Universidad de SanAndrés

Ciencia de Datos
Trabajo Práctico N° 4

Profesora: Maria Noelia Romero
Tutor: Ignacio Spiousas

Gabriel Alejandro Díaz y Gabriela Belen Sanchez

28 de noviembre de 2024

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

1. Posibles variables predictivas de la desocupación

Tras analizar el diseño de registro de la base de hogares se considera que las variables que pueden ser predictivas de la desocupación son:

IV1 -tipo de vivienda- y II7 -Régimen de tenencia-

- Por lo general, el tipo de vivienda en la que una persona vive se encuentra relacionado con las condiciones socioeconómicas de esa persona y/o su familia. De esta forma, el vivir en una casa o departamento puede reflejar una mayor estabilidad, mientras que una habitación puede reflejar un nivel más bajo de ingresos.

II1 -¿Cuántos ambientes/habitaciones tiene este hogar para su uso exclusivo?-

- El número de habitaciones puede ser un indicador del tamaño de la vivienda, por lo que aquellos hogares con un menor espacio pueden relacionarse con una mayor inestabilidad laboral y económica.

V13 -Gastar de lo que tenían ahorrado-, V14 -Pedir préstamos a familiares/amigos-, V15 -Pedir préstamos a bancos, financieras, etc.- y; V17 -¿Han tenido que vender alguna de sus pertenencias?-

- Todas estas pueden ser estrategias empleadas cuando los miembros de un hogar están pasando por dificultades económicas recientes, ya que hacen referencia al mes anterior. Estos cambios económicos relativamente recientes pueden ser causados, entre otras cosas, por situaciones de falta de un empleo estable, desempleo o subempleo en el hogar.

2. Descarga, filtrado y unión de bases

Se utilizaron distintas bases de datos para analizar los determinantes de la desocupación en los aglomerados de Ciudad Autónoma de Buenos Aires (CABA) y Gran Buenos Aires (GBA). Para ello, se trabajó con tres bases: una base de hogares correspondiente al año 2004 (Hogar_t104.dta), una base de hogares del año 2024 (usu_hogar_T124.xlsx) y una base individual (respondieron_limpio.csv), correspondiente al TP3.

Se cargaron las bases de datos a través de la biblioteca pandas para manejar eficientemente los diferentes formatos de archivo (.dta, .xlsx, .csv). Para evitar inconsistencias en las operaciones posteriores, se normalizaron los nombres de las columnas y se las convirtio a minúsculas. Posteriormente, se realizó un filtrado geográfico para incluir únicamente los registros correspondientes a CABA y GBA. En la base de 2004, se empleó el campo textual aglomerado, donde se seleccionó únicamente las categorías Ciudad de Buenos Aires y Partidos del GBA. Por su parte, en la base de 2024, se utilizaron los códigos numéricos 32 (CABA) y 33 (GBA) en el campo AGLOMERADO. Este paso permitió restringir el análisis a los aglomerados solicitados.

Una vez que se filtraron las bases de hogares, estas fueron concatenadas para formar una única base de datos de hogares que abarcara ambos años. Luego, esta base combinada de hogares se integró con la base individual mediante una operación de *merge*, donde se utilizó las claves codusu y nro_hogar. Estas claves buscan garantizar que cada registro individual esté vinculado a la información de su hogar correspondiente.

3. Limpieza de la Base

Valores Faltantes

Observamos, en primer lugar, la cantidad de valores con no NaN que había en cada columna y, en segundo lugar, la cantidad de NaN por columna en las variables que resultaron con diferencias en la primera parte (iv1_esp, iv3_esp, iv7_esp, ii7_esp, ii8_esp, ix_men10, idecifr, pdecifr, ideccfr, pdeccfr, idimph, pondih). Luego se utilizó la función *value.counts()* para cada una de estas variables y se analizó los casos donde era conveniente agrupar respuestas (como en el caso de ii8_esp, donde se agruparon las respuestas relacionadas a electricidad). Se descartaron las variables que no resultaban de interés para el análisis predictivo.

