Minería De Datos

INFORMACIÓN SOBRE VEHÍCULOS ELÉCTRICOS

Diego Cordero Contreras

Enrique Albalate Prieto

Lucía De Ancos Villa

Pablo Del Hoyo Abad

Mohamed Essalhi Ahamyan

Índice:

1. Introducción

En este segundo entregable de la práctica de minería de datos, debemos tratar los dataframes que hemos obtenido en el primer entregable.
Primero, preprocesaremos los dataframes, debemos limpiar datos que no utilizaremos y tomar decisiones en cuanto a los valores nulos del dataset. A continuación, tomaremos los datasets con el objetivo de crear tarjetas de datos con la información necesaria para desarrollar las hipótesis definidas.

En nuestro caso, la manera más cómoda de trabajar será creando cuatro tarjetas de datos, una por hipótesis. Esto se debe al hecho de que tenemos una gran cantidad de columnas en nuestros datos, muchos de ellas inconexas, por lo que puede hacer que el proceso de creación de una tarjeta de datos única para todo el análisis sea complejo y haga que la tarjeta sea incomprensible.

2. Limpieza y transformación de dataframes

En todos los dataframes hemos tenido que estandarizar ciertos datos que serían de uso común, como es el caso de los países. Hemos elegido como idioma común el inglés, de manera que el nombre de los países deberá traducirse al inglés en caso de ser necesario. Además, cada dataframe en concreto ha tenido su proceso de limpieza y estandarización particular:

2.1. Data_CO2

Se han limpiado los datos relacionados con la emisión de CO2 con el objetivo de mantener exclusivamente los que están dentro de nuestro alcance, es decir, los comprendidos entre 2015 y 2023. Además, se han eliminado las ultimas líneas, ya que contenían información irrelevante en nuestro caso, porque se trataba de datos globales en lugar de clasificados por países. Finalmente, conseguimos los datos de CO2 emitidos por país y por año.

2.2. Electricity_car_data_clean

Este dataframe contiene la información de todos los vehículos eléctricos a la venta en los últimos tiempos, es decir, no contamos con los vehículos eléctricos de todos los tiempos pues hay muchos que llevan sin fabricarse mucho tiempo, solo hemos tenido en cuenta los actuales. Los datos en el estado que queríamos. Esto es debido a que hemos usado el api de Kaggle, lo cual nos ha ahorrado problemas de datos nulos o mal recogidos por el webscrapping; y no es un dataframe que venga dado por años, con lo que no hemos tenido que ajustarlo.

2.3. Electricity_all_countries

El conjunto de “Electricity_all_countries”, contiene datos relativos al coste de la electricidad por megavatio/hora. Al contener los datos de cada país por meses y nuestro alcance estar definido en años, hemos decidido hacer la limpieza calculando la media del precio por año. De esta forma, conseguimos obtener los precios de la electricidad por megavatio/hora clasificados por país y año.

2.4. Gasoline_Diesel_prices

Este dataframe contiene los precios de la gasolina y diésel en los países que mencionamos en nuestro alcance durante los años comprendidos entre 2015 y 2023. Ha sido necesaria una limpieza pues al haber utilizado la técnica de webscrapping, algunos caracteres estaban codificados en utf 8 mientras que nuestra bbdd está en utf-8, causando incompatibilidades con el resto. Hemos eliminado estos caracteres y cambiado el tipo de datos a los caracteres numéricos.

2.5. Level_Of_Studies

En este dataframe tenemos el nivel de estudios por países, indicando el máximo nivel de estudios en porcentaje de la población. Hemos eliminado columnas que no podremos diferenciar, en este caso la distinción entre hombres y mujeres en sus estudios, ya que no sabemos el sexo del comprador de un vehículo.
Hemos eliminado el símbolo de porcentaje que aparecía y hemos transformado los valores de las columnas a enteros para poder trabajar de manera más fácil.
Por último, se ha transformado el nombre de las columnas eliminando el carácter '__' que aparecía al final del nombre de las columnas.

