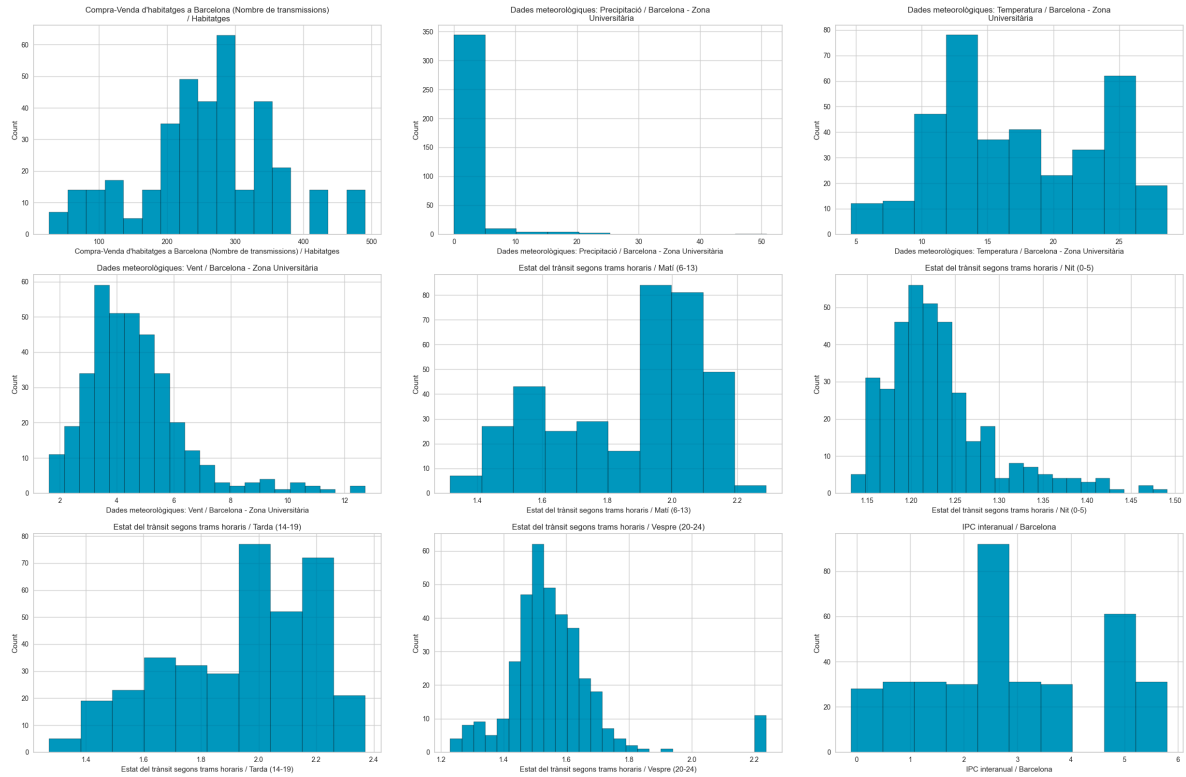
Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

La predicción bursátil es un problema complejo, pero a veces se pueden observar relaciones con variables que aparentemente no deberían influenciar. El portal de datos abiertos del ayuntamiento de Barcelona recoge informaciones diarias sobre la ciudad y esto nos ofrece la oportunidad de averiguar si lo que pasa en Barcelona tiene alguna influencia en el mercado del IBEX. Vamos a trabajar con un extracto de esos datos para el año 2021, con un subconjunto de variables que hemos elegido según nuestro criterio *experto* desentrañar esa influencia. El objetivo es aproximar el valor de la cotización del IBEX a partir de las otras variables.

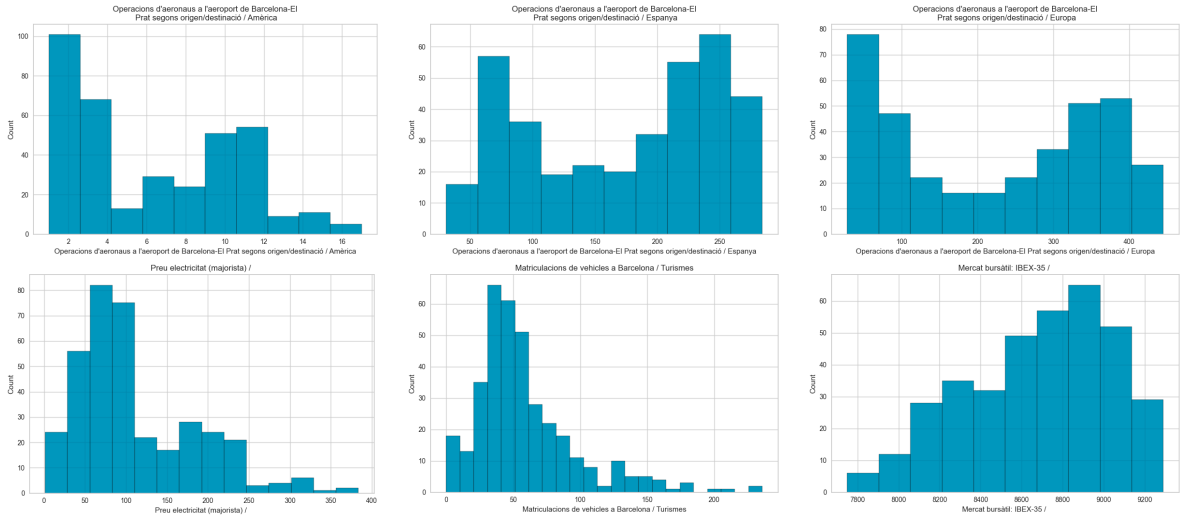
Puedes obtener estos datos mediante la función load_BCN_IBEX de la librería apafib. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

