15.- Feature Importance_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 *** > # 15.- Feature Importance

Feature Importance is the process that assigns scores to the input characteristics to a model, which indicate the relative importance of each characteristic, in order, for example, to be able to select the most important ones.

0.1 Tasks

Perform Feature importance from model coefficients
- Linear Regression Feature importance
- Logistic Regression Feature importance

Perform Feature importance from Decision Tree
- CART Feature Importance
- Random Forest Regression Feature Importance

Perform Feature importance from Permutation testing

Evaluate a Logistic Regression model with feature selection

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.

[18]:
```
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

[19]:
```
library(readr)
library(dplyr)
library(tidyr)
library(forcats)
library(lubridate)
library(rpart)
```

0.4.2 Paths

[20]:
```
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

[21]:
```
# file_data <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[22]:
```
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, ...)

[23]:
```
data <- read.csv(file_data)
```

Estructura de los datos:

[24]:
```
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:

[25]:
```
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 Feature importance from model coefficients

0.6.1 Linear Regression Feature importance

[26]:
```
# Fit a linear regression model
model <- lm(Target ~ ., data = data)

# Extract the coefficients and their names
coefficients <- coef(model)
names <- names(coefficients)

# Calculate the absolute values of the coefficients
abs_coefficients <- abs(coefficients)

# Sort the coefficients in descending order
sorted_coefficients <- sort(abs_coefficients, decreasing = TRUE)

# Print the feature importance
feature_importance <- data.frame(Feature = names(sorted_coefficients),␣
 ↪Importance = sorted_coefficients)
print(feature_importance)
```

```
                                   Feature
(Intercept)
(Intercept)
HospitalHOSPITAL GENERAL DE VILLALBA
HospitalHOSPITAL GENERAL DE VILLALBA
HospitalHOSPITAL REY JUAN CARLOS
HospitalHOSPITAL REY JUAN CARLOS
HospitalHOSPITAL UNIVERSITARIO DE TORREJON
```

HospitalHOSPITAL UNIVERSITARIO DE TORREJON
HospitalHOSPITAL UNIVERSITARIO PUERTA DE HIERRO MAJADAHONDA HospitalHOSPITAL
UNIVERSITARIO PUERTA DE HIERRO MAJADAHONDA
HospitalHOSPITAL UNIVERSITARIO INFANTA CRISTINA
HospitalHOSPITAL UNIVERSITARIO INFANTA CRISTINA
HospitalHOSPITAL UNIVERSITARIO DEL TAJO
HospitalHOSPITAL UNIVERSITARIO DEL TAJO
HospitalHOSPITAL UNIVERSITARIO DEL SURESTE
HospitalHOSPITAL UNIVERSITARIO DEL SURESTE
HospitalHOSPITAL UNIVERSITARIO INFANTA LEONOR
HospitalHOSPITAL UNIVERSITARIO INFANTA LEONOR
HospitalHOSPITAL UNIVERSITARIO DEL HENARES
HospitalHOSPITAL UNIVERSITARIO DEL HENARES
HospitalHOSPITAL UNIVERSITARIO INFANTA SOFIA
HospitalHOSPITAL UNIVERSITARIO INFANTA SOFIA
HospitalHOSPITAL UNIVERSITARIO INFANTA ELENA
HospitalHOSPITAL UNIVERSITARIO INFANTA ELENA
HospitalHOSPITAL UNIVERSITARIO DE FUENLABRADA
HospitalHOSPITAL UNIVERSITARIO DE FUENLABRADA
HospitalHOSPITAL UNIVERSITARIO FUNDACION ALCORCON
HospitalHOSPITAL UNIVERSITARIO FUNDACION ALCORCON
HospitalHOSPITAL UNIVERSITARIO DE GETAFE
HospitalHOSPITAL UNIVERSITARIO DE GETAFE
HospitalHOSPITAL EL ESCORIAL
HospitalHOSPITAL EL ESCORIAL
HospitalHOSPITAL UNIVERSITARIO DE MOSTOLES
HospitalHOSPITAL UNIVERSITARIO DE MOSTOLES
HospitalHOSPITAL UNIVERSITARIO SEVERO OCHOA
HospitalHOSPITAL UNIVERSITARIO SEVERO OCHOA
HospitalHOSPITAL UNIVERSITARIO PRINCIPE DE ASTURIAS
HospitalHOSPITAL UNIVERSITARIO PRINCIPE DE ASTURIAS
HospitalHOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA
HospitalHOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA
HospitalHOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ
HospitalHOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ
HospitalHOSPITAL UNIVERSITARIO LA PAZ
HospitalHOSPITAL UNIVERSITARIO LA PAZ
HospitalHOSPITAL RAMON Y CAJAL
HospitalHOSPITAL RAMON Y CAJAL
HospitalHOSPITAL UNIVERSITARIO 12 DE OCTUBRE
HospitalHOSPITAL UNIVERSITARIO 12 DE OCTUBRE
HospitalHOSPITAL GENERAL UNIVERSITARIO GREGORIO MARAÑON HospitalHOSPITAL
GENERAL UNIVERSITARIO GREGORIO MARAÑON
HospitalHOSPITAL CLINICO SAN CARLOS
HospitalHOSPITAL CLINICO SAN CARLOS
HospitalHOSPITAL UNIVERSITARIO SANTA CRISTINA
HospitalHOSPITAL UNIVERSITARIO SANTA CRISTINA
HospitalHOSPITAL UNIVERSITARIO DE LA PRINCESA

