Limpieza y análisis del dataset TITANIC

Ander Elkoroaristizabal

Mayo 2021

Contents

Descripción del dataset

El conjunto de datos que estudiaremos es el *Titanic dataset*. El objetivo primordial de este *dataset* es la creación de un modelo que prediga que pasajeros sobrevivieron al hundimiento del Titanic.

Este dataset resulta especialmente atractivo por lo interesante del tema, la variedad en las variables y la cantidad de estudios y discusiones sobre él que podemos encontrar, incluso en la propia página de discusiones de la competición en Kaggle.

El conjunto de datos "completo" tal y como se incluye en Kaggle viene ya dividido en dos subconjuntos: uno de entrenamiento y otro de evaluación. Podemos utilizar ambos en el periodo de limpieza y evaluación, pero sólo el primero puede ser analizado, dado que al segundo le falta la columna `Survived` que se debe predecir. Esto se debe a que sobre este segundo conjunto está pensado para que efectuemos nuestras predicciones sobre él y subamos estas predicciones a la competición a la que pertenece.

El conjunto contiene 10 variables además de la variable respuesta `survived`, con 891 registros en el conjunto de entrenamiento y 418 registros en el conjunto de evaluación.

A continuación mostramos el diccionario de datos:

Variable	Definición	Claves
survival	Variable binaría indicadora de la supervivencia.	1 = Sí, 0 = No.
pclass	Clase del billete de embarque.	1 = Primera clase, 2 = Segunda clase, 3 = Tercera clase
sex	Sexo	
Age	Edad, en años	
sibsp	Número de hermanos/esposas también en el Titanic.	
parch	Número de padres/hijos también en el Titanic.	
ticket	Código alfanumérico del billete.	
fare	Precio del billete.	
cabin	Número de cabina.	

Variable	Definición	Claves
embarked	Puerto de embarque.	C = Cherbourg, Q = Queenstown, S = Southampton

Integración y selección de los datos de interés a analizar.