2.6. Model_Per_Year

La primera tarea que se tuvo que realizar fue la de establecer un valor común para los nulos. En el conjunto de datos, este tipo de valores aparecen representados tanto por el símbolo '-' como por 'N/A'. A continuación, eliminamos el separador de millares que contenían los números porque obligaba a que el tipo de datos fuera 'category' cuando, en realidad, estábamos interesados en que fuesen de un tipo de dato numérico. En este caso concreto, como la página web donde se han obtenido los datos es estadounidense, utilizaban la coma. Además, descubrimos que, para algunos modelos en un país y año determinados, el número de unidades vendidas era negativo. Como no tiene sentido, decidimos sustituirlos por nulos.

Existe una columna denominada PowerTrain que indica el tipo de motor que utiliza el modelo. Sin embargo, el nivel de granularidad utilizado es mayor al que necesitamos para comprobar la hipótesis. Por ejemplo, la clasificación diferenciaba entre híbrido enchufable e híbrido no enchufable, pero nosotros los tratamos como eléctricos. Es por ello que decidimos crear una nueva columna denominada "isEv" que indicase si el modelo es eléctrico o no.

Por último, decidimos sustituir los valores nulos por cero ya que, para este estudio, estamos interesados en la cantidad de vehículo vendidos en un país y en un año determinado y no tanto en los modelos concretos. Por lo tanto, al agregar los datos, el valor de cero permite que ignoremos los nulos.

2.7. PIB_per_capita

Tras la obtención de los datos relacionados con el producto interior bruto per capita de países y diversas zonas geográficas, se realizó una poda de columnas a modo de filtro de los años manejados. En concreto, únicamente se deseaba tratar datos desde 2015 hasta 2023, pero fue necesario tener en cuenta a su vez registros de los 5 años anteriores y las previsiones para los 5 siguientes, como se explicará a continuación.

Debido a que al estado inicial del dataframe le faltaban ciertas mediciones y estas estaban sustituidas por la cadena de caracteres 'no data', se pensó que una buena solución al problema sería en primer lugar establecer todas las apariciones de dicha cadena por el valor nulo de la librería de Python pandas (nan) y en segundo lugar cambiarlos de forma automática por la media del producto interior bruto del país al que se refiere la fila del dato nulo dentro del rango temporal más arriba mencionado (desde 2010 hasta 2028). El objetivo de todo este proceso fue la generación de valores lo más realistas posibles.

2.8. Puntos_De_Carga

En este dataset en el que almacenamos el número de puntos de carga que existen por país, se han eliminado algunas columnas que se utilizaban en la página web desde donde se descargó el dataset para su control pero que en nuestro caso no son de utilidad. Estas columnas son: category, parameter, mode, unit.

También se han eliminado las columnas 'value' con valores decimales, se atribuyen a errores en la recogida de datos.

Por último, se ha elegido para cada país el rango de años entre 2015 y 2021. Como existen países que no empezaron a recoger datos hasta años posteriores, se ha creado una simulación de los puntos de carga de los datasets. El procedimiento ha sido el siguiente, se ha elegido el primer año donde se tenían datos y desde el 2015 hasta el primer año donde se tenían datos se han ido generando números aleatorios que iban creciendo hasta llegar al valor del primer registro. Tras no encontrar más datos, creemos que simular el crecimiento es lógico ya que los puntos de carga de coches eléctricos están en construcción y nos permitirá tener más datos para obtener conclusiones.

3. Líneas de Trabajo

3.1. Tarjeta Hipótesis 1

Se venden más vehículos por el aumento del rendimiento de los coches eléctricos

Para la comprobación de nuestra hipótesis número 1, hemos creado esta tarjeta de datos que contiene todos los datos extraídos de los datasets anteriores que hemos considerado de más relevancia para comprobar esta hipótesis. Estos son los datos relativos a la venta de coches eléctricos exclusivamente y sus características.