1. Divide el conjunto de datos en entrenamiento y test (80% / 20%). Haz una exploración mínima del conjunto de datos de entrenamiento observando las relaciones entre las variables, especialmente con la variable objetivo. Describe las cosas que hayas visto que te parezcan interesantes. Transforma las variables adecuadamente para poder ajustar un modelo de regresión tanto para el conjunto de entrenamiento como para el de test.

Se empezó la exploración de los datos comprobando las dimensiones del conjunto de datos, las cuales eran 365 filas y 15 columnas. Una vez se ha hecho una ojeada a los valores de las distintas columnas se optó por representar gráficamente, mediante histogramas, las distintas columnas.



Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

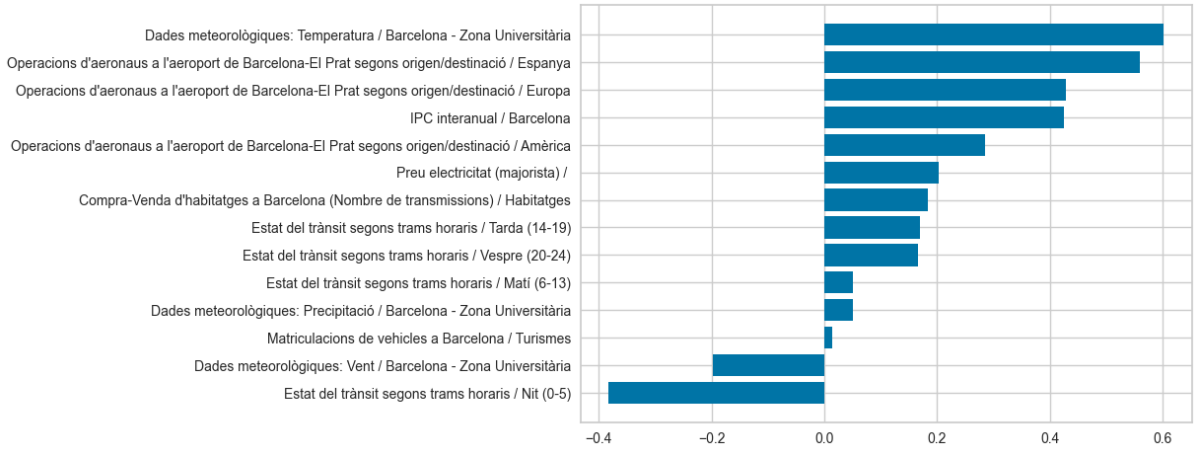


Mediante este primer análisis se pudo ver que los valores de las columnas no estaban normalizados, un hecho que nos podía llevar a problemas más adelante y, por eso, se decidió normalizar los valores.

El procedimiento que se describe a continuación se ha desarrollado, simultáneamente, con los valores normalizados y los no-normalizados. Se empezó calculando la matriz de correlación de las variables, que nos calcula la existencia de una relación lineal entre dos variables aleatorias.

Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

De esta matriz de correlación se analizó en profundidad las correlaciones con la variable objetivo, en este caso, se llama "Mercat bursàtil: IBEX-35". Se creo un gráfico de barras, con el que se pudo ver cuáles eran las variables con mayor dependencia lineal con la variable objetivo.

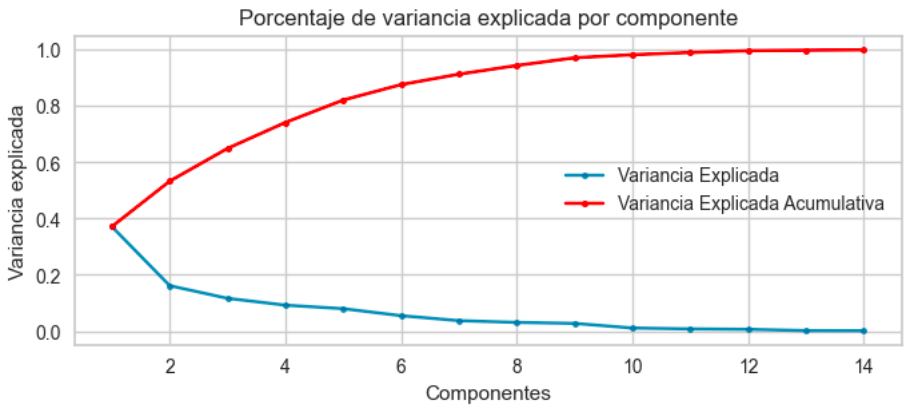


Una vez se había analizado, en detalle, el conjunto de datos, se decidió hacer la división entre el conjunto de entrenamiento y de test, mediante la función train_test_split. Además, una vez finalizada la división, se ha eliminado la variable objetivo de los conjuntos y se ha creado conjunto con los valores objetivos.

