EXAMEN DE PREDICCIÓN

Máster en Data Science, 2020-2021

Valentina Díaz Torres
22-01-2021
CUNEF

La presente práctica trata sobre un análisis predictivo de la base de datos de Craigslisty,en Estados unidos con información acerca de venta de automóviles.

El objetivo principal va a ser el estudio, mediante diferentes técnicas de predicción, de las variables más relevantes para el mismo, con el fin de formar un modelo predictivo que deduzca el precio, a partir del resto de variables seleccionadas.

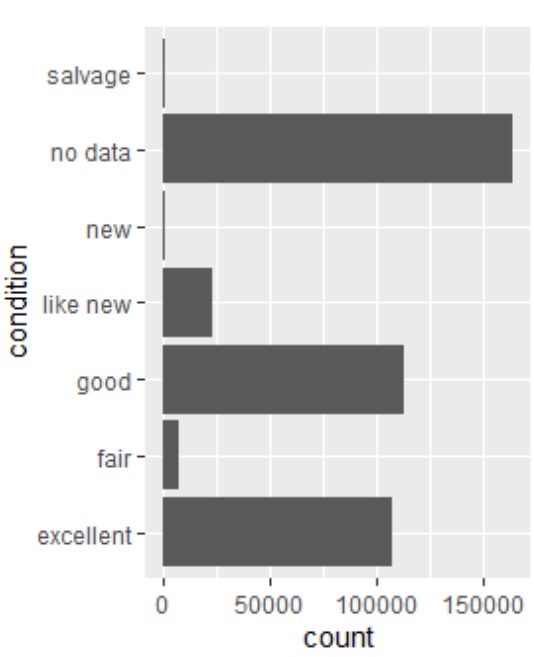
Sobre la base de datos

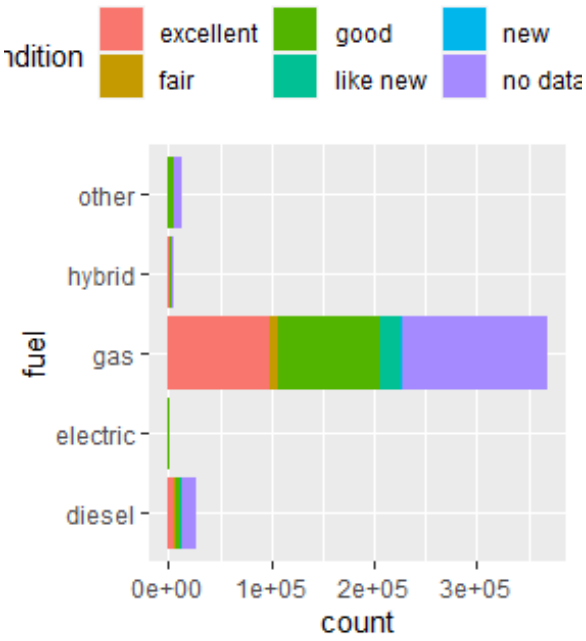
La base de datos utilizada está compuesta por un conjunto de 458213 registros de un total de 26 variables. Entre ellas, se encuentran estado, latitud, región, pero también otras como el número de cilindros del coche, el año o el precio.

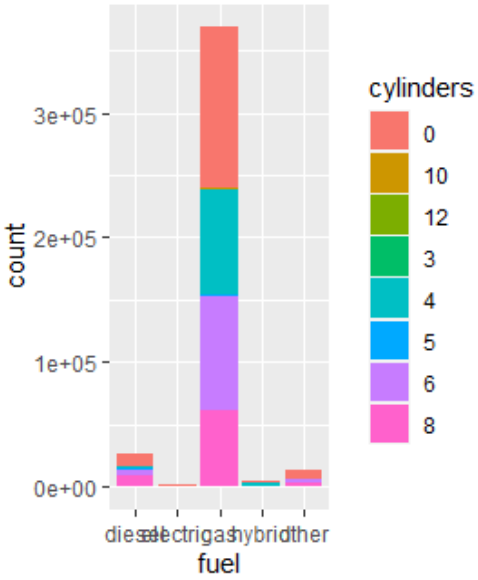
La naturaleza de estas variables es distinta, por un lado, se encuentran variables categóricas y por otro numéricas. Para el análisis pertinente, se ha decidido solo trabajar con variables de tipo numérico, es por esto que las que no lo eran han sido manipuladas mediante el método de One Hot Encoding, que consiste en utilizar las categorías de las variables como nuevas variables.

Por todo lo anterior, tras un análisis exploratorio previo, se han descartado aquellas variables que contenían información que alejaba el análisis objetivo, tales como las relacionadas con la localización, ya que se pretende, en este caso, predecir el precio de la venta de los coches, basándose solo en las características de los mismos, no obstante, se podría haber tomado un enfoque distinto.

Algunas de las variables estudiadas, se podrían visualizar del siguiente modo:







Así, se representarían algunas de las variables seleccionadas como el número de cilindros, fuel o la condición. Se detecta que hay un mayor número de cuatro cilindros y de la categoría fuel_gas, que las condiciones más habituales son las de buena condición y excelente y que el número de cilindros más frecuente también son los de 4, 6 y 8.

Cabe destacar, que en la base de datos con la que se trabaja, existen muchos valores nulos, por lo cual estos han sido sustituidos por una nueva variable llamada “no data” en el caso de las categóricas y por el valor de 0 en el caso de las numéricas.

En cuanto a la selección de variables, esta se ha hecho teniendo en muy presente el conocimiento del negocio y los objetivos de análisis claros. Como se ha dicho anteriormente, el enfoque es acerca de las características del propio coche y no por región o estado, por ejemplo.

De este modo, ya con los gráficos de exploración se ha podido observar qué variables iban a ser más importantes que otras, teniendo en cuenta que cada característica de ellas ha pasado a ser una nueva variable. Finalmente, las elegidas fueron: year, odometer, price, cylinders, fuel, condition y size, para luego ser transformadas, lo cual se quedó en un total de 28 variables.

Es importante resaltar que durante esta práctica las variables se han sometido a examen continuamente, por lo que, pese a que se ha hecho una previa elección, algunas de ellas fueron eliminadas posteriormente, por motivos de multicolinealidad entre las variables, por ejemplo.

Métodos empleados

Como se ha comentado anteriormente, para el tratado de las variables se ha llevado a cabo One Hot Encoding, formando un total de 28 variables. Además, se han realizado algunos tests con el fin de observar qué variables son las más relevantes.

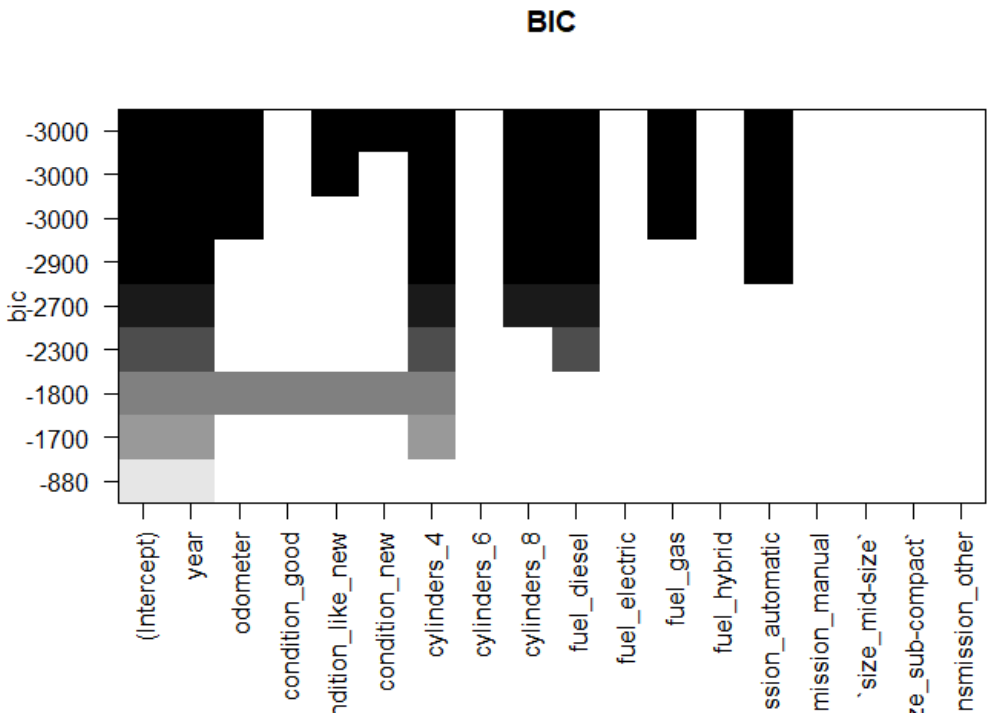
En primer lugar, se llevó a cabo la función stepAIC, que mide el AIC de las variables, descartando las no relevantes y ofreciendo un modelo final con el valor de AIC minimizado. Cuando esto se realizó, se ofreción que las variables de 8, 4 y 6 cilindros, años, condición buena, transmisión automática, Diesel y tamaño completo, son las más relevantes.

