Trabajo Práctico 4

Machine Learning para Economistas

Marotta, Salischiker, Zapiola

27 de Diciembre, 2024

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

Inciso 1

Para identificar variables predictivas de la desocupación y perfeccionar el análisis del TP3 a través del registro de la base de hogar, podríamos considerar las siguientes variables de la Encuesta Permanente de Hogares que, a priori, podrían ayudarnos a mejorar el ejercicio:

- **Características de la Vivienda**: Las condiciones habitacionales podrían ofrecernos información relevante sobre la situación laboral del jefe/a de hogar. Por un lado, el tipo de vivienda en sí ($IV1$) podría ser una variable indicativa de precariedad habitacional y reflejar una situación transitoria, asociada frecuentemente con cierta inestabilidad laboral. Esto es porque identifica si la vivienda corresponde a situaciones como "pieza de inquilinato", "pieza en hotel/pensión.º "local no construido para habitación". Este tipo de vivienda podría estar vinculado a mayores niveles de vulnerabilidad económica, lo que incrementa la probabilidad de desocupación. Por otro lado, las características de la vivienda (por ejemplo, $IV2 - IV12, II7$) describen aspectos relacionados con la calidad y las condiciones de la vivienda, como el número de habitaciones, tipo de materiales del piso y techo, acceso a agua corriente, presencia de cielorraso, ubicación cerca de basurales o zonas inundables, y régimen de tenencia (propietario, inquilino, ocupante de hecho, entre otros). Estas variables nos brindan información sobre infraestructura básica, deficiencias en el acceso a servicios corrientes y sobre la ubicación y vulnerabilidad ambiental de la vivienda, pudiendo indicar hogares con recursos limitados y expuestos a condiciones de marginalidad, asociados con mayor riesgo de desocupación debido a restricciones socioeconómicas y dificultades en la inserción laboral. Además, cuestiones sobre el régimen de tenencia ($II7$) que revelen por ejemplo ocupación de terrenos pueden asociarse a inseguridad económica, menor estabilidad y mayor exposición a shocks del mercado laboral para los/as jefes/as de hogar.
- **Estrategias del hogar**: Las preguntas sobre estrategias realizadas por los hogares en los últimos 3 meses podrían predecir bien la situación laboral de los/as jefes/as de hogar. Por ejemplo, las variables que nos informan si los hogares han vivido de

indemnización por despido o seguro de desempleo ($V3, V4$), de subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.; con mercaderías, ropa, alimentos ya sea de gobierno, iglesias, escuelas o de familiares, vecinos u otros individuos que no pertenezcan al hogar ($V6, V7$); de cuotas de alimentos o ayuda en dinero de personas que no viven en el hogar ($V12$), de gastar lo que tenían ahorrado o pedir préstamos a familiares, amigos, bancos o financieras ($V13 - V15$); si han tenido que vender alguna de sus pertenencias ($V17$) o si menores de 10 años ayudan con algún dinero trabajando o pidiendo ($V19_{A,B}$).

Inciso 2

Ver en el *Python Script*.

Inciso 3

En la limpieza de la base de datos se tomaron decisiones para garantizar la coherencia de los datos y robustecer el análisis. En primer lugar, se identificaron y se eliminaron los valores faltantes (*missing values*). En segundo lugar, se registraron las observaciones con valores faltantes o que no tengan sentido en variables críticas. Se eliminaron *missings* y observaciones con valores negativos en las variables de ingresos individuales y totales del hogar, así como edades negativas en los individuos, ya que no son posibles en la realidad y podría comprometer la calidad de las observaciones relacionadas. A diferencia de los valores negativos que son inconsistentes con la naturaleza de los datos, se decidió no eliminar valores positivos extremos (ingresos o edades demasiado altas). Estos pueden reflejar casos reales y relevantes para el análisis, como individuos u hogares con ingresos muy altos o personas mayores aún activas en el mercado laboral. No se excluyeron para evitar sesgos en la muestra y robustecer la capacidad de predicción. Por último, en relación a las variables categóricas, no se realizaron modificaciones ya que las categorías originales no presentan inconsistencias evidentes.