2. Aplica Análisis de Componentes Principales (PCA) al conjunto de entrenamiento y visualízalo en 2D representando la variable objetivo. ¿Crees que puede haber una relación entre las variables del conjunto de datos y la variable objetivo? ¿Por qué?

Se ha aplicado el análisis de componentes principales al conjunto de entrenamiento. Una vez ejecutado, se ha decidido estudiar la evolución de la variancia en función del número de componentes.

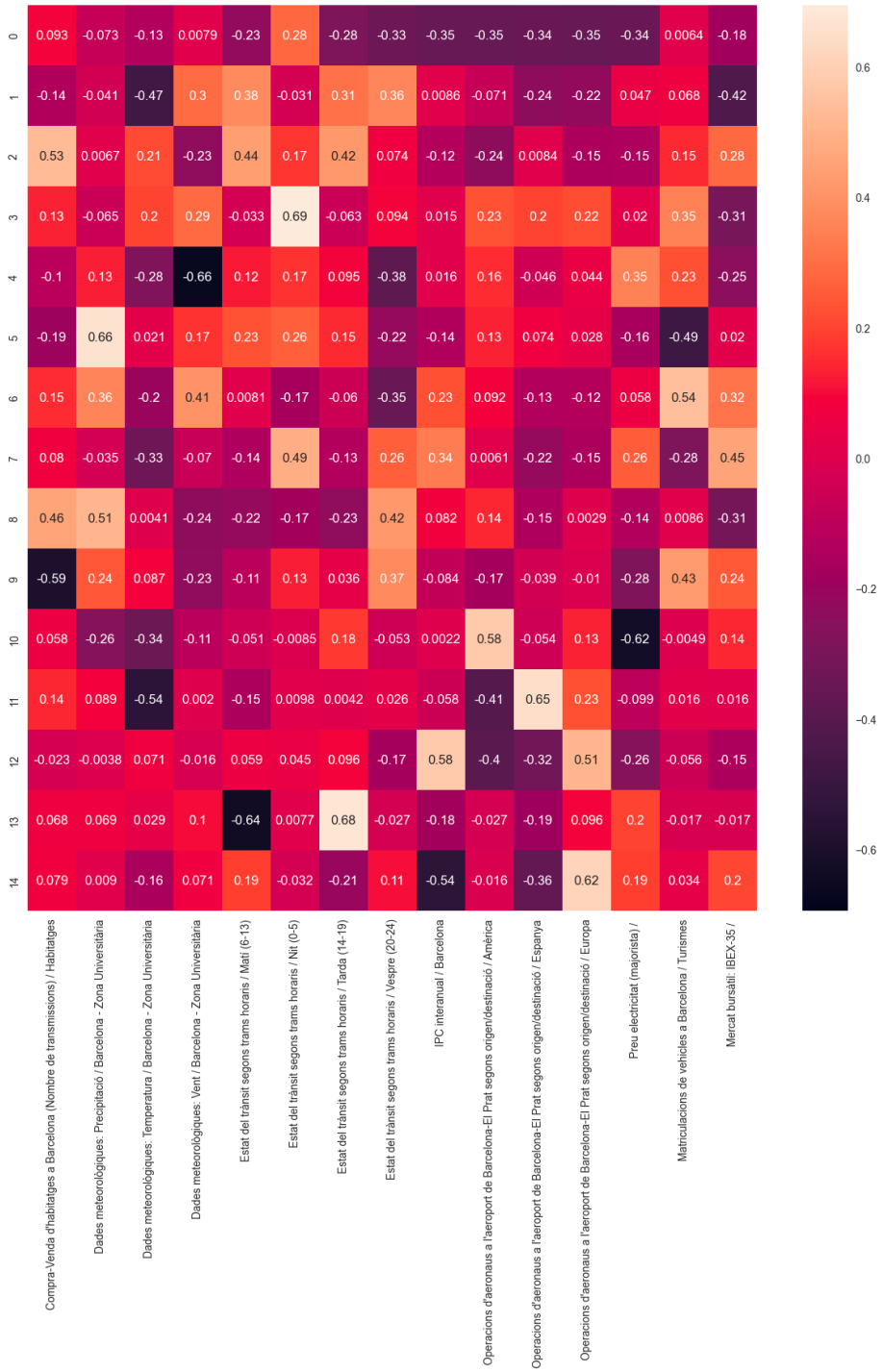
Como se puede ver en la gráfica, la variancia no aumenta drásticamente al reducir hasta 5 componentes. Si se quiere reducir por debajo de 5 componentes, se puede ver como la variancia aumenta y, por lo tanto, estamos perdiendo precisión de los datos.



