

Метрики качества Кросс-валидация

Классификация и регрессия

Арсен Абдулин

Дата-аналитик

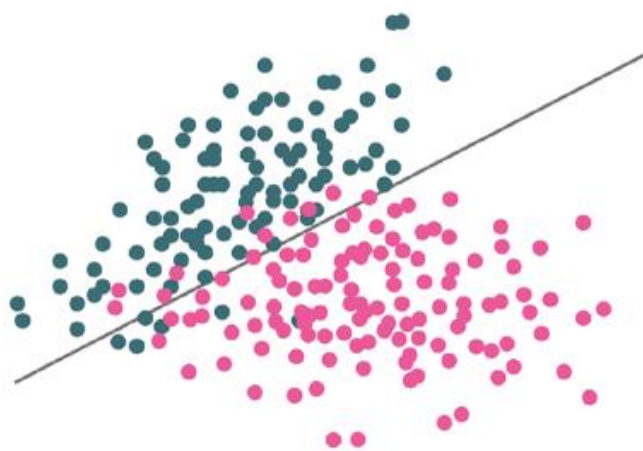
План занятия

- Типы задач ML
- Метрики задач регрессии
- MSE, RMSE, R2, MAPE, SMAPE
- Метрики задач классификации
- Confusion matrix
- Accuracy, precision, recall, F1
- FPR, TPR, AUC, ROC
- ROC-AUC

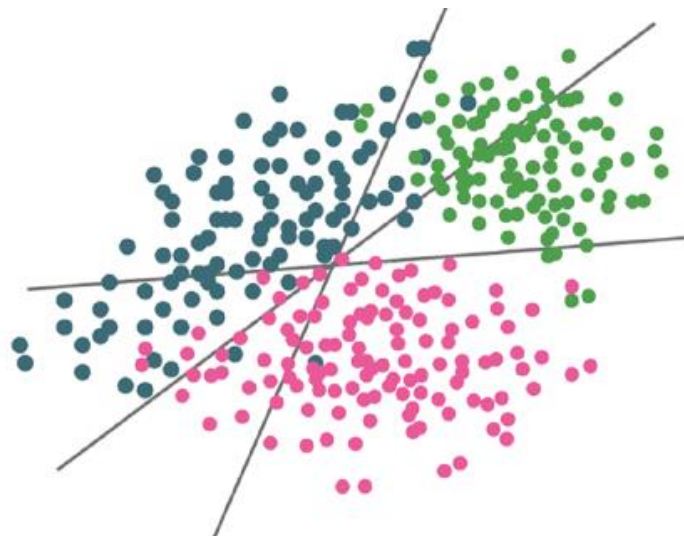
Типы задач в ML

Классификация:

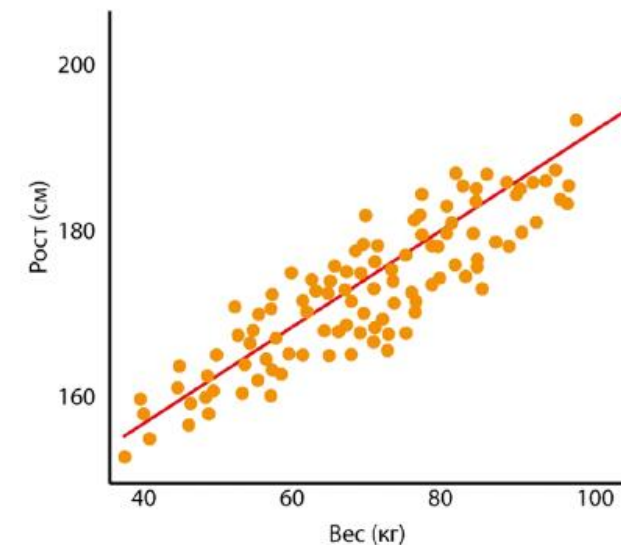
Бинарная



Многоклассовая



Регрессия



- Результат – метка класса
- Разделяет, а не описывает данные

- Результат – число
- Описывает природу данных

Метрики, в зависимости от их назначения

- A. Для задания функционала ошибки:
 - используются при обучении моделей
- B. Для подбора гиперпараметров:
 - используются при измерении качества на кросс-валидации
- C. Для оценивания итоговой модели:
 - пригодна ли модель для решения задачи? (бизнес-метрики)

Функция потерь и метрики качества

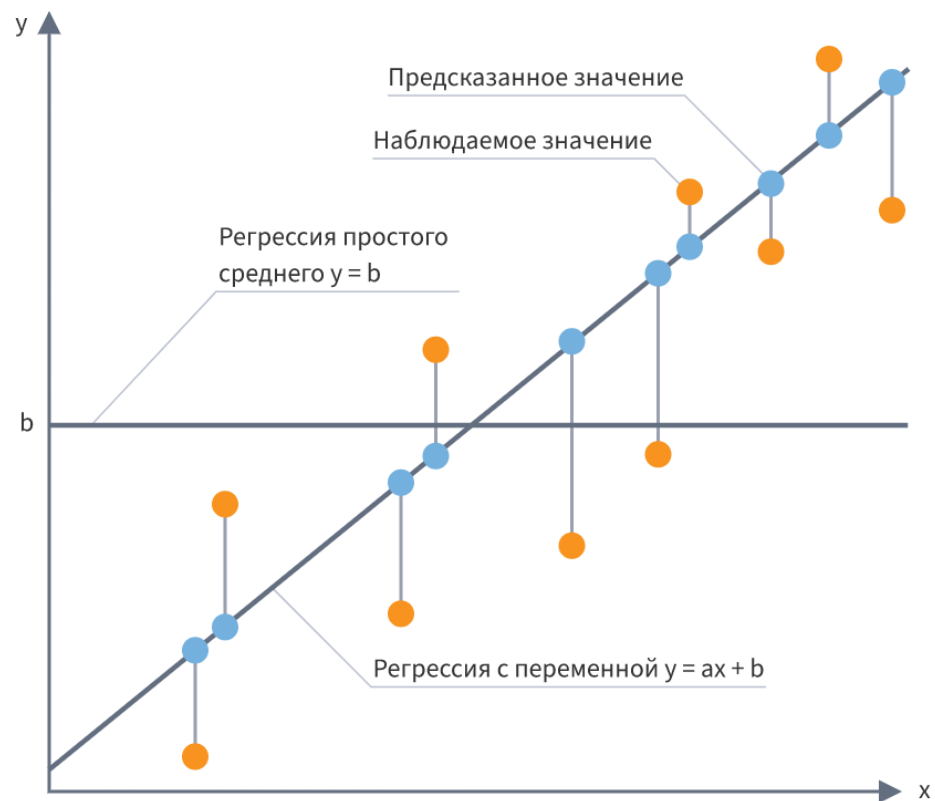
✚ Функция потерь \neq метрика качества!

Бинарная классификация:

$$\text{Accuracy}(y, y^{pred}) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[y_i = f(x_i)]$$

$$\text{Error rate} = 1 - \text{Accuracy}$$

Метрики регрессии



MSE, RMSE, R²

MSE

среднеквадратичная ошибка
штрафует за больше ошибки
неограничена сверху

$$MSE(y^{true}, y^{pred}) = \frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2$$

RMSE

корень из среднеквадратичной ошибки
- имеет размерность исходных данных

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

R-квадрат

коэффициент детерминации
макс. значение равно 1

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

MASE, MAPE, SMAPE

MAE

средняя абсолютная ошибка

устойчива к выбросам

сложнее минимизировать

$$MAE(y^{true}, y^{pred}) = \frac{1}{N} \sum_{i=1}^N |y_i - f(x_i)|$$

MAPE

средняя абсолютная
процентная ошибка

$$MAPE(y^{true}, y^{pred}) = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - f(x_i)|}{|y_i|}$$

SMAPE

симметричная средняя абсолютная
процентная ошибка

$$SMAPE(y^{true}, y^{pred}) = \frac{1}{N} \sum_{i=1}^N \frac{2|y_i - f(x_i)|}{y_i + f(x_i)}$$

Сравнение метрик регрессии

MSE

- + Позволяет подчеркнуть большие отклонения, простота вычисления
- Имеет тенденцию занижать качество модели, чувствительна к выбросам. Сложность интерпретации из-за квадратичной зависимости

MAPE

- + Является безразмерной величиной, поэтому её интерпретация не зависит от предметной области
- Нельзя использовать для наблюдений, в которых значения выходной переменной равны нулю

R²

- + Универсальность, простота интерпретации.
- Возрастает даже при включении в модель бесполезных переменных. Плохо работает когда входные переменные зависимы

RMSE

- + Простота интерпретации, поскольку измеряется в тех же единицах, что и целевая переменная
- Имеет тенденцию занижать качество модели, чувствительна к выбросам

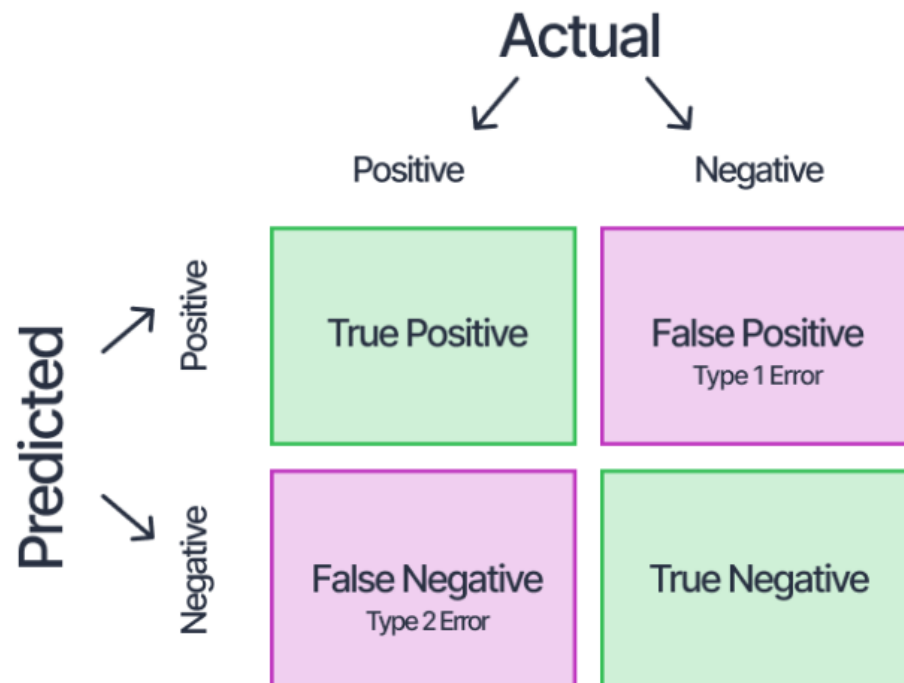
SMAPE

- + Позволяет корректно работать с предсказанными значениями независимо от того больше они фактического, или меньше
- Приближение к нулю фактического или предсказанного значения приводит к резкому росту ошибки, поскольку в знаменателе присутствует как фактическое, так и предсказанное значения

Задача классификации - Confusion matrix



Confusion matrix – Матрица ошибок



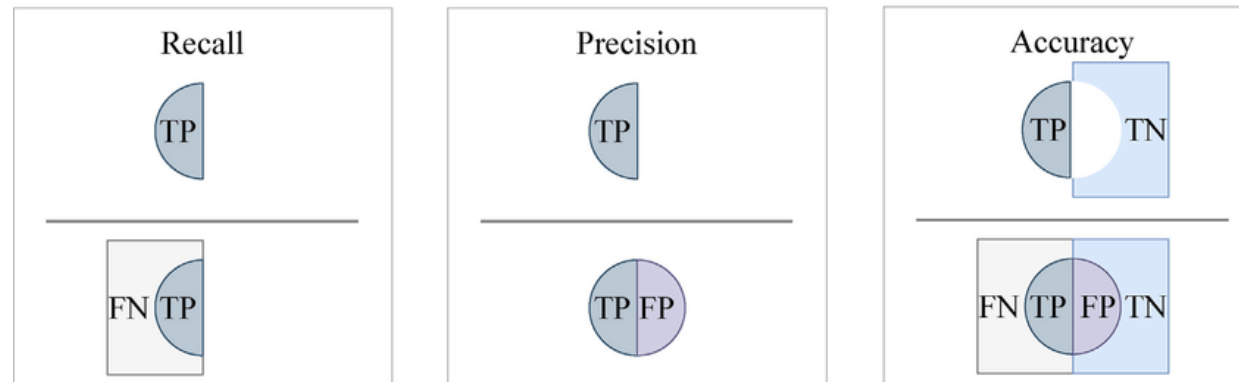
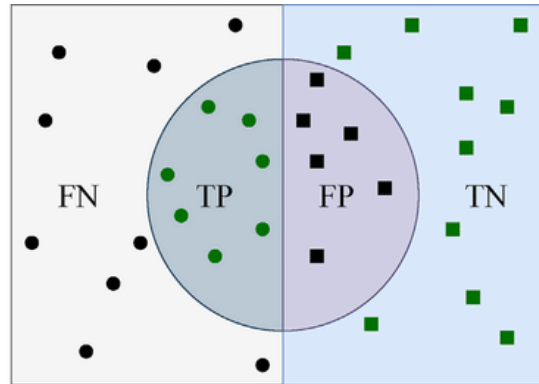
Confusion matrix

	Positive	Negative
Positive	True Positive	False Positive Type 1 Error
Negative	False Negative Type 2 Error	True Negative

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

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

Precision, recall, F1

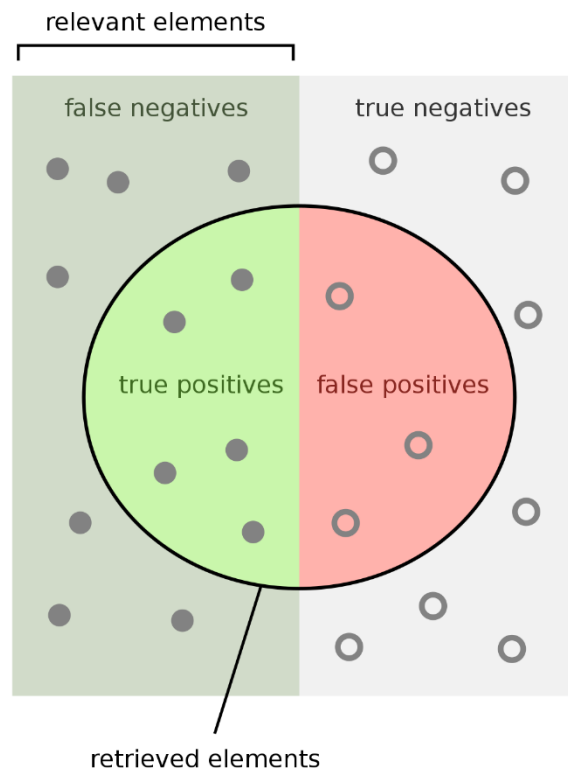


Precision, recall, F1

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} =$$
$$= 2 \frac{Recall \cdot Precision}{Recall + Precision} =$$



ШИФТ

Age Group	Don't know	Not a fan	Dislike	Like	Love
18-24	10%	10%	10%	40%	30%
25-34	10%	10%	10%	40%	30%
35-44	10%	10%	10%	40%	30%
45-54	10%	10%	40%	30%	10%
55-64	10%	10%	10%	40%	30%

FPR, TPR

TPR (true positive rate): полнота

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

FPR (false positive rate): (1-специфичность) $FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$

AUC ROC

ROC = receiver operating characteristic, «кривая ошибок»

AUC (area under the curve) – площадь под кривой ошибок

AUC ROC равен доле пар объектов вида (объект класса 1, объект класса 0), которые алгоритм верно упорядочил, т.е. первый объект идёт в упорядоченном списке раньше

$$\frac{\sum_{i=1}^q \sum_{j=1}^q I[y_i < y_j] I[a_i < a_j]}{\sum_{i=1}^q \sum_{j=1}^q I[y_i < y_j]},$$

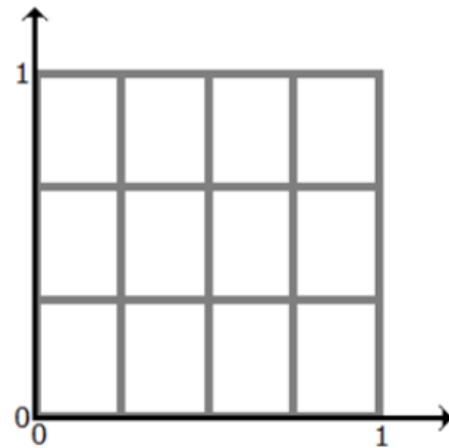
$$I[a_i < a_j] = \begin{cases} 0, & a_i > a_j, \\ 0.5 & a_i = a_j, \\ 1, & a_i < a_j, \end{cases} \quad I[y_i < y_j] = \begin{cases} 0, & y_i \geq y_j, \\ 1, & y_i < y_j, \end{cases}$$

a_i – ответ алгоритма на i -м объекте, y_i – его метка (класс), q – число объектов в тесте.

AUC ROC

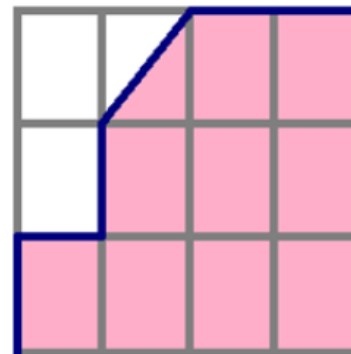
id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

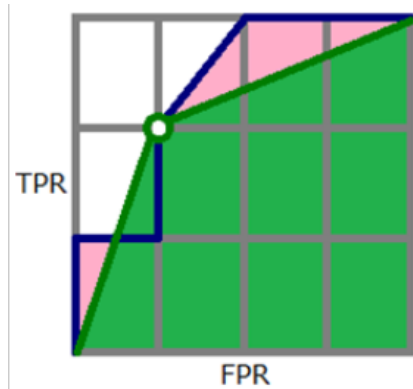


id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

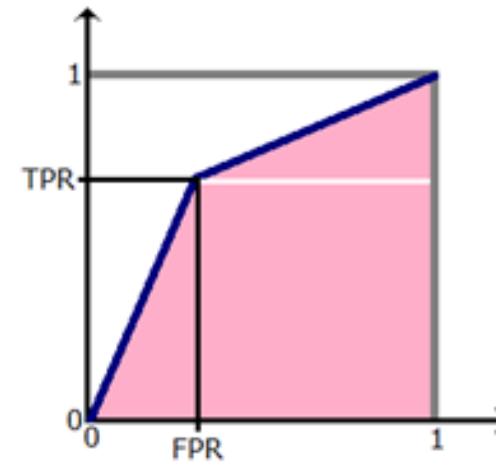
Табл. 2



AUC ROC



id	> 0.25	класс
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0



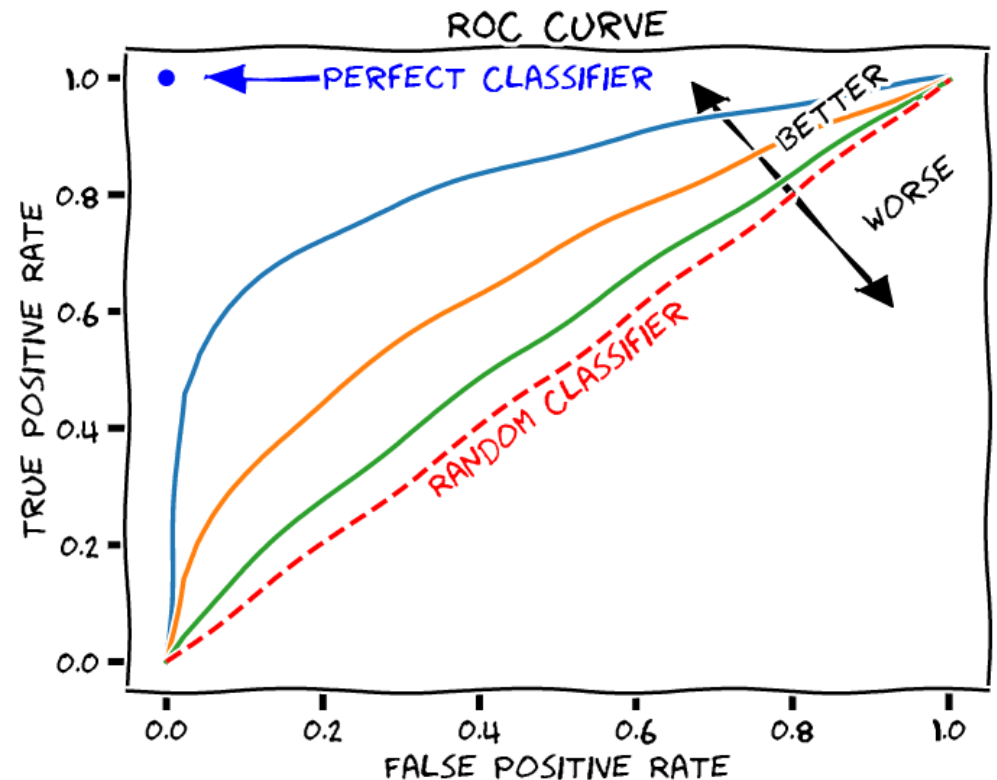
Вычисление AUC ROC в случае бинарных ответов

$$\frac{TPR \cdot FPR}{2} + TPR \cdot (1 - FPR) + \frac{(1 - TPR) \cdot (1 - FPR)}{2} = \frac{1 + TPR - FPR}{2}$$

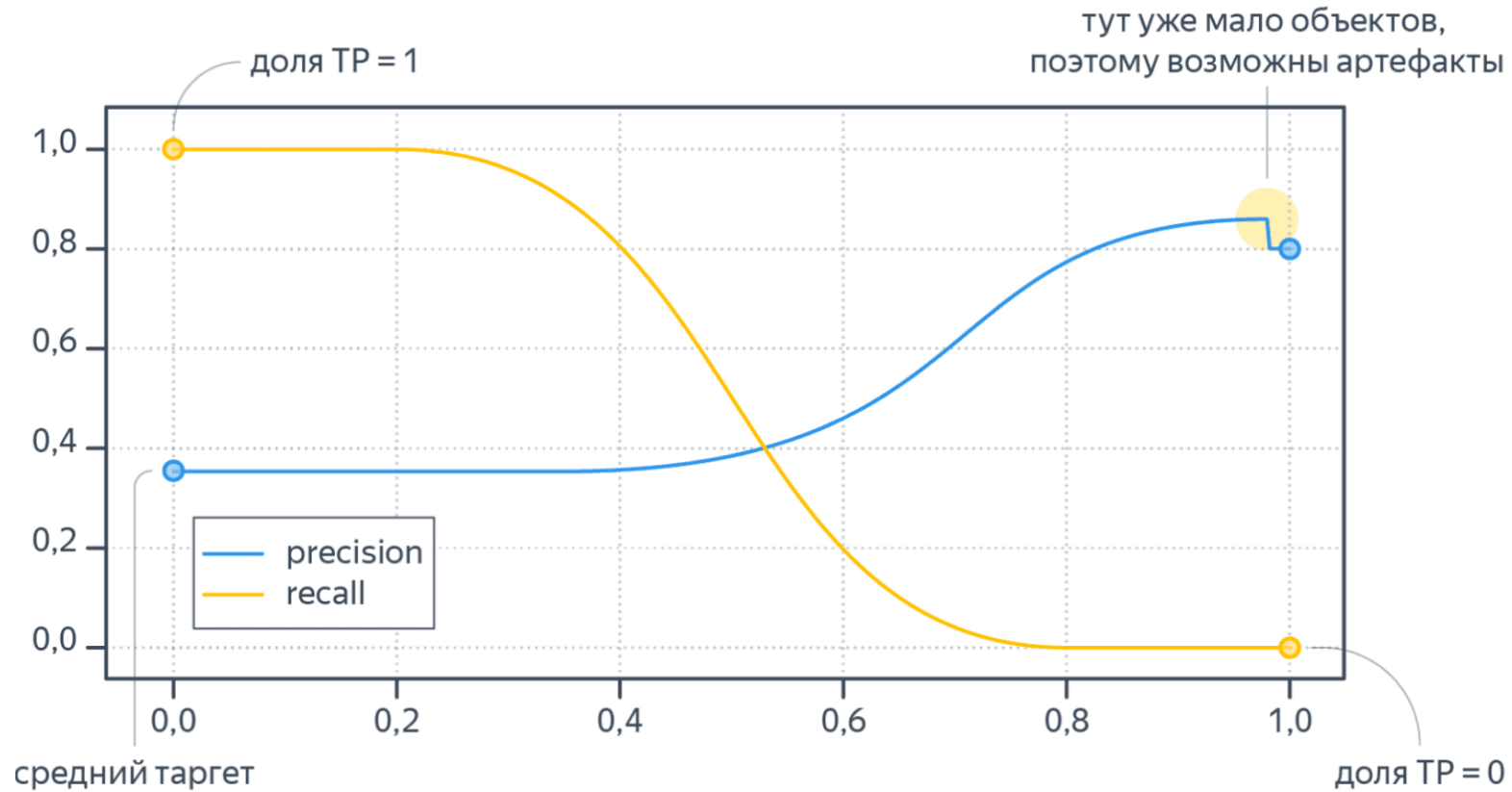
ROC curve

На ROC-кривой более высокое значение по оси X указывает на бóльшее количество ложных, чем истинно отрицательных "срабатываний". В то время как более высокое значение по оси Y указывает на большее количество истинно положительных, чем ложно отрицательных результатов

$0,8 \leq AUC \leq 1,0$ — модель работает превосходно;
 $0,6 \leq AUC < 0,8$ — модель работает хорошо;
 $0,5 < AUC < 0,6$ — модель работает удовлетворительно;
 $AUC \leq 0,5$ — модель не работает.



Оптимизация метрик классификации

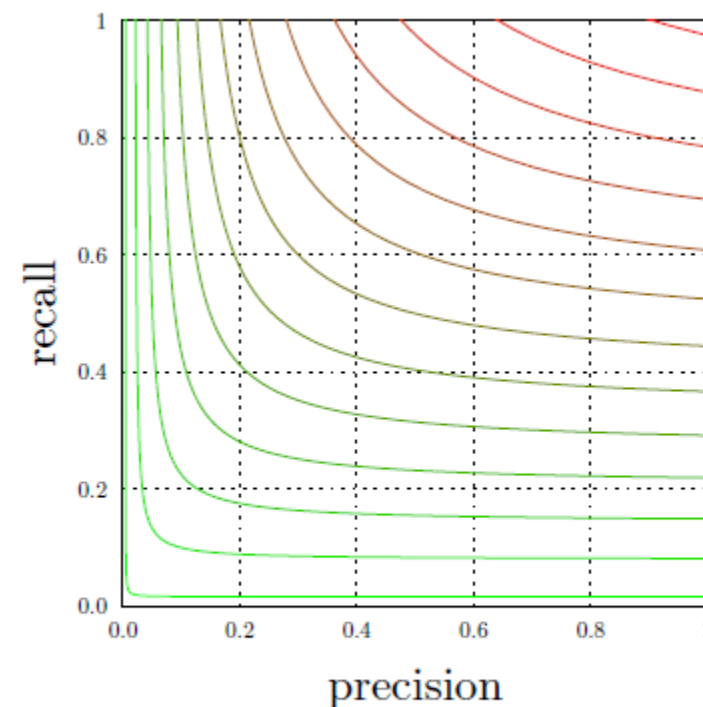
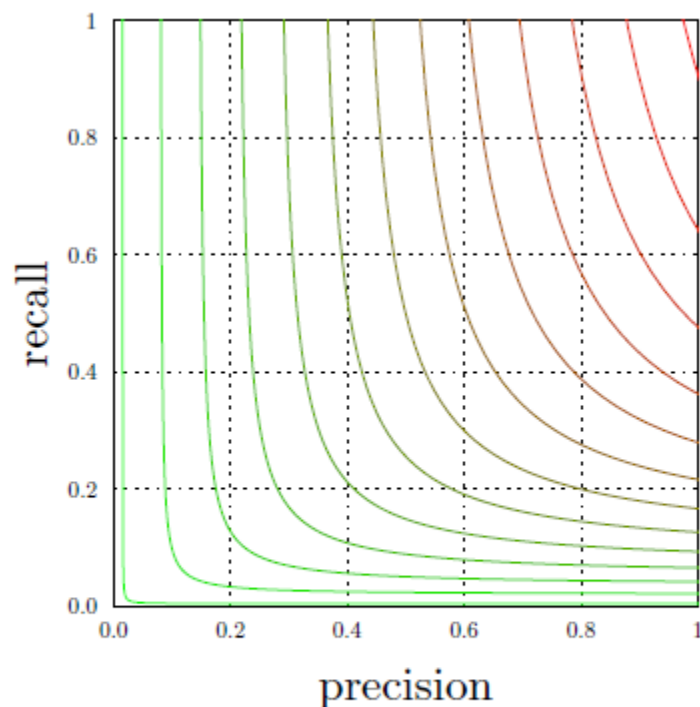
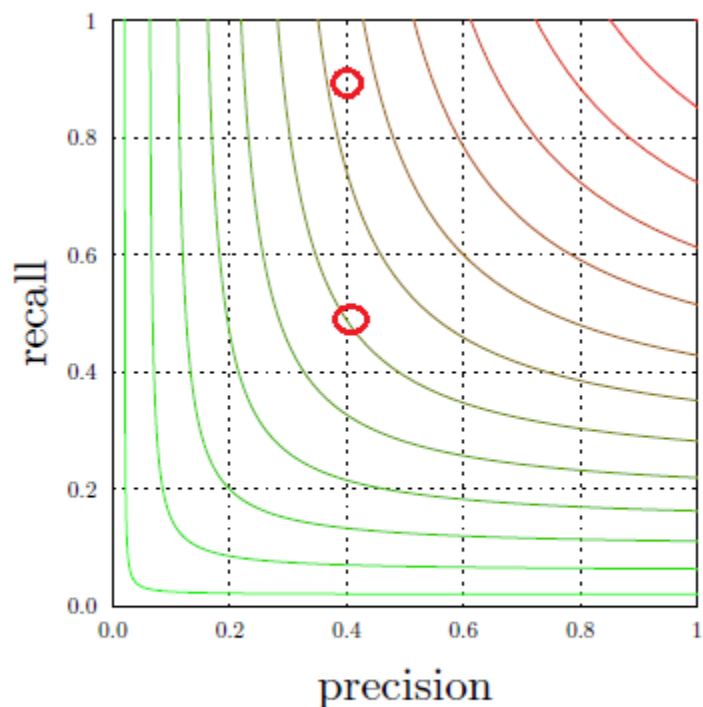


Объединение Precision и Recall

$$F = (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

$$\text{precision} = 0.4, \quad \text{recall} = 0.5 \quad \Rightarrow \quad F = 0.44,$$

$$\text{precision} = 0.4, \quad \text{recall} = 0.9 \quad \Rightarrow \quad F = 0.55.$$



Линии $F = \text{const}$ в координатах precision–recall при значениях $\beta = 1$, $\beta = 0.5$ и $\beta = 2$ соответственно

На этом пока всё!

Арсен Абдулин

email: a.abdulin@korona.net

tg: @GuitarBard

Полезные ссылки

[Хэндбуки Академии Яндекс](#)

[ODS Хабр. Метрики в задачах машинного обучения](#)

[Precision, recall. Статья на Википедии](#)

[ROC AUC. Статья Александра Дьяконова](#)

[Loginom. Статья про метрики качества.](#)

[Полезная лекция - ROC AUC по полочкам](#)