Informe Trabajo práctico N°4- Análisis de hogares y desocupación

Alumnos: Strejilevich Francisco, Giacobelli Francisco
Profesores: María Noelia Castillo, Ignacio Spiousas
Materia: Ciencia de Datos
3/11/2024

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

Variables que pueden predecir la desocupación

Como estilo de continuación al análisis desarrollado en el TP3 la incorporación de la base de hogares a los análisis sobre desocupación permite un análisis más complejo sobre los datos que realizamos en el trabajo práctico anterior. Mientras que en el TP3 pudimos identificar patrones más generales, estudiar esos fenómenos a nivel hogar es interesante para ver cómo las dinámicas familiares, domésticas o de convivencia se relacionan con el estado laboral de las personas.

Al explorar el diseño de la base de hogares, identificamos variables como la cantidad de integrantes (`IX_TOT`), el tipo de vivienda que habitan (`IV1`), el tipo de piso que tiene principalmente la casa (`IV3`), cómo se accede al agua (`IV6`), se decidió utilizar estas variables porque eran algunas de las que definían como era el estado de la vivienda que se estaba analizando, por lo que se podría considerar características demográficas del hogar analizado. Se consideró también utilizar la variable ITF que incluye el ingreso total familiar, pero un análisis demostró que tenía demasiados valores negativos en la base de datos de 2024, lo mismo con la variable IPCF de la misma base, por lo que se decidió no utilizarlas.

Bases de datos y filtrado por CABA y CBA

Para este análisis, se utilizaron las bases de microdatos de la Encuesta Permanente de Hogares (EPH) correspondientes al primer trimestre de 2004 y 2024. La base de hogares fue combinada con la base individual utilizando las claves `CODUSU` y `NRO_HOGAR`, lo que permitió unificar la información en una estructura de datos.

Como primer paso, se filtraron únicamente los datos de los aglomerados de Ciudad Autónoma de Buenos Aires (CABA) y Gran Buenos Aires (GBA) como áreas de interés de estudio. Posteriormente, se eliminaron columnas que consideramos irrelevantes para el análisis. Para esto,

utilizamos criterios cómo variables que tomen respuestas de lo que respondieron en otras variables, categorías derivadas de ingreso y variables de ponderación.

Limpieza de Base de datos

En el tratamiento de la base de datos primero identificamos y eliminamos observaciones con valores faltantes en variables clave como `CH06` (edad), `IX_TOT` (cantidad de personas en el hogar) y `IPCF` (ingreso per cápita familiar). Para los valores extremos (outliers), reemplazamos aquellos fuera de rango esperable (como 9, 99, 999 y 9999, que indican respuestas no válidas) por valores nulos (`NaN`), los cuales posteriormente fueron eliminados para evitar sesgos en los cálculos.

En el caso de las variables categóricas, transformamos aquellas con información textual, como `CH07` (estado civil), a valores numéricos mediante mapeos predefinidos. Por ejemplo, categorías que van desde "Casado" o "Unido" hasta “Soltero” se codificaron numéricamente para facilitar su análisis y construcción de variables adicionales. Este enfoque también lo tuvimos que aplicar a `ESTADO`, que presentaba diferencias en sus codificaciones entre los años 2004 y 2024, la normalizamos para garantizar consistencia en las comparaciones interanuales.

Por último, constantemente revisamos el tamaño de la muestra, las cantidades de valores nulos y distribuciones con las que contábamos para asegurarnos no tener valores atípicos ni modificaciones muy grandes que se alejen de la consigna.

Construcción de variables relevantes para predecir desocupación

Una vez limpia la base, se construyeron cinco nuevas variables que creemos que pueden predecir la desocupación de forma significativa. La primera de ellas fue la edad promedio del hogar, registrada como `EDAD_PROMEDIO_HOGAR`. Esta variable permite identificar una media de edad dentro del hogar. Por ejemplo, hogares predominantemente jóvenes podrían enfrentarse a tasas más altas de desocupación debido a la inexperiencia laboral o la inserción tardía en el mercado, asimismo, hogares con indices de edad más altos podrían reflejar una estabilidad laboral alta debido a la convivencia de personas con mayor edad aumenta la probabilidad de que al menos uno esté ocupado.