HospitalHOSPITAL UNIVERSITARIO DE LA PRINCESA
HospitalHOSPITAL INFANTIL UNIVERSITARIO NIÑO JESUS
HospitalHOSPITAL INFANTIL UNIVERSITARIO NIÑO JESUS
CODCNH
CODCNH
EspecialidadODONTOESTOMATOLOGÍA
EspecialidadODONTOESTOMATOLOGÍA
EspecialidadCIRUGÍA TORÁCICA
EspecialidadCIRUGÍA TORÁCICA
EspecialidadTRAUMATOLOGÍA
EspecialidadTRAUMATOLOGÍA
EspecialidadCIRUGÍA PEDIÁTRICA GENERAL
EspecialidadCIRUGÍA PEDIÁTRICA GENERAL
EspecialidadCIRUGÍA CARDIACA
EspecialidadCIRUGÍA CARDIACA
EspecialidadUROLOGÍA
EspecialidadUROLOGÍA
EspecialidadNEUROCIRUGÍA
EspecialidadNEUROCIRUGÍA
EspecialidadCIRUGÍA GENERAL Y DEL APARATO DIGESTIVO
EspecialidadCIRUGÍA GENERAL Y DEL APARATO DIGESTIVO
EspecialidadTOTAL
EspecialidadTOTAL
EspecialidadCIRUGÍA ORAL Y MAXILOFACIAL
EspecialidadCIRUGÍA ORAL Y MAXILOFACIAL
EspecialidadDERMATOLOGÍA
EspecialidadDERMATOLOGÍA
EspecialidadOTORRINOLARINGOLOGÍA
EspecialidadOTORRINOLARINGOLOGÍA
EspecialidadOFTALMOLOGÍA
EspecialidadOFTALMOLOGÍA
EspecialidadCIRUGÍA PLÁSTICA Y REPARADORA
EspecialidadCIRUGÍA PLÁSTICA Y REPARADORA
ano
ano
EspecialidadGINECOLOGÍA
EspecialidadGINECOLOGÍA
is_trainTRUE
is_trainTRUE
capacidad
capacidad
semana
semana
total_pacientes
total_pacientes
hospitalizaciones
hospitalizaciones
consultas

