Ejercicios clase: Logística

Regresión logística con dataset Titanic

1. Carga el dataset de Titanic

2. Ahora analizaremos más generalmente el dataset:
    a. Cuantos NA hay en el dataset y en que columnas?
    b. Que nos dice la variable *SibSp* y *Parch*? Cómo se distribuyen estas variables?
    c. Estudia la función countplot del paquete seaborn. Haz un countplots utilizando las columnas *Pclass* y *Sex*.
    d. Haz un histograma de la variable *Age*
    e. Qué columnas se podrían descartar “en principio” de un modelo solo con observar que significan?
    f. La columna Cabin tiene muchos missings, con que podría tener relación esta columna? Crea una columna para decir si esta variable está informada. Haz un ‘group by’ con esta columna junto a otras variables para encontrar alguna posible relación.
    g. Mira las relaciones que puede tener *Embarked* con *Survived.*

3. Ahora vamos a ajustar modelos logístico a partir de las columnas : *'Survived','Pclass','Sex','Age','SibSp','Parch','Fare'.*
    a. En las filas donde *Age* sea NA introduce la media total.
    b. Transforma la columna *Pclass, Sex* en strings.
    c. Convierte *Pclass, Sex* en dummies. Quita las columnas que escojas como variables base.
    d. Ajusta un modelo Logístico con todas las variables.
    e. Que Accuracy?
    f. Dibuja la curva ROC y calcula el AUC.
    g. Obten la confusion matrix. Haz un plot de ella.
    h. Observa los p-valores del modelo? Que variables podríamos descartar?
    i. Reentrena el modelo sin las columnas no significativas. Vuelve a obtener todas las métricas del modelo.