Perfecto el trabajo!

NOTA: 10

Machine Learning para Economistas

Trabajo Práctico N°4

Clasificación y regularización de desocupación usando la EPH

Manuel Díaz de la Fuente, Diego Fernández Meijide y Sofía Kastika

Profesor: Walter Sosa Escudero
Asistente: Tomás Pacheco

Diciembre 2024

Parte I: Análisis de la base de hogares y tipo de ocupación

El objetivo de esta sección es el de realizar una limpieza de datos y un análisis descriptivo utilizando la base de Hogares de la Encuesta Permanente de Hogares (EPH) correspondiente a Bahía Blanca para los primeros trimestres de los años 2004 y 2024.

Ejercicio 1

Si bien las características individuales como el sexo, la edad o el nivel de educación son relevantes para la predicción de la desocupación, no son las únicas que pueden influir en si una persona se encuentra desocupada. Sería relevante, también, agregar características a nivel de los hogares, ya que estas pueden ayudarnos a entender mejor la situación de un individuo particular y por ende perfeccionar las predicciones de desocupación. Para ello, complementamos el ejercicio realizado en el TP 3, en donde sólo se utilizaron variables de la base de personas de la EPH, con variables de la base de hogares de la misma.

Consideramos que, para la predicción de la desocupación, es importante incorporar variables que pueden clasificarse en cuatro grupos: fuentes alternativas del ingreso, condiciones físicas del hogar, espacio de trabajo y estigma social. Con respecto a las **fuentes alternativas de ingreso** de los hogares, en la EPH se hace una serie de preguntas sobre el origen de las fuentes de ingreso mediante las cuales las personas del hogar vivieron en los últimos 3 meses. Consideramos que el hecho de que los hogares hayan vivido a partir de fuentes de ingreso que no sean de lo que ganan en el trabajo puede reflejar estrategias de supervivencia de los hogares ante la falta de empleo. Dos ejemplos de variables que incorporamos son si los hogares vivieron de pedir préstamos a familiares / amigos y si vivieron de subsidios del gobierno o iglesias[1]. Cabe aclarar que dejamos de lado variables que hacen referencia a fuentes de ingreso que nos parecen triviales para la predicción de la desocupación. Estas son si el hogar vive de lo que ganan en el trabajo, de indemnización por despido y de seguro de desempleo.

El segundo grupo de variables a utilizar tiene que ver con **características físicas de la vivienda**. La ausencia de condiciones dignas en la vivienda está asociada con cuestiones de salud e higiene, lo cual podría limitar la inserción en el mercado laboral. Consideramos que personas que tienen poca salud es más probable que estén desocupadas ya que tienen imposibilitado realizar ciertos trabajos y porque al momento de la contratación se suele tener en cuenta el *status* de salud de las personas. Es por ello que incluímos como variables condiciones físicas del hogar que están estrechamente ligadas a cuestiones sanitarias. Estas variables abarcan el material de los pisos interiores y el techo, la proveniencia del agua y el sistema de desagüe del baño[2].

En tercer lugar, consideramos que si un hogar dispone de un **ambiente destinado exclusivamente al trabajo** como un consultorio, estudio, taller, negocio, etc. (II3), esto indica una menor probabilidad de que las personas en ese hogar se encuentren desempleadas, ya que tienen un espacio adecuado para desarrollar actividades laborales. Por último, consideramos relevante la variable que indica si **una persona vive en una villa de emergencia** (IV12_3). Dado el estigma social existente asociado a estas zonas, el hecho de vivir en una villa de emergencia genera una menor probabilidad de ser contratado.

[1] Las variables que agregamos de este grupo son si las familias en los últimos 3 meses vivieron de: alguna jubilación o pensión (V2), jubilación o pensión cobrada el mes anterior (V21), retroactivo de alguna jubilación que cobró el mes anterior (V22), subsidio o ayuda (en dinero) del gobierno o iglesias (V5), con mercaderías, ropa, alimentos (V6), con mercaderías, ropa, alimentos de familiares u otras personas que no viven en el hogar (V7), algún alquiler (V8), ganancias de algún negocio en el que no trabajan (V9), intereses o rentas por plazos fijos / inversiones (V10), beca de estudio (V11), cuitas de alimentos o ayuda en dinero de personas que no viven en el hogar (V12), gasto de ahorros (V13), préstamos de familiares / amigos (V14), préstamos de bancos (V15). Si han tenido que vender pertenencias (V17), si tuvieron otros ingresos en efectivo como limosnas (V18), si menores de 10 años ayudan con trabajo (V19_A), si menores de 10 años ayudan pidiendo plata (V19_B).