Se creó además la variable `NIVEL_EDUC_NUM`, que captura el nivel educativo en términos numéricos, facilitando así el cálculo de niveles educativos promedios tanto por hogar como generales. Básicamente permite tener una medida general del nivel de formación académica de los integrantes, partiendo del supuesto de que una mayor educación promedio está asociada con mejores oportunidades laborales y menores riesgos de desocupación. Complementando esta información, se añadió la variable `PROP_EDUC_SUPERIOR`, que mide la proporción de

personas con educación superior dentro del hogar. Que permite evaluar si la incidencia de educación superior dentro del hogar se correlaciona con una mayor estabilidad laboral y menores tasas de desempleo.

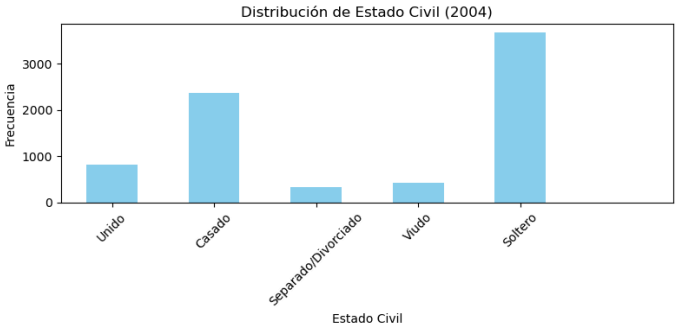
Finalmente, similar al caso anterior se transformó a américa la variable `CH07` que contiene el estado civil, creando `ESTADO_CIVIL_NUM`. Gracias a esto, luego se incluyó una variable binaria denominada `TIENE_PAREJA`, que distingue entre hogares con personas en pareja (casados o unidos) y aquellos que no se encuentran en esta condición. Esta variable facilita evaluar cómo la estructura familiar podría influir en el acceso al empleo, dado que los hogares con parejas pueden presentar dinámicas laborales y económicas más consolidadas.

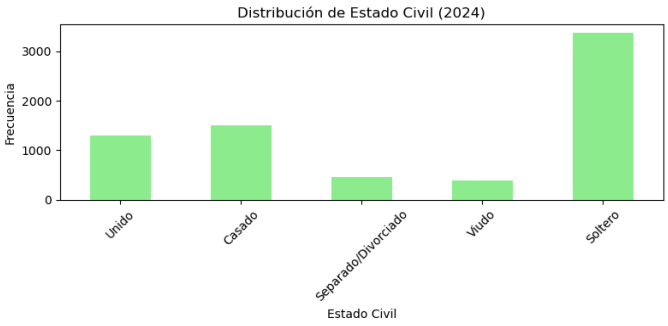
Análisis estadístico descriptivo de las variables

Se realizaron múltiples análisis descriptivos a partir de las variables construidas y algunas otras consideradas relevantes para predecir o analizar fenómenos de la desocupación. En primer lugar, realizamos algunas estadísticas descriptivas generales de las variables `EDAD_PROMEDIO_HOGAR`, `TIENE_PAREJA`, `PROP_EDUC_SUPERIOR` y `ESTADO_CIVIL_NUM`.

