Máster Universitario en Ciencia de Datos
CUNEF Universidad
Aplicaciones profesionales de la Ciencia de Datos

Trabajo realizado por:

Pablo Oceja Campanero

Profesor de la asignatura:

Andre Alves Portela Santos

Introducción:

En el siguiente trabajo, se va a analizar como utilizar los diferentes modelos de regresión propuestos por el docente para realizar predicciones acerca de los rendimientos de 119 fondos de inversión diferentes.

El principal objetivo del trabajo será utilizar los, también proporcionados por el docente, factores de riesgo más comunes que se suelen encontrar cuando se analizan los rendimientos de diferentes fondos de inversión, siendo estos los siguientes:

- **MKT-RF**: Representa el exceso de rendimiento del mercado sobre la tasa libre de riesgo.
- **SMB (Small Minus Big)**: Captura la prima relacionada con el tamaño de las empresas.
- **HML (High Minus Low)**: Refleja la prima de valor, es decir, la diferencia en rendimientos entre empresas de alto y bajo valor.
- **RMW (Robust Minus Weak)**: Mide la rentabilidad comparando empresas muy rentables con otras menos rentables.
- **CMA (Conservative Minus Aggressive)**: Se enfoca en la estrategia de inversión, diferenciando entre empresas conservadoras y agresivas.
- **Momentum (UMD: Up Minus Down)**: Captura la tendencia de los activos a continuar con su dirección de mercado reciente.

Una vez sabemos los diferentes factores de riesgo que van a influir en los rendimientos del fondo a lo largo del trabajo, hemos de hablar de los modelos de regresión que se van a estudiar, siendo estos los siguientes:

- **Regresión Lineal Ordinaria (OLS)**: Encuentra la línea que mejor ajusta los datos minimizando el error cuadrado.
- **Lasso**: Agrega una penalización que reduce algunos coeficientes a cero, realizando selección de variables.
- **Ridge**: Añade una penalización para evitar sobreajuste, disminuyendo el tamaño de todos los coeficientes.
- **Elastic Net**: Combina las penalizaciones de Lasso y Ridge, equilibrando selección de variables y regularización.

Encontramos entonces como el análisis va a depender de un dataset con 240 registros, correspondiendo cada registro a una fecha recogida por la columna *“Date”.*

1. Análisis de los datos

Previó al análisis de los rendimientos en función de los factores, hay que realizar un análisis exploratorio de los datos, con el fin de realizar las transformaciones necesarias para poder trabajar con los datos de forma correcta.

La primera acción, tras la carga de los datos, es verificar las dimensiones de los datasets caso de estudio, en este caso se va a trabajar con dos datasets a los que he denominado como “rendimientos” y “factores” siendo el dataset de rendimientos el que recoge los rendimientos para cada fecha de los diferentes fondos de inversión y el dataset factores el que recoge los factores de riesgo anteriormente mencionados siendo sus dimensiones de (240, 120) para rendimientos y (240, 7) el de factores siendo columnas y filas respectivamente.

También era necesario estudiar si existían valores nulos o faltantes, en este caso parece que los datos se encuentran limpios y no es necesario imputar valores.

La transformación que si es necesaria, debido a como se nos entregan los datos, es la de poner los datos de los factores de riesgo en términos proporcionales ya que se encuentran en términos porcentuales, a diferencia del dataset de rendimientos. Podría realizarse también la transformación de los rendimientos en términos porcentuales, pero al ser la variable caso de estudio veo más adecuado transformar los factores de riesgo.