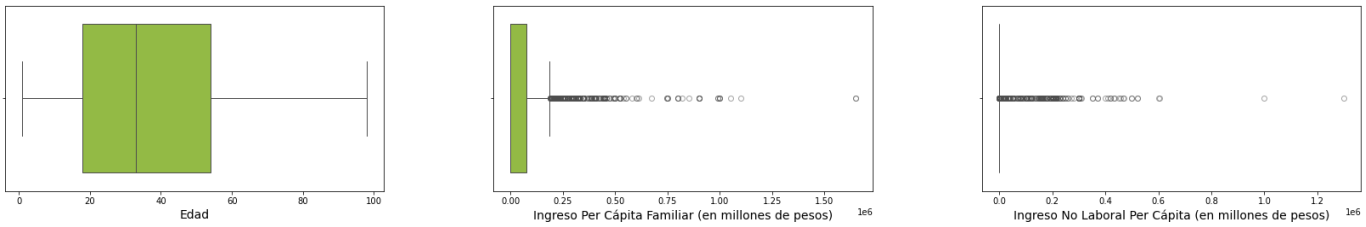
[2] Específicamente, las variables incluídas de este grupo son: material de los pisos interiores (IV3), material de la cubierta exterior del techo (V4), Si el techo tiene cielorraso (IV5), Si tiene agua adentro de la vivienda (IV6), De dónde proviene el agua (IV7), Tenencia de baño (IV8), Si tiene baño adentro de la vivienda (IV9), si el baño tiene inodoro con con botón / mochila / cadena y arrastre de agua (IV10) y el sistema de desagüe del baño (IV11).

Ejercicio 3

El objetivo de este ejercicio es el de realizar una limpieza de la base. Nuestra base contiene 2164 observaciones. Encontramos únicamente 74 missing values en el ingreso no laboral, por lo que eliminamos esas observaciones. A su vez, encontramos 22 edades negativas, por lo que eliminamos también esas observaciones. No hay valores negativos en ingreso per cápita e ingreso per cápita no laboral.

Con respecto a los outliers, observamos los *box plots* para nuestras tres variables continuas: edad, ingreso per cápita familiar e ingreso no laboral per cápita en la Figura 1. Como puede observarse, la edad no presenta outliers. Sin embargo, tanto el ingreso per cápita familiar como el ingreso no laboral per cápita sí presentan outliers. No obstante, nuestra decisión en estos casos es no eliminar outliers, ya que buscamos que nuestros modelos presentan la capacidad de predecir el estado de ocupación de todos los individuos de nuestro aglomerado urbano (nuestra población). La presencia de outliers en estas variables en nuestra muestra no es una característica propia de nuestra muestra sino que es propio de la distribución del ingreso en Argentina y, en particular, en nuestra región, Bahía Blanca. Esto es así, porque la muestra de indviduos de la EPH es representativa de la población de los aglomerados urbanos de la cual es tomada. Por lo tanto, para que nuestros modelos puedan predecir el estado de ocupación de todos los individuos de Bahía Blanca debemos considerar a todos los individuos independientemente de su nivel de ingresos (laborales o no laborales).

Figura 1: Box Plots para Edad, Ingreso Per Cápita Familiar e Ingreso No Laboral Per Cápita



a) Edad **b) IPCF** **c) Ingreso no laboral**

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares (EPH).

Finalmente, para las variables categóricas con más de dos categorías, generaremos una variable dummy por cada categoría. Por ejemplo, la variable Estado Civil tiene múltiples categorías. Para tratarla, crearemos una dummy para cada estado: una para personas casadas, otra para personas unidas, otra para para viudas, otra para personas divorciadas y finalmente una para personas solteras

Ejercicio 4

Para predecir individuos desocupados, pueden construirse nuevas variables en base a los datos ya existentes en la EPH. La primera variable que construímos es la de **hacinamiento**. Esta, según define el INDEC, resulta de realizar el ratio de personas que viven en el hogar sobre la cantidad de habitaciones en el hogar. Luego, si ese ratio es mayor a tres, se considera que hay hacinamiento crítico. Es por ello que primero construímos el ratio y luego definimos una dummy que indica si ese ratio es mayor o menor a 3.

