Метрики качества классификации

Пример: предсказание модели

id	Предсказанная вероятность	Правильный ответ	Предсказанный класс
1	0.6	-1	1
2	0.8	1	1
3	0.3	-1	-1
4	0.55	-1	1
5	0.1	-1	-1
6	0.96	1	1
7	0.33	1	-1
8	0.2	-1	-1
9	0.14	-1	-1
10	0.88	1	1

Accuracy

- **Accuracy** – это доля правильных ответов алгоритма

Accuracy = 0.7

id	Предсказанная вероятность	Правильный ответ	Предсказанный класс
1	0.6	-1	1
2	0.8	1	1
3	0.3	-1	-1
4	0.55	-1	1
5	0.1	-1	-1
6	0.96	1	1
7	0.33	1	-1
8	0.2	-1	-1
9	0.14	-1	-1
10	0.88	1	1

Accuracy

- 1000 объектов:

950 – не мошенники (класс 0)
50 – мошенники (класс +1)

- Модель: a(x) = 0

Accuracy?

Accuracy

- 1000 объектов:

950 – не мошенники (класс 0)
50 – мошенники (класс +1)

- Модель: a(x) = 0

Accuracy = 0.95

- *Если классы несбалансированы, то accuracy не надо использовать!*
- *Метрика не показывает какие классы между собой путаем*

Матрица ошибок

Predicted Value		Actual Value: positives	Actual Value: negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

Пример: кредитный скоринг

Модель 1: одобряет 100 кредитов

- 80 кредитов вернули
- 20 кредитов не вернули

Модель 2: одобряет 50 кредитов

- 48 кредитов вернули
- 2 кредита не вернули

На тестовой выборке, где 100 вернули, 100 не вернули

Какая модель лучше?

Точность (precision)

Точность показывает, насколько можно доверять классификатору в случае если он выдает положительный класс $a(x) = +1$

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

		Actual Value	
		positives	negatives
Predicted Value	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

Точность (precision)

Точность показывает, насколько можно доверять классификатору в случае если он выдает положительный класс $a(x) = +1$

$a_1(x) : precision = 0.8$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ Не получили кредит	20	80

$a_2(x) : precision = 0.96$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	48	2
$\mathbf{a(x) = -1}$ Не получили кредит	52	98

Полнота (recall)

Полнота показывает как много объектов положительного класса нашел классификатор

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

		Actual Value	
		positives	negatives
Predicted Value	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

Полнота (recall)

Полнота показывает как много объектов положительного класса нашел классификатор

$a_1(x) : recall = 0.8$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ Не получили кредит	20	80

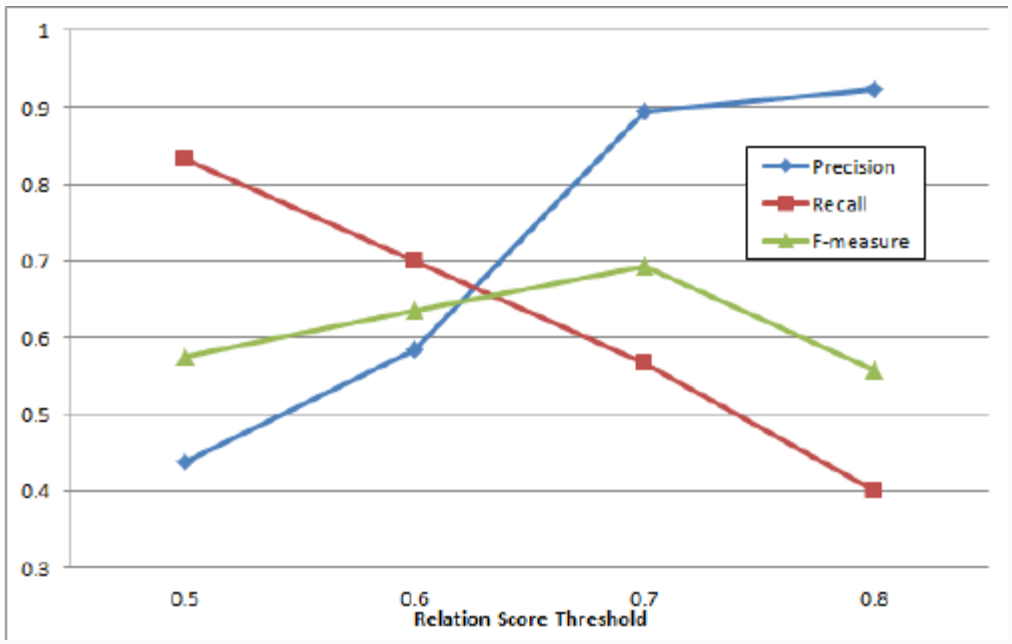
$a_2(x) : recall = 0.48$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	48	2
$\mathbf{a(x) = -1}$ Не получили кредит	52	98

F-мера

F-мера (F1-score) - среднее гармоническое точности и полноты

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$



Регулируем точность и полноту

Пусть $p(x)$ - уверенность классификатора в том, что объект x относится к классу +1, p(x) лежит на отрезке [0;1].

