TRABAJO PRÁCTICO N°4 - GRUPO N°2
CLASIFICACIÓN Y REGULARIZACIÓN DE DESOCUPACIÓN USANDO LA EPH

Big Data y Aprendizaje Automático para Economistas
Guadalupe Gorostiaga y Valentina Fuentes Mortensen

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

El objetivo de esta sección es analizar las tasas de desempleo a nivel de hogares, evaluando cómo las características de los hogares influyen en la desocupación.

1. Del diseño de registro de la base de hogares, consideramos que las siguientes variables pueden ser predictoras de la desocupación. A continuación, se mencionan y justifican las razones para su inclusión:
   - IV6: Indica de dónde accede el hogar al servicio de agua (dentro de la vivienda, fuera del terreno, pero dentro del lote, o fuera del lote). Justificación: un acceso más complicado al agua refleja condiciones de vida más desfavorables, lo que puede estar vinculado con una mayor probabilidad de desempleo.
   - IV8: Indica si el hogar tiene baño o letrina. Justificación: la ausencia de baño puede ser un indicador de precariedad habitacional, lo que podría incrementar las posibilidades de estar desempleado.
   - IV9: Indica dónde está el baño o letrina (dentro de la vivienda, fuera del terreno pero dentro del lote, o fuera del lote). Justificación: además de tener baño, la ubicación también es importante. Un baño fuera de la vivienda refleja mayores dificultades para cubrir esta necesidad, lo que puede estar relacionado con una mayor probabilidad de desempleo, similar al acceso al agua.
   - IV12_1: Indica si la vivienda está cerca de basurales. Justificación: la proximidad a basurales u otras zonas vulnerables podría indicar condiciones de vida desfavorables, lo que a su vez aumenta las chances de estar desempleado.
   - IV12_3: Indica si la vivienda está ubicada en una villa de emergencia. Justificación: Ídem IV12_1.
   - V14: Indica si en los últimos tres meses los miembros del hogar han solicitado préstamos a familiares o amigos. Justificación: la necesidad de pedir préstamos a redes informales puede ser un indicador de carencias económicas, lo que podría estar relacionado con una mayor probabilidad de estar desempleado.
   - V15: Indica si en los últimos tres meses los miembros del hogar han solicitado préstamos a bancos, financieras, etc. Justificación: el hecho de recurrir a préstamos formales puede ser un indicio de que la persona percibe ingresos y, por lo tanto, disminuye la probabilidad de desempleo.
   - ITF: Indica el monto del ingreso total familiar percibido en el mes de referencia. Justificación: ingresos más bajos, cercanos o por debajo de la línea de pobreza, pueden reflejar una situación económica vulnerable, incrementando la probabilidad de desempleo.

- IPCF: Indica el monto del ingreso per cápita familiar percibido en el mes de referencia. Justificación: Ídem ITF.

2. Para este trabajo, analizaremos las bases de datos de hogares e individuos de la EPH correspondientes al primer trimestre de 2004 y 2024. Al unificar estas cuatro bases en una sola y seleccionar únicamente el aglomerado de Gran Tucumán - Tafí Viejo, se obtiene una base de datos con 4656 filas y 248 columnas.

3. Para la limpieza de datos, se seleccionó un subconjunto de variables de interés, algunas de las cuales fueron mencionadas el el primer punto, mientras que otras son propias de la base de individuos. Se detectaron personas con edades negativas, por lo que estas observaciones sin sentido fueron eliminadas. También, se eliminaron outliers en las variables de ingreso total familiar e ingreso per cápita familiar, aplicando el criterio de 2 desviaciones estándar respecto a la media. No se encontraron missing values en las variables consideradas, por lo que no fue necesaria ninguna imputación. Tras la limpieza, la base quedó con 4436 filas y la misma cantidad de columnas.

