

# CONTROL MODELOS DE REGRESION CASO COMPASS MARITIME

ANGELO ANTUNEZ  
JUAN COLLANTES  
RONAL NOA

PEBIBA 2024-2025

# INDICE DE CONTENIDOS

1. SINOPSIS
2. LECTURA Y ARREGLO DE DATOS TÉCNICAS AVANZADAS
3. REGRESIÓN LINEAL
- 3.1. MODELO EN DETALLE
- 3.2. VISUALIZACIÓN DE BANDA DE CONFIANZA
4. REGULARIZACIÓN
- 4.1. REGRESIÓN RIDGE
- 4.2. REGRESIÓN LASSO

PROYECTO FIN DE GRADO

# 1 / SINOPSIS

## 1/ SINOPSIS

El caso se centra en Compass Maritime Services (CMS), una empresa de corretaje marítimo especializada en la compra, venta y valorización de barcos. En mayo de 2008, Basil Karatzas, director de finanzas y proyectos, enfrenta el desafío de determinar un precio razonable para el barco “Bet Performer”, un carguero tipo capesize que un cliente desea adquirir. En un mercado marítimo volátil y competitivo, Karatzas debe emplear métodos de valorización basados en datos históricos y características del barco, además de diseñar una estrategia de negociación efectiva.

El éxito de esta operación depende de ofrecer un precio competitivo sin comprometer la relación con el cliente ni la reputación de CMS como un actor clave en la industria.

# CONTROL I - MODELOS DE REGRESION CASO COMPASS MARITIME

# 2 / LECTURA Y ARREGLO DE DATOS

## 1/ TECNOLOGÍA

## 2 / ENSAYO-ERROR

## 3 / CORRECCIÓN DE CÓDIGO

## 2 / LECTURA Y ARREGLO DE DATOS

- Importar y leer los datos: El archivo Excel se carga en un DataFrame de pandas.
- Preparar las variables: Seleccionamos las variables predictoras (Age.at.Sale, Dead.Weight.Tons, Avg.Month.Balt, Flag.fecha, Age.at.Sale\_Flag.fecha) y la variable objetivo (Sale.Price).
- Dividir los datos: Los datos se dividen en conjuntos de entrenamiento y prueba.
- Crear y ajustar el pipeline: El pipeline escala los datos y realiza la regresión Ridge con validación cruzada para optimizar el valor de alpha.
- Evaluar el modelo: Calculamos el MSE y el R<sup>2</sup> para evaluar el rendimiento del modelo en el conjunto de prueba.
- Visualizar los resultados: Un gráfico de dispersión para comparar los valores reales y las predicciones.

# 3 / REGRESIÓN LINEAL MÚLTIPLE

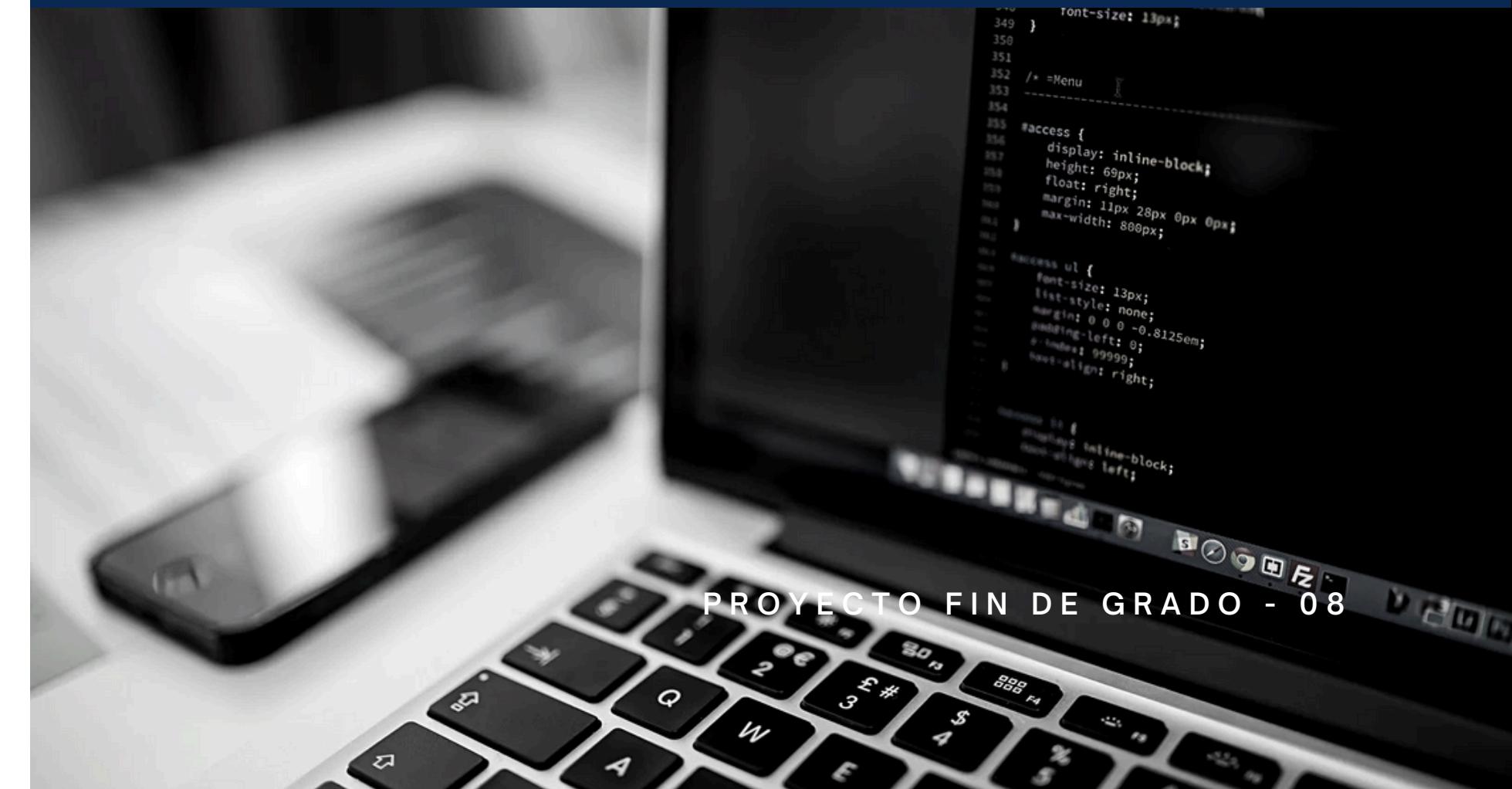
## 3 / REGRESION LINEAL MÚLTIPLE

- Sale.Date: Fecha de venta
- Vessel.Name: Nombre del barco
- Sale.Price: Precio de venta
- year.Built: Año en que el barco fue construido
- Age.at.Sale (Year of Sale – Year Built): Antigüedad del barco en el momento de su compra.
- Dead.Weight.Tons: Peso muerto
- Avg.Month.Balt: Promedio móvil de 1 año de los valores mensuales del índice Baltic Dry (BDI).

