



Связь бизнес-показателей и DS-метрик

2021



**Бизнес : измеряет деньги, время ,
товар**

**ML : измеряет процент и долю
верных решений**

Как измерить успех/неудачу?

В ЭТОМ УРОКЕ

- 1. train/test/val и cross-validation**
- 2. Переход от вероятностей к меткам в задаче бинарной классификации**
- 3. Какие бывают DS-метрики, их интерпретация**
- 4. Пример из практики бизнес-метрики**
- 5. Взаимосвязь DS и бизнес метрик**
- 6. Case1 - фильтрация токсичных комментариев**
- 7. Case2 - определение CC3**

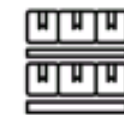
Метрики бизнеса



Результативность
(англ. Effectiveness)



Эффективность
(англ. Efficiency)



Емкость
(англ. Capacity)



Утилизация ресурсов
(англ. Resource utilization)



Качество
(англ. Quality)



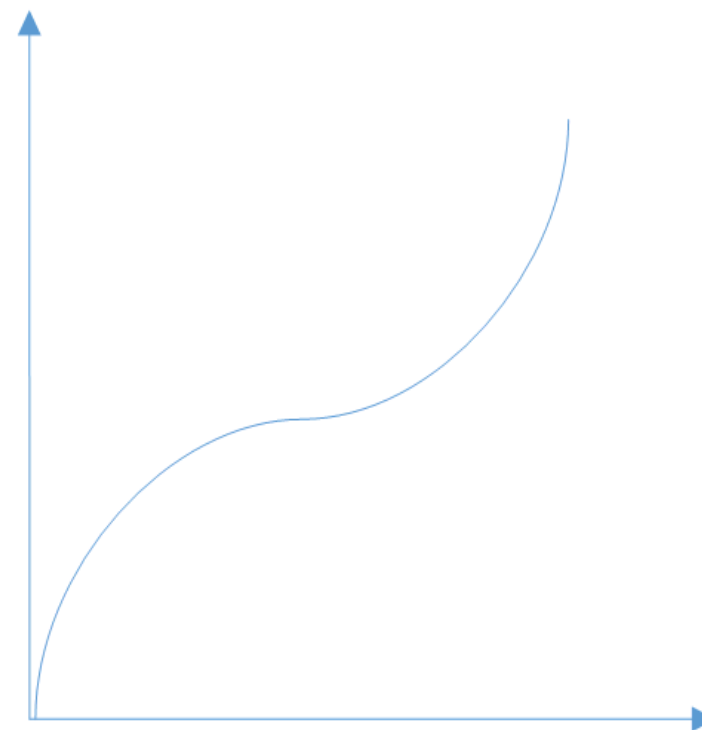
Ценность для клиента
(англ. Customer Experience)



