Universidad de
SanAndrés

Professor: Noelia Romero

Trabajo Práctico 2

Pacheco - Paganini - Starobinski

Fecha: 23/10/2023

Parte I: Analizando la base

Inciso 1

Segun el INDEC (2016), se clasifica como hogares pobres a aquellos que no superan el umbral establecido mediante la "Línea de Pobreza"(LP). Esta característica que se le atribuye se extiende posteriormente a los miembros de dicho hogar. Para definir dicho umbral, se toma en cuenta el valor de la Canasta Básica Total (CBT) y las necesidades energéticas según edad y sexo de las personas. La CBT está compuesta por el valor de la Canasta Básica Alimentaria, la cual se define como una canasta de alimentos capaz de satisfacer un umbral mínimo de necesidades energéticas y proteicas, y el valor de bienes y servicios básicos no alimentarios como pueden ser vestimenta, transporte, educación, salud, etc. Así, la línea de pobreza se construye por hogar considerando su tamaño y composición y se la compara con el Ingreso Total Familiar. Aquellos hogares que tengan un ingreso menor que el umbral serán considerados pobres; es decir, son clasificados como pobres aquellos hogares que no tienen un ingreso familiar capaz de cubrir las necesidades básicas de sus integrantes que les permita vivir dignamente en sociedad y desarrollarse personalmente.

Inciso 2

Literales a) y b)

Se trabajaron estos literales en el *Notebook* entregado.

Literal c)

En la figura 1, se puede observar que hay ligeramente más mujeres que varones en la muestra. Siendo más preciso, hay 3181 varones y 3589 mujeres.

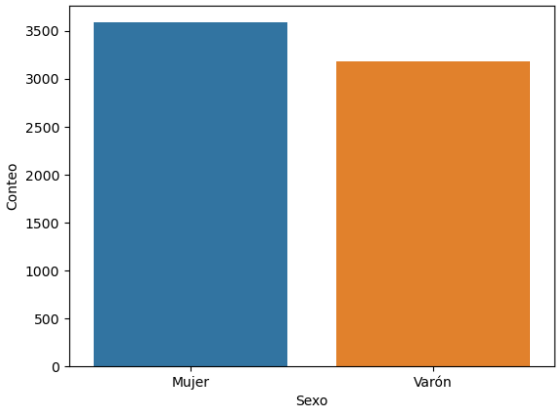

Figura 1: Composición de la muestra por sexo

Literal d)

La figura 2 muestra dos matrices de correlación de las variables CH04, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC y IPCF. Estas corresponden al sexo, estado civil, cobertura médica, nivel educativo, condición de actividad, categoría de inactividad y el ingreso per cápita familiar, respectivamente. La diferencia entre ambas matrices es que, dado que algunas variables de las variables mencionadas son categóricas, se construyeron otras a partir de estas para realizar un mejor análisis. Para ello, se generaron variables dummies para cada categoría y se construyeron otras variables a partir de ellas para que la matriz resultante no sea muy extensa. De esta manera, se obtiene una matriz que contiene nuevas variables que indican si la observación tiene o no pareja, tiene o no cobertura médica, si está ocupado, desocupado, inactivo y si se considera que lo más probable es que sea inactivo permanente dado que es jubilado o discapacitado.

Dicho esto, las correlaciones que se consideran más importantes a analizar son las que existen, por ejemplo, entre estar ocupado con el nivel educativo y estar en pareja. La matriz indica que existe una correlación positiva entre estas tres variables, lo cual puede ser una señal de que aquellos que alcanzan un mayor nivel educativo tienen mayor posibilidad de encontrar trabajo. Asimismo, parece haber cierta correlación positiva entre el ingreso per cápita familiar con la cobertura médica y el nivel educativo, lo cual puede indicar que aquellos con mayor capacidad adquisitiva tienen mayor posibilidad de contar con cobertura médica y alcanzar mayores niveles de educación.

