

# Несбалансированная классификация

## Обучение модели со взвешиванием классов

```
In model = LogisticRegression(class_weight='balanced', random_state=12345)
```

## Соединение таблиц

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

## Перемешивание объектов

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

## Удаление случайных объектов из таблицы

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

# frac – доля объектов исходной таблицы, которая будет возвращена методом
```

## Вычисление вероятностей классов

```
In probabilities = model.predict_proba(features)
```

## Цикл по значениям диапазона

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

# first, last, step – могут быть дробными
```

## Построение ROC-кривой

```
In from sklearn.metrics import roc_curve

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

# Возвращает значения FPR, TPR и thresholds
```

## Подсчёт метрики AUC-ROC

```
In from sklearn.metrics import roc_auc_score

auc_roc = roc_auc_score(target_valid, probabilities_one_valid)

# принимает на вход
# истинные значения классов объектов и предсказанную вероятность класса
```

# Словарь

## Upsampling

техника балансирования классов, которая заключается в увеличении числа объектов меньшего класса путём многократного копирования объектов меньшего класса

## Downsampling

техника балансирования классов, которая заключается в уменьшении числа объектов большего класса путём случайного удаления объектов большего класса

## Порог

граница вероятности, которая отделяет положительный и отрицательный классы

## PR-кривая

график зависимости точности от полноты

## Доля истинно положительных ответов (*True Positive Rate, TPR*)

отношение числа истинно положительных ответов к числу положительных ответов:

$$TPR = \frac{TP}{P}$$

## Доля ложноположительных ответов (*False Positive Rate, FPR*)

отношение числа ложноположительных ответов к числу положительных ответов:

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

## ROC-кривая (кривая ошибок)

график зависимости доли истинно положительных ответов от доли ложноположительных ответов

## AUC-ROC

метрика качества классификации, равная площади под ROC-кривой. Значения метрики изменяются от 0 до 1. AUC-ROC случайной модели равна 0.5