INFORME DE ANÁLISIS EXPLORATORIO DE DATOS (EDA)

NUCLIO DIGITAL SCHOOL – Data Science

Fecha	**:**	2 de enero de 2024	
Proyecto	**:**	Data Engineering: BMW pricing dataset	
Realizado por	**:**	Jorge Guiberteau Apellido2	Pedro Rincón Fernández
Grupo 6		Esther Sanz Apellido2	Toni Vargas Apellido2

1. Análisis inicial

Este proyecto trata el dataset *'bmw_pricing_v3.csv*', que contiene 4846 registros cada uno relacionado con la venta de un automóvil de la marca BMW, de segunda mano, con detalles del modelo, el tipo de combustible, los kilómetros recorridos, los extras, las fechas de registro y de venta y el precio de la venta. El objetivo, es realizar la limpieza y el preprocesamiento de los datos con el fin de generar un modelo de predicción del precio de venta.

2. Columnas eliminadas

Se comprueba que no existen registros duplicados. Se eliminan las siguientes columnas:

- Columna **marca**: Sus valores son 'BMW' o Nulo (20%). No aporta información.
- Columna **fecha_registro:** Tiene un 50% de nulos y no hay manera objetiva de imputarle datos fiables.
- Columna **asientos_traseros_plegables:** Tiene un 70% de nulos no imputables
- Columna **fecha_venta:** Las fechas no tienen continuidad y no aportan información relevante
- Columna **gps:** Analizando la correlación de la variable frente al target (precio), se observa que esta columna no es significativa.

3. Tratamiento de datos nulos

- Columna **color**: 9% nulos no imputables. Se imputa el valor '*sin_info_color*'.
- Columna **tipo_coche:** 30% nulos. La columna es significativa, pero hay demasiada variabilidad para imputar valores a los registros nulos. Se imputa el valor *'otros'*.
- Columna **aire_acondicionado:** 10% nulos. Se imputa el valor '*sin_info'*.
- Columna **bluetooth**: 15% nulos. Se imputa el valor '*sin_info*'.
- Columna **alerta_lim_velocidad:** 15% nulos. Se imputa el valor '*sin_info'*.
- Columna **modelo**: 3 registros nulos. Se imputan manualmente valores para los registros 174 (*modelo 318)* y 4766 (*modelo X1*), basándonos en la tendencia del dataset. Como el registro 4802 no tiene una imputación manual clara, se elimina.
- Columna **km**: 2 registros nulos. Se eliminan.
- Columna **potencia**: 1 registro nulo. Se imputa manualmente el valor *'160'.*

- Columna **tipo_gasolina:** 5 registros nulos. Se imputa el valor *'diesel'* después de analizar la probabilidad para cada registro respecto al dataset.
- Columna **volante_regulable:** 4 registros nulos no imputables. Se eliminan.
- Columna **camara_trasera**: 2 registros nulos. Se imputa el valor *False.*
- Columna **elevalunas_electrico:** 2 registros nulos. Se imputa el valor *True*.
- Columna **precio**: 6 registros nulos. Como son pocos y es el target, se eliminan.

4. Análisis univariable

Las variables **volante_regulable, camara_trasera** y **elevalunas_electrico** son tipo *object,* pero sus valores son *True* y *False*. Se convierten a tipo *boolean*.

Variables numéricas:

- **km**: Se elimina el valor negativo y los outliers por encima de 400.000. Se convierte a *int32* para optimizar el cálculo y almacenamiento.
- **potencia**: Se elimina el registro con valor 0 y los outliers por encima de 400. Se convierte a *int16*.
- **precio (Target)**: Hay 3 registros por encima de 70.000, que se consideran outliers y se eliminan.

Variables categóricas:

- **modelo:** Se agrupan los modelos por su categoría general (por ejemplo *318 Gran Turismo* pasa a ser *318*). Los modelos con una representación inferior al 0.2% se agrupan en la categoría 'otros'. A partir de los 75 modelos iniciales, se obtienen 28 categorías relevantes.
- **tipo_gasolina:** Se agrupan las categorías *'Diesel'* y *'diesel'*. Se eliminan las categorías *'hybrid_petrol'* y *'electro'* ya que tienen muy pocos registros (8 y 3). Finalmente, se obtienen solamente dos categorías: *'diesel'* y *'petrol'*.
- **color:** No se realizan modificaciones. 11 categorías (10 colores y '*sin_info_color'*).
- **tipo_coche:** No se realizan modificaciones. 9 categorías (11 tipos y *'otros'*).
- **aire_acondicionado, bluetooth y alerta_lim_velocidad:** Son categóricas porque se inputó el valor '*sin_info'* a los registros nulos, pero sus valores son *True*, *False* y '*sin_info'.* No se realizan modificaciones.

5. Análisis de correlación inicial

Se observa que las variables con mayor influencia (con mucha diferencia) sobre el precio son la potencia y el kilometraje. La relación de la potencia con el target es directa, la del kilometraje inversa.

6. Análisis Variable vs Target

- **modelo:** Distribución única para cada modelo. Variable significativa.
- **tipo_gasolina:** Diferencias mínimas entre *diesel* y *petrol.*
- **color:** Distribución bastante uniforme. Los colores que se alejan de la tendencia general tienen muy pocos registros.
- **tipo_coche:** Al igual que modelo, parece ser una variable significativa.
- **potencia y km:** Se aprecia claramente la relación de estas variables con el precio de venta, directa e indirecta respectivamente.
- **Extras:** Se observa como el precio tiende a ser mayor cuando el vehículo tiene algún tipo de extra que cuando no lo tiene.

7. Transformación de variables categóricas

Se transforman las variables categóricas (**modelo**, **tipo_gasolina, color**, **tipo_coche, aire_acondicionado**, **bluetooth** y **alerta_lim_velocidad**) a booleanas utilizando el método **One_Hot-Encoding.** Esto hace que se obtenga una columna por cada valor único de las variables categóricas representado por valor 0 (False) o 1 (True) para cada registro. Así se puede procesar el dataset para la construcción del modelo.

8. Normalización de variables numéricas

Se normalizan las variables numéricas (km y potencia) utilizando **MinMaxScaler**. De esta forma, se ponderan los valores de las columnas numéricas asignando valores comprendidos entre 0 y 1, siendo 0 el valor mínimo de la muestra y 1 el valor máximo.

9. Análisis de correlación final

- Se reafirma que la potencia y kilometraje son las variables con mayor relación con el target.
- Se observa la diversa relación de cada modelo y cada tipo de coche con el target, destacando la alta relación directa que tiene el modelo X5.
- También se observa cómo afecta directa o inversamente al precio el hecho de que el automóvil tenga extras o no los tenga.
- Por otro lado, se observa que el tipo de combustible tiene una correlación mínima, por lo tanto, se podría eliminar esta variable.