Inciso 4

Se construyen tres variables que no están en la base pero que son relevantes para predecir si los individuos están desocupados o no:

I. Proporción de personas ocupadas en el hogar

II. Proporción de personas en edad laboral en el hogar

III. Proporción de niños en el hogar.

IV. Proporción de personas con secundaria completa o mas.

Ver en el *Python Script*.

Inciso 5

Como ya se mencionó en el Inciso 1, las variables sobre las estrategias del hogar para para afrontar sus necesidades pueden capturar diferentes dimensiones de la vulnerabilidad

Tabla 1: Estadísticas descriptivas sobre las estrategias del hogar para vivir.

	2004		2024	
Variable	**Cantidad**	**Prop. (%)**	**Cantidad**	**Prop. (%)**
Subsidios o Ayuda Social	63	8.06	104	13.37
Mercaderías de Gobierno/Iglesias/Escuelas	175	22.38	81	10.41
Ahorros	153	19.57	345	44.34
Préstamos de Familias/Vecinos	169	21.61	233	29.95
Venta de Pertenencias	111	14.19	119	15.30
Préstamos a Bancos	43	5.50	86	11.05

económica y social y ser relevantes para la predicción del desempleo. En la Tabla 1 se presentan las estadísticas descriptivas obtenidas para los años 2004 y 2024.

En primer lugar, la proporción de hogares que reciben subsidios incrementó de un 8,06 % en 2004 a un 13,37 % en 2024, lo que podría reflejar un aumento en la cobertura de programas sociales o una mayor dependencia de los mismos debido a condiciones económicas desfavorables. En cambio, la proporción de hogares que recibe mercaderías proporcionadas por el gobierno, iglesias o escuelas disminuyó significativamente, pasando de un 22,38 % en 2004 a un 10,41 % en 2024. Esta caída podría estar relacionada con cambios en las políticas de asistencia o con una disminución de la disponibilidad de este tipo de ayuda.

Por otro lado, la proporción de hogares que viven de gastar lo que tenían ahorrado aumentó considerablemente, de un 19,57 % en 2004 a un 44,34 % en 2024, y también los que solicitan préstamos a familiares o vecinos, creciendo de un 21,61 % a un 29,95 %. Este incremento podría reflejar un mayor estrés económico y ser indicativo de una mayor fragilidad financiera, ya que más hogares recurren a sus ahorros y a su red social más cercana como estrategia para enfrentar dificultades económicas. En cuanto al porcentaje de hogares que recurre a la venta de pertenencias, éste se mantuvo relativamente constante, con un 14,19 % en 2004 y un 15,30 % en 2024. Esto sugiere que esta estrategia sigue siendo utilizada por un segmento constante de la población como un mecanismo para afrontar crisis económicas.

Finalmente, la proporción de hogares que solicita préstamos a bancos o financieras se duplicó, pasando del 5,50 % en 2004 al 11,05 % en 2024. Este incremento podría estar asociado con un mayor acceso al sistema financiero, aunque también podría reflejar un incremento en la necesidad de recurrir a estas instituciones para cubrir necesidades básicas.

Inciso 6

En el *Python Script* se presenta el cálculo de la tasa de hogares con desocupación para el aglomerado de Gran Córdoba para el 1er trimestre de 2024. Primero, filtramos las observaciones pertenecientes al período de interés. Segundo, creamos una variable para identificar hogares con desocupación, agrupando los datos por hogar ($CODUSU$) y asignando un valor de 1 al hogar que cuente con al menos un miembro desocupado (es decir, algún miembro que cuente con un estado laboral tal que $ESTADO = 2$). En tercer lugar, seleccionamos sólo una observación por hogar para eliminar duplicados dentro de un hogar. Luego, hacemos una suma ponderada de todos los hogares desocupados (con al menos un desocupado) para expandir la muestra de la EPH al total de la población. Por último, calculamos la tasa de hogares desocupados sobre el total. Según nuestro cálculo,

la tasa de hogares con desocupación en el aglomerado de Gran Córdoba es de 9,99 %.