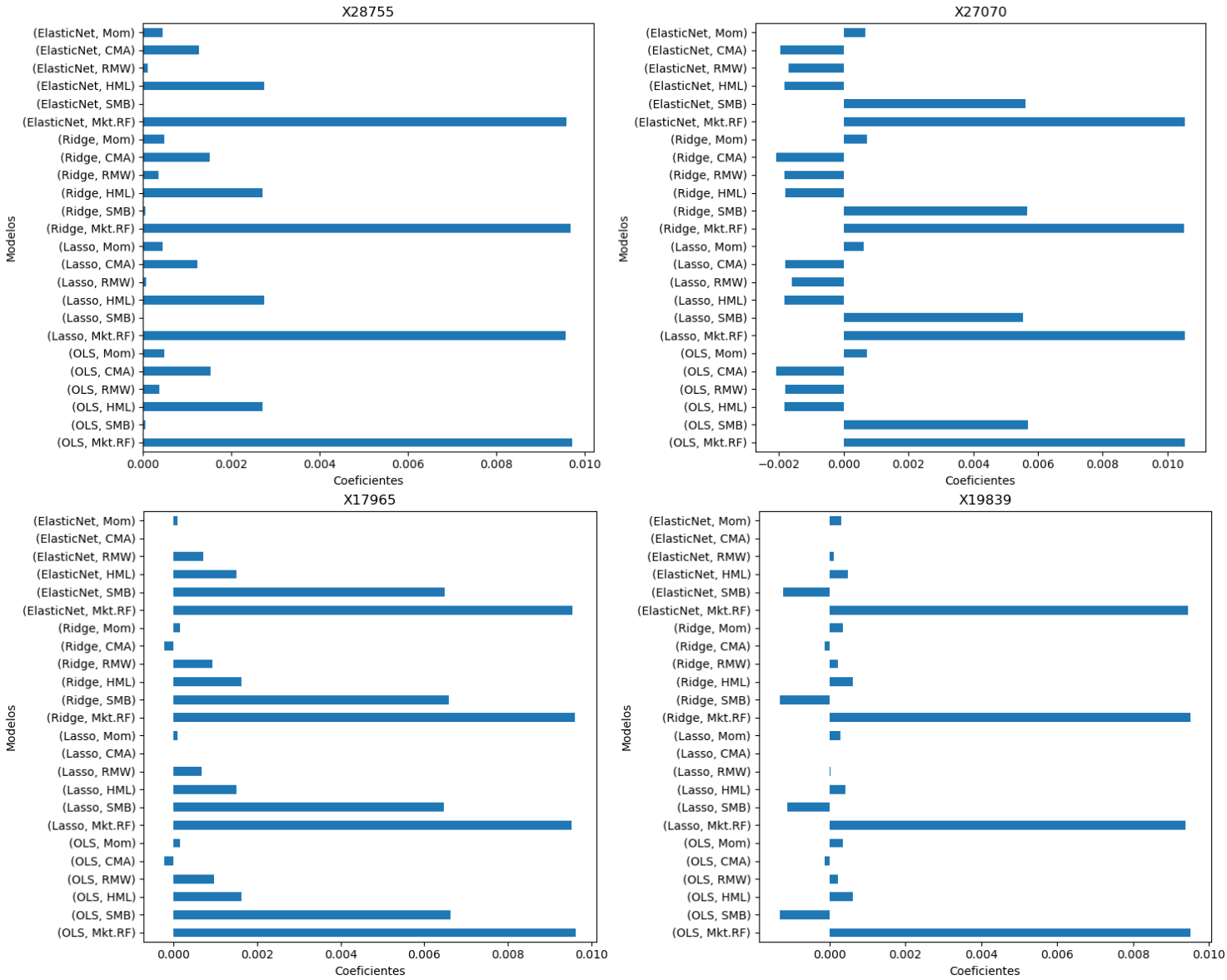
Una vez se encuentran en la misma escala, queda realizar una última transformación que es la estandarización de los datos de las variables explicativas. Esto se debe a que al realizar modelos de regresión los datos deben de estar estandarizados para poder trabajar correctamente con ellos. A continuación se muestra una tabla que indica el por qué es necesaria la estandarización de los valores según el modelo que estamos realizando:

Modelo	¿Necesita estandarización?	Razón
Regresión Lineal (OLS)	Recomendable	Facilita la interpretación y comparación de coeficientes.
Lasso	Sí	La penalización L1 es sensible a la escala de las variables.
Ridge	Sí	La penalización L2 es sensible a la escala de las variables.
Elastic Net	Sí	Combina L1 y L2, por lo que requiere estandarización.

Al haber realizado las transformaciones mencionadas en el dataset de “factores” se procede a guardarlo para posteriormente trabajar con el estando ya transformado.

