BMW DATA ENGINEERING

Esta es la documentación adjunta para el notebook con el proyecto de data engineering de un dataset de bmw con el objetivo de predecir el precio de estos vehículos. Explicaremos paso por paso el proceso que se ha llevado a cabo y el motivo por el cual se han tomado cada una de las decisiones.

Exploración y limpieza de valores nulos

Antes de empezar con la limpieza de nulos lo que hemos hecho es convertir las variables temporales a datetime para poder trabajar adecuadamente con ellas, estas eran 'fecha_registro' y 'fecha_venta'. Con ellas extraeremos una nueva variable llamada antigüedad y seguidamente desglosaremos las anteriores en otras como el mes, el dia y el año. Para terminar con las variables temporales generaremos la última columna temporal que será 'año_registro' y sus valores serán el resultado de restar el año de la venta con la antigüedad.

Para seguir con la exploración ya podemos darnos cuenta que hay una serie de variables que no nos interesan para nuestra predicción así como 'marca' (no nos interesa debido a que el dataset es solo de BMW), 'fecha_registro', 'fecha_venta' (las dos variables temporales ya las hemos desglosado, así que seguimos teniendo los valores necesarios) y 'asientos_traseros_plegables' (considero que no puede tener mucho peso a la hora de decidir el precio de un vehículo ya que la mayoría de los vehículos lo tienen, además tenían demasiados valores nulos como para imputar algún valor o eliminarlos, nos quedaríamos con un dataset muy pobre).

Nos encontramos con un número de variables que tienen muy pocos nulos así que podemos proceder a eliminar estos nulos sin temer por la calidad de datos de nuestro dataset. Estas variables son: 'modelo', 'km', 'potencia', 'tipo_gasolina', 'volante_regulable', 'camara_trasera', 'elevalunas_electrico', 'precio'.
Hay que tener en cuenta que la variable precio será la escogida a predecir, entonces no hay otra alternativa que eliminar los nulos ya que sino tendremos datos falsos y el modelo falseará

Ahora ya nos disponemos a lidiar con los valores grandes de nulos, después de mirar muy bien todas las variables existentes, decidí generar una función que lo que hace es coger el porcentaje de la distribución de la variable, y a cada atributo se le aplica el mismo valor porcentual del total de nulos, así conseguimos mantener la misma distribución.
Este método lo he aplicado a las variables: 'color', 'tipo_coche', 'aire_acondicionado', 'bluetooth', 'alerta_lim_velocidad' y 'año_registro.

Una vez teniendo todas estas variables limpias, puedo rellenar los nulos de la variable antigüedad realizando la resta entre 'fecha_venta_año' y 'año_registro'.

Eliminamos las 3 columnas con muchos nulos restantes ya que considero complicado una imputación correcta ya que son temporales y seguidamente convertimos todas las variables a la categoría que les corresponde.

Exploración e imputación de outliers

Empezaremos con la exploración de outliers representando todas las distribuciones, tanto visualizándolas como usando analíticos descriptivos.

En primera instancia vemos que hay que corregir la variable 'tipo_gasolina', ya que tiene un atributo repetido, el diesel, y los agruparemos en un solo atributo. Seguidamente apreciamos un kilometraje negativo, y lo que hacemos es eliminarlo.
Nos encontramos con algo parecido en 'potencia', que hay un vehículo con potencia 0, que al igual que la anterior la eliminaremos, pero esta vez eliminaremos todos los valores inferiores a 50 ya que si hablamos de cv de potencia, es muy raro encontrar vehículos con potencias inferiores a 50cv.
Después lo mismo pasa con antigüedad, que tiene valores negativos, también los eliminamos, solo pueden existir valores positivos cuando hablamos de años.

Volvemos a repetir el análisis y vemos que ahora nos gusta más, pero hay dos variables temporales que no difieren en valores, solo tienen 1 mismo valor para toda la distribución, las eliminamos, estas son: 'fecha_venta_año' y 'fecha_venta_dia'.