Al compararla con la tasa reportada por el INDEC en el informe de Mercado de trabajo (1T 2024), notamos un punto relevante. La tasa que presenta el INDEC muestra el porcentaje de individuos desocupados respecto a la población económicamente activa (PEA), mientras que nuestro cálculo es la proporción de hogares con al menos un miembro desocupado, por lo que la comparación no es del todo directa ya que miden fenómenos distintos. La tasa de desocupación muestra el porcentaje de personas que buscan activamente trabajo y no lo encuentran sobre la PEA, midiendo el desempleo en términos de individuos. Según el INDEC, el Gran Córdoba tuvo una tasa de desocupación del 7,6 %. En cambio, nuestra medida a nivel hogar muestra la distribución del desempleo entre el total de hogares en la población (9,99 %). Es decir, un hogar puede ser clasificado como 'hogar desocupado' aunque tenga otros miembros empleados. Por lo tanto, en hogares más grandes con más cantidad de miembros que están dentro de la PEA, la probabilidad de tener al menos un desempleado es mayor, lo que podría inflar la tasa de hogares con desocupación. El hecho de que nuestra tasa a nivel hogar es más alta que la tasa individual puede sugerir que el desempleo está concentrado en los hogares grandes o sobre la heterogeneidad de la composición de los hogares (si hay más miembros en cada hogar, esos hogares tienden a contribuir más a la tasa), elevando la tasa de hogares con desocupación porque es más probable que al menos un miembro del hogar se encuentre desocupado. Por esto es que, aunque la desocupación individual sea baja, ésta afecta a un mayor número de hogares, incrementando la proporción. De todos modos, la medida por hogares obtenida es similar en magnitud (+2,3 %) a la tasa individual de desempleo reportada por el INDEC, pero nos permite sacar algunas conclusiones sobre la distribución del desempleo y la composición de los hogares.

Parte II: Clasificación y Regularización

Inciso 1

Ver en el *Python Script*.

Inciso 2

Para ajustar un modelo hay distintos métodos para poder mejorar la precisión de la predicción y la interpretabilidad del modelo. El primer enfoque es identificar un grupo de variables predicadores p que creemos que están relacionados a la variable que queremos predecir, y de ahí ajustar el modelo usando MCO. El segundo enfoque involucra ajustar el modelo usando todos los p, pero cada uno de los estimadores se regularizan a 0 relativo a los estimadores por MCO. Esta regularización (*Shrinkage* o *Regularization*) tiene un efecto en la varianza, dependiendo que tipo de regularización se implementa (Lasso o Ridge). Y el tercer enfoque implica proyectar los p predicadores en un subespacio con dimensión M, donde $M < p$.

Dentro del método de regularización, elegir un λ óptimo es fundamental. El λ es un parámetro que afecta cómo se penalizan los coeficientes del modelo durante el entrenamiento, con el objetivo de prevenir el sobreajuste y mejorar la generalización. La validación cruzada nos permite encontrar ese λ óptimo. A partir de un set de λs que elegimos, computamos el error de predicción de cada valor de λ. Luego seleccionamos aquel λ con el menor error de predicción, y reajustamos el modelo con este λ óptimo.

Con validación cruzada, para computar el error de predicción de cada valor de λ, aleatoriamente dividimos nuestras observaciones en k grupos (usamos $k = 10$) del mismo tamaño. El primer grupo es el set de validación y el resto el set de entrenamiento. Esto se repite k veces, donde cada grupo se alterna en ser el set de validación. El modelo se va ajustando sobre el test de entrenamiento, y se usa para predecir las respuestas para las observaciones del set de validación. El resultado del set de validación computa la tasa de error k veces hecho para cada una de los λ. La validación cruzada promedia la tasa de error para cada λ. Aquel λ con menor tasa de error es el óptimo a utilizar.

Se utiliza esta técnica antes que usar el conjunto de prueba para su elección porque garantiza una selección del modelo más generalizable al alternar cada grupo como set de validación. Esto genera una estimación mas robusta. Al mismo tiempo, si se usa solamente el conjunto de prueba puede llevar a sesgos en la evaluación si este grupo no es representativo.

Inciso 3

En validación cruzada, utilizar un k pequeño, como $k = 2$, implica dividir aleatoriamente las observaciones en dos partes iguales, donde una se utiliza para entrenamiento y la otra para validación, alternando sus roles en cada iteración. Este enfoque genera una alta variabilidad en el cálculo del error de validación, ya que las particiones pueden no ser representativas de la población completa y, al repetir el proceso aleatorizando los datos, los valores del error pueden cambiar significativamente. Además, con$k = 2$, el modelo se entrena con menos datos, lo que reduce su capacidad de generalización y afecta la estabilidad de las métricas calculadas.