- Error Cuadrático Medio (MSE) 54.26
- R<sup>2</sup>: 0.95

Conclusión:

- Modelo Preciso: El alto R<sup>2</sup> indican que el modelo de regresión lineal es preciso y explica bien las variaciones en la variable target.
- Buen Desempeño: Estos valores sugieren que las predicciones del modelo son confiables y que la mayor parte de la variación en los precios de venta se puede atribuir a las variables que has incluido en el modelo.



# 3.1. / MODELO EN DETALLE: OLS

OLS Regression Results						
Dep. Variable:	Sale_Price	R-squared:	0.952			
Model:	OLS	Adj. R-squared:	0.946			
Method:	Least Squares	F-statistic:	165.7			
Date:	Thu, 19 Dec 2024	Prob (F-statistic):	1.65e-26			
Time:	20:42:11	Log-Likelihood:	-163.96			
No. Observations:	48	AIC:	339.9			
Df Residuals:	42	BIC:	351.2			
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	40.1337	14.454	2.777	0.008	10.963	69.304
Flag_fecha[T.1]	40.3582	7.744	5.212	0.000	24.731	55.986
Age_at_Sale	-3.4041	0.359	-9.477	0.000	-4.129	-2.679
Dead_Weight_Tons	0.2722	0.073	3.713	0.001	0.124	0.420
Avg_Month_Balt	0.0041	0.001	4.138	0.000	0.002	0.006
Age_at_Sale_Flag_fecha	-1.7267	0.410	-4.213	0.000	-2.554	-0.900
Omnibus:	23.596	Durbin-Watson:	2.039			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.011			
Skew:	-1.410	Prob(JB):	1.68e-10			
Kurtosis:	6.815	Cond. No.	1.02e+05			
...						
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
[2] The condition number is large, 1.02e+05. This might indicate that there are strong multicollinearity or other numerical problems.						

