EJERCICIO ENTREGABLE BMW – Benjamin Monrabal Orts

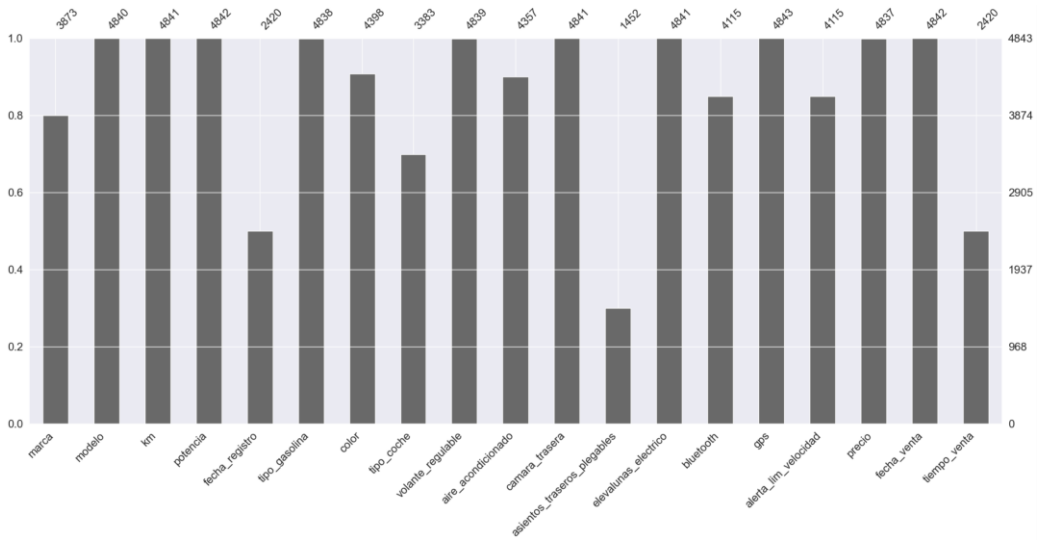
1. **Qué columnas eliminaron**

- Columna “marca”, debido a que se presupone que es un dataset de BMW, y todos los registros son marca BMW.
- A la hora de imputar nulos, se detecta que la columna “asientos_traseros_plegables”, tiene un 70% de nulos, por lo que se considera que esta variable no aporta valor al futuro modelo, y por lo tanto, se ha decidido eliminarla.
- Al hacer la normalización del target (“log_precio”), se elimina la columna original (“precio”).

2. **Qué se hizo con los nulos y cómo se limpiaron las columnas**

Se han tomado diferentes criterios, según la variable:

- **“modelo”:** se detectan 3 nulos en el modelo, y debido a que la cantidad de nulos es menos a un 0,062%, se decide eliminarlos.
- **“km”:** se detectan 2 nulos en el modelo, y debido a que la cantidad de nulos es menos a un 0,04%, se decide hacer una mediana.
- **“potencia”:** se detectan 1 nulo en el modelo, y debido a que la cantidad de nulos es menos a un 0,02%, se decide hacer una mediana. **“tipo_gasolina”:** se detectan 5 nulos en el modelo, y se decide no eliminarlos, e imputarlos a “sin_tipo_gasolina”.
- **“color”:** se detectan 445 nulos en el modelo, y se decide no eliminarlos, e imputarlos a “sin_color”.
- **“tipo_coche”:** se detectan 1460 nulos en el modelo, y se decide no eliminarlos, e imputarlos a “sin_tipo_coche”.
- **“volante_regulable”:** se detectan 4 nulos en el modelo, y debido a que la cantidad de nulos es menos a un 0,08%, se decide eliminarlos.
- **“aire_acondicionado”:** se detectan 484 nulos en el modelo, y debido a que esta columna tiene registros booelanos, se decide imputarlos a “-1”.
- **“camara_trasera”:** se detectan 2 nulos en el modelo, y debido a que la cantidad de nulos es menos a un 0,04%, se decide eliminarlos.
- **“asientos_traseros_plegables”:** se detectan 3391 nulos en el modelo, y debido a que eso es un 70% de toda la columna, se decide eliminar la columna entera, debido a que no aporta al valor al modelo.
- **“elevalunas_electrico”:** se detectan 2 nulos en el modelo, y debido a que la cantidad de nulos es menos a 0,04%, se decide eliminarlos.
- **“bluetooth”:** se detectan 725 nulos en el modelo, y se decide no eliminarlos, e imputarlos a “-1”.
- **“alerta_lim_velocidad”:** se detectan 725 nulos en el modelo, y se decide no eliminarlos, e imputarlos a “-1”.
- **“tiempo_venta”:** columna creada diferencia entre fecha_venta y fecha_registro”, se detectan 2243 nulos y se decide imputar los nulos al valor de la mediana.
- **“precio”:** se detectan 6 nulos, y se decide imputar los nulos al valor de la mediana.