Con un k muy grande, la variabilidad en el cálculo del error de validación disminuye considerablemente. Sin embargo, lo ideal es utilizar valores como $k = 5$ o $k = 10$, que suelen ofrecer un buen equilibrio entre estabilidad y eficiencia. Cuando k es extremadamente grande, llegando a $k = n$ (leave-one-out cross-validation), el sesgo se reduce significativamente, ya que el modelo se ajusta utilizando todas las observaciones excepto una en cada iteración. No obstante, este enfoque presenta un inconveniente principal de tipo computacional: el tiempo requerido para completarlo aumenta considerablemente a medida que n crece, lo que puede volverlo impráctico para conjuntos de datos grandes.

Inciso 4

A partir de la regresión logística, utilizando los dos métodos de regularización: Lasso y Ridge, la Tabla 3 presenta la matriz de confusión, los valores de AUC y Accuracy para los años 2004 y 2024. Al mismo tiempo, la Figura 2 y 3 presentan las curvas ROC de estos respectivos años.

Para evaluar el desempeño predictivo de cada modelo, se presentan las métricas de *Accuracy* (exactitud), AUC y las matrices de confusión de cada modelo para cada año. La medida de *Accuracy* mide la proporción de predicciones correctas sobre el total de predicciones. A pesar de ser una medida simple y clara, si la proporción de desocupados y ocupados no está balanceada, la exactitud puede ser engañosa y poco ilustrativa del ejercicio de predicción. La medida de AUC se refiere al área debajo de la Curva ROC de los *scores* predichos. La Curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para todos los umbrales del modelo. Por lo que, al medir el área por debajo, muestra qué tan bien el modelo predice como desocupados

a los verdaderos desocupados: a mayor AUC, mayor precisión predictiva. Por úlitmo, se muestran las matrices de confusión correspondientes. Éstas resumen el rendimiento del clasificador mostrando el número de predicciones correctas e incorrectas para cada categoría. En nuestro ejemplo se presentan en las filas las Y reales y en las columnas las Y predichas; la intuición de las matrices es la siguiente:

$$\begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix}$$

Tabla 2: Resultados de la Regresión Logística con Regularización Ridge y Lasso para los años 2004 y 2024.

Año	**Regularización**	**Métrica**	**Valor**
2024	Ridge	Accuracy	0.9601
		AUC	0.8502
		Matriz de Confusión	$\begin{bmatrix} 599 & 6 \\ 19 & 2 \end{bmatrix}$
	Lasso	Accuracy	0.9585
		AUC	0.8566
		Matriz de Confusión	$\begin{bmatrix} 598 & 7 \\ 19 & 2 \end{bmatrix}$
2004	Ridge	Accuracy	0.9333
		AUC	0.7950
		Matriz de Confusión	$\begin{bmatrix} 764 & 8 \\ 47 & 6 \end{bmatrix}$
	Lasso	Accuracy	0.9345
		AUC	0.8012
		Matriz de Confusión	$\begin{bmatrix} 765 & 7 \\ 47 & 6 \end{bmatrix}$

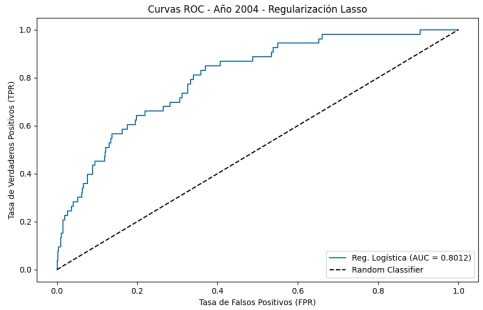
A simple vista, podemos ver que para el año 2024 la regresión logística con Ridge presenta una mayor *Accuracy* que con Lasso. Sin embargo, el AUC es mayor con Lasso que con Ridge. En comparación con el Trabajo Práctico 3, el cual realiza las predicciones con regresión logística pero sin regularizar los coeficientes estimados, tanto el *Accuracy* como el AUC son más chicos (0.9489 y 0.7395 respectivamente). Para el año 2004, la regresión logística con Lasso presenta una mayor *Accuracy* y AUC que con Ridge. En comparación con el Trabajo Práctico 3, para el año 2004, tanto el *Accuracy* como el AUC son más chicos (0.9321 y 0.6342 respectivamente).

