Regresión Logística Múltiple

Link a la info de este dataset: https://www.kaggle.com/c/titanic/data?select=train.csv

Las variables a usar son:

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

# Exploración de datos

Empezamos viendo las medidas representativas de cada variable

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sobrevivio | clase | sexo | edad | her\_esp | padre\_hijo | ticket |
| Min. :0.0000 | Min. :1.000 | female:314 | Min. : 0.42 | Min. :0.000 | Min. :0.0000 | 1601 : 7 |
| 1st Qu.:0.0000 | 1st Qu.:2.000 | male :577 | 1st Qu.:20.12 | 1st Qu.:0.000 | 1st Qu.:0.0000 | 347082 : 7 |
| Median :0.0000 | Median :3.000 |  | Median :28.00 | Median :0.000 | Median :0.0000 | CA. 2343: 7 |
| Mean :0.3838 | Mean :2.309 |  | Mean :29.70 | Mean :0.523 | Mean :0.3816 | 3101295 : 6 |
| 3rd Qu.:1.0000 | 3rd Qu.:3.000 |  | 3rd Qu.:38.00 | 3rd Qu.:1.000 | 3rd Qu.:0.0000 | 347088 : 6 |
| Max. :1.0000 | Max. :3.000 |  | Max. :80.00 | Max. :8.000 | Max. :6.0000 | CA 2144 : 6 |
|  |  |  | NA's :177 |  |  | (Other) :852 |

|  |  |  |
| --- | --- | --- |
| precio\_ticket | nro\_cabina | puerto\_embarcacion |
| Min. : 0.00 | :687 | : 2 |
| 1st Qu.: 7.91 | B96 B98 : 4 | C:168 |
| Median : 14.45 | C23 C25 C27: 4 | Q: 77 |
| Mean : 32.20 | G6 : 4 | S:644 |
| 3rd Qu.: 31.00 | C22 C26 : 3 |  |
| Max. :512.33 | D : 3 |  |
|  | (Other) :186 |  |

Lo que se observa es:

* sobrevivio no lo trata como un factor
* clase no lo trata como un factor
* En edad hay 117 NAs
* En puerto\_embarcacion hay 2 valores sin nivel

Para las variables que no se tratan como factores las convertimos a factores. Luego, para los valores de edad con NA lo que hacemos es aproximarlo con la media de los valores restantes. No usamos la mediana porque no se presentan muchos outliers entonces no es necesario usar la mediana. Por último, para los valores sin nivel de puerto\_embarcacion decidimos no modificarlos ya que no parece haber inconsistencia en los datos.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sobrevivio | clase | sexo | edad | her\_esp | padre\_hijo | ticket |
| 0:549 | 1:216 | female:314 | Min. : 0.42 | Min. :0.000 | Min. :0.0000 | 1601 : 7 |
| 1:342 | 2:184 | male :577 | 1st Qu.:22.00 | 1st Qu.:0.000 | 1st Qu.:0.0000 | 347082 : 7 |
|  | 3:491 |  | Median :29.70 | Median :0.000 | Median :0.0000 | CA. 2343: 7 |
|  |  |  | Mean :29.70 | Mean :0.523 | Mean :0.3816 | 3101295 : 6 |
|  |  |  | 3rd Qu.:35.00 | 3rd Qu.:1.000 | 3rd Qu.:0.0000 | 347088 : 6 |
|  |  |  | Max. :80.00 | Max. :8.000 | Max. :6.0000 | CA 2144 : 6 |
|  |  |  |  |  |  | (Other) :852 |

|  |  |  |
| --- | --- | --- |
| precio\_ticket | nro\_cabina | puerto\_embarcacion |
| Min. : 0.00 | :687 | : 2 |
| 1st Qu.: 7.91 | B96 B98 : 4 | C:168 |
| Median : 14.45 | C23 C25 C27: 4 | Q: 77 |
| Mean : 32.20 | G6 : 4 | S:644 |
| 3rd Qu.: 31.00 | C22 C26 : 3 |  |
| Max. :512.33 | D : 3 |  |
|  | (Other) :186 |  |