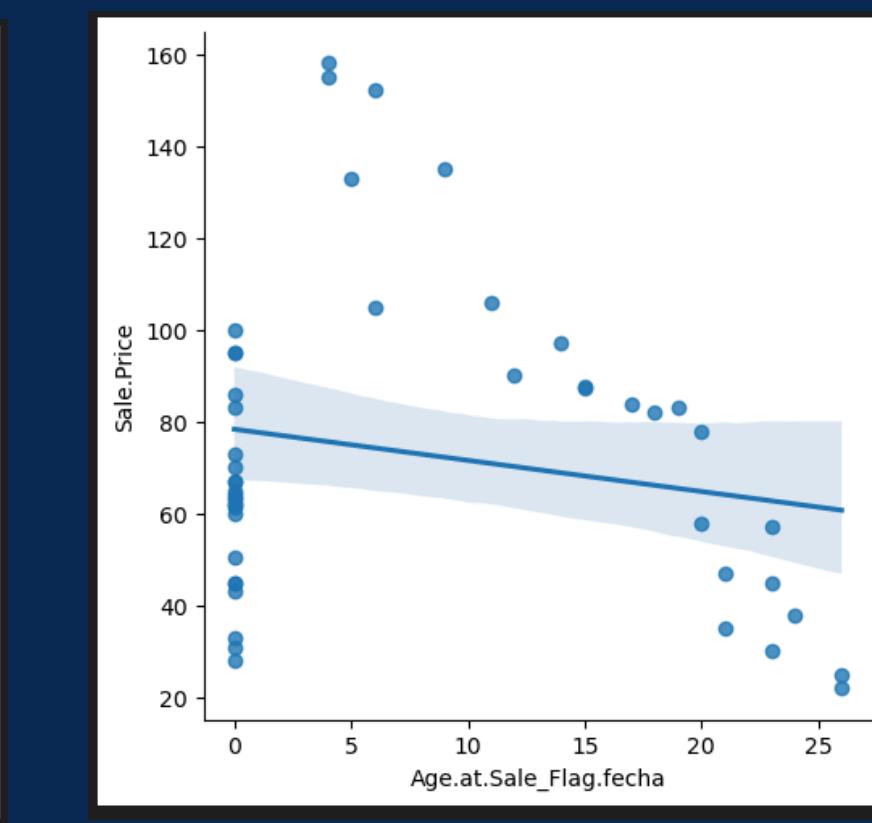
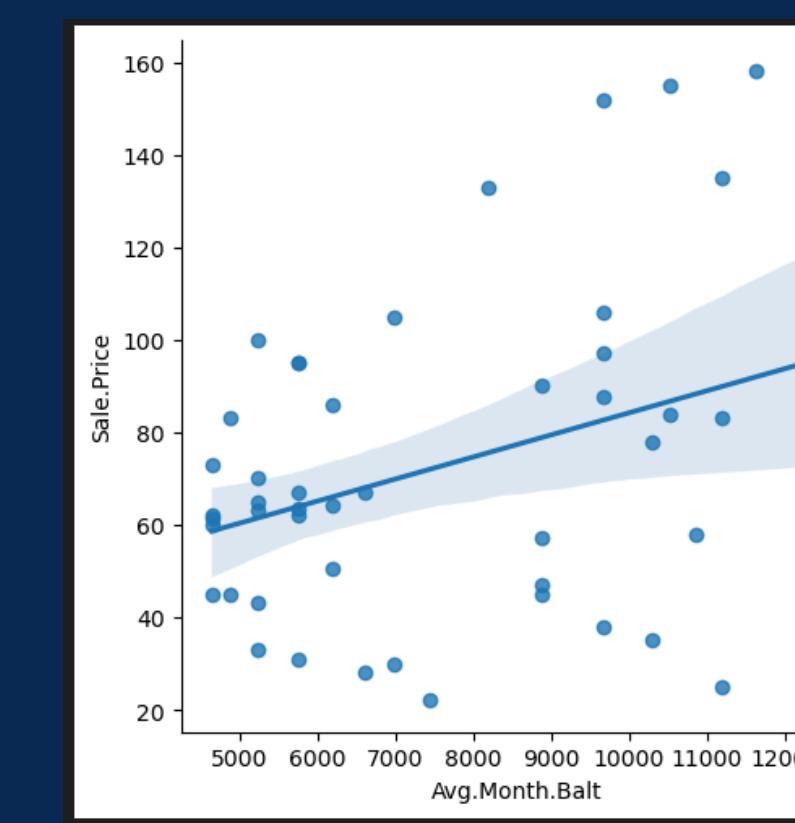
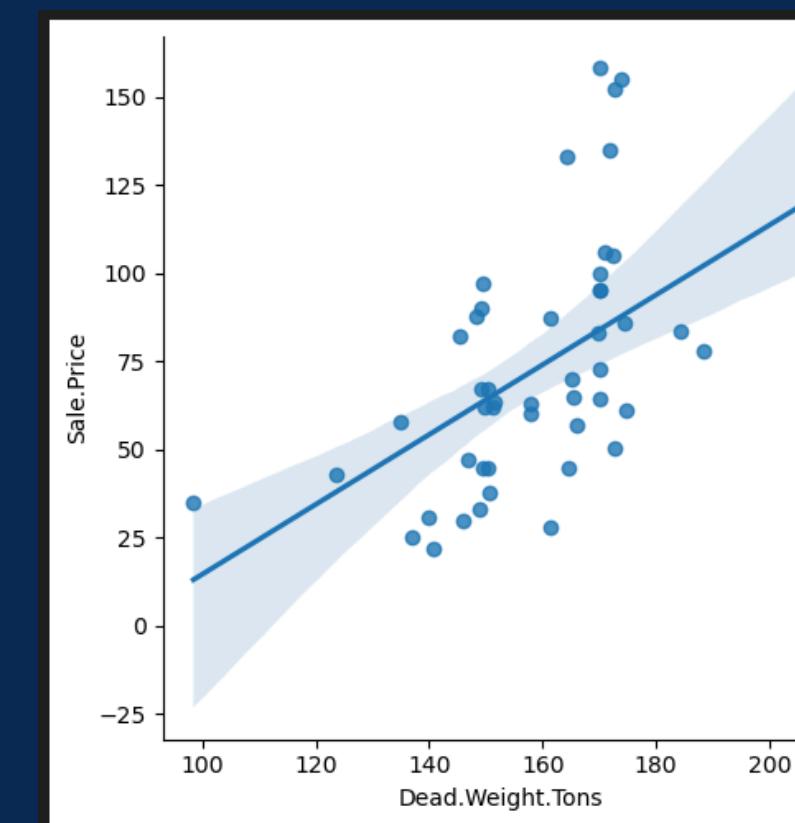
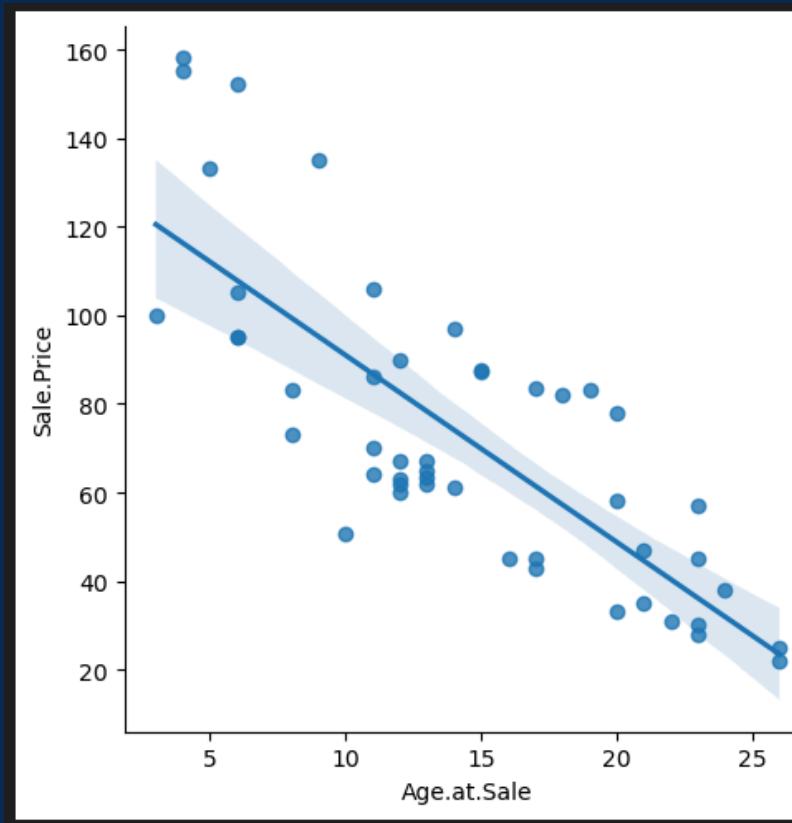
### 3.1 / MODELO EN DETALLE

```
*****Parameters*****
Intercept          40.133748
Flag_fecha[T.1]    40.358234
Age_at_Sale        -3.404078
Dead_Weight_Tons   0.272237
Avg_Month_Balt    0.004147
Age_at_Sale_Flag_fecha -1.726679
dtype: float64
*****P-Values*****
Intercept          8.171974e-03
Flag_fecha[T.1]    5.339060e-06
Age_at_Sale        5.439860e-12
Dead_Weight_Tons   5.966289e-04
Avg_Month_Balt    1.646078e-04
Age_at_Sale_Flag_fecha 1.305537e-04
dtype: float64
*****Standard Errors*****
Intercept          14.454446
Flag_fecha[T.1]    7.743667
Age_at_Sale        0.359193
Dead_Weight_Tons   0.073313
Avg_Month_Balt    0.001002
Age_at_Sale_Flag_fecha 0.409858
dtype: float64
*****Confidence Interval*****
...
Age_at_Sale        -0.119354
Dead_Weight_Tons   -0.002696
Avg_Month_Balt    0.000034
Age_at_Sale_Flag_fecha 0.167983
```

*Output is truncated. View as a [scrollable element](#) or open in a [text editor](#). Adjust cell output [settings](#)...*

