Muy buen trabajo!
Código prolijo.
Nota: 9,7

Universidad de
SanAndrés

Maestría en Economía

Machine Learning para Economistas

Profesor: Walter Sosa Escudero
Asistente: Tomás Pacheco

Trabajo Práctico 4: Clasificación y Regularización de Desocupación usando la EPH

De Boeck, Carolina
Hausvirth, Martina
Hayduk, Gaspar

Diciembre 2024

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

Inciso 1

La base de hogares de la EPH (Encuesta Permanente de Hogares) del INDEC es una de las bases de datos fundamentales para el análisis socioeconómico en Argentina. Proporciona información sobre las características habitacionales y socioeconómicas de los hogares relevados en la encuesta. Explorando el diseño de registro de la base hogar, hay tres tipos de variables que pensamos que pueden ser predictivas de la desocupación: características de la vivienda y el hogar, de qué dinero viven las personas del hogar y miembros del hogar.

En cuanto a las características del hogar, decidimos incluir el tipo de vivienda, de qué son los pisos de la vivienda, de dónde viene el agua, si tiene baño o no, dónde está el baño, si la vivienda está ubicada en una villa y si la vivienda está ubicada cerca de un basural.

En cuanto a de qué dinero viven los miembros del hogar, decidimos incluir si los miembros del hogar han vivido en los últimos 3 meses de lo que ganan en el trabajo (esta pregunta no es colineal con el estatus de desocupación de una persona pues es una pregunta referida al hogar y dentro de un hogar puede haber desocupados y no desocupados); de alguna jubilación o pensión; de aguinaldo de alguna jubilación o pensión cobrada el mes anterior; de retroactivo de alguna jubilación o cobró el mes anterior; de indemnización por despido; de seguro de desempleo; de subsidio o ayuda social (en dinero) del gobierno, iglesias, etc; con mercaderías, ropa, alimentos gobierno, iglesias, escuelas, etc; algún alquiler (por una vivienda, terreno, oficina, etc.) de su propiedad; ganancias de algún negocio en el que no trabajan; intereses o rentas por plazos fijos / inversiones; gastar lo que tenían ahorrado; pedir préstamos a familiares / amigos; menores de 10 años ayudan con algún dinero trabajando; menores de 10 años ayudan con algún menores pidiendo. Pensamos que estas variables pueden ser útiles ya que reflejarán una mala situación económica a pesar de trabajar (tengo trabajo e igual debo quemar mis ahorros) o reflejan directamente ausencia de ingresos laborales.

Por último, en cuanto a miembros del hogar, nos quedamos con la cantidad de miembros del hogar y cantidad de miembros del hogar menores de 10 años.

Inciso 2

En el archivo .py

Inciso 3

Las variables que seleccionamos de la base hogares son las descriptas en el Inciso 1. En cuanto a la base de personas, decidimos quedarnos con el sexo; la edad; el estado civil; tipo de cobertura médica; nivel educativo; ingreso per cápita familiar; y categoría de inactividad. Estos fueron los predictores del TP3. Además, mirando la variable 'CH03' que indica la relación de parentesco (Jefe, Hijo, Pareja, Yerno, Nieto, etc.) decidimos crear una variable que indique si el individuo es Jefe de Hogar o no. Esta es la primer variable que creamos.

Hay muchas variables como estado civil, cobertura médica, nivel educativo, categoría de inactividad, tipo de vivienda, de qué son los pisos interiores, de dónde viene el agua, dónde está el baño, etc. donde cada valor indica una categoría. Para estas variables se realiza One-Hot-Encoding y se crea una dummy por categoría.

Para las variables continuas como la edad, el ingreso per capita familiar y cantidad de miembros del hogar se inspeccionó sus valores y se eliminaron los missing values y los valores negativos.

Inciso 4

Aprovechando que tenemos datos de individuos y datos acerca de las características del hogar al que pertenecen los individuos, decidimos crear las siguientes variables:

- Proporción de ocupados por hogar: mide la proporción de miembros ocupados en el hogar. Su propósito es capturar el contexto económico del hogar, bajo la hipótesis de que hogares con mayor proporción de personas empleadas podrían estar en una mejor situación económica y, por lo tanto, reducir la probabilidad de que un individuo específico esté desocupado.

