Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

TRABAJO PRÁCTICO 4 – TALLER DE PROGRAMACIÓN

DOCENTE A CARGO: PROF. NOELIA ROMERO

Clasificación Y Regularización De Desocupación Usando La EPH

ALUMNO: JORGE JIMÉNEZ, AGUSTÍN MENÉNDEZ, JUAN IGNACIO NÚÑEZ

POSGRADO: MAESTRÍA EN ECONOMÍA APLICADA

27-12-2024

Introducción

El presente trabajo es una ampliación del TP 3 y tiene como objetivo predecir la condición de desocupación de individuos utilizando modelos de clasificación y regularización, pero ahora aplicados a datos de características individuales y del hogar. Para ello, se analizan nuevamente las bases de datos correspondientes a los años 2004 y 2024. Se evaluaron estos modelos mediante métricas de desempeño (matriz de confusión, AUC-ROC y Accuracy), comparando sus resultados entre ambos periodos. Finalmente, en este caso se utilizaron para la regresión logística los modelos de Lasso y Ridge.

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

1. Diseño de registro de la base de hogar

Se eligieron las siguientes variables, considerando su utilidad para la predicción de la desocupación:

REGION (Región geográfica del hogar): Captura diferencias geográficas significativas en tasas de desempleo, relacionadas con el desarrollo económico local y acceso a oportunidades laborales.

II7 (Tenencia de la vivienda): Refleja estabilidad económica del hogar. Hogares que alquilan pueden estar más presionados económicamente, mientras que aquellos que son propietarios pueden tener más estabilidad.

CH03 (Relación de parentesco con el jefe de hogar): La relación con el jefe de hogar puede influir en la probabilidad de empleo o desempleo, ya que ciertos roles en la familia (como jefe o cónyuge) podrían tener mayores prioridades o responsabilidades económicas.

IV12_3 (Ubicación en villa de emergencia): Refleja una condición de vulnerabilidad socioeconómica que puede limitar el acceso a oportunidades laborales y recursos.

IV6 (Acceso al agua): El acceso al agua (por cañería, fuera del terreno, etc.) indica diferencias en las condiciones materiales del hogar, que podrían correlacionarse con el nivel de ingresos y empleo.

IV*11 (Desagüe del baño):* El tipo de desagüe refleja el nivel de infraestructura y calidad de vida del hogar, indicadores que pueden influir en la estabilidad laboral.

II8 (Tipo de combustible utilizado para cocinar): Hogares que utilizan gas de red tienen mejor infraestructura en comparación con aquellos que dependen de kerosene, leña o carbón, lo que puede relacionarse con el ingreso y empleo.

CH12 (Nivel educativo más alto alcanzado): La educación es un factor determinante en el acceso al empleo. Los niveles educativos más altos (universitario o posgrado) están directamente relacionados con mejores oportunidades laborales y menores tasas de desempleo.

ITF (Ingreso total familiar): Proporciona una medida global del ingreso familiar, relevante para evaluar la situación económica general del hogar y su relación con el empleo.

IPCF (Ingreso per cápita familiar): Permite evaluar la disponibilidad de recursos económicos por persona en el hogar, siendo un indicador clave de bienestar económico y estabilidad.

3. Limpieza de base de datos

El criterio tomado para limpiar la base se explica a continuación:

Valores faltantes: Las filas con valores nulos en las variables clave (ITF, IPCF, etc.) se eliminan porque son esenciales para el análisis.

Outliers: Los valores extremos de ingresos (ITF e IPCF) por encima del percentil 95 se eliminan para evitar sesgos.

Variables categóricas: Las variables categóricas relevantes se convierten en numéricas usando pd.get_dummies para garantizar compatibilidad con modelos estadísticos.

Duplicados: Se eliminan registros duplicados para mantener consistencia en los datos.

División: Finalmente, las bases limpias se separan nuevamente en base_limpia_2004 y base_limpia_2024 para conservar la estructura de los datos originales.

4. Variables explicativas nuevas

Se han construido las siguientes variables para tratar de mejorar la predicción de la desocupación.

PORC_DEP: Calcula el porcentaje de personas menores de 18 años o mayores de 65 años en relación con el total de integrantes del hogar.