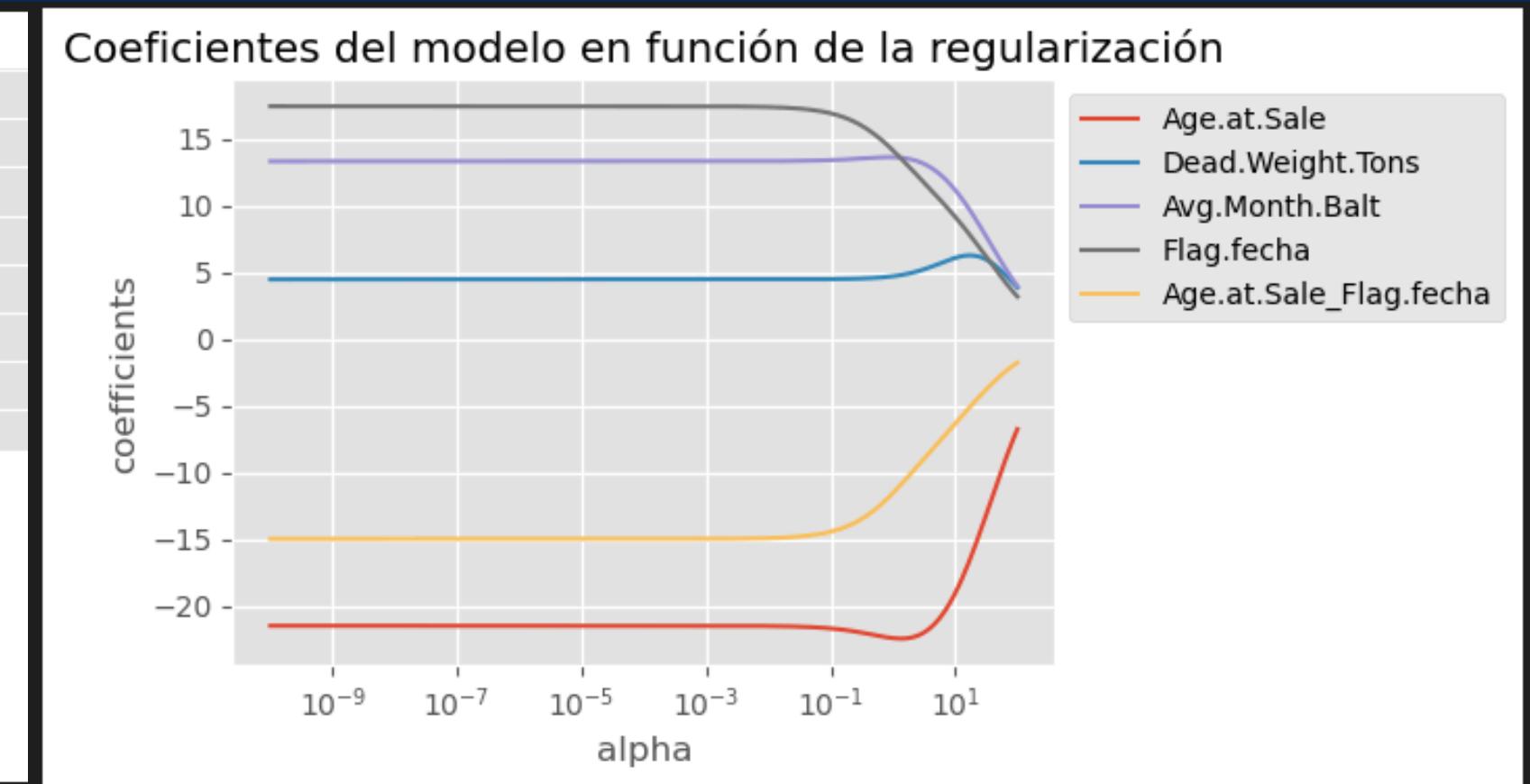
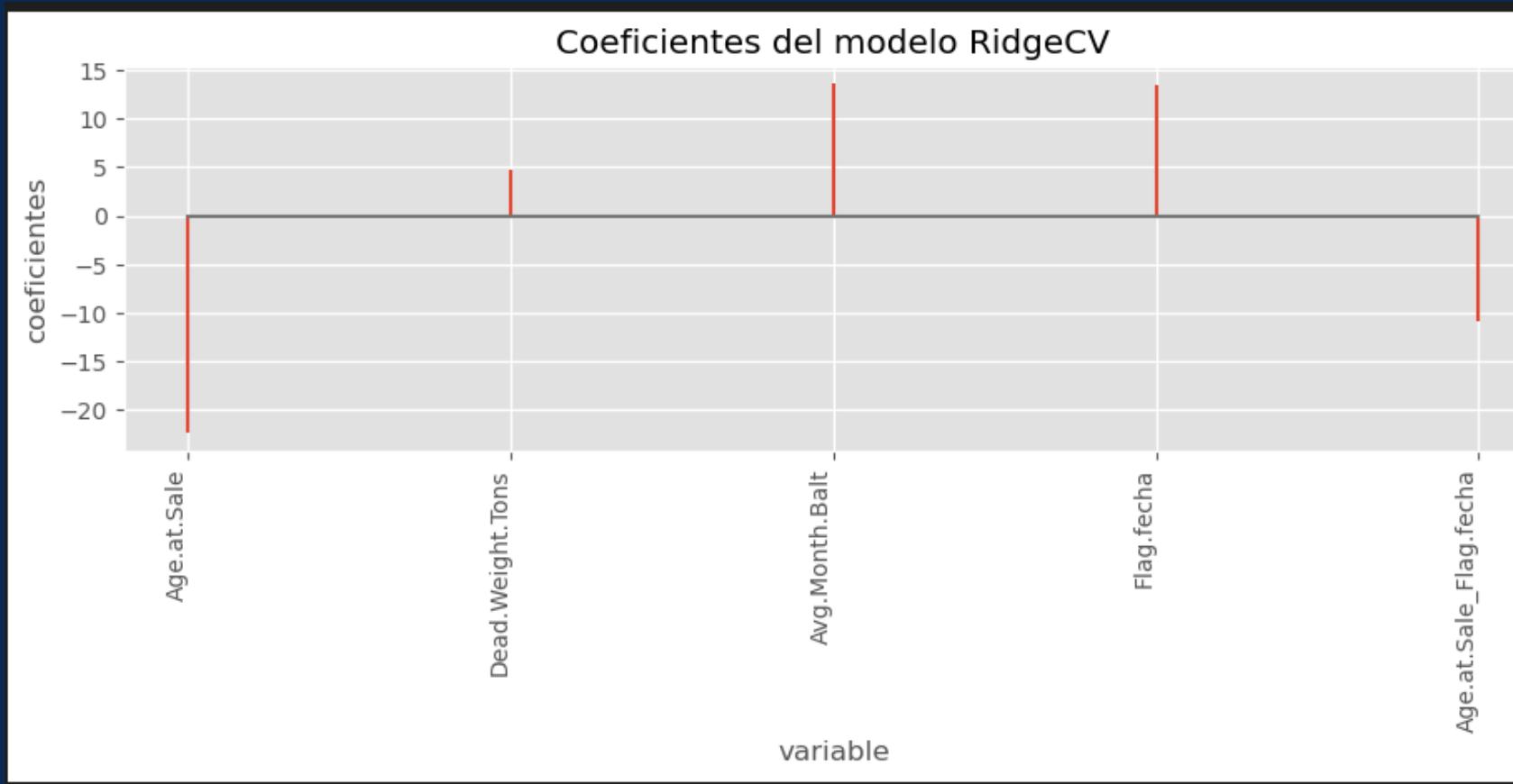
- Calidad del Ajuste del Modelo:
  - El alto R<sup>2</sup> y los coeficientes significativos indican que el modelo es robusto y explica bien la variabilidad en los precios de venta.
- Importancia de las Variables:
  - Todas las variables incluidas tienen un impacto significativo, con Flag\_fecha y Age\_at\_Sale siendo especialmente influyentes.
  - Las interacciones y la antigüedad del barco tienen impactos notables, lo que puede ayudar a ajustar mejor los precios.
- Revisión de Multicolinealidad:
  - El alto número de condición (Cond. No. 1.02e+05) **sugiere posible multicolinealidad**, probablemente debido a la inclusión de la interacción de la fecha con el año de venta del barco.