- Jefe de Hogar Ocupado: mide si el jefe de hogar está ocupado y puede actuar como un factor protector contra la desocupación individual. Si el jefe de hogar trabaja, podría ser un indicador de estabilidad económica o acceso a recursos, reduciendo el riesgo de que otros miembros del hogar estén desocupados.

- Jefe de Hogar Universitario: se supone que un jefe de hogar con mayor nivel educativo podría ser un factor protector frente a la desocupación.

Inciso 5: gráficos exploratorios de por dónde puede variar la desocupación

En lo que sigue se presentarán una serie de gráficos exploratorios indicando por dónde quizás podría variar la desocupación y por qué algunas variables son importantes como predictores.

Tasa de Desocupación según Sexo

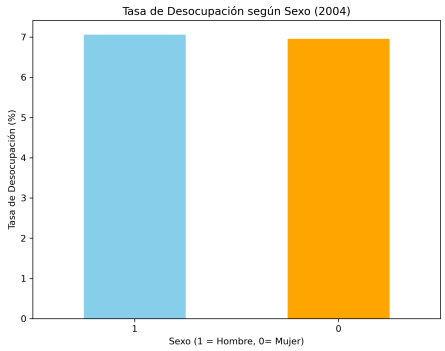

Figura 1: Tasa de Desocupación según Sexo para 2004

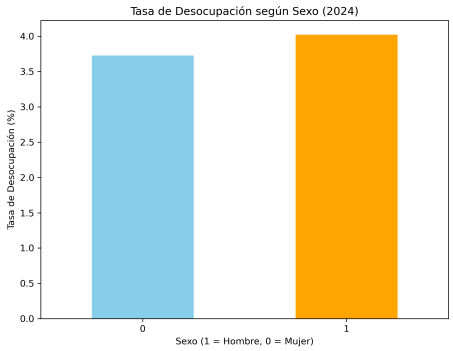

Figura 2: Tasa de Desocupación según Sexo para 2024

Distribución de la cantidad de menores de 10 años en el Hogar según desocupados y no desocupados

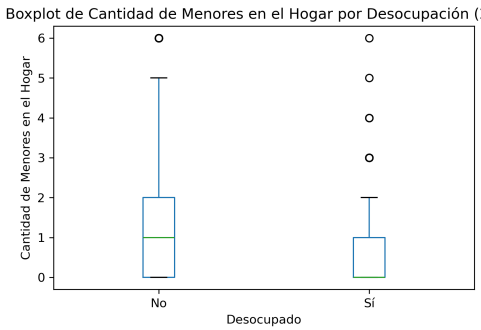

Figura 3: Distribución de la cantidad de menores de 10 años en el Hogar según desocupados y no desocupados para 2004

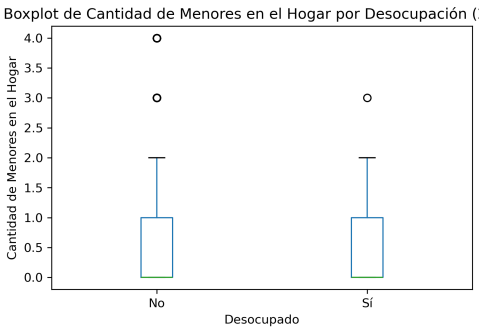

Figura 4: Distribución de la cantidad de menores de 10 años en el Hogar según desocupados y no desocupados para 2024