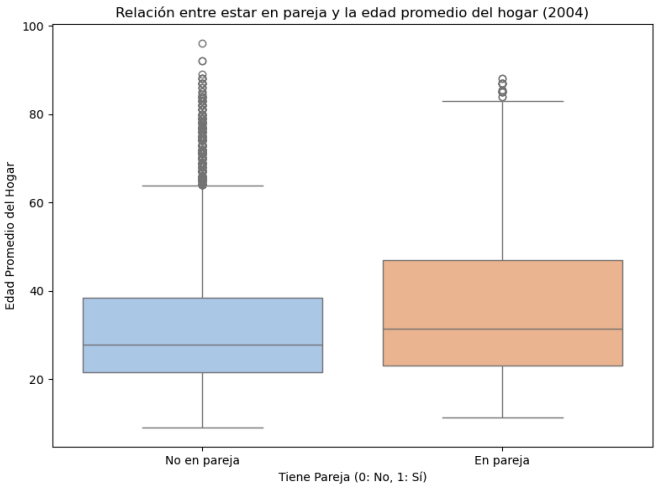
```
Estadísticas descriptivas (2004):
       EDAD_PROMEDIO_HOGAR  PROP_EDUC_SUPERIOR  TIENE_PAREJA
count          7647.000000         7647.000000   7647.000000
mean             34.301643            0.281548      0.417549
std              16.813181            0.298382      0.493187
min               9.000000            0.000000      0.000000
25%              22.142857            0.000000      0.000000
50%              29.250000            0.250000      0.000000
75%              42.000000            0.500000      1.000000
max              96.000000            1.000000      1.000000

Estadísticas descriptivas (2024):
       EDAD_PROMEDIO_HOGAR  PROP_EDUC_SUPERIOR  ESTADO_CIVIL_NUM  TIENE_PAREJA
count          7051.000000         7051.000000       7038.000000   7051.000000
mean             38.187693            0.342079          3.433220      0.397674
std              17.297581            0.345266          1.654983      0.489452
min              10.000000            0.000000          1.000000      0.000000
25%              25.250000            0.000000          2.000000      0.000000
50%              33.333333            0.250000          4.000000      0.000000
75%              47.000000            0.500000          5.000000      1.000000
max              97.000000            1.000000          5.000000      1.000000
```

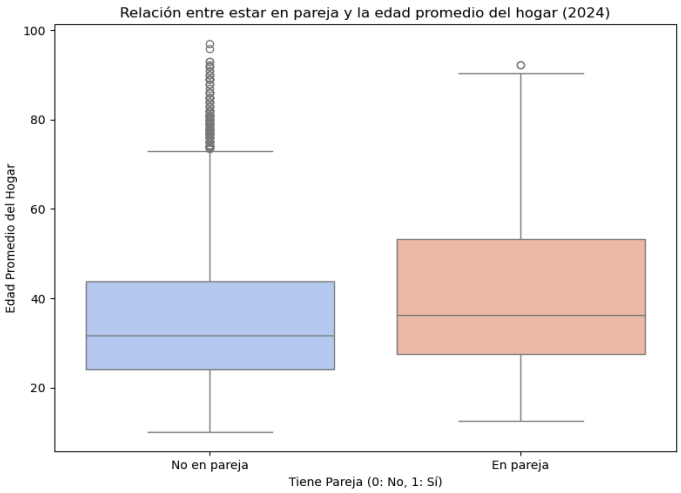
Estas estadísticas descriptivas permiten ver medias generales de las variables que realizamos previamente. Si bien los valores se mantienen bastante estables, puede ser interesante en casos donde se quieran estudiar diferencias entre años o cambios productos de algún fenómeno o suceso social que pueda afectar lo que estamos estudiando.