## 3.2 / VISUALIZACIÓN DE BANDA DE CONFIANZA



# 4. REGULARIZACIÓN

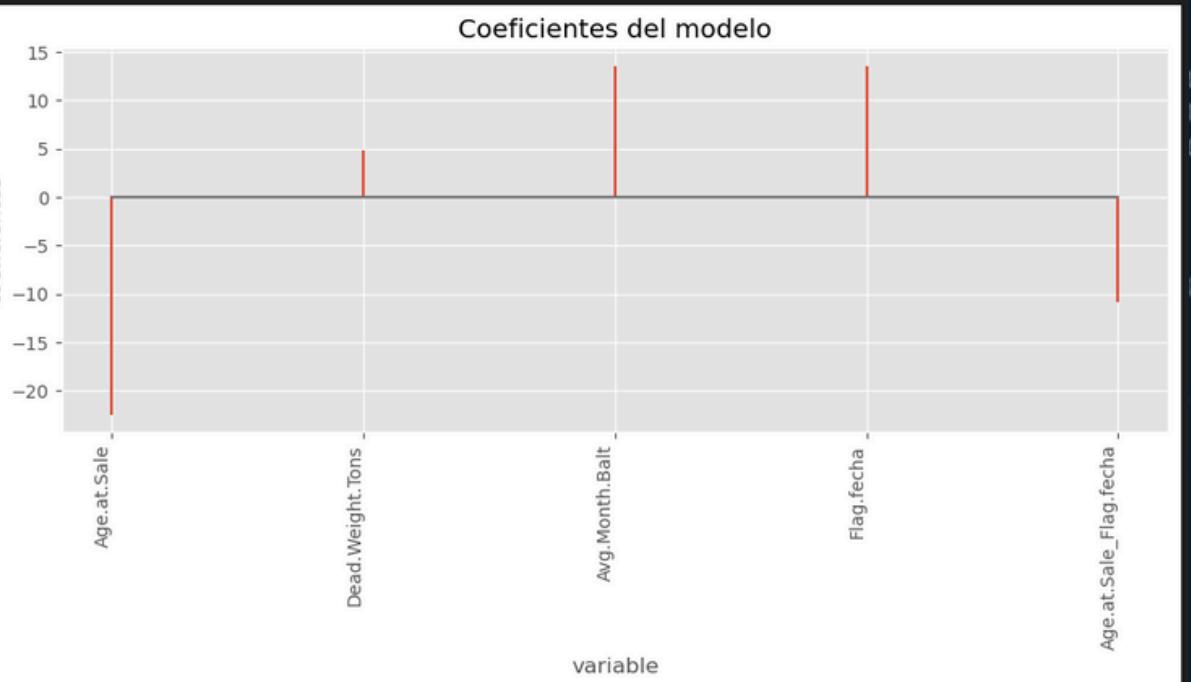
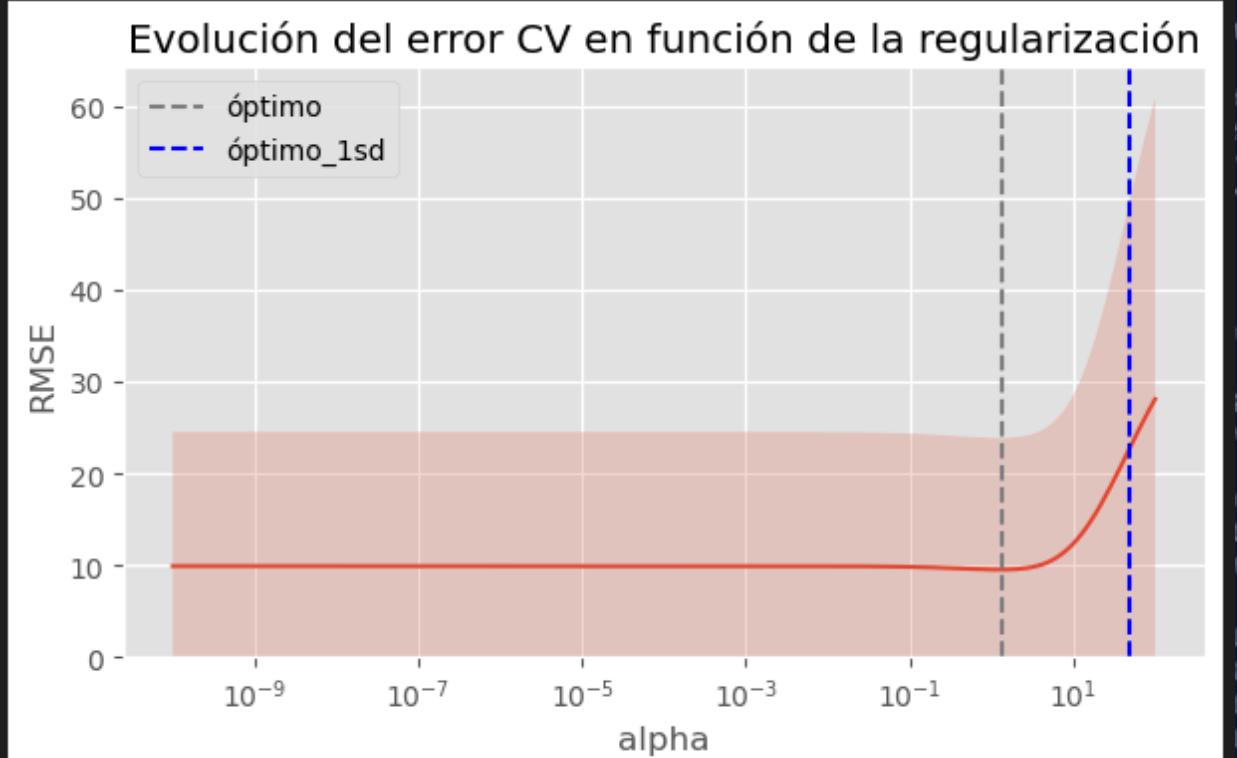
## 4.1. REGRESIÓN RIDGE