Метрики бизнеса



Количество
людей



деньги



Метрики ML



→ **MCE**

→ **Precision**

→ **Recall**

→ **F-score**

→ **ROC-AUC**

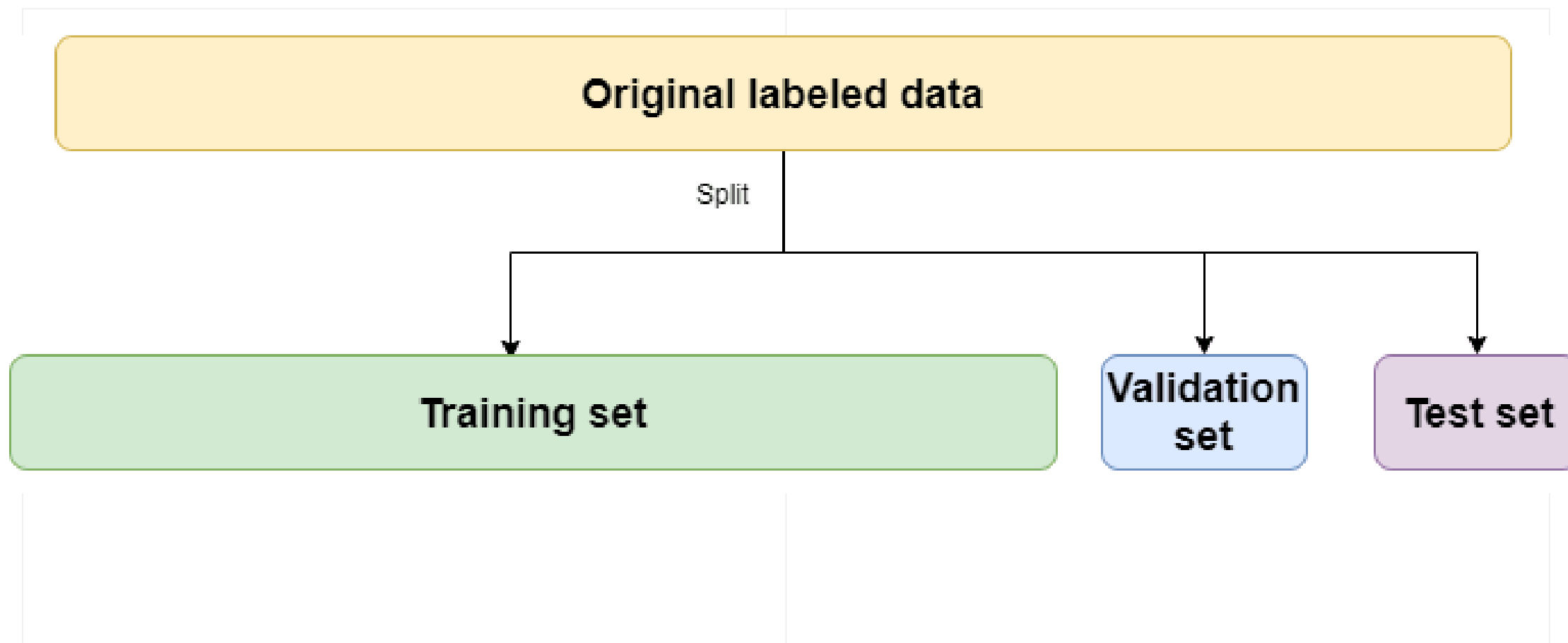
→ **MSE**

→ **MAE**

→ **MAPE**

