Regresión Logística Múltiple

Link al dataset: https://www.kaggle.com/c/titanic/data?select=train.csv

Las variables para usar son:

|  |  |  |
| --- | --- | --- |
| **Nombres variables** | **Definición** | **Valores** |
| 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 | edad | her\_esp | padre\_hijo | precio\_ticket |
| Min | 0 | 1 | 0.42 | 0 | 0 | 0 |
| 1st Qu. | 0 | 2 | 20.12 | 0 | 0 | 7.91 |
| Median | 0 | 3 | 28.00 | 0 | 0 | 14.45 |
| Mean | 0.3838 | 2.309 | 29.70 | 0.523 | 0.3816 | 32.20 |
| 3rd Qu. | 1 | 3 | 38.00 | 1 | 0.0000 | 31.00 |
| Max. | 1 | 3 | 80.00 | 8.000 | 6.0000 | 512.33 |
|  |  |  | NA's :177 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| sexo | ticket | nro\_cabina | puerto\_embarcacion |
| Length:891 | Length:891 | Length:891 | Length:891 |
| Class :character | Class :character | Class :character | Class :character |
| Mode :character | Mode :character | Mode :character | Mode :character |

Lo que se observa es:

* sobrevivio no lo trata como un factor.
* clase no lo trata como un factor.
* sexo no lo trata como un factor.
* nro\_cabina no lo trata como un factor.
* Ticket no lo trata como un factor.
* Puerto\_embarcacion no lo trata como un factor.
* En edad hay 117 NAs

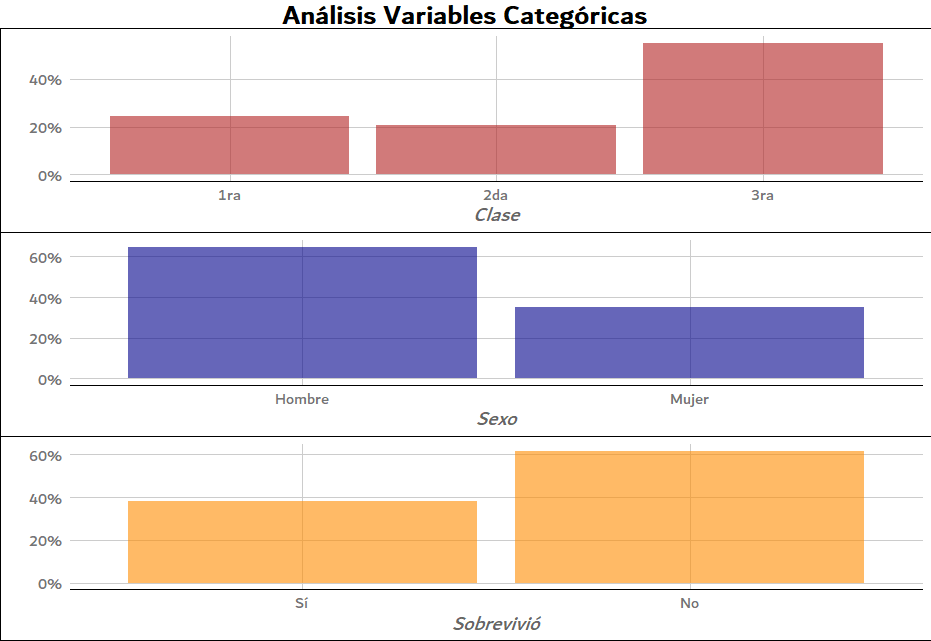
Para las variables que no se tratan como factores las convertimos a factores. Luego, para los valores de edad con NA lo que hacemos es aproximarlo con la media de los valores restantes. No usamos la mediana porque no se presentan muchos outliers entonces no es necesario usar la mediana. Por último, para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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | edad | her\_esp | padre\_hijo | precio\_ticket |
| Min | 0.42 | 0 | 0 | 0 |
| 1st Qu. | 22 | 0 | 0 | 7.91 |
| Median | 29.7 | 0 | 0 | 14.45 |
| Mean | 29.7 | 0.523 | 0.3816 | 32.2 |
| 3rd Qu. | 35 | 1 | 0 | 31 |
| Max. | 80 | 8 | 6 | 512.33 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sobrevivio | clase | sexo | puerto\_embarcacion | nro\_cabina | ticket |
| 0: 549 | 1: 216 | female: 314 | : 2 | : 687 | 1601: 7 |
| 1: 342 | 2: 184 | male : 577 | C: 168 | B96 B98: 4 | 347082: 7 |
|  | 3: 491 |  | Q: 77 | C23 C25 C27: 4 | CA. 2343: 7 |
|  |  |  | S: 644 | G6: 4 | 3101295: 6 |
|  |  |  |  | C22 C26: 3 | 347088: 6 |
|  |  |  |  | D: 3 | CA 2144: 6 |
|  |  |  |  | (Other): 186 | (Other): 852 |