Detección y Tratamiento de Outliers

Para la detección de outliers primero se hizo una tabla de las variables numéricas con la base EPH_completo_H y la función describe().T, para observar los valores mínimos y máximos y la concentración de los datos. Se realizaron boxplots en aquellas variables con mínimos y/o máximos extremos y donde las concentraciones de los datos estuvieran sesgadas en algún cuartil. También se realizaron boxplots en las variables que usamos en el TP3 para revisar outliers. Se utilizó el método de Desviación Absoluta Mediana (MAD) para eliminar valores atípicos en la variable ipcf_I e ix_tot. Se establecieron límites basados en la mediana ± 3*MAD, donde se elimina las observaciones que se encontrasen fuera de este rango. Solo se sacaron los outliers en las variables más relevantes (codusu, estado, nro_hogar, desocupado, ch04, ch06, pareja_actual, cobertura_m, ch09, ch15, nivel_ed_or, ipcf, iv1, iv2, iv3, iv4 , iv6 , iv8 , iv11 , iv12_1 , iv12_2 , iv12_3 , ii1 , ii2 , ii4_1 , ii7 , ii8 , ii9 , v1 , v2 , v3 , v5, v7, v12, v13, v14, v15, v16, v17, ix_tot, ix_men10, ix_mayeq10 , ipcf_I , vii1_1, ano4)

Mapeo y dummies

Debido a que la base de 2004 y 2024 tienen codificaciones distintas, se realizó un mapeo de las variables seleccionadas para que tuvieran los mismos valores de entrada en la base. A su vez, se transformaron las variables categóricas a dummies. Luego de esto, nos quedamos con 128 variables en total para los modelos.

4. Creación de nuevas variables

La primera variable creada fue el número de habitaciones por persona en el hogar (HpP), calculada como la relación entre el número de habitaciones disponibles (ii1) y el total de personas que residen en el hogar (ix_tot). Este indicador busca medir el nivel de hacinamiento, una condición que frecuentemente se asocia a vulnerabilidad económica. En hogares donde ii1 es igual a cero, se asignó un valor de cero a esta variable para evitar resultados indefinidos.

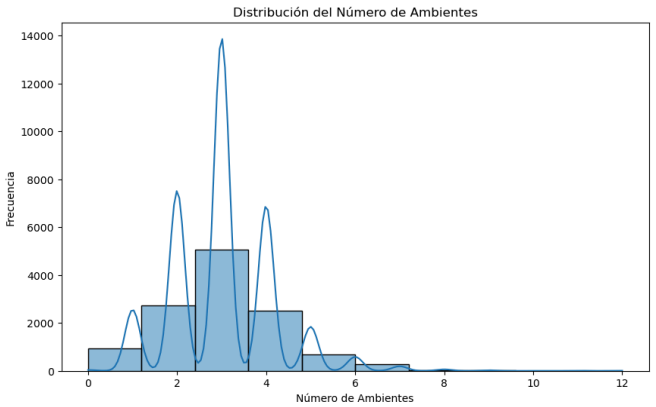
La segunda variable fue la proporción de personas ocupadas en el hogar (trabajadores_hogar). Para obtener esta información, se identificaron las personas ocupadas dentro del hogar (estado=1) y se agruparon por hogar para contar la cantidad de trabajadores. Esta información se unió al conjunto de datos original, donde se asigna un valor de cero a aquellos hogares sin trabajadores. Este indicador es relevante porque podría reflejar la capacidad del hogar para generar ingresos y mantener su estabilidad económica.

La tercera variable creada fue el ingreso familiar per cápita ajustado por niño (ipcf_niño), calculado como la relación entre el ingreso familiar per cápita (ipcf_I) y el número de niños menores de 10 años (ix_men10). Este indicador busca evaluar cómo se distribuyen los recursos disponibles entre los integrantes más jóvenes del hogar. En los hogares donde no hay niños, el valor de esta variable se asignó como cero.

5. Estadística descriptiva

Figura 1.

Distribucion del numero de ambientes por hogar

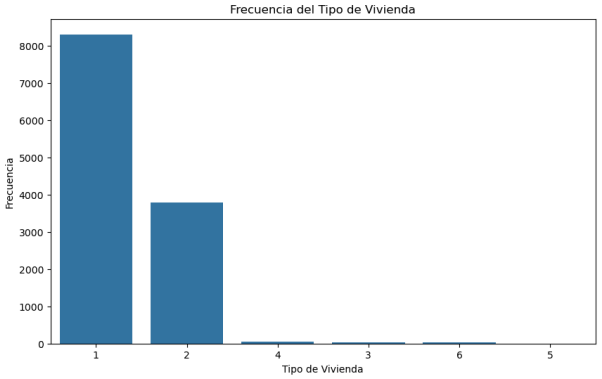


Nota. El histograma muestra la cantidad de ambientes por hogar.