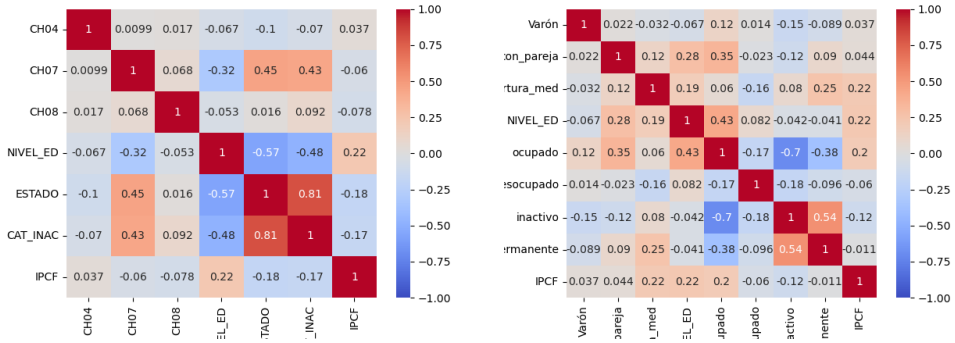

Figura 2: Matriz de correlación

Literal e)

Hay en la muestra 286 desocupados y 2826 inactivos. Asimismo, la media de ingreso per cápita familiar de los ocupados es 76055.79, la de los desocupados es 25536.02 y la de los inactivos es 40089.14

Literal f)

Se trabajó este literal en el *Notebook* entregado.

Inciso 3

2597 personas no respondieron cuál es su ingreso total familiar

Inciso 4

Se trabajó esta sección en el *Notebook* entregado.

Inciso 5

Se identificó a 1555 personas pobres.

Parte II: Clasificación

Inciso 1

Para la consigna de este ejercicio referirse al código. Las variables fueron eliminadas tanto de la base *respondieron* como de *norespondieron*, para garantizar la consistencia de las dos bases al momento de predecir la variable *pobre* para la segunda.

Además de lo que estipula explícitamente la consigna, en esta parte del código se realizaron manipulaciones de las bases que se juzgaron apropiadas para el ejercicio empírico a realizar en los siguientes incisos. Se eliminaron las variables que permanecen constantes en una muestra del Gran Buenos Aires en el primer trimestre de 2023, lo que justificó la eliminación de ambas bases de las variables identificadoras de año, trimestre, región, y la identificadora de territorios con más de medio millón de habitantes. Tampoco parece apropiado incluir como variables explicativas aquellas que reflejan formalidades del proceso de encuestamiento, como *CODUSU* y *NRO_HOGAR* que identifican viviendas y hogares, respectivamente, así como *COMPONENTE*, que identifica individuos dentro de cada hogar. Asimismo, se eliminaron variables como *Edad*, importada de otra base en la PARTE I, CH05 (una variable *string* que indica la fecha de nacimiento) y *Sexo*, creada de manera instrumental durante un *merge* en la Parte I, por considerarse redundantes al existir variables de sexo y edad en años en la base original.

Se añadió además a ambas una variable llamada *edad_2* que se corresponde al cuadrado de la edad (*CH06*), dado que es habitualmente útil en modelos para captar relaciones no lineales con la edad. La variable *IMPUTA* refiere a individuos para los que se tuvo que imputar algún dato. Considerándose que esto puede guardar relación con dimesiones socialmente relevantes (invalidez, estatus en el hogar), para esto se reemplazaron sus valores faltantes por ceros (la variable solamente tiene unos en la base original) par incluirla.

Otras manipulaciones a las bases fueron: eliminar (de ambas bases) las variables cuyos datos faltantes superen al 50 % de las observaciones en *respondieron* (con eso se eliminaron todos los datos faltantes de la muestra). Las variables categóricas no deben ser tratadas como numéricas en una estimación, por lo que cada una de ellas fue reemplazada por un conjunto *dummies* para cada categoría (exceptuando una). Para asegurar una consistencia entre las dos bases (algunas variables tomaban valores en *respondieron* que nunca tomaban en *norespondieron* y viceversa), nuevamente con el objetivo de permitir realizar la predicción del ejercicio 5, se procedió a incluir *norespondieron* como variables con valores siempre nulos a aquellas que pertenecían a *respondieron* pero no a su contraparte, y a eliminar de *norespondieron* las variables que pertenecían solamente a esta última.

