Метрики и валидация

Лекция 3

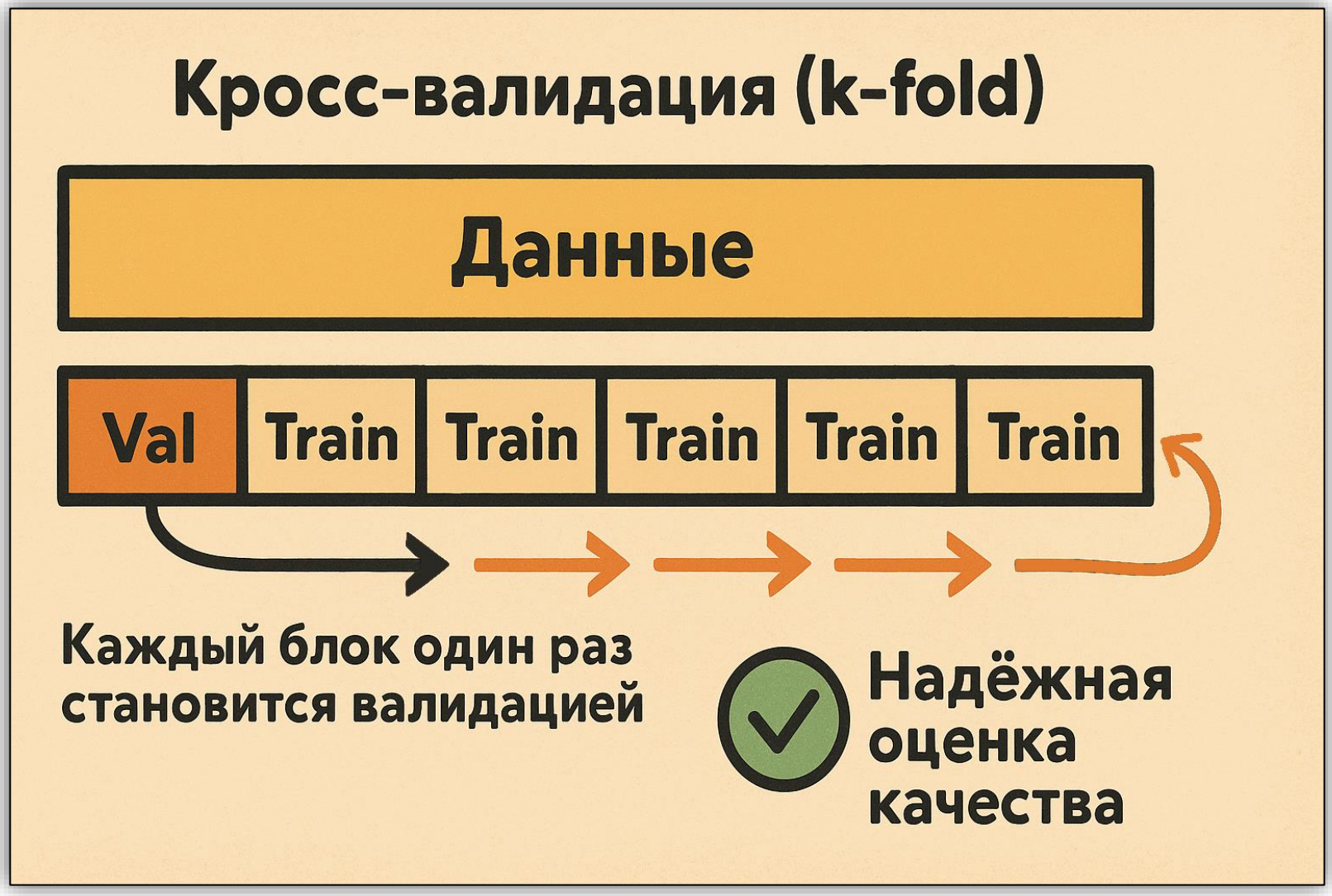
Валидация и метрики: зачем они нужны

Train / Validation / Test

Data Leakage: как испортить метрику

- **Что это:** использование информации из теста или валидации при обучении.
- **Последствия:** метрики становятся нереалистично высокими, модель «подсмотрела ответы».
- **Примеры:** нормализация по всему датасету, признаки с целевой переменной, перемешивание временных рядов.
- **Главное правило:** Train и Test полностью изолированы.