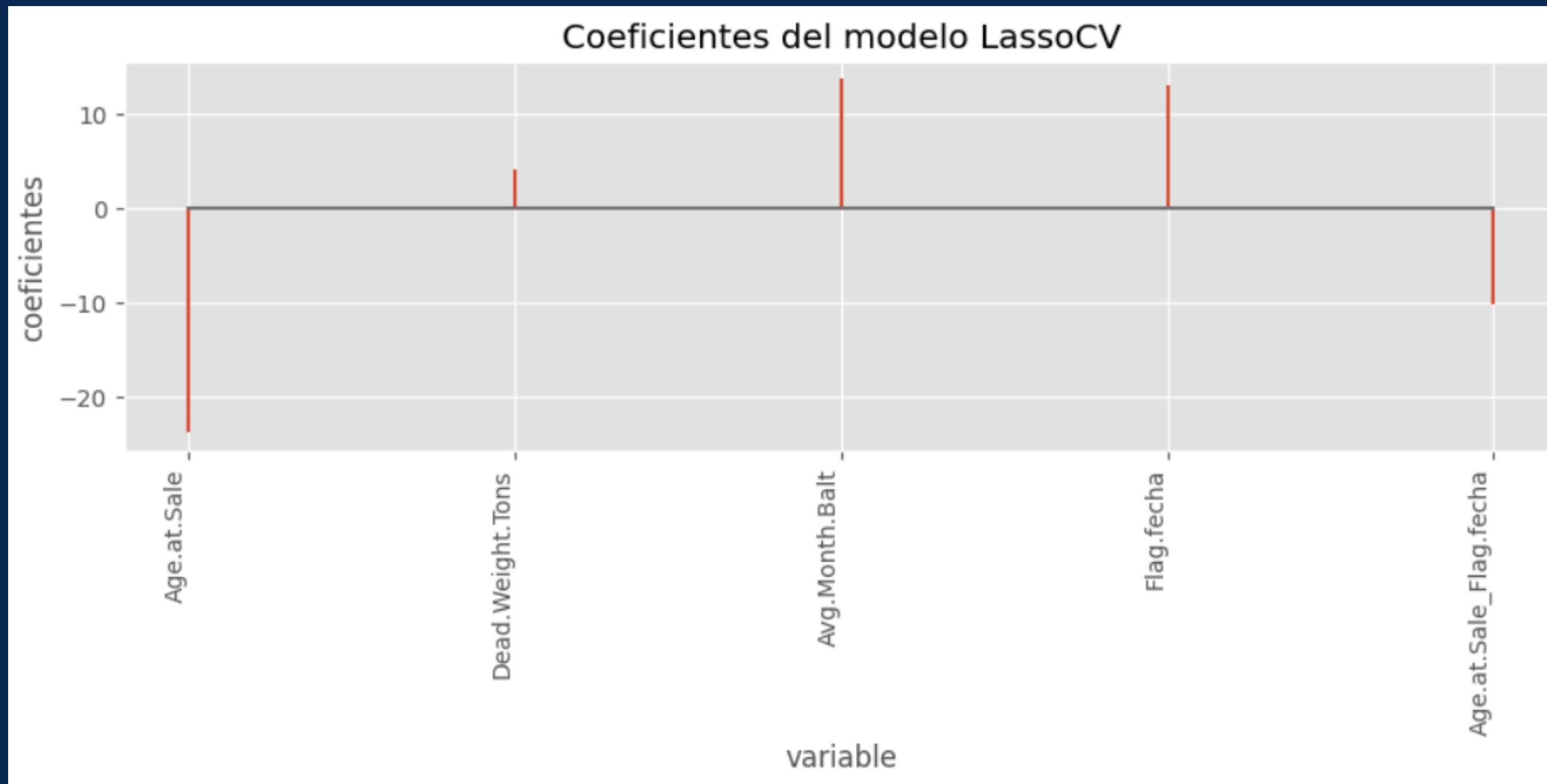
- Los coeficientes no se ven afectados por la regularización hasta  $\alpha = 1/10$ . Esta situación puede ocurrir debido a que las variables del modelo tienen un impacto significativo y constante en la predicción, independientemente del nivel de regularización que se aplique. Es posible que estas variables sean fundamentales para explicar la relación entre las características y la variable objetivo, por lo que el modelo no encuentra necesario ajustar sus coeficientes, incluso cuando se introduce regularización.
- Decaimiento después de 10: A partir de un valor de  $\alpha$  de aproximadamente  $1/10$ , algunos coeficientes empiezan a disminuir rápidamente (por ejemplo, la línea correspondiente a 'Age.at.Sale' en rojo y otras), indicando que la penalización de la regularización Ridge comienza a tener un impacto importante. Estas características empiezan a ser consideradas menos relevantes para el modelo bajo una regularización más estricta.
- Impactos contrarios en variables: Algunas líneas, como 'Age.at.Sale\_Flag.fecha', muestran un cambio hacia un crecimiento positivo tras un decaimiento inicial. Esto puede reflejar que la interacción o combinación de características se ajusta a niveles altos del coeficiente, haciendo que el modelo revalorice su importancia relativa.

# 1/ RESULTADOS

- Mejor valor de alpha encontrado: 1.35
- El error (RMSE) de test es: 7.50
- Evaluación del Modelo: El RMSE proporciona una manera fácil de entender la precisión de tu modelo en unidades reales de la variable dependiente. Un RMSE de 7.50 sugiere que el modelo tiene un buen desempeño, ya que los errores de predicción son relativamente pequeños.
- Contexto: Comparar el RMSE con la escala de tus datos puede darte una mejor idea de qué tan buenos son estos errores. Si los precios de venta varían ampliamente, un RMSE de 7.50 (comparado con el 54.2 del modelo sin ajustar) indica una mejora en la capacidad de predicción del modelo.



