Cuarto Trabajo Práctico
Informe

Alumnas:
Angela Navajas McCormick
Legajo: 33542
Tiziana Sol Paciente
Legajo: 33377

Materia:
Ciencia de Datos
Lic. en Ciencias del Comportamiento

Docentes:
Maria Noelia Romero
Ignacio Spiousas

Fecha de entrega:
28 de Noviembre, 2024

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

PUNTO 1

Dentro de la base de hogar de EPH, las variables que pueden ser predictivas de la desocupación (sumadas a las variables utilizadas en el TP3 que se comparten en ambas bases: *ano4* (año de relevamiento), *ch03* (relacion de parentesco), *ch04* (sexo), *ch06* (años cumplidos), *ch07* (estado civil), *ch08* (cobertura médica), *nivel_ed* (nivel educativo), *estado* (condición de actividad), *cat_inac* (categoría de inactividad), *componente* (número de orden que se asigna a las personas que conforman cada hogar de la vivienda), *h15 (*entrevista individual realizada*)*, *mas_500 (*aglomerados segun tamaño*)*, *aglomerado* (codigo de aglometado) y e *ipcf* (monto de ingreso per cápita familiar percibido en el mes de referencia)) son: *V5* (si en los últimos 3 meses vivieron de subsidios o ayuda social), *V6* (si en los últimos 3 meses vivieron con mercaderia, ropa, alimentos del gobierno, iglesias, etc), *V7* (si en los últimos 3 meses vivieron con mercaderia, ropa, alimentos de personas externas al hogar), *V8* (si en los últimos 3 meses vivieron con un alquiler de su propiedad), *V11* (si en los últimos 3 meses vivieron con una beca de estudio), *V12* (si en los ultimos 3 meses vivieron con cuotas de alimentos o ayuda economica de personas externas al hogar), *V13* (si en los ultimos 3 meses vivieron gastando lo que tenian ahorrado), *V17* (si tuvieron que vender alguna pertenencia), *V19_a* (si en los ultimos 3 meses los menores de 10 años del hogar ayudaron económicamente trabajando), *V19_b* (si en los ultimos 3 meses los menores de 10 años del hogar ayudaron pidiendo recursos), *PP02H* (si en los ultimos 12 meses busco trabajo), *PP02I* (si en los ultimos 12 meses trabajo en algun momento), "*ix_tot*" (cantidad de miembros del hogar)

Estas variables mencionadas anteriormente pueden ser predictoras de la desocupación, ya que se relacionan con la capacidad económica del hogar, la relación de los individuos con el trabajo durante el último año y se conecta con las oportunidades y necesidades que posee cada hogar.

PUNTO 2

Para comenzar con el análisis de los datos y con la predicción, preparamos la base sobre la cual trabajamos a lo largo del estudio. Luego de descargar la base de hogar y la base de individuo tanto de 2004 como de 2024, filtramos las observaciones de interés (de GBA y CABA) y unimos las bases en un mismo dataframe tomando como referencia las variables "*codusu*" y "*nro_hogar*", al mergear las variables que compartían las dos bases se duplicaban aunque poseían los mismos valores, por lo que las eliminamos luego de procurar que los datos que tenían ambas columnas sean lo mismo.

PUNTO 3

En tanto a la limpieza de la base, comenzamos tratando las variables categóricas al modificar su valor de la palabra a la etiqueta que poseía esa respuesta en el diccionario de variables, con el fin de tener las variables con las que trabajaremos en un formato numérico.

Luego se observaron sí existían valores negativos en variables que no tendría sentido que los tuvieran, como por ejemplo *CH06* (edad) y *IPCF* (ingreso per cápita familiar), en la primera de estas se encontraron 51 valores negativos y estas observaciones fueron

eliminadas, pasando de 14698 datos a 14647 observaciones; en IPCF no se observaron valores negativos.

Como siguiente paso, se trataron los missing values. Observando patrones de respuesta en la base de datos, se encontraron 33 individuos que pusieron como respuesta "Ns./Nr." o "9" (etiqueta que representa "No sabe / No responde") en todas las variables de interés de la base de hogar ("*v5*", "*v6*", "*v7*", "*v8",* "*v11*", "*v12*", "*v13*", "*v17*", "*v19_a*", "*v19_b*"). Por esta razón, fueron renombrados como valores faltantes y eliminados de la base de datos que vamos a utilizar, dejándonos con 14614 observaciones. Por otro lado, se identificaron respuestas que no se entraban dentro de las opciones señaladas, esos datos también fueron eliminados.

