Universidad de Buenos Aires

Facultad de Ciencias Económicas

Maestría en Economía Aplicada

Big data y Aprendizaje Automático Aplicado

Trabajo Práctico 4

Alumnos:

Oscar Cuentas Sandy

Jhonatan Saúl Elguera Quispe

Rodrigo Hermoso

Profesora:

María Noelia Romero

3° Trimestre 2024

Índice general

Introducción

En este informe se presenta un análisis comparativo entre los datos de la Encuesta Permanente de Hogares (EPH) de Argentina correspondientes a los años 2004 y 2024, con un enfoque principal en el desempleo y sus determinantes. El estudio tiene como objetivo identificar las diferencias en las características del mercado laboral en dos períodos distantes, tomando en cuenta factores individuales y del hogar que pudieron haber influido en las tasas de desempleo a lo largo de estas dos décadas.

A través de este análisis, se pretende construir un conjunto de variables que permitan explicar las variaciones en las tasas de desempleo, utilizando datos sociodemográficos y económicos disponibles en la EPH.

Para abordar el objetivo del informe y seleccionar las variables más relevantes, se utilizará técnicas de regularización, es especifico, los modelos Ridge y Lasso. Esta metodología es particularmente útil cuando se trabaja con un gran número de variables predictoras y busca evitar el sobreajuste (*overfitting*) al penalizar la inclusión de variables innecesarias. A través de la regularización Lasso, se identificará qué variables tienen un mayor impacto en la determinación del desempleo en Argentina, comparando las tendencias de 2004 con las de 2024.

El análisis y la comparación de estos dos períodos permitirá arrojar conclusiones sobre cómo el contexto socioeconómico y las políticas laborales han influido en la evolución del desempleo, contribuyendo al entendimiento de los cambios en la estructura del mercado laboral en Argentina en las últimas dos décadas.

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

¿Puede la configuración del hogar determinar la pertenencia al mercado laboral de sus miembros?

La literatura económica identifica diferentes fuentes que determinan as decisiones económicas de los individuos, una de las principales suele estar asociada a las características del hogar donde se reside. La condición de ocupación no escapa de esta lógica. Existe un cuerpo teórico considerable que respalda el hecho de que dependiendo de cómo es el hogar donde se vive sus miembros pertenecerán o no al mercado laboral.

Por otro lado, la base de datos de la Encuesta Permanente de Hogares (EPH) cuentan con múltiples variables que caracterizan el tipo de hogar y pueden ayudar a perfeccionar la estimación de desocupados (u ocupados) al funcionar como un proxy de estatus socioeconómico. A priori, puede considerarse que una de estas métricas son las que buscan evaluar diferentes dimensiones de las Necesidades Básicas Insatisfechas (baño, número de habitaciones, condiciones de hacinamiento).

Adicionalmente, las variables de la sección "¿En los últimos tres meses, las personas de este hogar han vivido...?", que identifican las diversas fuentes de ingresos de los hogares también pueden ayudar a este objetivo. En la misma línea, la configuración de los miembros de un hogar puede determinar la pertenencia al mercado laboral, por ejemplo, contar con un mayor número de menores de edad en el hogar hace más necesario que más

residentes del hogar se encuentren empleados para financiar las necesidades de los menores.

Considerando lo anterior, se seleccionaron tres variables para mejorar la estimación de la población desocupada. Estas son, el número de adultos mayores en el hogar, la cantidad de menores en primera infancia y si el hogar recibe transferencias del Estado.