Una vez se había hecho un pequeño estudio a la evolución de la variancia con el número de componentes después de ejecutar el análisis de componentes principales, se ha decidido ver

Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

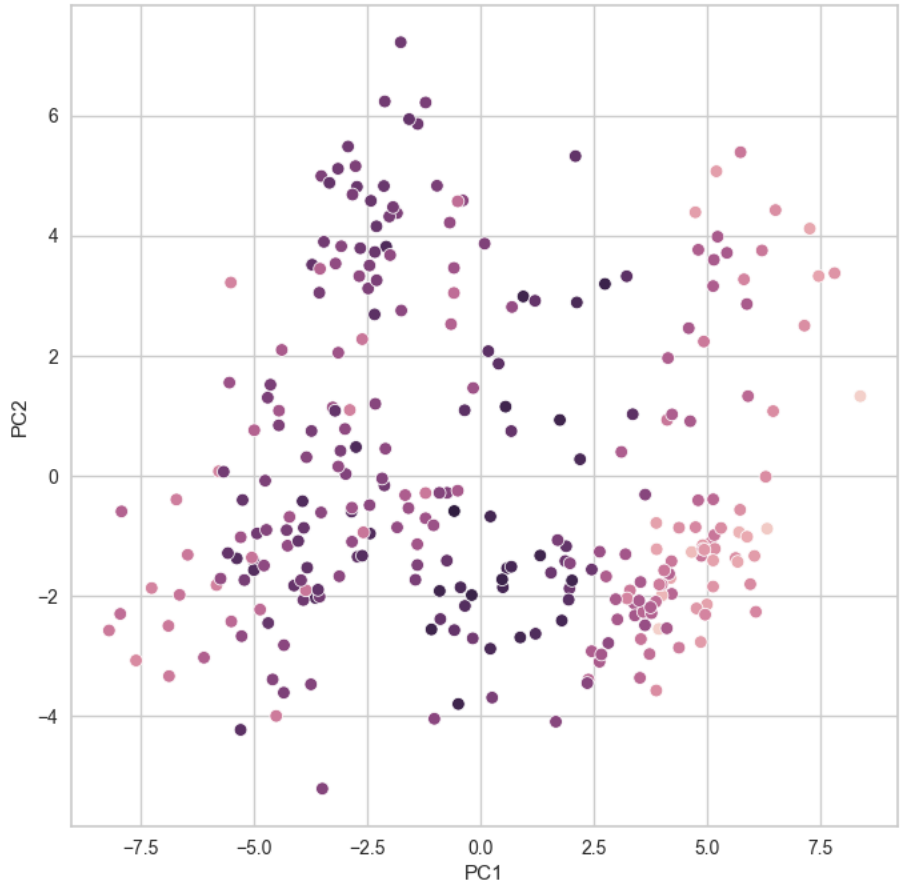
la relación de la variable objetivo. Esta información se ha sacado como una matriz de correlaciones que se puede ver a continuación.



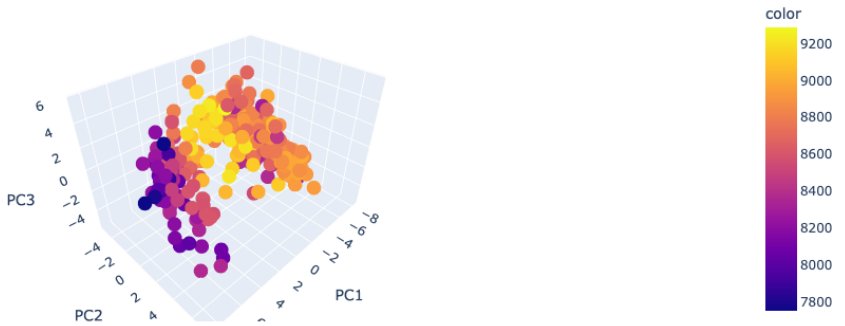
Hay que destacar que, para hacer la matriz de correlaciones se ha escogido con el mismo número de componentes que de variables iniciales y, se puede ver como se ha perdido relación lineal con la variable objetivo si lo comparamos con la matriz de correlación de los datos iniciales.

Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

Por otra parte, se decidió hacer una representación bidimensional del conjunto de datos en PCA, para ver si existía algún tipo de relación entre las variables, que no permitiese sacar resultados más acurados con una complejidad en los datos mucho menor.



Como se puede ver, existe patrón en la representación bidimensional de los datos. Se puede apreciar que los valores más oscuros se encuentran centrados en el eje de PC1 y los valores más claros se encuentran en los extremos de esa misma coordenada. Para sacar conclusiones más acertadas se creyó oportuno hacer una representación tridimensional de los datos, obteniendo la siguiente gráfica:

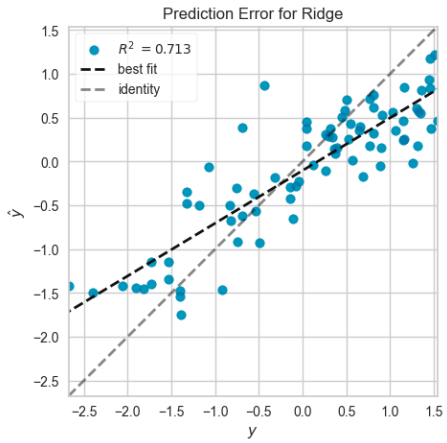


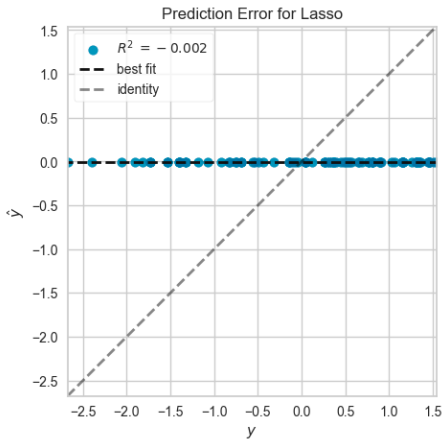
Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