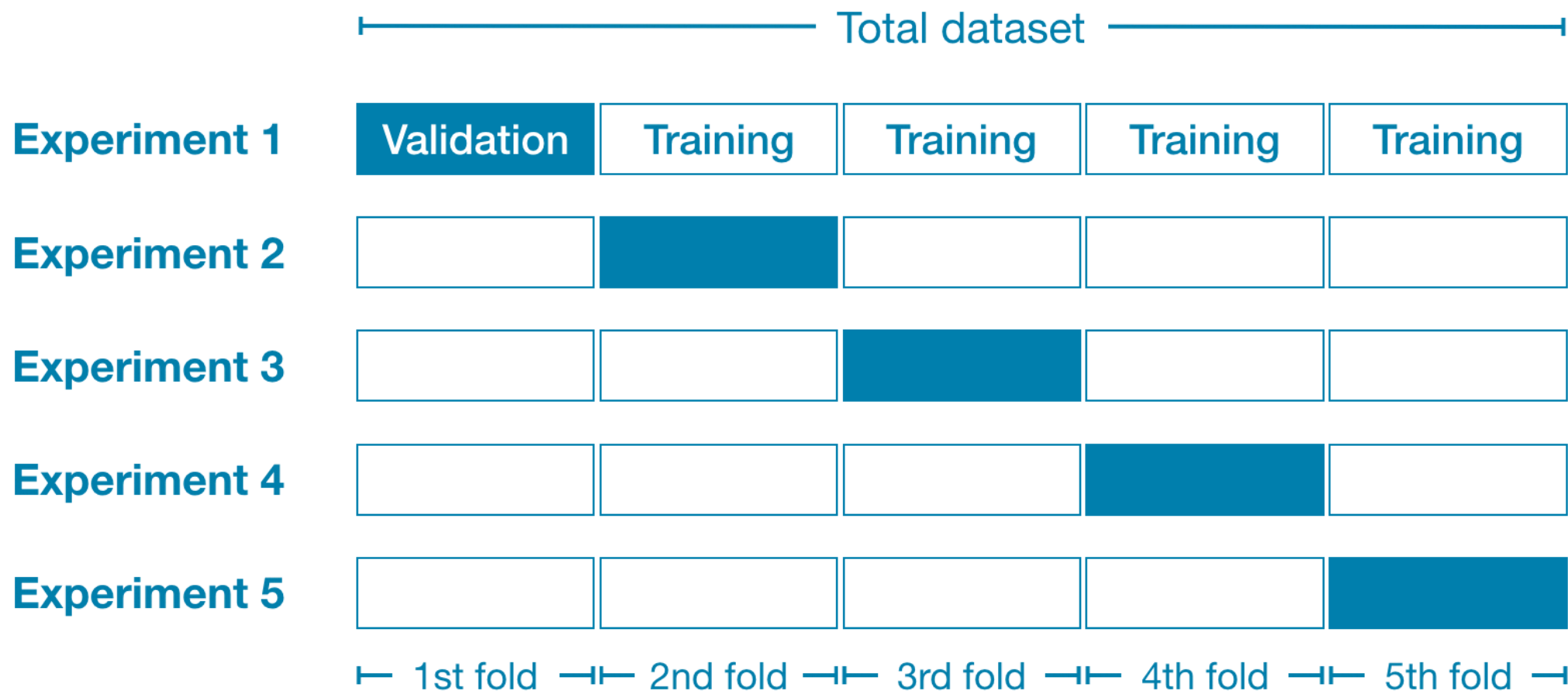


train/test/val



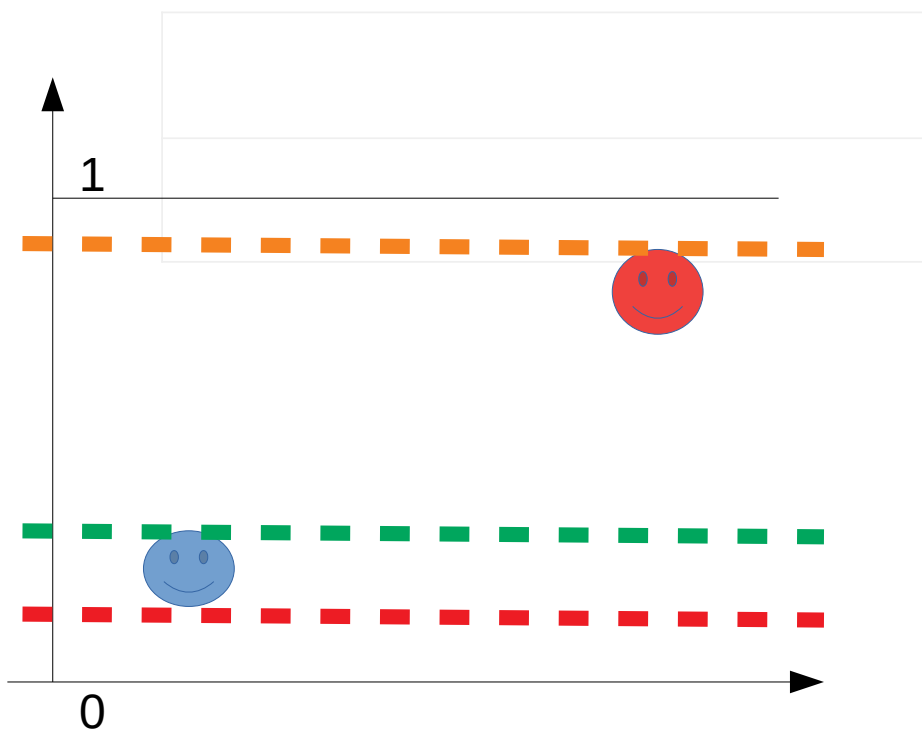


cross-validation



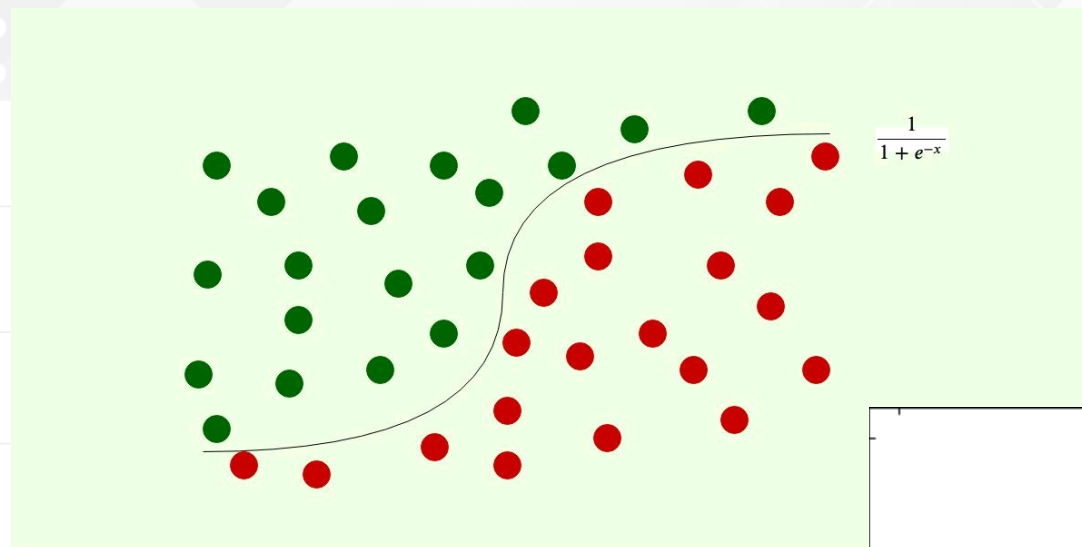
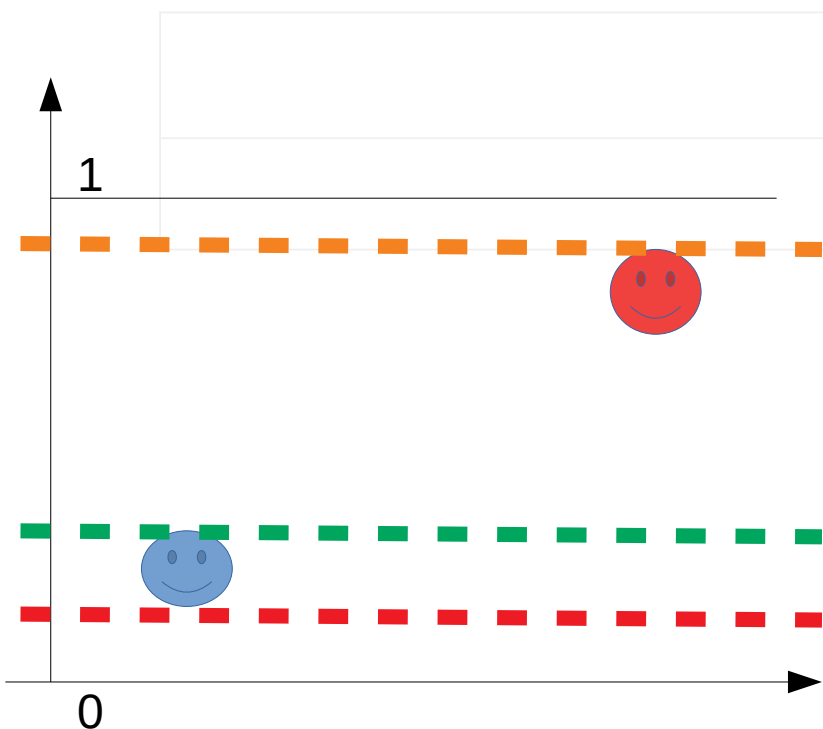