Inciso 2

Para este inciso referirse al código. Para poder utilizar las ponderaciones de muestreo en el ejercicio clasificación (al menos para el ajuste con un modelo Logit, donde esto es posible), se mantuvo anteriormente la variable *PONDERA*, y en este inciso se separó tanto de la muestra de variables explicativas como de la variable dependiente, dividiéndose también para esta variable las observaciones en entrenamiento y testeo de la misma forma que lo hecho para el resto de las variables de *respondieron*.

Inciso 3

Las imágenes de matrices de confusión y curvas ROC se muestran al final del documento.

El cálculo de AUC y Accuracy se muestra a continuación:

AUC Logit: 0.749
Accuracy Logit: 0.772

AUC ADL: 0.729
Accuracy ADL: 0.756

AUC KNN: 0.659
Accuracy KNN: 0.694

Inciso 4

Esta pregunta puede abordarse de diferentes formas, pero todas apuntan a que el mejor desempeño predictivo fue obtenido con el modelo logit.

Analizando la Matriz de Confusión de cada modelo notamos que en la diagonal de falsos positivos y falsos negativos quién obtuvo un número menor es el modelo logit con tal solo 162 falsos positivos y 123 falsos negativos.

Esto se ve reflejado en el cálculo de Accuracy donde vemos que el modelo logit supera tanto a ADL como a KNN obteniendo 0.772 por encima del 0.756 y 0.694 de los otros dos modelos respectivamente.

Por otro lado podemos ver como la curva ROC de Logit está más cerca del ideal (donde la

tasa de verdaderos positivos es 1 para todo valor de la tasa de falsos positivos) en comparación a los otros dos modelos. Si bien es cierto que es por muy poco en comparación con ADL, la diferencia con KNN es marcada y visible.

Si se nos dificulta ver gráficamente porque las curvas ROC de Logit y ADL son bastante similares, lo que puede hacerse es calcular el área por debajo de la curva ROC. Lo que se busca es que esta área esté lo más cerca de 1 posible. Los resultados presentaos en el inciso anterior permiten ver que el AUC Logit de 0.749 es mayor al de ADL que es 0.729 (y ambos son sustancialmente mayores al de KNN, lo que se corresponde con lo que puede apreciarse con más facilidad en el gráfico).

Inciso 5

A partir del ajuste a un modelo *Logit* realizado en el Inciso 3, se clasificó a las personas de la base *norespondieron* como pobres o no pobres siguientedo la regla de Bayes. Se predijo que un total de 987 personas, que representan un 38,0 % de la muestra, son pobres. La proporción es mayor a la observada para la base *respondieron* (1555 entre un total de 4173, o 37,3 %), pero la diferencia es muy ligera.

Inciso 6

Un modelo que incluye todas las variables disponibles es bueno para eliminar la mayor parte posible del sesgo de los estimadores, pero puede inducir demasiada varianza, tornándolo menos útil que otros modelos algo más sesgados para el propósito de predecir qué individuos son pobres en base a sus características observables. Se ha visto que el problema, relacionado, de *overfitting* puede llevar a modelos que predicen muy bien adentro de la muestra pero mal fuera de ella.

Aunque un ejercicio más exhaustivo permitiría elegir variables en función del poder predictivo del modelo, en principio tiene sentido conservar las variables que la evidencia empírica o consideraciones teóricas justifican asumir como más correlacionadas con la pobreza. Esto implica excluir tanto variables irrelevantes como aquellas potencialmente relevantes cuyo poder explicativo no se espera que sea particularmente significativo. En este sentido, para este inciso se conservaron las *dummies* contruidas a partir de las siguientes variables categóricas: *CH04* (sexo), *CH06* (edad en años), *CH07* (estado civil), *CH08* (tipo de cobertura médica), *CH09* (si sabe leer y escribir), *NIVEL_ED* (máximo nivel educativo alcanzado), *ESTADO* (si está ocupado, desocupado, inactivo o es menor de 10 años), *AGLOMERADO* (aglomerado urbano; si vive en CABA o a GBA), *CAT_OCUP* (categoría ocupacional), *CAT_INAC* (categoría de

inactividad) y *PP02I* (si trabajó en algún momento en los últimos 12 meses). También se añadieron dos *dummies* para categorías puntuales, sin incluir todas las *dummies* generadas a partir de la misma variable categórica original: *CH11_2* (si asistió a un establecimiento educativo privado) y *CH15_4* (si nació en un país limítrofe). Se observó que la inclusión de *edad_2* reducía el desempeño del modelo, por lo que no fue incluida.