2. Regresión con distintos modelos

Para cada uno de los fondos, se nos pide entrenar un modelo para cada uno de los cuatro modelos diferentes de regresión y posteriormente comparar los coeficientes resultantes de cada modelo y que tan diferentes son estos coeficientes respecto a los diferentes factores y la relevancia de estos últimos en la regresión.

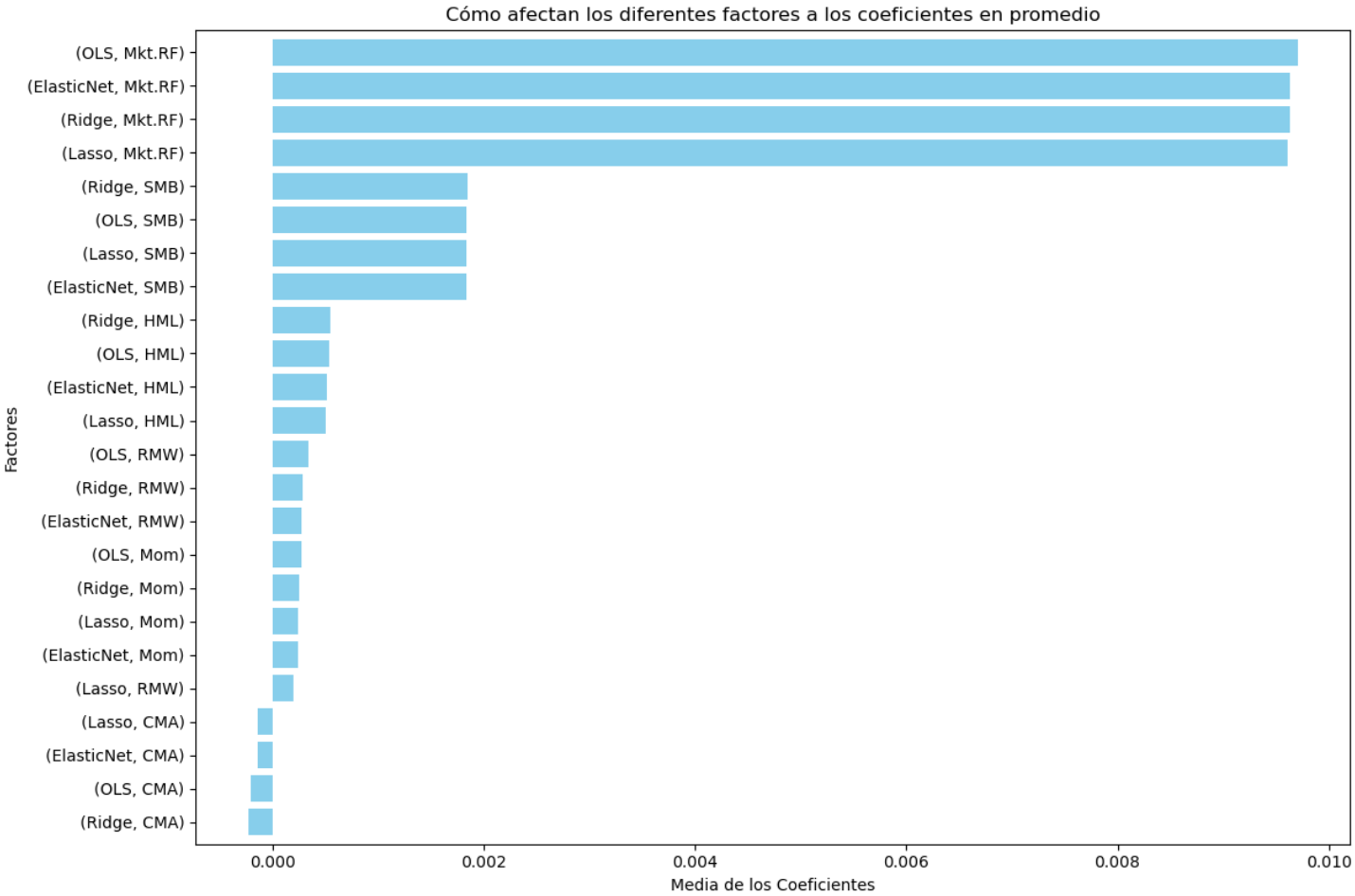


En las imágenes anteriores se pueden observar como afectan los diferentes factores de riesgo por cada uno de los modelos de regresión para 4 de los 119 fondos escogidos de forma aleatoria, en estos 4 gráficos podemos ya observar como afectan los diferentes factores de riesgo a los modelos, siendo que se comportan de manera muy similar entre sí. Teniendo en cuenta que los resultados de los modelos han sido los siguientes:

- OLS - MSE: 0.00021832405353841152, R^2: 0.9171874929509088
- Lasso - MSE: 0.0002194489426253634, R^2: 0.916827292288419
- Ridge - MSE: 0.00021842942592873352, R^2: 0.9171591960306581
- Elastic Net - MSE: 0.00021943438791276236, R^2: 0.9168327557071497

tiene bastante sentido debido a las similitudes en los resultados para los diferentes modelos de regresión.

Además, se ha realizado un gráfico que nos muestra cómo afectan los diferentes factores a los coeficientes promedio de los modelos en todos los fondos, y qué factores tienen un mayor o menor impacto en el modelo.



Podemos observar como el factor que más impacto tiene en los modelos es el **Mkt.Rf** seguido de **SMB** siendo este primero bastante más representativo. El resto de los factores tienen un impacto más reducido, por debajo del 0.001 de impacto en media e incluso encontramos como el factor **CMA** incluso tiene un impacto negativo en los modelos de regresión.

Como conclusiones, factor **Mkt.RF** es el que tiene el mayor impacto en los modelos de regresión, destacándose como el más representativo para explicar los resultados. Le sigue **SMB**, que aunque tiene un impacto significativo, es considerablemente menor en comparación con Mkt.RF. El resto de los factores, como **HML**, **RMW** y **Mom**, muestran un impacto mucho más reducido, con medias por debajo de 0.001. Además, el factor **CMA** presenta un impacto negativo en los modelos, lo que sugiere que este factor podría estar contribuyendo de manera inversa a la predicción en comparación con los demás. Esto indica que, en promedio, los modelos priorizan el exceso de retorno del mercado y el tamaño de la empresa como las variables más influyentes, mientras que los otros factores tienen un peso marginal o incluso contraproducente.

3. División de los datos

En esta parte del trabajo se va a realizar la división de los datos en train y test, se nos propone que sea 80-20 o 70-30 de train y test respectivamente, en mi caso decidí hacerlo de las dos formas para ver si existía diferencia en los modelos al entrenarlos con muestras de diferentes tamaños y cuales eran estas diferencias si es que las había.

En cuanto a los resultados para train y test cuando la muestra se dividía en 80-20 obtuve los siguientes resultados:

	R^2	
	TRAIN	TEST
OLS	0.9175	0.9158
LASSO	0.9163	0.9141
RIDGE	0.9175	0.9158
ELASTICNET	0.9162	0.9139

Podemos observar como el rendimiento en los conjuntos de entrenamiento son ligeramente mejores que en el conjunto de prueba, esto es normal ya que los modelos se ajustan específicamente a los datos de entrenamiento.
Esto último, también es indicativo de que no se esta realizando sobreajuste ya que en caso de que las diferencias entre el R^2 del conjunto de train y el conjunto de test fueran muy grandes se estaría realizando sobreajuste al adaptarse muy bien a los datos de train pero mal a los de test y viceversa.

En el caso de una división 70-30, se obtuvieron los siguientes resultados:

	R^2	
	TRAIN	TEST
OLS	0.9179	0.9126
LASSO	0.9011	0.8976
RIDGE	0.9179	0.9126
ELASTICNET	0.9161	0.9114

Vemos como los resultados son muy similares, pero observamos que la diferencia de R^2, aunque sigue dentro de lo aceptable, es mayor que en la división 80-20. Esto puede deberse a datos que ahora “sesgan” el conjunto de test y ya no sesgan el conjunto del train dentro de ese 10% que cambia de una comparación a la otra.

