Regresión Logística

Efrén Jiménez

1 de setiembre de 2016

Análisis del Problema

Cuando sucede un accidente en un medio de transporte colectivo de gran tamaño como un barco o un avión,una de las situaciones más sensibles es cómo saber las posibilidades de que un pasajero sobreviva. En el caso del Titanic, ya no hay nada que se pueda hacer sobre los pasajeros, no hay decisión que se pueda tomar. Sin embargo, la forma en que se van a manejar los datos para determinar si un pasajero sobrevivió o no se puede transportar a otros escenarios actuales.

Al momento en que sucede un accidente, conforme se empiecen a encontrar los primeros sobrevivientes y/o los primeros cuerpos, podríamos comenzar a crear un modelo que permita predecir las probabilidades de otros pasajeros de haber sobrevivido. Esto podría ayudar en el momento a los cuerpos de rescate a saber qué es necesario tener en la escena del accidente, para poder reaccionar inmediatamente ante cualquier situación.

Entendimiento de los Datos

El conjunto de datos que se va a analizar cuenta con 891 observaciones y 12 variables: - PassengerID: Número de pasajero; numérico, rango de 1 a 891. - Survived: Indica si el pasajero sobrevivió o no; valores posibles: 0 (no), 1 (sí). - Pclass: Clase en la cual viajaba el pasajero; valores posibles: 1, 2 ó 3. - Name: Nombre del pasajero; variable cualitativa categórica. - Sex: Género del pasajero; valores posibles: male (hombre) y female (mujer). - Age: Edad del pasajero; rango: 0.42 a 80 años, con 177 valores faltantes. - SibSp: Cantidad de hermanos o cónyuges en el barco; numérica, rango de 0 a 8. - Parch: Cantidad de hij@s o padres a bordo; numérica, rango de 0 a 6. - Ticket: Número de tiquete; variable cualitativa categórica. - Fare: Monto pagado por el pasajero por su tiquete; rango: de 0 a 512.33. - Cabin: Cabina en la cual esaba hospedado el pasajero; variable cualitiva categórica. - Embarked: Puerto en el cual embarcó el pasajero; variable cualitiva categórica.

Exploración de los Datos

```
# librerías utilizadas
library(titanic)
library(lattice)
library(caTools)
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
data(titanic_train)
data(titanic_test)
# crear factores
titanic_train$Survived <- factor(titanic_train$Survived)
titanic_train$Pclass <- factor(titanic_train$Pclass)
titanic_train$Sex <- factor(titanic_train$Sex)
titanic_train$Cabin <- factor(titanic_train$Cabin)
titanic_train$Embarked <- factor(titanic_train$Embarked)
str(titanic_train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```r
# Dividir el conjunto de datos en entrenamiento y prueba
set.seed(351)
splt <- sample.split(titanic_train$Survived, SplitRatio = 0.7)
datos.entrenamiento <- titanic_train[splt, ]
datos.prueba <- titanic_train[!splt, ]
```

Una vez cargados los datos, podemos comenzar a explorarlos. Para comenzar, podemos analizar la distribución de la variable Edad, en el contexto de si el pasajero sobrevivió o no: