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.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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.

Chart

Description automatically generated

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.

Chart

Description automatically generated with medium confidence

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.

1. 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?
2. 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?
3. 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 con- junto 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?