Por todo esto, podemos decir que los cuatro modelos tienen una buena capacidad de generalización y no se encuentra sobreajuste en su entrenamiento.

4. Análisis de inversión

En este último apartado del proyecto se va a realizar un análisis de los resultados obtenidos de los diferentes modelos realizados. Se nos pide que obtengamos los fondos, respecto a los rendimientos obtenidos, según su pertenencia al "top decile" o al "bottom decile", es decir, los fondos pertenecientes al 10% mejor y peor de entre los 119 fondos analizados.

Al igual que en el apartado anterior, se ha realizado tanto para una división del 80-20 como del 70-30 para de esta forma tener diferentes puntos de vista a la hora de seleccionar los mejores fondos caso de estudio.

En cuanto a los fondos pertenecientes al "top decile" y "bottom decile" para la división 80-20 encontramos:

		OLS											
TRAIN	TOP DECILE	X24795	X8313	X30745	X3563	X12510	X7365	X27124	X11042	X7353	X29211	X8316	X12504
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009
TEST	TOP DECILE	X24795	X16970	X17583	X10637	X30745	X11042	X11225	X24797	X26995	X14571	X30931	X20359
	BOTTOM DECILE	X15109	X26333	X6718	X7145	X27498	X26334	X17666	X32452	X27521	X17985	X26981	X22632

		LASSO											
TRAIN	TOP DECILE	X24795	X8313	X30745	X3563	X12510	X7365	X27124	X11042	X7353	X29211	X8316	X12504
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009
TEST	TOP DECILE	X24795	X8313	X30745	X3563	X12510	X7365	X27124	X11042	X7353	X29211	X8316	X12504
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009

		RIDGE											
TRAIN	TOP DECILE	X24795	X8313	X30745	X3563	X12510	X7365	X27124	X11042	X7353	X29211	X8316	X12504
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009
TEST	TOP DECILE	X24795	X16970	X17583	X10637	X30745	X11042	X11225	X24797	X26995	X14571	X30931	X20359
	BOTTOM DECILE	X15109	X26333	X6718	X7145	X27498	X26334	X17666	X32452	X27521	X17985	X26981	X22632

		ELASTIC NET											
TRAIN	TOP DECILE	X24795	X8313	X30745	X3563	X12510	X7365	X27124	X11042	X7353	X29211	X8316	X12504
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009
TEST	TOP DECILE	X24795	X30745	X3563	X7365	X27124	X11042	X7353	X29211	X8313	X6188	X20359	X12090
	BOTTOM DECILE	X27521	X17579	X11865	X4976	X27498	X6718	X11838	X3720	X17985	X34452	X5009	X22632

Se Observar como muchos de los fondos, sobre todo los del "top decile", son similares entre los 4 modelos, siendo que el cambio se encuentra en el test donde los fondos pertenecientes al top decile y bottom decile tiene más cambios, aunque muchos de estos fondos son comunes entre el top y el bottom decile tanto del train como del test, como pueden ser el X24795 para el top decile y el X17985 para el bottom decile.

En cuanto a la división 70-30, encontramos algunos cambios respecto a la división anterior siendo que obtenemos los siguientes resultados:

		OLS											
TRAIN	TOP DECILE	X24795	X30745	X11042	X12510	X7365	X8313	X7353	X3943	X3563	X24797	X17583	X20359
	BOTTOM DECILE	X28143	X11865	X17579	X32453	X10987	X27498	X3720	X32452	X11838	X10637	X22632	X5009
TEST	TOP DECILE	X24795	X30745	X11225	X20359	X11042	X19609	X16970	X7365	X7353	X3563	X27126	X31270
	BOTTOM DECILE	X28143	X27498	X17666	X27517	X8030	X7145	X26333	X6718	X26334	X17985	X27521	X22632