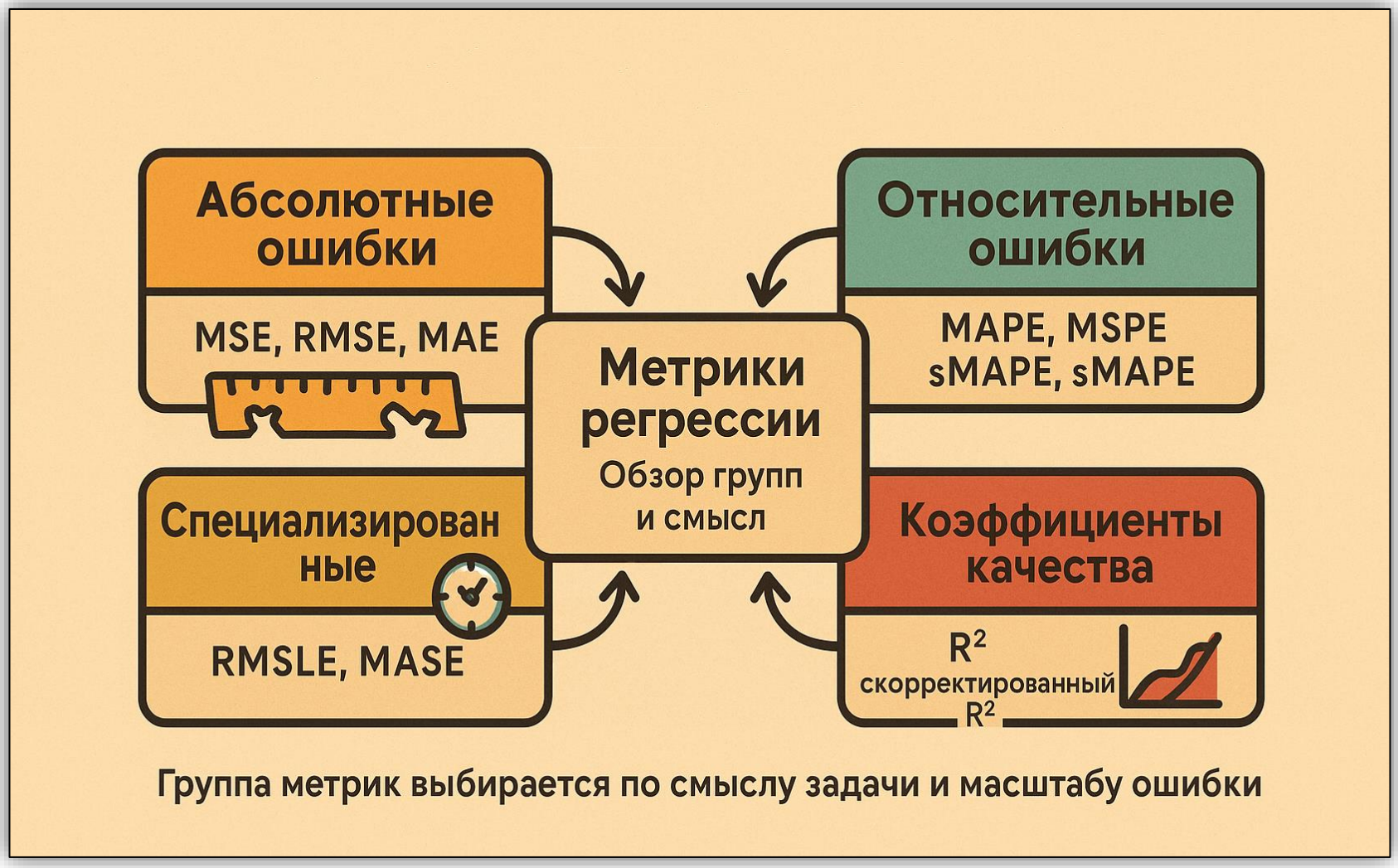
Кросс-валидация



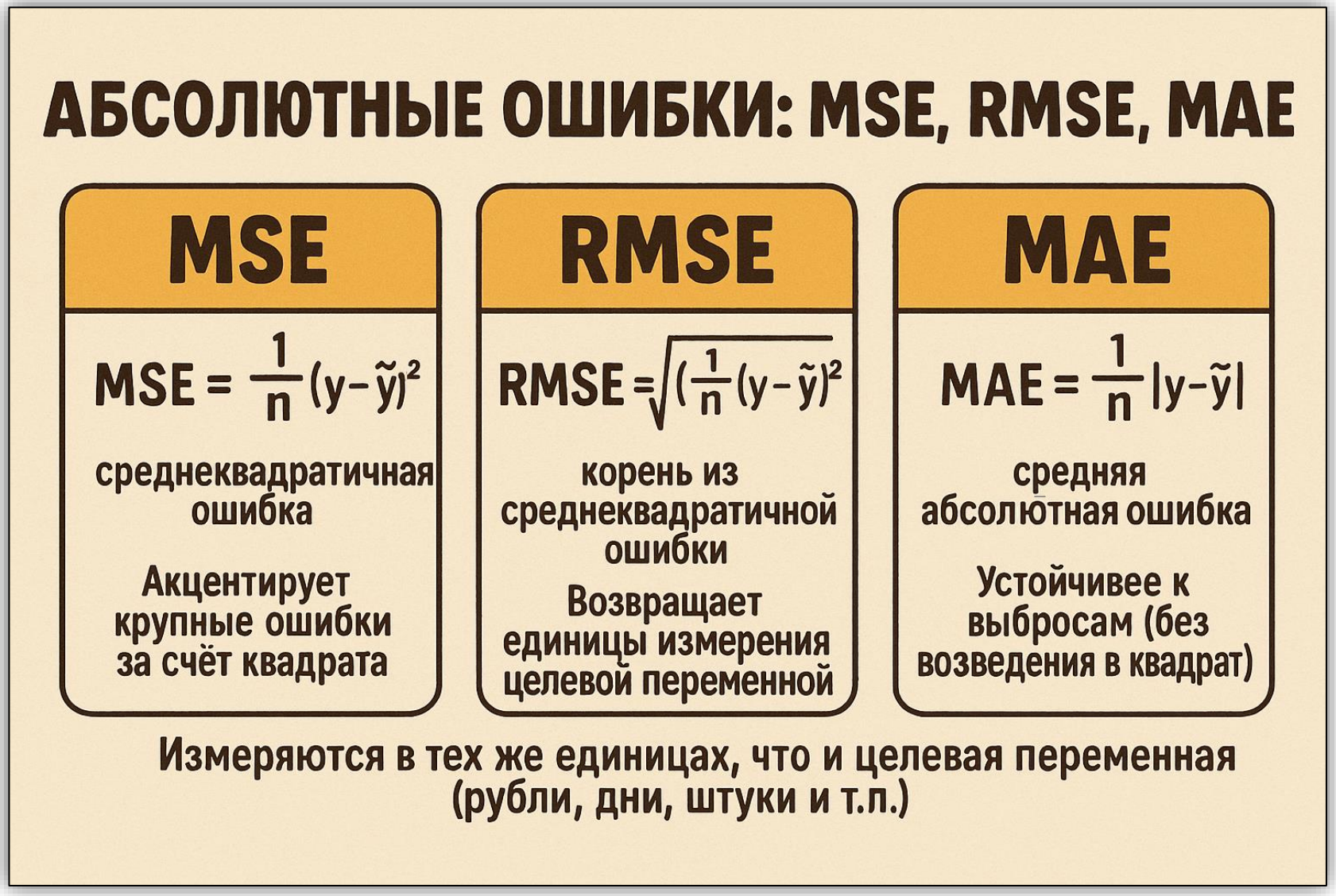
Baseline и выбор порога в классификации

- Бейзлайн: простая стратегия для сравнения (например, всегда предсказывать самый частый класс).
- Порог классификации: разделяет вероятности на классы.
- Сдвиг порога меняет баланс: ↑ Precision ⇔ ↓ Recall.
- Правильный порог зависит от задачи: медицина → Recall, финансы → Precision.