3. Análisis univariable: comentarios, outliers, agrupar…

- Se ha decidido agrupar el índice con la columna “modelo”, y eliminar la columna “modelo”

```
nuevo_indice = bmw6.index.astype(str) + '_' + bmw6['modelo'].astype(str)
bmw6.index = nuevo_indice
```

- Se ha realizado la diferencia entre “fecha_registro” y “fecha_venta” (y eliminados esas dos columnas), para obtener la columna tiempo_venta.
- En la columna tipo_gasolina, cambiamos el registro Diesel por diesel

```
bmw["tipo_gasolina"]=bmw["tipo_gasolina"].replace("Diesel", "diesel", regex=True)
```

- Se detectan 87 outliers en la columna “km” (incluido valores negativos) que se imputan al valor de la mediana.
- Se detectan 591 outliers en la columna “potencia” que se imputan al valor de la mediana.
- En la columna “precio” se consideran precios por encima de 100000€ o por debajo de 2000€, como outliers, un total de 100 outliers, que se imputan al valor de la mediana.

```
outliers2_precio = bmw3[(bmw3["precio"]<= 2000) | (bmw3["precio"] > 100000)]
```

- En la columna tipo_gasolina: se han agrupado aquellos valores que tienen menos de 100 registros.
- En la columna color: se han agrupado aquellos valores que tienen menos de 400 registros.
- En la columna tipo_coche: se han agrupado aquellos valores que tienen menos de 100 registros.

4. Análisis de Correlación inicial

Se ve cierta correlación entre precio y potencia (63.9%) pero no produce alerta roja de que sean variables altamente correlacionadas.

	km	potencia	gps	precio
km	1.000000	-0.050141	0.154815	-0.410189
potencia	-0.050141	1.000000	0.008862	0.639254
gps	0.154815	0.008862	1.000000	-0.005227
precio	-0.410189	0.639254	-0.005227	1.000000

5. Análisis variable vs target

A mayor kilometraje, hay una tendencia descendente del precio, y viceversa, coches con menos km, mayor precio. Con potencia 120, hay una gran variedad de rango de kilometraje y de precio. Los coches de muy baja o muy alta potencia, hay menos rango de precio. Mucha variedad de precio con coches diesel, seguido por coches de petróleo (que en general son de precios bajos cuando hay algo de kilometraje).

6. Transformación de categóricas a numéricas

- La variable tiempo_venta, con formato timedelta64[ns], la convierto a numérico.

```
bmw8["tiempo_venta_int"] = bmw8["tiempo_venta"]/np.timedelta64(1, 's')
```

- Para pasar de valores true/false a 1 y 0, utilizamos la función np.where:

```
bmw9["volante_regulable_int"]=np.where(bmw9["volante_regulable"]==True,1,0 )
```

- Utilizo la función get_dummies de pandas, y se transforman 3 variables dentro de las categóricas, lcat = ['tipo_gasolina', 'color', 'tipo_coche'], y obtener valores 0 y 1.

```
bmw12 = pd.get_dummies(data=bmw11, columns=lcat2)*1
```

7. Normalizar variables numéricas

Se utiliza un MinMaxScaler para las tres variables numéricas: “km”, “potencia” y “tiempo_venta_int”. Se hace cada normalización por separado. Ejemplo de km:

```
bmw13[lnum2]= minMaxResultado_km.fit_transform(bmw13[lnum2])
```

8. Análisis de correlación final, hay alguna variable correlacionada?

Ningún tipo de correlación importante. Algo de correlación tipo_gasolina_electrica y modelo_i3 (70%). Cierta correlación entre el log_precio y la columna “alerta_lim_velocidad_int” (42%).

