Segunda entrega de proyecto

Por:

Luis David Morales Aguilar

Materia:
Introducción a la inteligencia artificial

Profesor:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA

MEDELLÍN

1. Importación de datos y preparación.

En esta sección se realizó la importación de los datos directamente desde Kaggle utilizando el archivo kaggle.json, y su preparación para el trabajo posterior.

2. Descripción y preparación del dataframe.

Luego se realiza una eliminación de datos de las columnas 'aspiration' , 'drivewheel', 'fuelsystem' y 'doornumber'. Estas columnas se seleccionan debido a que proporcionan datos muy específicos que no afectan en gran medida el valor de los vehículos. El objetivo es eliminar el 5% de los datos, por lo que se hace el cálculo en función de los datos disponibles, el cual es 205 (cantidad de filas)x 30 (cantidad de columnas) dando un valor de 308 datos. De estas columnas se eliminan de forma aleatoria valores con el fin de completar este requisito.

El dataset original sólo contenía 25 columnas, por lo que adicionan columnas basadas en las existentes, pero realizando operaciones matemáticas indicadas en el notebook. A continuación se muestra una captura de las líneas de código utilizadas.

```
ds1= dc.assign(Log_peakrpm=log_pr)
ds2= ds1.assign(Log_wheelbase=log_wb)
ds3= ds2.assign(Square_root_wp=sqrt_hp)
ds4= ds3.assign(Exp2_stroke=exp_st)
dsf= ds4.assign(Compressionratio_norm=comp_nor)
```

En estas ección también se describe el df original, con la cantidad de datos y el tipo de los mismos.

3. Limpieza de datos.

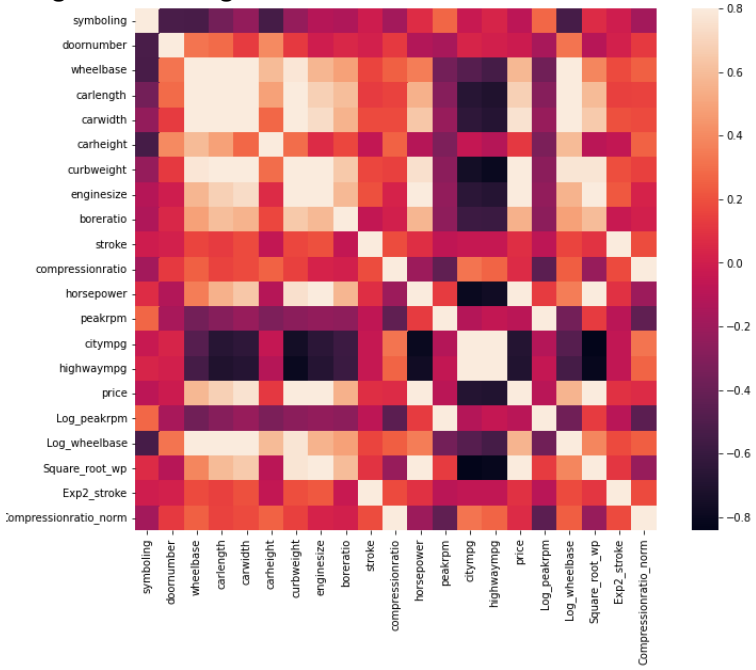
Antes de comenzar a manipular el dataframe, se realiza una limpieza de datos con el fin de llenar los datos faltantes con otros que puedan llegar a ser útiles. La siguiente captura describe las líneas de comando utilizadas para dicho fin:

```
ds["aspiration"]=ds.aspiration.fillna("Unk")
ds["drivewheel"]=ds.aspiration.fillna("Unk")
ds["fuelsystem"]=ds.aspiration.fillna("Unk")
ds["doornumber"]=ds.doornumber.fillna(int(np.mean(ds["doornumber"]))+1)
ds
```

Los datos de las columnas "aspiration", "drivewheel" y "fuelsystem" contienen información muy específica de los autos que no necesariamente afecten o tengan una corre lación con su precio, por lo cual los datos faltantes se llenaron con el dato "UnK" que siganifica Unknown o desconocido en español. Por esta razón se utilizarán otras variables para predecir el precio de los vehículos. En la columna "doornumber" los datos faltantes se completaron con su promedio, el cual dió de 3.14, pero debido a que el dato debe ser entero (ya que un carro comúnmente tiene un numero entero de puertas) se convirtió en entero, danto un valor de 3, a lo que se realiza una suma de 1 para así obtener un valor de 4. Esta decisión se debe a que los vehículos de 4 puertas son más comunes, por ello el promedio estaba más cerca a 4 que a 2. Luego se convierten los datos de la columna doornumber de letras a números, los four y two por 4 y 2.

