MINERÍA DE DATOS

BlaBlaCar Pooling

María Garrido Arcos

Sergio Jiménez Fernández

Ismael Pérez Nieves

David Pinto Camacho

CONSIDERACIONES PREVIAS

Previamente a explicar los avances para este hito 2, es necesario tener en cuenta los siguientes cambios realizados sobre la entrega anterior. Las consideraciones son las siguientes:

1. **Nuevos datasets**: Con el fin de calcular nuevos datos para los fines del estudio, hemos incluido dos nuevos datasets. Los datasets en cuestión son los siguientes:
    a. Latitud y longitud de los pueblos de España. Este dataset contiene información acerca de la latitud y longitud de los diferentes pueblos de España, como el nombre indica. Este dataset será usado para calcular posteriormente (mediante la fórmula del semiverseno) la distancia entre el pueblo de origen y de destino.
    b. World Cities (Versión Básica). En este dataset se incluyen datos como (principalmente) el país al que pertenece un pueblo/ciudad, la latitud y longitud de dicho pueblo. Este nos será también útil para el cálculo de la distancia entre ciudades así como el país a la que pertenece dicha ciudad.
    c. Transporte Interurbano Autobús en España. En este dataset se incluyen más datos acerca del volumen de viajeros en España que usan el autobús interurbano (*Tabla 3: Viajeros transportados comparación anual, Pág 19*).

2. **Reformulación de las hipótesis**: Hemos decidido reformular las hipótesis para tener una estructuración más formal. Quedando tal que:
    a. Analizar si el Carpooling de BlaBlaCar supone un porcentaje considerable (5%) del volumen de movilizaciones por tierra.
    Más formalmente:

$$H_0: \frac{Viajeros_i}{ViajerosTP_i + Viajeros_i} > 0,05 \; / \; i \in [2017, 2019]$$

$$H_a: \frac{Viajeros_i}{ViajerosTP_i + Viajeros_i} <= 0,05 \; / \; i \in [2017, 2019]$$

    Donde $Viajeros_i$ se refiere a la variable que contabiliza las personas desplazadas por cada año y $ViajerosTP_i$ el número de desplazamientos por transporte público (ferroviario y autobús) por año.

    Consideramos que un 5% es una cifra aceptable, teniendo en cuenta el gran impacto que supone el transporte público en este país. Además, de la capacidad del transporte público para mover más gente que el carpooling.

b. Tratar de analizar el impacto medioambiental que ha supuesto el uso de carpooling dentro de las emisiones por vehículos terrestres a lo largo de estos años.
Formalmente:

$$H_0 : \frac{Emisiones_i}{EmisionesN_i} < 0,1 / i \in [2017, 2019]$$

$$H_a : \frac{Emisiones_i}{EmisionesN_i} >= 0,1 / i \in [2017, 2019]$$

Donde $Emisiones_i$ se refiere a las emisiones (en kilogramos) de CO_2 por las movilizaciones de Carpooling por año, y $EmisionesN_i$ se refiere a las emisiones (en kilogramos) del transporte público nacional por cada año.

Consideramos que un 10% es una cifra interesante pues quizá el ahorro de compartir coche con 5 personas en lugar de 5 coches, lo que debería suponer una influencia considerable.

SELECCIÓN DE CARACTERÍSTICAS

Para llevar a cabo las hipótesis planteadas anteriormente, haremos uso de las siguientes variables obtenidas tras varias iteraciones en los conjuntos de datos especificados anteriormente:

1. BlaBlaCar
    - *Día*. Es necesaria para agrupar los datos cronológicamente.
    - *ORIGEN*. Es necesaria para obtener el país de origen.
    - *DESTINO*. Es necesaria para obtener el país de destino.
    - *Asientos_confirmados*. Es necesaria para evaluar la cantidad de personas que hacen uso de la plataforma BlaBlaCar.
    - *Viajes_confirmados*. De igual manera que la variable anterior, esta nos permitirá evaluar las personas que hacen uso de la plataforma.
2. Movilidad
    - *Modo*. Nos permite diferenciar el modo de transporte. Los diferentes modos que tiene en cuenta son: carretera, ferroviario, aéreo y marítimo.
    - *Year*. Es necesaria para obtener un orden cronológico.
    - *Millones_viajeros_kilómetro*. Nos permitirá comparar las tendencias de viajes.