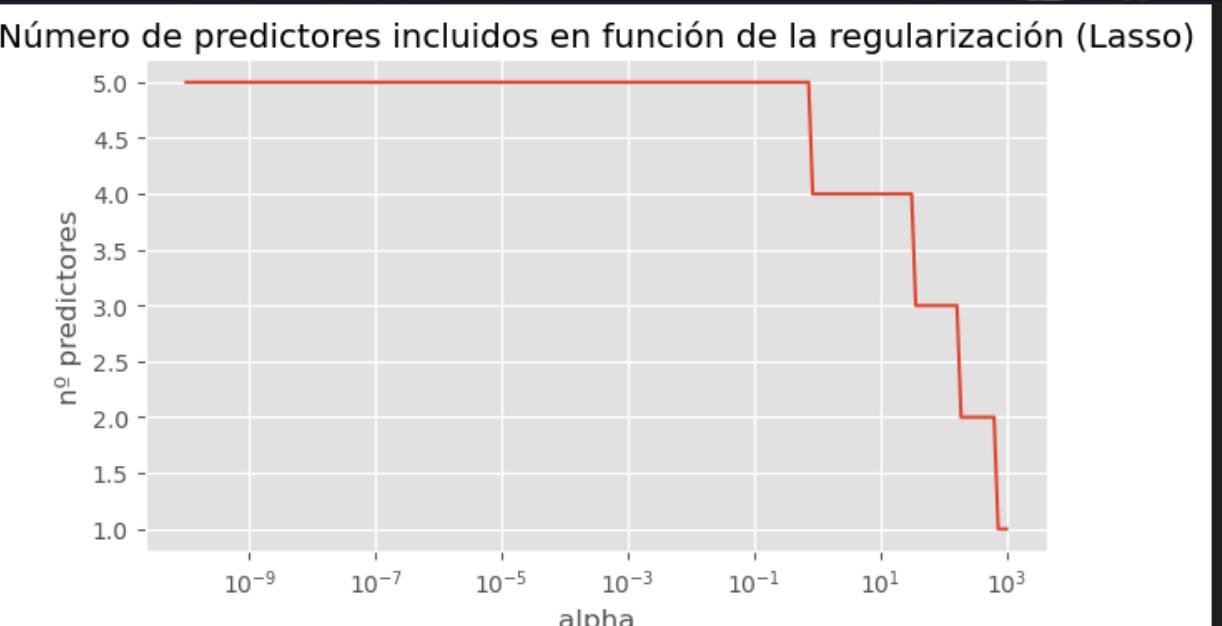
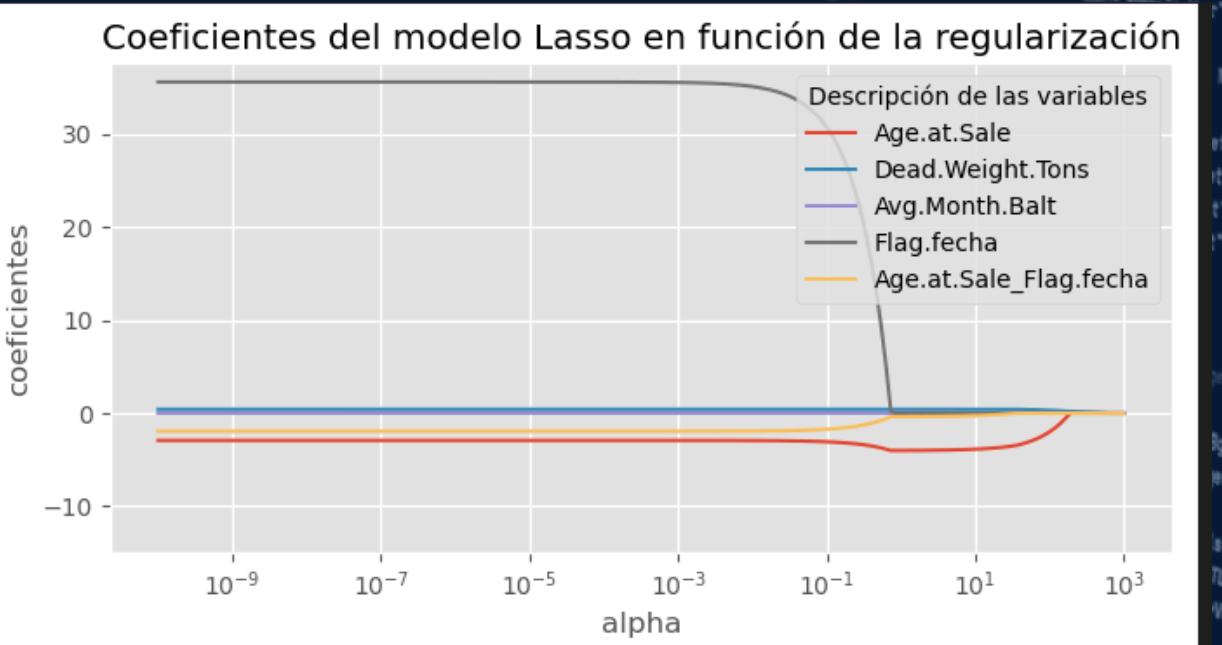
## 4 . 2 . R E G R E S I Ó N L A S S O



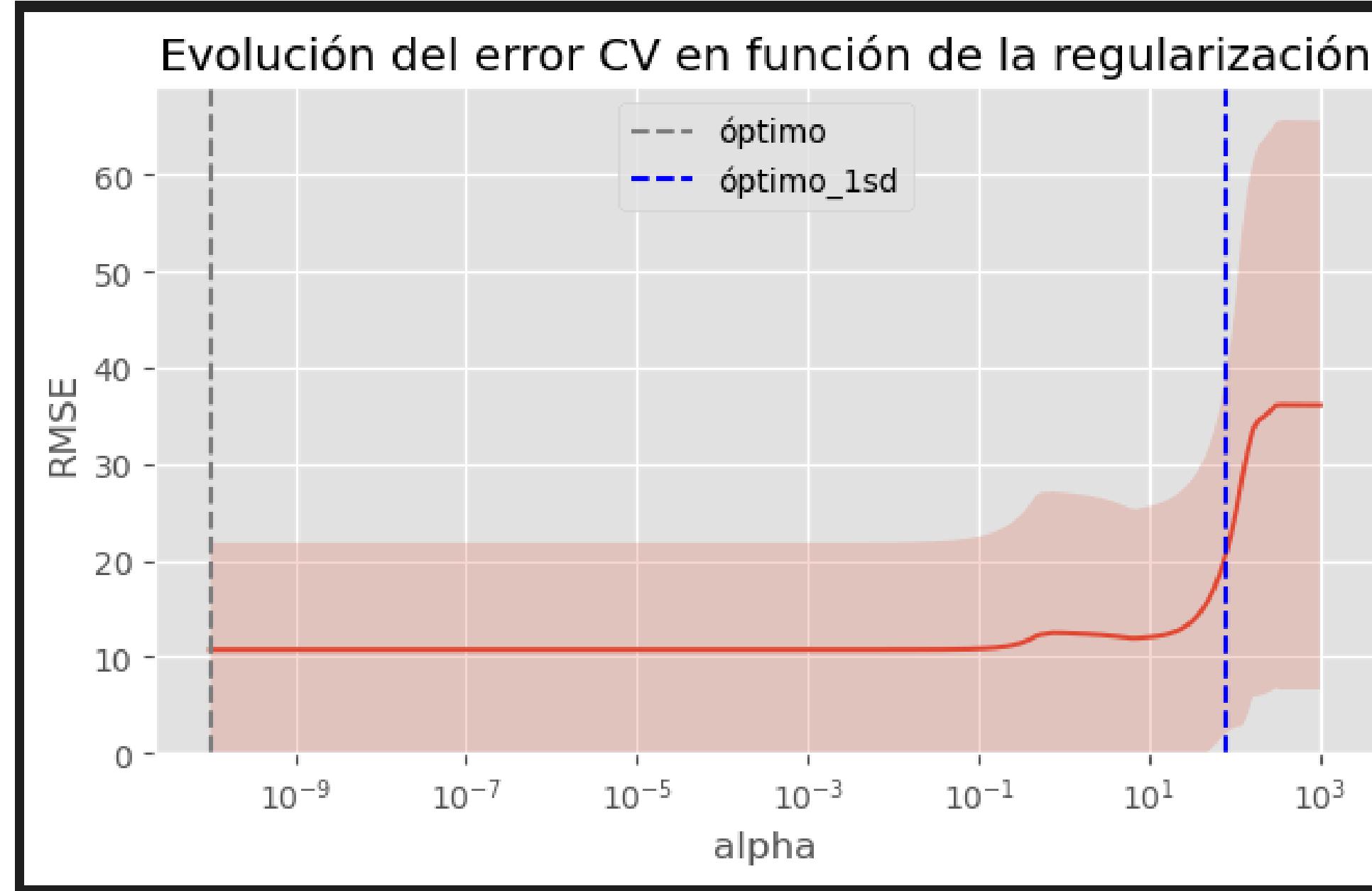
- La regresión Lasso, al tener una función de costo más estricta, suele eliminar variables predictoras si no aportan al análisis. Además, suele restar importancia a algunas variables como en este caso a Flag.fecha

