Nombre: Leandro Lopez Totti
Enlace solución: https://github.com/lltotti/Limpieza_Datos_BMW

Proyecto Limpieza de Datos - Fichero BMW

Resolución

1. **Qué columnas eliminaron (en caso se haya eliminado)**

   R: Fueron eliminadas las siguientes columnas:
   - **“Asientos_traseros_plegables”** : por contener 70% de sus valores nulos;
   - **“Marca”:** Por contener proactivamente apenas 1 categoria (“BMW”) - baja varianza/cardinalidad
   - **“Fecha_registro”**: por aportar 50% de nulos que no pueden ser estimados por medio de otras variables. Los demás valores se concentran apneas en 3 categorías básicamente (2012, 2013 y 2014).
   - **“Fecha_venta”**: La variable prácticamente contiene datos referente al año de 2018 y de forma parcial en relación a sus meses. Baja varianza;cardinalidad (año). En respeto al precio, la distribución es equilibrada, lo que lleva a creer que no impacta de modo relevante en la variable precio.
   - “**Gps**”: por baja cardinalidad/varianza y no aportar valor en la predicción.

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

   R: fueron realizadas las siguientes operaciones en las variables:
   - **"precio, tipo_gasolina, volante_regulable, modelo, camara_trasera, elevalunas_electrico,km, potencia, fecha_venta"**: Eliminación de las filas con contenido nulo en las variables por su baja incidencia (<0.2 %) ;
   - **“Asientos_traseros_plegables”**: eliminación de la columna, por su alto contenido de nulos (70%);
   - “**Marca**” : En principio, podríamos remplazar los valores nulos de "marca" por "BMW", que es de que se trata el Dataset. Pero, para garantizar la integridad de los datos, reemplazó por el valor de “Marca” de los mismos modelos en el dataset. Restaron sólo 6 filas , para las cuales fueron imputadas el mismo valor “BMW” por deducción;
   - “**Tipo_coche**”: reemplazo por la moda “tipo_coche” de mismos modelos;
   - **“Color**”: reemplazo por valor “SIN COLOR”;
   - **“Bluetooth”, “alerta_lim_velocidad” y “aire_condicionado”**: por seren variables dicotómicas booleanas (1, 0), los nulos fueron reemplazados por el valor “-1”;
   - **“Fecha_registro”**: la variable fué eliminada por elevado contenido de nulos (50%)

3. **Comentarios del análisis univariable, están todas ok? Hay alguna con outliers? Hay alguna por agrupar?**

   R: Fueran encontrados valores *“Diesel”* y “*diesel*” para la variable “**color**”, que fueron padronizados como “*diesel*”, mismo formado de los demás valores.

También fueron encontrados outliers en las variables “km”, “precio” y “potencia”, para las cuales fueron definidos valores de “piso” y “techo” según parámetros de cuartiles: **“precio”**: piso= 2000 y techo = 50000; “**km**”: piso = 0 y techo = 400000; “**potencia**”: piso= 70 y techo = 300.
. Los valores inferiores al piso y superiores al techo fueron eliminados.

Analizados los histogramas de las variables, fueron verificadas las posibilidades de agrupar contenidos en las siguientes variables, como forma de reducción de la cardinalidad:

- **“Tipo_gasolina”:** agrupación de los valores *"petrol”,”electro”, “hybrid_petrol"* como *“otras”*
- **“Color”:** agrupación de los valores *"red”,”silver”,”beige”, “brown”, “green”, “orange"* como *“otras”*

4. **Análisis de Correlación inicial, hay alguna variable correlacionada?**

R: Hay alguna correlación débil entre “**Potencia**” y “**volante_regulable_BOOL”, “camara_trasera_BOOL”, “elevalauna_electrico_BOOL” y “alerta_lim_velocidad_BOOL**", pero sin sentido aparente.

5. **Análisis variable vs target, hay algún insight interesante?**

   R: Indícios de correlación entre las variables **"km” y “precio"** (inversamente proporcionales) y **"potencia” y “precio"** (directamente proporcionales).

6. **Transformación de categóricas a numéricas, que variables van a transformar? que técnica se va usar?**

   R: Fueron transformadas las variables categóricas **'volante_regulable', 'aire_acondicionado', 'camara_trasera', 'elevalunas_electrico', 'bluetooth', 'alerta_lim_velocidad'** a numéricas (“int8”), por conversion directa de tipo, ya que, en ese momento, contenían categorías entre “1” , “0” y “-1”.

