Regresión Logística Múltiple

Tomás Sánchez Grigioni

6/9/2020

Para acceder al data set usar este [link](https://www.kaggle.com/c/titanic/data?select=train.csv)

# Exploración de Datos

Primero nombramos las variables a usar y sus respectivos significados

|  |  |  |  |
| --- | --- | --- | --- |
| Nombres Variables | Definición | Valores | Tipo |
| sobrevivio | Sobrevivio al accidente | 0 = No, 1 = Si | Categórica |
| clase | Clase del ticket | 1 = 1st, 2 = 2nd, 3 = 3rd | Categórica |
| sexo | Sexo | - | Categórica |
| edad | Edad en años | - | Continua |
| her\_esp | Cantidad de hermanos/conyuges en el Titanic | - | Discreta |
| padre\_hijo | Cantidad de padres/hijos en el Titanic | - | Discreta |
| ticket | Numero de ticket | - | Categórica |
| precio\_ticket | Precio del ticket | - | Continua |
| nro\_cabina | Número de cabina | - | Categórica |
| puerto\_embarcacion | Puerto de embarcación | C = Cherbourg, Q = Queenstown, S = Southampton | Categórica |

Analizamos como identifica R los datos para ver si es necesario realizar alguna modificación

## 'data.frame': 891 obs. of 10 variables:  
## $ sobrevivio : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ clase : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ sexo : chr "male" "female" "female" "female" ...  
## $ edad : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ her\_esp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ padre\_hijo : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ precio\_ticket : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ nro\_cabina : chr "" "C85" "" "C123" ...  
## $ puerto\_embarcacion: chr "S" "C" "S" "S" ...

Como se observa a las variables **sobrevivio**, **clase**, **sexo**, **ticket**, **nro\_cabina** y **puerto\_embarcación** no las esta tratando como un factor, entonces necesitamos modificarlas

## 'data.frame': 891 obs. of 10 variables:  
## $ sobrevivio : Factor w/ 2 levels "Sí","No": 2 1 1 1 2 2 2 2 1 1 ...  
## $ clase : Factor w/ 3 levels "1ra","2da","3ra": 3 1 3 1 3 3 1 3 3 2 ...  
## $ sexo : Factor w/ 2 levels "Hombre","Mujer": 1 2 2 2 1 1 1 1 2 2 ...  
## $ edad : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ her\_esp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ padre\_hijo : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...  
## $ precio\_ticket : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ nro\_cabina : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...  
## $ puerto\_embarcacion: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

Ahora las variables las trabaja de forma correcta. El siguiente paso es analizar las medidas resumen, empezamos con **las variables cuantitativas**

Medidas Resumen de Variables Cuantitativas

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Media | Mediana | Primer\_Cuartil | Tercer\_Cuartil | Desvio | NAs |
| edad | 29.6991176 | 28.0000 | 20.1250 | 38 | 14.5264973 | 177 |
| her\_esp | 0.5230079 | 0.0000 | 0.0000 | 1 | 1.1027434 | 0 |
| padre\_hijo | 0.3815937 | 0.0000 | 0.0000 | 0 | 0.8060572 | 0 |
| precio\_ticket | 32.2042080 | 14.4542 | 7.9104 | 31 | 49.6934286 | 0 |

Como se observa, estamos en presencia de 177 NAs en la variable edad. Como en el data set se presentan 891 observaciones, podemos eleminar todas estas observaciones y quedarnos con el resto.