# Занятие № 4

## Оценка точности модели, переобучение, регуляризация

**Сапрыкин Артур**
Data Scientist

# Содержание

# Обучающая и тестовая выборка, кросс-валидация

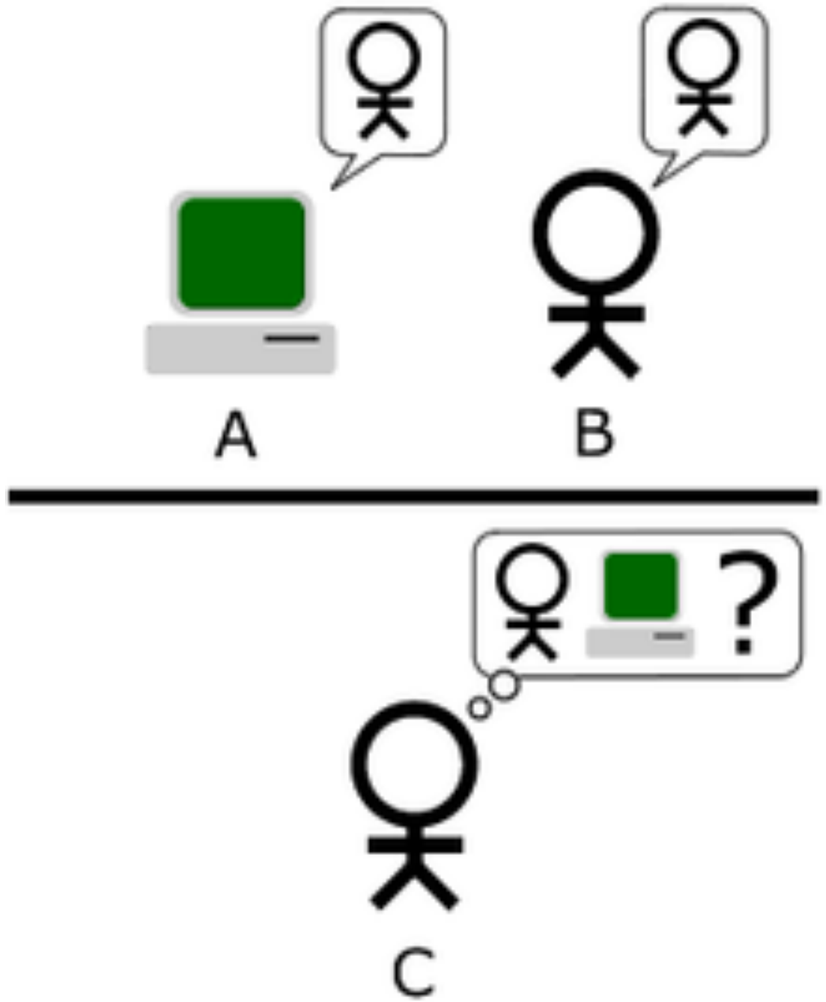
**Обучающая выборка** содержит значения признаков и целевой переменной.

На обучающей выборке строим модель.

# Обучающая и тестовая выборка, кросс-валидация

**Тестовая выборка** содержит значения признаков, по которым необходимо предсказать значение целевой переменной.

Оцениваем качество различных вариантов модели.



# Обучающая и тестовая выборка, кросс-валидация

**Проблемы:**
Модель может хорошо работать на обучающей выборке, однако сильно терять в качестве на тестовой (один из вариантов - переобучение).

Преобразования данных на обучающей выборке должны быть повторены и иметь смысл для тестовой.