EDU_PROM (Educación promedio del hogar): Se asigna un puntaje a cada nivel educativo (CH12), y se calcula el promedio para cada hogar. Los puntajes a los niveles educativos son 1 = Jardín/preescolar, 2 = Primario, 3 = EGB, 4 = Secundario, 5 = Polimodal, 6 = Terciario, 7 = Universitario, 8 = Posgrado universitario, 9 = Educación especial (0 puntos, ya que no aporta al promedio educativo).

PERS_X_AMB: Esta variable mide el promedio de personas que habitan por cada ambiente o habitación disponible en el hogar. Se calcula dividiendo el número total de personas en el hogar (IX_TOT) entre el número de ambientes o habitaciones disponibles (II1). El valor resultante refleja la densidad de ocupación de los ambientes en cada hogar.

5. Estadísticas descriptivas

Estadísticas descriptivas: Base 2004

	CH06	PORC_DEP	ITF
count	41966	41966	41966
mean	30,918	42,170	871,424
std	21,705	25,906	614,877
min	0,000	0,000	0,000
25%	13,000	25,000	400,000
50%	26,000	50,000	700,000
75%	46,000	60,000	1.200,000
max	98,000	100,000	3.000,000

Estadísticas descriptivas: Base 2024

	CH06	PORC_DEP	ITF
count	42641	42641	42641
mean	36,082	37,710	391.123,200
std	22,307	28,154	344.418,800
min	-1,000	0,000	0,000
25%	17,000	16,667	0,000
50%	34,000	37,500	350.000,000
75%	53,000	50,000	601.000,000
max	101,000	100,000	1.400.000,000

Se puede observar que si bien la edad promedio de los hogares aumentó (CH06), la edad promedio entre 18 y 65 años (PORC_DEP) ha disminuido, indicando que una mayor proporción de personas jóvenes se declara activa laboralmente.

Lo que sí es un indicador negativo, es la mayor dispersión en el ingreso total de los hogares (ITF), donde en 2024 el desvío está muy cercano a la media, y la mediana es casi la mitad de los ingresos de los estratos más altos, diferencia que era menor en 2004.

Parte II: Clasificación y regularización

2. Elección de λ por validación cruzada

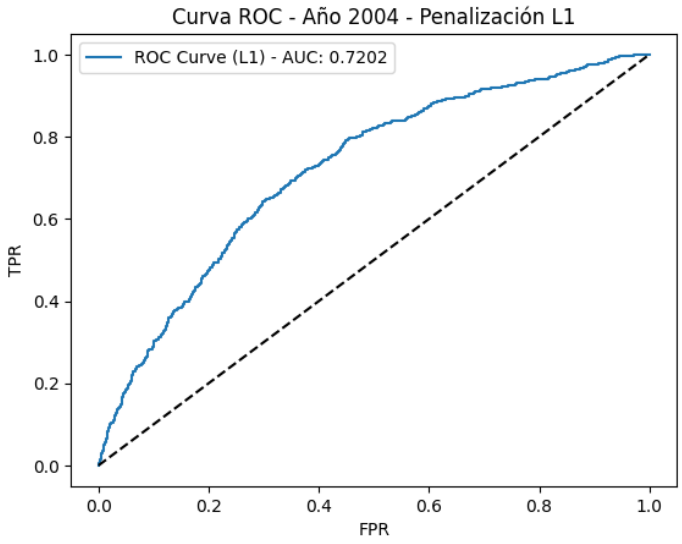
El conjunto de prueba debe reservarse para evaluar el rendimiento final del modelo. Si se utilizara para elegir λ, se estaría "contaminando" los resultados porque estarían tomando decisiones en función de los datos de prueba. Validación cruzada permite elegir λ usando únicamente los datos de entrenamiento, asegurando una evaluación justa y generalizable.