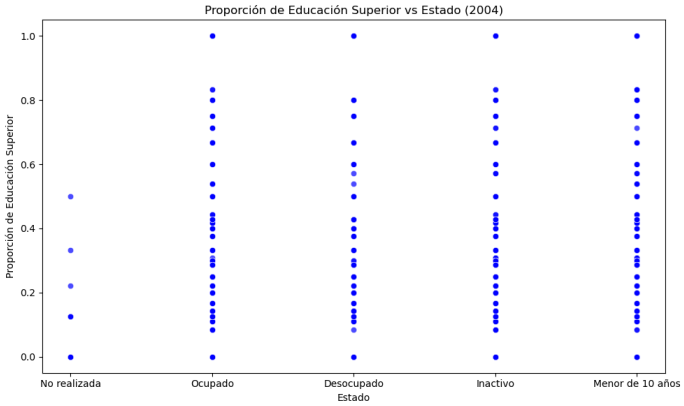


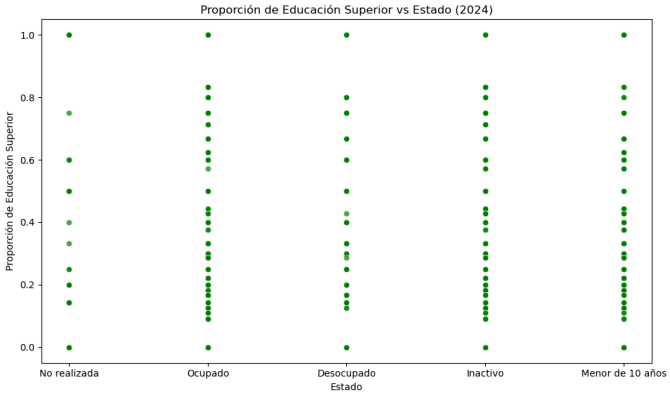
Se utilizaron histogramas para observar las distribuciones de los diferentes estados civiles que se presentan en los datasets de ambos años, buscando captar alguna diferencia que pueda darnos a conocer alguna inferencia interesante sobre las distribuciones con las que contamos y los cambios anuales observados. Al igual que en este caso también graficamos las distribuciones de aquellos que están en pareja y los que no.





El boxplot realizado para comparar la distribución de la edad promedio del hogar y el estar en pareja o no permitió observar patrones de que aquellos que se encuentran en pareja tienen un promedio de edad del hogar mayor a aquellos que no. Esto se relaciona con la desocupación en la medida en que es posible que aquellos en pareja, al tener un promedio de edad más alto, cuenten con mayores probabilidades de ocupación dentro del hogar, dado que es poco razonable mantener un hogar en pareja cuando ninguno de los dos se encuentra ejerciendo una ocupación.



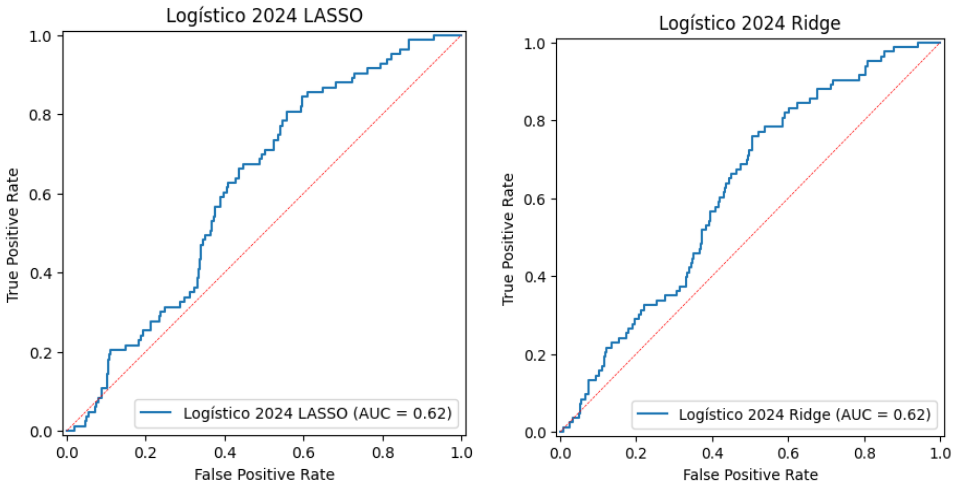


Por último, resultan interesantes los resultados de este scatter plot entre Prop. de educación superior y Estado ya que respaldan la idea de que los hogares con bajas proporciones de personas con educación superior tienden a presentar mayores niveles de desocupación o inactividad. Es razonable suponer que los hogares con una alta proporción de personas con educación superior completa tienen menores probabilidades de encontrarse en condición de desocupación, dado su mayor acceso a oportunidades laborales. Además, se plantea la hipótesis de que la ocupación puede estar influida por dinámicas de red dentro del hogar: un miembro ocupado podría facilitar la inserción laboral de otros, ya sea mediante recomendaciones o contactos, lo que refuerza la idea de que la ocupación es un fenómeno "contagioso". Es una posible buena observación filtrar del análisis a los menores de 10 años para ver este índice más limpio. Incluso filtrar a los menores de 18 sería interesante.

