

# Конспект по теме "Метрики регрессии"

## Коэффициент детерминации

**Коэффициент детерминации**, или **метрика R2**, вычисляет долю среднеквадратичной ошибки модели от *MSE* среднего, а затем вычитает эту величину из единицы. Увеличение метрики означает прирост качества модели.

Формула расчёта *R2* выглядит так:

$$R^2 = 1 - \frac{\text{MSE модели}}{\text{MSE среднего}}$$

- Значение метрики *R2* равно единице только в одном случае, если *MSE* нулевое. Такая модель предсказывает все ответы идеально.
- *R2* равно нулю: модель работает так же, как и среднее.
- Если метрика *R2* отрицательна, качество модели очень низкое.
- Значения *R2* больше единицы быть не может.

В библиотеке *sklearn.metrics* есть функция для подсчёта этой метрики — *r2\_score()*:

```
from sklearn.metrics import r2_score

print("R2 =", r2_score(target, predicted))
```

## Среднее абсолютное отклонение

Дадим общепринятые в *Data Science* обозначения:

$$y_i$$

- Значение целевого признака для объекта с порядковым номером  $i$  в выборке, на которой измеряется качество. Нижний индекс показывает номер объекта.

$$\hat{y}_i$$

- Значение предсказания для объекта с порядковым номером  $i$ , например, в тестовой выборке.

Ещё одна метрика качества — **MAE** (*mean absolute error*). Она похожа на *MSE*, но в ней нет возведения в квадрат. Запишем метрику в обозначениях, принятых в Data Science.

Отклонение объекта:

$$\text{Отклонение} = y_i - \hat{y}_i$$

Чтобы в новой метрике избавиться от разницы между недооценкой и переоценкой, вычисляется **абсолютное отклонение**. Это модуль от отклонения:

$$\text{Абсолютное отклонение} = |y_i - \hat{y}_i|$$

Чтобы собрать отклонения по всей выборке, дополним обозначения:

$$N$$

- Количество объектов в выборке.

$$\sum_{i=1}^N$$

- Суммирование по всем объектам выборки ( $i$  меняется от 1 до  $N$ ).

Формула **среднего абсолютного отклонения**, или **MAE**:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

В библиотеке `sklearn.metrics` есть функция для подсчёта этой метрики — `mean_absolute_error()`:

```
from sklearn.metrics import mean_absolute_error

mae = mean_absolute_error(target, predicted))
```

Чтобы рассчитать  $MSE$ , за константу мы принимали среднее значение.

Константная модель выбирается так, чтобы значение метрики  $MAE$  было предельно низким. Нужно найти такое значение  $a$ , при котором достигается минимум:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - a|$$

Минимум получается, когда ***a*** равно медиане целевого признака.

В отличие от *MAE*, метрика *RMSE* **чувствительнее к большим значениям**: значимые ошибки сильно влияют на итоговое значение квадратного корня из среднеквадратичной ошибки. Таким образом, можно менять значение *RMSE*, не меняя значения *MAE*.