Luego, se observaron valores outliers de *ipcf* tomando en cuenta el año en el que se hizo el relevamiento de datos, ya que el ingreso en 2004 no es comparable con el ingreso de 2024. Para establecer los límites que definen los outliers, calculamos el rango intercuartílico (IQR), Calculamos los percentiles 0.25 y 0.75, que representan los cuartiles superiores e inferiores de la variable de interés, así calculamos la diferencia entre estos (IQR) y capturamos la dispersión central de los datos. Utilizando el IQR, definimos como outliers los valores que se encuentran por debajo del límite inferior ($Q1-1.5\times IQR$) o por encima del límite superior ($Q3+1.5\times IQR$). De esta forma identificamos los datos extremos que podrían distorsionar los resultados de los análisis futuros. De esta variable se eliminaron 1033 datos atípicos, haciendo que la base de datos quede con 13581 observaciones.

Por último, se generó una nueva base "*base_final*" con las 27 variables que proponemos que son de interés ("*ano4*", "*codusu*", "*ix_tot*", "*ipcf*", "*ch03*", "*ch04*", "*ch06*", "*ch07*", "*ch08*", "*nivel_ed*", "*estado*", "*cat_inac*", "*componente*", "*h15*", "*mas_500*", "*aglomerado*", "*v5*", "*v6*", "*v7*", "*v8*", "*v11*", "*v12*", "*v13*", "*v17*", "*v19_a*", "*v19_b*", "*pp02h*") para la parte II de este trabajo y las columnas que tomaban valores binarios fueron transformadas a dummies, haciendo que la base de datos "*base_final_dummies*" termine con 67 columnas.

PUNTO 4

Con el objetivo de mejorar el modelo predictivo, se crearon cuatro variables a partir de relaciones entre variables ya existentes en la base de datos. En primer lugar, se construyó una variable que representa la proporción de trabajadores activos dentro de un hogar "*proporcion_trabajadores*", este resultado se obtuvo agrupando las respuestas por cada hogar y dividiendo la cantidad de trabajadores activos que hay por el total de personas que viven en el hogar.

Después, se calculó la proporción de ayudas externas en una variable llamada "*ingreso_externo*", que representa la suma de las variables utilizando las variables *v5_1*, *v6_1*, *v7_1*, *v12_1* (las cuales indican diferentes formas de ayudas externas en los hogares). Al calcular la proporción en función del número total de personas en el hogar, se logra una medida más precisa que permite comparar de manera justa la recepción de ayudas entre hogares de diferentes tamaños. Para construir esta variable se sumaron las respuestas de las variables agrupando por código de vivienda y después dividiendo por el número total de personas en el hogar.

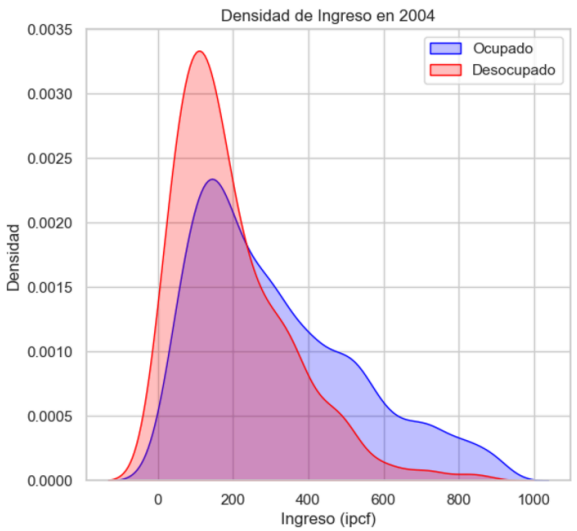
Para otra de las variables se propuso la variable de trabajo infantil en el hogar "*ingreso_infante*", para ello se tomaron las variables *v19_a_1* y *v19_b_1* (indican que menores de 10 años ayudaron economicamente trabajando y pidiendo), se sumaron estas variables y se la dividio por la cantidad de individuos en el hogar "*ix_tot*".

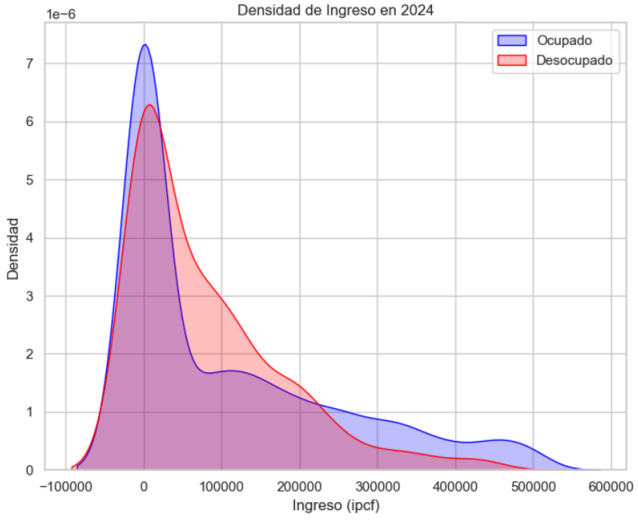
Por último, construimos fue la proporción de venta de objetos materiales y el gasto de ahorros en el hogar, llamada "*ingreso_ventas*". Se sumaron las variables *v13_1* y *v17_1* (indican que en los últimos 3 meses gastaron ahorros y vendieron pertenencias) y luego se lo dividió por *ix_tot* (la cantidad de miembros en el hogar).