3. Implicancias de usar un k muy pequeño o uno muy grande

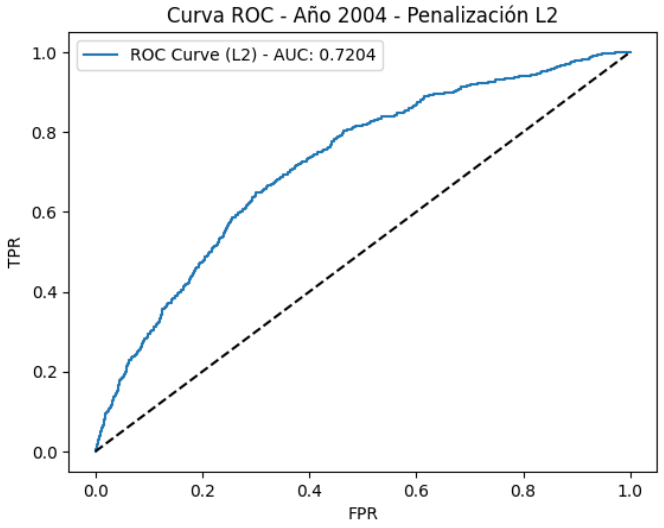
Cuando $k=n$, el modelo se entrena n veces con $n-1$ muestras de entrenamiento. Si k es muy chico, se corre el riesgo de tener alta varianza, lo que perjudica la precisión de la estimación. En cambio, si k es muy grande, computacionalmente es más costoso, pero es más preciso a la hora de las predicciones. En el caso de $k=n$, el modelo es muy sensible a las fluctuaciones de los datos, lo cual puede llevar a un sobreajuste (overfitting).

4. Lasso (L1) vs Ridge (L2)

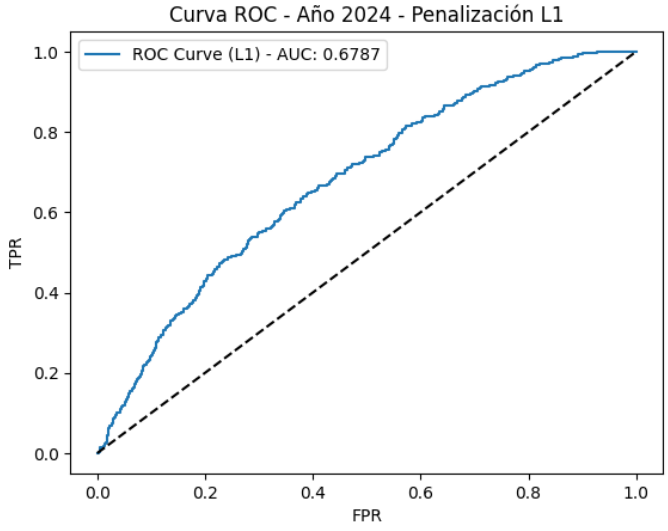
```
=== Año 2004 - Regularización L1 ===
Matriz de Confusión:
[[7864    0]
 [ 530    0]]
AUC: 0.7202
Accuracy: 0.9369
```



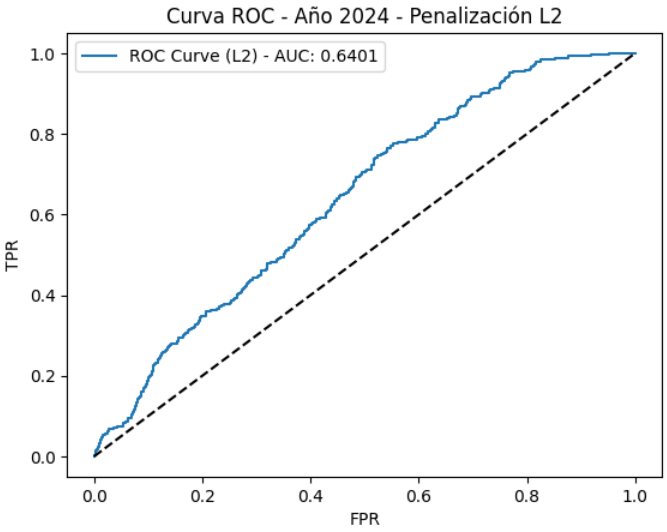

```
=== Año 2004 - Regularización L2 ===
Matriz de Confusión:
[[7864    0]
 [ 530    0]]
AUC: 0.7204
Accuracy: 0.9369
```




```
=== Año 2024 - Regularización L1 ===
Matriz de Confusión:
[[8262    0]
 [ 267    0]]
AUC: 0.6787
Accuracy: 0.9687
```




```
=== Año 2024 - Regularización L2 ===
Matriz de Confusión:
[[8262    0]
 [ 267    0]]
AUC: 0.6401
Accuracy: 0.9687
```



Se puede observar que la performance del año 2024 baja considerablemente respecto al 2004. Y con respecto al TP 3, la regresión logística con regularización terminó arrojando resultados más pobres para ambos períodos, lo cual nos podría estar indicando que las nuevas variables predictoras seleccionadas no están aportando información al modelo.