Es importante tener en cuenta que los valores más oscuros de la representación bidimensional son aquellos que se acercan más al color amarillo. Por el contrario, los valores con representación más clara son aquellos que se acercan al color azul. De la misma manera que en las conclusiones anteriores. En esta representación se puede apreciar la misma relación, los valores más altos se encuentran en el centro y los valores más bajos se encuentran en los extremos.

3. Ajusta una regresión lineal, una regresión Ridge y una regresión LASSO a los datos. ¿Te parece suficientemente bueno el resultado? Representa los valores de la variable objetivo para el conjunto de test contra las predicciones y el qqplot. ¿Qué modelo te parece mejor? ¿Tienen sentido las variables con más peso que aparecen en los modelos para la variable que queremos predecir? Elimina las variables que tienen menos peso en los modelos del conjunto de datos y reajusta el modelo de regresión lineal ¿Cómo ha cambiado el peso de las variables que quedan?

Después de entrenar los modelos para el conjunto de datos de entrenamiento, se han logrado los siguientes resultados en las predicciones:





Como podemos ver, el modelo de Ridge ha obtenido unos resultados mucho más satisfactorios que el de Lasso. Para cuantificarlo analíticamente se ha ejecutado la función score en ambas regresiones y se han obtenido 0.713 y 0.002 respectivamente. Con los resultados en mano y este pequeño análisis, se puede afirmar, sin temor a equivocarnos, que hay aun bastante camino de mejora. Con este objetivo, se decidió comprobar cuales son las variables que tienen más peso en las regresiones.

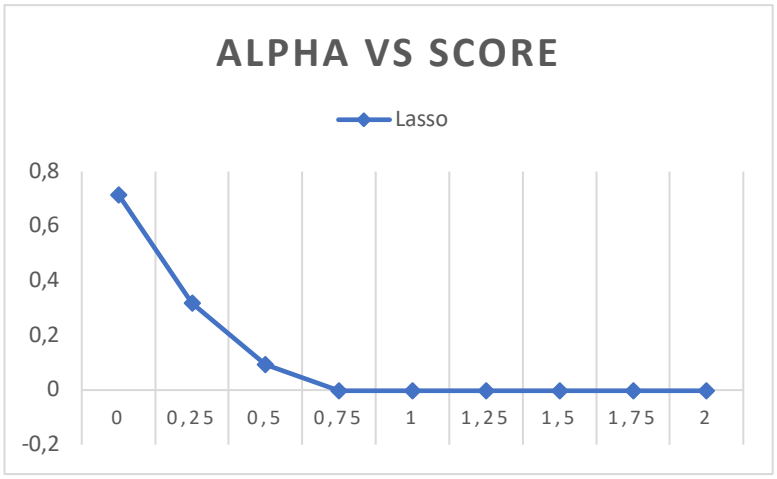
Idx	**Name**	**Ridge**	**$Lasso_1$**
0	Compra-Venda	0.039	0.0
1	Dades meteo	0.059	0.0
2	Dades meteo	0.434	0.0
3	Dades meteo	-0.108	-0.0
4	Estat del trànsit	-0.323	0.0
5	Estat del trànsit	-0.118	-0.0
6	Estat del trànsit	0.326	0.0
7	Estat del trànsit	-0.127	0.0
8	IPC interanual	1.06	0.0

9	Operacions	-0.121	0.0
10	Operacions	0.681	0.0
11	Operacions	-0.91	0.0
12	Preu elec	-0.485	0.0
13	Matriculacions	-0.081	0.0

Con los resultados en mano, se ha visto que existe algún tipo de error en la ejecución de la regresión lineal de Lasso. A raíz de esto se decide hacer un estudió de esta regresión y encontrar cual es el error que está causando problemas en la ejecución. Después de hacer una investigación del funcionamiento de esta regresión se ve que existe un parámetro Alpha que viene definido a 1, pero que se puede modificar en la creación de la regresión. Como se enuncia en la documentación de la clase:

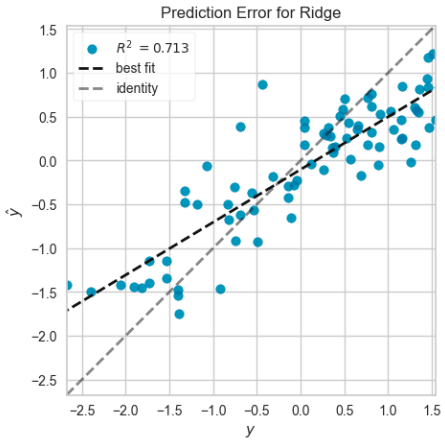
> Constant that multiplies the L1 term, controlling regularization strength. `alpha` must be a non-negative float i.e. in `[0, inf)`.
>
> When `alpha = 0`, the objective is equivalent to ordinary least squares, solved by the **`LinearRegression`** object. For numerical reasons, using `alpha = 0` with the `Lasso` object is not advised. Instead, you should use the **`LinearRegression`** object.