En segundo lugar, construímos una medida de **cantidad de inactivos en el hogar**. Consideramos que a una mayor cantidad de personas inactivas en el hogar (como personas mayores o niños), va a caer el salario de reserva de la persona en edad para trabajar. Esto es, cae el salario mínimo que está dispuesta a aceptar para comenzar a trabajar ya que tiene que hacerse cargo de las personas inactivas dentro del hogar. De esta manera, mayor cantidad de personas inactivas en el hogar está correlacionado negativamente con la probabilidad de estar desempleado.

Por último, construimos una variable que representa el **ingreso no laboral familiar per cápita**. Esta es el resultado de sumar todos los ingresos no laborales per cápita por hogar y dividir a esa sumatoria por los miembros del hogar. Consideramos que, si el ingreso no laboral per cápita es mayor, el salario de reserva del individuo también aumenta, lo que puede llevarla a mantener su estado de desempleo, ya que no está en tanta situación de urgencia ya que puede depender de estos ingresos no laborales para el sustento familiar. Por lo tanto, un mayor ingreso no laboral per cápita se correlaciona positivamente con la probabilidad de estar desempleado.

Ejercicio 5

En esta sección se reportan estadísticas descriptivas de variables que consideramos relevantes para la predicción de la desocupación para los primeros trimestres del 2004 y 2024. Como mencionamos en el Ejercicio 1, consideramos que las variables relevantes para la predicción de la desocupación podrían agruparse en cuatro categorías: fuentes alternativas de ingreso, características físicas de la vivienda, ambiente destinado exclusivamente al trabajo y si la persona vive en una villa de emergencia. Es por ello que elegimos al menos una variable por categoría para presentar su estadística descriptiva.

Como puede observarse en el El Cuadro 1, si se compara el 2004 y el 2024, hubo un aumento de los hogares que al momento de la encuesta reportan haber vivido en los últimos 3 meses tanto de subsidios del gobierno o iglesias como de ahorros. Mientras que en el 2004 únicamente el 0.61 % de los hogares de Bahía Blanca reporta haber vivido de subsidios, en 2024 el 14.83 % reporta haberlo hecho. Por su parte, mientras que en 2004 un 6.29 % de los hogares reporta haber tenido que gastar ahorros durante los últimos 3 meses, en 2024 el porcentaje de hogares aumenta a un 17.32 %. Por otro lado, notamos que la proporción de hogares que poseen al menos un ambiente que usan exclusivamente para el trabajo (como para consultorio, estudio, taller o negocio) se mantuvo constante, con aproximadamente 5 % de los hogares que lo poseen. Por último, mientras que en 2004 el 1.23 % de los hogares de Bahía Blanca reporta vivir en una villa de emergencia, en 2024 no hay hogares que lo reporten.

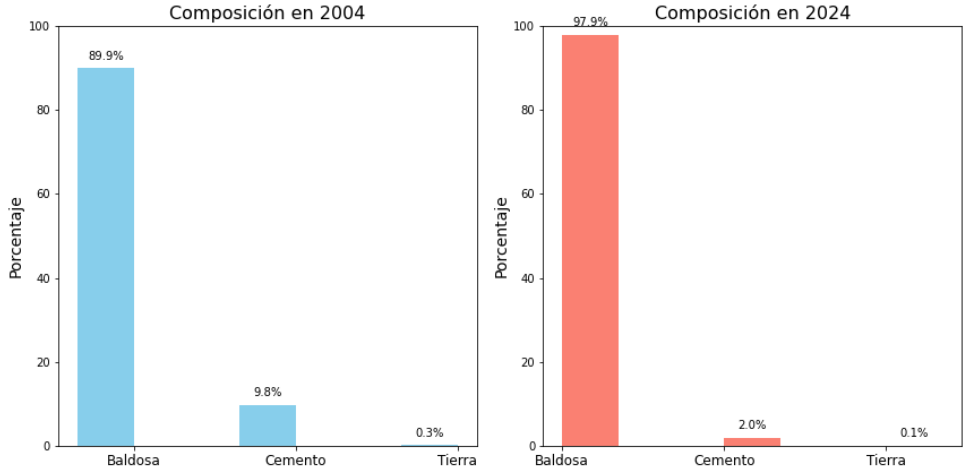
Cuadro 1: Estadísticas descriptivas 2004 y 2024

Año	Subsidios	Ahorros	Lugar de trabajo	Villa de Emergencia
2004	0.61	6.29	4.12	1.23
2024	14.83	17.32	5.52	0.00

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares (EPH).