4. Dado que el objetivo de este trabajo es predecir la desocupación, hemos construido cuatro variables a partir de la información disponible, las cuales consideramos pueden ser relevantes para dicha predicción:

   A. Proporción de personas ocupadas en el hogar: se define como la cantidad de personas ocupadas en el hogar dividida por el total de miembros. En promedio, el 37% de los miembros del hogar están ocupados. Un valor relativamente bajo en esta variable sugiere que hay un número considerable de personas en situación de desocupación en los hogares analizados.
   B. Nivel educativo alcanzado: creamos tres variables dummies que indican si el nivel educativo del individuo es bajo (sin instrucción, primaria incompleta o no sabe/no responde), medio (primaria completa o secundaria incompleta) o alto (secundaria completa, educación superior universitaria incompleta o completa), según corresponda. El 28% de los individuos en la muestra tiene un nivel educativo bajo, lo que sugiere que una parte significativa de la población tiene un acceso limitado a la educación, mientras que un 34% de los individuos tiene un nivel educativo alto.
   C. Hacinamiento: calculada como la cantidad de miembros del hogar sobre la cantidad de ambientes o habitaciones de uso exclusivo. Para una correcta construcción de esta variable, se eliminaron aquellas observaciones en las que no se declaró ningún ambiente de uso exclusivo en el hogar. En promedio, hay 1,73 miembros del hogar por cada ambiente o habitación de uso exclusivo. Un valor superior a 3 sugiere un nivel significativo de hacinamiento, lo que puede implicar condiciones de vida precarias.
   D. Índice de carga familiar: se define como el ratio entre la cantidad de personas dependientes en el hogar (incluye menores de 6 años, personas con discapacidad y adultos mayores jubilados) y el total de miembros del hogar. En promedio, el 16% de los miembros del hogar son personas dependientes. Un índice de carga familiar relativamente bajo puede indicar una dependencia moderada a baja, lo que podría facilitar la capacidad económica del hogar.

Tabla 1

	A	B_bajo	B_medio	B_alto	C	D
Observaciones	4422	4422	4422	4422	4422	4422
Media	0,37	0,28	0,39	0,34	1,73	0,16
Desvío Estándar	0,25	0,45	0,49	0,47	1,14	0,20
Mínimo	0	0	0	0	0,11	0
Máximo	1	1	1	1	9	1
Missing values	0	0	0	0	0	0

5. Las cinco variables de la encuesta de hogares que consideramos que pueden ser relevantes para predecir desocupación son: IV6 (acceso al agua), IV8 (tenencia de baño o letrina), ITF (ingreso total familiar), V14 (préstamos familiares) y V15 (préstamos de instituciones). El 97,85% de los hogares tiene baño o letrina, y el 13,55% accede al agua desde fuera de la vivienda, ya sea dentro o fuera del terreno. En promedio, los hogares se endeudan más con préstamos familiares que con entidades financieras. El ingreso promedio de los hogares es de $189.966,66, aunque es importante señalar que este valor no fue calculado por año.

Tabla 2

	acceso_agua	tenencia_baño	ITF	prestamos_ flia	prestamos_ instituciones
Observaciones	4422	4422	4422	4422	4422
Media	1,14	1,02	189.966,66	1,76	1,89
Desv. Est.	0,36	0,15	315.672,79	0,42	0,32
Mínimo	1	0	0	1	1
Máximo	3	2	2.360.000	2	2

6. La tasa de desocupación para el primer trimestre de 2024, calculada utilizando el ponderador de la EPH (PONDERA), es igual a 8,84%, mientras que la reportada por el INDEC para el mismo período fue del 7,7%. Esta diferencia puede deberse a que el INDEC calcula su medida a partir de la base de individuos, mientras que en este ejercicio se consideró un individuo aleatorio para tener una observación por hogar, siempre que pertenezca a la PEA.

Parte II: Clasificación y regularización

El objetivo de esta sección es predecir si una persona está desocupada o no, utilizando datos a nivel de individuo y de hogar.