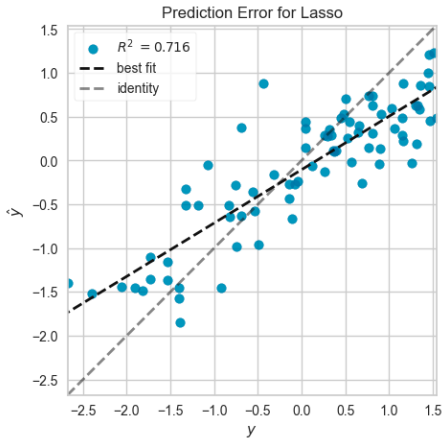
Con este parámetro en mente, se decide hacer una pequeña experimentación y así encontrar el valor de Alpha que nos permite un mejor ajuste de los datos de prueba. Se ejecuto con múltiples valores desde 0.0 hasta 2.0. Los resultados obtenidos fueron los siguientes:



Como se puede ver, se obtienen los mejores resultados con un valor de Alpha cercano a 0. Por esto, se decidió repetir la experimentación, con un Alpha equivalente a 0, aunque no era recomendado en la documentación. Se pueden ver gráficamente a continuación:

Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu





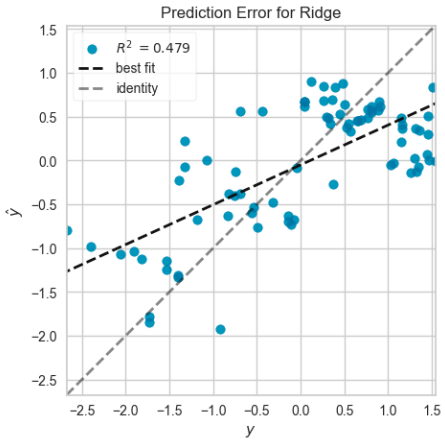
Con estos nuevos resultados, se han comprobado cuales son los componentes que más influyen en el cómputo de la predicción:

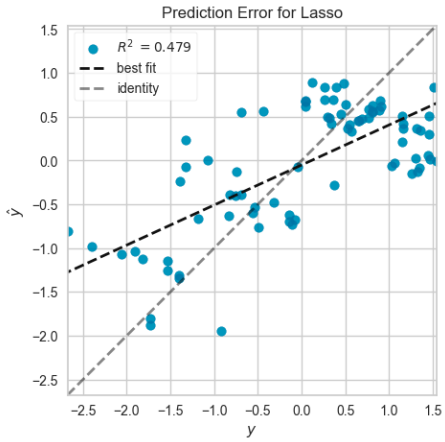
Idx	**Name**	**Ridge**	**Lasso$_2$**
0	Compra-Venda	0.039	0.032
1	Dades meteo	0.059	0.059
2	Dades meteo	0.434	0.429
3	Dades meteo	-0.108	-0.11
4	Estat del trànsit	-0.323	-0.366
5	Estat del trànsit	-0.118	-0.144
6	Estat del trànsit	0.326	0.363
7	Estat del trànsit	-0.127	-0.13
8	IPC interanual	1.06	1.099
9	Operacions	-0.121	-0.089
10	Operacions	0.681	0.761
11	Operacions	-0.91	-1.031
12	Preu elec	-0.485	-0.514
13	Matriculacions	-0.081	-0.083

Con estos datos, ahora se puede decidir eliminar del conjunto de datos de entrada todos aquellos que tienen un coeficiente negativo y así reducir el número de datos. Los cuales son aquellos que están soberados de color naranja.

Con el nuevo conjunto de datos, se han entrenado ambas regresiones lineales y se han obtenido los siguientes resultados:

Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu



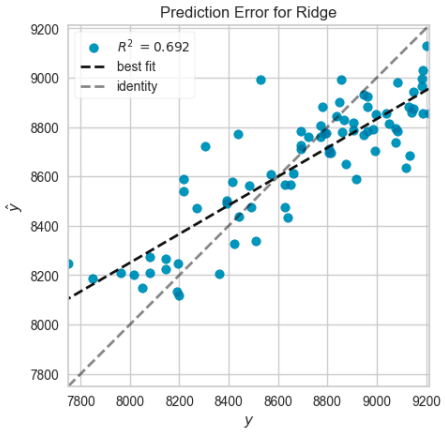


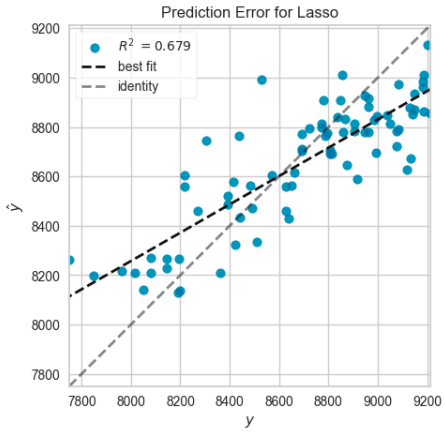
Como se puede ver, los resultados han empeorado y se han obtenido, en ambos casos, un score de 0.479, inferior al anterior. Con estos nuevos modelos, se ha mirado que variables tienen más peso en la regresión.