# 1 / RESULTADOS

- Por lo general, las variables predictoras tienen coeficientes pequeños a excepción de Flag.fecha, el cual ve una caída drástica en su parámetro estimado luego de  $\alpha=1e-1$ . Hasta poro después de ese valor, el modelo Lasso está manteniendo consistentemente 5 predictores en el modelo. Es probable que estos predictores sean considerados importantes o relevantes para el modelo, ya que sus coeficientes no se reducen a cero incluso con una regularización más fuerte. Pero después de se reduce la cantidad de predictores progresivamente.
- Cuando el valor de  $\alpha$  aumenta significativamente, el modelo Lasso aplica una regularización más fuerte. Esta penalización adicional puede hacer que el modelo reduzca drásticamente el número de predictores incluidos, llegando incluso a eliminar casi todos los predictores, dejando únicamente 1 predictor en el modelo cuando  $\alpha$  se acerca a 1000.

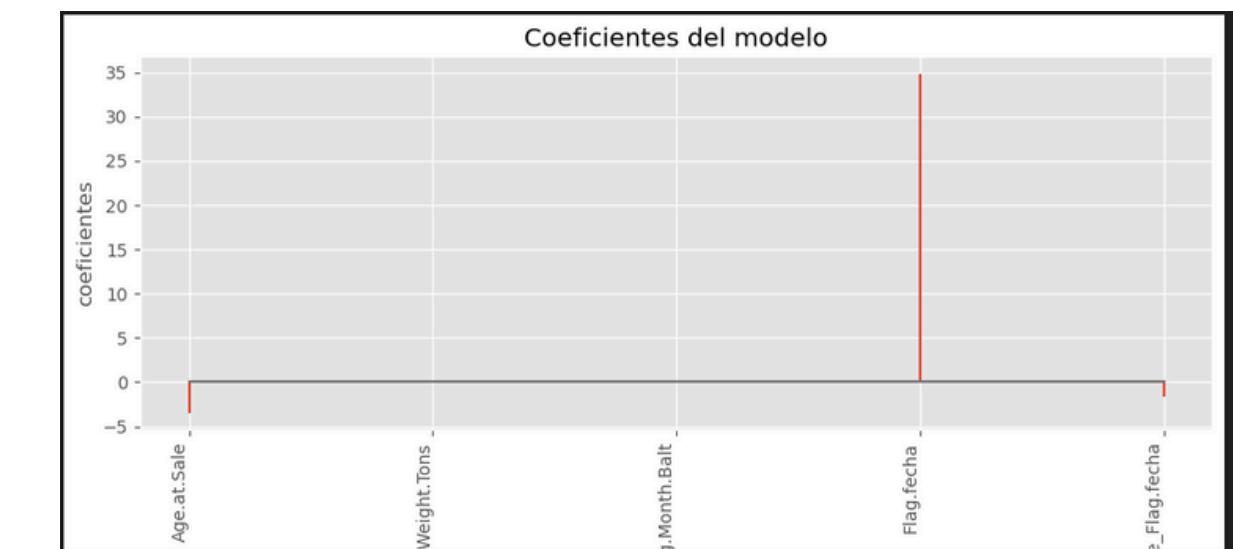


# MODELO LASSO AJUSTADO CON ÉXITO.



- Mejor valor de alpha encontrado: 1e-10
- Valor de alpha óptimo + 1SD: 77.52
- Se entrena de nuevo el modelo, esta vez empleando el mayor valor de alpha cuyo error está a menos de una desviación típica del mínimo encontrado en la validación cruzada.

```
array([ 61.7971572 , 103.78207062, 89.79563611, 62.40366925,
       49.56273936, 34.30353406, 57.50551152, 44.36080414,
       89.79563611, 27.20111576])
```



# COMPARACIÓN DE MODELOS



- Se obtienen mejores predicciones aplicando regularizaciones. Ambos con un error considerablemente menor al del modelo OLS. Ya que todos los modelos conservan los 5 predictores, bajo el único criterio de comparar el error cuadrático medio, la conclusión sería optar por el modelo LASSO.

# COMPARACIÓN DE MODELOS

Variable	OLS	Ridge	LASSO
Age.at.Sale	-21.452	-22.416	-3.421
Dead.Weight.Tons	4.457	4.783	0.259
Avg.Month.Balt	13.295	13.563	0.006
Flag.fecha	17.410	13.509	34.836
Age.at.Sale_Flag.fecha	-14.957	-10.890	-1.670

- Sin embargo, al revisar como se está redistribuyendo la importancia de las variables en los distintos métodos, se nota que se le da mucha importancia a la dummy de fecha, lo cual podría generar problemas para explicar a un público no técnico el modelo y a futuro (porque en los nuevos datos la dummy siempre tomará el valor de 1). En ese sentido, la decisión es quedarse con el modelo Ridge.

MUCHAS GRACIAS

GRUPO 4  
BUSINESS ANALYTICS INTRODUCTION

```
render() {
  return (
    <React.Fragment>
      <div className="py-5">
        <div className="container">
          <Title name="our" title="product">
          <div className="row">
            <ProductConsumers>
              {(value) => {
                |   |   |   console.log(value)
              }}
            </ProductConsumers>
          </div>
        </div>
      </React.Fragment>
```