9. Dataset limpio y preprocesado

```
<class 'pandas.core.frame.DataFrame'>
Index: 4832 entries, 0_118 to 4842_525
Data columns (total 27 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   km                              4832 non-null   float64
 1   potencia                        4832 non-null   float64
 2   tiempo_venta_int                4832 non-null   float64
 3   volante_regulable_int           4832 non-null   int32
 4   aire_acondicionado_int          4832 non-null   int32
 5   camara_trasera_int              4832 non-null   int32
 6   elevalunas_electrico_int        4832 non-null   int32
 7   bluetooth_int                   4832 non-null   int32
 8   gps_int                         4832 non-null   int32
 9   alerta_lim_velocidad_int        4832 non-null   int32
 10  log_precio                      4832 non-null   float64
 11  tipo_gasolina_diesel            4832 non-null   int32
 12  tipo_gasolina_otro_gasolina     4832 non-null   int32
 13  tipo_gasolina_petrol            4832 non-null   int32
 14  tipo_gasolina_sin_tipo_gasolina 4832 non-null   int32
 15  color_black                     4832 non-null   int32
 16  color_blue                      4832 non-null   int32
 17  color_grey                      4832 non-null   int32
 18  color_otro_color                4832 non-null   int32
 19  color_sin_color                 4832 non-null   int32
 20  color_white                     4832 non-null   int32
 21  tipo_coche_estate               4832 non-null   int32
 22  tipo_coche_hatchback            4832 non-null   int32
 23  tipo_coche_otro_coche           4832 non-null   int32
 24  tipo_coche_sedan                4832 non-null   int32
 25  tipo_coche_sin_tipo_coche       4832 non-null   int32
 26  tipo_coche_suv                  4832 non-null   int32
dtypes: float64(4), int32(23)
memory usage: 751.9+ KB
```

	0_118	1_M4	2_320	3_420	4_425
km	0.495195	0.049327	0.646375	0.451568	0.342507
potencia	0.290598	0.461538	0.461538	0.589744	0.803419
tiempo_venta_int	0.339936	0.307818	0.337572	0.307818	0.307818
volante_regulable_int	1.000000	1.000000	0.000000	1.000000	1.000000
aire_acondicionado_int	1.000000	1.000000	0.000000	1.000000	1.000000
camara_trasera_int	0.000000	0.000000	0.000000	0.000000	0.000000
elevalunas_electrico_int	1.000000	0.000000	1.000000	1.000000	0.000000
bluetooth_int	0.000000	1.000000	0.000000	1.000000	1.000000
gps_int	1.000000	1.000000	1.000000	1.000000	1.000000
alerta_lim_velocidad_int	0.000000	1.000000	0.000000	0.000000	1.000000
log_precio	4.053078	4.843233	4.008600	4.399674	4.523746
tipo_gasolina_diesel	1.000000	0.000000	1.000000	1.000000	1.000000
tipo_gasolina_otro_gasolina	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_gasolina_petrol	0.000000	1.000000	0.000000	0.000000	0.000000
tipo_gasolina_sin_tipo_gasolina	0.000000	0.000000	0.000000	0.000000	0.000000
color_black	1.000000	0.000000	0.000000	0.000000	0.000000
color_blue	0.000000	0.000000	0.000000	0.000000	0.000000
color_grey	0.000000	1.000000	0.000000	0.000000	0.000000
color_otro_color	0.000000	0.000000	0.000000	1.000000	1.000000
color_sin_color	0.000000	0.000000	0.000000	0.000000	0.000000
color_white	0.000000	0.000000	1.000000	0.000000	0.000000
tipo_coche_estate	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_hatchback	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_otro_coche	0.000000	1.000000	0.000000	1.000000	0.000000
tipo_coche_sedan	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_sin_tipo_coche	1.000000	0.000000	1.000000	0.000000	1.000000
tipo_coche_suv	0.000000	0.000000	0.000000	0.000000	0.000000