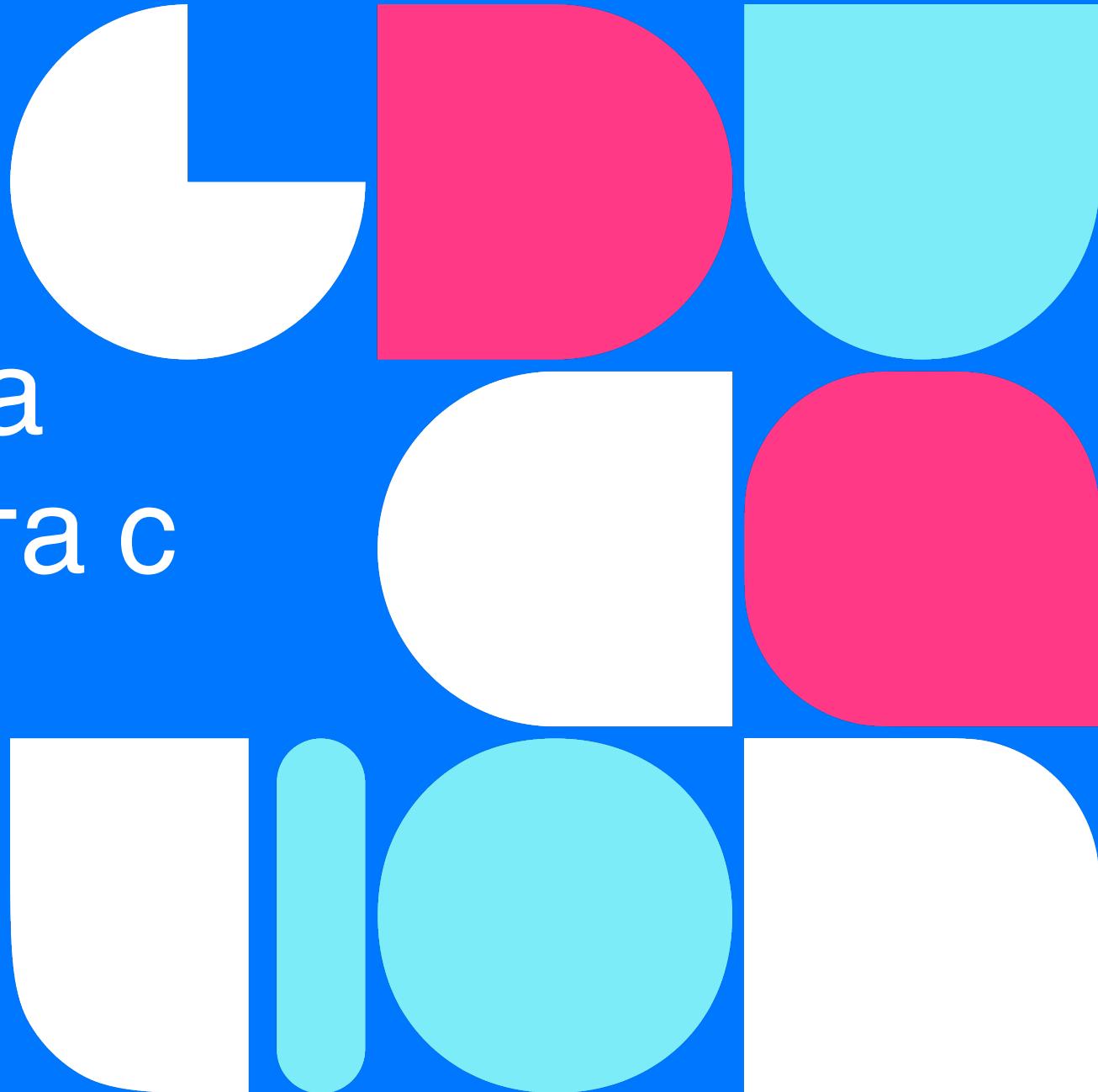




Оценка качества моделей и работа с признаками

Занятие №4

Ярошенко Ангелина



To Do

Показывать	Дисциплина	Тип события
Ближайшие две недели	Весь семестр	Все типы
2 марта	18:00 — 21:00	Введение в ML (ML-23) Смешанное занятие 1
четверг		онлайн ML-11, 12, 13
15 марта	18:00 — 21:00	Введение в ML (ML-23) Смешанное занятие 2
среда		онлайн ML-11, 12, 13
22 марта	18:00 — 21:00	Введение в ML (ML-23) Смешанное занятие 3
среда		онлайн ML-11, 12, 13
29 марта	18:00 — 21:00	Введение в ML (ML-23) Смешанное занятие 4
среда		онлайн ML-11, 12, 13
5 апреля	18:00 — 21:00	Введение в ML (ML-23) Смешанное занятие 5
среда		онлайн ML-11, 12, 13