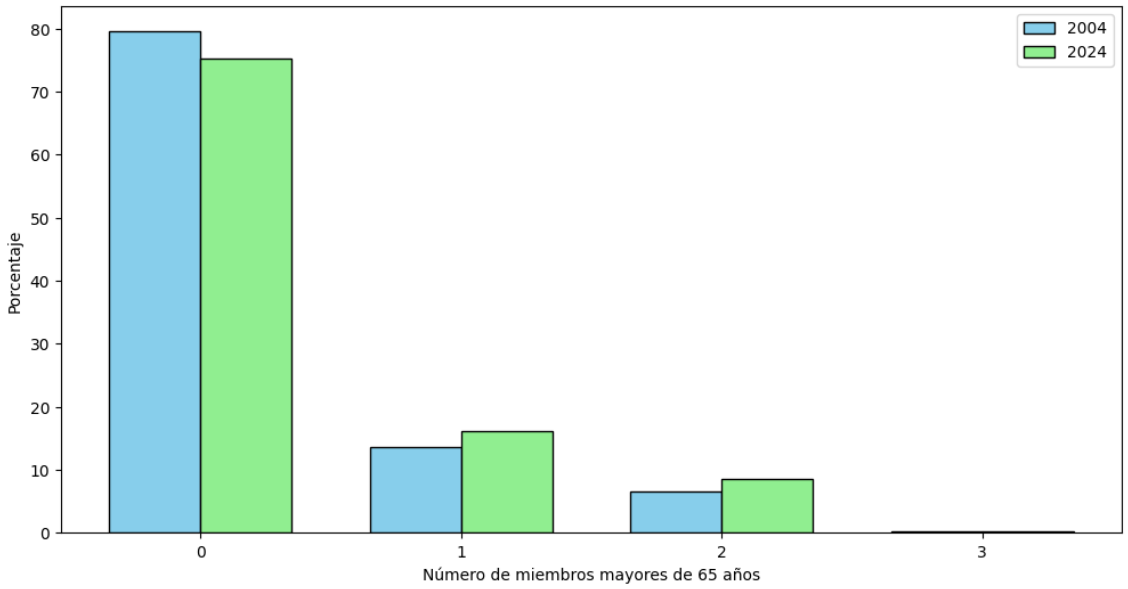
Ante de comprobar la capacidad de estas variables para mejorar la presión de la clasificación de los modelos, se realizan algunos filtros. En primer lugar, no se consideran a aquellos hogares que cuentan con una cantidad de ambientes mayores de 11, debido a que pueden representar a espacios multifamiliares que sesguen los resultados. Siguiendo esta misma razón, tampoco se tiene en cuenta a las observaciones que registren una cantidad de miembros del hogar mayores a 10. Además, se eliminaron los casos donde los individuos registran edades negativas o iguales cero, ya que representarían errores de registro o a recién nacidos.

Análisis descriptivo de las circunstancias del hogar que predicen la desocupación

Con relación a los adultos mayores en el hogar, si se compara la muestra de 2004 y 2024 encontramos que la proporción de encuestados que reside con un mayor número de adultos mayores ha experimentado un aumento con el tiempo. De esta forma es que para 2024 los encuestados que residían con uno o dos adultos mayores se incrementaron en 3 y 1 punto porcentual, respectivamente.

Gráfico 1

Distribución de encuestados según número de adultos mayores en el hogar, 2004 y 2024



Como se puede ver en la Tabla 1, al desagregar los resultados por condición de ocupación en ambos grupos, ocupados y desocupados, se ha incrementado, entre 2004 y 2024, la proporción de individuos que residen con uno o dos adultos mayores. Es importante resaltar que en ambos años se registra que, en comparación con los desocupados, existe un mayor número de ocupados que no reside con personas mayores.

En relación con la segunda variable, la proporción de encuestados que viven en hogares que reciben, se observa en la tabla 2 que entre 2004 y 2024 el porcentaje que percibía alguna transferencia del Estado paso del 6.8 % al 18 %. Al desagregar los resultados por la situación de ocupación, se encuentra que en ambos años la proporción de desocupados que recibían algún financiamiento del gobierno era mayor que los ocupados que también percibirán esta fuente de ingreso. Asimismo, se encuentra que, en el tiempo se ha duplicado el porcentaje de desocupados que perciben una transferencia estatal (ver Tabla 3).

Tabla 1

Proporción de encuestados según situación de ocupación y número de adultos mayores con los que reside, 2004 y 2024

	2004		2024	
	Número de miembros mayores de 65 años	%	Número de miembros mayores de 65 años	%
Ocupados	0	86.3 %	0	84.2 %
	1	10.5 %	1	11.6 %
	2	3.1 %	2	4.1 %
	3	0.1 %	3	0.0 %
Desocupados	0	85.7 %	0	82.6 %
	1	11.8 %	1	12.5 %
	2	2.1 %	2	4.8 %
	3	0.4 %	3	-

Tabla 2

Proporción de encuestados según situación de ocupación y número de adultos mayores con los que reside, 2004 y 2024