Задача бинарной классификации

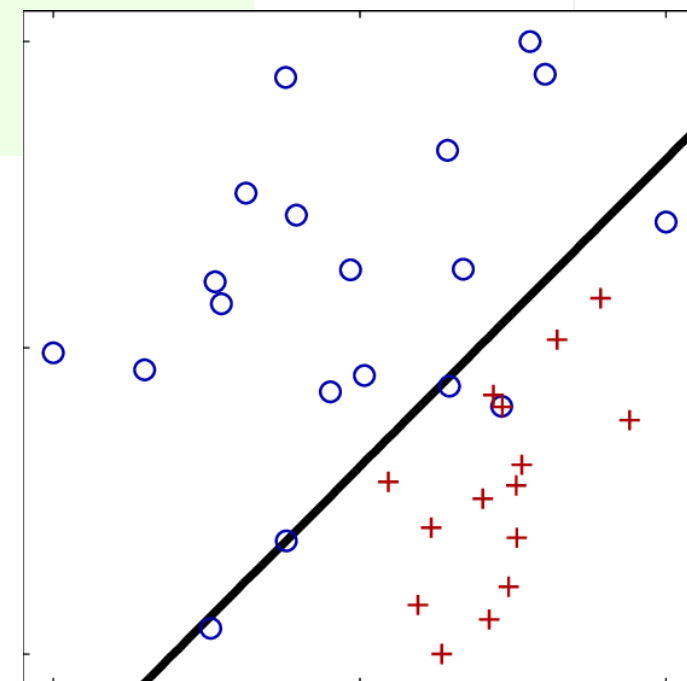




Задача бинарной классификации



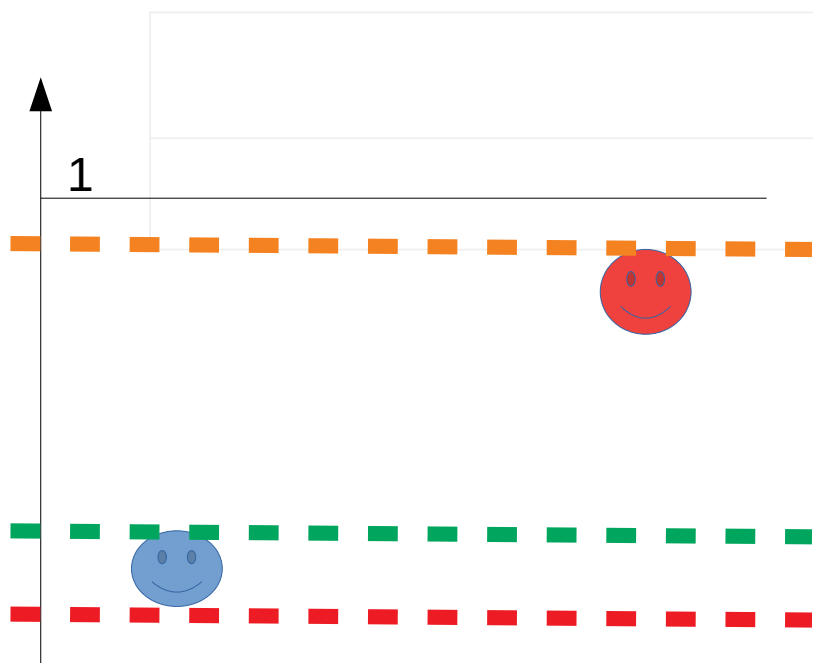
$$\text{MCE} = \frac{1}{m} \sum_{i=1}^m I[a_i \neq y_i]$$