Обычно

- если p(x) > 0.5, то мы относим объект к положительному классу
- а иначе - к отрицательному

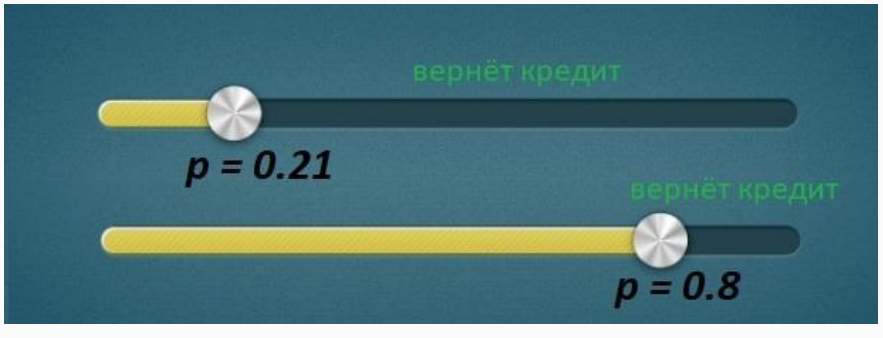
Регулируем точность и полноту

Пусть $p(x)$ - уверенность классификатора в том, что объект x относится к классу +1, p(x) лежит на отрезке [0;1].

Обычно

- если p(x) > 0.5, то мы относим объект к положительному классу
- а иначе - к отрицательному

Можно изменять этот порог, то есть вместо 0.5 брать другое число из отрезка [0;1].



Интегральные метрики

Хотим измерить качество всего семейства классификаторов независимо от выбранного порога.

Для этого будем использовать метрику AUC

AUC *– Area Under ROC Curve (площадь под ROC-кривой)*

ROC-AUC: интуиция

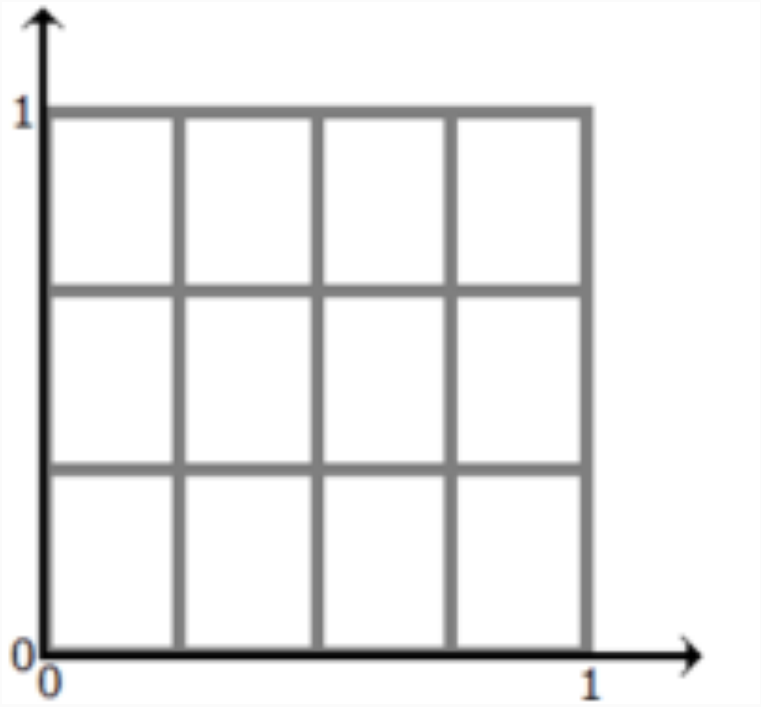
- Пример:

p	класс
0.5	0
0.1	0
0.25	0
0.6	1
0.2	1
0.3	1
0.0	0

p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0

ROC-AUC: алгоритм

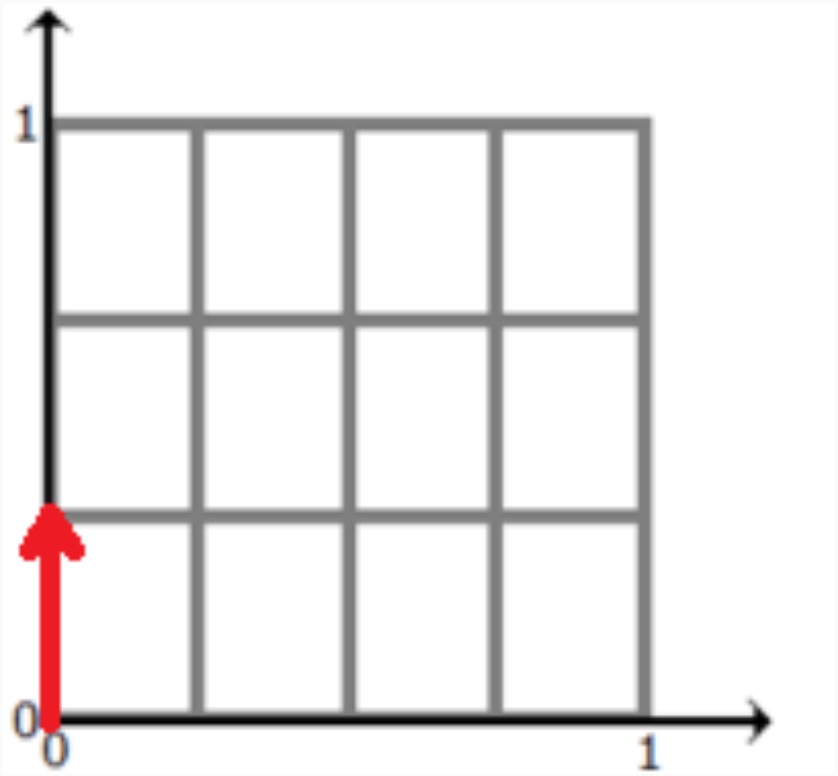
- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1