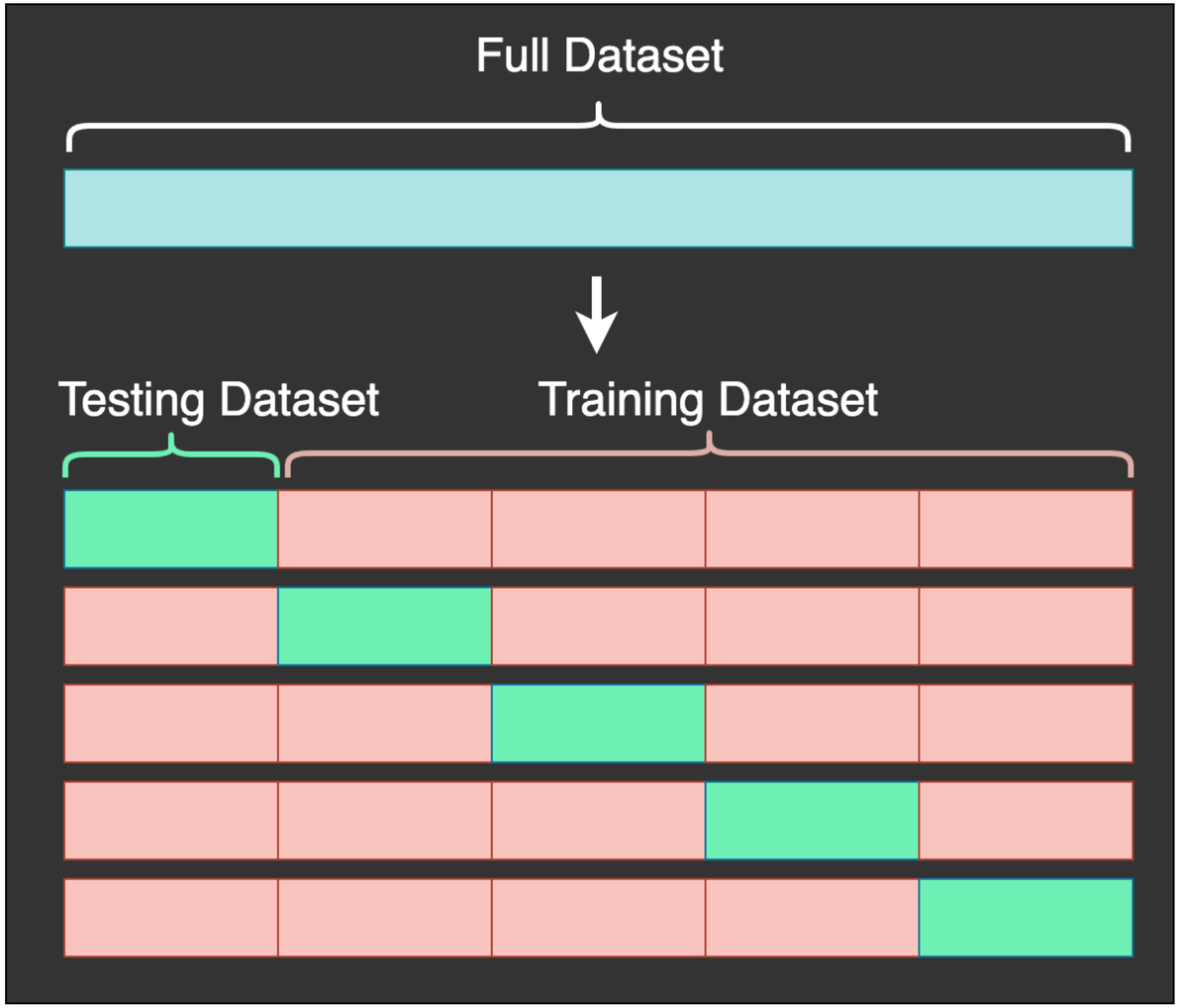
# Обучающая и тестовая выборка, кросс-валидация

Разбиваем обучающую выборку на 2 части.

На одной будем тренировать модель, на другой – проверять (т. е. использовать в качестве тестовой, только с известной целевой переменной)