Tasa de Desocupación según rango etario

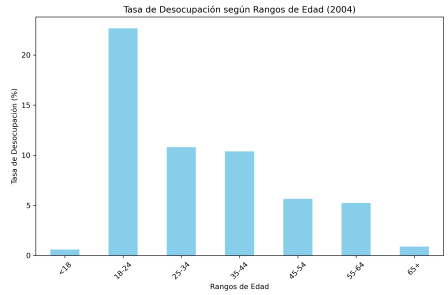

Figura 5: Tasa de Desocupación según rango etario para el 2004

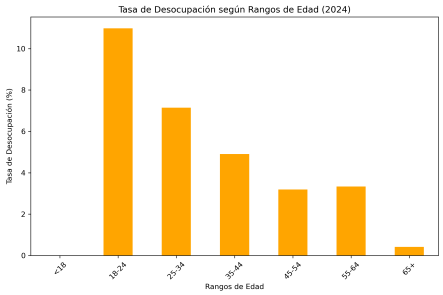

Figura 6: Tasa de Desocupación según rango etario para el 2024

Proporción de Ocupados por Hogar

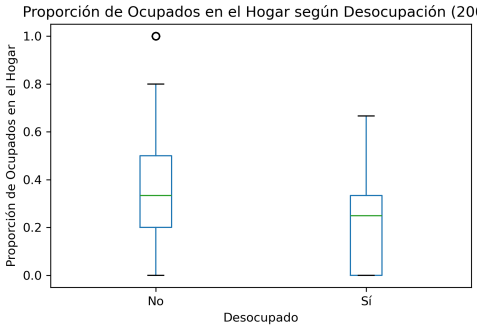

Figura 7: Distribución de la Proporción de Ocupados por Hogar para el 2004

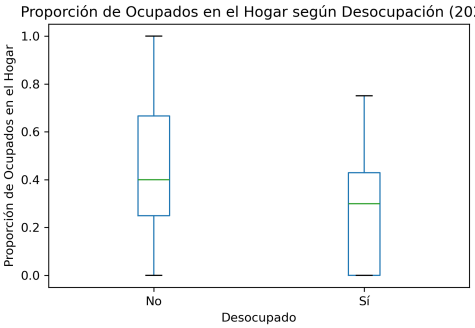

Figura 8: Distribución de la Proporción de Ocupados por Hogar para el 2024

Tasa de Desocupación y vivienda cercana a un basural

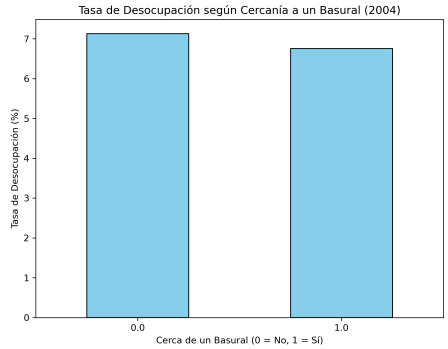

Figura 9: Tasa de Desocupación según vivienda cercana a un basural para 2004

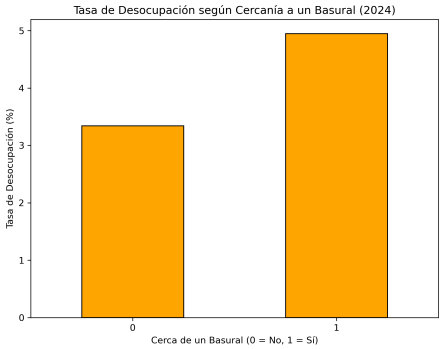

Figura 10: Tasa de Desocupación según vivienda cercana a un basural para 2024

Inciso 6

Utilizando una observación por edad y considerando a un hogar como desocupado si al menor un integrante del hogar es desocupado, encontramos que la tasa de hogares con desocupación para

Tucumán para el primer trimestre del 2024 es del 10.88 %, mientras que la tasa reportada por en INDEC en sus informes para dicho período es de 7.9 %.

Parte II: Clasificación y regularización

Inciso 1

En el archivo .py.

Inciso 2

En los métodos de regularización, el parametro λ controla la cantidad de penalización aplicada a los coeficientes del modelo. Un λ pequeño permiten coeficientes grandes, lo que puede llevar a overfitting; mientras que un λ alto reducen los coeficientes, lo que puede llevar a underfitting. El objetivo es encontrar un valor óptimo de λ que minimice el error de validación y generalice bien en datos nuevos. Podemos seleccionar el λ óptimo que minimice el error de predicción usando K-fold Cross-Validation. Para ello, primero se divide el training set en K subconjuntos (folds) aleatorios. Luego, el modelo se entrena repetidamente en K-1 folds, se valida en el fold restante y se calcula un error de predicción (puede ser el Error Cuadrático Medio o Log-Loss) en cada iteración. Al repetir esto K veces, obtenemos un error promedio de predicción para cada λ. Este proceso se repite para una barrida de λs y se selecciona el λ con menor error promedio de predicción. Con este enfoque, cada dato es utilizado para entrenar el modelo y para validarlo.

El testing set debe simular datos no observados y representar la capacidad del modelo para generalizar, si se utiliza el conjunto de prueba para elegir λ, el modelo se ajustará a ese conjunto específico, perdiendo su capacidad de generalización. La idea es seleccionar el mejor modelo (seleccionar el λ óptimo) usando el training set usando K-Fold Cross-Validation y evaluar el desempeño de ese mejor modelo en el testing set.