La figura 1 muestra la distribución del número de ambientes (II1) en un hogar. La mayoría de las observaciones (5057 registros) están concentradas en valores bajos, principalmente entre 2 y 4 ambientes (*M*: 1.93, *SD*: 1.62). Los valores extremos (0 a 12) son muy poco frecuentes y representan menos del 1% de los datos. La asimetría a la izquierda sugiere que las viviendas con menos ambientes predominan ampliamente en el conjunto.

Figura 2.

Frecuencia de tipo de vivienda



Nota. El gráfico describe la frecuencia del tipo de vivienda de los hogares, donde 1 es casa, 2 departamento , 3 pieza de inquilinato, 4 pieza en hotel/pensión, 5 local no construido para habitación y 6 otros.

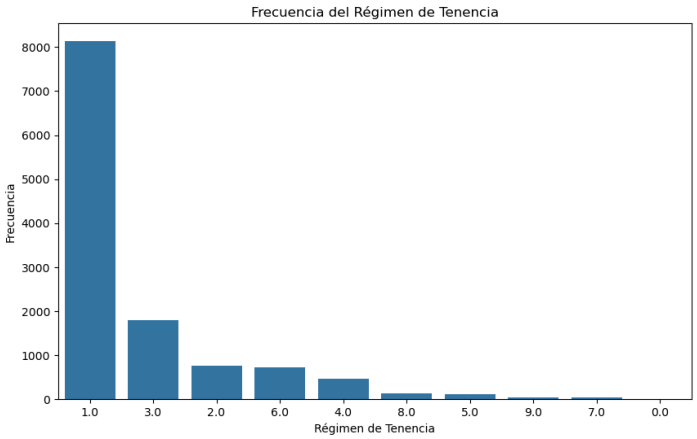
La figura 2 muestra la frecuencia de 6 tipos de vivienda en la que vive el hogar. Los valores varían desde 1 (independiente, casa propia) hasta 6, según el tipo de vivienda. La media es de 1.34, lo que podría indicar que, en promedio, los hogares encuestados viven en viviendas con un tipo clasificado como 1 (ver que significaba 1). La desviación estándar es de 0.55, lo que muestra una dispersión moderada en los tipos de viviendas.

La gran mayoría de los hogares (8314) reportan tener una vivienda de tipo 1. Hay una proporción significativa de hogares (3799) con tipo 2. Las categorías de viviendas 4, 5 y 6 tienen muy pocos casos (menos de 50 en total), lo que sugiere que estos tipos de viviendas son menos comunes en los hogares encuestados. Vivir en casas propias (tipo 1) podría reflejar una mayor estabilidad económica

en comparación con aquellos que alquilan. La prevalencia de hogares con tipo 1 (casa propia) sugiere una fuerte tendencia hacia la propiedad en este conjunto de datos, aunque hay una presencia significativa de alquileres.

Figura 3.

Frecuencia del régimen de tenencia



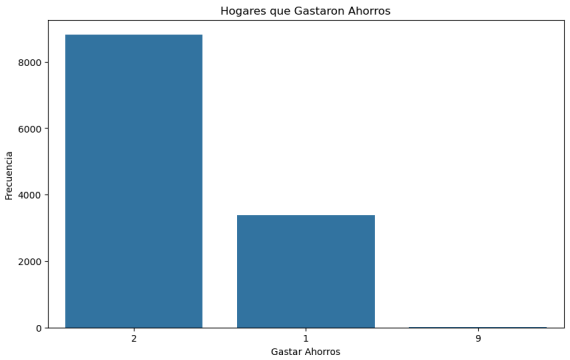
Nota. El gráfico describe la frecuencia del régimen de tenencia, donde 1 es propietario de la vivienda y el terreno, 2 propietario de la vivienda solamente, 3 inquilino / arrendatario de la vivienda, 4 ocupante por pago de impuestos / expensas, 5 ocupante en relación de dependencia, 6 ocupante gratuito (con permiso), 7 ocupante de hecho (sin permiso), 8 está en sucesión y 9 otra situación.

La figura 3 describe la frecuencia del régimen de tenencia de los hogares, donde este último refleja si el hogar es propietario, alquilado o tiene otro tipo de relación con su vivienda. Las personas que viven en los hogares son en su gran mayoría propietarios (8132), pero hay una amplia variabilidad en los regímenes de tenencia (*M:* 1.93, *SD*: 1.62). En este sentido, un porcentaje menor vive en alquiler u otros regímenes (1791 con régimen 3, 765 con régimen 2, etc.). Hay una pequeña proporción de hogares que están en situaciones más inusuales o no especificadas, como el valor 9 (37 hogares).