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size = 0.3, random_state = 0 )
```

# Обучающая и тестовая выборка, кросс-валидация



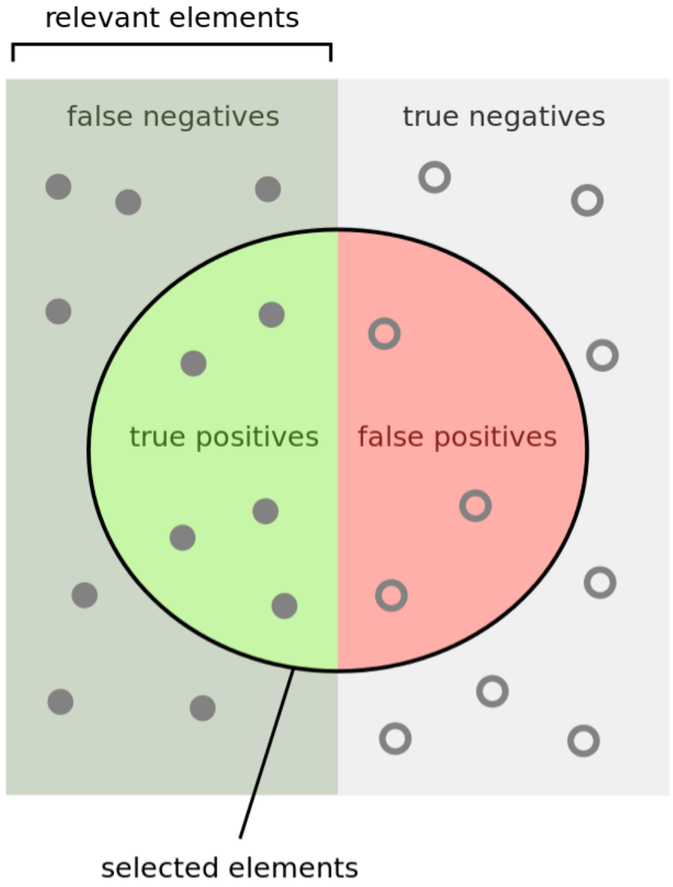
# Метрики качества: accuracy, precision, recall

| n=165 | Predicted: NO | Predicted: YES | |
|---|---|---|---|
| Actual: NO | TN = 50 | FP = 10 | 60 |
| Actual: YES | FN = 5 | TP = 100 | 105 |
| | 55 | 110 | |

# Метрики качества: accuracy, precision, recall

$$accuracy = \frac{correct}{correct + incorrect}$$

# Метрики качества: accuracy, precision, recall



# Метрики качества: accuracy, precision, recall

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

