14.- Feature Data Transform_25_01_listas_espera_v_01

June 10, 2023

#

CU25_Modelo de gestión de Lista de Espera Quirúrgica

Citizenlab Data Science Methodology > III - Feature Engineering Domain *** > # 14.- Feature Data Transform

Feature Data Transform is the process that allows change (if is required) the type and/or distribution of data features (e.g. scaling, normalizing o standardizing data features).

0.1 Tasks

Perform Basic Data Transforms

Perform Categorical Variable Transformation
- Encode Transformation
- One-hot encoding
- Ordinal encoding
- Dummy encoding
- Evaluate a Logistic Regression model
- Consider Embedding if text mining context

Perform Numeric Variable Transformation
- Scale Transformation
- Normalization
- Standardization
- IQR Robust Scaler Transform
- Evaluate a KNN model
- Distribution Transformation
- Discretization
- Uniform
- Clustered(k-Means)
- Quantile
- Normal Quantile
- Uniform Quantile
- Evaluate a KNN model
- Evaluate a KNN model
- Power transforms (Make Distributions More Gaussian)
- Box-Cox Transform
- Yeo-Johnson Transform
- Evaluate a KNN model

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_25_09.2_01_lista_espera_completo_clean_v_01.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.

[51]:
```
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

[52]:
```
library(readr)
library(dplyr)
library(tidyr)
library(forcats)
library(lubridate)
```

0.4.2 Paths

[53]:
```
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

[54]:
```
# file_data <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[55]:
```
iFile <- "CU_25_09.2_01_lista_espera_completo_clean_v_01.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_25_09.2_01_lista_espera_completo_clean_v_01.csv
```

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

[56]:
```
data <- read.csv(file_data)
```

Estructura de los datos:

[57]:
```
data |> glimpse()
```

```
Rows: 55,216
Columns: 46
$ Hospital          <chr> "HOSPITAL REY JUAN CARLOS",
"HOSPITAL CENTRAL DE LA …
$ Especialidad      <chr> "UROLOGÍA", "ODONTOESTOMATOLOGÍA",
"GINECOLOGÍA", "D…
$ total_pacientes   <int> 344, 0, 52, 37, 0, 4, 0, 718, 0,
271, 108, 0, 34, 86…
$ ano               <int> 2021, 2020, 2021, 2021, 2021, 2020,
2021, 2020, 2021…
$ semana            <int> 30, 36, 49, 23, 3, 5, 50, 7, 35, 1,
42, 10, 21, 33, …
$ CODCNH            <int> 281348, 280724, 281292, 281292,
281236, 280724, 2807…
$ id_area           <int> 8, 7, 11, 11, 11, 7, 3, 6, 1, 2, 2,
8, 11, 11, 1, 3,…
$ nombre_area       <chr> "SUR-OESTE I", "CENTRO-OESTE", "SUR
II", "SUR II", "…
$ cmunicipio        <int> 280920, 280796, 280133, 280133,
281610, 280796, 2800…
$ Municipio         <chr> "MÓSTOLES", "MADRID", "ARANJUEZ",
"ARANJUEZ", "VALDE…
$ CAMAS             <int> 382, 475, 98, 98, 182, 475, 507,
613, 269, 1143, 156…
```