Inciso 3

En validación cruzada, el valor de K afecta directamente al sesgo y la varianza de las estimaciones. Con un K muy pequeño (pocos folds), cada conjunto de entrenamiento es grande y el conjunto de validación es pequeño. Esto hace que las métricas de errores sean sensibles a qué datos caen en el fold de validación, generando varianza. Un K muy grande (muchos folds) maximiza los datos para validar el modelo, por lo que el modelo aprovecha la máxima cantidad de datos en cada iteración y se reduce el sesgo.

Cuando $K = n$, el modelo se estima n veces con n-1 datos.

Inciso 4

A continuación se reporta el AUC, la curva ROC, la matriz de confusión y el Accuracy para la penalización L1 (LASSO) para el método de Regresión Logística para los años 2004 y 2024.

2004 (Accuracy: 0.927)

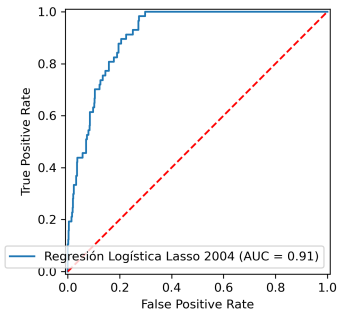

Figura 11: Curva ROC de la RL Lasso para 2004. El AUC es de 0.9114

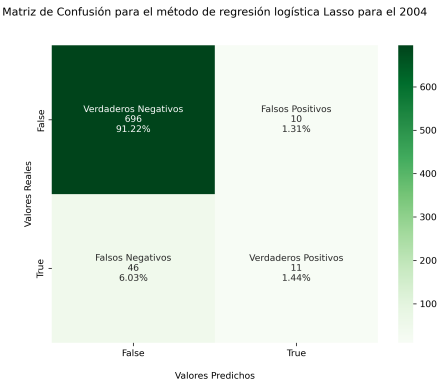

Figura 12: Matriz de Confusión de la RL Lasso para 2004

2024 (Accuracy: 0.976)

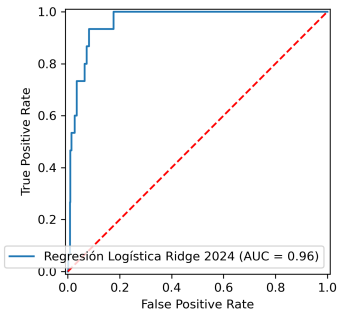

Figura 13: Curva ROC de la RL Lasso para 2024. El AUC es de 0.9643

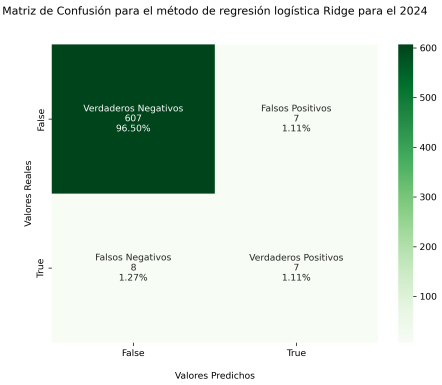

Figura 14: Matriz de Confusión de la RL Lasso para 2024

A continuación se reporta el AUC, la curva ROC, la matriz de confusión y el Accuracy para la penalización L2 (RIDGE) para el método de Regresión Logística para los años 2004 y 2024.

2004 (Accuracy: 0.928)

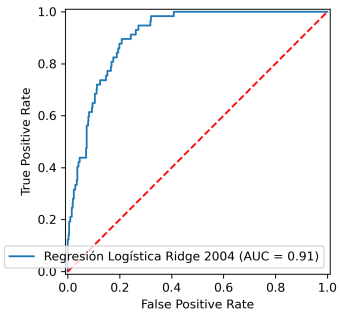

Figura 15: Curva ROC de la RL Ridge para 2004. El AUC es de 0.9068

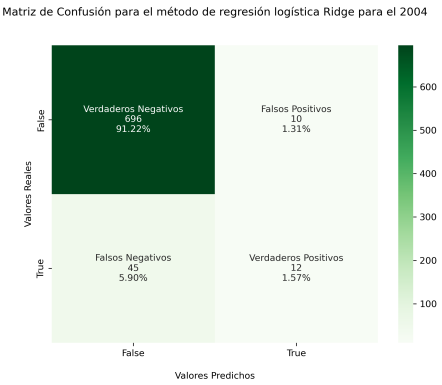

Figura 16: Matriz de Confusión de la RL Ridge para 2004

2024 (Accuracy: 0.976)