Antes habíamos representado con histogramas, ahora dividiremos las variables en diferentes listas según su categoría y realizaremos visualizaciones de las variables numéricas con gráficos de cajas. La variable 'km', tiene muchos outliers pero sólo eliminaremos 1, el que más resalta en la distribución, no hay indicios que los demás no sean reales, son valores que pueden ser tranquilamente kilometrajes de vehículos.

Preprocesamiento

Para empezar con el preprocesamiento del dataset generamos una nueva variable llamada 'log_precio', ya que haciendo la logarítmica de precio conseguimos normalizar la distribución de la variable. Una vez realizado esto, visualizaremos las correlaciones en un mapa de calor y apreciamos que no existen correlaciones pero si una correlación inversa y es la de 'precio' y 'año_registro', entendemos que el precio estará muy condicionado a esta variable.

Volveremos a realizar una exploración de variables, ahora toca hacerlas todas numéricas para que el modelo pueda entenderlo y veremos la relación de cada una de ellas con la variable precio.

En las categóricas vemos que el modelo tiene impacto sobre el precio, pero hay tantos que es difícil de leer, en 'tipo_gasolina' también se aprecia que el diesel tiene un precio superior, a parte de tener muchos más valores que los demás atributos de su variable. Hablando del color de los coches, hay dos que salen de la media, el gris y el azul. 'Tipo_coche', el sur es el más caro y en dia de la semana el miércoles.
Se aprecian diferencias, en cada categoría hay uno o dos atributos que el precio se dispara más, pero yo no las consideraría muy importantes, el resto son muy lineales entre ellas.

Al ver que la variable 'modelo' tenia demasiados atributos me propuse realizar un cambio para poder tener todo un poco más organizado, aunque vi que el impacto no fue tampoco positivo la eliminé al momento pero generé una variable agrupando todos los modelos en subgrupos y volví a hacer unas representaciones para poder ver como actuaban.

Continuamos con las representaciones de las variables numéricas por precio, los inputs más destacables son que a menos kms el precio es más alto, igual que a más potencia también.
Hay alguna variable que no se aprecia muy bien por el tipo de representación así que realizaremos alguna extra.
Apreciamos que el miércoles es el dia que se vendieron los vehículos más caros, pero no creo que sea muy representativo, hablando de meses, el pico más alto se encuentra en Agosto y el más bajo en enero. Lo más remarcable podría decirse que es la relación entre precio y año_registro junto con antigüedad, dónde vemos que a más nuevo más caro y para terminar también es de interés la relación entre antigüedad y km.

Seguimos con las booleanas y no hay datos relevantes para ver que aumenten el precio por tener según que extras, ya que lo más lógico seria que por tener x extra el precio final aumentara.
Entonces eliminamos las variables: 'alerta_lim_velocidad', 'gps', 'bluetooth', 'elevalunas_electrico', ‘camara_trasera’.

Ahora ya empezaremos con el verdadero preprocesamiento, generaremos de nuevo unas listas para cada tipo de variable, una para numéricas, otra para booleanas y una categórica, excluyendo la variable a predecir, es decir, el target (precio).
Primero de todo, hay que convertir las booleanas a numéricas, pasándolas a int ellas solas se convierten en 0 y 1, siendo False y True respectivamente.
Para las categóricas usaremos la función get.dummies, que convierte cada atributo de las variables categóricas en columnas y asigna un 1 o un 0 según sea cierto que cumplen esa característica o no.
Y para las numéricas usaremos el minmaxscaler, que lo que hace es de todas las distribuciones numéricas las redistribuye entre 0 y 1.

Una vez hecho todo esto, el dataset estará limpio y preprocesado, volveremos a ver las correlaciones, y vemos que sigue habiendo solo dos variables que se correlacionan inversamente entre sí, precio y antigüedad. Así que el precio estará firmemente condicionado por la antigüedad.