Un alto porcentaje de hogares siendo propietarios podría indicar una base económica más sólida. Sin embargo, el hecho de que un número significativo viva en régimen de alquiler refleja una segmentación social que podría tener implicaciones para la predicción de la desocupación.

Figura 4.

Cantidad de hogares que gastaron ahorros



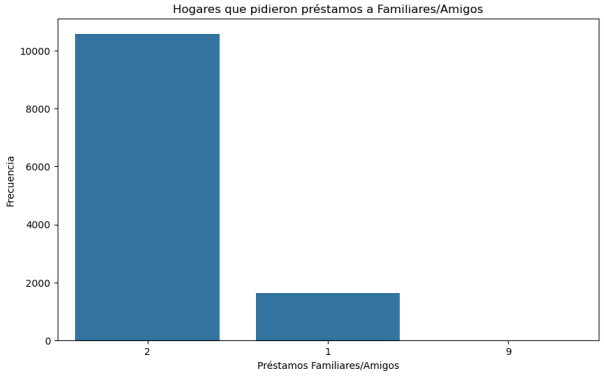
Nota. El gráfico describe la cantidad de hogares que gastaron sus ahorros, donde 2 indica que no gastaron sus ahorros y 1 que si.

La figura 4 nos indica si los hogares han tenido que usar sus ahorros recientemente. Gran parte de los hogares no ha gastado sus ahorros (8822) y la mayoría de las respuestas se agrupan alrededor de 2 (*M*: 1.73, *SD:* 0.49). Un número menor (3392) no ha ahorrado (valor 1).

Gastar ahorros podría reflejar un posible signo de vulnerabilidad económica. Los hogares que gastaron sus ahorros podrían haberse enfrentado con situaciones de inestabilidad económica y/o desempleo reciente. Esta variable puede correlacionarse con el riesgo de desocupación, ya que las personas podrían recurrir a sus ahorros cuando enfrentan dificultades laborales.

Figura 5

Cantidad de hogares que pidieron préstamos a familiares y/o amigos

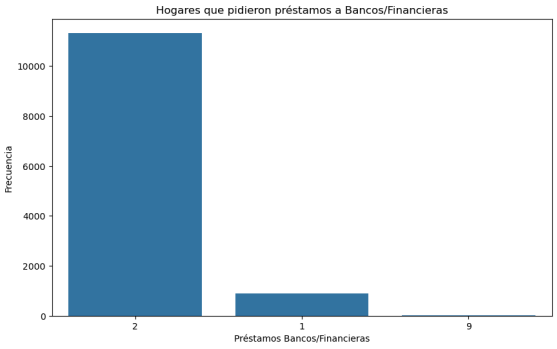


Nota. El gráfico describe la cantidad de hogares que recurrieron a préstamos a familiares o amigos, donde 1 indica que pidieron préstamos, 2 que no lo hicieron y 9 no sabe/no responde.

La figura 5 nos muestra si los hogares han tenido que recurrir a pedir préstamos a familiares o amigos. Gran parte de los hogares no han pedido préstamos (10571) y hay baja variabilidad en las respuestas (*M:* 1.87, *SD:* 0.38). Un número mucho menor (1646) ha tenido que pedir préstamos (valor 1). La mayoría de los hogares no han tenido que recurrir a esta opción, pero aquellos que lo hacen probablemente enfrentan dificultades laborales o de ingresos.

Figura 6

Cantidad de hogares que pidieron préstamos a Bancos/Financieras

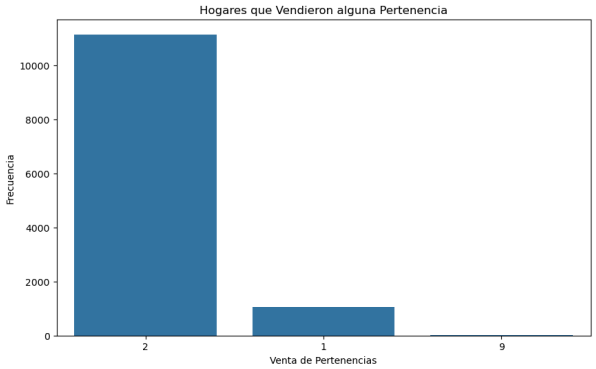


Nota. El gráfico describe la cantidad de hogares que recurrieron a préstamos a Bancos/Financieras, donde 1 indica que pidieron préstamos, 2 que no lo hicieron y 9 no sabe/no responde.