consultas
pacientes
pacientes

	Importance
(Intercept)	2.957680e+11
HospitalHOSPITAL GENERAL DE VILLALBA	1.278521e+09
HospitalHOSPITAL REY JUAN CARLOS	1.266908e+09
HospitalHOSPITAL UNIVERSITARIO DE TORREJON	1.255294e+09
HospitalHOSPITAL UNIVERSITARIO PUERTA DE HIERRO MAJADAHONDA	1.232068e+09
HospitalHOSPITAL UNIVERSITARIO INFANTA CRISTINA	1.220454e+09
HospitalHOSPITAL UNIVERSITARIO DEL TAJO	1.207785e+09
HospitalHOSPITAL UNIVERSITARIO DEL SURESTE	1.196172e+09
HospitalHOSPITAL UNIVERSITARIO INFANTA LEONOR	1.184559e+09
HospitalHOSPITAL UNIVERSITARIO DEL HENARES	1.183503e+09
HospitalHOSPITAL UNIVERSITARIO INFANTA SOFIA	1.171890e+09
HospitalHOSPITAL UNIVERSITARIO INFANTA ELENA	1.148663e+09
HospitalHOSPITAL UNIVERSITARIO DE FUENLABRADA	1.053645e+09
HospitalHOSPITAL UNIVERSITARIO FUNDACION ALCORCON	9.744631e+08
HospitalHOSPITAL UNIVERSITARIO DE GETAFE	8.878911e+08
HospitalHOSPITAL EL ESCORIAL	8.150439e+08
HospitalHOSPITAL UNIVERSITARIO DE MOSTOLES	7.875942e+08
HospitalHOSPITAL UNIVERSITARIO SEVERO OCHOA	7.284719e+08
HospitalHOSPITAL UNIVERSITARIO PRINCIPE DE ASTURIAS	6.302865e+08
HospitalHOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	6.081157e+08
HospitalHOSPITAL UNIVERSITARIO FUNDACION JIMENEZ DIAZ	2.882215e+08
HospitalHOSPITAL UNIVERSITARIO LA PAZ	1.414713e+08
HospitalHOSPITAL RAMON Y CAJAL	1.256350e+08
HospitalHOSPITAL UNIVERSITARIO 12 DE OCTUBRE	1.193004e+08
HospitalHOSPITAL GENERAL UNIVERSITARIO GREGORIO MARAÑON	1.034641e+08
HospitalHOSPITAL CLINICO SAN CARLOS	8.023744e+07
HospitalHOSPITAL UNIVERSITARIO SANTA CRISTINA	3.800724e+07
HospitalHOSPITAL UNIVERSITARIO DE LA PRINCESA	2.217086e+07
HospitalHOSPITAL INFANTIL UNIVERSITARIO NIÑO JESUS	1.583634e+07
CODCNH	1.055756e+06
EspecialidadODONTOESTOMATOLOGÍA	4.312146e+01
EspecialidadCIRUGÍA TORÁCICA	3.441241e+01
EspecialidadTRAUMATOLOGÍA	2.932081e+01
EspecialidadCIRUGÍA PEDIÁTRICA GENERAL	2.789507e+01
EspecialidadCIRUGÍA CARDIACA	2.699811e+01
EspecialidadUROLOGÍA	1.617903e+01
EspecialidadNEUROCIRUGÍA	1.594099e+01
EspecialidadCIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	1.318964e+01
EspecialidadTOTAL	1.238248e+01
EspecialidadCIRUGÍA ORAL Y MAXILOFACIAL	1.087416e+01
EspecialidadDERMATOLOGÍA	1.065110e+01
EspecialidadOTORRINOLARINGOLOGÍA	1.027569e+01
EspecialidadOFTALMOLOGÍA	6.426765e+00
EspecialidadCIRUGÍA PLÁSTICA Y REPARADORA	3.770782e+00

ano	2.670217e+00
EspecialidadGINECOLOGÍA	1.918536e+00
is_trainTRUE	1.665356e-01
capacidad	1.227245e-01
semana	1.053269e-01
total_pacientes	2.182061e-02
hospitalizaciones	1.834939e-02
consultas	9.237181e-03
pacientes	1.540184e-03

0.6.2 Logistic Regression Feature importance

No aplica

0.7 Decision Tree Feature Importance

0.7.1 CART Regression Feature Importance

[27]:
```
# Fit a CART regression model
model <- rpart(Target ~ ., data = data, method = "anova")

# Calculate feature importance
importance <- round(model$variable.importance, 2)

# Sort the feature importance in descending order
sorted_importance <- sort(importance, decreasing = TRUE)

# Print the feature importance
feature_importance <- data.frame(Feature = names(sorted_importance), Importance
  ↪= sorted_importance)
print(feature_importance)
```

	Feature	Importance
total_pacientes	total_pacientes	46585351.57
consultas	consultas	42984648.71
hospitalizaciones	hospitalizaciones	36939615.15
Hospital	Hospital	24889013.04
Especialidad	Especialidad	24414140.39
capacidad	capacidad	23312800.43
DIAL	DIAL	11881062.31
Municipio	Municipio	10285219.66
TAC	TAC	9819026.12
CAMAS	CAMAS	8508724.45
ASD	ASD	6396715.74
HEM	HEM	2251712.22
DO	DO	1863984.85
RM	RM	1700792.14
nombre_area	nombre_area	1546078.05
Y	Y	1244093.28

```
id_area                    id_area  1122111.35
CODCNH                     CODCNH   804724.89
t12_1                      t12_1    598653.71
X                              X    498569.19
Dependencia            Dependencia  302727.42
pacientes                pacientes    1248.58
```

0.7.2 CART Classification Feature Importance

[28]:
```
# # Fit a CART classification model
# library(rpart)
# model <- rpart(Target ~ ., data = data, method = "class")

# # Calculate feature importance
# importance <- round(model$variable.importance, 2)

# # Sort the feature importance in descending order
# sorted_importance <- sort(importance, decreasing = TRUE)

# # Print the feature importance
# feature_importance <- data.frame(Feature = names(sorted_importance),␣
 ↪Importance = sorted_importance)
# print(feature_importance)
```

0.7.3 Random Forest Regression Feature Importance

[29]:
```
# # Fit a Random Forest regression model
# library(randomForest)
# model <- randomForest(Target ~ ., data = data)

# # Get feature importance
# importance <- importance(model)

# # Print the feature importance
# print(importance)
```