```
$ Clase             <chr> "HOSPITALES GENERALES", "HOSPITALES
GENERALES", "HOS…
$ Dependencia       <chr> "SERVICIOS E INSTITUTOS DE SALUD DE
LAS COMUNIDADES …
$ TAC               <int> 2, 2, 1, 1, 1, 2, 3, 3, 0, 0, 1, 2,
6, 6, 1, 3, 4, 1…
$ RM                <int> 3, 2, 1, 1, 2, 2, 2, 3, 0, 0, 0, 2,
5, 5, 1, 2, 4, 1…
$ GAM               <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
2, 2, 0, 0, 2, 0…
$ HEM               <int> 1, 2, 0, 0, 1, 2, 1, 2, 0, 0, 0, 1,
3, 3, 0, 1, 1, 0…
$ ASD               <int> 2, 1, 1, 1, 1, 1, 1, 3, 0, 0, 0, 1,
2, 2, 0, 1, 2, 1…
$ ALI               <int> 1, 2, 0, 0, 0, 2, 0, 4, 0, 0, 0, 0,
3, 3, 0, 2, 2, 0…
$ SPECT             <int> 1, 1, 0, 0, 0, 1, 0, 4, 0, 0, 0, 0,
3, 3, 0, 0, 0, 0…
$ MAMOS             <int> 2, 1, 1, 1, 1, 1, 2, 2, 0, 0, 1, 2,
3, 3, 1, 1, 3, 1…
$ DO                <int> 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1,
2, 2, 0, 1, 2, 0…
$ DIAL              <int> 20, 24, 13, 13, 17, 24, 28, 31, 0,
0, 0, 28, 43, 43,…
$ X                 <dbl> -3.870412, -3.745529, -3.610795,
-3.610795, -3.69744…
$ Y                 <dbl> 40.33920, 40.38791, 40.05726,
40.05726, 40.19884, 40…
$ t3_1              <dbl> 42.34715, 45.37878, 42.06149,
42.06149, 42.06149, 45…
$ t1_1              <int> 532487, 511605, 899702, 899702,
899702, 511605, 3830…
$ t2_1              <dbl> 0.5122493, 0.5296804, 0.5240445,
0.5240445, 0.524044…
$ t2_2              <dbl> 0.4877507, 0.4703198, 0.4759555,
0.4759555, 0.475955…
$ t4_1              <dbl> 0.1659665, 0.1054260, 0.1540793,
0.1540793, 0.154079…
$ t4_2              <dbl> 0.6371549, 0.6742432, 0.6753787,
0.6753787, 0.675378…
$ t4_3              <dbl> 0.1968769, 0.2203341, 0.1705449,
0.1705449, 0.170544…
$ t5_1              <dbl> 0.1137647, 0.1744493, 0.1747059,
0.1747059, 0.174705…
$ t6_1              <dbl> 0.1604646, 0.2629599, 0.2641879,
0.2641879, 0.264187…
$ t7_1              <dbl> 0.05422176, 0.05481008, 0.04898547,
0.04898547, 0.04…
```

