

Метрики и валидация

Лекция 3

Валидация и метрики: зачем они нужны



**МЕТРИКА = КРИТЕРИЙ
УСПЕХА**



ПРОТОКОЛ ОЦЕНКИ:
Train / Val / Test, K-Fold



**МЕТРИКУ И ПРАВИЛА
ОЦЕНКИ ВЫБИРАЕМ
ЗАРАНЕЕ**



**ЧЕСТНОЕ СРАВНЕНИЕ
МОДЕЛЕЙ**



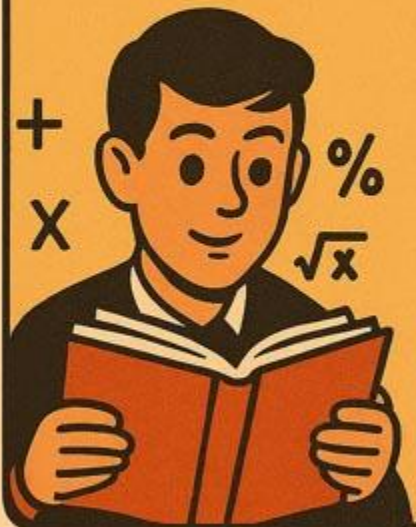
СНАЧАЛА ПРАВИЛА — ПОТОМ ИГРА

Train / Validation / Test

РАЗДЕЛЕНИЕ ДАННЫХ = ЧЕСТНАЯ ОЦЕНКА

TRAIN

(обучающая выборка)



VALIDATION

(валидационная выборка)



TEST

(тестовая выборка)



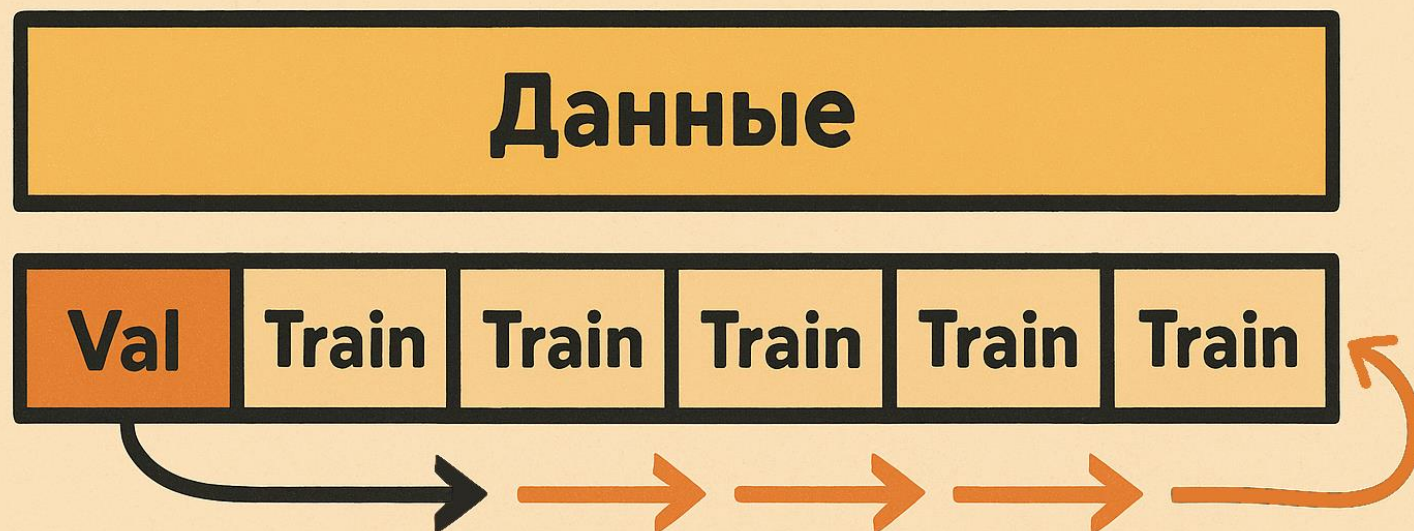
Train — учим; Validation — настраиваем; Test — проверяем результат

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

Кросс-валидация (k-fold)