Esta figura refleja si los hogares han obtenido préstamos a través de entidades financieras. La mayoría de los hogares (11325) no han tenido que recurrir a préstamos bancarios (valor 2) (*M:* 1.93, *SD:* 0.31). Un pequeño número de hogares (892) ha pedido préstamos a bancos/financieras (valor 1).

Figura 7

Cantidad de hogares que vendieron alguna pertenencia



Nota. El gráfico describe la cantidad de hogares que vendieron alguna pertenencia, donde 1 indica que vendieron, 2 que no lo hicieron y 9 no sabe/no responde.

La figura 7 muestra si los hogares han tenido que vender alguna de sus pertenencias debido a dificultades económicas. La mayoría de los hogares (11150) no han tenido que vender pertenencias (valor 2) y hay baja variabilidad en las respuestas (*M:*1.92, *SD*: 0.33). Un pequeño número (1067) ha vendido pertenencias (valor 1).

Parte II: Clasificación y regularización

Parte 1: Selección de modelos LASSO y Ridge mediante validación cruzada

A continuación, con el fin de predecir la variable "desempleo" de un individuo, se construyeron tanto un modelo LASSO (*Least Absolute Shrinkage and Selection Operator*), como un modelo Ridge, mediante un proceso de validación cruzada. Ambos modelos se caracterizan por ser métodos de regularización, donde se imponen restricciones sobre los estimadores de las variables X, con el fin de reducir el sobreajuste y mejorar la capacidad de generalización. La principal diferencia entre ambos se trata del tipo específico de penalización: LASSO utiliza una penalización L1 que fuerza a algunos coeficientes a ser 0, es decir, hace una selección de las variables independientes; mientras que el penalizador de Ridge (L2), simplemente los reduce lo más posible, pero nunca llegan a 0. La fuerza que toman estas regularizaciones viene dada por el hiper-parámetro λ (lambda), cuyo valor óptimo se puede determinar mediante una validación cruzada.

Parte 2 y 3: Elección de λ mediante *cross-validation*

Para elegir λ, en primer lugar se dividió a la base de datos compuesta por la encuesta hogares e individual de la EPH según el año (2004 y 2024). Luego, por cada año, se realizó una segunda subdivisión, donde se generaron datos de entrenamiento (*n_2004* = 2379: *n_2024* = 1689) y de testeo (*n_2004* = 1020: *n_2024* = 725). Cada base de entrenamiento, a su vez, se separó en 10 (valor *k*, número de particiones) subconjuntos, entrenando el modelo en 9 subconjuntos y probándolo en el restante.

De esta forma, el valor del parámetro *k* toma una gran relevancia, ya que determina cuantas veces se entrena y evalúa el modelo. Un valor de *k* pequeño (2 o 3) implica que cada partición de la muestra de entrenamiento es mayor (contiene más observaciones, en términos absolutos) y, a su vez, la cantidad de iteraciones es menor, debido a que se presentan pocas combinaciones de las particiones de prueba y

entrenamiento. Esto, aunque favorece su costo computacional, puede llevar a una estimación menos precisa debido al aumento de la varianza que puede ser causado por posibles sesgos en las particiones. Por otro lado, cuando *k* es muy grande (20 o más), implica que el conjunto de prueba será más grande, ya que cada partición contendrá una menor cantidad de datos. Esto, aunque aumenta la estabilidad del modelo, su precisión y robustez, también incrementa el cómputo del modelo, ya que se incrementa la cantidad de iteraciones que realiza.

En el caso de que k = n (n = número de muestras), es decir, que se tengan un total de particiones igual a la cantidad de datos, cada muestra es utilizada una vez como conjunto de prueba, mientras el resto es utilizado como conjunto de entrenamiento (metodo de validacion conocido como LOO, *Leave One Out*). De esta forma, el modelo se estima n veces, lo que da lugar a una estimación mucho más precisa, a costas de un elevado costo computacional y más susceptible a una variabilidad alta por parte de los datos.