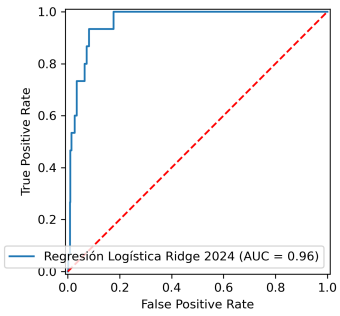

Figura 17: Curva ROC de la RL Ridge para 2024. El AUC es de 0.9626

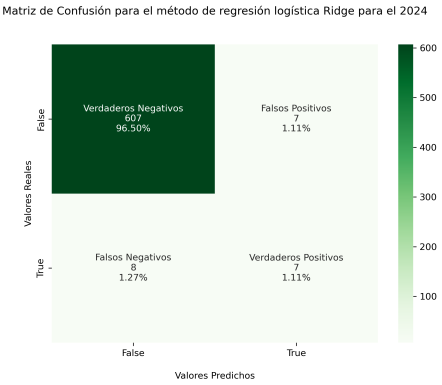

Figura 18: Matriz de Confusión de la RL Ridge para 2024

Como dijimos en el TP3, dado que la clase a predecir está desbalanceada, la métrica adecuada a mirar es el AUC (que mide la capacidad del modelopara distinguir entre las dos clases, independientemente de la distribución de clase). Recordemos que en el TP3 el AUC para el método de RL había sido de 0.92 para el 2024 y 0.89 para el 2004. Pudiendo incorporar data de la base de hogares, logramos mejorar nuestro AUC a 0.9626 usando la penalización Ridge y a 0.9643 usando la penalización Lasso para el 2024; mientras que para el 2004 pudimos mejorar nuestro AUC a 0.9068 usando la penalización Ridge y a 0.9114 usando la penalización Lasso. Como conclusión, mejoró la performance de Regresión Logística incorporando regularización.

Inciso 5

Dado que no se especifica con detalle qué métrica de error de predicción usar y dado que estamos en un problema de clasificación donde vamos a predecir probabilidades, nos pareció adecuado usar la métrica de Log-Loss dado que la misma penaliza fuertemente las predicciones mal calibradas, especialmente si un modelo está muy seguro de algo incorrecto. Log-loss aumenta su penalización cuando un modelo predice una alta probabilidad para la clase equivocada.

2024

A continuación se reporta un boxplot mostrando la distribución del error de predicción para Ridge y Lasso para cada λ para el año 2024.

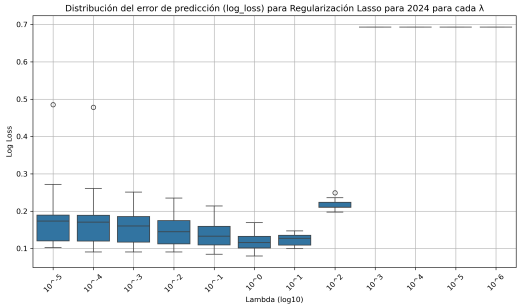

Figura 19: Distribución del Error de Predicción para Regularización Lasso

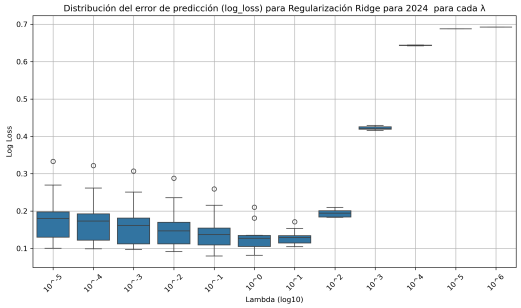

Figura 20: Distribución del Error de Predicción para Regularización Ridge

2004

A continuación se reporta un boxplot mostrando la distribución del error de predicción para Ridge y Lasso para cada λ para el año 2004.