2004		2024	
Reside en un hogar que percibe transferencia del Estado	%	Reside en un hogar que percibe transferencia del Estado	%
No	93.2 %	No	82.0 %
Si	6.8 %	Si	18.0 %

Por otro lado, en cuanto a la cantidad de niños en primera infancia que residen en el hogar de los encuestados, se observa en el Gráfico 2 que esta ha disminuido significativamente en últimos 20 años. En 2004, el 65.7 % de los encuestados no residía con ningún menor en el hogar, este porcentaje se incrementó a 80.7 % en 2024. Este resultado se encuentra

alineado con lo registrado en los demás escalones, el porcentaje que declara residir con al menos un niño menor de 6 años a disminuido entre 2004 y 2024. Este resultado junto con el observado con lo observado para la variable de adultos mayores es una señal del envejecimiento poblacional: cada vez hay más adultos mayores, pero menos niños menores de 6 años.

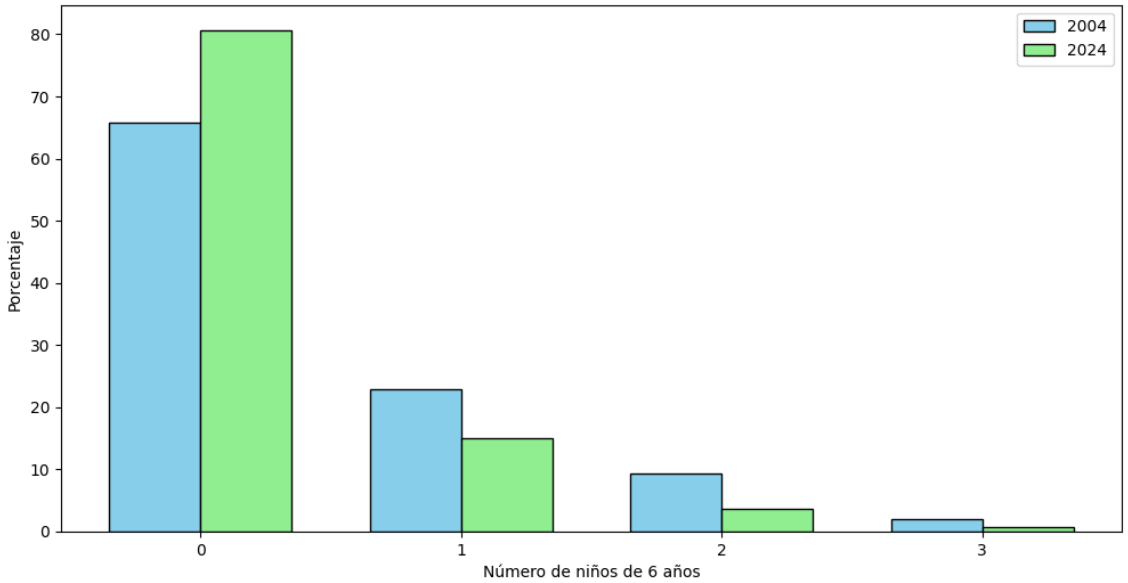
Tabla 3

Proporción de encuestados según situación de ocupación y número de adultos mayores con los que reside, 2004 y 2024

	2004		2024	
	Reside en un hogar que percibe transferencia del Estado	%	Reside en un hogar que percibe transferencia del Estado	%
Ocupados	No	95.3 %	No	86.3 %
	Si	4.7 %	Si	13.7 %
Desocupados	No	88.0 %	No	77.8 %
	Si	12.0 %	Si	22.2 %

Gráfico 2

Distribución de encuestados según número menores en primera infancia en el hogar, 2004 y 2024



Asimismo, al igual que las variables anteriores, se desagregan los resultados según la condición de ocupación de los encuestados. Según la Tabla 4, tanto para los ocupados como para los desocupados se ha incrementado el porcentaje que no viven con menores de edad en sus hogares. Por otro lado, entre los ocupados se registra una mayor proporción que declara residir con al menos un niño en primera infancia en el primer hogar. Esto

sugiere que la presencia de menores en el hogar y la necesidad de financiar las necesidades de estos miembros del hogar incentiva que los miembros del hogar se encuentren ocupados.