PUNTO 5

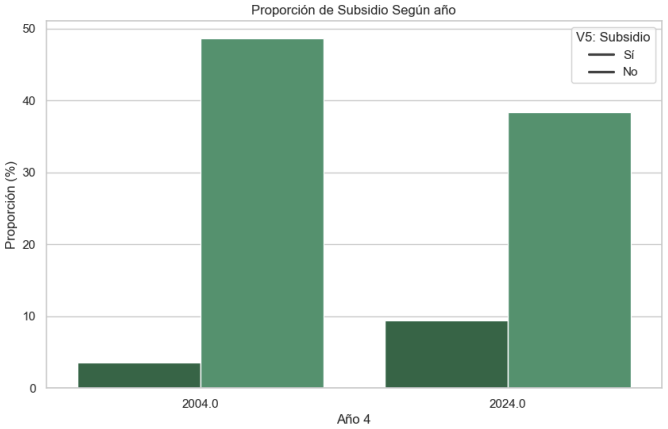
La media del ingreso por año fue: para 2004 $264,96 con un desvío estándar de $201.43, para 2024 $99542.24 con una desviación estándar de $126879.84. En tanto al porcentaje de ocupados, en 2004 fue de un 38.4% y en 2024 de 44%; el porcentaje de desocupados en 2004 fue de 7.2%, en cambio en 2024 disminuyó a 4.7%

Se realizó un gráfico de densidad del monto de ingreso per cápita familiar (*ipcf*), para las personas desocupadas y ocupadas, separando por 2004 y 2024. Para poder observar con más precisión la dispersión de los datos se realizó un recorte en el máximo de ingreso. Para 2004, como es de esperar, la distribución de los ingresos muestra como los ingresos de los desocupados se distribuyen a lo largo del rango mientras que los desocupados se concentran más cerca del cero. En 2024 se ve una mayor dispersión de los datos de los desocupados y hay una alta concentración de datos cerca del cero. Esto demuestra una mayor desigualdad en la distribución de los ingresos. Respecto a los ingresos de los desocupados en este año parecen mostrar una distribución similar a los ingresos de los de su mismo estado en 2004.

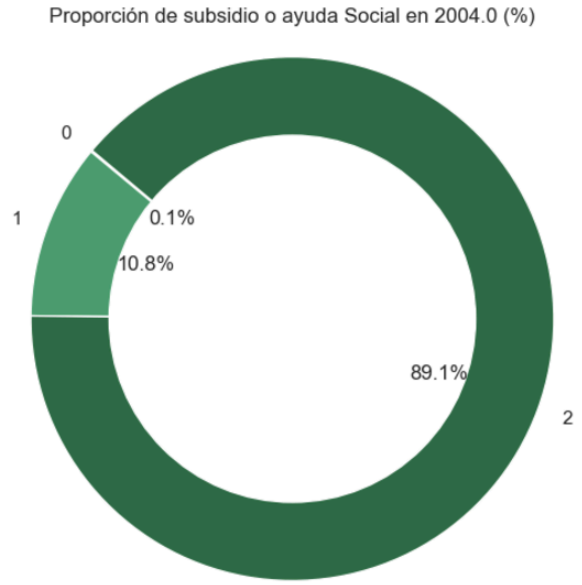


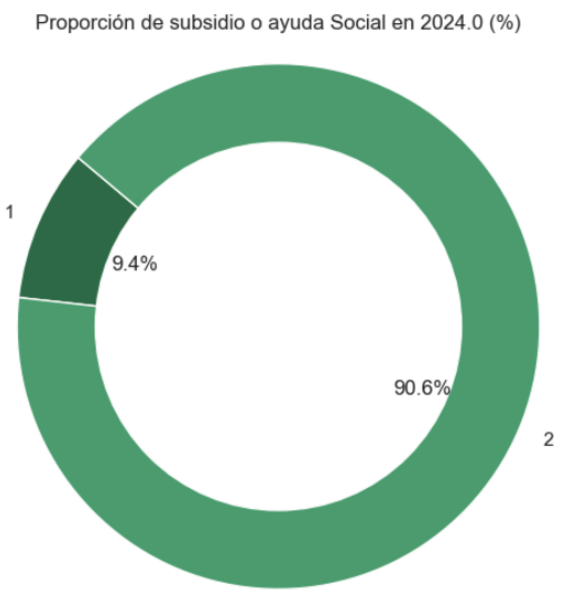


Por otro lado, se realizó un gráfico de barras sobre la proporción de subsidios que se dieron por año. Se observa que en ambos años la mayoría de los hogares no recibieron subsidios, pero hubo un crecimiento en la cantidad de subsidios dados entre 2004 y 2024.