ROC-AUC: алгоритм

- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1

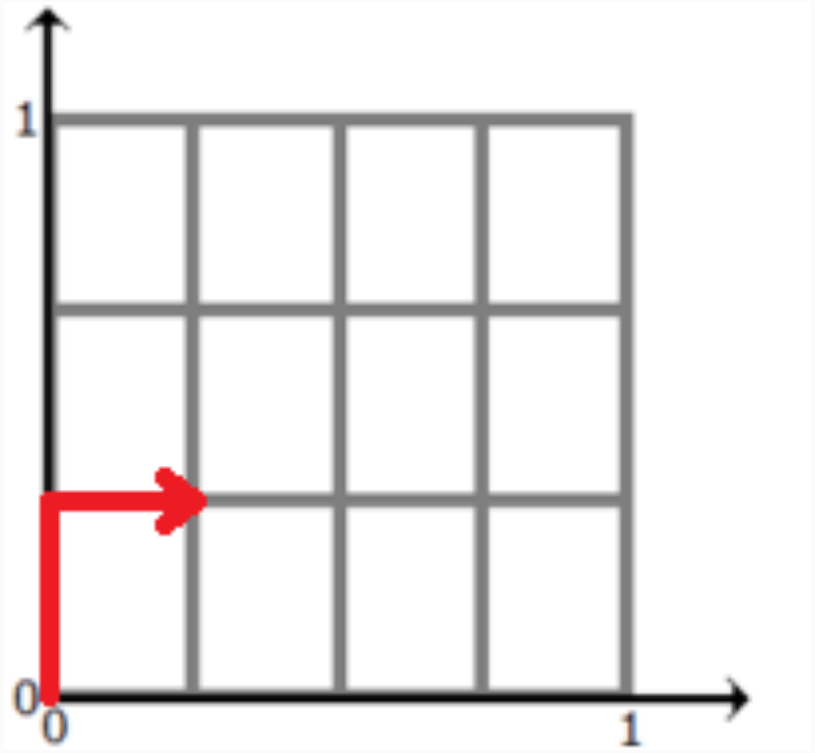
p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



ROC-AUC: алгоритм

- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1

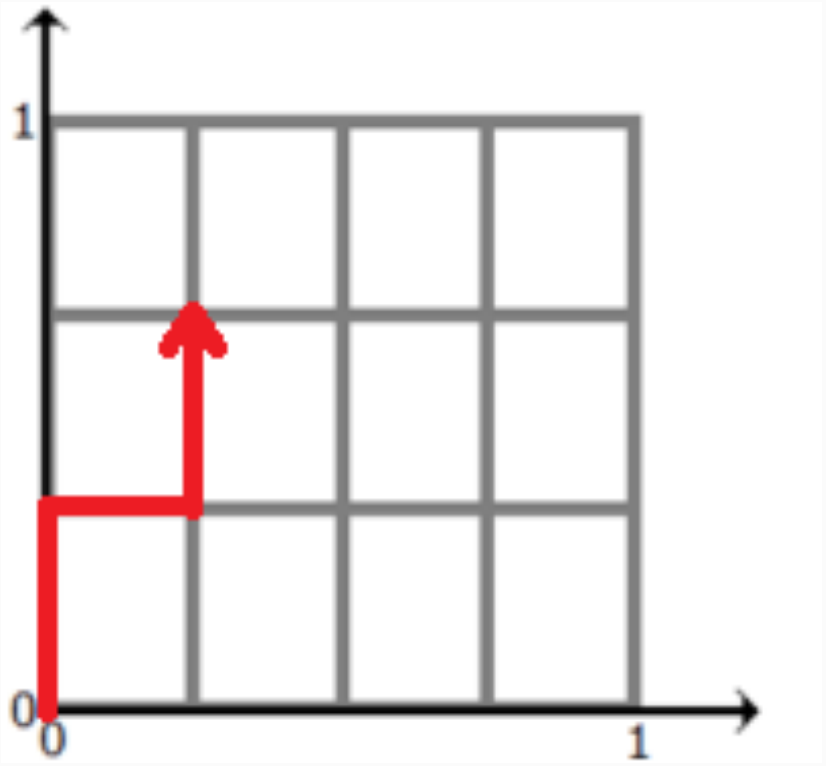
p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