```
$ t8_1              <dbl> 0.04120012, 0.04653221, 0.03679912,
0.03679912, 0.03…
$ t9_1              <dbl> 0.3348780, 0.4914365, 0.3346063,
0.3346063, 0.334606…
$ t10_1             <dbl> 0.13692541, 0.12170996, 0.15173209,
0.15173209, 0.15…
$ t11_1             <dbl> 0.5072726, 0.4915713, 0.5024130,
0.5024130, 0.502413…
$ t12_1             <dbl> 0.5849309, 0.5597213, 0.5900028,
0.5900028, 0.590002…
$ capacidad         <int> 17, 0, 8, 5, 0, 5, 1, 24, 6, 6, 30,
4, 2, 15, 20, 6,…
$ pacientes         <int> 1447, 1211, 1293, 1501, 1240, 1504,
1502, 1533, 1463…
$ consultas         <int> 573, 45, 108, 103, 44, 42, 36,
1119, 34, 466, 220, 6…
$ hospitalizaciones <int> 12, 0, 2, 2, 0, 1, 0, 4, 0, 12, 3,
0, 2, 4, 1, 2, 15…
$ Target            <dbl> 54.45, 0.00, 37.96, 23.14, 0.00,
6.25, 0.00, 78.20, …
$ is_train          <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,
TRUE, TRUE, TRUE…
```

Muestra de los primeros datos:

[58]:
```
data |> slice_head(n = 5)
```

A data.frame: 5 × 46

Hospital <chr>	Especialidad <chr>
HOSPITAL REY JUAN CARLOS	UROLOGÍA
HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	ODONTOESTOMATOLOG
HOSPITAL UNIVERSITARIO DEL TAJO	GINECOLOGÍA
HOSPITAL UNIVERSITARIO DEL TAJO	DERMATOLOGÍA
HOSPITAL UNIVERSITARIO INFANTA ELENA	ODONTOESTOMATOLOG

0.6 Basic Data Transforms

0.6.1 Data Selecting

[59]:
```
data |> select(1)
```

	Hospital <chr>
	HOSPITAL REY JUAN CARLOS
	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO INFANTA ELENA
	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA
	HOSPITAL UNIVERSITARIO PRINCIPE DE ASTURIAS
	HOSPITAL UNIVERSITARIO PUERTA DE HIERRO MAJADAHONDA
	HOSPITAL UNIVERSITARIO INFANTA LEONOR
	HOSPITAL GENERAL UNIVERSITARIO GREGORIO MARAÑON
	HOSPITAL UNIVERSITARIO SANTA CRISTINA
	HOSPITAL UNIVERSITARIO FUNDACION ALCORCON
	HOSPITAL UNIVERSITARIO 12 DE OCTUBRE
	HOSPITAL UNIVERSITARIO 12 DE OCTUBRE
	HOSPITAL UNIVERSITARIO DEL SURESTE
	HOSPITAL UNIVERSITARIO DE TORREJON
	HOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ
	HOSPITAL UNIVERSITARIO 12 DE OCTUBRE
	HOSPITAL RAMON Y CAJAL
	HOSPITAL UNIVERSITARIO DE TORREJON
	HOSPITAL EL ESCORIAL
	HOSPITAL UNIVERSITARIO DE GETAFE
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO SANTA CRISTINA
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO INFANTA SOFIA
A data.frame: 55216×1	HOSPITAL CENTRAL DE LA CRUZ ROJA SAN JOSE Y SANTA ADELA
	HOSPITAL UNIVERSITARIO SEVERO OCHO

0.6.2 Data Filtering

[60]:
```
data |> filter(ano == 2021) |> select(1)
```

	Hospital <chr>
	HOSPITAL REY JUAN CARLOS
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO INFANTA ELENA
	HOSPITAL UNIVERSITARIO PRINCIPE DE ASTURIAS
	HOSPITAL UNIVERSITARIO INFANTA LEONOR
	HOSPITAL UNIVERSITARIO SANTA CRISTINA
	HOSPITAL UNIVERSITARIO 12 DE OCTUBRE
	HOSPITAL UNIVERSITARIO DEL SURESTE
	HOSPITAL UNIVERSITARIO DE TORREJON
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO DEL TAJO
	HOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ
	HOSPITAL RAMON Y CAJAL
	HOSPITAL UNIVERSITARIO DE TORREJON
	HOSPITAL UNIVERSITARIO SANTA CRISTINA
	HOSPITAL UNIVERSITARIO INFANTA SOFIA
	HOSPITAL UNIVERSITARIO INFANTA LEONOR
	HOSPITAL EL ESCORIAL
	HOSPITAL REY JUAN CARLOS
	HOSPITAL UNIVERSITARIO DEL SURESTE
	HOSPITAL GENERAL UNIVERSITARIO GREGORIO MARAÑON
	HOSPITAL UNIVERSITARIO LA PAZ
	HOSPITAL UNIVERSITARIO LA PAZ
	HOSPITAL INFANTIL UNIVERSITARIO NIÑO JESUS
	HOSPITAL GENERAL DE VILLALBA
	HOSPITAL UNIVERSITARIO INFANTA LEONOR
	HOSPITAL UNIVERSITARIO DE LA PRINCESA
	HOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ
A data.frame: 22736 × 1	HOSPITAL UNIVERSITARIO DE LA PRINCESA
	HOSPITAL UNIVERSITARIO SEVERO OCHOA

0.6.3 Insert New Column

[61]:
```
data |>
    mutate(x = TRUE) |> head()
```

A data.frame: 6 × 47

	Hospital <chr>	Especialidad <chr>
1	HOSPITAL REY JUAN CARLOS	UROLOGÍA
2	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	ODONTOESTOMATOI
3	HOSPITAL UNIVERSITARIO DEL TAJO	GINECOLOGÍA
4	HOSPITAL UNIVERSITARIO DEL TAJO	DERMATOLOGÍA
5	HOSPITAL UNIVERSITARIO INFANTA ELENA	ODONTOESTOMATOI
6	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	CIRUGÍA TORÁCICA

0.6.4 Delete Column

[62]:
```
data |> select(-t4_2) |> head()
```

A data.frame: 6 × 45

	Hospital <chr>	Especialidad <chr>
1	HOSPITAL REY JUAN CARLOS	UROLOGÍA
2	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	ODONTOESTOMATOI
3	HOSPITAL UNIVERSITARIO DEL TAJO	GINECOLOGÍA
4	HOSPITAL UNIVERSITARIO DEL TAJO	DERMATOLOGÍA
5	HOSPITAL UNIVERSITARIO INFANTA ELENA	ODONTOESTOMATOI
6	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	CIRUGÍA TORÁCICA

0.6.5 Rank Data

Operation

[63]:
```
data |> mutate(rank = order(capacidad)) |> head()
```

A data.frame: 6 × 47

	Hospital <chr>	Especialidad <chr>
1	HOSPITAL REY JUAN CARLOS	UROLOGÍA
2	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	ODONTOESTOMATOI
3	HOSPITAL UNIVERSITARIO DEL TAJO	GINECOLOGÍA
4	HOSPITAL UNIVERSITARIO DEL TAJO	DERMATOLOGÍA
5	HOSPITAL UNIVERSITARIO INFANTA ELENA	ODONTOESTOMATOI
6	HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	CIRUGÍA TORÁCICA

0.7 Categorical Variable Transformation

0.7.1 Encode Transformation

No aplica

Ordinal Encoding Transform No aplica

One Hot Encoding Transform

Dummy Variable Encoding Transform

0.7.2 Embedding Transformation

Specific encode for text mining context. No code here.

0.8 Numeric Variable Transformation: Scale

0.8.1 Data to Transform

Evaluating Normalization Tranform

Evaluating Standarization Transform

0.8.2 Normalization Transform

No aplica

0.8.3 Standarization Transform

No aplica

0.9 Numeric Variable Transformation: Distribution

No aplica

0.9.1 Discretization Transform

Evaluating Discretization Transformations

Uniform Discretization Transform No aplica

0.9.2 Power Transform

No aplica

Data to Transform No aplica

Evaluating Box-Cox tranform

Evaluating Yeo-Johnson tranform

Box-Cox Transform No aplica

Yeo-Johnson Transform No aplica

0.10 Data Save

- No aplica

Identificamos los datos a guardar

[64]:
```
data_to_save <- data
```

Estructura de nombre de archivos:

- Código del caso de uso, por ejemplo "CU_04"
- Número del proceso que lo genera, por ejemplo "_06".
- Resto del nombre del archivo de entrada
- Extensión del archivo

Ejemplo: "CU_04_06_01_01_zonasgeo.json, primer fichero que se genera en la tarea 01 del proceso 05 (Data Collection) para el caso de uso 04 (vacunas) y que se ha transformado en el proceso 06

Importante mantener los guiones bajos antes de proceso, tarea, archivo y nombre

0.10.1 Proceso 14

[65]:
```
# caso <- "CU_XX"
# proceso <- '_09.2'
# tarea <- "_XX"
# archivo <- ""
# proper <- "_xxxxx"
# extension <- ".csv"
```

OPCION A: Uso del paquete "tcltk" para mayor comodidad

- Buscar carpeta, escribir nombre de archivo SIN extensión (se especifica en el código)
- Especificar sufijo2 si es necesario
- Cambiar datos por datos_xx si es necesario

[66]:
```
# file_save <- paste0(caso, proceso, tarea, tcltk::tkgetSaveFile(), proper,␣
 ↪extension)
# path_out <- paste0(oPath, file_save)
# write_csv(data_to_save_xxxx, path_out)