Las Figuras 2 y 3, acompañado con los resultados previos, presentan las Curvas ROC para los años 2004 y 2024 con regularización Lasso y Ridge, respectivamente. A partir de estos resultados y su comparación con el previo trabajo, regularizando los coeficientes estimados tanto por Lasso como Ridge mejoran el *Accuracy* como el AUC.

Inciso 5

Para 2004, el λ^* óptimo en regresión logística con Ridge es de 355.65, y con Lasso es de 0.0028. Para 2024, el λ^* óptimo en regresión logística con Ridge es de 568.98, y con Lasso es de 0.0045.

Figura 1: Curvas ROC para el 2004 y 2024 con regularización Lasso.



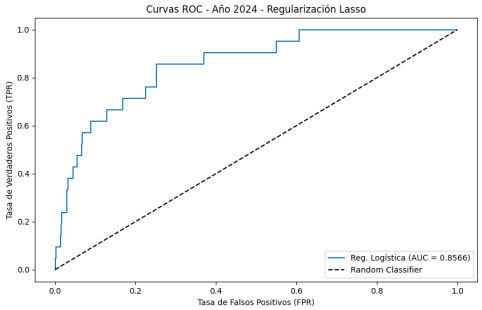
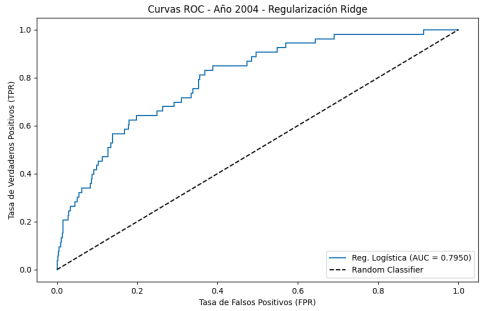
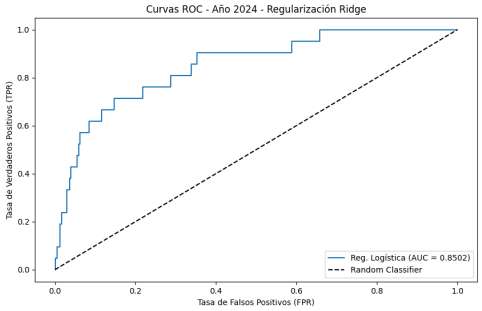


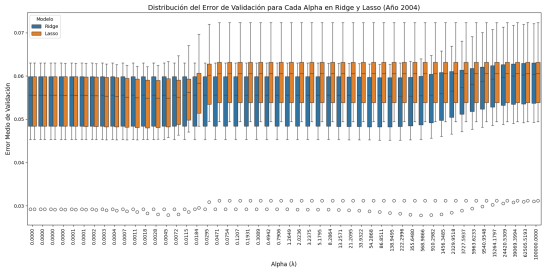
Figura 2: Curvas ROC para el 2004 y 2024 con regularización Ridge.



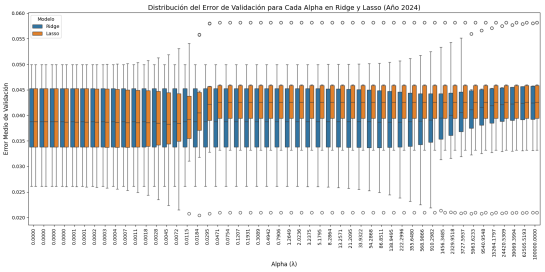


La Figura 4 y 5 presentan la distribución del error de predicción para cada λ usando boxplots para los años 2004 y 2024. A su vez, la Figura 6 presenta el promedio de la proporción de variables ignoradas por el modelo en función de λ.

Figura 3: Distribución del error de predicción para cada λ.



(a) 2004.