ROC-AUC: алгоритм

- Нарисуем квадрат 1 на 1.
- Горизонтальную сторону квадрата разобъем на равные отрезки, число которых равно числу 0 в данных
- Вертикальную сторону разобъем на равные отрезки, число которых равно числу 1

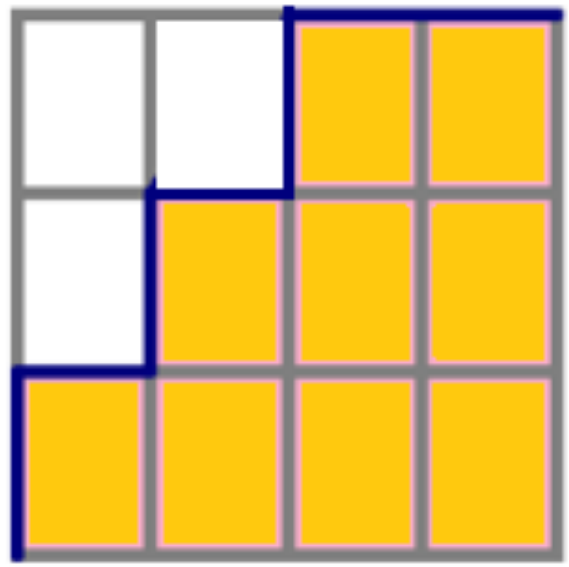
p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



ROC-AUC: алгоритм

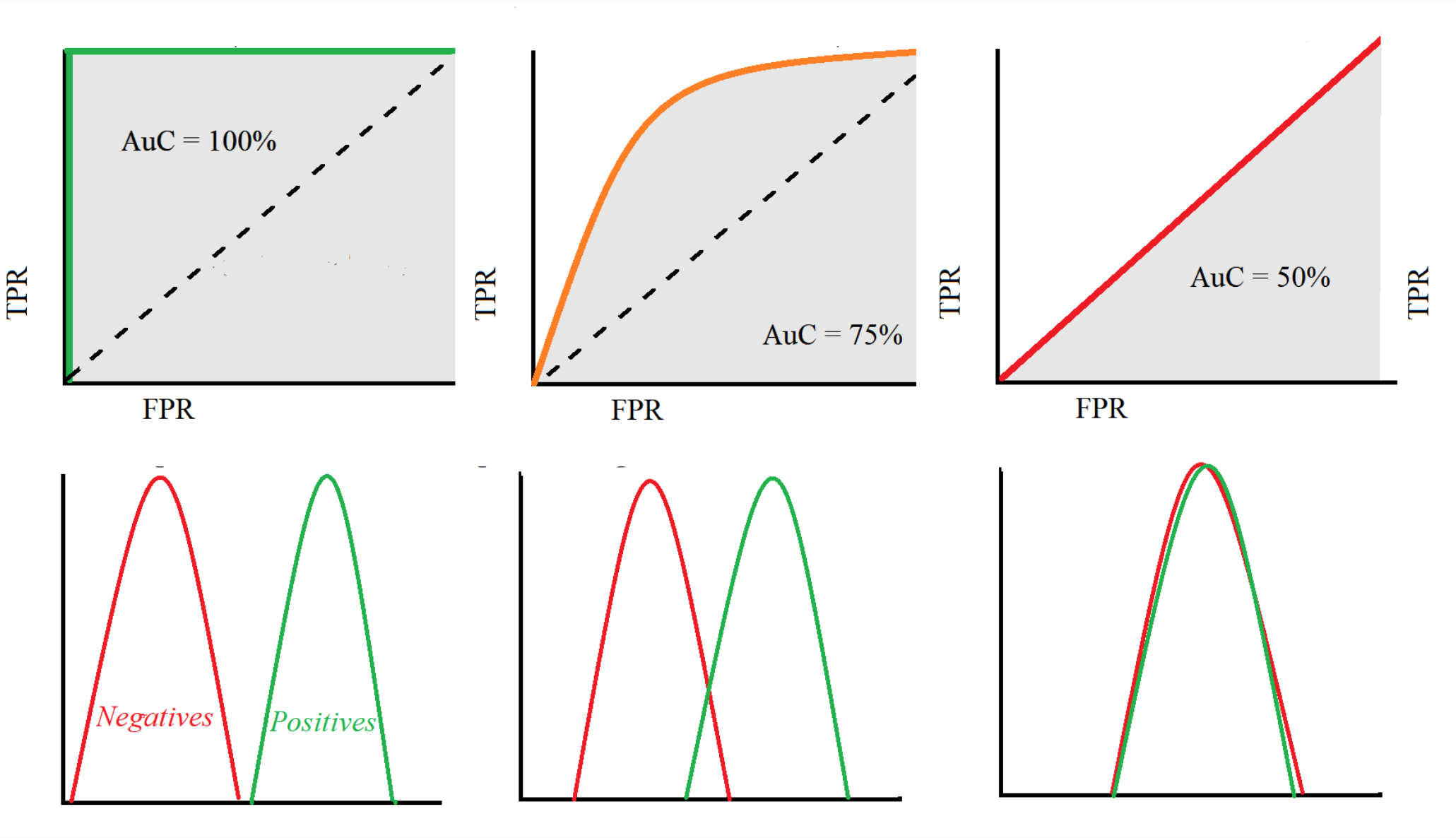
- Пойдем по отсортированной таблице по столбцу класс сверху вниз
- Будем стартовать из точки (0,0) на квадрате. И если мы встречаем 1, сдвигаемся на одну клеточку вверх, а если 0 - то вправо
- В итоге мы придём в точку (1,1).