ANEXO

Lista de columnas

km float64
potencia float64
volante_regulable int64
aire_acondicionado int64
precio int64
antiguedad float64
fecha_venta_mes float64
año_registro float64
log_precio float64
modelo_ Active Tourer uint8
modelo_114 uint8
modelo_116 uint8
modelo_118 uint8
modelo_120 uint8
modelo_123 uint8
modelo_125 uint8
modelo_135 uint8
modelo_214 Gran Tourer uint8
modelo_216 uint8
modelo_216 Active Tourer uint8
modelo_216 Gran Tourer uint8
modelo_218 uint8
modelo_218 Active Tourer uint8
modelo_218 Gran Tourer uint8
modelo_220 uint8
modelo_220 Active Tourer uint8
modelo_225 uint8
modelo_225 Active Tourer uint8
modelo_316 uint8
modelo_318 uint8
modelo_318 Gran Turismo uint8
modelo_320 uint8
modelo_320 Gran Turismo uint8
modelo_325 uint8
modelo_325 Gran Turismo uint8
modelo_328 uint8
modelo_330 uint8
modelo_330 Gran Turismo uint8
modelo_335 uint8
modelo_335 Gran Turismo uint8
modelo_418 Gran Coupé uint8
modelo_420 uint8
modelo_420 Gran Coupé uint8
modelo_425 uint8
modelo_430 uint8
modelo_430 Gran Coupé uint8
modelo_435 uint8
modelo_435 Gran Coupé uint8
modelo_518 uint8
modelo_520 uint8
modelo_520 Gran Turismo uint8
modelo_523 uint8
modelo_525 uint8
modelo_528 uint8
modelo_530 uint8
modelo_530 Gran Turismo uint8
modelo_535 uint8
modelo_535 Gran Turismo uint8
modelo_630 uint8
modelo_635 uint8
modelo_640 uint8
modelo_640 Gran Coupé uint8
modelo_650 uint8
modelo_730 uint8
modelo_735 uint8
modelo_740 uint8
modelo_750 uint8
modelo_ActiveHybrid 5 uint8
modelo_M135 uint8
modelo_M235 uint8
modelo_M3 uint8
modelo_M4 uint8
modelo_M5 uint8
modelo_M550 uint8
modelo_X1 uint8
modelo_X3 uint8
modelo_X4 uint8
modelo_X5 uint8
modelo_X5 M uint8
modelo_X5 M50 uint8
modelo_X6 uint8
modelo_X6 M uint8
modelo_Z4 uint8
modelo_i3 uint8
modelo_i8 uint8
tipo_gasolina_diesel uint8
tipo_gasolina_electro uint8
tipo_gasolina_hybrid_petrol uint8
tipo_gasolina_petrol uint8
color_beige uint8
color_black uint8
color_blue uint8
color_brown uint8
color_green uint8
color_grey uint8
color_orange uint8
color_red uint8
color_silver uint8
color_white uint8
tipo_coche_convertible uint8
tipo_coche_coupe uint8
tipo_coche_estate uint8
tipo_coche_hatchback uint8
tipo_coche_sedan uint8
tipo_coche_subcompact uint8
tipo_coche_suv uint8
tipo_coche_van uint8
fecha_venta_nombredia_Friday uint8
fecha_venta_nombredia_Monday uint8
fecha_venta_nombredia_Saturday uint8
fecha_venta_nombredia_Sunday uint8
fecha_venta_nombredia_Thursday uint8
fecha_venta_nombredia_Tuesday uint8
fecha_venta_nombredia_Wednesday uint8
dtype: object

	km	potencia	volante_regulable	aire_acondicionado	precio	antiguedad	fecha_venta_mes	año_registro	log_precio	_ Active Tourer	...	tipo_coche_subcompact
0	0.289039	0.095238	1	1	11300	0.153846	0.000	0.846154	4.053078	0	...	0
1	0.027787	0.703081	1	1	69700	0.115385	0.125	0.884615	4.843233	0	...	0
2	0.377621	0.151261	0	0	10200	0.153846	0.125	0.846154	4.008600	0	...	0
3	0.263476	0.193277	1	1	25100	0.115385	0.125	0.884615	4.399674	0	...	0
4	0.199573	0.263305	1	1	33400	0.115385	0.375	0.884615	4.523746	0	...	0

5 rows × 114 columns