4. Visualización de datos.

En esta sección se da una descripción gráfica de los datos, haciendo uso de caja de bigotes, diagrama de distribución y diagramas de correlación. Primero se observan algunos datos atípicos de la variable de precios, y posteriormente se observa en un gráfico de distribución que los datos poseen una distribución un tanto estrecha, determinándose así que la mayoría de los precios están por debajo de 18000.
El siguiente es un gráfico de correlación:



El gráfico anterior muestra la correlación entre las variables numéricas del dataframe, en el cual se evidencias varias cosas:
- La distancia entre ejes posee una alta relación con la longitud del vehículo, el ancho y la altura. En el siguiente gráfico se puede observar que entre mayor la separación de los ejes, mayores las demás dimensiones del vehículo, algo que tiene sentido ya que se requiere proporcionalidad en el vehículo, tanto por temas de funcionalidad como estéticos. El logaritmo de este valor posee las mismas correlaciones.
- El tamaño del motor también posee una correlación con las dimensiones del vehículo, pero su mayor relación esta con el peso y los caballos de fuerza. En el siguiente gráfico se evidencia que entre mayor el tamaño del motor, mayor peso y también una tendencia a incrementar la cantidad de caballos de fuerza.
- El precio, que es la variable de interés, está fuertemente relacionada con los caballos de fuerza, el tamaño del motor y las dimensiones del vehículo. En el siguiente gráfico, se nota una tendencia de aumento de la variable precio en función a las anteriormente descritas. Mayores dimensiones del vehículo requieren un motor más grande debido a un mayor peso, y por ende mayor cantidad de caballos de fuerza y a su vez un incremento en el precio.

5. Conversión y visualización de variables categóricas.

En esta sección se eliminan las columnas:
'fueltype','carbody','drivewheel','enginetype','cylindernumber','Log_wheelbase','Log_peakrpm','Sq

uare_root_wp','Exp2_stroke','Compressionratio_norm','symboling','aspiration','drivewheel','fuelsystem','doornumber','carheight','stroke','compressionratio','peakrpm','citympg','highwaympg','Car Name','enginelocation'

Ya que algunas columnas contienen información que es muy específica y no aporta mucho a la predicción y otras no tienen información relacionada directamente con la variable 'price'.

Posteriormente, las columnas: 'fueltype','carbody','drivewheel','enginetype','cylindernumber'

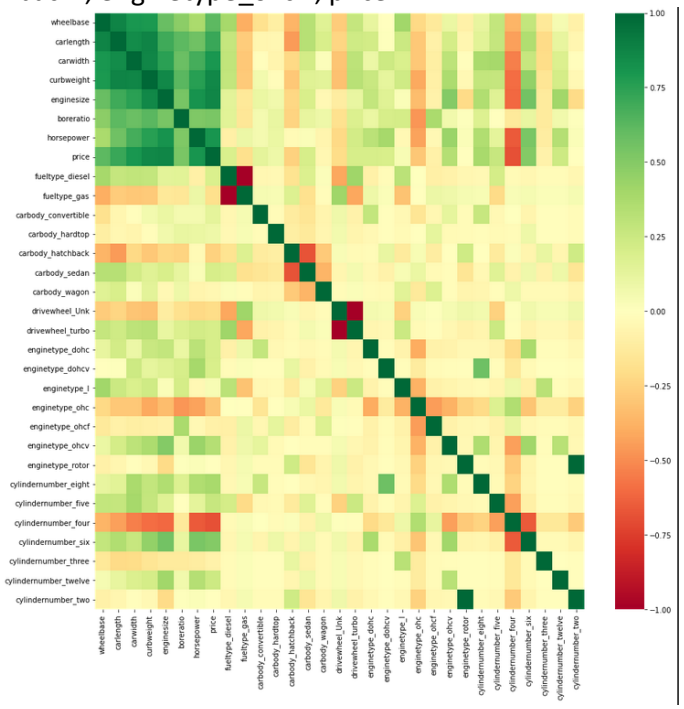
Se convierten de columnas con datos categóricos a columnas con información de cada uno de los descriptores, tal como se puede apreciar en la siguiente imagen:

	fueltype_diesel	fueltype_gas	carbody_convertible	carbody_hardtop	carbody_hatchback	carbody_sedan	carbody_wagon	drivewheel_Unk	drivewheel_turbo
0	0	1	1	0	0	0	0	1	0
1	0	1	1	0	0	0	0	1	0
2	0	1	0	0	1	0	0	1	0
3	0	1	0	0	0	1	0	1	0
4	0	1	0	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...

6. Separación y normalización de datos.

En esta sección se normalizan los datos entre 0 y 1, con el fin de eliminar los problemas de escala entre los datos y obtener así una mejor comparabilidad entre estos. Luego se hace la separación de los datos de train y test de forma aleatoria. Con los datos de train se realiza otro mapa de correlación, fijando la variable Price como la de interés. De este gráfico se determina que las columnas más importantes para predecir el valor de los vehículos son las siguientes:

'wheelbase','carlength','carwidth','curbweight','enginesize','boreratio','horsepower','carbody_hatchback' ,'enginetype_ohcv','price'



Por esto se reacondiciona el df a estar solo con estas 10 columnas.

7. Modelos predictivos.

En esta sección se definen los datos Xtr, Xts, ytr, yts, que son los necesarios para utilizar en los modelos predictivos a trabajar.

El primero a utilizar es un modelo lineal proporcionado por algunas de las librerías de Python.

Con este primer modelo se obtuvo un buen coeficiente de error cuadrático de 0.81 para los datos de test.

```
lr = LinearRegression()
lr.fit(Xtr, ytr)
lr.score(Xtr, ytr), lr.score(Xts, yts)
```

```
(0.8462975550082843, 0.8182875710609149)
```

El objetivo es mejorar este valor y así obtener mejores predicciones y a su vez cumplir de forma satisfactorias con las métricas de trabajo.