p	класс
0.6	1
0.5	0
0.3	1
0.25	0
0.2	1
0.1	0
0.0	0



Полученная кривая называется ROC-кривой, а метрика, равная площади под ней - AUC-ROC.

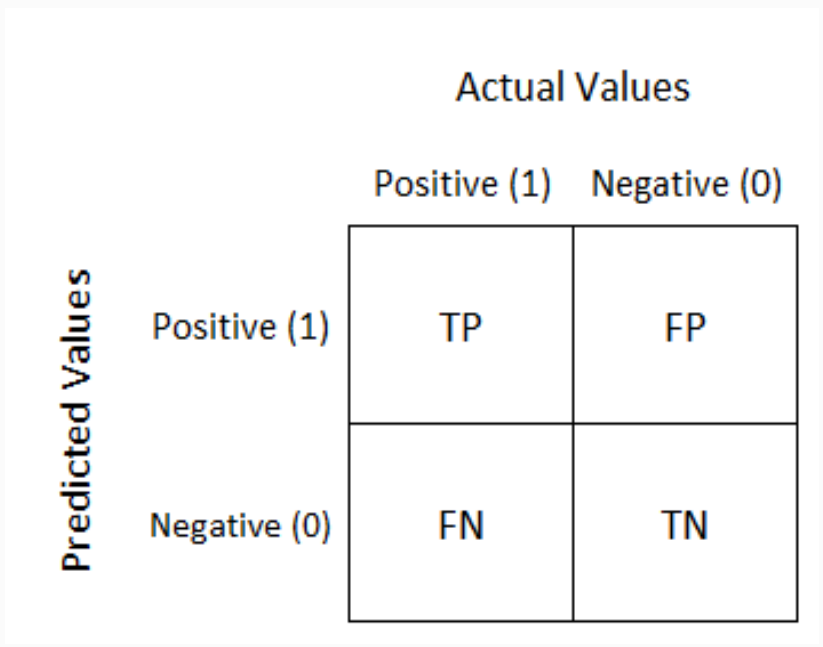
ROC-AUC: примеры



ROC-AUC: формализация

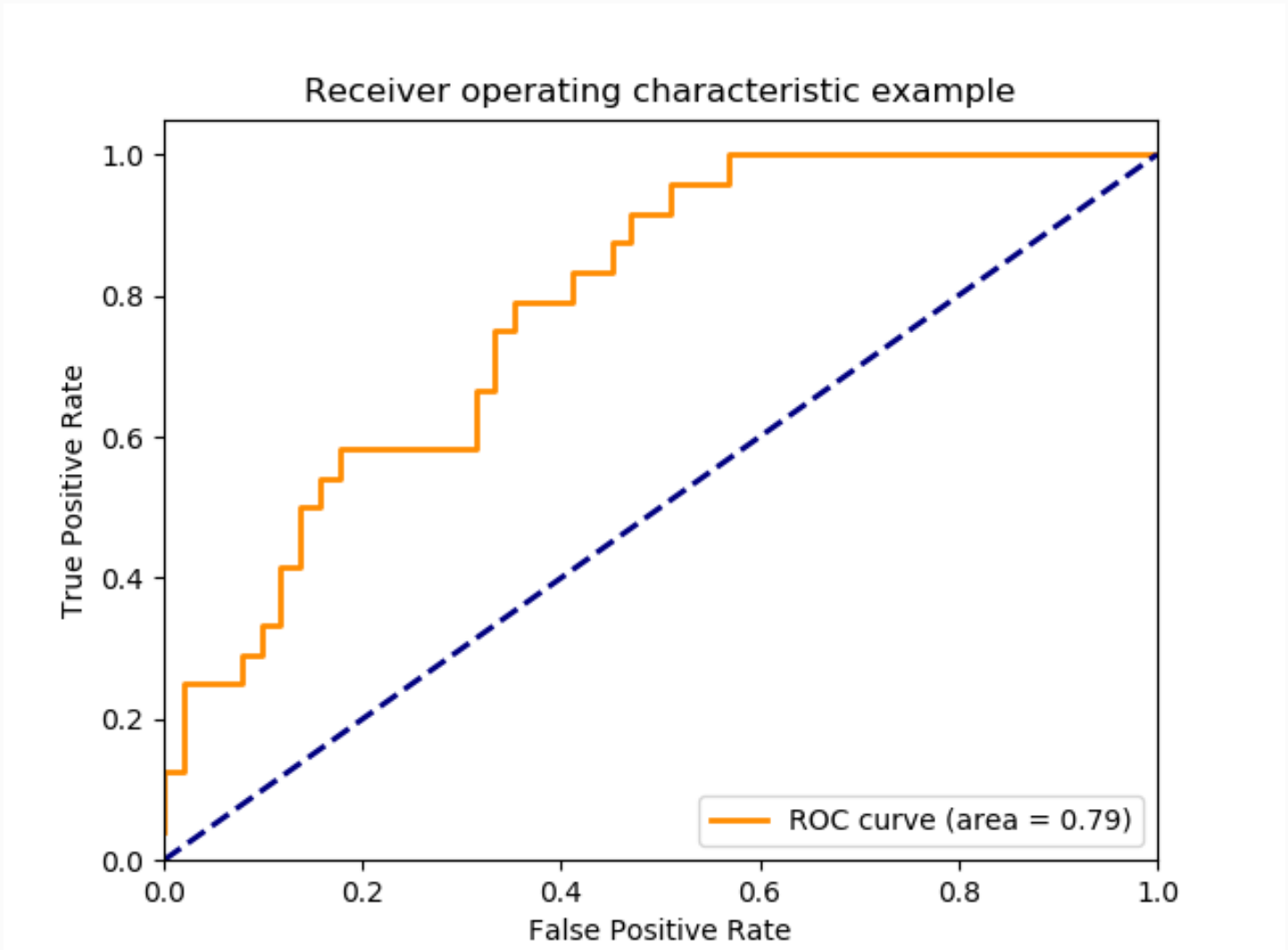
Для каждого порога посчитаем False Positive Rate (FPR) и True Positive Rate (TPR)