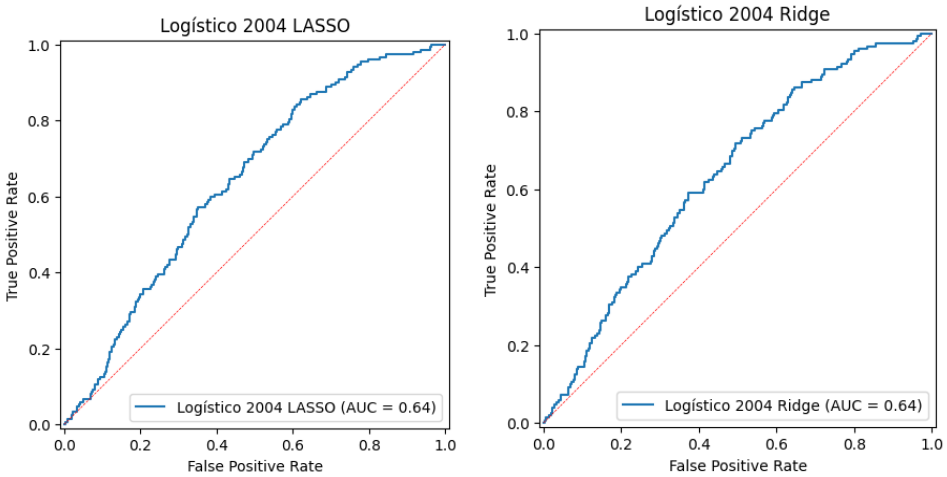
Analisis estadístico

Inicialmente se realizaron dos modelos logísticos diferentes por año, uno con una restricción de Ridge y otra de Lasso, ambas utilizando un alpha de 1. Para el modelo se utilizaron las variables planteadas al inicio del informe y las variables utilizadas durante el trabajo anterior ('CH04', 'CH06', 'CH07', 'CH08', 'NIVEL_ED'), además se incluyó las variables creadas que se consideraban relevantes para el estudio ('EDAD_PROMEDIO_HOGAR','PROP_EDUC_SUPERIOR','TIENE_PAREJA'). Luego de la

creación de los modelo se analizó su eficiencia en comparación a los planteados en el trabajo anterior mediante la curva ROC, el AUC y el accuracy con los siguiente resultados



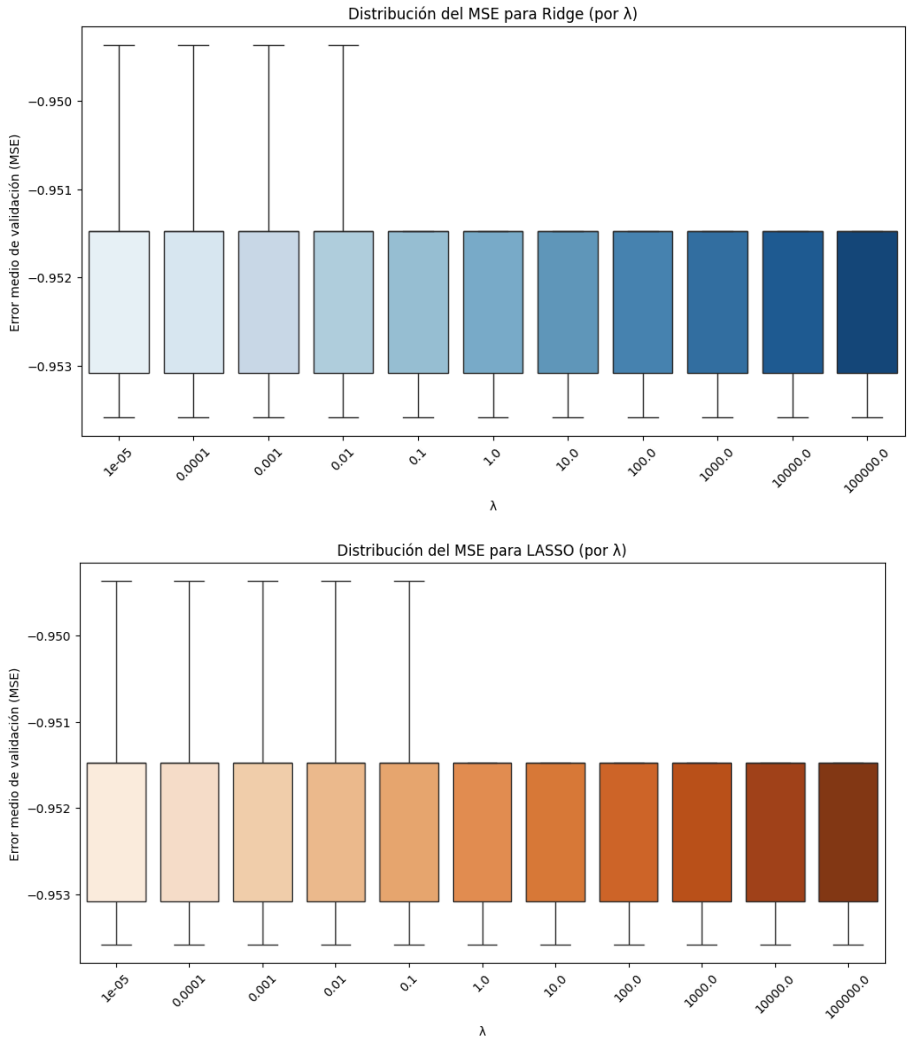
(*Modelos logísticos para los datos de 2024*)