Каждый блок один раз
становится валидацией

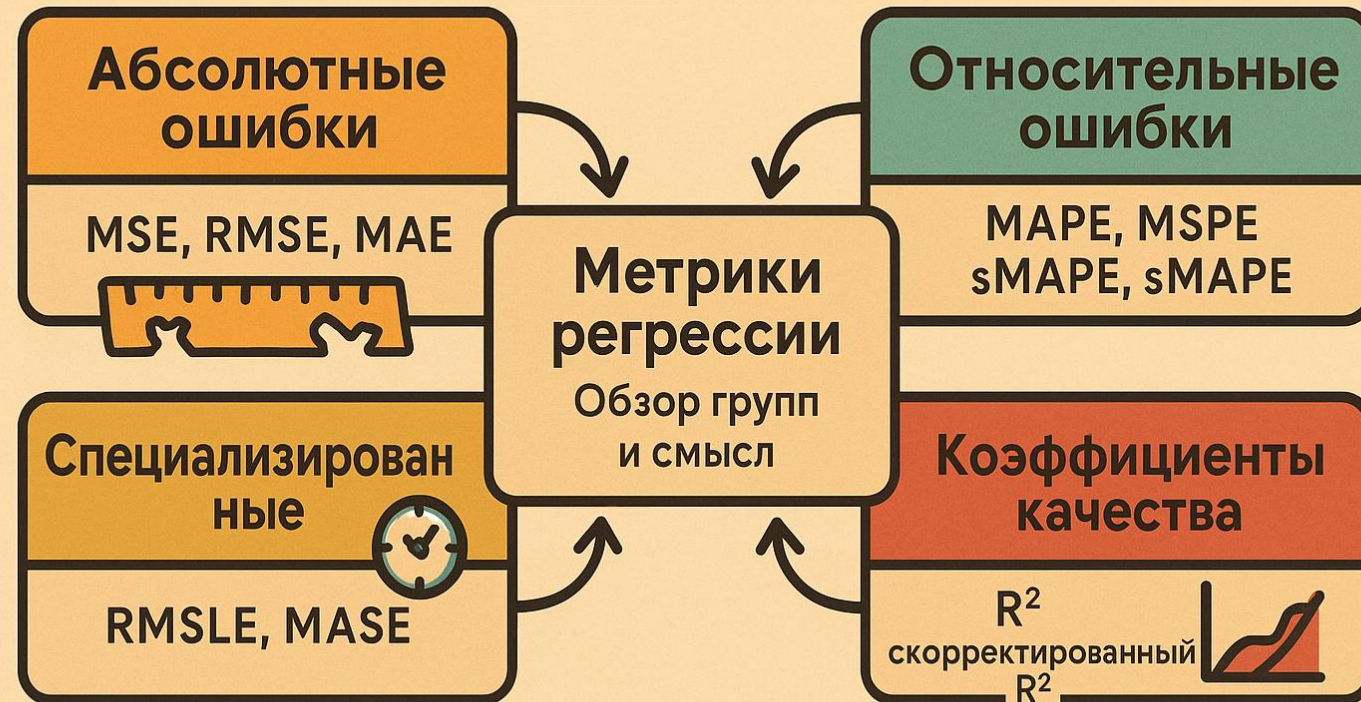


Надёжная
оценка
качества

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: \uparrow Precision \Leftrightarrow \downarrow Recall.
- Правильный порог зависит от задачи: медицина \rightarrow Recall, финансы \rightarrow Precision.

Карта метрик для регрессии



Группа метрик выбирается по смыслу задачи и масштабу ошибки

Абсолютные ошибки

АБСОЛЮТНЫЕ ОШИБКИ: MSE, RMSE, MAE

MSE

$$MSE = \frac{1}{n} (y - \tilde{y})^2$$

среднеквадратичная
ошибка

Акцентирует
крупные ошибки
за счёт квадрата

RMSE

$$RMSE = \sqrt{\frac{1}{n} (y - \tilde{y})^2}$$

корень из
среднеквадратичной
ошибки

Возвращает
единицы измерения
целевой переменной

MAE

$$MAE = \frac{1}{n} |y - \tilde{y}|$$

средняя
абсолютная ошибка

Устойчивее к
выбросам (без
возведения в квадрат)

Измеряются в тех же единицах, что и целевая переменная
(рубли, дни, штуки и т.п.)

Относительные ошибки

MAPE, sMAPE, MSPE, MRE

MAPE

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i| + \varepsilon}$$

чувствительна к нулям

sMAPE

$$\text{sMAPE} = \frac{100\%}{n} \sum_{i=1}^n \frac{2 \cdot |y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i| + \varepsilon}$$

симметризует шкалу

MSPE

$$\text{MSPE} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{|y_i| + \varepsilon} \right)^2$$

квадрат относительной ошибки

MRE

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i| + \varepsilon}$$

в долях (0...1)

ε — малая константа для $y = 0$

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

ТОЧНОСТЬ = 99%

F1=0

(МОДЕЛЬ БЕСПОЛЕЗНА)

МАТРИЦА ОШИБОК

TN = 99	FP = 0
FN = 1	TP = 0

**ВЫСОКАЯ ТОЧНОСТЬ
≠ ПОЛЕЗНАЯ МОДЕЛЬ**



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

Precision и Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

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 = $40 / (40 + 50) \approx 0.44$

Precision → «Сколько из предсказанных положительных действительно положительные?»

Recall → «Сколько из настоящих положительных мы нашли?»

F1-мера: баланс точности и полноты

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

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

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

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

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

Итоги: валидация и метрики