* edad presenta valores con decimales, esto puede ser porque son bebes menores a un año o porque se está aproximando su edad, en cuyo caso se presenta en la forma de xx.5. Además, se reemplazaron los valores NA con la media de los valores restantes, y por esto es recurrente el valor 29.7.
* puerto\_embarcacion presenta dos valores sin categoría.
* nro\_cabina presenta 687 valores sin categoría.

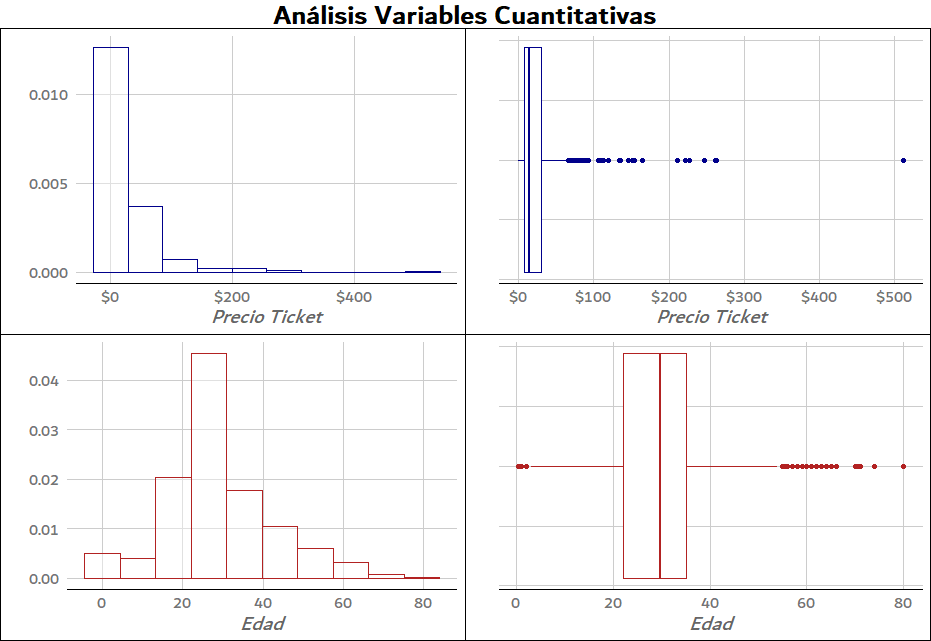
Ahora vamos a analizar las variables **clase,** **sexo**, **sobrevivio**, **edad** y **precio\_ticket**.



Con respecto a **clase**, la mayoría de los pasajeros pertenecían a la tercera clase, siendo estos más del 50%. Luego, le siguen los pasajeros de primera clase y por último los de segunda clase.

Con respecto a **sexo**, la mayoría de los pasajeros eran masculino, con mas del 60% del total.

Con respecto a **sobrevivió**, la mayoría de las personas no sobrevivieron, con aproximadamente el 60% del total.