   Las variables “**modelo”, “color”, “tipo_gasolina”, y “tipo_coche”** fueron transformadas a numéricas con utilización de “*LabelEncoder*”, para asignar valores numéricos a cada categoría.

   La variable **“tipo_coche**” fue posteriormente dividida en 8 variables “dummies” (int8), una para cada categoría de coche, con aplicación de la función “*pd.get_dummies()*” .

7. **Normalizar variables numéricas**

   R: La variable “**precio**” estaba afectada por la distribución "right-skweed", por lo que se le aplicó una normalización logarítmica (np.log10()). Para las variables “potencia” y “km” , fue aplicada una normalización “*MinMaxScaler*”

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

   R: Las correlaciones anteriores se mantuvieron, con destaque para “precio” (target) y “km” y “potencia”, aunque con una intensidad un poco más baja.

Surgieron nuevas correlaciones, como entre tipos de coche (ejemplo, ***"tipo_coche_estate"*** *y* ***"tipo_coche_sedan", "tipo_coche_suv", "tipo_coche_hatchback"***) alrededor de -0.40 ; y entre "**modelo**" y "**tipo_coche_suv**" (0.64) , "**modelo**" y "**precio**" (0.49) y "**modelo**" y "**potencia**" (0.55), por ejemplo.

El análisis, todavia, no parece proporcionar una base sólida que justifique eliminar alguna otra variable del conjunto.

9. ****Y finalmente deben poner la lista de columnas completa que tendría su dataset limpio y preprocesado (además del tipo de dato de cada columna) y un pantallazo de las 5 primeras líneas (si esto no entra en 3 páginas lo pueden agregar como anexo)****

```
Data columns (total 20 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   modelo                     4677 non-null   int8
 1   km                         4677 non-null   float64
 2   potencia                   4677 non-null   float64
 3   tipo_gasolina              4677 non-null   int8
 4   color                      4677 non-null   int8
 5   volante_regulable_BOOL     4677 non-null   int8
 6   aire_acondicionado_BOOL    4677 non-null   int8
 7   camara_trasera_BOOL        4677 non-null   int8
 8   elevalunas_electrico_BOOL  4677 non-null   int8
 9   bluetooth_BOOL             4677 non-null   int8
 10  alerta_lim_velocidad_BOOL  4677 non-null   int8
 11  precio_LOG                 4677 non-null   float64
 12  tipo_coche_convertible     4677 non-null   int8
 13  tipo_coche_coupe           4677 non-null   int8
 14  tipo_coche_estate          4677 non-null   int8
 15  tipo_coche_hatchback       4677 non-null   int8
 16  tipo_coche_sedan           4677 non-null   int8
 17  tipo_coche_subcompact      4677 non-null   int8
 18  tipo_coche_suv             4677 non-null   int8
 19  tipo_coche_van             4677 non-null   int8
dtypes: float64(3), int8(17)
memory usage: 223.8 KB
```

	0	2	3	4	5
modelo	3.000000	22.000000	32.000000	34.000000	29.000000
km	0.351978	0.459850	0.320849	0.243031	0.382014
potencia	0.130435	0.217391	0.282609	0.391304	0.673913
tipo_gasolina	0.000000	0.000000	0.000000	0.000000	1.000000
color	1.000000	5.000000	4.000000	4.000000	1.000000
volante_regulable_BOOL	1.000000	0.000000	1.000000	1.000000	1.000000
aire_acondicionado_BOOL	1.000000	0.000000	1.000000	1.000000	1.000000
camara_trasera_BOOL	0.000000	0.000000	0.000000	0.000000	0.000000
elevalunas_electrico_BOOL	1.000000	1.000000	1.000000	0.000000	1.000000
bluetooth_BOOL	-1.000000	0.000000	1.000000	1.000000	1.000000
alerta_lim_velocidad_BOOL	-1.000000	0.000000	-1.000000	1.000000	1.000000
precio_LOG	4.053078	4.008600	4.399674	4.523746	4.232996
tipo_coche_convertible	0.000000	0.000000	1.000000	0.000000	0.000000
tipo_coche_coupe	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_estate	1.000000	1.000000	0.000000	1.000000	1.000000
tipo_coche_hatchback	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_sedan	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_subcompact	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_suv	0.000000	0.000000	0.000000	0.000000	0.000000
tipo_coche_van	0.000000	0.000000	0.000000	0.000000	0.000000