De esta forma, el proceso de selección de λ es el siguiente:

1. Se genera una lista de posibles valores de λ (en este caso, $\lambda = 10^n$ con $n \in \{-5, -4, -3 ..., +4, +5\}$
2. Para cada valor de **λ**, se entrena el modelo utilizando validación cruzada (k =10) para obtener una estimación del error de generalización en el subconjunto de entrenamiento.
3. Se elige el valor de **λ** que minimice el error promedio sobre los diferentes pliegues de la validación cruzada (ECM)
4. Se usa este valor óptimo de **λ** para ajustar el modelo final.

En este punto es de suma relevancia no utilizar el conjunto de prueba (*test*) debido a que son estos los valores que se busca predecir. Si se lo incluye en la sección de **λ,** incluso aunque no sea considerado en el modelado en sí, esta información llevaría a una evaluación sesgada y sobre-optimista del modelo. En otras palabras, el modelo final, aunque tenga un buen desempeño en la muestra de testeo, cuando se busque predecir nuevos datos perderá efectividad, ya que el conjunto de prueba está influenciado por la selección de los hiperparámetros. De esta manera, el conjunto de prueba sirve como un último control de rendimiento del modelo en datos "nuevos", donde se evalúa si el modelo final es capaz de generalizarse, en lugar de predecir datos ya evaluados.

Parte 4: Ridge y LASSO con λ = 1 y comparación con TP3

Como parte del análisis para la generación de un modelo predictivo, se realizaron modelos de LASSO y Ridge con un valor de λ = 1 para cada año. En la tabla 1 se presentan las métricas obtenidas respecto de su desempeño y en la figura 8 se presentan las curvas ROC respectivas.

Tabla 1

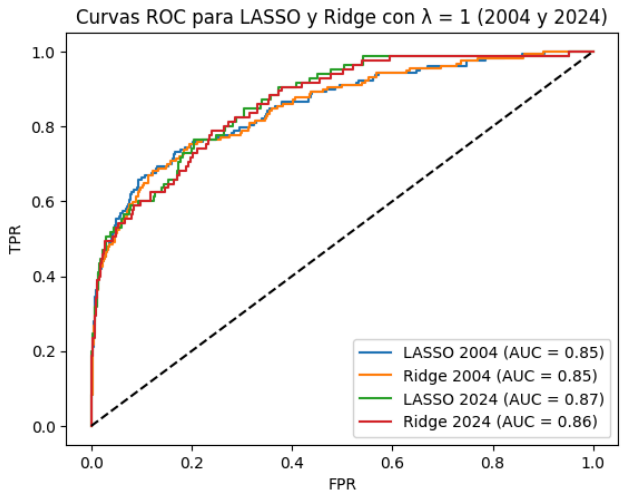
Métricas de evaluación de los modelos con λ = 1

Año	**2004**				**2024**			
Modelo	**LASSO**		**Ridge**		**LASSO**		**Ridge**	
Accuracy	0.8950		0.8941		0.9186		0.9158	
Matriz de	849	14	848	15	626	14	624	16

Confusión	93	64	93	64	45	40	45	40
AUC	0.8535		0.8508		0.8675		0.8632	
ECM	0.1049		0.1058		0.0813		0.0841	

Figura 8

Curvas ROC para LASSO y Ridge con λ = 1



Como se puede apreciar, tanto LASSO como Ridge muestran un rendimiento similar en ambos conjuntos de datos, con el primero ligeramente superando al segundo en términos de exactitud y AUC en ambos casos. Para los datos de 2004, se logra una exactitud cercana al 89.5%, mientras que en 2024, ambos mejoran cerca del 91.5%. Por otro lado, presentan un ECM similar, lo que sugiere que tienen un desempeño comparable en este término, pero LASSO muestra una leve ventaja en cuanto a la discriminación, como lo indica su mayor AUC.

Al comparar estos resultados con el modelo de regresión logística del TP3 se observan diferencias significativas. En primer lugar, aunque la cantidad de falsos negativos sigue siendo significativa, en los modelos regularizados es menor para ambos años en comparación con el modelo anterior (93 en 2004 y 45 en 2024 [LASSO y Ridge] vs. 152 y 82 [regresión logística]). Por otro lado, en cuanto al rendimiento discriminativo, la curva ROC muestra un AUC superior en los modelos actuales (0.85 para 2004 y 0.87 para 2024) en comparación con los valores del TP3 (0.76 y 0.73). Sin embargo, la precisión es bastante similar en ambos análisis, siendo 89.5% en 2004 y 91.5% en 2024 en los modelos regularizados, mientras que en el modelo anterior es de 93% en 2004 y 95% en 2024, lo que indica un rendimiento ligeramente superior de este último.