3. Emisiones
    - *Modo*. Nos permite diferenciar el modo de transporte. Los diferentes modos que tiene en cuenta son: carretera, ferroviario, aéreo y marítimo.
    - *Year*. Es necesaria para obtener un orden cronológico.
    - *NombreGas*. Nos permite escoger los gases necesarios para comparar las emisiones. Los diferentes gases son: Dióxido de carbono, Metano, Óxido nitroso, Óxido de azufre, Óxidos de nitrógeno, Amoníaco, Compuestos orgánicos volátiles no metánicos, Monóxido de carbono, Material particulado <= 2.5 µm, Material particulado <= µm.
    - *Valor*. Nos permitirá comparar diferentes emisiones.
    - *TipoTráfico*. Nos permitirá saber si se trata de datos de movilizaciones internacionales o nacionales.
4. Latitud y longitud de los pueblos en España
    - *Población*. Nos permite ubicar el pueblo en cuestión.
    - *Latitud*. Nos permitirá localizar una ubicación.
    - *Longitud*. Nos permitirá localizar una ubicación.
5. World Cities
    - *Country*. Obtenemos el país de la ubicación buscada.
    - *Lat*. Nos permitirá localizar una ubicación.
    - *Lng*. Nos permitirá localizar una ubicación.
6. Transporte Interurbano Autobús en España
    - *Year*. Es necesaria para obtener un orden cronológico.
    - *Total anual*. Nos permitirá comparar las tendencias de viajes.

PREPROCESAMIENTO Y TRANSFORMACIÓN

El siguiente paso consiste en el preprocesamiento de los datos. En nuestro caso, el resultado del preprocesado se resumen en:

- **Limpieza de datos**: Principalmente sobre el dataset original de BlaBlaCar, en el que se han filtrado varios errores de escritura en los valores de las columnas que no nos permitían transformar los datos en pasos posteriores (registros que le añaden sufijos al nombre de la ciudad, como ‘CP’; nombres de ciudades mal escritos, sobre todo silabas rusas como ‘iy’ que debería ser ‘y’; además de caracteres inadmisibles, como letras rusas).
- **Codificación de valores categóricos**: Como fueron los casos de la variable *Modo* del dataset de emisiones y de movilidad, y las variables *TipoTráfico* y *NombreGas* del dataset de emisiones.

- **Comprobación de valores nulos**: Como ocurrió con la variable Millones_viajeros_kilometro en el dataset de movilidad. Como se trata de una variable numérica, se procedió a emplear la media de los valores anteriores para evitar valores nulos.

Con lo que respecta a la transformación, se han realizado varios procesos:

- **Enriquecimiento con otras fuentes**: Para los casos de estudio mencionados en el apartado de consideraciones previas, creímos conveniente calcular nuevas columnas. Distinguimos principalmente:
    - *Pais_Origen*, *Pais_Destino, Latitud y Longitud*: Estas variables serían útiles para poder quedarnos así solo con los viajes que entren o salgan a España, Portugal y Francia así como posteriores cálculos. Para calcular esto, empleando 2 primeros datasets mencionados en el apartado de consideraciones previas. Buscando primero, los pueblos españoles (pues son los que más hay en el dataset original), buscando en el dataset del apartado a. Si no apareciera en este dataset (por no ser de España, principalmente) buscaríamos en el del apartado b. Si pese a esto diera error, usamos la API de Nominatim.
    - *Distancia*: Para calcularla, empleamos la distancia del semiverseno[1] para calcular la distancia a partir de las columnas de *Latitud* y *Longitud* previamente obtenidas.
    - *Emisiones*: Esta columna nos será útil para poder hacer el estudio de poluciones y los efectos del carpooling en esto. Para obtener estos valores, empleamos la librería de Python *transport_co2*[2] para calcular este valor en función de la distancia del viaje y el número de viajeros.
- **Creación de nuevos datasets**: A partir de otros más pequeños, como es el caso del dataset de movilidad y el de autobuses. De este modo obtendremos un nuevo dataset con dos columnas: el año y el número de viajeros en el transporte público (trenes y autobuses).