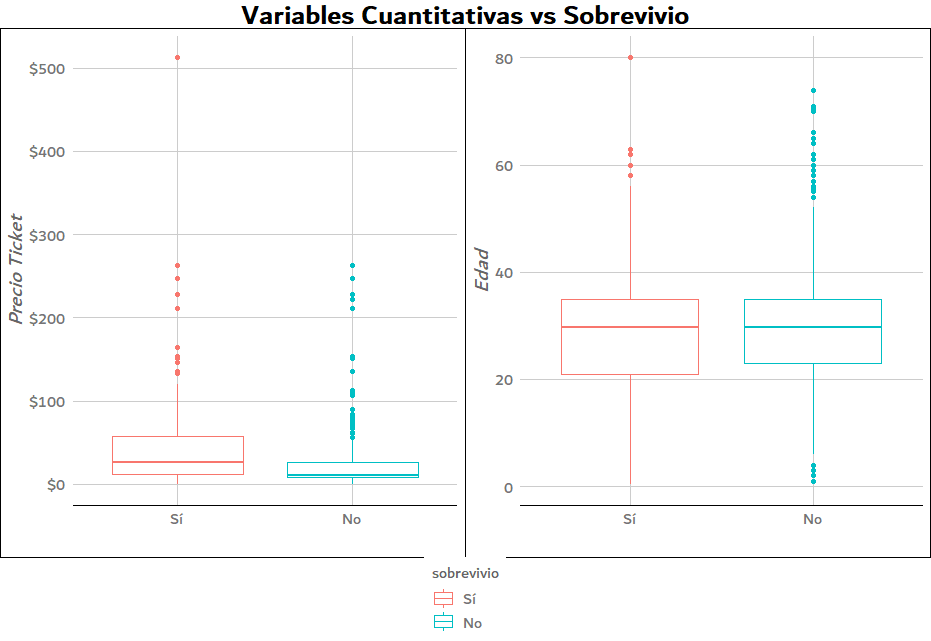
**precio\_ticket** presenta una gran cantidad de muestras en los primeros grupos y estas disminuyen a medida que aumenta el precio. Además, en el boxplot se puede ver la existencia de varios outliers, siendo el más significativo la observación que se encuentra por encima de los $500.

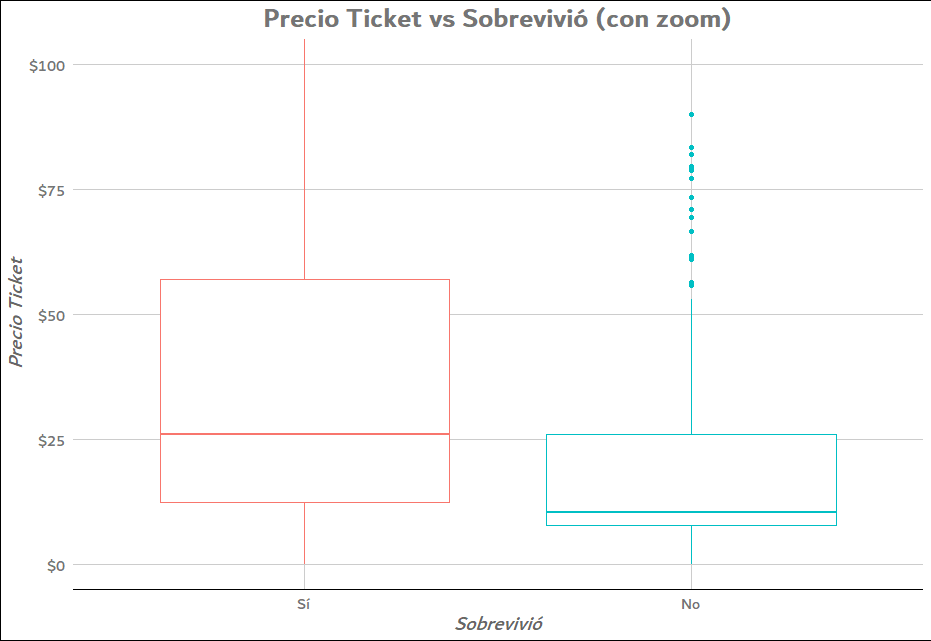
**edad** presenta una distribución que pareciera ser normal, teniendo gran cantidad de muestras cerca de las edades que se encuentran próximas a los 30 años. En el boxplot se puede ver como hay presencia de outliers, siendo el más alejado un pasajero de 80 años.