Para crear esta tarjeta hemos tenido que unir información que estaba en concreto en dos datasets anteriores. Los dos datasets hablaban de vehículos, uno de todas las ventas de vehículos en países y otro de las características exclusivamente de coches eléctricos. En el primer datasets agrupamos todos los datos en función del modelo, quedando solo el modelo y las ventas que tuve en sus respectivos años; mientras que en el segundo datasets tan solo eliminamos la marca y nos quedamos con el modelo y los datos que más nos interesaban del mismo.

Para juntarlos ha habido problemas ya que en que en cada dataset, las versiones de los propios modelos se escribían de formas distintas, es por ello que para cada modelo de coche que tuviera más de una versión, se han hecho la media de sus características y se ha añadido únicamente el modelo como tal, como es en el caso del tesla model 3.

De esta manera nos quedamos con el siguiente diccionario de datos:

Nombre del campo	Tipo de dato	Descripción
Model	String	Indica el modelo de coche del que hablan las características.
Year	Int	Es el número de ventas que el modelo tuvo ese año mundialmente.
AccelSec	Float	El tiempo en segundos que el coche tarda en acelerar desde 0kmh hasta 100kmh.
TopSpeed_Kmh	Int	La velocidad máxima alcanzable con el coche en kmh.
Range_Km	Int	Autonomía de la batería en Km bao el ciclo WLTP.
Efficiency_Kwh	Float	Indica la carga del vehículo (con ello el tiempo de recarga) en kwh.
Seats	Int	Número de asientos en el vehículo.
PriceEuro	Int	Precio de venta al público en euros.
FastCharge_KwH	Float	Indica, si tiene, los watios a los que el coche puede cargar con un fast charger.
RapidCharge	Boolean	Indica si tiene carga rápida.
PowerTrain	String	Indica las ruedas del coche que traccionan, delanteras, traseras o 4x4.
PlugType	String	Indica el tipo de enchufe que tiene para cargar.
BodyStyle	String	Indica la carrocería del coche.
Segment	String	Indica el segmento en el que está el coche.

3.2. Tarjeta Hipótesis 2

Se venden más vehículos por el aumento de los precios del carburante

Se ha creado una tarjeta de datos con la finalidad de agrupar la información relevante para aceptar o refutar esta hipótesis. Para ello, hemos seleccionado los datasets correspondientes a los precios de la gasolina y diésel y a las ventas de coches eléctricos y sus modelos en diferentes países del mundo, ambos del esquema SILVER.

El proceso consistió en unir las dos tablas mencionadas, por medio de la columna que contiene los países. Además, solo hemos conservado el periodo de años común en ambas tablas (2017-2022).

Finalmente, obtenemos el siguiente diccionario de datos correspondiente a la tarjeta creada para la hipótesis 2:

Nombre del Campo	Tipo de Dato	Descripción
Country	String	Contiene los nombres de los países manejados y se utilizará para identificarlos.
CochesVendidos_[year]	Double	Contiene la cantidad de vehículos oficialmente vendidos para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”.
Gasolina_[year]	Double	Almacena el valor del precio de la gasolina para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”
Diesel _[year]	Double	Funciona como las columnas Gasolina_[year], pero en este caso almacenando el coste de Diesel.

3.3. Tarjeta Hipótesis 3

Predicción del número de puntos de carga en base a las ventas de vehículos eléctricos

Esta hipótesis ha variado ligeramente, para introducir distintos tipos de hipótesis, queremos tratar de crear un modelo que prediga el número de puntos de carga que se instalarán en un país dependiendo de las ventas de vehículos eléctricos el año anterior.