1. Nuestra variable de interés a clasificar es si la persona está desocupada o no. El set de variables explicativas que tomamos está compuesto por:

    - Proporción de personas ocupadas en el hogar
    - Nivel educativo (bajo, medio o alto)
    - La densidad de personas por habitación en el hogar (hacinamiento)
    - La proporción de personas dependientes en el hogar (niños, adultos mayores y discapacitados)
    - Sexo
    - Edad
    - Si tiene acceso al agua dentro de la vivienda
    - Si tiene baño
    - Si el baño está dentro de la vivienda
    - Ingreso Total Familiar
    - Cobertura médica
    - Si vive cerca de un basural
    - Si vive en una villa de emergencia
    - Si pidió préstamos a familia o amigos
    - Si pidió préstamos a instituciones

2. La validación cruzada (cross-validation) es una técnica utilizada para evaluar la capacidad de generalización de un modelo. Para elegir el parámetro de regularización λ (lambda) en modelos como Ridge y LASSO, seguiríamos estos pasos:
    1. **Dividir el conjunto de datos**: Se divide el conjunto de datos en k pliegues (folds) de igual tamaño. Esto significa que si tenemos nnn muestras, cada pliegue tendrá aproximadamente n/k muestras.
    2. **Entrenamiento y validación**: Para cada valor de λ en un rango predefinido, el modelo se entrena en $k-1$ pliegues y se valida en el pliegue restante. Este proceso se repite k veces, cada vez utilizando un pliegue diferente como conjunto de validación.
    3. **Promediar los resultados**: Se calcula la métrica de rendimiento (por ejemplo, error cuadrático medio o log-loss) para cada combinación de λ y pliegue. Luego, se promedia el rendimiento para cada λ.
    4. **Selección del mejor λ**: Finalmente, se selecciona el valor de λ que minimiza el error promedio. Este valor se considera el más adecuado para el modelo, ya que ha sido evaluado utilizando múltiples subconjuntos de datos.

No se utiliza el conjunto de prueba (test) porque su propósito es medir el rendimiento final del modelo. Si usamos el conjunto de prueba durante el proceso de selección de λ, corremos el riesgo de overfitting al

conjunto de prueba, lo que puede dar una falsa sensación de rendimiento. Esto significa que el modelo podría no generalizar bien a datos no vistos. Al mantener el conjunto de prueba separado, aseguramos que nuestra evaluación sea objetiva y refleje la capacidad del modelo para generalizar a nuevos datos.

3. **Si k muy pequeño**:

- **Sobreajuste**: Si k es muy pequeño (por ejemplo, k=2), cada conjunto de entrenamiento será muy similar al conjunto de prueba. Esto puede llevar a un sobreajuste, ya que el modelo se ajustará demasiado a los datos de entrenamiento y no generalizará bien.
- **Alta varianza**: Con pocos pliegues, la estimación de rendimiento puede ser inestable y tener alta varianza, ya que depende en gran medida de cómo se dividen los datos.

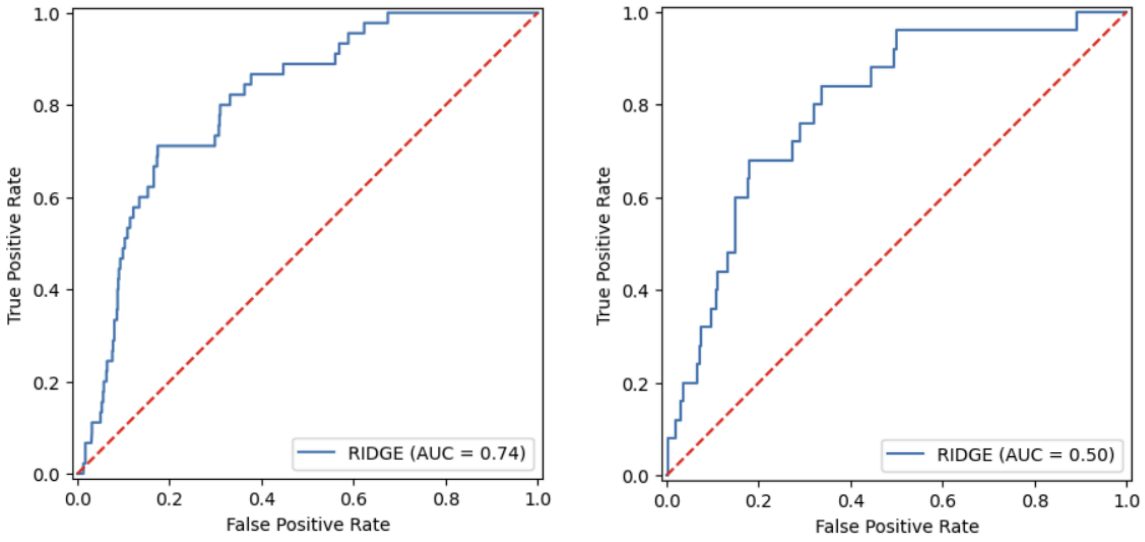
Si k muy grande:

- **Bajo sesgo**: Un k grande (cercano a n) significa que cada pliegue es pequeño, lo que resulta en conjuntos de entrenamiento que son casi completos. Esto generalmente produce estimaciones de rendimiento con bajo sesgo.
- **Costo computacional**: Sin embargo, el costo computacional aumenta, ya que el modelo se entrena casi n veces, una vez para cada muestra. Esto puede ser ineficiente, sobre todo para modelos costosos de entrenar.
- **k = n**: Cuando k=n, se trata de una validación cruzada de "dejar uno fuera" (Leave-One-Out Cross-Validation). En este caso, se estima el modelo n veces, lo que permite evaluar cada observación individualmente. Aunque puede proporcionar una evaluación precisa, puede ser muy costosa computacionalmente y no ser práctica para conjuntos de datos grandes.

4.

Gráfico 1

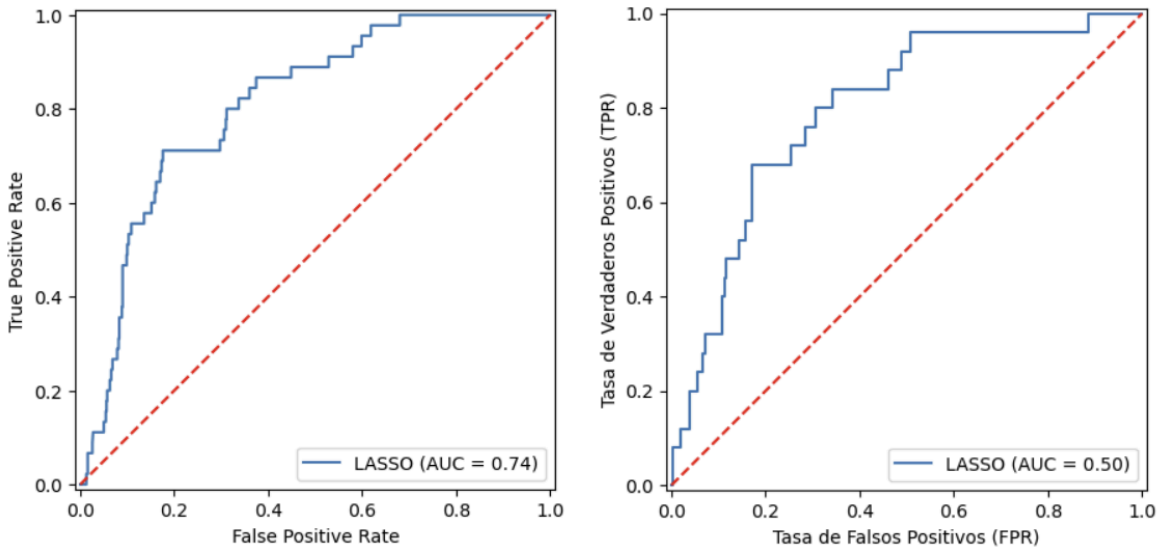
Curvas ROC para RIDGE



a.2004 b. 2024

Gráfico 2

Curvas ROC para LASSO



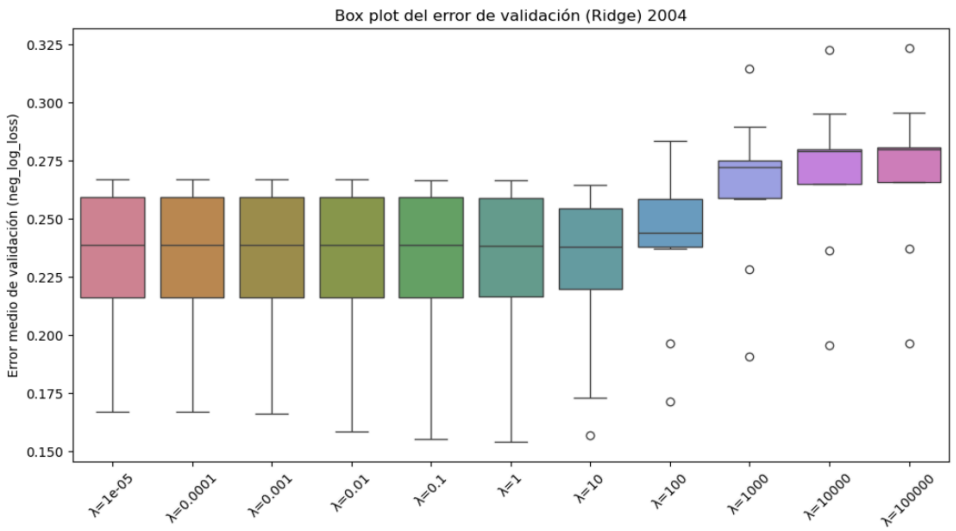
a.2004 b. 2024

Siguiendo la medida de AUC, estas regresiones no se desempeñan mejor que la regresión logística presentada en el Trabajo Práctico 3.

Las regresiones de LASSO y RIDGE se desempeñan peor que la regresión logística en la medida de accuracy para 2004, pero son prácticamente iguales en esta dimensión para las estimaciones de 2024.

5. Los parámetros de penalización seleccionados fueron:
   - λ óptimo para Ridge 2004: 10
   - λ óptimo para LASSO 2004: 1
   - λ óptimo para Ridge 2024: 10
   - λ óptimo para LASSO 2024: 1

Gráfico 3



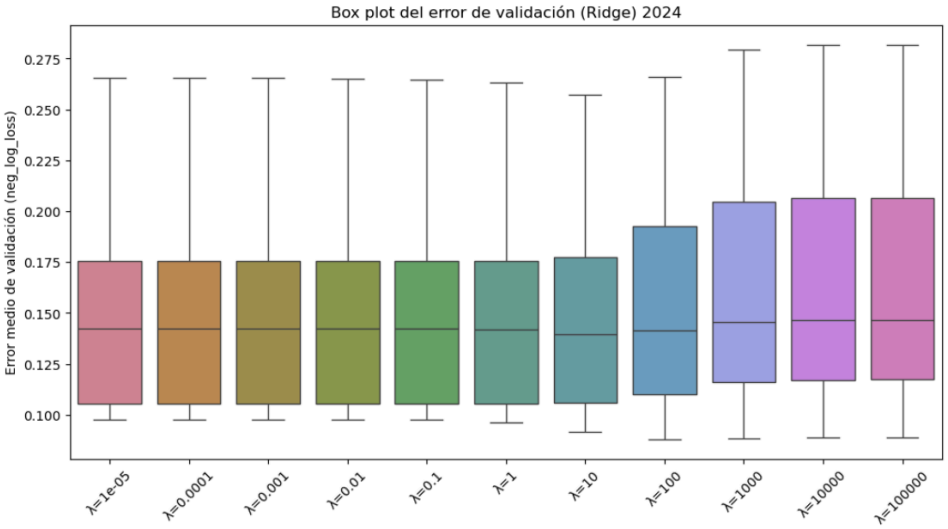
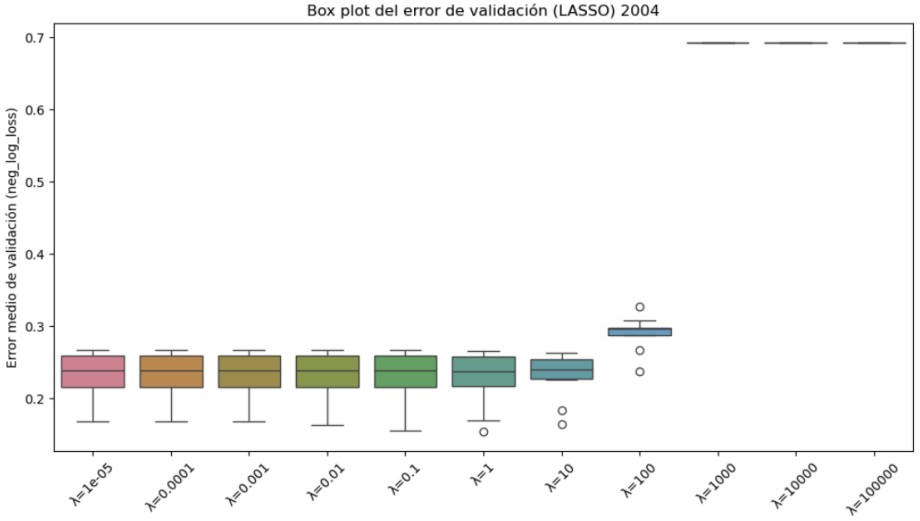