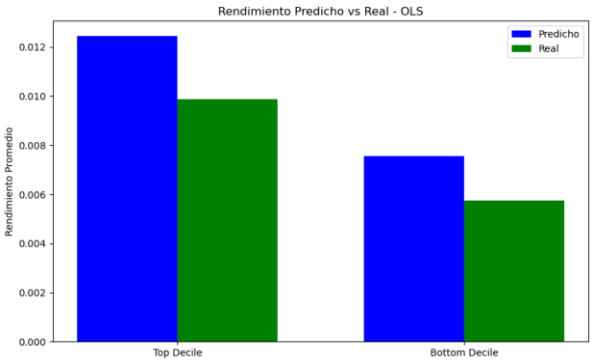
		LASSO											
TRAIN	TOP DECILE	X24795	X30745	X11042	X12510	X7365	X8313	X7353	X3943	X3563	X24797	X17583	X20359
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009
TEST	TOP DECILE	X24795	X30745	X11042	X12510	X7365	X8313	X7353	X3943	X3563	X24797	X17583	X20359
	BOTTOM DECILE	X28143	X11865	X17579	X32453	X10987	X27498	X3720	X32452	X11838	X10637	X22632	X5009

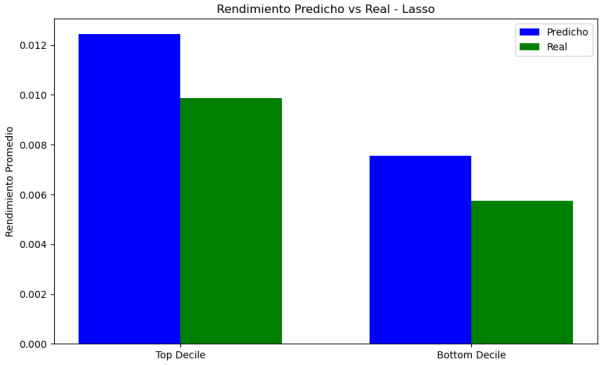
		RIDGE											
TRAIN	TOP DECILE	X24795	X30745	X11042	X12510	X7365	X8313	X7353	X3943	X3563	X24797	X17583	X20359
	BOTTOM DECILE	X28143	X11865	X17579	X32453	X10987	X27498	X3720	X32452	X11838	X10637	X22632	X5009
TEST	TOP DECILE	X24795	X30745	X11225	X20359	X11042	X19609	X16970	X7365	X7353	X3563	X27126	X31270
	BOTTOM DECILE	X32452	X27498	X17666	X27517	X8030	X7145	X26333	X6718	X26334	X17985	X27521	X22632

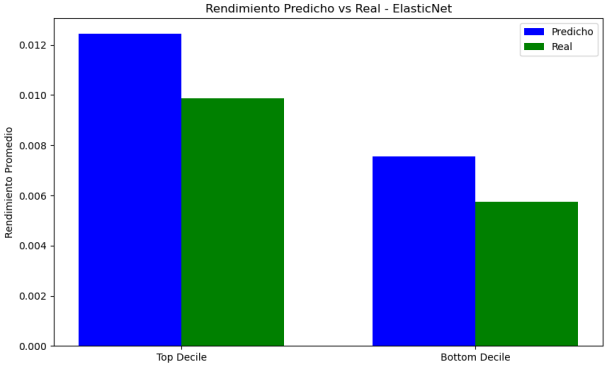
		ELASTIC NET											
TRAIN	TOP DECILE	X24795	X30745	X11042	X12510	X7365	X8313	X7353	X3943	X3563	X24797	X17583	X20359
	BOTTOM DECILE	X28143	X17579	X11865	X17985	X4976	X22632	X27498	X6718	X11838	X3720	X32452	X5009
TEST	TOP DECILE	X24795	X30745	X11042	X12510	X7365	X8313	X7353	X3943	X3563	X24797	X17583	X20359
	BOTTOM DECILE	X28143	X11865	X17579	X32453	X10987	X27498	X3720	X32452	X11838	X10637	X22632	X5009