Idx	Name	Ridge	$Lasso_2$
0	Compra-Venda	0.116	0.115
1	Dades meteo	0.042	0.043
2	Dades meteo	0.583	0.593
6	Estat del trànsit	0.020	0.020
8	IPC interanual	0.456	0.466
10	Operacions	-0.131	-0.146

Como se ha podido ver en este informe, el modelo de regresión lineal de Lasso no era capaz de entrenarse con el conjunto de datos facilitado, si estaba estandarizado. Esto ha provocado que, durante la experimentación, la regresión lineal de Lasso tomaba como valor de Alpha 0.0, hecho que no es recomendable.

En este apartado final, se ha optado por repetir la experimentación con un conjunto de datos no escalado, hecho que nos ha permitido obtener los siguientes resultados.





Como se puede ver, si los datos no están estandarizados la regresión de Ridge tiene un desempeño superior a la de Lasso. Pero como el objetivo de este apartada es compararlo con la regresión de Lasso estandarizada se obviará de ahora en adelante.

Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

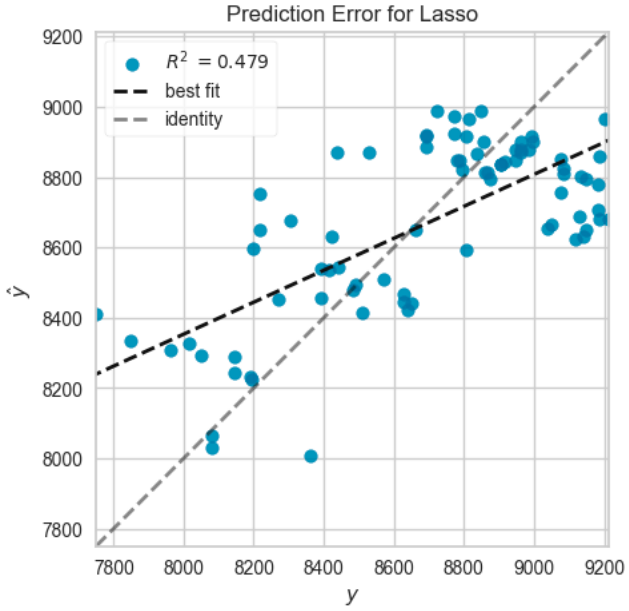
De la misma manera que en el apartado c, se pide limpiar aquellos parámetros que influyen poco en la predicción, en este caso tenemos los siguientes coeficientes.

Idx	Name	Ridge	$Lasso_2$	$Lasso_3$
0	Compra-Venda	0.039	0.032	0.117
1	Dades meteo	0.059	0.059	5.351
2	Dades meteo	0.434	0.429	22.885
3	Dades meteo	-0.108	-0.11	-24.668
4	Estat del trànsit	-0.323	-0.366	-61.811
5	Estat del trànsit	-0.118	-0.144	-93.504
6	Estat del trànsit	0.326	0.363	0.0
7	Estat del trànsit	-0.127	-0.13	-128.419
8	IPC interanual	1.06	1.099	222.255
9	Operacions	-0.121	-0.089	-9.955
10	Operacions	0.681	0.761	4.078
11	Operacions	-0.91	-1.031	-2.683
12	Preu elec	-0.485	-0.514	-2.436
13	Matriculacions	-0.081	-0.083	-0.642

Podemos afirmar, que los coeficientes que menos influencia tienen en la predicción final siguen siendo los mismo que en los datos estandarizados.

El estudio de la regresión lineal con datos no estandarizados continua con la predicción con los datos limpiados, o en otras palabras, eliminando del conjunto de datos esas variables que tienen menor importancia en el resultado final de la predicción.

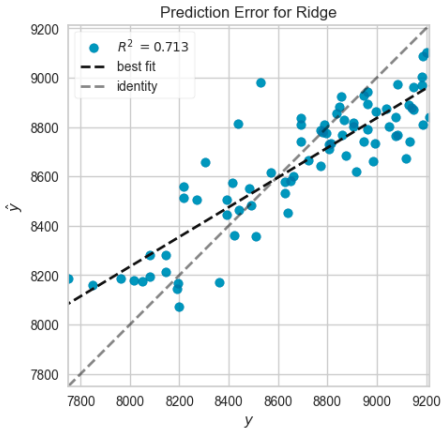
Con este nuevo conjunto de datos, se han reentrenado los modelos, pero no se ha mejorado el score del conjunto de datos de test.