Realizamos un análisis bivariado de las variables cuantitativas contra sobrevivio.

En **edad** no hay diferencia entre el grupo de pasajeros que sobrevivieron y que no.

En **precio\_ticket** parece existir una diferencia pero no se logra distinguir bien debido a que por la presencia de los outliers se toma una escala que no es conveniente. Entonces, se realiza un zoom para poder apreciar las diferencias.





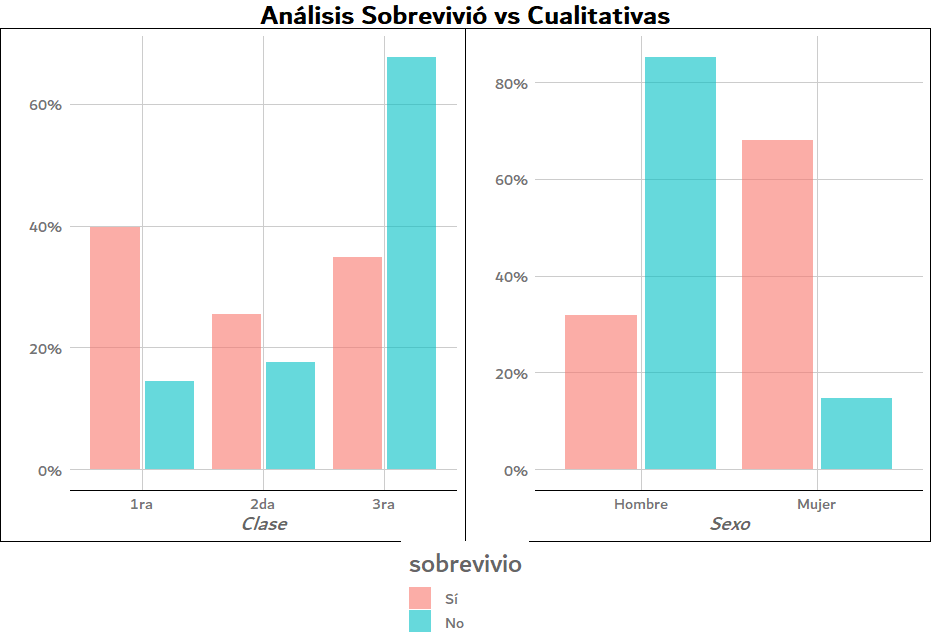
El 75% superior, es decir los datos que se encuentran por encima del primer cuartil, de las personas que sobrevivieron están por encima de la mediana del precio del ticket de las personas que no sobrevivieron. Esto quiere decir que precios grandes de los tickets se asocian con personas que sobrevivieron, y precios chicos con personas que no sobrevivieron.

Realizamos un análisis bivariado de sobrevivió contra el resto de las variables cualitativas. Primero presentamos las tablas de contingencias

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Clase** | | |
| 1 | 2 | 3 |
| **Sobrevivió** | 0 | 80 | 97 | 372 |
| 1 | 136 | 87 | 119 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Sexo** | |
| female | male |
| **Sobrevivió** | 0 | 81 | 468 |
| 1 | 233 | 109 |

No encontramos celdas con pocas observaciones en ninguna de las tablas, por lo que no será necesario realizar ningún tipo de agrupamiento.



Para **clase**: dentro de grupo de pasajeros que no sobrevivieron, la mayoría pertenece a la tercera clase, teniendo un 60% de este grupo. Luego sigue la segunda y la primera clase en porcentaje de pasajeros. Por el otro lado, dentro del grupo que sobrevivió, se presenta una distribución más homogénea, teniendo primera clase con un mayor porcentaje, y luego tercera y segunda clase.

Para **sexo**: el grupo que no sobrevivió estaba integrado en su mayoría por hombres, con más del 80%, y una minoría de mujeres. Para el grupo que sobrevivió sucede lo inverso, está compuesto por más del 60% de mujeres y el resto de los hombres.