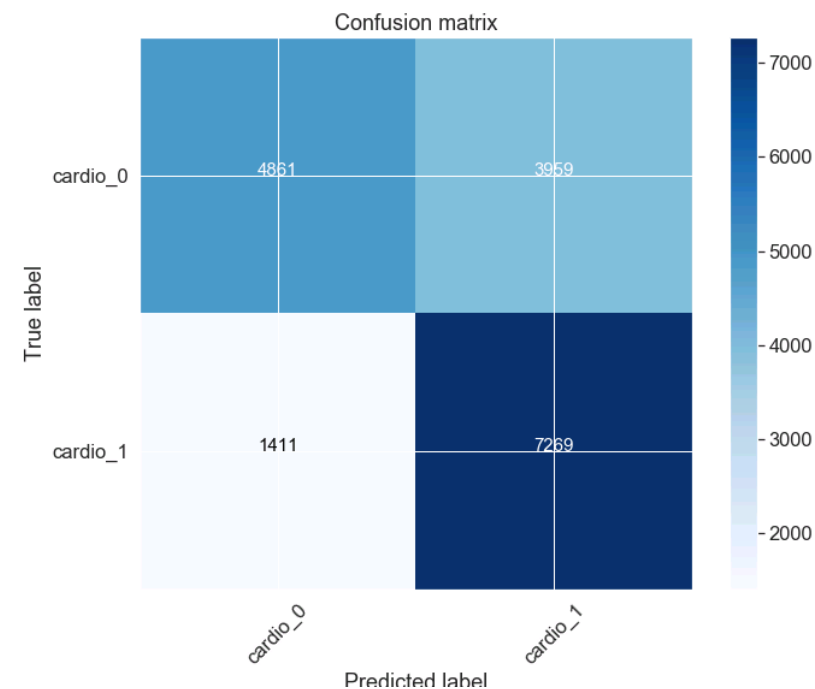


Задача бинарной классификации

<https://habr.com/ru/company/ods/blog/328372/>

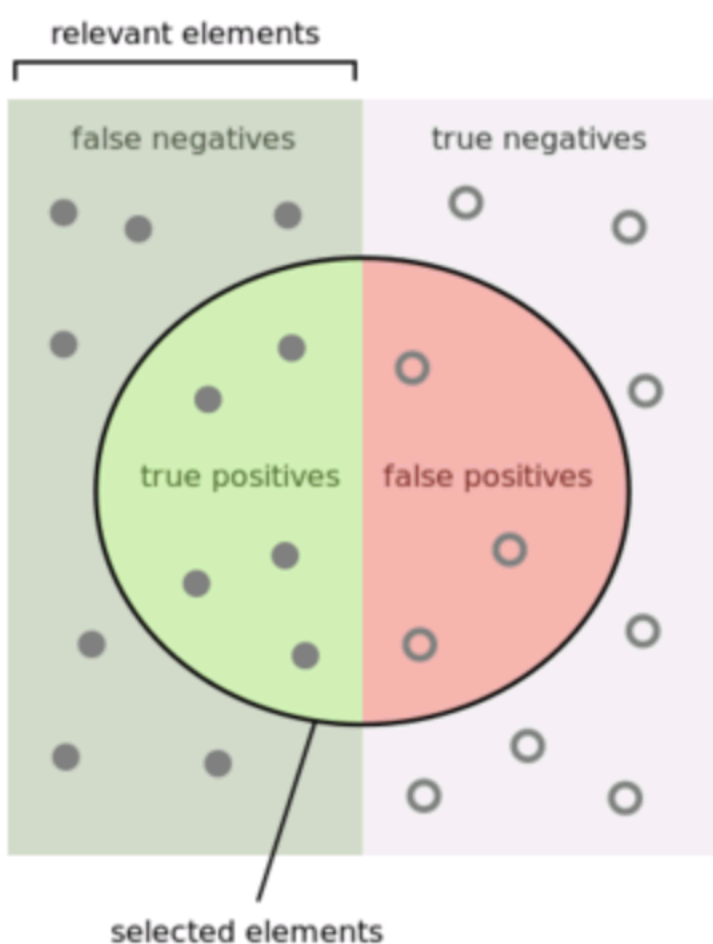


	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)



$$\text{MCE} = \frac{1}{m} \sum_{i=1}^m I[a_i = y_i]$$


Метрики классификации



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

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

How many selected items are relevant?

Precision = 

How many relevant items are selected?

