Regresión Logística para Estimar Probabilidad de Supervivencia de Pasajeros del Titanic

Miguel González Borja
Andrea Marín Alarcón

9 de abril de 2019

Introducción

Se utilizo una base de datos de pasajeros del Titanic, con 891 pasajeros, de los cuales 342 sobrevivieron. Para cada pasajero, se registráron las siguientes variables:

- passengerID - ID del pasajero
- survived - 1 si sobrevivió el pasajero, 0 si no
- pclass - Clase del pasajero
- name - Nombre del pasajero
- sex - Sexo del pasajero
- age - Edad del pasajero
- sibsp - Número de hermanos/pareja abordo del Titani
- parch - Número de padres/hijos abordo del Titanic
- ticket - ID del boleto del pasajero
- fare - Costo del boleto del pasajero
- cabin - Número de cabina
- embarked - Puerto donde embarcó el pasajero

En la base de datos original habían 177 datos faltantes de edad, 687 de la cabina y 2 del puerto de embarque. Debido a que faltaba más del 50 % de los datos de la cabina, se decidió no utilizar esta variable para el modelo de regresión. Los valores faltantes de edad se sustituyeron con la media de la clase a la que pertenece el pasajero; y, los dos valores faltantes del puerto de embarque se sustituyeron por el puerto de dónde salieron más personas, el puerto S.

Asimismo se creo una nueva variable *companion*, que combina las variables *sibsp* y *parch*, la cual es 1 si el pasajero estaba acompañado, es decir tenía algún hermano/pareja/padre/hijo en el barco y 0 si no.

Se tomó como grupo control a los pasajeros de clase 1, mujeres, menores de 18 y que salieron del puerto Q. Debido a esto, se creó una nueva variable *is_adult*, la cual vale 1 si el pasajero tiene 18 años o más y 0 en otro caso. Además, para las variables categóricas (*pclass, sex, embarked*) se crearon las variables *dummies*: *class_2, class_3, is_male, port_C, port_S*

Regresión Logística

Se tomó una muestra aleatoria del 65 % para el entrenamiento del modelo y 35 % para la prueba.

Se hizo un primer modelo tomando en cuenta todas las variables, el cual tuvo un *score* de 77.56 %, con 242 aciertos.

	No sobrevivió (predicción)	Sobrevivió (predicción)
No sobrevivió	172	26
Sobrevivió	44	70

Cuadro 1: Tabla cruzada del modelo 1

Sin embargo, los valores p para las variables *fare* y *companion* eran mayores a 0.05, por lo que se hizo un modelo, sin estas que también obtuvo un *score* de 77.56 % y 242 aciertos, pero en este caso predijo mejor los pasajeros que sobrevivieron.

	No sobrevivió (predicción)	Sobrevivió (predicción)
No sobrevivió	167	31
Sobrevivió	39	75

Cuadro 2: Tabla cruzada del modelo 2

Finalmente se hizo un modelo únicamente con las variables *is_male, class_2, class_3, is_adult* que tuvo un mejor desempeño con un *score* de 78.21 % y 244 aciertos.

	No sobrevivió (predicción)	Sobrevivió (predicción)
No sobrevivió	167	31
Sobrevivió	37	77

Cuadro 3: Tabla cruzada del modelo 3

Probabilidad de Supervivencia

Tomando en cuenta el último modelo, nuestra función *logit(Y|X=x)* es:

$$
\begin{aligned}
logit(Y|X=x) &= ln\Big(\frac{P(Y|X=x)}{1-(Y|X=x)}\Big) \\
&= 2{,}637 - 0{,}41I_{class_2}(x) - 1{,}82I_{class_3}(x) - 2{,}48I_{is\_male}(x) - 0{,}64I_{is\_adult}(x)
\end{aligned}
$$

Con esto, podemos obtener la probabilidad de supervivencia como:

$$
\begin{aligned}
\frac{P(Y|X=x)}{1-(Y|X=x)} &= e^{logit(Y|X=x)} \\
\Rightarrow P(Y|X=x) &= \frac{e^{logit(Y|X=x)}}{1+e^{logit(Y|X=x)}}
\end{aligned}
$$

Esto nos da las siguientes probabilidades de supervivencia:

Sexo	Edad	Clase 1	Clase 2	Clase 3
Mujer	Niño	0.93323611	0.90233503	0.69335371
Mujer	Adulto	0.87952708	0.82833979	0.54148164
Hombre	Niño	0.53909583	0.4360153	0.15909899
Hombre	Adulto	0.37922634	0.28763742	0.0899305

Podemos observar que las mujeres menores de 18 en primera clase son las que tienen mayor probabilidad de sobrevivir (0.9332) y los hombres mayores de 18 en tercera clase tienen la menor probabilidad de sobrevivir (0.08993).