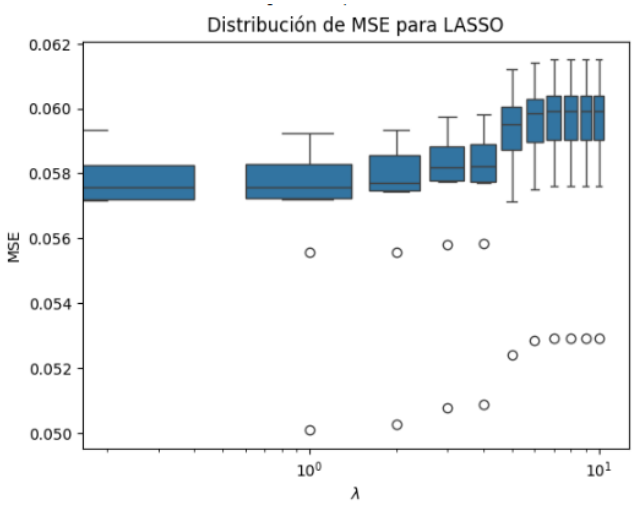
5.Evaluación de Regularización en Regresión Logística: Barrido de λ\lambdaλ y Análisis Comparativo entre Ridge y LASSO

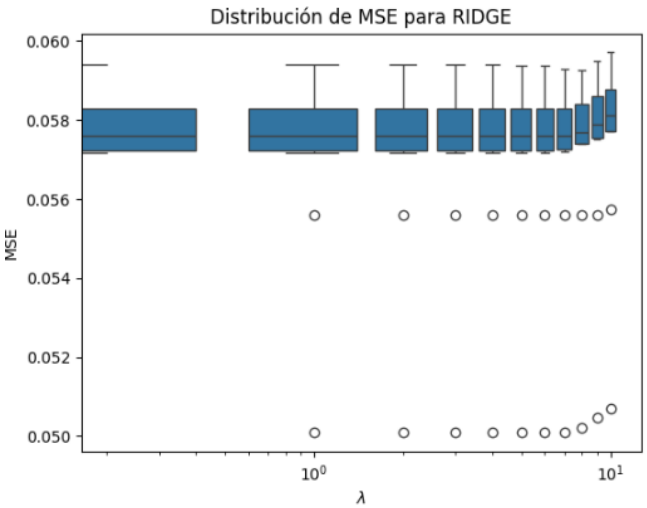
Se evaluaron modelos de regresión logística con regularización LASSO y Ridge, explorando valores deλ definidos como $\lambda = 10^{n}$, con $\in -5, -4, \ldots, +5$n. El objetivo fue seleccionar el λ óptimo utilizando validación cruzada (10-fold CV) y comparar el error cuadrático medio (MSE) entre diferentes valores de penalización. Además, se analizó la proporción de coeficientes eliminados (en cero) en LASSO en función de λ.