Por último, la Figura 2 muestra las proporciones de hogares que tienen pisos con distintos materiales, lo cual es una de las variables que representa las condiciones físicas de la vivienda. Como puede observarse, hubo una disminución de hogares que poseen un piso de tierra o ladrillo suelto, pasando del 0.3 % en 2004 de los hogares a un 0.1 % en 2024. Por su parte, también hubo una reducción de la cantidad de hogares que viven con pisos de cemento o ladrillo fijo, pasando de representar un 9.8 % de los hogares en 2004 a un 2 % en 2024. Notamos que hubo una transición hacia pisos de baldosa, madera, mosaico, cerámica o alfombra. Mientras que en 2004 el 89.9 % tenía pisos de estos materiales, en 2024 casi la totalidad de los hogares pasaron a tener pisos de estos materiales, representando el 97.9 % de los hogares.

Figura 2: Composición Material de Pisos por año



Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares (EPH). La etiqueta 'Baldosa' hace referencia a hogares en donde el material del piso es de mosaico, baldosa, madera, cerámica o alfombra. Por otro lado, la etiqueta 'Cemento' hace referencia a hogares cuyo piso está hecho de cemento o ladrillo fijo. Por último, la etiqueta 'Tierra' hace referencia a hogares cuyo piso está hecho de tierra o ladrillo suelto.

Ejercicio 6

A diferencia del Censo Nacional de Población, Hogares y Viviendas en donde se recolectan los datos de toda la población de Argentina, la Encuesta Permanente de Hogares (EPH) trabaja únicamente con una muestra de hogares. Sin embargo, se utiliza un ponderador llamado PONDERA que, según el INDEC, es un factor de ponderación que ajusta los datos de la muestra para que sean representativos del área geográfica de estudio. En nuestro caso, si utilizamos el ponderador, nos permite poder expandir la muestra para que sea representativa de toda la población de Bahía Blanca - Cerri.

Por ende, para calcular la tasa de desocupación lo que se realizó fue sumar la cantidad de jefes de hogares desocupados y dividirlo por la totalidad de jefes de hogares, utilizando el PONDERA para ampliar la muestra. Se obtuvo que la tasa de desocupación de Bahía Blanca - Cerri es de 3.03 %. No obstante, nuestro cálculo nos da diferente a la tasa reportada en el informe del INDEC, la cual es de un 7.5 %.

Parte II

El objetivo de esta sección es el de predecir si una persona está desocupada o no utilizando tanto la parte individual como la de hogares de la EPH. Además, realizaremos ejercicios de regularización y validación cruzada.

Ejercicio 1

Dividimos a la muestra en entrenamiento y prueba para los respectivos años. El 30 % de la muestra fue asignada al grupo de prueba aleatoriamente utilizando la semilla 101, el resto de la muestra fue asignada al grupo de entrenamiento. Los predictores de interés fueron separados en una matriz X y la variable a predecir -desocupación- fue separada en otro vector. El procedimiento puede observarse en el código adjunto en el repositorio.

Ejercicio 2

Luego de separar los datos en un conjunto de entrenamiento y en un conjunto de prueba, se define una grilla de valores de λ. Estos valores son seleccionados buscando que representen una escala detallada de pérdida por complejidad de los modelos -cantidad de variables dentro del modelo-. De esta forma, la grilla presenta valores de muy pequeños, con muy bajas penalidades por complejidad, a muy altos, con muy altas penalidad por complejidad. Para cada uno de estos valores, se realiza *cross validation* con la muestra de entrenamiento. De esta manera, podemos computar el error cuadrático medio del modelo que predice mejor fuera de la muestra para cada valor posible de λ. Por último se elige el valor de λ que minimiza el error cuadrático medio.

Ejercicio 3

En el contexto de *cross validation* (*k-fold*), utilizar un valor de k pequeño es menos costoso computacionalmente, ya que implica realizar menos estimaciones del modelo. Sin embargo, esta elección puede afectar negativamente la capacidad predictiva del modelo fuera de la muestra de entrenamiento, ya que, para $k \ll n$, se emplean pocas observaciones para entrenar el modelo. Por ejemplo, si $n = 100$ y $k = 2$, el modelo se estima solo dos veces, con muestras de entrenamiento de tamaño $n_e = 50$.