Tabla 4

Proporción de encuestados según situación de ocupación y número de menores en primera infancia con los que reside, 2004 y 2024

	2004		2024	
	Número de miembros menores de 6 años	%	Número de miembros menores de 6 años	%
Ocupados	0	71.0 %	0	85.2 %
	1	20.9 %	1	12.1 %
	2	7.1 %	2	2.2 %
	3	1.0 %	3	0.4 %
Desocupados	0	73.4 %	0	83.6 %
	1	18.8 %	1	13.5 %
	2	6.3 %	2	2.6 %
	3	1.5 %	3	0.3 %

Parte I. Clasificación y regularización

Una vez analizadas algunas características descriptivas de las nuevas variables que se consideran el ejercicio de clasificación de los desocupados, se responden algunas cuestiones teóricas relacionadas con los modelos regularización.

Por un lado, tanto el modelo Ridge como Lasso requieren determinar el valor de un hiperparámetro de penalización (λ). Para obtener esta magnitud se utiliza el procedimiento de validación cruzada.

El método de validación cruzada requiere definir un número K de *splits* o subconjuntos que se extraerán de la muestra total. Es decir, si a K se le asigna el valor de 10, entonces, el método dividirá la muestra de trabaja en 10 subconjuntos donde cada subconjunto tendrá una misma proporción de muestra de entrenamiento y prueba.

Una vez selecciona la cantidad de *splits*, se debe determinar si se emplea un Grid Search o un Random Search. En el caso del primero, se definirá un conjunto discreto de posibles valores que puede tomar el hiperparámetro. Una vez hecho esto, se tomará cada uno de estos valores y cada subconjunto construido para estimar el modelo de clasificación y obtener una medida de resultado (MSE, accuracy, etc.). De esta manera, a cada valor del hiperparámetro de le asociará K medidas de resultado, a partir de las cuales se puede obtener un promedio y seleccionar como mejor valor del hiperparámetro a aquel que optimice la magnitud de este promedio. Para el Random Search, el procedimiento es similar con la diferencia de que en vez de definir un conjunto discreto de valores del hiperparámetro se establece un rango continuo de números dentro del cual se seleccionará aleatoriamente la magnitud del hiperparámetro que se utilizará para completar el resto del procedimiento.

Es importante mencionar que, el conjunto de prueba se debe reservar únicamente para evaluar el rendimiento final del modelo después de seleccionar todos los hiperparámetros, incluido λ. Si se utilizara el conjunto de prueba para elegir λ, estaríamos introduciendo sesgo en el proceso de selección de parámetros, ya que el modelo estaría "viendo" los datos de prueba durante el ajuste de hiperparámetros. Esto haría que el desempeño del modelo en el conjunto de prueba no reflejara realmente su capacidad de generalización a datos nuevos, pues el modelo ya habría sido ajustado indirectamente para tener un buen rendimiento en esos datos.

Con relación a la validación cruzada y la pregunta ¿cuáles son las implicancias de usar un K muy pequeño o uno muy grande? Respondemos que el valor de K (subconjuntos de muestreo) es importante porque un valor reducido puede generar un mayor riesgo de sobreajuste, ya que puede ajustarse a las particularidades de los datos de entrenamiento. Sin embargo, esto tiene asociado un menor costo computacional. En contraste, elegir un valor elevado de K, disminuye las probabilidades de sobreajuste, pero usa de forma intensiva recursos computacionales.

Asimismo, en el extremo, podría considerarse K=N (siendo N el tamaño de muestra), es decir, que se realicen tantos subconjuntos como cantidad de datos con los que se cuenta. De esta forma, se estima el modelo N veces con N-1 datos asegurando que al menos un dato haya sido considerado, alguna vez, como parte del conjunto de prueba o lo que es lo mismo que, alguna vez, un dato no haya formado parte del conjunto de prueba.