Con esto en mente, es posible concluir que la regresión logística con regularización (LASSO y Ridge) muestra un rendimiento ligeramente superior en términos de AUC y precisión en comparación con los resultados previos.

Parte 5: 10-fold CV para seleccionar el mejor λ

Como próximo paso del análisis se realizó un barrido sobre el parámetro de regularización λ utilizando una validación cruzada 10-fold (k=10) para determinar su valor óptimo. De esta forma, el objetivo principal es analizar el rendimiento de ambos modelos bajo diferentes valores de λ y seleccionar aquélla que presente el menor error de predicción.

Tras realizar este proceso, se obtuvieron los siguientes modelos:

Tabla 2

Métricas de evaluación de los modelos con k=10

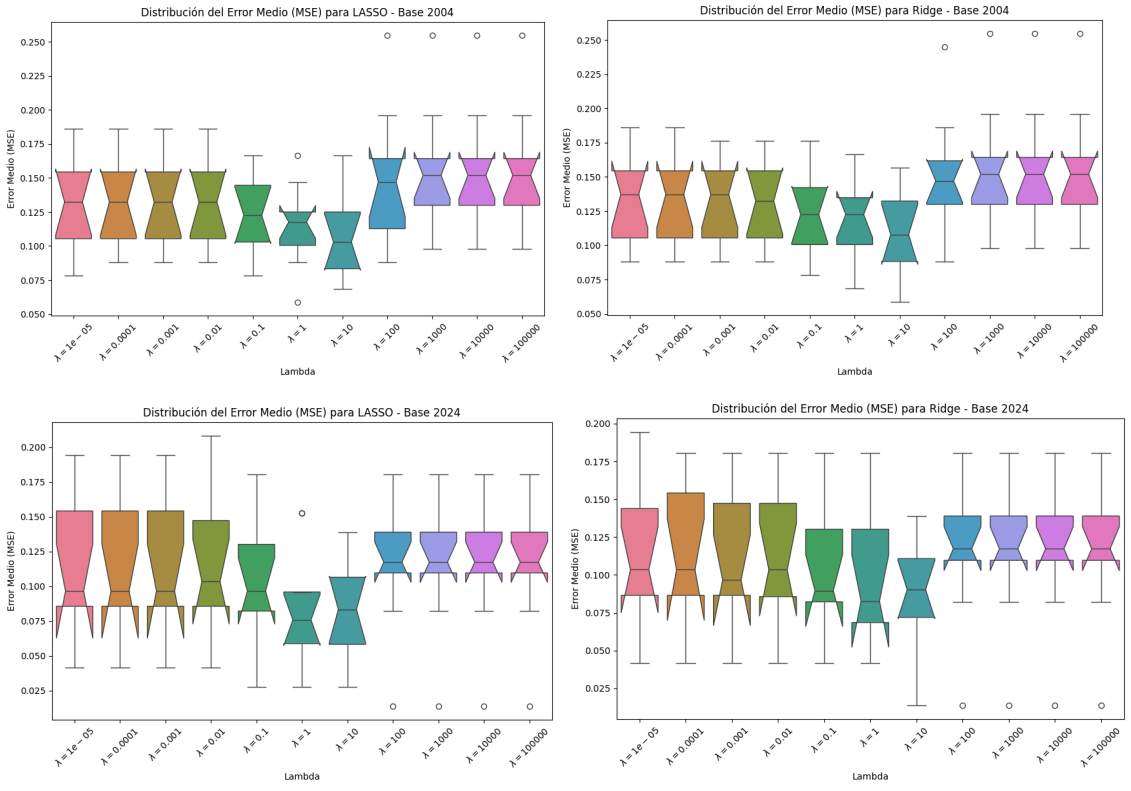
Año	**2004**	**2024**

Modelo	LASSO	Ridge	LASSO	Ridge
λ (Lambda)	**10**	**10**	**10**	**10**
Accuracy	0.8970	0.8980	0.9186	0.9144
AUC	0.6864	0.6948	0.6937	0.6812
ECM	0.1029	0.1019	0.0813	0.0855

Como se puede apreciar, para todos los modelos se seleccionó λ = 10. En todos estos casos, se presentaron distribuciones de los errores de predicción extrañas, como se puede observar a continuación.