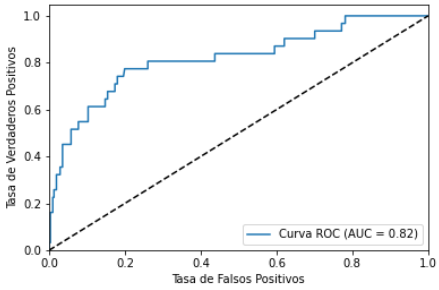
En contraste, a medida que k aumenta, la complejidad computacional también incrementa debido a que el modelo se estima más veces. Adicionalmente, surge un mayor riesgo de *overfitting*, ya que las muestras de prueba se vuelven más chicas, dificultando una evaluación robusta de la capacidad predictiva del modelo. En el extremo, cuando $n = k$ (*leave-one-out*), el modelo se estima n veces, entrenándose con $n_e = n-1$ observaciones en cada iteración y evaluándose con la única observación restante. Por ejemplo, si $n = k = 100$, se estima el modelo 100 veces, cada muestra de entrenamiento tendrá 99 observaciones y el conjunto de prueba tendrá una única observación.

Ejercicio 4

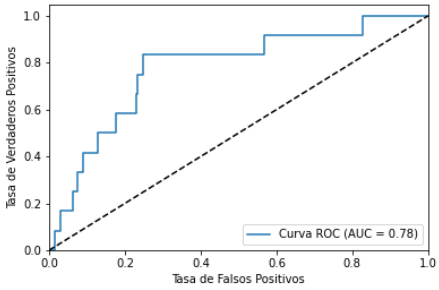
En el trabajo anterior el modelo de regresión logística predecía que todos los indviduos se encontraban ocupados para los datos correspondientes al año 2004 y para los datos correspondientes al año 2024. El *accuracy* correspondiente al modelo estimado para el año 2004 era de 90.7 % y para el año 2024 era de 96.7 %. En lo que respecta a esta medida de performance tanto el modelo LASSO como el Ridge presentan una mejor performance para el año 2004 -92.1 % y 91.2 % de accuracy respectivamente- y una peor performance para el año 2024 -aproximadamente 94 % de accuracy en ambos modelos-. Sin embargo, estas diferencias podrían explicarse porque para este trabajo el conjunto de variables explicativas seleccionado para la estimación es distinto al conjunto seleccionado para el trabajo anterior. Resulta llamativo que, a pesar de haber seleccionado un mayor número de variables, los modelos estimados en este trabajo práctico presentan un peor rendimiento para 2024 que los modelos ajustados en el trabajo anterior.

En lo que respecta a la medida *Area Under the Curve* (AUC), las regresiones logísticas estimadas en el trabajo anterior presentan valores de 0.8 para 2004 y 0.74 para 2024. Para ambos años los modelos con penalidades por complejidad estimados presentan medidades de AUC más altas. Esto indica que los modelos regularizados presentan una mejor performance predictiva porque, en promedio, presentan mayores tasas de verdaderos positivos para cada tasa de falsos positivos. Esto también se verifica en la comparación con las curvas ROC del trabajo anterior, que se encontraban más alejadas del vértice izquiero del gráfico en el cual la tasa de verdaderos positivos es 100 % independientemente de la tasa de falsos positivos.

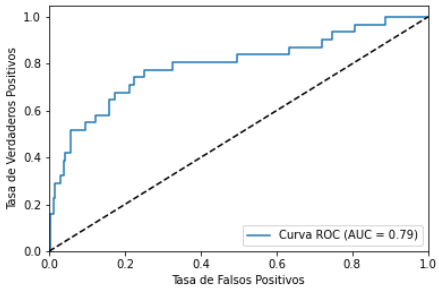
Figura 3: Curvas ROC para LASSO y Ridge en los años 2004 y 2024



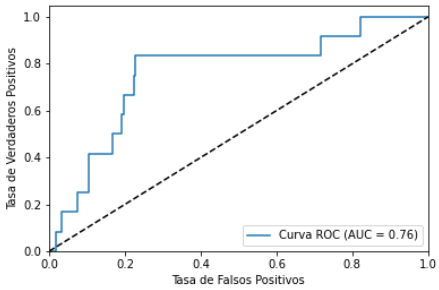
(a) Curva de ROC. LASSO 2004



(b) Curva de ROC. LASSO 2024



(c) Curva de ROC. Ridge 2004



(d) Curva de ROC. Ridge 2024

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares (EPH).