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

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

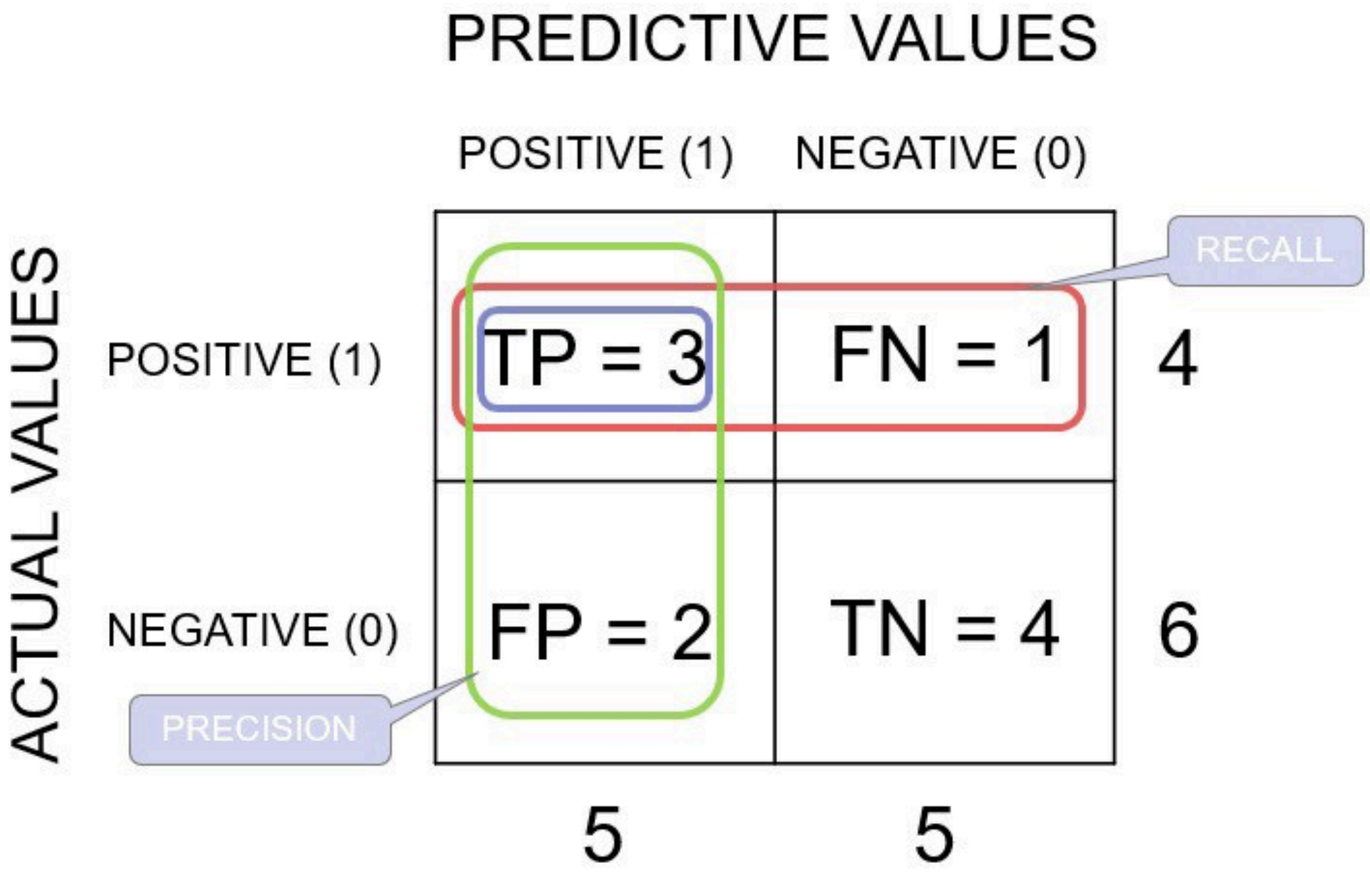
$TP$ = True positive

$TN$ = True negative

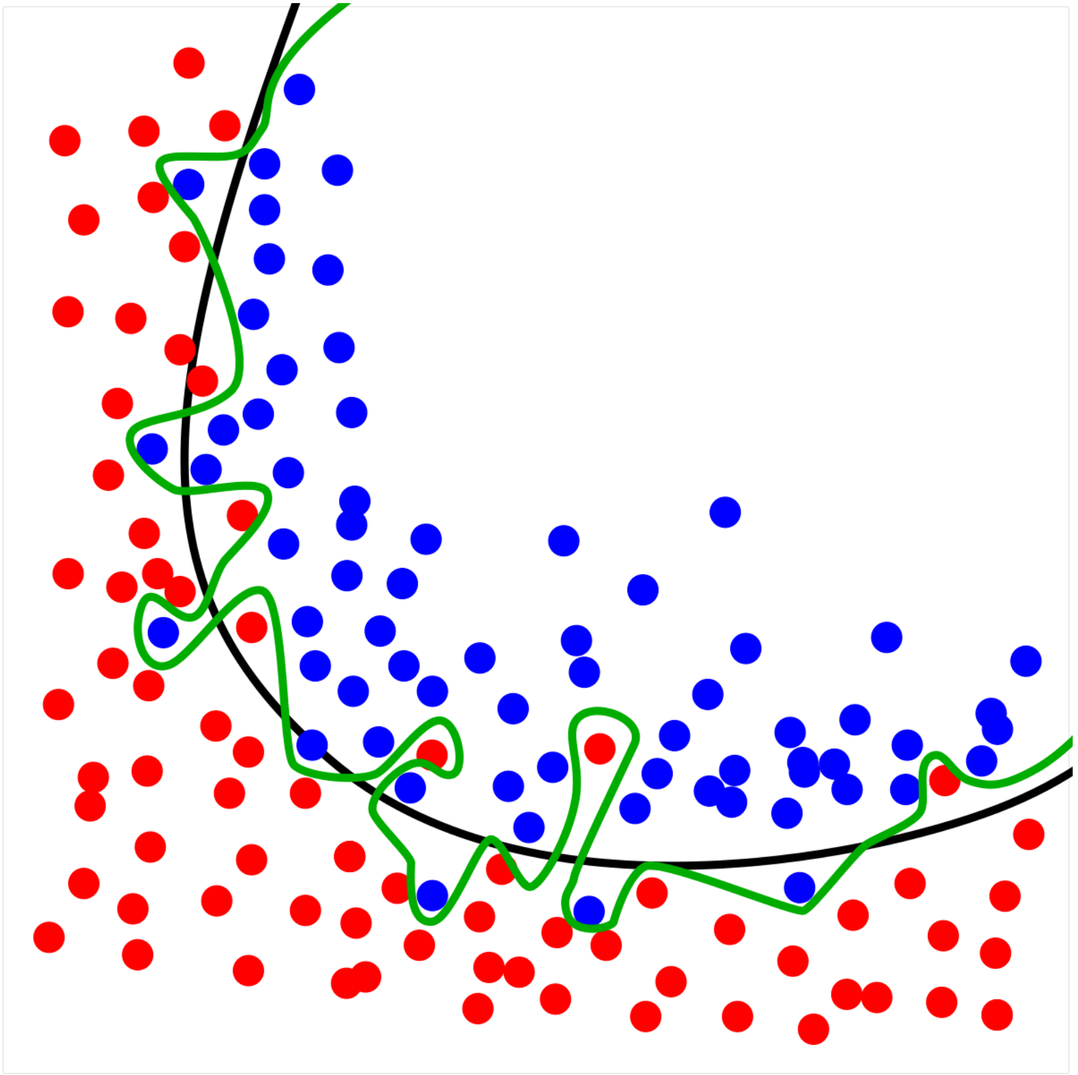
$FP$ = False positive

$FN$ = False negative

# Метрики качества: accuracy, precision, recall



# Признаки переобучения и регуляризация



# Признаки переобучения и регуляризация

# Признаки переобучения и регуляризация

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

# ПРАКТИКА

# Спасибо за внимание!

**Сапрыкин Артур**
Data Scientist

fb.com/asaprykin92

asaprykin92@gmail.com