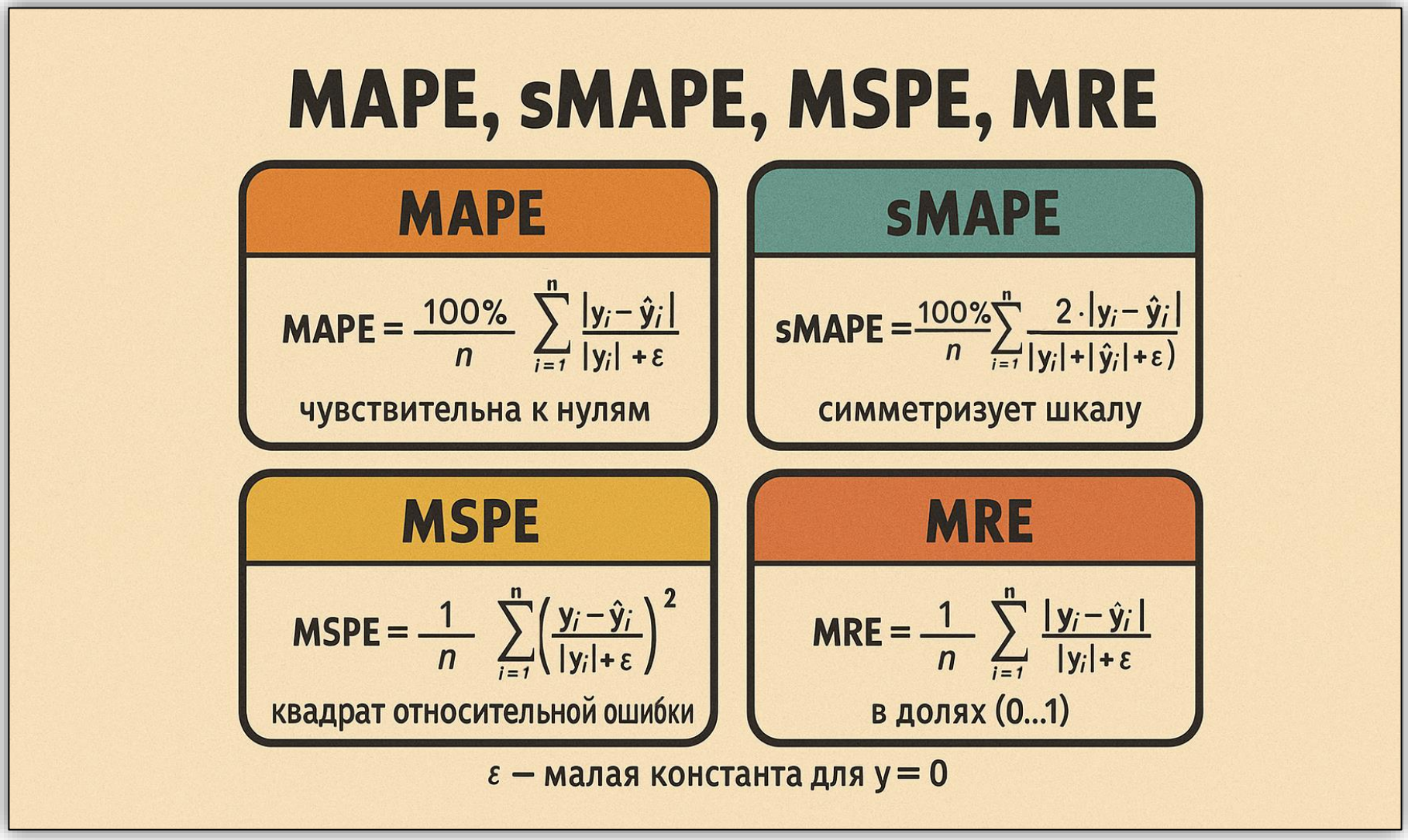
Карта метрик для регрессии



Абсолютные ошибки



Относительные ошибки



Специализированные метрики

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(\hat{y}_i + 1) - log(y_i + 1))^2}$$

Константы, равные 1, добавляемые в скобках, необходимы чтобы не допустить обращения в 0 выражения под логарифмом, поскольку логарифм нуля не существует.

$$MASE = \frac{MAE_i}{MAE_j}$$

MASE симметрична и устойчива к выбросам.