Se han unido las tablas que contienen las ventas de coches y de los puntos de carga. Primero, se han agrupado las ventas creando una nueva tabla con 2 tipos: Hybrid y Electric que sustituye a la columna PowerTrain en el dataset de ventas de coches. Además, se ha tenido que pivotar ambas tablas: en la tabla de ventas de coches los años aparecían como columnas, ahora se han cambiado creando una nueva columna llamada year donde aparece el año estudiado y se ha creado otra columna llamada Sells donde aparece el número de ventas.

Finalmente, se han agrupado los dos datasets utilizando como nexo las columnas Country y year (país y año). La tarjeta de datos pertenece al intervalo de años entre 2017 y 2022. El diccionario de la tarjeta de datos para la hipótesis 3 es la siguiente:

Nombre del Campo	Tipo de Dato	Descripción
Country	String	Almacenará el nombre de todos los países manejados y que servirá para identificarlos
Type_Vehicle	String	Contiene el tipo de vehículo que se ha vendido, puede ser Hybrid o Electric
year	Int	Contiene el año de estudio en la fila
Sells	Int	Es el número de ventas de coches
Fast Charging Point	Int	Es el número de puntos de carga rápidos distribuidos en el país
Slow Charging Point	Int	Es el número de puntos de carga lenta distribuidos en el país

3.4. Tarjeta Hipótesis 4

Se venden más vehículos en países con más PIB per cápita

Con el objetivo de contar únicamente con los datos necesarios para contrastar la hipótesis de manera cómoda, clara y rápida se decidió unir la tabla que contiene datos referentes a ventas de coches eléctricos y sus modelos en diferentes países del mundo junto con aquella destinada a reflejar la evolución del producto interior bruto de dichas zonas.

La columna que sirvió de nexo entre ambas fue en la que se detallaban los países a los que los datos de una determinada fila pertenecían. Además, se recortó finalmente el intervalo de años entre los que se puso el foco, desde 2017 hasta 2022. La estructura de la tarjeta de datos desarrollada para el contraste de la hipótesis 4 es la siguiente:

Nombre del Campo	Tipo de Dato	Descripción
Country	String	Almacenará el nombre de todos los países manejados y que servirá para identificarlos
CochesVendidos_[year]	Double	Contiene la cantidad de vehículos oficialmente vendidos para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”

PIB_[year]	Double	Contiene el valor del producto interior bruto oficial para un año concreto ([year]) y un determinado país, es decir, el que aparece en la columna “Country”

- [year] representa a cualquier año comprendido entre 2017 y 2022

En adición se requirió un tratamiento adicional que afectó al contenido de la nueva tarjeta de datos, pero no a su estructura. Fue detectada la ausencia en ella de ciertos países que eran comunes a las dos tablas que sirvieron de fuente de información, por lo que acabó comprobándose que tenían ciertos detalles diferentes en su redacción dentro de la columna “Country” de dichas tablas. Esto se resolvió usando un diccionario que aplicara reemplazos una de ellas en la columna referida a países (por ejemplo, desde "China, People's Republic of" a “China”).

Finalmente, se rellenaron ciertas celdas de filas de la tarjeta de datos final en las que aparecía un 0 representando los coches vendidos de algunos años concretos (únicamente ocurría para Puerto Rico y Uzbekistán). Como valor para completar la cada respectiva sustitución se empleó la media de coches vendidos no nulos (en este caso marcados con 0) de esa misma fila.

4. Profiling Tarjetas de Datos

El profiling es una técnica para analizar una aplicación, en este caso medimos la calidad de los datos de las tarjetas de datos generados durante el proceso KDD hasta el momento. Este tipo de reportes pueden ser interesantes de cara a conocer métricas que nos serán útiles para obtener conclusiones en el siguiente entregable del proyecto.

Lo más interesante del profiling que nos ofrece la librería pandas-profiling es la serie de alertas indicándonos atributos de los datos. Estas alertas serán las que analizaremos en este punto.

Todos los profiling de las tarjetas de datos se pueden encontrar en la siguiente carpeta del repositorio de github.