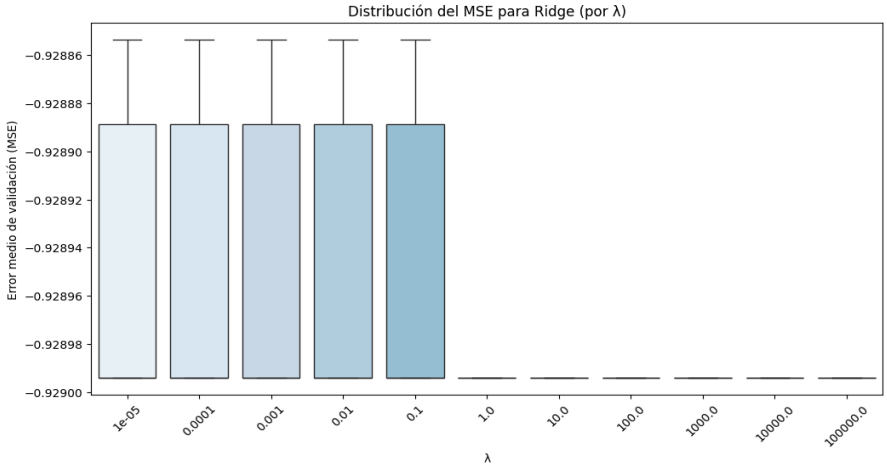
(*Modelos logísticos para los datos de 2004*)

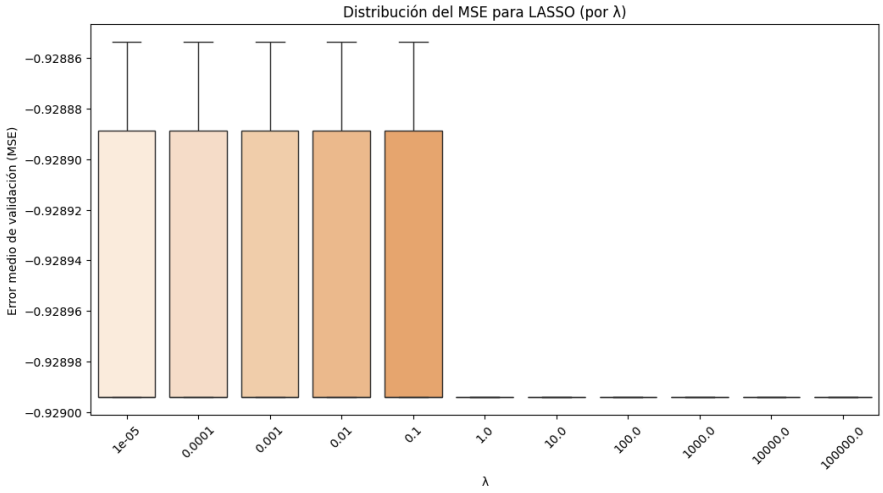
Los cuatro modelos presentaron un accuracy score de 0.96 y en comparación a los modelos del trabajo anterior el modelo logístico de 2024 resultó mejor que los modelos penalizados teniendo un AUC de 0.72 en comparación a los penalizados y teniendo un accuracy score de 0.95, 0.01 menos que los modelos penalizados. Por otro lado, el modelo de 2004 sin penalizaciones resultó inferior debido a que el AUC es igual que en ambos modelos sin embargo el accuracy es inferior teniendo un accuracy score de 0.93. Debido a errores a la hora de realizar el modelo original (no hubo creación de variables dummy en el trabajo anterior) su efectividad no se puede validar y por ende no se deberían comparar con los modelos resultantes de este estudio.