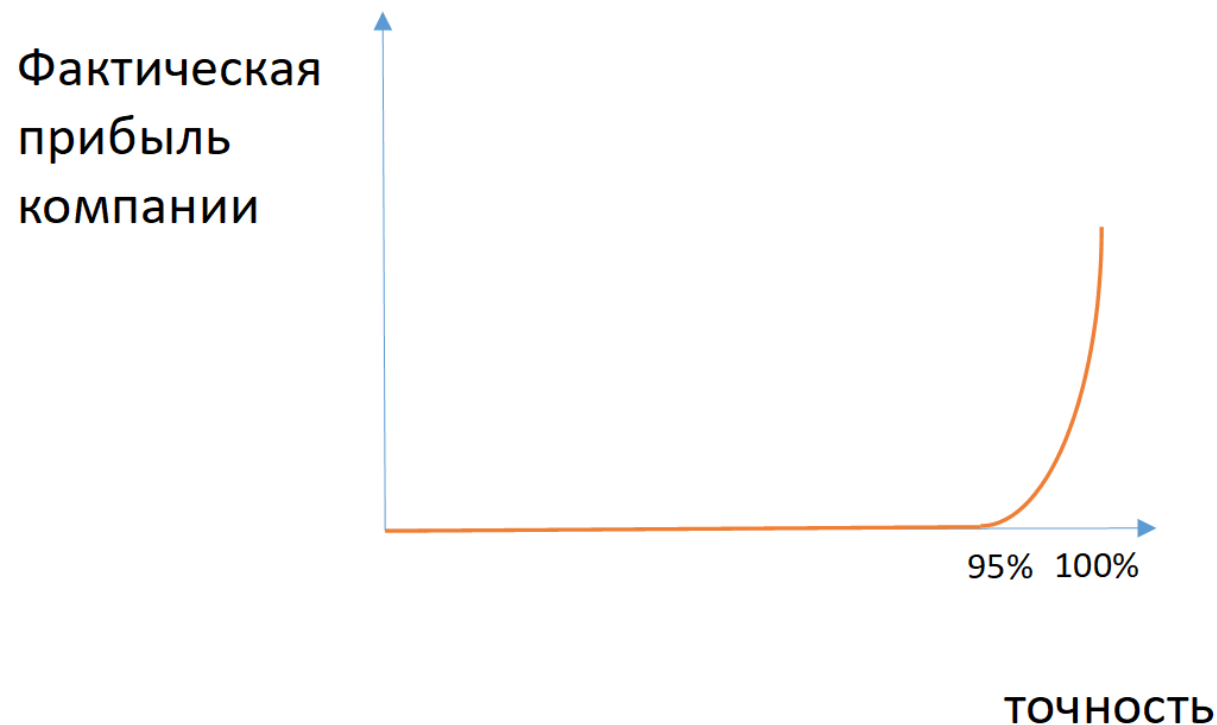
Recall = 

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



Метрики классификации

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



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

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

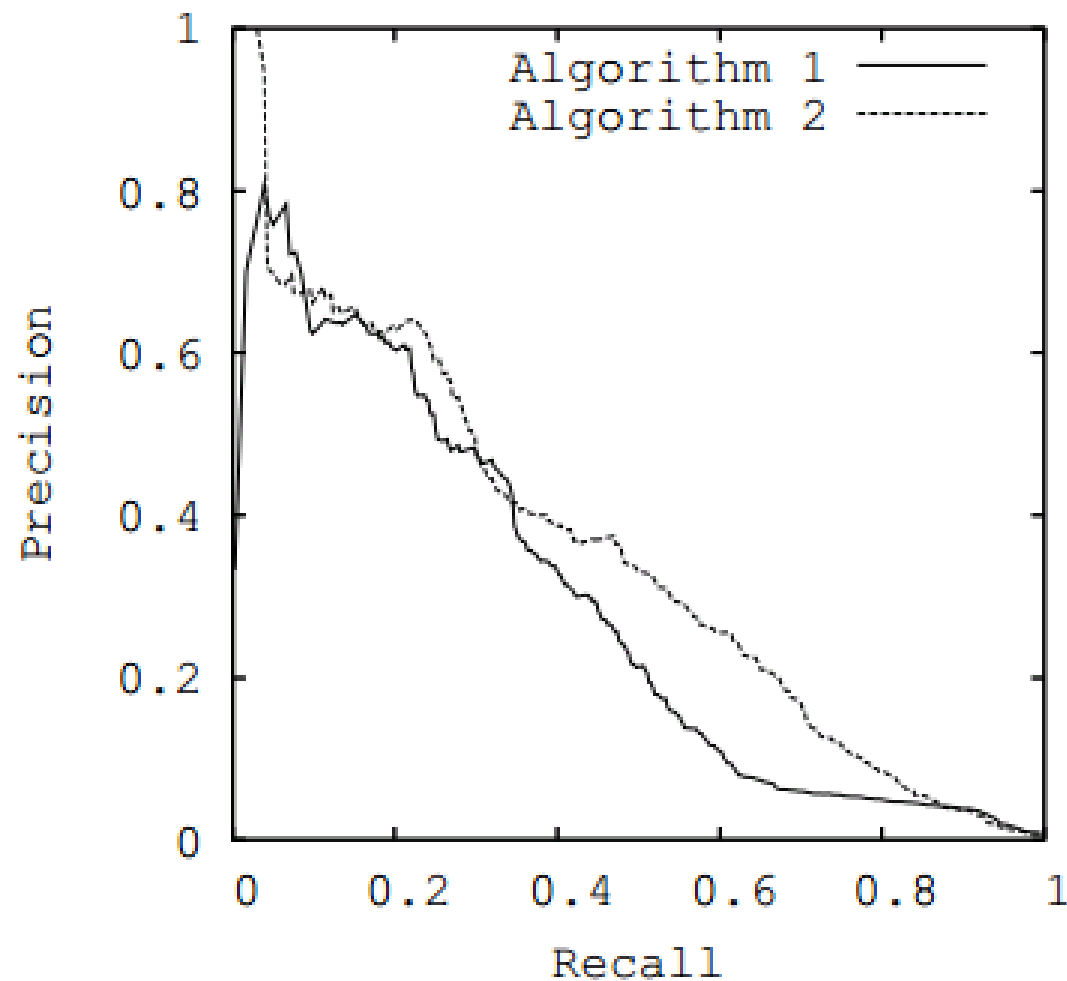


Метрики классификации

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

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

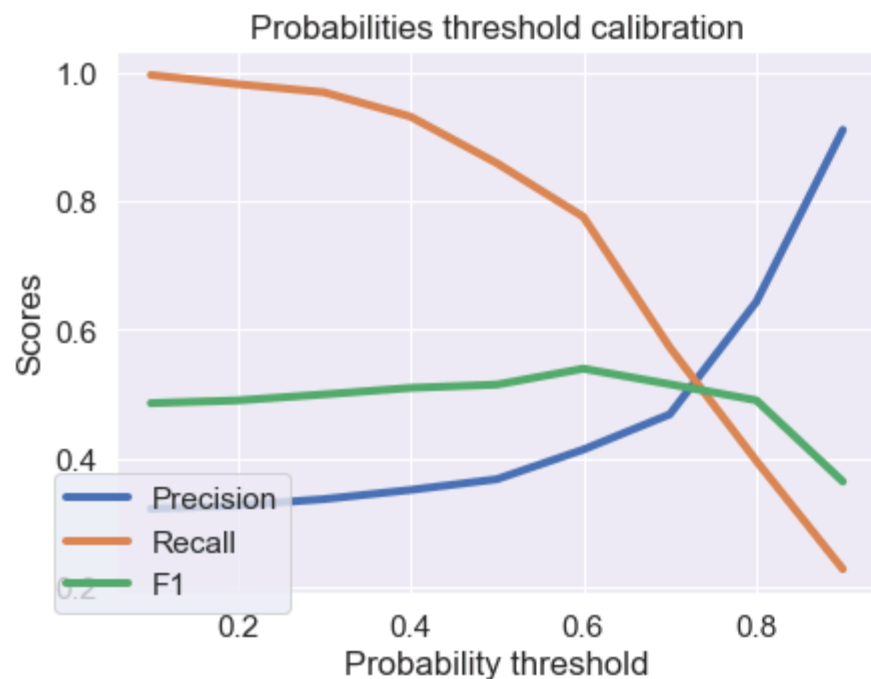
$$\text{recall} = \frac{TP}{TP + FN}$$