Por último, se realizó un gráfico de torta que representa la proporción de hogares que recibieron ayuda en los últimos 3 meses con mercaderías, ropa, alimentos por parte del gobierno, iglesias, escuelas, entre otros. Se observa que de 2004 a 2024 disminuyó en 1 punto porcentual la cantidad de hogares que recibieron ayuda social.





Parte II: Categorización y regularización

PUNTO 1

Tras haber terminado con la limpieza de la base de datos, se separó la base en cuatro subconjuntos, dos de entrenamiento (70% de la base) y dos de prueba, uno de cada uno para cada año. Se utilizó una semilla (random state instance) de 101 y se usó la variable desocupado (*estado_1*) como variable dependiente. Como último paso dentro de este punto, se estandarizaron las variables para que los métodos de regularización que vamos a usar en los siguientes pasos no penalicen los coeficientes de las variables erróneamente y así no sesgar la selección y reducción de coeficientes.

PUNTO 2

Se utiliza validación cruzada (CV) para evaluar diferentes valores posibles de λ y así seleccionar aquel que minimice el error de validación. Para realizar este procedimiento se dividen los datos en k particiones, donde se usa una parte de los datos para entrenar y otra para validar. Para cada λ se ajusta el modelo k veces, en cada iteración se entrena con k -1 particiones y se evalúa con las particiones restantes, después se calcula el error en los datos de validación. Luego, se promedian los errores obtenidos en cada partición, obteniendo así el CV. Este proceso se repite para las diferentes opciones de λ y se selecciona aquel que minimice el error cuadrático medio (MSE) de CV. Para realizar este procedimiento se utiliza exclusivamente el conjunto de entrenamiento ya que, si se utiliza el conjunto de prueba, estaríamos sesgando el error estimado.

PUNTO 3

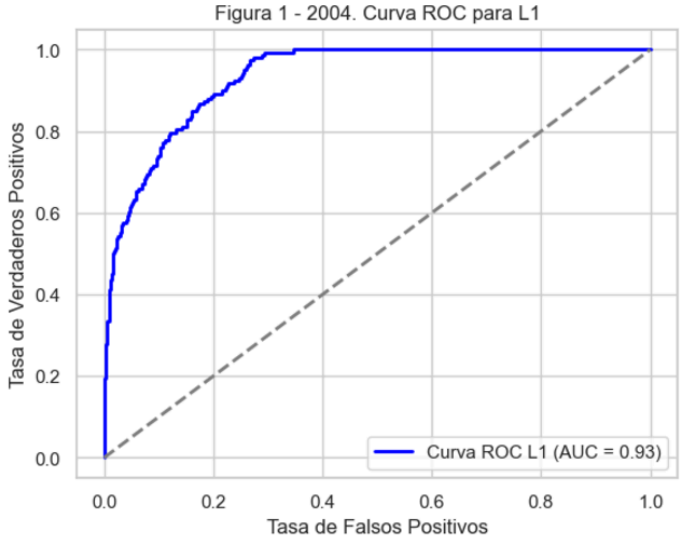
A la hora de elegir el tamaño de *k*, nos encontramos con un problema recurrente, un trade-off entre el sesgo y la varianza, un *k* muy grande genera una varianza demasiado grande en la estimación del error, mientras que un *k* muy chico elevaría el sesgo. Cuando se elige que $k = n$ (n representa el número de muestras), es decir al usar Leave-one-out CV, cada iteración utiliza n-1 observaciones para entrenar el el método de aprendizaje, que se repite *n* veces, una iteración para cada observación. De esta forma, como resultado se producirían *n* MSE.

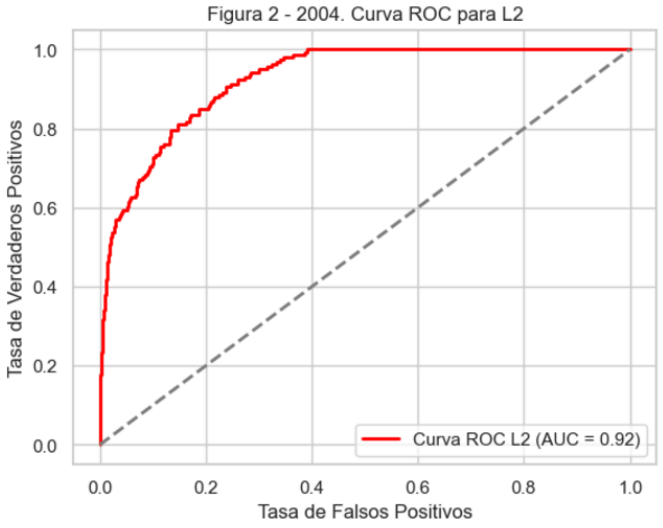
PUNTO 4

Para cada año, se realizaron regresiones logísticas con dos tipos de penalizaciones, Lasso (l1) y Ridge (l2) con $\lambda = 1$ (*model_l1_04*, *model_l2_04*, *model_l1_24*, *model_l2_24*). A continuación se va a reportar la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy para cada uno de los modelos.