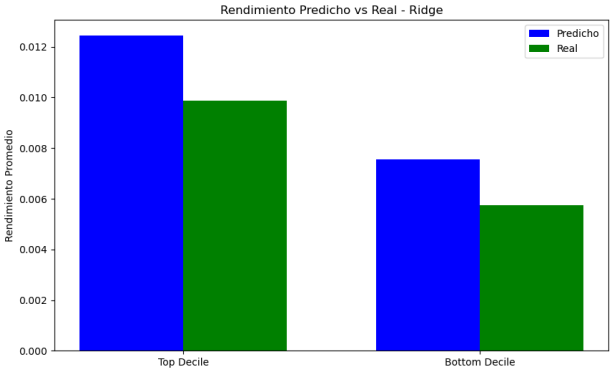
Vemos como muchos de los fondos, tanto para el top decile como para el bottom decile, son comunes entre la división 80-20 como la división 70-30. Esto resulta en una estrategia de inversión interesante que puede ser invertir en aquellos fondos comunes para ambas divisiones ya que quieren decir que, independientemente de la división realizada, siempre tienen buenos rendimientos (teniendo en cuenta que solo se han hecho dos divisiones de datos).

Pese a todo esto, encontramos como al analizar más a fondo el rendimiento real versus el rendimiento predicho, encontramos como la muestra que divide los datos en 80-20 obtiene unos resultados más cercanos a los reales como podemos ver a continuación

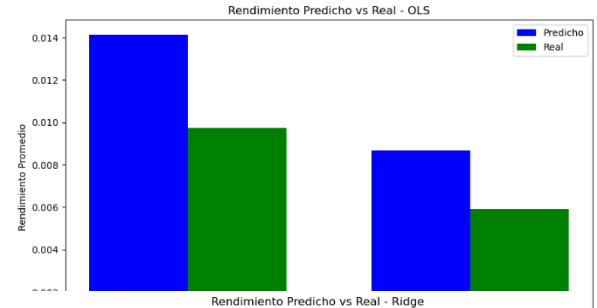


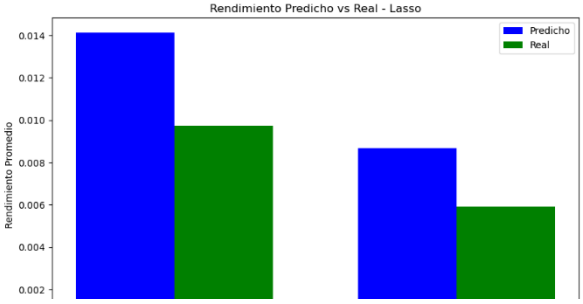


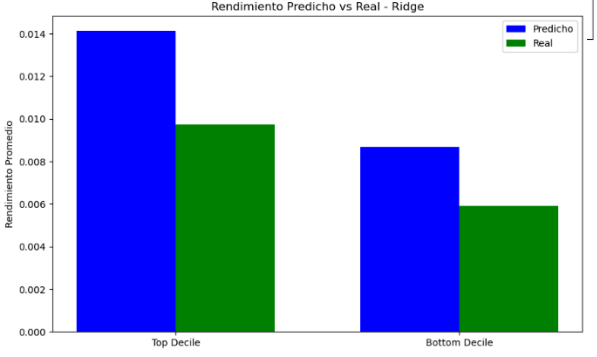


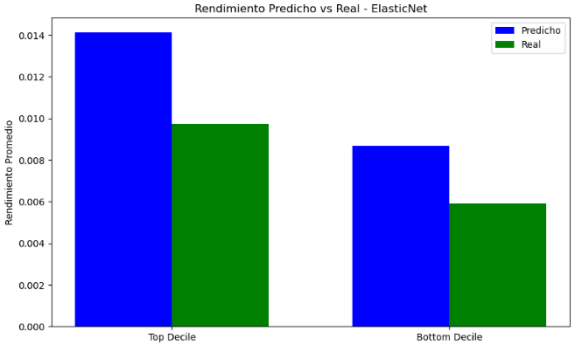


A diferencia de división 70-30 donde, aunque también cercanos, el rendimiento real versus el predicho tiene un *gap* mayor como podemos ver a continuación:









Al ser el valor real el mismo para ambos, esto nos indica que la división 70-30 proporciona en el entrenamiento un rendimiento bastante superior al real a diferencia de la división 80-20, por lo que otra forma interesante de inversión podría ser utilizar los fondos del “top decile” de la división 80-20, ya que parecen más fiables al ser más cercanos al rendimiento real.