# cat('File saved as: ')
# path_out
```

OPCION B: Especificar el nombre de archivo

- Los ficheros de salida del proceso van siempre a Data/Output/.

[ ]:
```
# file_save <- paste0(caso, proceso, tarea, archivo, proper, extension)
# path_out <- paste0(oPath, file_save)
# write_csv(data_to_save_xxxx, path_out)
```

```
# cat('File saved as: ')
# path_out
```

Copia del fichero a Input Si el archivo se va a usar en otros notebooks, copiar a la carpeta Input

[ ]:
```
# path_in <- paste0(iPath, file_save)
# file.copy(path_out, path_in, overwrite = TRUE)
```

0.11 REPORT

A continuación se realizará un informe de las acciones realizadas

0.12 Main Actions Carried Out

- Si eran necesarias se han realizado en el proceso 05 por cuestiones de eficiencia
- O bien se hacen en el dominio IV o V para integrar en el pipeline de modelización

0.13 Main Conclusions

- Los datos están listos para la modelización y despliegue

0.14 CODE TO DEPLOY (PILOT)

A continuación se incluirá el código que deba ser llevado a despliegue para producción, dado que se entiende efectúa operaciones necesarias sobre los datos en la ejecución del prototipo

Description

- No hay nada que desplegar en el piloto, ya que estos datos son estáticos o en todo caso cambian con muy poca frecuencia, altamente improbable durante el proyecto.

CODE

[ ]:
```
```