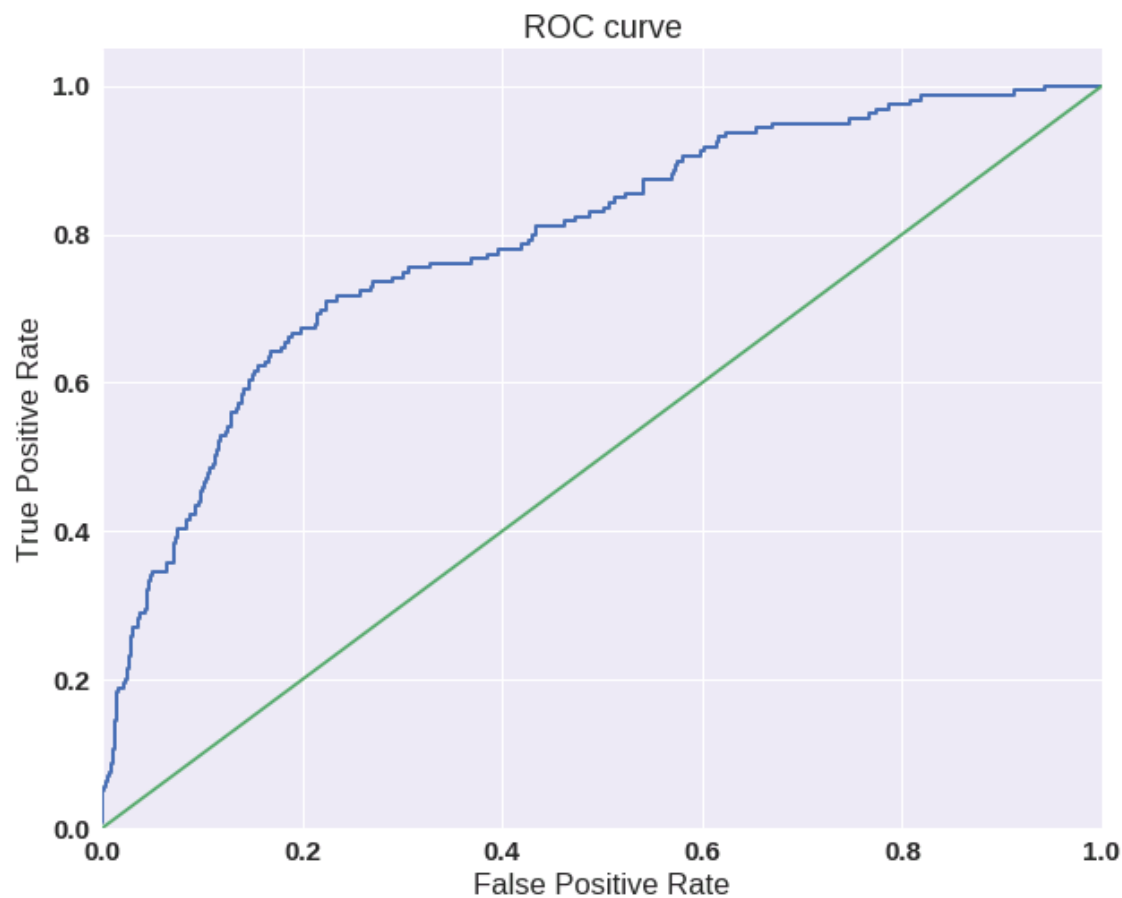
Метрики классификации

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





Метрики классификации



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

$$FPR = \frac{FP}{FP + TN}$$

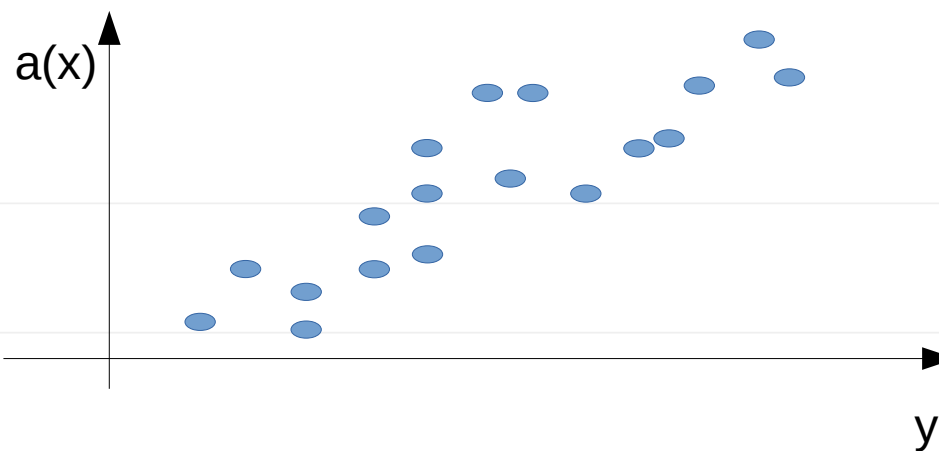
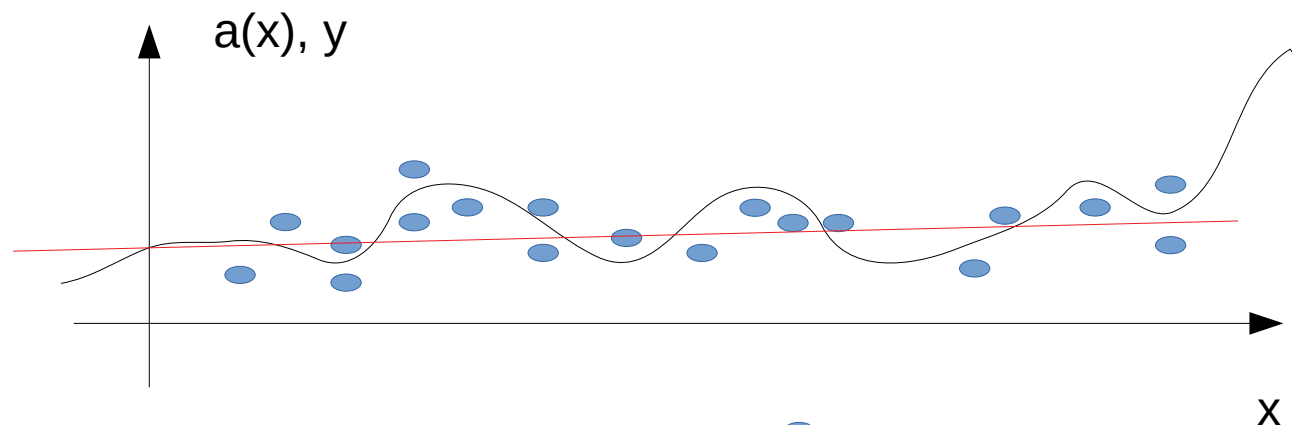


Регрессия

$$\text{MSE} = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2;$$

$$\text{MAE} = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|.$$

$$R^2 = 1 - \frac{\sum_{i=1}^{\ell} (a(x_i) - y_i)^2}{\sum_{i=1}^{\ell} (y_i - \bar{y})^2}$$