'model_l1_04' tiene una matriz de confusión con 1951 verdaderos positivos, 16 falsos negativos, 105 falsos positivos y 53 verdaderos negativos. La precisión es de 0.943 y el AUC es de 0.934, indicando un buen desempeño en la clasificación y capacidad de discriminación. El buen desempeño que indica este último valor, se puede ver graficado en la figura 1

'model_l2_04' presenta una matriz de confusión con 1954 verdaderos positivos, 13 falsos negativos, 108 falsos positivos y 50 verdaderos negativos. La precisión es de 0.943 y el AUC es de 0.925, lo que refleja un buen desempeño en la clasificación y una capacidad de discriminación sólida, tal como se puede ver en la Figura 2 con la curva ROC.





'model_l1_24' (Figura 3) muestra una matriz de confusión con 1826 verdaderos positivos, 8 falsos negativos, 68 falsos positivos y 38 verdaderos negativos. La precisión es de 0.961 y el AUC es de 0.950, indicando un excelente rendimiento en la clasificación y una alta capacidad de discriminación.

'model_l2_24' (Figura 4) presenta una matriz de confusión con 1828 verdaderos positivos, 6 falsos negativos, 72 falsos positivos y 34 verdaderos negativos. La precisión es de 0.960 y el AUC es de 0.942, lo que indica un buen desempeño en la clasificación y una capacidad de discriminación sólida.

Figura 3 - 2024. Curva ROC para L1

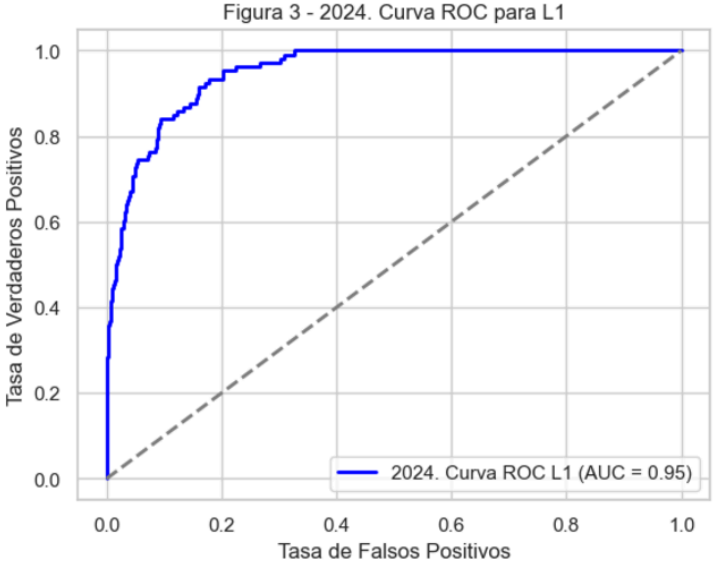
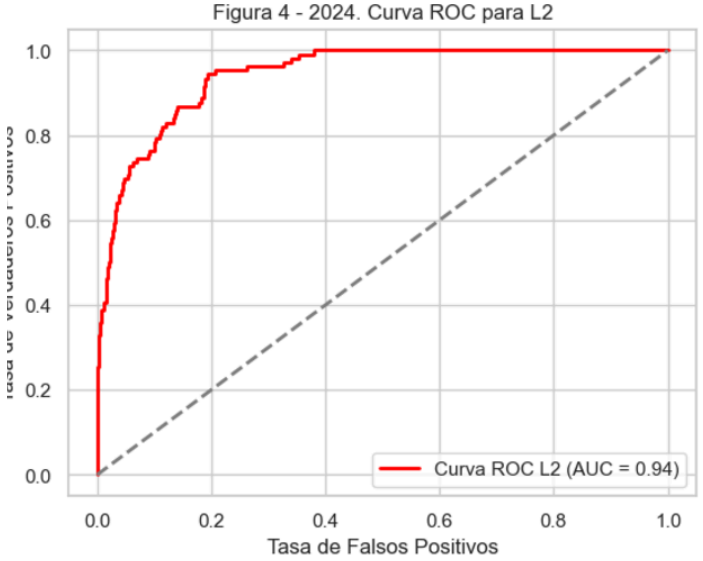