Cuadro 2: Resultados de la evaluación: matriz de confusión, AUC y Accuracy

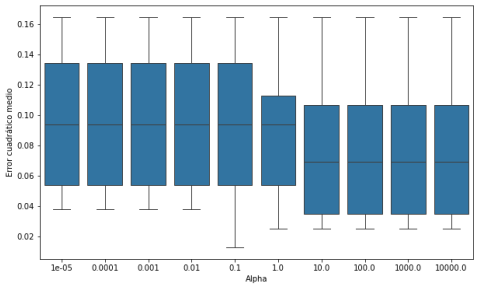
Año	Penalidad	TN	FP	FN	TP	AUC	Accuracy
2004	LASSO	310	1	26	5	0.818	0.921
2004	Ridge	307	4	26	5	0.793	0.912
2024	LASSO	262	3	12	0	0.776	0.946
2024	Ridge	261	4	12	0	0.760	0.942

Ejercicio 5

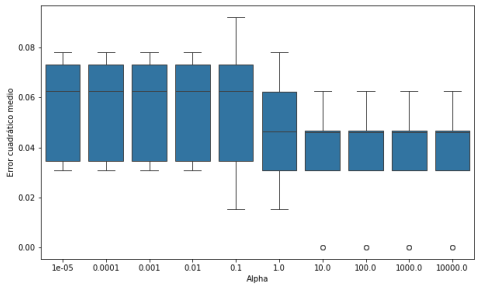
La Figura 4 muestra la distribución de Errores Cuadráticos Medios (ECM) calculados para cada modelo estimado en el proceso de CV para cada hiperparámetro λ. Los cuatro gráficos muestran que, a partir de un λ dado, todos los modelos reducen su ECM medio. En el caso de los modelos LASSO, el λ que minimiza el ECM es igual a 10 y en el caso de los modelos Ridge estimados el valor del hiperpárametro λ que minimiza el ECM para el año 2004 es 100 y para el año 2024 es 10.

La Figura 5 muestra la proporción de coeficientes estimados iguales a 0 en el modelo LASSO para los datos de 2004 y para los datos de 2024 para cada λ dado. En ambos casos el gráfico muestra una forma de ”S”porque mientras más alta es la penalidad definida para modelos más complejos, mayor es la proporción de coeficientes a los que LASSO les asigna un valor de 0.

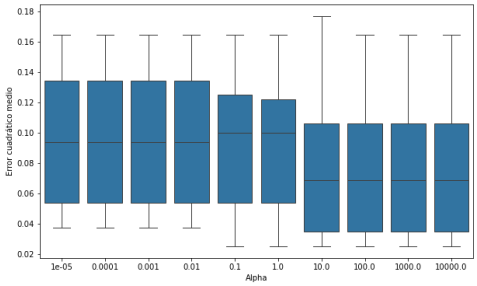
Figura 4: Error Cuadrático Medio para LASSO y Ridge en los años 2004 y 2024 para cada valor de λ



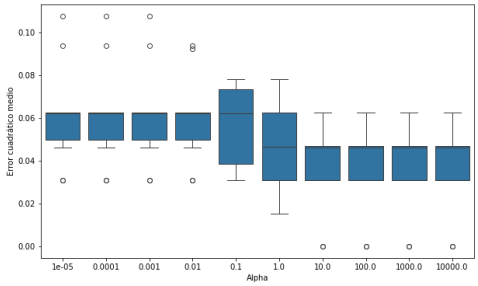
(a) Curva de ROC. LASSO 2004



(b) Curva de ROC. LASSO 2024



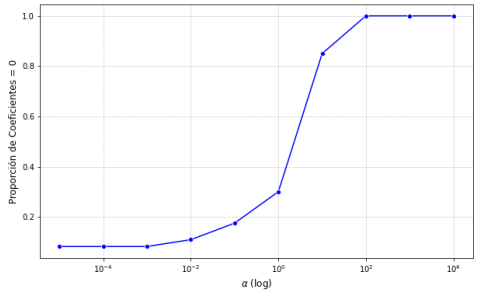
(c) Curva de ROC. Ridge 2004



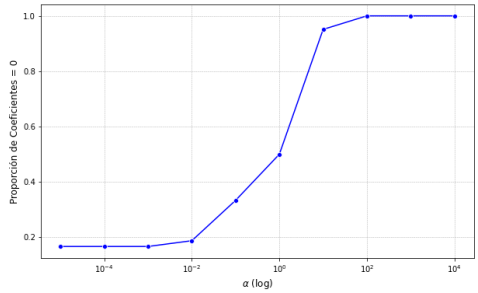
(d) Curva de ROC. Ridge 2024

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares (EPH).

