Hoja informativa: clasificación desequilibrada

Práctica

```
#entrenamiento de modelo con ajuste de peso de clase
model = LogisticRegression(class_weight='balanced', random_state=12345)
```

```
# concatenación de tablas

pd.concat([table1, table2])
```

```
# mezcla de observaciones

features, target = shuffle(features, target, random_state=12345)
```

```
# Eliminar observaciones aleatorias de la tabla
# frac es una fracción de las observaciones de la tabla inicial que devolverá el método

features_sample = features_train.sample(frac=0.1, random_state=12345)
```

```
# Calcula las probabilidades de clases

probabilities = model.predict_proba(features)
```

```
# itera sobre elementos del rango
# los pasos primero y último pueden ser números fraccionarios

for value in np.arange(first, last, step):
```

```
# trazado de curva ROC
# Devuelve TFP, TVP y umbrales

from sklearn.metrics import roc_curve

fpr, tpr, thresholds = roc_curve(target, probabilities)
```

```
# cálculo de AUR-ROC
# toma los valores verdaderos de las clases de observación y la predicción de probabilidad de la clase

from sklearn.metrics import roc_auc_score
auc_roc = roc_auc_score(target_valid, probabilities_one_valid)
```

Teoría

El **sobremuestreo** es una técnica de equilibrio de clase que aumenta el número de observaciones al duplicar varias veces las observaciones de clase más raras.

El submuestreo es una técnica de equilibrio de clase que disminuye el número de observaciones eliminando de forma aleatoria las observaciones de la clase mayoritaria.

Un umbral es el límite de probabilidad que separa a las clases positivas y negativas..

La curva PR es un diagrama que muestra la precisión contra recall.

Tasa de **verdaderos positivos, o TVP** es el resultado de dividir las respuestas VP entre todas las respuestas positivas:

$$TVP = \frac{VP}{P}$$

Tasa de **falsos positivos, o TFP** es el resultado de dividir las respuestas FP entre todas las respuestas negativas.

$$TFP = \frac{FP}{N}$$

Curva ROC es un gráfico que muestra la tasa de verdaderos positivos frente a la tasa de falsos positivos.

AUC-ROC es una métrica de evaluación para tareas de clasificación que equivale al área bajo la curva ROC. Los valores de la métrica están en el rango de 0 a 1. El valor AUC-ROC para un

modelo aleatorio es 0.5.