Además, se ha implementado la prueba de VIF (Variance Inflation Factor), que mide la entidad de la multicolinealidad en un análisis de regresión de mínimos cuadrados. Aquí se detectó que no existía problema alguno de multicolinealidad entre las variables y los coeficientes más altos de significancia fueron otorgados a las variables fuel_diesel, fuel_gas, transmission_automatic, transimission_manual, lo cual encaja con los resultados obtenidos anteriormente.

Cabe resaltar que los datos fueron separados en un subconjunto de training, del 70% y otro de test del resto, con el fin de medir la eficacia de los modelos implementados.

Tras ello, se usaron tres coeficientes diferentes para crear modelos en base a la puntuación que estos otorgaban a las variables. En primer lugar, el coeficiente de determinación ajustado (adjusted r-squared), el cp o Mallo's Cp y por último el BIC (Schwartz's information criterion, BIC). Mediante la función regsubsets, se selecciona el modelo mejor, en base a la información obtenida por los estadísticos anteriores.

Los valores obtenidos en la selección del modelo según BIC son representados de este modo:



En él, las variables más representativas quedan con color más oscuro, siendo negro el máximo de la escala, máxima relevancia y el gris claro el de menor. Siguiendo esto, se puede observar como year, odometer, condition_good, condition new, cylinders_4, cylinders_8, fuel_diesel, fuel_gas y transmission_automatic, son las más importantes, siendo year, la que parece tener más relevancia. Todo esto tendría sentido, ya que el número de cilindros podría subir o bajar el precio fácilmente, el año del coche también, el cuentakilómetros, la buena condición, la transmisión automática y qué tipo de combustible consuma.

Tras ello, teniendo en cuenta a los tres coeficientes, las variables que tenían mayores valores son las de fuel_gas, condition_good, cylinders_4, cylinders_8, full_size y fuel_diesel. Esto parece tener coherencia con los datos obtenidos ya anteriormente.

Tras ello, se han realizado cada uno de estos coeficientes, para ver qué variables eran las que ofrecían cada uno y después con esos datos formar un modelo para cada uno de ellos el resultado de los modelos fue, el primero, el de Adj.R2, formado por year, odometer y condition Good, el segundo, el Cp, formado por cylinders 8 y fuel diesel, y por último, el de BIC, solo recomendaba el de cylinders 8 como la mejor opción. Una vez que estos fueron formados, se evaluaron, según el test de AIC. Los resultados mostraron que el mejor modelo es el de Cp, pues tiene un menor valor de tal coeficiente.

Estos son los valores obtenidos por los tres coeficientes medidos:

Adj.R2	CP	BIC
0.002904060	2465.1191	-884.0499
0.005532127	1636.4045	-1696.7074
0.007576619	991.9346	-2327.8313
0.008706791	636.1272	-2672.0671
0.009439293	405.8695	-2891.2944
0.003298654	2345.5041	-949.4188
0.009906810	260.2694	-3015.4381
0.010089747	203.5147	-3061.5007