- $FPR = \frac{FP}{FP + TN}$ - доля неверно принятых объектов отрицательного класса
- $TPR = \frac{TP}{TP + FN}$ - доля верно принятых объектов положительного класса



ROC-AUC: формализация

Кривая, состоящая из точек с координатами (FPR, TPR) для всех возможных порогов – это и есть ROC-кривая.

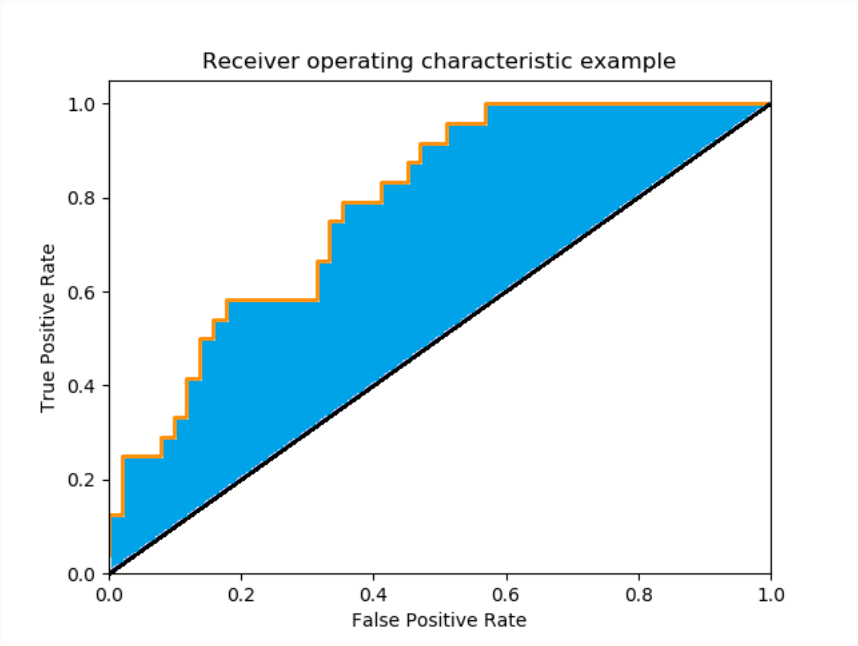


Индекс Джини

Индекс Джини:

$$Gini = 2 \cdot AUC - 1$$

- Индекс Джини - это удвоенная площадь между главной диагональю и ROC-кривой



ROC-AUC при дисбаланс классов

Рассмотрим следующую задачу: нам необходимо выбрать 100 релевантных документов из 1 миллиона документов.

- **Алгоритм 1** возвращает 100 документов, 90 из которых релевантны. Таким образом,

$$TPR = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{10}{10 + 999890} = 0.00001$$

- **Алгоритм 2** возвращает 2000 документов, 90 из которых релевантны. Таким образом,

$$TPR = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{1910}{1910 + 997990} = 0.00191$$

ROC-AUC при дисбаланс классов

- Разница в False Positive Rate между этими двумя алгоритмами *крайне* мала — всего 0.0019
- Это следствие того, что AUC-ROC измеряет долю False Positive относительно True Negative и в задачах, где нам не так важен второй (больший) класс, может давать не совсем адекватную картину при сравнении алгоритмов.

