18.- Feature Construction_04_19_06_turismo_gasto_completo_v_01

June 15, 2023

#

CU55_Modelo agregado de estimación del gasto medio por turista

Citizenlab Data Science Methodology > III - Feature Engineering Domain *** > # 18.- Feature Construction

Feature Construction is the process related to create new features from your existing ones to improve model performance.

0.1 Tasks

Feature Construction - Create Interaction Features - Create derived variables - Combine Sparse Classes - Explore Binning for Feature Construction

0.2 Consideraciones casos CitizenLab programados en R

- Algunas de las tareas de este proceso se han realizado en los notebooks del proceso 05 Data Collection porque eran necesarias para las tareas ETL. En esos casos, en este notebook se referencia al notebook del proceso 05 correspondiente
- Otras tareas típicas de este proceso se realizan en los notebooks del dominio IV al ser más eficiente realizarlas en el propio pipeline de modelización.
- Por tanto en los notebooks de este proceso de manera general se incluyen las comprobaciones necesarias, y comentarios si procede
- Las tareas del proceso se van a aplicar solo a los archivos que forman parte del despliegue, ya que hay muchos archivos intermedios que no procede pasar por este proceso
- El nombre de archivo del notebook hace referencia al nombre de archivo del proceso 05 al que se aplica este proceso, por eso pueden no ser correlativa la numeración
- Las comprobaciones se van a realizar teniendo en cuenta que el lenguaje utilizado en el despliegue de este caso es R

0.3 File

- Input File: CU_55_08_03_gasto_municipio.csv
- Sampled Input File: CU_55_07_03_gasto_municipio.csv
- Output File: No aplica

0.3.1 Encoding

Con la siguiente expresión se evitan problemas con el encoding al ejecutar el notebook. Es posible que deba ser eliminada o adaptada a la máquina en la que se ejecute el código.

```
In [1]: Sys.setlocale(category = "LC_ALL", locale = "es_ES.UTF-8")
```

'LC_COLLATE=es_ES.UTF-8;LC_CTYPE=es_ES.UTF-8;LC_MONETARY=es_ES.UTF-8;LC_NUMERIC=C;LC_TIME=es_ES.UTF-8'

0.4 Settings

0.4.1 Libraries to use

```
In [2]: library(readr)
        library(dplyr)
        library(tidyr)
        library(forcats)
        library(lubridate)
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union



Attaching package: 'lubridate'


The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

0.4.2 Paths

```
In [3]: iPath <- "Data/Input/"
        oPath <- "Data/Output/"
```

0.5 Data Load

OPCION A: Seleccionar fichero en ventana para mayor comodidad
Data load using the {tcltk} package. Ucomment the line if using this option

```
In [4]: # file_data <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

```
In [5]: iFile <- "CU_55_08_03_gasto_municipio.csv"
        file_data <- paste0(iPath, iFile)

        if(file.exists(file_data)){
            cat("Se leerán datos del archivo: ", file_data)
        } else{
            warning("Cuidado: el archivo no existe.")
        }
```

```
Se leerán datos del archivo:  Data/Input/CU_55_08_03_gasto_municipio.csv
```

Data file to dataframe Usar la función adecuada según el formato de entrada (xlsx, csv, json, ...)

```
In [6]: data <- read_csv(file_data)
```

```
Rows: 50294 Columns: 10
 Column specification
Delimiter: ","
chr (5): mes, pais_orig_cod, pais_orig, mun_dest, CMUN
dbl (4): mun_dest_cod, turistas, gasto, Target
lgl (1): is_train

 Use `spec()` to retrieve the full column specification for this data.
 Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Estructura de los datos:

```
In [7]: data |> glimpse()
```

```
Rows: 50,294
Columns: 10
$ mes           <chr> "2019-08", "2021-07", "2021-07", "2022-01", "2019-08", "
$ pais_orig_cod <chr> "110", "010", "010", "000", "128", "000", "011", "126",
$ pais_orig     <chr> "Francia", "Total Europa", "Total Europa", "Total", "Rum
$ mun_dest_cod  <dbl> 28161, 28176, 28132, 28141, 28130, 28126, 28075, 28005,
$ mun_dest      <chr> "Valdemoro", "Villanueva de la Cañada", "San Martín de l
$ turistas      <dbl> 466, 1375, 465, 54, 135, 30, 285, 768, 31, 1646, 116, 36
$ CMUN          <chr> "161", "176", "132", "141", "130", "126", "075", "005",
$ gasto         <dbl> 76.360, 99.650, 99.650, 107.820, 109.210, 118.230, 118.2
$ Target        <dbl> 76.360, 99.650, 99.650, 107.820, 109.210, 118.230, 118.2
$ is_train      <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TR
```

Muestra de los primeros datos:

```
In [8]: data |> slice_head(n = 5)
```

	mes <chr>	pais_orig_cod <chr>	pais_orig <chr>	mun_dest_cod <dbl>	mun_dest <chr>
A spec_tbl_df: 5 Œ 10	2019-08	110	Francia	28161	Valdemoro
	2021-07	010	Total Europa	28176	Villanueva de la Cañada
	2021-07	010	Total Europa	28132	San Martín de la Vega
	2022-01	000	Total	28141	Sevilla la Nueva
	2019-08	128	Rumania	28130	San Fernando de Henares

0.6 Creating Interaction Features

Ver notebooks del proceso 05 Data Collectio

0.7 Creating derived variables

Ver notebooks del proceso 05 Data Collectio

0.8 Combining Sparse Classes

Ver notebooks del proceso 05 Data Collectio

0.9 Binning for Feature Construction

Ver notebooks del proceso 05 Data Collectio