Con estas variables se ajustó nuevamente un modelo *logit* a los datos de entrenamiento de *respondieron* y se evaluó a partir de su desempeño para los datos de prueba. Las Figuras 9 y 10 permiten ver la matriz de confusión y la curva ROC para la estimación. Los valores de AUC y *Accuracy* son de 0,743 y 0,768, respectivamente.

Los resultados son sumamente similares a los obtenidos en el inciso 3, lo que implica una relevancia moderada de los cambios realizados. Los valores de AUC y *Accuracy* son algo menores, de hecho, algo que llama la atención, puesto que implica que la reducción en la varianza no compensó la pérdida de poder explicativo al dejar variables *a priori* poco importantes afuera del modelo. Sin embargo, debe notarse que la página de *scikit-learn* explicita que *LogisticRegression()* emplea métodos de regularización por defecto, de manera que es posible que este sorprendente resultado se deba a que el ajuste del inciso 3 ya corregía de manera adecuada la excesiva complejidad del modelo planteado.

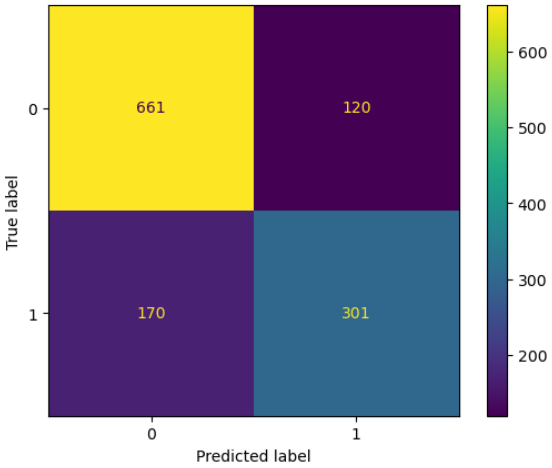

Figura 9: Matriz de Confusión Logit

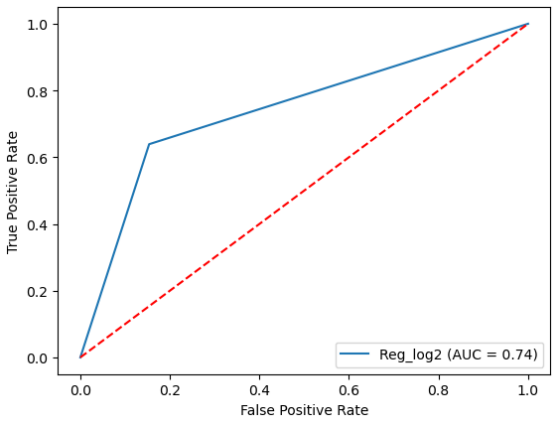

Figura 10: Curva ROC Logit

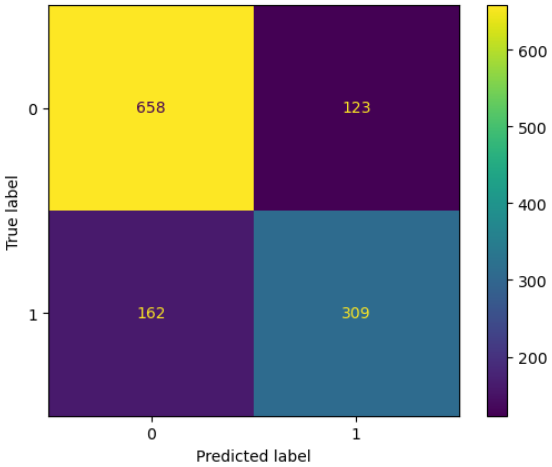

Figura 3: Matriz de Confusión Logit

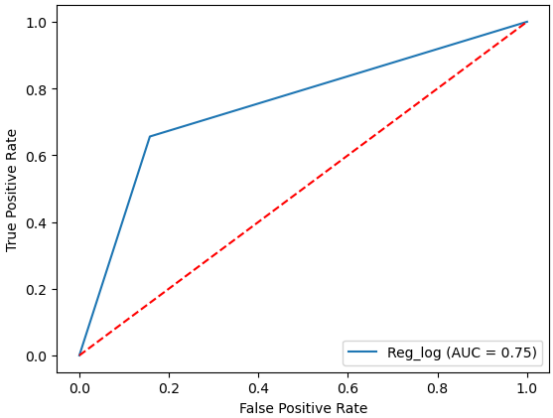

Figura 4: Curva ROC Logit

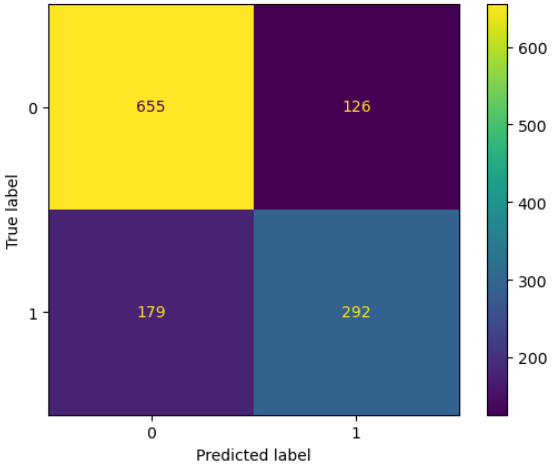

Figura 5: Matriz de Confusión ADL

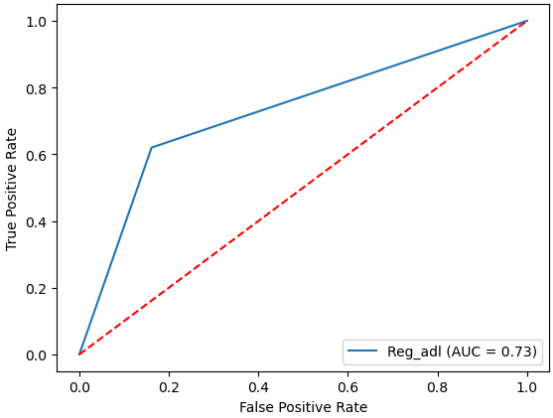

Figura 6: Curva ROC ADL

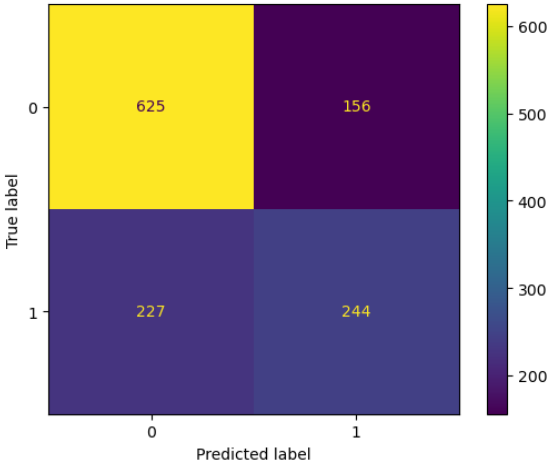

Figura 7: Matriz de Confusión KNN

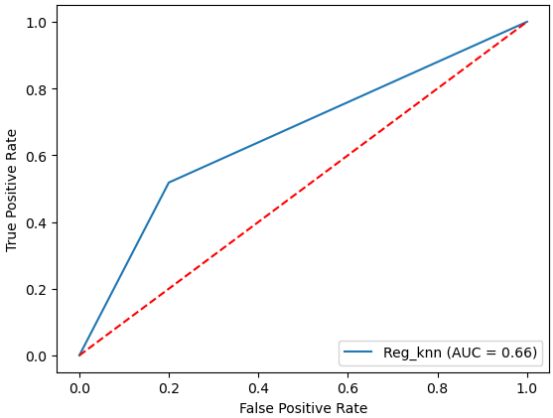

Figura 8: Curva ROC KNN