Бизнес-метрики: ROI

$$\text{ROI} = \frac{\text{Доход} - \text{Затраты}}{\text{Затраты}} \times 100\%$$

<https://ciox.ru/return-on-investment>

<https://checkroi.ru/instrumenty-marketologa/raschet-roi/>



Бизнес-метрики: ROI

$$\text{ROI} = \frac{\text{Доход} - \text{Затраты}}{\text{Затраты}} \times 100\%$$

	Доход	Расход	Прибыль	ROI
Кампания 1	7200	3500	3700	105%
Кампания 2	17600	11250	6450	56%



Бизнес-метрики: ROI

$$\text{ROI} = \frac{\text{Доход} - \text{Затраты}}{\text{Затраты}} \times 100\%$$

ROMI (Return on Marketing Investment)

$$\text{ROMI} = \frac{\text{Прибыль} - \text{Маркетинговый бюджет}}{\text{Маркетинговый бюджет}} \times 100$$



Бизнес-метрики:

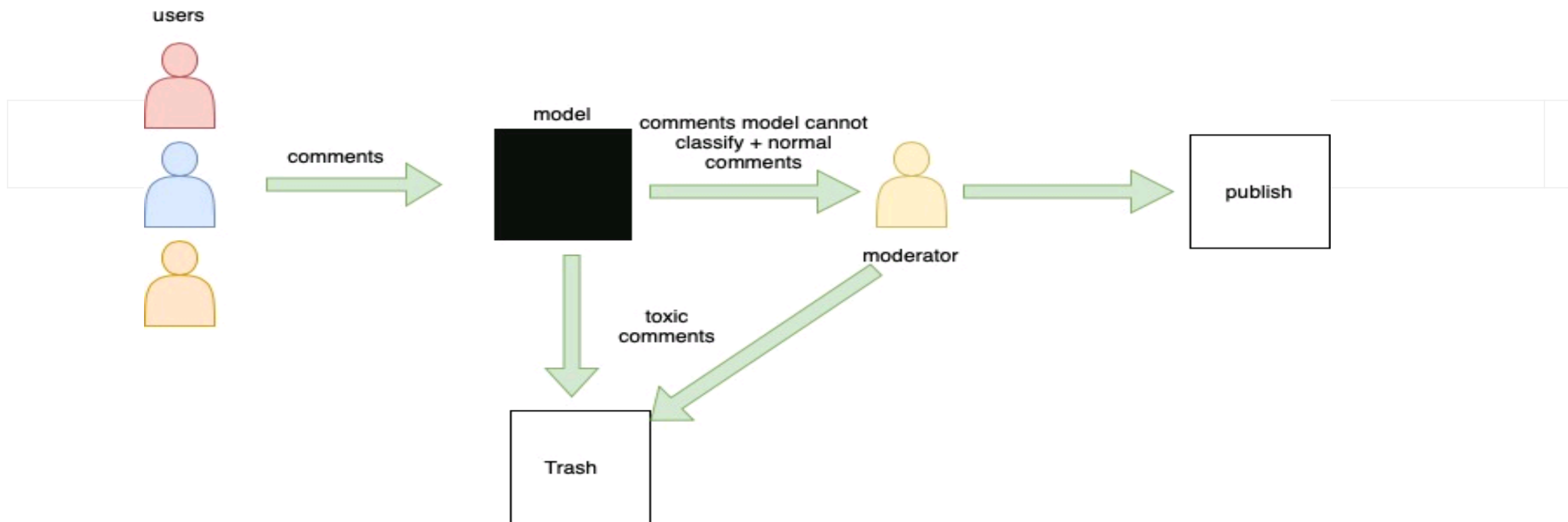
ROAS (Return on Advertising Spend)

$$\text{ROAS} = \frac{\text{Доход от рекламной кампании}}{\text{Стоимость рекламной кампании}}$$



Задача

Токсичные комментарии



A horizontal decorative bar with a pattern of small white dots on a light gray background, transitioning to a solid white background on the right.

Практика





Задача

Токсичные комментарии

автоматическая фильтрация токсичных комментариев снизит нагрузку на модераторов и повысит их пропускную способность на 10%, при этом доля FP не превысит 10%

- 1) векторное представление слов
- 2) Классификатор
- 3) Метрика



Задача

Диагностика ССЗ

Определим долю пациентов , которым нужны
дополнительные исследования

- 1) векторное представление пациентов
- 2) Классификатор
- 3) Метрика



ДОМАШНЕЕ ЗАДАНИЕ

- 1) обучить несколько разных моделей на наборе данных CC3 (train_case2.csv): логрег, бустинг, лес и т.д - на ваш выбор 2-3 варианта
- 2) при обучении моделей обязательно использовать кроссвалидацию
- 3) вывести сравнение полученных моделей по основным метрикам классификации: pr/rec/auc/f_score (можно в виде таблицы, где строки - модели, а столбцы - метрики)
- 4) сделать выводы о том, какая модель справилась с задачей лучше других
- 5) (опциональный вопрос) какая метрика (precision_recall_curve или roc_auc_curve) больше подходит в случае сильного дисбаланса классов? (когда объектов одного из классов намного больше чем другого).

p.s. В вопросе проще разобраться, если вспомнить оси на графике roc auc curve и рассмотреть такой пример:

Имеется 100000 объектов, из которых только 100 - класс "1" (99900 - класс "0", соответственно). Допустим, у нас две модели:

- первая помечает 100 объектов как класс 1, но $TP = 90$
- вторая помечает 1000 объектов как класс 1, но TP такой же - 90

Какая модель лучше и почему? И что позволяет легче сделать вывод - roc_auc_curve или precision_recall_curve?

ВАШИ ВОПРОСЫ