# Regresión Logística Múltiple

Se va a realizar una regresión logística múltiple para **sobrevivio** usando las variables **clase,** **sexo**, **edad** y **precio\_ticket** como predictoras.

Como conclusiones del análisis anterior:

* En **clase** el grupo que no sobrevivió este compuesto en su mayoría por la tercera clase, seguido por la segunda y primera clase. El grupo que sobrevivió presenta una distribución más homogénea, en comparación con el grupo anterior, con primera, tercera y segunda clase.
* En **sexo** el grupo que no sobrevivió este compuesto en su mayoría por hombres y una parte de mujeres. Por otro lado, para el grupo que sobrevivió este compuesto en su mayoría por mujeres y una parte de hombres.
* En **precio\_ticket** se asocia precios chicos con el grupo que no sobrevivió y precios grandes con el grupo que si sobrevivió.
* En **edad** no hay diferencia entre ambos grupos, pero no se descarta detectar alguna cuando se combina con alguna de las restantes variables.

Se plantea un modelo inicial con todas las variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 3.47494004 | 0.41779078 | 8.31741671 | 8.99E-17 |
| clase2 | -1.07954917 | 0.2859962 | 3.77469756 | 0.0001602 |
| clase3 | -2.28007833 | 0.28002451 | -8.1424241 | 3.87E-16 |
| sexomale | -2.60485871 | 0.18764725 | 13.8816782 | 8.18E-44 |
| edad | -0.03302175 | 0.00741459 | 4.45361677 | 8.44E-06 |
| precio\_ticket | 0.00071977 | 0.00210017 | 0.34271861 | 0.73181016 |

Se obtuvo que casi todos los valores estimados son significativos excepto por el de **precio\_ticket** con un pvalor = 0.731. Por ende se piensa descartar esta variable del modelo. El AIC del modelo es 817.17.

Interpretamos dos valores:

* Por cada año de **edad**, los odds de **sobrevivir** se reducen un 3,25%, manteniendo el resto de las variables constantes .
* Los odds de que un **hombre** sobreviva son un 92.61% inferiores que los de las **mujeres** .
* Los odds de que una persona de **tercera clase** sobreviva son un 89.77% inferiores que las de una persona **de primera clase** .
* Los odds de que una persona de **segunda clase** sobreviva son un 66.025% inferiores que las de una persona **de primera clase** .

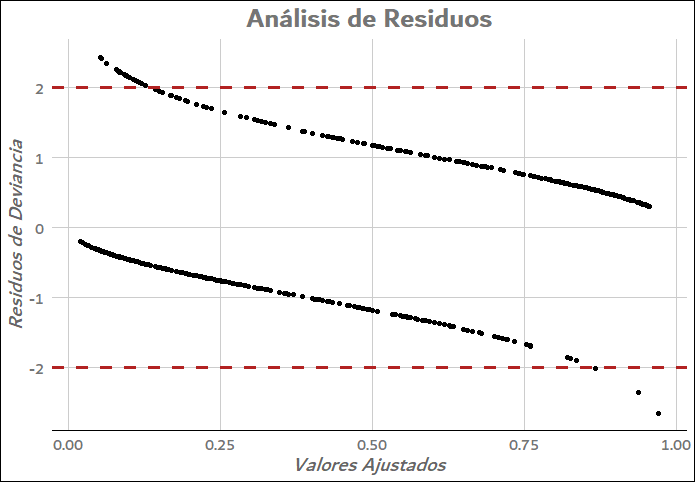
Volvemos a realizar un modelo sin la variable **precio\_ticket** porque el valor estimado no resulto ser significativo.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 3.54473662 | 0.36537184 | 9.70172362 | 2.96E-22 |
| clase2 | -1.12216121 | 0.25773493 | -4.35393536 | 1.34E-05 |
| clase3 | -2.32917186 | 0.2408897 | -9.66903873 | 4.08E-22 |
| sexomale | -2.61131471 | 0.18670877 | -13.9860315 | 1.90E-44 |
| edad | -0.03330413 | 0.00736971 | -4.51905405 | 6.21E-06 |