Figura 4 - 2024. Curva ROC para L2



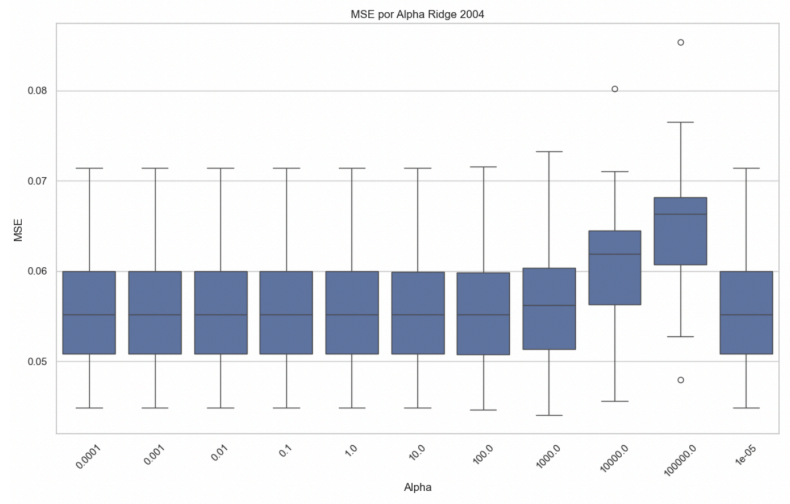
Cuando comparamos los resultados de las regresiones con los resultados de las que hicimos para el TP3, podemos ver algunos cambios. El AUC de 2004 y 2024 de las regresión fue de 0.8692 y 0.948645, respectivamente. Entonces podemos ver que si bien la regresión de 2004 si tuvo una mejora con respecto al AUC, en 2024 muestra valores similares. Si bien las diferencias no aparentan ser tan grandes, errores cometidos a lo largo del desarrollo del TP3 y que fueron corregidos para el TP actual, nos hacen sospechar que los resultados de este pueden no ser exactamente los que deberían. Por esta razón, podemos decir que la performance de regresión logística con regularización, tanto de Lasso como de Ridge, es mejor que sin.

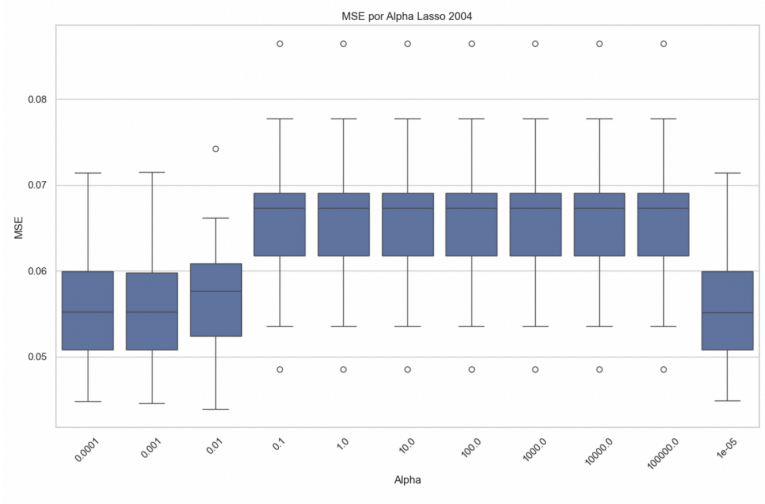
PUNTO 5

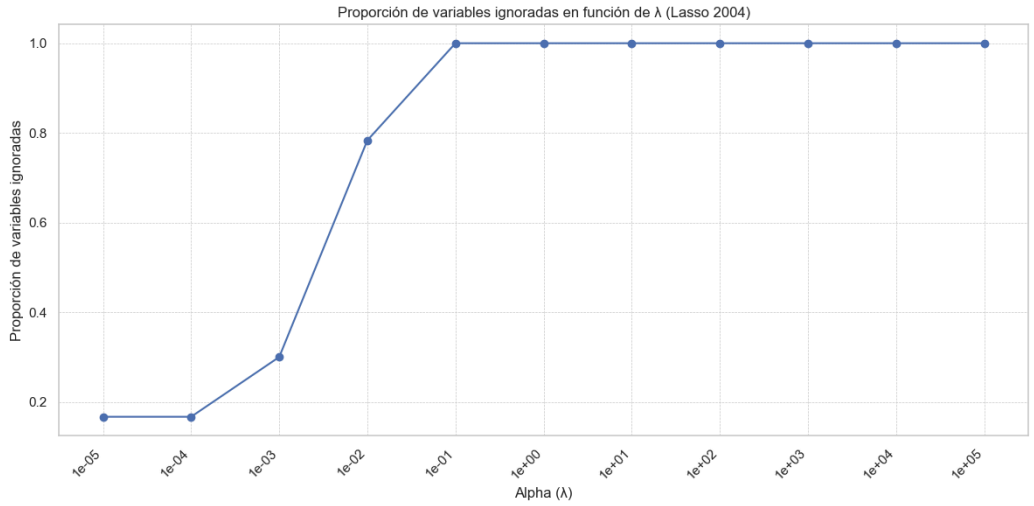
En este punto vamos a realizar una validación cruzada para elegir el lambda óptimo para Lasso y Ridge, se graficaron los MSE para cada partición y se analizará el resultado de la regularización en los coeficientes. Para comenzar con la validación cruzada, tanto de Lasso como de Rindge, se generó un parámetro (α) con rango de valores utilizando un $n \in \{-5, -4, -3 \ldots, +4, +5\}$ utilizando la fórmula 10^n.