Finalmente, tras aplicar las técnicas de transformación previamente mencionadas, consideramos necesario volver a seleccionar variables, pues como algunas serían útiles sólo para cálculos, podrían ser eliminadas. Como conclusión, las variables finalmente elegidas para la ficha de datos serán:

[1] https://es.wikipedia.org/wiki/F%C3%B3rmula_del_semiverseno
[2] https://pypi.org/project/transport-co2/

1. BlaBlaCar
    - *Día*. Esta variable es necesaria para agrupar los datos cronológicamente.
    - *Viajeros*. Esta variable es necesaria para evaluar la cantidad de personas, incluyendo conductores, que hacen uso de los servicios de carpooling de BlaBlaCar.
    - *Emisión de CO_2*. Esta variable es necesaria para evaluar el impacto ambiental teniendo en cuenta la cantidad de kilómetros y pasajeros.
2. Movilidad + Transporte Interurbano Autobús en España
    - *Año*. Es necesaria para obtener un orden cronológico.
    - *Millones_viajeros*. Nos permitirá comparar las tendencias de viajes.
3. Emisiones:
    - *Año*. Necesaria para obtener el orden cronológico.
    - *Valor*. Nos permitirá comparar diferentes emisiones.

FICHA DE DATOS

Una vez realizada la selección de variables, el preprocesado y la transformación de datos, es el momento de realizar la fusión de estos para obtener la ficha de datos definitiva.

Encontramos una diferencia en la granularidad de los datasets, puesto que el nivel de precisión en los registros varía. Debido a que la granularidad de los datos cronológicos del dataset de BlaBlaCar viene dado en días mientras que para los datos cronológicos de los otros datasets (emisiones y transporte público) viene dado en años, tenemos que decidir cómo fusionar estos datos.

Para ello, tenemos dos vías de trabajo:

1. Reducir el dataset principal hasta tener el mismo tamaño que el dataset más pequeño y fusionarlo. Esto es, agrupar las columnas según el año y hacer la suma de las columnas. De esta manera, se obtiene un dataset derivado del original de BlaBlaCar con solo 3 filas, correspondiente a los años. Como consecuencia encontramos el inconveniente de perder información durante la agrupación de datos, y esto puede derivar en la pérdida de precisión en el resultado final.
2. Añadir nuevas columnas en las que los valores se repitan en función del año del registro. De esta manera, se obtienen valores repetidos, pero no encontramos la desventaja de pérdida de información.

PROPUESTAS DE TRABAJO

Para obtener algo más pragmático de este estudio, proponemos las siguientes líneas de trabajo:

1. Modelo de predicción del porcentaje de uso de BlaBlaCar en siguientes años.
2. Modelo de predicción del porcentaje de emisiones que supone el carpooling de BlaBlaCar dentro del transporte terrestre en años posteriores.
3. Sistema de recomendación de fecha de reserva de viaje. Dado un rango de tiempo (fecha de inicio y fin de periodo de interés) devuelve una fecha (o lista de fechas) en las cuales es más recomendable pedir un BlaBlaCar en función del mes del año y el día de la semana, por ejemplo.

ANEXO

Todo el desarrollo completo está disponible en el siguiente enlace al Google Colab:
https://colab.research.google.com/github/mga06/MineriaMSID/blob/master/BlaBlaCar.ipynb