Posteriormente se volvieron a realizar los modelos utilizando un kfold cross validation de 10 con lambda elevado por una lista de valores de -5 a 5. Se presentó posteriormente cuáles fueron los mejores lambda para cada modelo, un boxplot que presenta los errores cuadráticos medios de cada lambda, y un análisis de los modelos Lasso de como fueron penalizando las variables conforme avanzó el proceso mediante un line plot. Para todos los modelos salvo el ridge de 2024 resultó mejor el lambda de 1 mientras que para el ridge de 2024 resultó el más óptimo el de 10

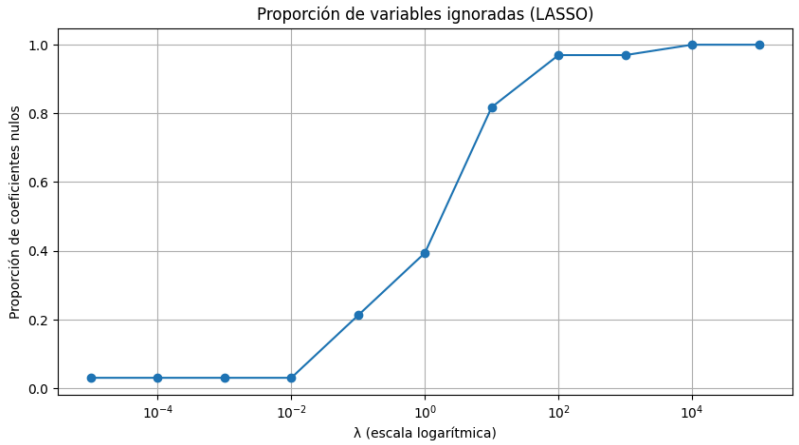


(*Distribución de MSE para los modelos de 2024)*

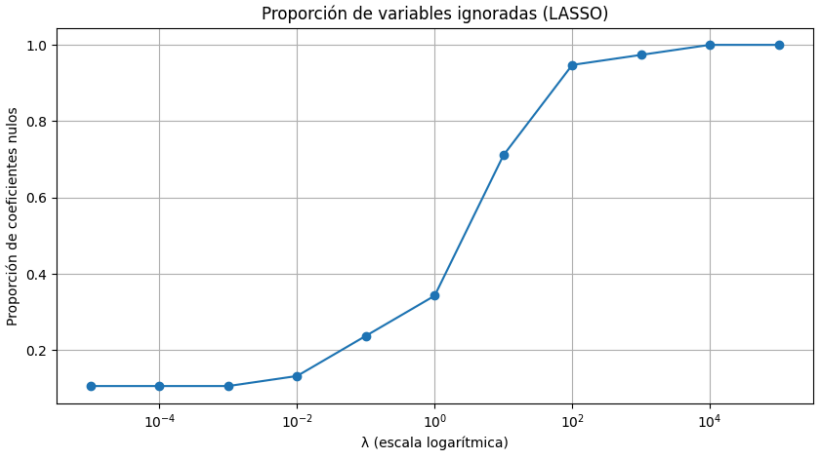




(*Distribución de MSE para los modelos de 2004)*



(*Proporción de coeficientes nulos (variables eliminadas) en el modelo de 2024)*



(*Proporción de coeficientes nulos (variables eliminadas) en el modelo de 2004)*

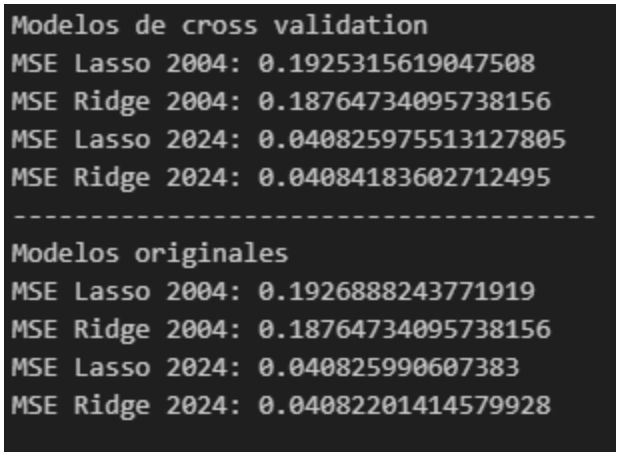
Cuando se observó que variables habían eliminado los modelos resultó evidente que ambos eliminaron las mismas variables, de la base de entrevista de individuo se eliminó partes de CH08 y CH07 (estado de seguro médico y estado civil) y del hogar se eliminó partes de IV1, 3 y 6 (que tipo de hogar, el tipo de piso que tiene y el acceso al agua).

	Variable	Coeficiente
8	CH07_3.0	0.0
11	CH08_2.0	0.0
12	CH08_3.0	0.0
14	CH08_12.0	0.0
15	CH08_13.0	0.0
16	CH08_23.0	0.0
24	IV1_3	0.0
25	IV1_4	0.0
26	IV1_5	0.0
27	IV1_6	0.0
29	IV3_3	0.0
30	IV3_4	0.0
32	IV6_3	0.0

	Variable	Coeficiente