Comenzando por RidgeCV para 2004, se obtuvo como mejor alpha 100 y el error cuadrático fue de 0.05834. Con respecto a LassoCV para el mismo año se obtuvo un alpha de 0.001 y el error cuadrático de este fue 0.058251. Se graficaron en box plots los errores de cada lambda para ambos modelos, lo que se puede observar en los siguientes gráficos. Los valores del MSE para Ridge se mantienen relativamente estables para los diferentes valores de α, con ligeras variaciones hacia valores más altos conforme aumenta α. Podemos ver en los box plots baja variabilidad en el MSE entre las diferentes ejecuciones del modelo. Sin embargo, para valores de α 10.000 y 100.000 la dispersión aumenta notablemente, y hay valores atípicos hacia el extremo superior, lo que sugiere menor consistencia en el rendimiento del modelo. En el box plot de lasso podemos ver que si bien en 0.0001 y 0.001 los valores son bajos, a partir de 0.01 empieza a aumentar el rango de valores reflejando una mayor variabilidad, aparte empiezan a aparecer outliers. Para Lasso se graficó la proporción

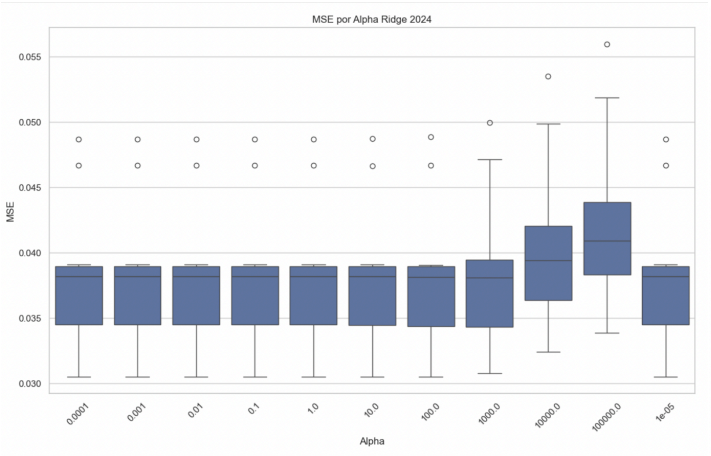
de variables ignoradas para cada lambda, en el gráfico podemos observar que a partir de 0.1 el modelo ya elimina todas las variables por lo que estaría ya penalizando demasiado.