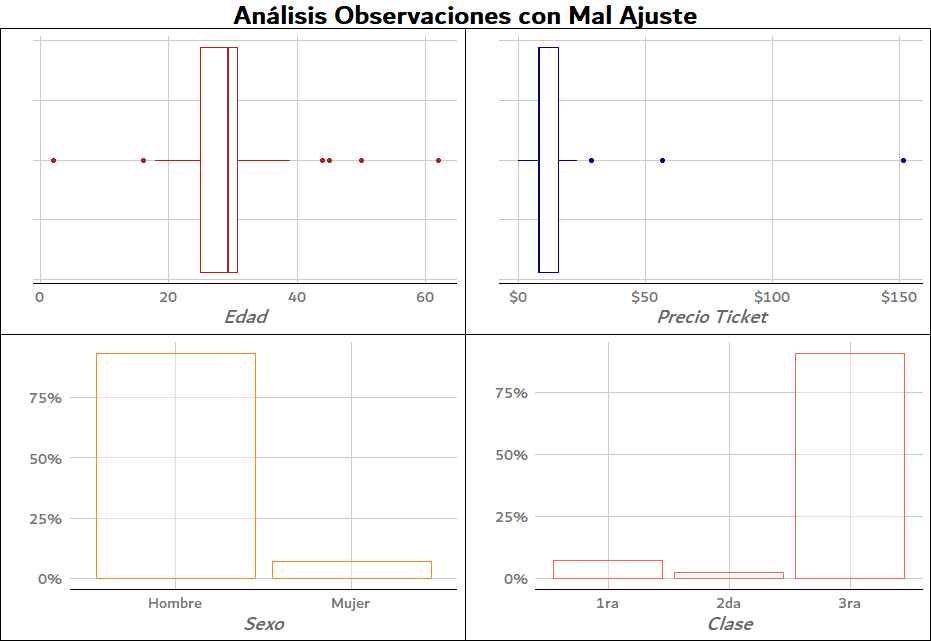
En este modelo todas las variables presentan valores estimados resultan ser significativos. El AIC en este caso es 815.29.

Ahora, seleccionamos variables usando la técnica de stepwise. El resultado fue el mismo modelo.

Ahora intentamos identificar puntos de mal ajuste usando los resuidales de deviancia



Se observan varios puntos cuyos residuos en valor absoluto son mayores a dos. Vemos si logramos encontrar algún patrón en estos datos.



En **sexo** la mayor parte de las observaciones corresponden a hombres, con más del 80%.

En **clase** la mayor parte de observaciones corresponden a la tercera clase, con más del 80%.

En **precio\_ticket** se distingue un outlier que es significativamente mayor que el resto. Esta observación ya se había detectado en el análisis inicial de los datos.

En **edad** los datos se encuentran centrados alrededor de los 30 años de edad. También se distinguen varios outliers, la observación con edad mayor a 60 años se había detectado al inicio del análisis como un outlier.

Ahora buscamos los outliers, usando el leverage, y los puntos influyentes, usando la distancia de cook. Los resultados se resumen en la siguiente tabla, nos quedamos únicamente con los primeros tres valores porque son los que presentan una distancia de cook mayor con respecto a los demás.

|  |  |  |
| --- | --- | --- |
| Nro. observación | leverage | distancia\_cook |
| 298 | 0.00363595 | 0.02373252 |
| 631 | 0.01612686 | 0.0188116 |
| 571 | 0.00845336 | 0.01637408 |

Las observaciones **298** y **571** presentan mal ajuste, residuos de deviancia en valor absoluto mayor a dos. Por esto vamos a analizarlas, estas observaciones son

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nro. Observación | sobrevivio | clase | sexo | edad | precio\_ticket |
| 298 | 0 | 1 | female | 2 | 151.55 |
| 571 | 1 | 2 | male | 62 | 10.5 |

La observación **571** corresponde con el outlier detectado de **edad** en el análisis inicial de los datos y el de los valores con un ajuste pobre.