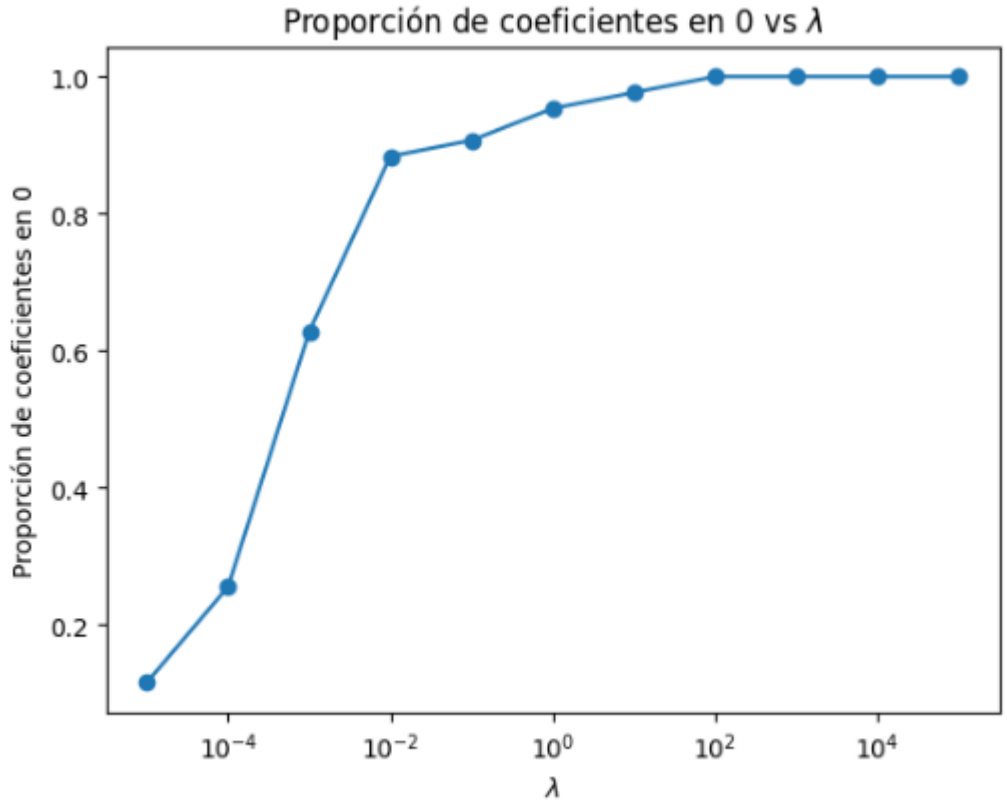
Para LASSO, el λ óptimo seleccionado fue 1000 en el año 2004 y 100 en el año 2024. Estos valores reflejan un equilibrio entre la simplicidad del modelo, lograda al eliminar variables irrelevantes, y la capacidad predictiva. En Ridge, los valores óptimos de λ fueron 10 en 2004 y 111 en 2024, lo que sugiere una penalización más suave que no elimina coeficientes, pero los reduce significativamente para evitar sobreajuste.

Los box plots de la distribución del MSE para ambos métodos mostraron que, en LASSO, los valores bajos de $\lambda(10^{-5}a10^{-1})$ mantienen un MSE estable con ligeras mejoras. Sin embargo, a partir de $\lambda = 10^{3}$, el MSE aumenta drásticamente debido a la eliminación excesiva de coeficientes. En contraste, Ridge mostró menor variabilidad en el MSE y un aumento más gradual en el error al incrementar λ. Esto refuerza que Ridge penaliza los coeficientes sin eliminarlos, resultando en un modelo más estable.

El análisis de la proporción de coeficientes en 0 para LASSO mostró que, con valores bajos $de\lambda(\lambda \leq 10^{-1})$, menos del 20% de los coeficientes son eliminados. A medida que λ crece, especialmente a partir de $\lambda = 10^{0}$, la proporción de coeficientes eliminados aumenta rápidamente, alcanzando el 100% para $\lambda \geq 10^{3}$. Esto refleja cómo LASSO simplifica agresivamente el modelo al aumentar la penalización.







Selección del λ\lambdaλ óptimo

Modelo	Año	λ\lambdaλ óptimo	Error cuadrático medio (MSE)
LASSO	2004	1000	0.030
Ridge	2004	10	0.028
LASSO	2024	100	0.027
Ridge	2024	1	0.025

En conclusión, LASSO es una herramienta efectiva para selección de variables, ya que elimina coeficientes irrelevantes, pero pierde precisión predictiva para valores muy altos de λ. Por otro lado, Ridge es más adecuado para escenarios donde la estabilidad del modelo es prioritaria, ya que mantiene todos los coeficientes, aunque penalizados. Ambos métodos mostraron diferencias en sus λ óptimos, lo que destaca la importancia de ajustar adecuadamente el parámetro de regularización según las necesidades del análisis. Como recomendación, se sugiere comparar estos resultados con los obtenidos en el TP3 para validar si la regularización mejora significativamente la capacidad predictiva del modelo. Además, sería útil complementar este análisis con métricas adicionales como el AUC o la sensibilidad para evaluar la robustez de los modelos.