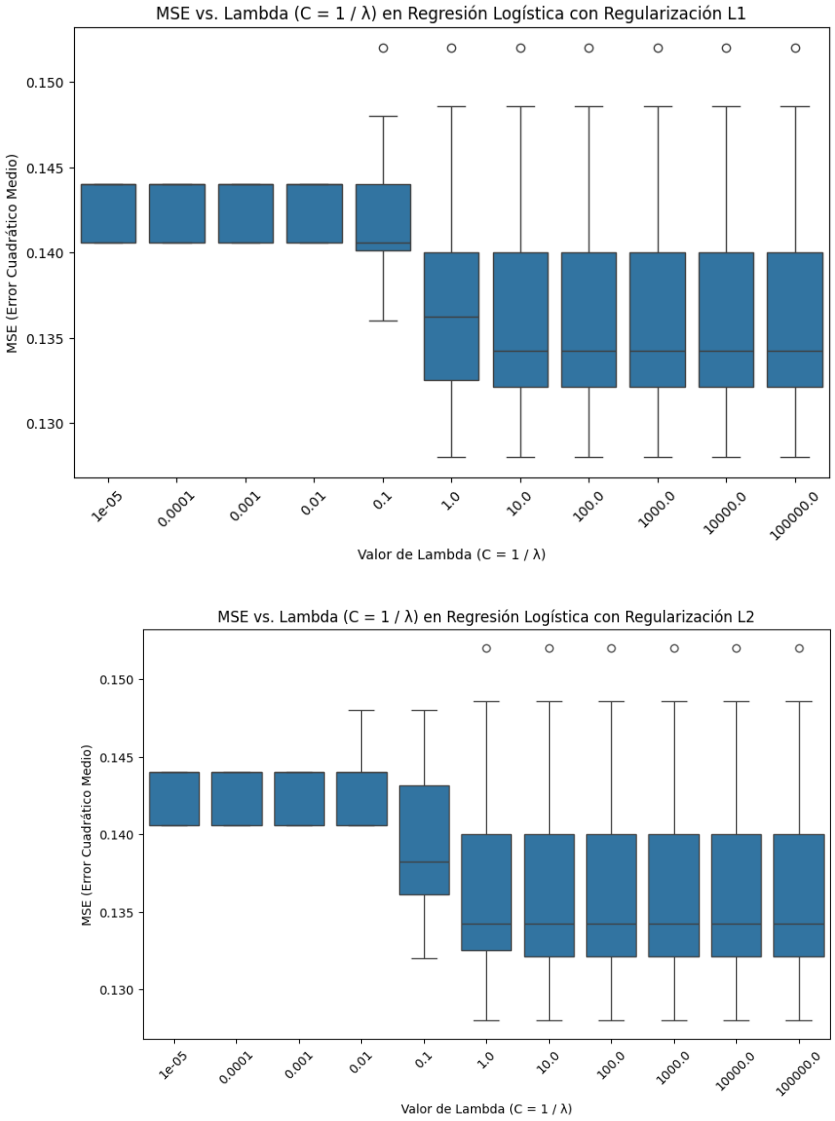
Una vez respondidas estos aspectos teóricos, se procede a presentar los resultados de los modelos estimados. Inicialmente se presenta los resultados del modelo de clasificación donde el hiperparámetro de penalización es seleccionado de manera ad hoc ($\lambda = 1$), posteriormente se muestran los resultados de los modelos obtenidos cuando el valor del hiperparámetro es seleccionado por validación cruzada.

Indiferentemente del método de regularización empleado, al comparar los resultados de los años 2004 y 2024, con los obtenidos cuando no se aplicaba algún método de regularización y solo se consideraban las características individuales de los encuestados, se obtienen resultados más realistas ya que los anteriores mostraban un *accuracy* de 1. En línea con esto, el nivel de precisión para las estimaciones del 2004 se encuentra alrededor del 0.84, mientras que para el 2024 se encuentra cercano al 0.91.

Cuando se utiliza el procedimiento de validación cruzada se encuentra que para el año 2004, el mejor valor que λ puede asumir es igual a 0.1 tanto en el modelo Ridge como Lasso. En el caso del modelo Lasso, la única variable descartada fue la constante, asimismo, las variables añadidas presentan las direcciones esperadas. La presencia de población vulnerable como adultos mayores y menores en primera infancia, reduce las posibilidades de encontrarse desocupado. Esto puede estar asociada por la necesidad familiar de cubrir las necesidades económicas de estos miembros del hogar. Asimismo, el hecho de residir en un hogar que cuente como una fuente de ingresos a alguna transferencia del Estado tiene una menor posibilidad de encontrarse laboralmente ocupado. De forma complementaria, se presenta en el Gráfico 3, la distribución de los MSE calculados para cada uno de los valores que se definieron como posibles valores del hiperparámetro. Se observa como a medida que se incrementa el valor de hiperparámetro el valor del MSE cae.