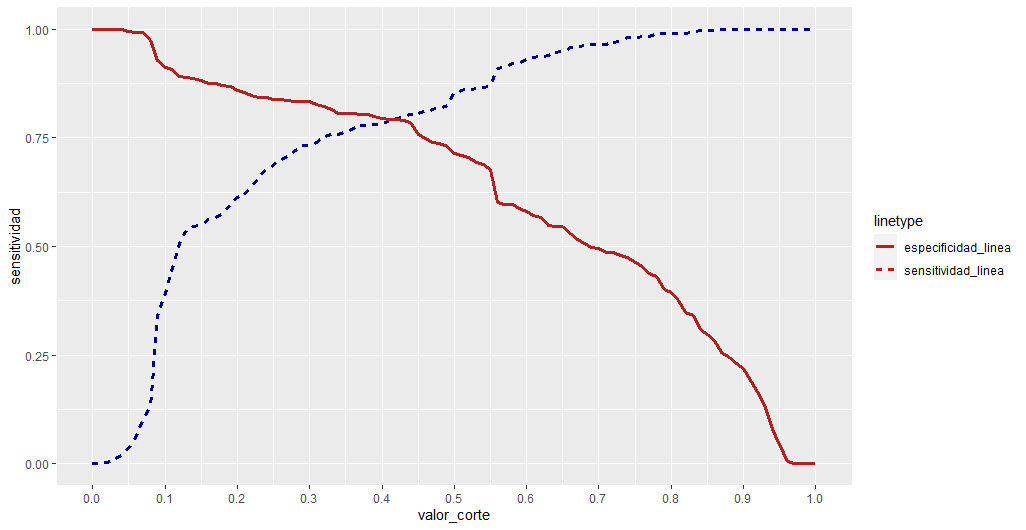
La observación **298** corresponde con el outlier detectado de **precio\_ticket** en el análisis inicial de los datos y el de los valores con un ajusto pobre.

Se realiza un tercer modelo sin estas observaciones para probar si se consigue un mejor ajuste, aplicando una selección de variables con stepwise.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 3.75068467 | 0.37513532 | 9.99821797 | 1.55E-23 |
| clase2 | -1.22194324 | 0.26149867 | -4.6728468 | 2.97E-06 |
| clase3 | -2.42183071 | 0.24509539 | -9.88117629 | 5.02E-23 |
| sexomale | -2.65714059 | 0.18901167 | -14.0580767 | 6.87E-45 |
| edad | -0.03702353 | 0.00751522 | -4.92647638 | 8.37E-07 |

El modelo obtenido emplea las mismas variables predictoras que el original, siendo todas significativas. El AIC del modelo es de 803.22, siendo notablemente menor al original.

Ahora que contamos con un modelo vamos a buscar el valor de corte. Para esto optamos graficar las curvas de sensitividad y especificad para los posibles valores de corte y elegir aquel que haga máxima ambas curvas.



Como se observa el valor crítico se encuentra entre 0.4 y 0.45, hace que tanto la sensitividad y la especificidad se encuentren por encima de 0.75. En principio no existe un valor que haga ambos parámetros iguales. Sin embargo, podemos buscar el valor que haga que la diferencia entre ambos parámetros se encuentra por debajo de un umbral, en nuestro caso a 0.01 como umbral. Se obtuvieron estos posibles valores:

|  |  |  |  |
| --- | --- | --- | --- |
| valor\_corte | sensitividad | especificidad | accuracy |
| 0.41 | 0.78649635 | 0.79178886 | 0.78852643 |
| 0.42 | 0.79379562 | 0.79178886 | 0.79302587 |
| 0.43 | 0.79744526 | 0.7888563 | 0.79415073 |

Se opta por el valor de corte de 0.43 porque es el que maximiza la accuracy. A continuación presentamos la matriz de confusión.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Observados** | |
| Si | No |
| **Predicciones** | Si | 266 | 110 |
| No | 76 | 439 |

Como se observa, con este modelo se logra clasificar 705 observaciones correctamente y 186 de forma incorrecta, con una accuracy del 79,44%