8	CH07_Separado o divorciado	0.0
11	CH07_Ns./Nr.	0.0
15	CH08_Ns./Nr.	0.0
17	CH08_Obra social y planes y seguros públicos	0.0
18	CH08_Mutual/prepaga/servicio de emergencia/pla...	0.0
19	CH08_Obra social, mutual/prepaga/servicio de e...	0.0
28	IV1_Pieza en hotel/pensión	0.0
29	IV1_Local no construido para habitación	0.0
32	IV3_Cemento/ladrillo fijo	0.0
33	IV3_Ladrillo suelto/tierra	0.0
34	IV3_Otro	0.0
36	IV6_Fuera de la vivienda pero dentro del terreno	0.0
37	IV6_Fuera del terreno	0.0

(*Variables eliminadas en los modelos Lasso de 2024 [Derecha] y 2004 [izquierda]*)

Por último, se comparó el error cuadrático medio (MSE) de todos los modelos realizados en este estudio para observar cual era el mejor para cada año.



(*MSE* de cada uno de los modelos realizados en este estudio)

Como se puede observar, los modelos resultaron teniendo un MSE similar en todos los casos, esperable debido a que el cross validation realizado presentó que los mejores modelos resultaron de la penalización de Lambda=1 por lo que es esperable que los MSE sean similares salvo para el caso de ridge 2024 donde el óptimo resultó ser de lambda = 10.

Para 2004 el mejor modelo resultó ser el ridge si se analiza mediante el MSE debido a que tiene un error cuadrático medio (MSE) de 0.187 en contraste al Lasso de 0.192. Por otro lado, el mejor modelo de 2024 fue el Lasso que fue inducido a cross validation con un MSE de 0.040825.

Bibliografía

- *INDEC, 2024 Mercado de trabajo. Tasas e indicadores socioeconómicos (EPH), Segundo trimestre, Trabajo e Ingreso vol 8 N°7*