Gráfico 3

Distribución de MSE obtenidos para los diferentes valores del hiperparámetro y método de regularización, 2004



Para el 2024, el mejor valor que λ puede asumir es igual a 10^5 tanto en el modelo Ridge como Lasso. Esto genera que en el modelo Lasso, los coeficientes asociados cada una de las variables sean cercanos a cero. Asimismo, se presenta en el Gráfico 4, la distribución de los MSE calculados para cada uno de los valores que se definieron como posibles valores del hiperparámetro.

En resumen, los métodos de regularización parecen ser más apropiados para la muestra del 2004. Para el caso de 2024, la estimación, a pesar de considerar valores muy elevados del hiperparámetro de penalización, no llega a un punto de saturación que optimice el proceso y, por lo tanto, no genera una estimación confiable. Asimismo, para los resultados del 2004 ambos métodos, Ridge y Lasso, no presentan diferencias significativas en terminamos de su performance, los dos procedimientos presentan el mismo nivel de MSE (ver Tabla 5). Es importante mencionar que, si bien los métodos de regularización considerados no resultaron apropiados para el año 2024, la dirección de los coeficientes

tanto para ese año como para el 2004 resultan acorde a lo esperado. Ser mujer, no presentar una unión, no tener una cobertura médica, tener menor instrucción educativa y recibir transferencias del Estado incrementan las posibilidades de estar desocupado. Por otro lado, ser más joven, registrar menores ingresos familiares per cápita, residir con adultos mayores o menores en primera infancia reduce las probabilidades de no estar ocupado.

Gráfico 4

Distribución de MSE obtenidos para los diferentes valores del hiperparámetro y método de regularización, 2024

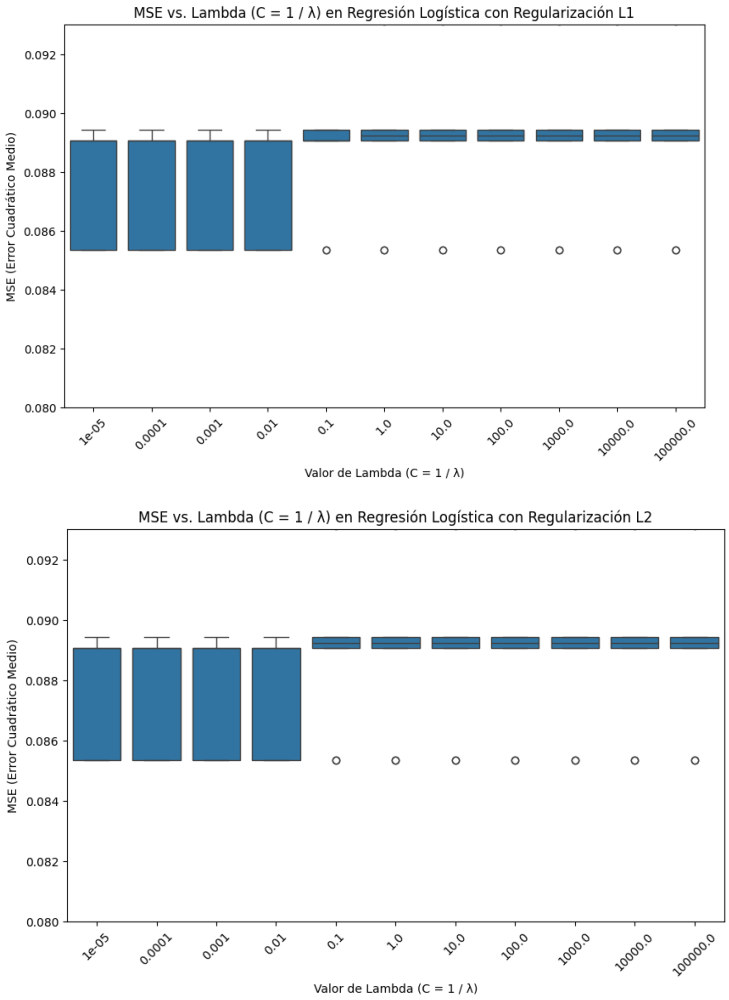


Tabla 5

Resultados de resumen de los modelos de regularización

	2004		2024	
	Ridge	Lasso	Ridge	Lasso
Lambda	0.1	0.1	10^5	10^5
MSE	0.137	0.137	0.09	0.09