Коэффициенты детерминации: R^2 и скорректированный R^2

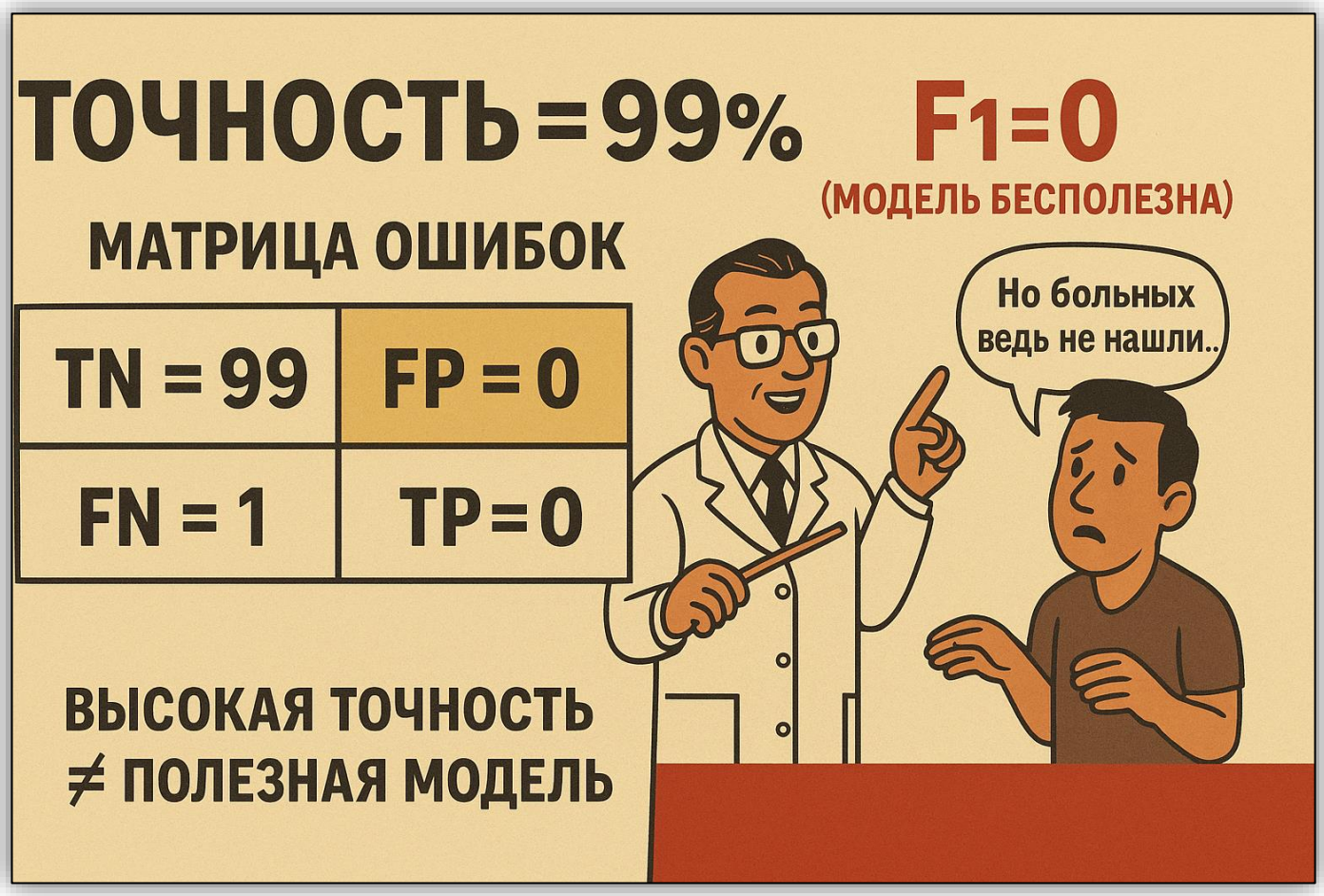
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Практически, в числителе данного выражения стоит среднеквадратическая ошибка оцениваемой модели, а в знаменателе — модели, в которой присутствует только константа.

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,(n-k)}{\sum_{i=1}^{n}(y_i - \bar{y})^2\,(n-1)}$$

где n — число наблюдений, на основе которых строится модель, k — количество переменных в модели.

От регрессии к классификации: матрица ошибок



$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision и Recall

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

Precision — *точность среди найденных*: как часто мы правы, когда говорим «да».
Recall — *доля найденных среди всех*: сколько из реальных «да» мы нашли.

TP = 40, FP = 10, FN = 50
Precision = 40 / (40+10) = 0.8
Recall = 40 / (40+50) ≈ 0.44

Precision → «Сколько из предсказанных положительных действительно положительные?»
Recall → «Сколько из настоящих положительных мы нашли?»

F1-мера: баланс точности и полноты

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision + Recall)}$$

$$F1_{macro} = \frac{1}{K} \sum_{i=1}^{K} F1_i$$

$$F1_{weighted} = \sum_{i=1}^{K} \omega_i \cdot F1_i \qquad \omega_i = \frac{n_i}{\sum_{j=1}^{K} n_j}$$

K — количество классов в задаче классификации (например, при трёхклассовой задаче K = 3).

n_i — количество объектов именно в классе i (например, 200 примеров "здоровых").

$\sum_{j=1}^{K} n_j$ - сумма по всем классам, то есть общее число примеров (например, 1000 пациентов).

ω_i — вес класса *i*,

β — задаёт баланс между точностью (*Precision*) и полнотой (*Recall*).

- при β = 1 получаем обычный F1 (равный вес);
- при β > 1 больше взвешивается полнота (Recall);
- при β < 1 больше взвешивается точность (Precision).

Итоги: валидация и метрики