Figura 5: Proporción de coeficientes nulos en 2004 y en 2024 para cada λ dado



(a) Proporción de coeficientes iguales a 0. LASSO 2004.



(b) Proporción de coeficientes iguales a 0. LASSO 2024.

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares (EPH).

Ejercicio 6

Los coeficientes estimados para el modelo LASSO en ambos años (respectivamente) indican que las variables que resultan relevantes para predecir la desocupación son el sexo del individuo; el ingreso per cápita familiar, en el caso del año 2004; la cantidad de inactivos que habitan en el hogar; el estado civil del individuo; si presenta una cobertura médica; el nivel educativo del individuo; si el hogar vive principalmente de ahorros, venta de pertenencias, u otros ingresos, en el caso del año 2004; y el material de los techos, en el caso del año 2024.

En este sentido, obtenemos evidencia a favor de nuestra hipótesis inicial de que las fuentes alternativas de ingresos del hogar se configuran como un conjunto de variables relevantes para predecir la desocupación. Esta

información sugiere que las personas que habitan en hogares en los cuales se vive de ahorros o de venta de pertenencias presentan una mayor probabilidad de estar desocupados.

En lo que refiere a las características físicas de la vivienda, en cambio, éstas no parecen presentar poder predictivo del estado ocupacional del individuo. Nuestros modelos LASSO sugieren que, ni en 2004 ni en 2024, las características físicas de los hogares correlacionaron de manera significativa con el estado ocupacional del individuo. Lo mismo sucede con el acceso a un ambiente destinado exclusivamente al trabajo, y con la variable que indica si una persona vive en una villa de emergencia. Ninguna de estas variables presenta poder predictivo del estado ocupacional del individuo. Estos resultados señalan que el estado de pobreza no determina el estado de ocupación de los individuos.

En contraste, las variables socioeconómicas tradicionales, como el sexo, el nivel educativo, la cobertura médica, el estado civil y el ingreso per cápita de la familia si presentan un poder predictivo significativo del estado ocupacional del individuo.

En el caso de las tres variables construidas en este trabajo, hacinamiento, cantidad de inactivos en el hogar, e ingreso no laboral familiar per cápita, únicamente la cantidad de inactivos en el hogar presenta poder predictivo referido al estado ocupacional del individuo. En ambos años, personas que viven en hogares con una mayor cantidad de habitantes inactivos presentan una menor probabilidad de estar desocupados. Esto se condice con nuestra hipótesis inicial en la cual aquellas personas que habitan en hogares con mayor cantidad de personas que no pueden valerse por si mismas económicamente presentan mayores incentivos a aceptar menores salarios y, por lo tanto, presentan una mayor probabilidad de estar ocupadas.

En síntesis, nuestros hallazgos indican que las características de los individuos presentan un mayor poder predictivo de su estado ocupacional que las características de los hogares en los que habitan. En particular, su sexo, su nivel económico, la cantidad de personas que no trabajan en el hogar, su nivel educativo, su estado civil y su cobertura médica parecerían ser las variables más indicadas para realizar estas predicciones.

Ejercicio 7

El Cuadro 3 muestra el ECM de las estimaciones de los modelos LASSO y Ridge para ambos años utilizando los hiperparámetros óptimos obtenidos en el ejercicio 5. Puede observarse que el rendimiento de los modelos en ambos años es muy similar, con la diferencia de que para el año 2004 el modelo Ridge presenta un mejor rendimiento. Para el año 2024 el rendimiento de ambos modelos es prácticamente idéntico. En otras palabras, el modelo LASSO presenta un ECM ligeramente superior al modelo Ridge en el año 2004, mientras que esta medida es idéntica para ambos modelos en el año 2024 .

Cuadro 3: Error Cuadrático Medio (ECM) para LASSO y Ridge en los años 2004 y 2024

Modelo	Año	ECM
LASSO	2004	0.09
LASSO	2024	0.04
Ridge	2004	0.08
Ridge	2024	0.04

Nota: Elaboración propia en base a datos de la EPH.