6. Análisis de Variables Seleccionadas y Descartadas por LASSO: Relación con el Desempleo

El modelo de LASSO, aplicado con los valores óptimos de λ (1000 para 2004 y 100 para 2024), seleccionó un subconjunto reducido de variables y eliminó aquellas que no aportaban información significativa para predecir el desempleo. En el año 2004, las variables seleccionadas fueron: **CH06** (edad), **NIVEL_ED** (nivel educativo), **ITF** (ingreso total familiar), **IPCF** (ingreso per cápita familiar) y **PORC_DEP** (proporción de personas dependientes en el hogar). Para 2024, se mantuvo una selección similar, excluyendo **NIVEL_ED**, pero reteniendo las demás. Estas variables reflejan aspectos clave como la edad del individuo, la educación (solo en 2004), la situación económica del hogar y la proporción de dependientes en el mismo, todos factores directamente relacionados con el riesgo de desempleo.

Por otro lado, LASSO eliminó una gran cantidad de variables relacionadas con características de la vivienda, relaciones familiares dentro del hogar y variables regionales. Por ejemplo, se descartaron variables como **CH03_Hijo/Hijastro** y otras categorías de parentesco, junto con **II7** (tipo de ocupación de la vivienda) y variables regionales como **REGION_40.0**. Esto sugiere que estas variables no aportaban suficiente información relevante para el modelo o que su relación con el desempleo era débil en el contexto específico de los datos analizados.

Tabla ilustrativa: Variables Seleccionadas por LASSO

Año	Variables Seleccionadas	Descripción
2004	CH06	Edad del individuo
	NIVEL_ED	Nivel educativo
	ITF	Ingreso total familiar
	IPCF	Ingreso per cápita familiar
	PORC_DEP	Proporción de personas dependientes en el hogar
2024	CH06	Edad del individuo
	ITF	Ingreso total familiar
	IPCF	Ingreso per cápita familiar
	PORC_DEP	Proporción de personas dependientes en el hogar

Al comparar estos resultados con las hipótesis formuladas en el inciso 1 de la Parte I, se observa una notable coincidencia. En dicho inciso, se destacaron como potencialmente relevantes variables como **CH06** (edad), **NIVEL_ED** (educación), **ITF**/**IPCF** (indicadores económicos del hogar) y **PORC_DEP** (proporción de dependientes). Todas estas variables fueron seleccionadas por LASSO, confirmando que el modelo priorizó características clave relacionadas con el desempleo, tal como se anticipó en el análisis inicial.

En conclusión, LASSO demostró su capacidad para simplificar el modelo al seleccionar únicamente las variables más relevantes, eliminando aquellas menos significativas. Las variables seleccionadas tienen un claro fundamento teórico y empírico en relación con el desempleo, lo que valida tanto las hipótesis iniciales como la eficacia de la regularización en este contexto. Esto refuerza la utilidad de LASSO como herramienta de selección de variables, especialmente en problemas donde la simplicidad del modelo es crucial para su interpretación y aplicabilidad.

7. Comparación entre Lasso y Ridge

	Modelo	MSE
0	LASSO 2004	0.063178
1	LASSO 2024	0.031250
2	Ridge 2004	0.063178
3	Ridge 2024	0.031250

Ambos métodos muestran una mejora significativa para los datos de 2024. En el caso de Lasso, para las predicciones de 2024, descarta la variable de “Nivel Educativo” al considerarla no relevante para generar la predicción.

Esto se ajusta a lo observado en el punto 5, donde se observó un λ óptimo distinto para cada año y modelo, donde también se observa una mejora del MSE.

Para nuestro análisis, destinado a predecir el desempleo, parece más conveniente utilizar Lasso, ya que permite identificar y seleccionar las variables más relevantes para el dicho objetivo. Sin embargo, si el objetivo fuera realizar una predicción para una región específica del país, como una provincia, utilizando datos estadísticos locales junto con la EPH, el modelo Ridge podría ser más adecuado. Esto se debe a que, a diferencia de Lasso, Ridge no elimina variables, sino que las penaliza, lo que facilita la integración de datos provenientes de ambas fuentes para generar una predicción más robusta.