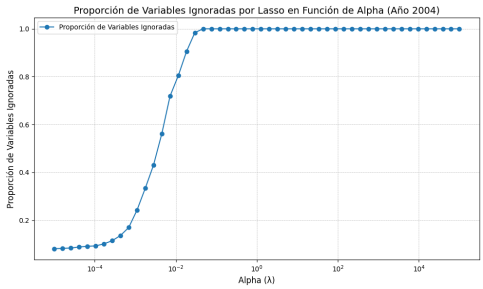
(b) 2024.

Inciso 6

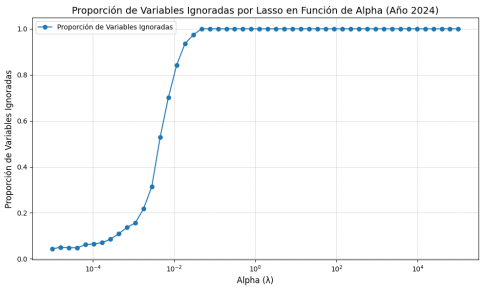
En 2004, las variables con coeficientes iguales a 0 incluyen el sexo del individuo, el ingreso total familiar, la proporción de personas con educación secundaria o más, y relaciones con el jefe del hogar (hermano/a, suegro/a o yerno/nuera). También se descartan indicadores educativos como nunca haber asistido a la escuela, actualmente asistir a la escuela, y el tipo de establecimiento educativo (privado o público). Además, variables relacionadas con el lugar de nacimiento (como otra localidad u otro país) y ayudas sociales, como la recepción de ayuda social o la participación en programas de empleo.

En 2024, aparecen nuevas variables con coeficientes iguales a 0, como el nivel educativo alcanzado, la proporción de niños en el hogar, y categorías adicionales dentro de

Figura 4: Promedio de la proporción de variables ignoradas por el modelo en función de λ.



(a) 2004.



(b) 2024.

variables existentes, como nuevas relaciones con el jefe del hogar y nuevas categorías de edad. También se incorporan indicadores sociales y subsidios adicionales, junto con nuevas categorías relacionadas con el lugar de nacimiento y la situación previa. Sin embargo, variables como la proporción de personas con educación secundaria o más, las relaciones familiares específicas mencionadas anteriormente, y ciertas ayudas sociales como la recepción de ayuda social o la participación en programas de empleo, ya no figuran entre las de coeficiente igual a 0 en 2024.

Hay varias variables descartadas que no fueron las que esperábamos, como aquellas relacionadas con el sexo del individuo, que podría haber sido un factor relevante dada la disparidad de género en el acceso al empleo en muchos contextos. También se descartaron variables asociadas al nivel educativo alcanzado, lo cual sorprende considerando que la educación suele estar estrechamente vinculada con las oportunidades laborales. Además, las relaciones familiares dentro del hogar, como hermano/a, suegro/a o yerno/nuera, también fueron descartadas, a pesar de que podrían influir en la dinámica económica del hogar y, en consecuencia, en el estado de desocupación. Por último, es llamativo que variables relacionadas con la participación en programas sociales y ayudas económicas, como la recepción de subsidios o asistencia estatal, no hayan tenido un peso significativo en el modelo, a pesar de su posible impacto en la capacidad de búsqueda de empleo.

Inciso 7

Al comparar los modelos de regularización Ridge y Lasso para los años 2004 y 2024, se observa que Lasso obtiene un mejor desempeño en ambos casos. En 2004, el error cuadrático medio (ECM) de Lasso fue ligeramente menor (0.0555) que el de Ridge (0.0562), indicando una ventaja sutil de Lasso en la capacidad de generalización. En este caso, aunque la diferencia es pequeña, Lasso se beneficia de su capacidad para reducir a cero los coeficientes de características menos relevantes, simplificando el modelo.

En 2024, la diferencia entre ambos métodos es mucho más significativa. Lasso obtuvo un ECM de 0.0302, muy inferior al de Ridge, que fue 0.0562. Esto sugiere que Lasso fue mucho más eficaz en este año, probablemente debido a su capacidad para manejar datos complejos mediante la selección de características irrelevantes o redundantes. En general, Lasso demostró ser el método de regularización más adecuado para ambos años, especialmente en 2024, donde su capacidad de simplificación y ajuste adecuado fue clave para su mejor desempeño.