**EJERCICIO BMW – Benjamin Monrabal Orts**

Enunciado: Objetivo es predecir el precio del vehículo, para ello en este primer entregable se necesitará realizar todo el data cleaning y data preprocessing de los datos (ojo que si bien el objetivo es predecir el precio no será necesario hacer la predicción).

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

Se ha decidido eliminar dos columnas.

* 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.

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

Se han tomado diferente 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 xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide hacer una mediana.
* **“potencia”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide hacer una mediana. **“tipo\_gasolina”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide no eliminarlos, e imputarlos a “sin\_tipo\_gasolina”.
* **“color”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide no eliminarlos, e imputarlos a “color”.
* **“tipo\_coche”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide no eliminarlos, e imputarlos a “sin\_tipo\_coche”.
* **“volante\_regulable”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide eliminarlos.
* **“aire\_acondicionado”:** se detectan 3 nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide eliminarlos.
* **“camara\_trasera”:** se detectan x nulos en el modelo, y debido a que la cantidad de nulos es menos a un 0,062%, se decide eliminarlos.
* **“asientos\_traseros\_plegables”:** se detectan x nulos en el modelo, y debido a que la cantidad de nulos es menos a un 70%, se decide eliminar la columna entera, debido a que no aporta al valor al modelo.
* **“elevalunas\_electrico”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a xxx%, se decide eliminarlos.
* **“bluetooth”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide no eliminarlos, e imputarlos a “-1”.
* **“alerta\_lim\_velocidad”:** se detectan xx nulos en el modelo, y debido a que la cantidad de nulos es menos a un xx%, se decide no eliminarlos, e imputarlos a “-1”.
* **“tiempo\_venta”:** (columna creada diferencia entre fecha\_venta y fecha\_registro”, se decide imputar los nulos a la mediana.
* **“precio”:** se decide imputar los nulos a la mediana.

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

* **“km”:**
* **“potencia”:**
* **“fecha\_registro” y “fecha\_venta”:**
* **“precio:**

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

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

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

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

Utilizo la función get\_dummies de pandas.

Se transforman 4 variables categóricas,

lcat = ['modelo', 'tipo\_gasolina', 'color', 'tipo\_coche']

**Normalizar variables numéricas**

Se utiliza un MinMaxScaler para las tres variables numéricas: “km”, “potencia” y “tiempo\_venta\_int”.

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

Correlación muy alta entre precio y log\_precio (91.9)%

Correlación tipo\_gasolina\_electrica y modelo\_i3 (70%)

**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)**