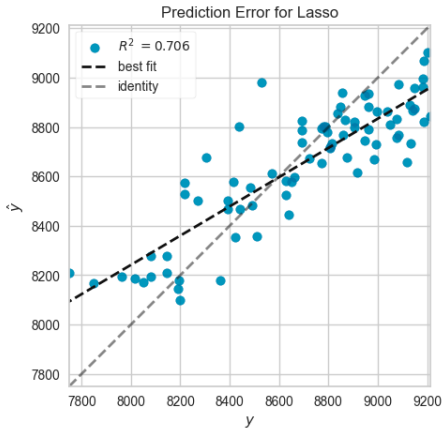


Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

Después de hacer una búsqueda intensiva por internet, se ha visto que la regresión no funcionaba como esperado ya que se estaban usando unos valores normalizados en la variable a predecir. Por esta causa, se ha decidido repetir la experimentación de todo el apartado 3.

Se empieza ejecutando las regresiones lineales de Ridge y Lasso, obteniendo los siguientes resultados:



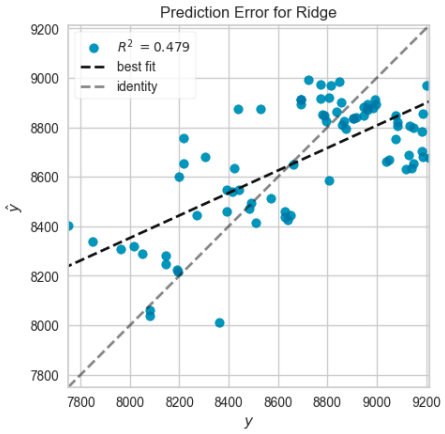


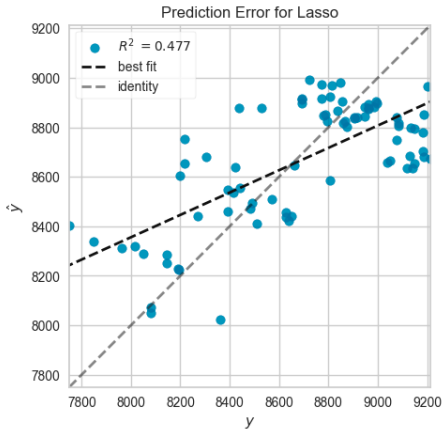
Podemos ver, que utilizando las variables objetivas no normalizadas se obtienen resultados satisfactorios. Las variables han obtenido los siguientes coeficientes:

Idx	**Name**	**Ridge**	**Lasso$_1$**
0	Compra-Venda	13.535	12.917
1	Dades meteo	20.497	19.817
2	Dades meteo	150.602	149.089
3	Dades meteo	-37.343	-36.921
4	Estat del trànsit	-112.246	-86.799
5	Estat del trànsit	-40.905	-38.879
6	Estat del trànsit	113.159	83.708
7	Estat del trànsit	-44.172	-39.753
8	IPC interanual	367.783	368.359
9	Operacions	-41.914	-38.362
10	Operacions	236.202	246.458
11	Operacions	-315.907	-325.051
12	Preu elec	-168.413	-168.059
13	Matriculacions	-28.19	-25.862

Después de limpiar el conjunto de datos, se ha vuelto a ejecutar el entrenamiento de las regresiones obteniendo los siguientes resultados.

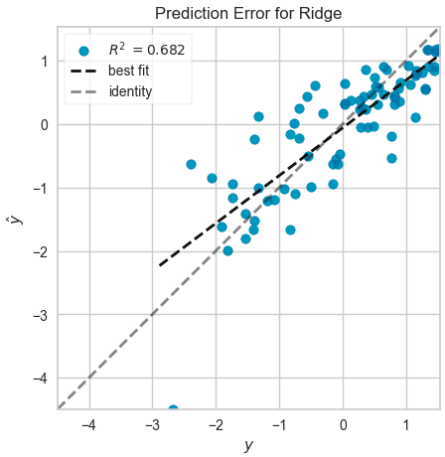
Ignasi Fibla Figuerola
47442872J
ignasi.fibla@estudiantat.upc.edu

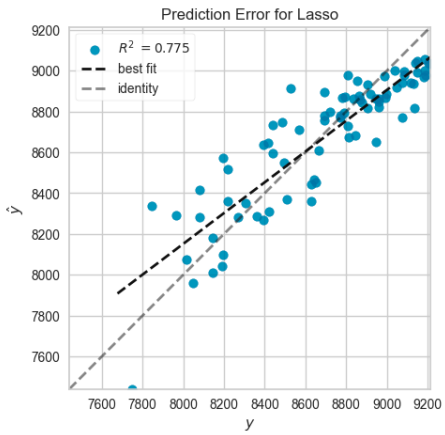




4. Al ser un problema complejo, igual hay interacciones entre variables que explican mejor la variable objetivo. Usa la función PolynomialFeatures de scikit-learn para añadir al conjunto de datos original características que correspondan a polinomios de grado 2. Vuelve a ajustar la regresión Ridge y la regresión LASSO. ¿Han mejorado los modelos? Fíjate en las variables a las que LASSO no les ha dado un peso 0. ¿Se corresponden con interacciones entre variables?

Después de la ejecución, de la función PolynomialFeatures se ha vuelto a ejecutar el entrenamiento de ambas regresiones.





Después de ejecutar el PolynomialFeatures se ha podido ver como los resultados mejoran. Si hacemos una inspección a los coeficientes de Lasso, no encontramos una clara relación entre los coeficientes y las interacciones entre variables