Figura 9

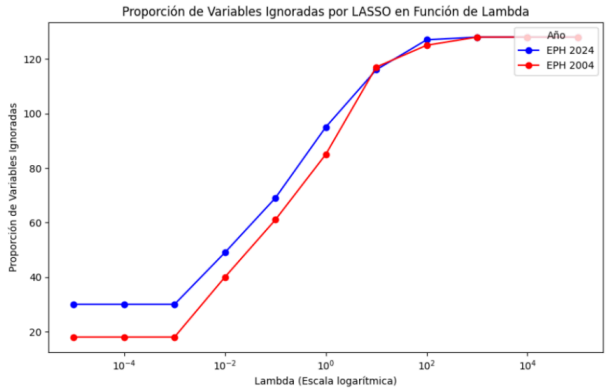
Distribución del Error Medio para LASSO y Ridge



Para los modelos de 2024, se puede notar una elevada sensibilidad a los valores de λ bajos, de forma tal que el MSE aumenta y se encuentra muy disperso en estos casos, tanto para Ridge como para LASSO. En el caso de 2004, los errores muestran un comportamiento similar, pero mucho menos disperso en los valores extremos de λ. En todos los casos se puede observar como el MSE mínimo se encuentra entre λ = 1 y λ=10.

Figura 10

Proporción de variables ignoradas por LASSO en función de Lambda



Por otro lado, al poner foco sobre los modelos de LASSO generados, se puede apreciar como la proporción de variables ignoradas en función de λ es muy similar para ambos años. En ambos casos se observa un incremento significativo conforme aumenta la regulación de las variables. La principal diferencia observable se presenta en que, al menos en los valores de λ más bajos, el modelo de LASSO de 2024 descarta una mayor cantidad de variables.

Al profundizar en el caso del valor óptimo de λ para estos modelos, se vio que una gran cantidad de variables fueron descartadas. Específicamente, se descartaron un total de 111 variables para el modelo de 2004 y 117 para el modelo de 2024. Entre ellas, las más llamativas fueron II1 (¿Cuántos ambientes/habitaciones tiene este hogar para su uso exclusivo?) y II7 (Régimen de tenencia), las cuales se descartaron en ambos modelos. Finalmente, también fueron descartadas las variables V13 y V14 (gasto ahorro y pedir préstamos a familiares/amigos, respectivamente). Esto va en contra de las predicciones realizadas en la parte 1 de este trabajo, ya que se descarta la mayoría de las variables que se consideraban relevantes. Entre estas, solo IV1 (“Tipo de vivienda”), v15_2 (“¿Pedir préstamos a bancos, financieras, etc.?”) y V17_2 (“¿Han tenido que vender alguna de sus pertenencias?”) se mantuvieron en por lo menos uno de los dos modelos.

Parte 6: Comparación entre años 2004 y 2024 de los mejores modelos.

Al observar el resto de métricas (Tabla 2), se puede notar que el modelo Ridge de 2004 tiene el mejor AUC (0.6948), pero no la mejor precisión, premio que se lleva LASSO de 2024 (0.9186). Si nos enfocamos en cada año en particular, LASSO 2004 y Ridge 2004 son muy similares, pero este último es ligeramente superior en los tres aspectos: ECM (0.1019 vs. 0.1029), AUC (0.6948 vs. 0.6864) y precisión (0.8980 vs. 0.8970). En 2024, ocurre lo contrario, con LASSO siendo ligeramente superior a Ridge: ECM (0.0813 vs. 0.0855), AUC (0.6937 vs. 0.6812) y precisión (0.9186 vs. 0.9144).

Por otro lado, como se mencionó anteriormente, los modelos de LASSO hicieron una selección distinta de variables para predecir el desempleo. Específicamente tuvieron una diferencia de 11 variables, entre las que se destacan: “ipcf_i” (Ingreso per cápita familiar) y “ipcf_i_niño” (Ingreso per cápita familiar por niño), las cuales, a pesar de su aparente relevancia para el desempleo, solo son consideradas por el modelo de 2004. Similarmente, en 2024 se consideran las variables “V2” (“Las personas de este hogar han vivido de alguna jubilación o pensión”) e “IV6” (“Si tiene agua en el terreno”), entre otras, que no son consideradas en 2004.

En resumen, el análisis de las métricas y la selección de variables revela que LASSO y Ridge presentan un desempeño similar, con LASSO destacándose en precisión y Ridge en AUC. Si bien ambos modelos consideran variables relevantes para el desempleo.