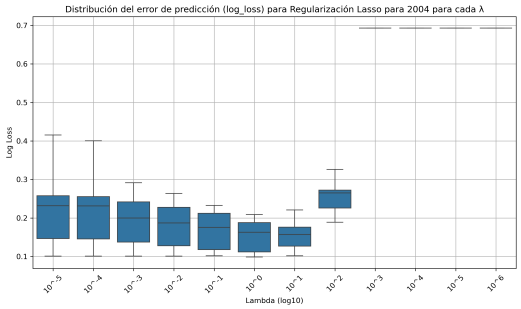

Figura 21: Distribución del Error de Predicción para Regularización Lasso

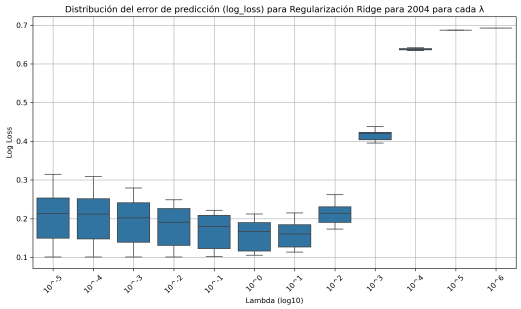

Figura 22: Distribución del Error de Predicción para Regularización Ridge

Podemos observar que el λ con menor error de predicción promedio para Lasso 2004 y 2024 y para Ridge 2004 es λ=0, mientras que para Ridge 2024 es λ=1

Promedio de Proporción de variables ignoradas para Regularización Lasso para el 2004

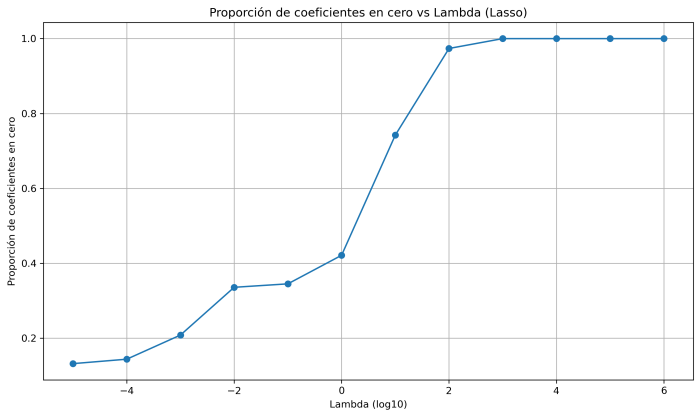

Figura 23: Proporción de coeficientes iguales a cero en Lasso para cada λ para el 2004

Promedio de Proporción de variables ignoradas para Regularización Lasso para el 2024

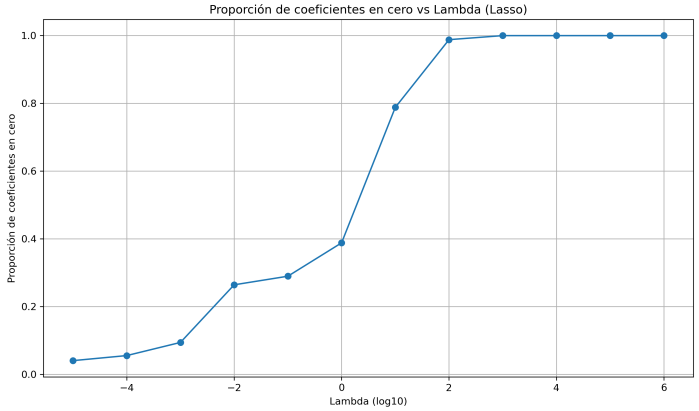

Figura 24: Proporción de coeficientes iguales a cero en Lasso para cada λ para el 2024

Inciso 6

Como dijimos en el inciso anterior, el λ con menor error de predicción promedio es λ=1 para la Regularización Lasso tanto para 2004 como para 2024. Las siguientes tablas muestran las variables con coeficientes iguales a cero en la Regularización Lasso Óptima para los años 2024 y 2004:

Cuadro 1: Variables con coeficientes iguales a cero en Lasso 2024 y 2004

Lasso 2024

Variable	Coeficiente
IV8	0.0000
IV12_1	0.0000
V5	0.0000
V8	0.0000
V19_A	0.0000
V19_B	0.0000
IPCF	0.0000
civ_2	0.0000
civ_3	0.0000
cob_2	0.0000
cob_3	0.0000
cob_23	0.0000
nivel_ed_1	0.0000
nivel_ed_2	0.0000
nivel_ed_7	0.0000
incap_1	0.0000
incap_2	0.0000
incap_3	0.0000
incap_4	0.0000
incap_5	0.0000
incap_6	0.0000
incap_7	0.0000
piso_2	0.0000
agua_2	0.0000
agua_3	0.0000
baño_0	0.0000
baño_1	0.0000