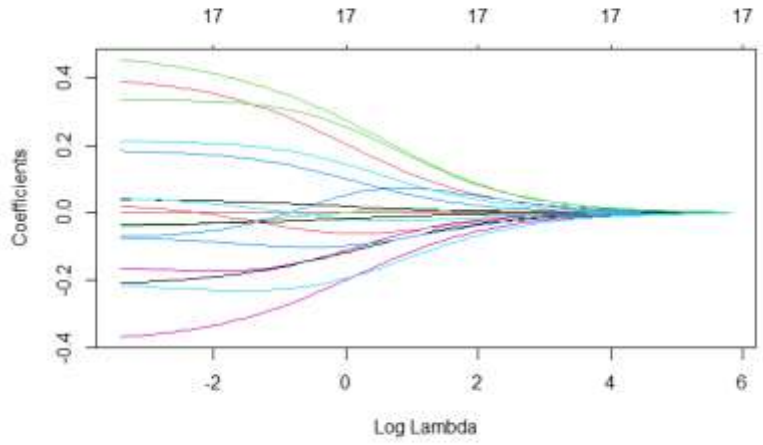
Además, se ha realizado también la predicción de ambos modelos, ofreciendo un valor de media y de desviación típica cada uno. El mayor valor de media es ofrecido por el modelo Cp.

Más tarde, tras haber comprobado los tres coeficientes anteriores, se realizó, mediante cross validation, en este caso, se crearon tres modelos, en primer lugar, uno formado por las todas las variables, el segundo con las variables que habían resultado significativas a lo largo del análisis en alguno de los oeficientes, un total de 7 y por último un modelo con las dos variables que habían sido más recomendadas, fuel_diesel y cylinders_8.

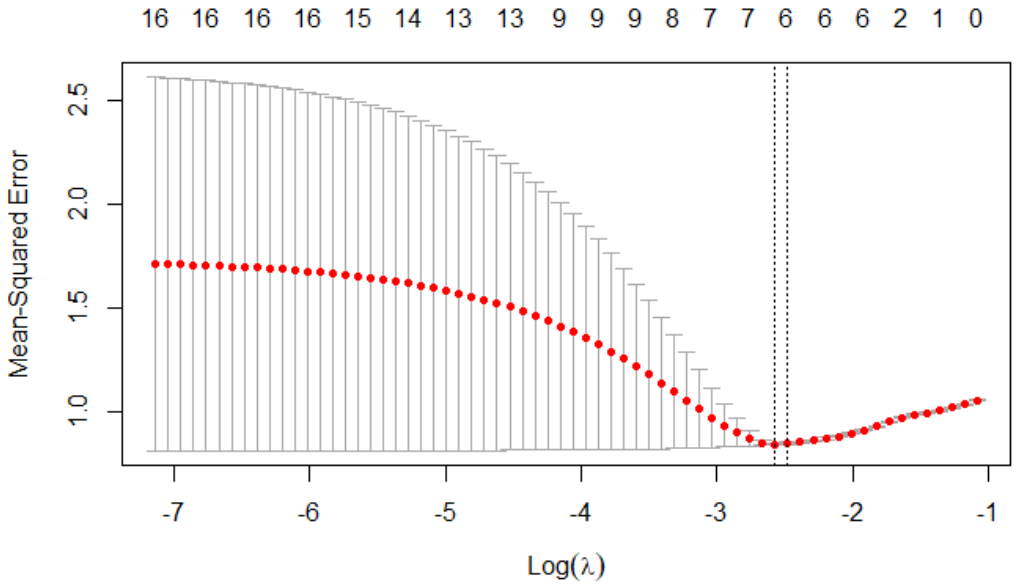
Tras inspeccionar y validar los modelos, se obtiene que el que tiene un menor error cuadrático (MSE) es el segundo, el que tiene aquellas variables más repetidas a lo largo de la práctica.

Por último, las técnicas finales empleadas han sido La regresión de Ridge y el modelo Lasso. La regresión de Ride penaliza la contracción cuando los coeficientes se acercan a cero, por lo que usa el parámetro lambda, cuyo valor debe ser estimado por cross validation.

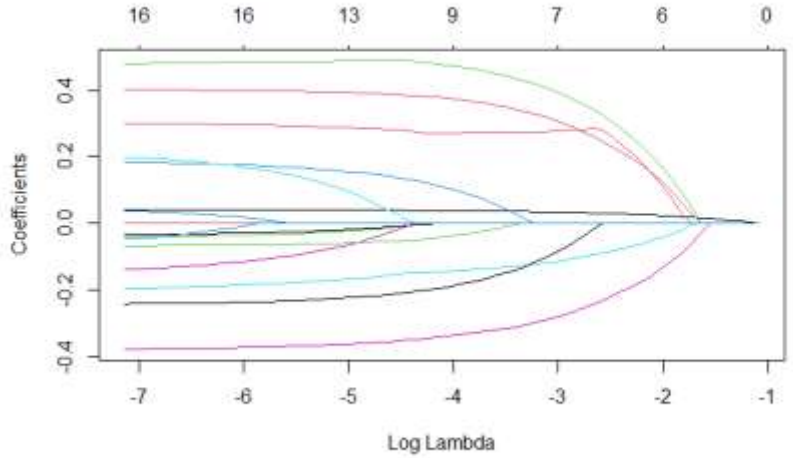
Esta es la representación obtenida por la regresión Ridge:



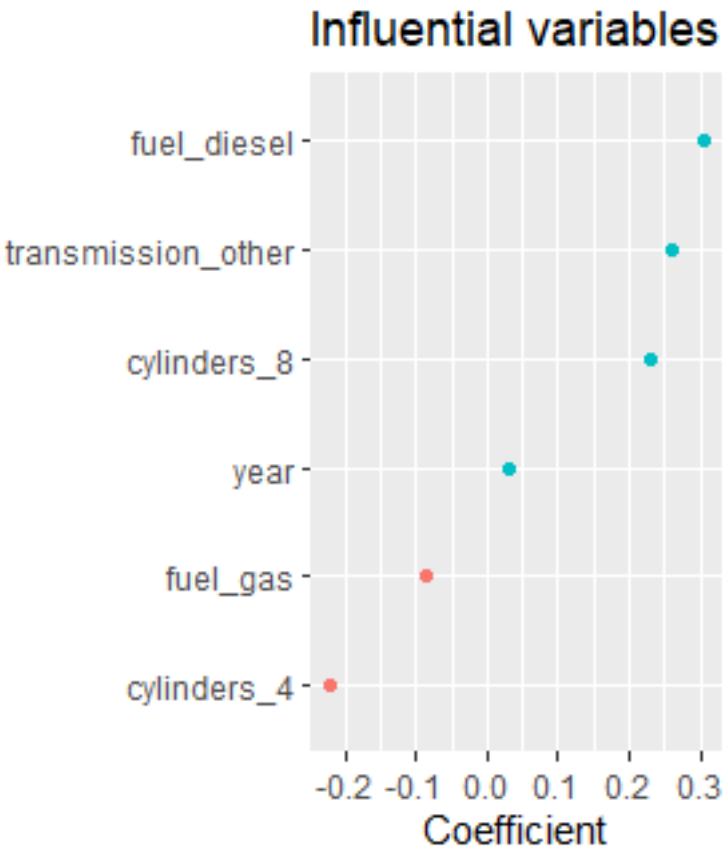
Además, el resultado de aplicar cross validation a Ridge es el siguiente:



En el caso del Lasso el mínimo error cuadrático ha sido 0.846 y la lambda correspondiente para este valor 0.076. El Alpha en este caso es 1 a diferencia de en Ridge que se estableció como 0.



Además, Lasso es empleado para la selección de variables, por lo que, después de aplicar cross validation, las variables recomendadas han sido las siguientes:



Conclusión

Como conclusión cabría destacar que todos los modelos y técnicas empleadas han llegado casi a las mismas conclusiones. Las variables más significativas han sido las de fuel diesel, cylinders_8, year, fuel_gas ,cylinders 4 y transmission automatic.

Esto querría decir que son las variables del modelo planteado que aportarían mayor precisión a la hora de predecir el precio de los vehículos en Estados Unidos, según los datos proporcionados.

Se han realizado abundantes pruebas para medir estas variables que formen el mejor modelo y algunos otros, en total uno para BIC, otro para Cp, el de AdjR2 y por último Lasso. El mejor de los tres primeros, según sus coeficientes y valores obtenidos, así como el mejor recomendado por AIC es el de Cp. Sin embargo, este modelo englobaba tan solo tres variables, que se consideran en cierta medida insuficientes para la predicción.

Teniendo en cuenta Lasso, este no ha aportado variables diferentes a las ya nombradas anteriormente, pero si tiene en cuenta algunas más. Además, este método es mejor, ya que se reduce el error mínimo cuadrático o MSE.

Por tanto, los precios de los coches vendrían definido por su combustión, por si tiene 8 o 4 cilindros, lo cual quiere decir que su precio será mayor o menor sustancialmente, si estos son automáticos o no y el año del que sea el coche.