4.1. Profiling Tarjeta de Datos 1

El profiling del primer dataset nos muestra principalmente los siguientes avisos

- Hay mucha correlación entre variables, 16 alertas son de gran correlación entre variables, algunas referidas a los años, lo cual es lógico ya que se presupone que los datos de un año a otro no tendrán una gran variación. También tenemos correlación entre algunas variables que nos indican las características de los coches, por ejemplo la aceleración de 0 a 100 (accelSec) del coche con el precio o con el tipo (segment) del coche

- También nos avisa de que hay varias columnas con ceros en las ventas del año, esto se debe a que no tenemos datos de las ventas en algunos casos por la complicación de obtener estos.

4.2. Profiling Tarjeta de Datos 2

El profiling del segundo dataset nos muestra principalmente los siguientes avisos

- Hay mucha correlación entre variables, 18 alertas son de gran correlación entre variables, algunas referidas a los años, lo cual es lógico ya que se presupone que los datos de un año a otro no tendrán una gran variación. Esto se presenta en los coches vendidos, en el precio de la gasolina y en el precio del diesel
- El segundo tipo de alerta que obtenemos es de tipo Unique en distintas columnas del dataset. Esta alerta nos indica que hay varias columnas cuyos valores son todos únicos, se puede entender debido a la variación del precio de la gasolina, del diésel y de las ventas ya que sería extraño que fuesen iguales

4.3. Profiling Tarjeta de Datos 3

El profiling del tercer dataset nos muestra principalmente los siguientes avisos

- Hay correlación entre los puntos de carga rápidos con las ventas y con el número de puntos de carga lentos. En una primera exploración podríamos pensar que es debido a que cuantos más puntos de carga rápidos haya, más infraestructura, más concienciación con los vehículos eléctricos y por tanto más ventas.
- También nos avisa de que hay varias columnas con ceros en las ventas y en el número de puntos de carga rápidos, esto puede ser que se deba a las restricciones de los países, se ha comprobado a mano que en muchos casos son países en los que suele haber pocos recursos para este tipo de vehículos (India, Tailandia…)

4.4. Profiling Tarjeta de Datos 4

- Hay mucha correlación entre variables, 12 alertas son de gran correlación entre variables, algunas referidas a los años, lo cual es lógico ya que se presupone que los datos de un año a otro no tendrán una gran variación. Esto se presenta en los coches vendidos y en el PIB.
- El segundo tipo de alerta que obtenemos es de tipo Unique en distintas columnas del dataset. Esta alerta nos indica que hay varias columnas cuyos valores son todos únicos, se puede entender debido a la variación de las ventas de coches por año y del pib por año.

5. Metodología Seguida

En este segundo entregable la metodología seguida fue la siguiente. Se eligió una base de datos donde almacenaríamos los datasets encontrados en el primer entregable, en este caso, se contrató un servidor VPS donde se ha creado un contenedor de docker que contiene una imagen de la base de datos MySQL 8.0. En esta base de datos el preprocesamiento se hace de la siguiente forma:

- [RAW]: en este esquema de la base de datos almacenamos los datasets de la misma forma en la que se descargaron.
- [SILVER]: tomamos los datasets de [RAW] y los limpiamos y transformamos según las directrices especificadas en el punto 2.
- [GOLD]: tomamos los datasets de [GOLD] y creamos las tarjetas de datos según las directrices especificadas en el punto 3.

Además, con las tarjetas de datos creadas, se ha hecho un profiling para analizar la calidad de los datos, de este modo, en el siguiente punto podremos partir con un conocimiento que no teníamos a la hora de desarrollar nuestras hipótesis.

6. Participación

Alumno	Participación (%)
Diego Cordero Contreras	20
Enrique Albalate Prieto	20
Lucía De Ancos Villa	20
Pablo Del Hoyo Abad	20
Mohamed Essalhi Ahamyan	20