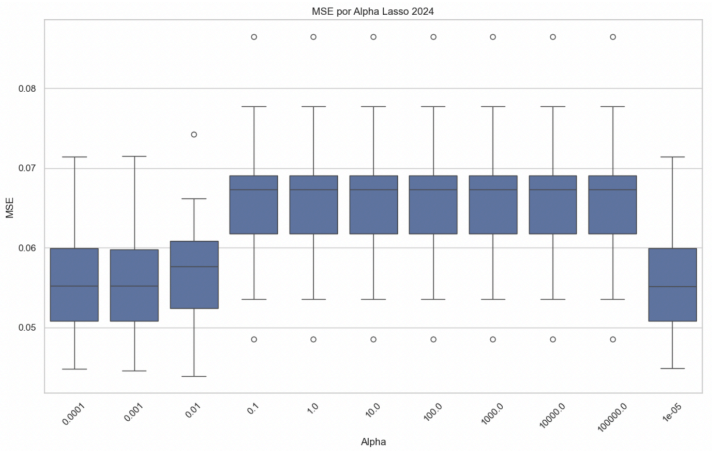


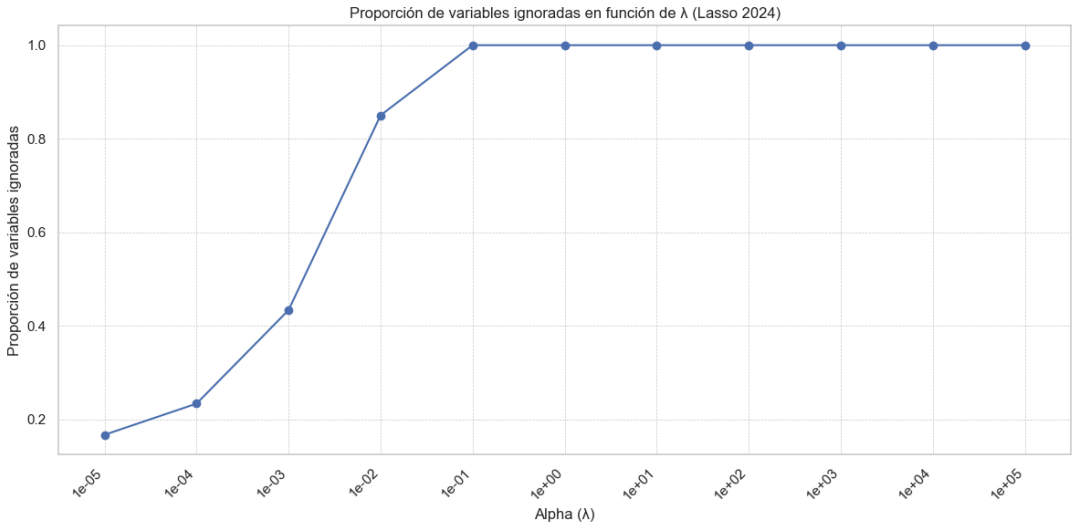




Para el año 2024, se obtuvo un mejor alpha de 100.0 para el modelo Ridge CV, con un error cuadrático medio de 0.0446. En cuanto al modelo Lasso CV, el mejor alpha fue de 0.001, y su error cuadrático medio fue de 0.0462. De la misma forma que antes, los errores de cada lambda para ambos modelos se graficaron en box plots. Para Ridge, podemos observar que a partir de 1.000 el MSE aumenta y muestra también una mayor dispersión. Para Lasso en cambio podemos ver en general una mayor dispersión de los errores a lo largo de todos los alphas, para 0.0001 y 0.001 el modelo se mantiene bajo y a partir de 0.1 comienza a aumentar ligeramente. De la misma forma que para 2024, se gráfico la proporción de variables ignoradas y se obtuvo a medida de que el alpha aumenta, como es de esperar, una mayor cantidad de variables son eliminadas y que a partir de 0.1 el modelo elimina el 100% de ellas.







PUNTO 6

En Lasso, el valor de λ óptimo para tanto para 2004 como para 2023 fue de λ = 0.001. Lasso selecciona automáticamente variables que son relevantes para el modelo y lo demuestra con el coeficiente de la variable predictora igual a 0. En el modelo realizado para 2004, se eliminaron 17 variables de las cuales algunas esperábamos que no fueran significativas para la predicción de desocupación como por ejemplo *ch03* (describe la relación de parentesco de los integrantes del hogar), *ch07* (indica estado civil) pero otras no eran esperadas a ser eliminadas como por ejemplo *v7* (si vivieron en los últimos 3 meses con mercaderías, ropa, alimentos de familiares, vecinos u otras personas externas al hogar), *v13* (si en los últimos 3 meses gato lo que tiene ahorrado), *v19_b* (si en los últimos 3 meses los menores de 10 años ayudaron al hogar pidiendo).

En tanto al análisis de 2024, se eliminaron 25 variables. Algunos ejemplos de variables que no se esperaban que sean eliminadas son: algunos valores de respuesta de *ch08* (tipo de cobertura médica) como *ch08_2* (Mutual / prepaga / servicio de emergencia), *ch08_3* (planes y seguros públicos); *v19_a* y *v19_b* que indican trabajo de menores en el hogar con el fin de ayudar económicamente al hogar, entre otros.

La variables que agregamos como de interés en el inciso 1 del punto 1 como posibles predictoras coinciden con lo que obtuvimos con Lasso de variables relevantes que estaban en

la base de hogar, *v7* fue eliminada en la predicción de ambos años, para 2004 se eliminaron *v13* y *v19_b*; en cambio para 2024 se eliminaron *v8*, *v11*, *v12*, *v17, v19_a* y *v19_b*. El resto de variables eliminadas eran variables en común entre la base de individuo y la base de hogar.

PUNTO 7

Para determinar qué método de regularización de Lasso o Ridge funcionó mejor, esto se puede determinar observando el error cuadrático medio (MSE) del alpha elegido por Cross Validation de cada método para cada año. Un valor de MSE más bajo implica que el modelo tiene un mejor desempeño en el conjunto de prueba.

Por un lado, en el año 2004 tuvo mejor desempeño Ridge con un MSE = 0.0583, Lasso obtuvo un MSE = 0.0582. En tanto al año 2024, Ridge también tuvo un mejor desempeño con un MSE = 0.0446, en cambio en Lasso se obtuvo un MSE = 0.0445.

Al comparar los resultados de cada uno de los modelos entre años, se observa que en Ridge, el modelo del año 2004 predijo mejor que en el año 2024; de la misma manera, en Lasso se predijo mejor en 2004 que en 2024.

Lasso realizó una selección distinta de predictores entre 2004 y 2024, en el primero de estos años seleccionó 42 variables, en cambio en 2024 se seleccionaron 34 características. Dado que posee menos variables, esta puede ser la razón por la que quizás el MSE es mayor.