Box plot del error de validación (Ridge) 2024
Error medio de validación (neg_log_loss)
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
λ=1e-05
λ=0.0001
λ=0.001
λ=0.01
λ=0.1
λ=1
λ=10
λ=100
λ=1000
λ=10000
λ=100000

Gráfico 4


Box plot del error de validación (LASSO) 2004
Error medio de validación (neg_log_loss)
0.7
0.6
0.5
0.4
0.3
0.2
λ=1e-05
λ=0.0001
λ=0.001
λ=0.01
λ=0.1
λ=1
λ=10
λ=100
λ=1000
λ=10000
λ=100000

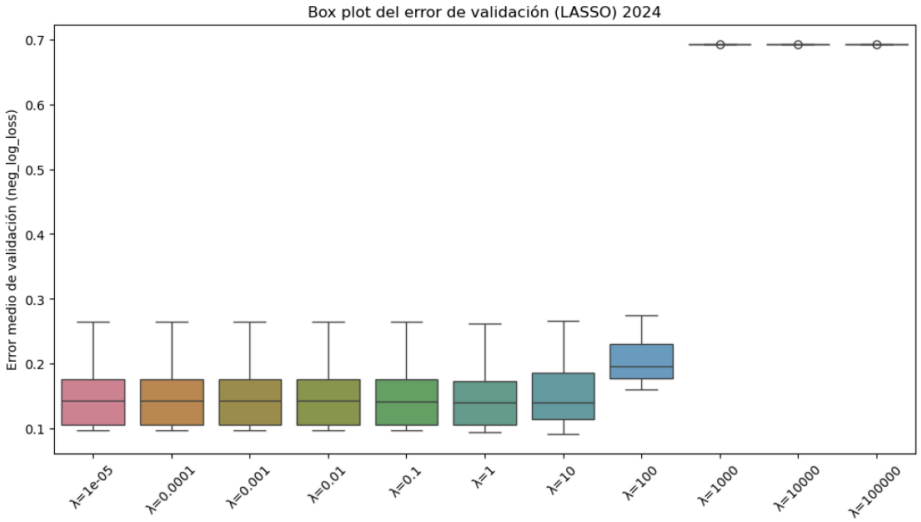
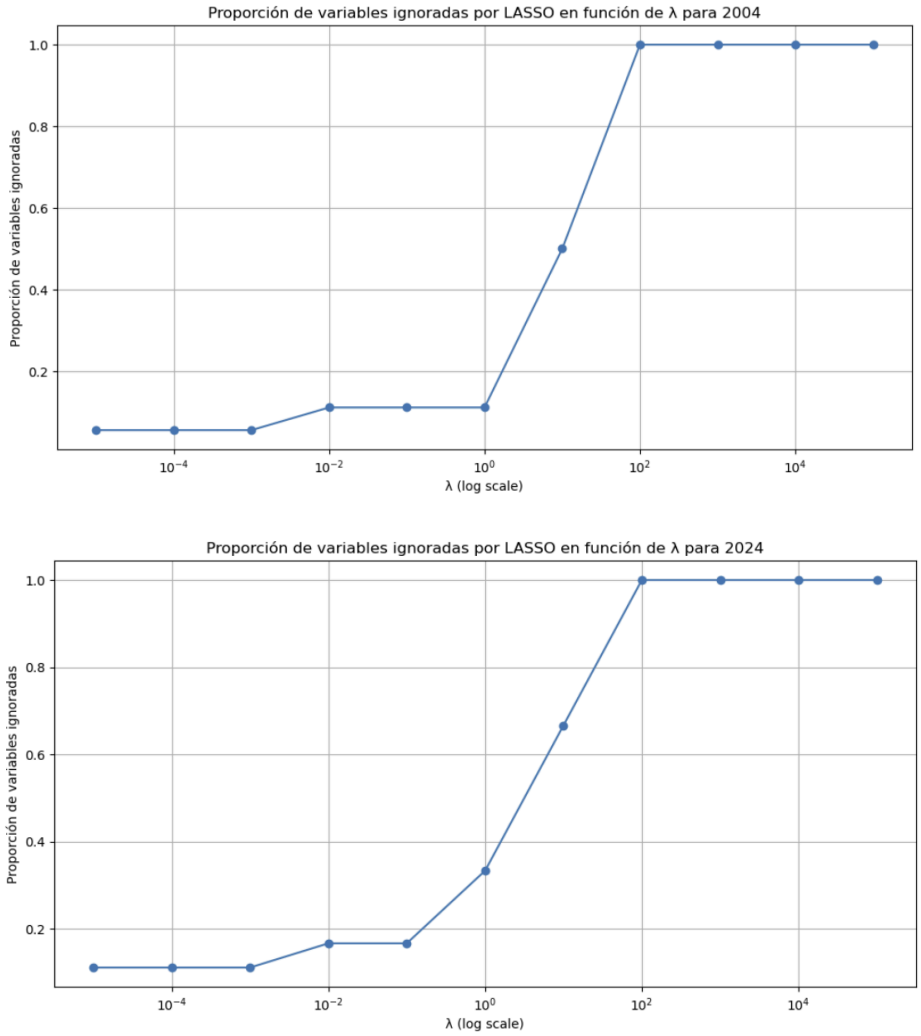

Box plot del error de validación (LASSO) 2024
Error medio de validación (neg_log_loss)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
λ=1e-05
λ=0.0001
λ=0.001
λ=0.01
λ=0.1
λ=1
λ=10
λ=100
λ=1000
λ=10000
λ=100000

Grafico 5



6. Para los valores óptimo de λ en el caso de LASSO, en cada año ignora las siguientes variables:

- 2004
  - Variable binaria que indica si la persona tiene un nivel de educación intermedio (primaria completa, secundario incompleta)
- 2024
  - Variable binaria que indica si la persona tiene un nivel de educación intermedio (primaria completa, secundario incompleta)
  - Si tiene baño
  - Si vive cerca de un basural
  - Si pidió préstamos a familia o amigos
  - Si pidió préstamos a instituciones

Tiene sentido que estas variables no necesariamente sean predictoras de desocupación, aunque nos sorprendió la falta de potencia de las variables de préstamos, ya que préstamos de instituciones podría

indicar que la persona percibe ingresos mientras que préstamos a familia o amigos puede indicar que se encuentra en el entramado informal o desocupada.