Lasso 2004

Variable	Coeficiente
v4	0.0000
v6	0.0000
v8	0.0000
v10	0.0000
v14	0.0000
v19_a	0.0000
v19_b	0.0000
civ_1	0.0000
civ_9	0.0000
cob_1	0.0000
cob_9	0.0000
nivel_ed_2	0.0000
incap_1.0	0.0000
incap_2.0	0.0000
incap_4.0	0.0000
incap_5.0	0.0000
incap_6.0	0.0000
incap_7.0	0.0000
vivienda_1	0.0000
vivienda_3	0.0000
vivienda_4	0.0000
vivienda_Otro	0.0000
piso_0.0	0.0000
piso_1.0	0.0000
piso_4.0	0.0000
agua_0.0	0.0000
agua_1.0	0.0000
redagua_0.0	0.0000
redagua_3	0.0000
redagua_4	0.0000
redagua_Ns./Nr.	0.0000
baño_1.0	0.0000

En el Inciso 1 propusimos incorporar variables referidas a las características de la vivienda (estas variables empiezan con prefijo IV) y variables referidas de a qué dinero viven las personas del hogar (estas variables empiezan con V). Para el 2024, se descartaron las variables que indicaban si el hogar vivió de ingresos referidos a subsidio de gobiernos e iglesias (V5), de ingresos referidos a alquileres de propiedades ni tampoco si menores de 10 años colaboraron pidiendo dinero o trabajando (V19A y V19B). También se descartó la variable que indica si la vivienda tiene agua fuera de la vivienda pero dentro del terreno o fuera del terreno (agua2 y agua3). Se descartaron las variables que indican la categoría de inactividad. Se descartaron las dummies que indican Primario incompleto, primario completo y Sin Instrucción.

Para el 2004, se descartaron las dummies que indican si el hogar vivió de seguro de desempleo; con mercaderías, ropa, alimentos de gobierno, iglesias, escuelas, etc.; de algún alquiler (por una vivienda, terreno, oficina, etc.) de su propiedad; de intereses o rentas por plazos fijos / inversiones; de pedir préstamos a familiares / amigos. No se descararon las dummies que indican si el hogar vivió de ahorros pasados; de alguna jubilación o pensión. También hubo dummies como si la vivienda es un departamente o local no construido para habitación que tuvieron un coeficiente distinto de cero.

Dado que la performance del método de Regresión Logística mejoró al incorporar variables referidos al hogar, vivienda e ingresos del hogar y al incorporar regularización, no perdimos nada en sofisticar el modelo del TP3. También podemos observar que las tres variables que creamos en el Inciso 4 no obtuvieron coeficientes iguales a cero, por lo que sirvieron para predecir el estatus de desocupación.

Inciso 7

En primer lugar, aprovechamos que el λ optimo para la regularización Ridge es diferente para el 2004 (λ=1) y 2024 (λ=10) y hacemos una comparación entre ambos modelos. El ECM para el Ridge Óptimo 2024 aplicado sobre el test set es de 0.021, mientras que el ECM para el Ridge Óptimo 2004 aplicado sobre el test set es de 0.052. Si usamos log-loss como métrica, los resultados son de 0.086 para 2024 y 0.175 para el 2004. Por tanto, el Ridge Óptimo de 2024 predice mejor que el Ridge Óptimo de 2004.

Ahora, vamos a comparar el Lasso Óptimo 2024 contra el Ridge Óptimo 2024 y ver qué método predice mejor. El ECM aplicado sobre el test set para el Lasso Óptimo 2024 es de 0.021, mientras que el ECM aplicado sobre el test set del Ridge Óptimo 2024 es de 0.052. Si miramos el log-loss como métrica, el log-loss del Lasso Óptimo para 2024 sobre el test set es de 0.073, mientras que el log-loss del Ridge Óptimo para 2024 aplicado sobre el test set es de 0.086. Analizando ambas métricas, concluimos que Lasso predice mejor para 2024 que Ridge y además preselecciona variables y simplifica el modelo.