- **Алгоритм 1** возвращает 100 документов, 90 из которых релевантны. Таким образом,

$$TPR = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{10}{10 + 999890} = 0.00001$$

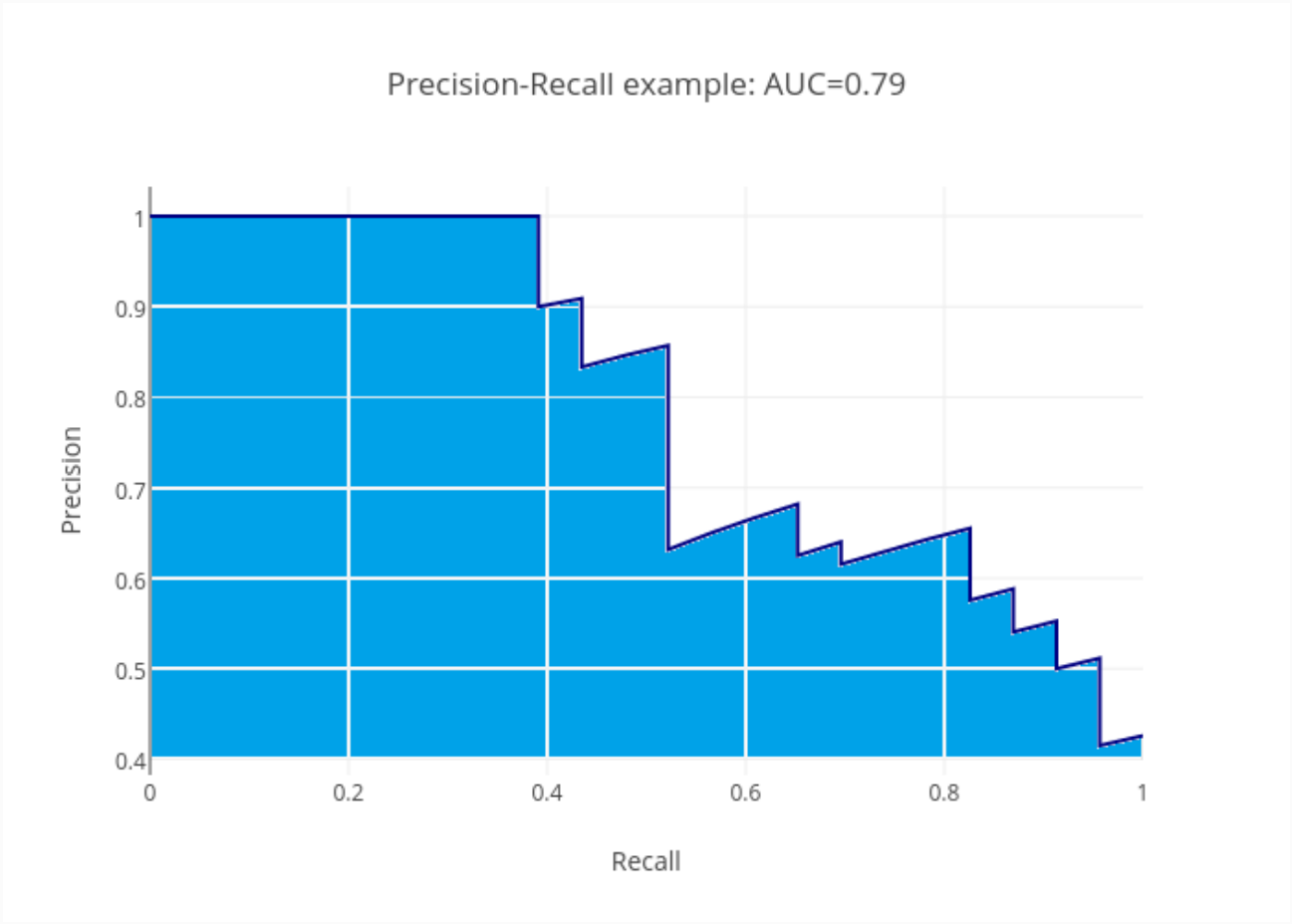
- **Алгоритм 2** возвращает 2000 документов, 90 из которых релевантны. Таким образом,

$$TPR = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{1910}{1910 + 997990} = 0.00191$$

PR-AUC

- В случае малой доли объектов положительного класса AUC-ROC может давать неадекватно хороший результат
- В таких задачах можно использовать Average Precision (Precision-Recall кривую)



PR-AUC при дисбаланс классов

Рассмотрим следующую задачу: нам необходимо выбрать 100 релевантных документов из 1 миллиона документов.