0.7.4 Random Forest Classification Feature Importance

No aplica

0.8 Permutation Feature Importance

0.8.1 Permutation Feature Importance for Regression

[30]:
```
# # Install and load the 'iml' package

# # Fit a regression model
# model <- lm(Target ~ ., data = data)
```

```r
# # Create an instance of the 'FeatureImp' class
# feature_imp <- FeatureImp$new(model, data = data, y = data$Target)

# # Calculate permutation feature importance
# perm_importance <- feature_imp$permutation()

# # Print the permutation feature importance
# print(perm_importance)
```

0.8.2 Permutation Feature Importance for Classification

No aplica

0.9 Evaluating a Regression model with feature selection.

0.9.1 Evaluating with all selected features.

[56]:
```r
library(rpart)
library(dplyr)
library(caTools)

# Split the data into training and testing sets based on the "is_train" column
train_data <- filter(data, is_train == 1)
test_data <- filter(data, is_train == 0)


# Select the top 5 most important features
selected_features <- colnames(data)


train_data_selected <- select(train_data, all_of(selected_features), Target)

# Train a Linear Regression model with the selected features
model_lm <- lm(Target ~ ., data = train_data_selected)

# Select the same features from the test data
test_data_selected <- select(test_data, all_of(selected_features), Target)
# Make predictions on the test data
predictions <- predict(model_lm, newdata = test_data_selected)

# Calculate Mean Squared Error (MSE)
mse <- mean((predictions - test_data_selected$Target)^2)
print(paste("Mean Squared Error:", round(mse, 4)))
```

```
[1] "Mean Squared Error: 917.7152"
```

0.9.2 Evaluating with feature selection performed using feature importance.

Select type of Feature Importance to use

Las 5 variables con mayor importancia

Operation

[62]:
```
colnames(data)
```

1. 'Hospital' 2. 'Especialidad' 3. 'total_pacientes' 4. 'ano' 5. 'semana' 6. 'CODCNH' 7. 'id_area' 8. 'nombre_area' 9. 'cmunicipio' 10. 'Municipio' 11. 'CAMAS' 12. 'Clase' 13. 'Dependencia' 14. 'TAC' 15. 'RM' 16. 'GAM' 17. 'HEM' 18. 'ASD' 19. 'ALI' 20. 'SPECT' 21. 'MAMOS' 22. 'DO' 23. 'DIAL' 24. 'X' 25. 'Y' 26. 't3_1' 27. 't1_1' 28. 't2_1' 29. 't2_2' 30. 't4_1' 31. 't4_2' 32. 't4_3' 33. 't5_1' 34. 't6_1' 35. 't7_1' 36. 't8_1' 37. 't9_1' 38. 't10_1' 39. 't11_1' 40. 't12_1' 41. 'capacidad' 42. 'pacientes' 43. 'consultas' 44. 'hospitalizaciones' 45. 'Target' 46. 'is_train'

[66]:
```
library(rpart)
library(dplyr)
library(caTools)

# Split the data into training and testing sets based on the "is_train" column
train_data <- filter(data, is_train == 1)
test_data <- filter(data, is_train == 0)

# Train a CART regression model to calculate feature importance
model_cart <- rpart(Target ~ ., data = train_data, method = "anova")

# Get the feature importance
importance <- round(model_cart$variable.importance, 2)

# Select the top 5 most important features
selected_features <- names(importance)[order(importance, decreasing = TRUE)][1:
 ↪18]
selected_features

train_data_selected <- select(train_data, all_of(selected_features), Target)

# Train a Linear Regression model with the selected features
model_lm <- lm(Target ~ ., data = train_data_selected)

# Select the same features from the test data
test_data_selected <- select(test_data, all_of(selected_features), Target)
# Make predictions on the test data
predictions <- predict(model_lm, newdata = test_data_selected)

# Calculate Mean Squared Error (MSE)
mse <- mean((predictions - test_data_selected$Target)^2)
print(paste("Mean Squared Error:", round(mse, 4)))
```

1. 'total_pacientes' 2. 'consultas' 3. 'hospitalizaciones' 4. 'Hospital' 5. 'Especialidad' 6. 'capacidad' 7. 'DIAL' 8. 'Municipio' 9. 'TAC' 10. 'CAMAS' 11. 'ASD' 12. 'HEM' 13. 'DO' 14. 'RM' 15. 'nombre_area' 16. 'Y' 17. 'id_area' 18. 'CODCNH'

```
[1] "Mean Squared Error: 923.2858"
```

0.10 Data Save

- No aplica

Identificamos los datos a guardar

[ ]:
```
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 15

[ ]:
```
# 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

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

[ ]:
```

```