Ejercicios clase 25: Logística

11/02/2020

Regresión logística con dataset Titanic

1. Entra en Kaggle.com y busca por el dataset Titanic. Acepta las reglas y descárgate el dataset. Carga en Python, a través de pandas, el csv llamado “train” y realiza un head para comprobar que has cargado correctamente los datos.

2. Primero vamos a analizar la variable *Pclass* a través de una regresión logística:
    a. Cuenta cuantos pasajeros de cada clase hay en el dataset. Haz un barplot.
    b. Calcula el tanto % de pasajeros de cada clase hay en el dataset. Haz un barplot.
    c. Haz una tabla de contingencia con la relación *Pclass/Survived*, calcula el porcentaje de personas que están en cada celda.
    d. Haz una tabla de porcentajes relativa, según la *Pclass* y según la variable *Survived*. De los pasajeros que sobrevivieron que % eran de la clase 2? De los pasajeros de la clase 3 que % sobrevivió? Haz un barplot de cada tabla de contingencia. Que nos dicen estos gráficos? Que nos dicen sobre las *Pclass*?
    e. Calcula los odds de cada clase, interpreta los resultados.
    f. Calcula los odds ratio escogiendo como base la classe 3. Interpreta los resultados.
    g. Haz una regresión logística con la variable *Pclass.* Coinciden los valores con los odds ratio estudiados anteriormente? Interpreta el incremento de odds ratio.

3. 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 packete 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.*

4. 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.