- **Алгоритм 1**

$$precision = \frac{TP}{TP + FP} = 90/(90 + 10) = 0.9$$

$$recall = \frac{TP}{TP + FN} = 90/(90 + 10) = 0.9$$

- **Алгоритм 2**

$$precision = \frac{TP}{TP + FP} = \frac{90}{90 + 1910} = 0.045$$

$$recall = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.9$$

Пример ROC-AUC и PR-AUC при сильном дисбаланс

Таблица 1: AUC-ROC

	annthyroid	cardio	mammography	PageBlocks	vowels	Wilt	yeast
IForest	**0.826**	0.944	0.850	0.890	0.781	0.430	0.430
OCSVM	0.606	0.940	0.855	0.893	0.533	0.301	0.448
CBLOF	0.674	0.851	0.848	0.895	0.895	0.328	0.477
COF	0.705	0.545	0.792	0.673	0.850	0.544	0.429
COPOD	0.796	0.928	0.899	0.873	0.501	0.331	0.406
ECOD	0.804	0.943	**0.907**	**0.912**	0.564	0.341	0.477
HBOS	0.692	0.865	0.872	0.789	0.646	0.281	0.410
KNN	0.730	0.742	0.860	0.770	**0.972**	0.472	0.414
LODA	0.306	0.893	0.815	0.753	0.656	0.408	**0.493**
LOF	0.706	0.628	0.765	0.702	0.953	0.474	0.472
PCA	0.693	**0.961**	0.894	0.894	0.566	0.173	0.444
SOD	0.791	0.673	0.810	0.755	0.843	**0.583**	0.470

Таблица 2: AUC-PR

	annthyroid	cardio	mammography	PageBlocks	vowels	Wilt	yeast
IForest	**0.353**	0.616	0.189	0.469	0.390	0.043	0.333
OCSVM	0.131	0.573	0.116	0.497	0.114	0.035	0.323
CBLOF	0.236	0.487	0.111	**0.567**	0.223	0.036	0.328
COF	0.183	0.126	0.117	0.338	0.489	0.055	0.305
COPOD	0.193	0.604	0.404	0.384	0.049	0.037	0.327
ECOD	0.298	0.593	**0.415**	0.537	0.162	0.038	0.365
HBOS	0.229	0.517	0.177	0.334	0.195	0.037	0.344
KNN	0.226	0.330	0.173	0.380	**0.702**	0.046	0.321
LODA	0.054	0.444	0.144	0.408	0.099	0.041	**0.368**
LOF	0.194	0.145	0.118	0.304	0.388	0.047	0.330
PCA	0.207	**0.652**	0.206	0.483	0.127	0.030	0.316
SOD	0.222	0.216	0.149	0.342	0.201	**0.062**	0.322

Связь с бизнесом

Связь с бизнесом

Показатели бизнеса

Например:

- Lifetime value
- Прибыль
- Расходы
- Доля аудитории
- Цена акций

Мы хотели бы смотреть, как модель влияет на них, но не можем

Измеряются месяцами

Связаны с показателями бизнеса
Можно сделать быстрый тест

Например:

- Конверсия в клик
- Оценка сервиса
- Средний чек
- MAU, DAU, WAU

Мы можем оценить эти метрики, проведя A/B-тест

Измеряются неделями

Являются приближением онлайн-метрик

Считаются на исторических данных

Например:

- Precision, recall
- Accuracy

Считаются минуты-часы

Можем почти бесплатно проверить наши модели

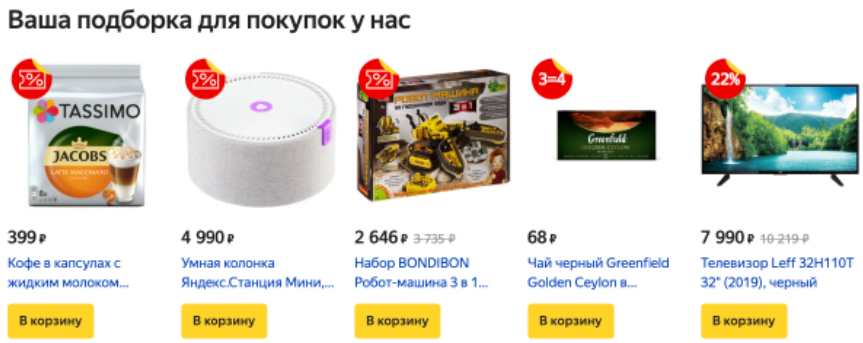
Свойства метрик

- Чувствительность
- Шум
- Интерпретация
- Иерархия

Пример

Иерархия метрик для задачи построения рекомендаций

- Хотим внедрить новое ML-ранжирование рекомендаций товаров
- Находимся в ситуации, когда этот элемент уже есть на сайте



Что измеряем?

Иерархия метрик

Бизнес-метрика

- Выручка

Онлайн-метрики

- Средний чек / Число купивших пользователей
- Выручка проданных товаров, через наш элемент
- CTR элемента

Офлайн-метрики

- Оффлайн метрики ранжирования
- Accuracy на валидационной выборке

Правильность иерархии метрик

- Необходимо проверять правильность иерархии метрик, чтобы по более чувствительным (менее важным для бизнеса) метрикам аппроксимировать более шумные (более важные для бизнеса)