida de nuestra predicción con árbol de regresión es del 96.82%. Si bien consideramos que mejora en relación a nuestro trabajo anterior, nos resulta extraño que sea un valor tan alto.