План лекции

1. Метрики качества

1.1 Метрики качества

1.2 Оценка качества моделей

2. Работа с признаками

2.1 Извлечение признаков

2.2 Преобразование признаков

2.3 Работа с пропущенными данными

2.4 Отбор признаков

Метрики



ff

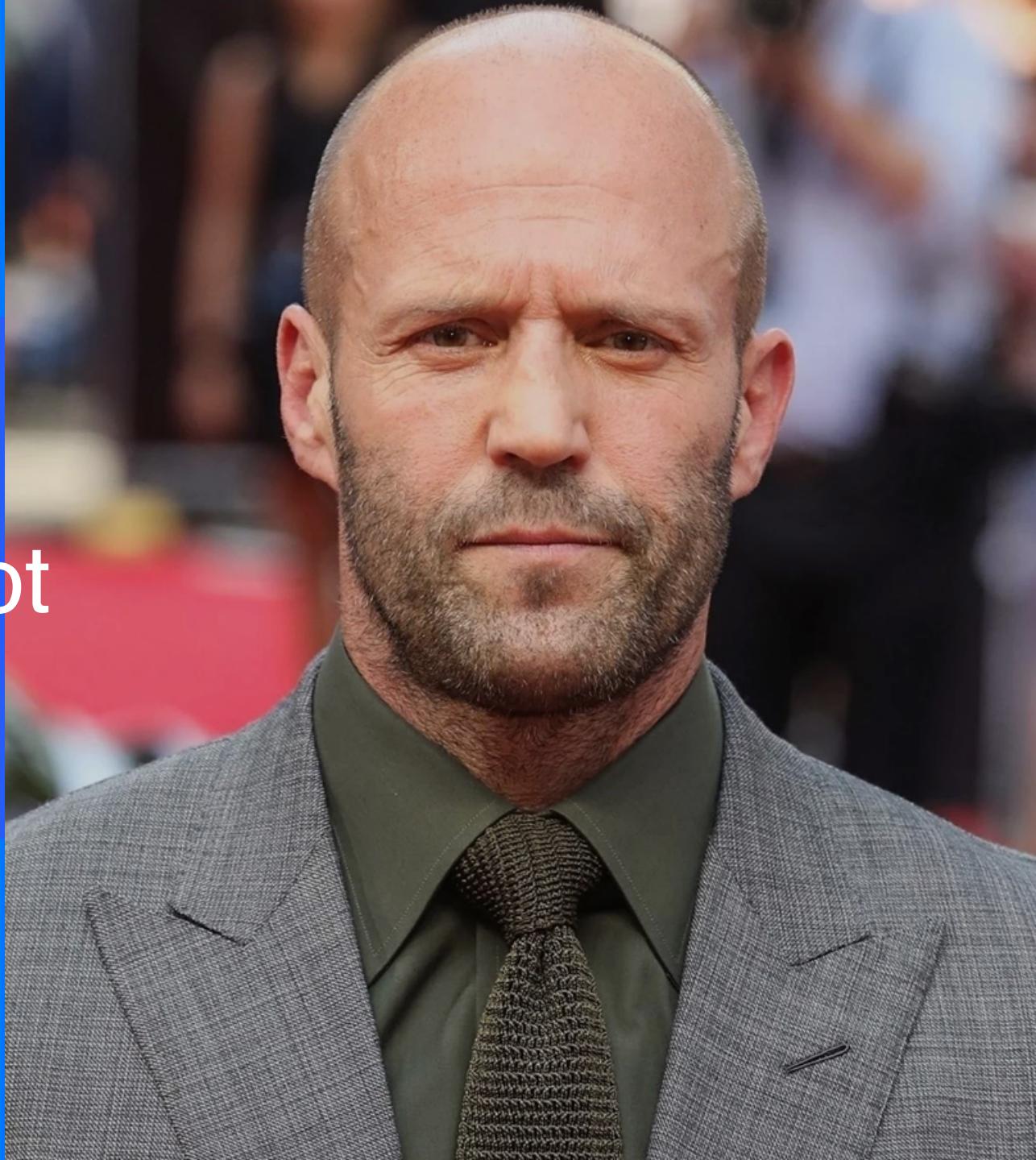
“If you can not
measure it, you can not
improve it.”

- Lord Kelvin

”



education



Зачем нужны метрики?

- Для оценки качества работы моделей
- Для сравнения моделей
- Для интерпретации результатов

Постановка задачи

X – множество **объектов**

$Y \in \mathbb{R}$ – множество **ответов**

$\{x_1, \dots, x_\ell\} \subset X$ – обучающая выборка

$y_i = y(x)$, $i = 1, \dots, \ell$ – известное множество **ответов**

$a : X \rightarrow Y$ – алгоритм, ставящий в соответствие **объекту x** некоторый **ответ y**

$a_i = a(x)$, $i = 1, \dots, \ell$ – **ответы** (предсказания) нашего алгоритма на выборке X

Задача регрессии



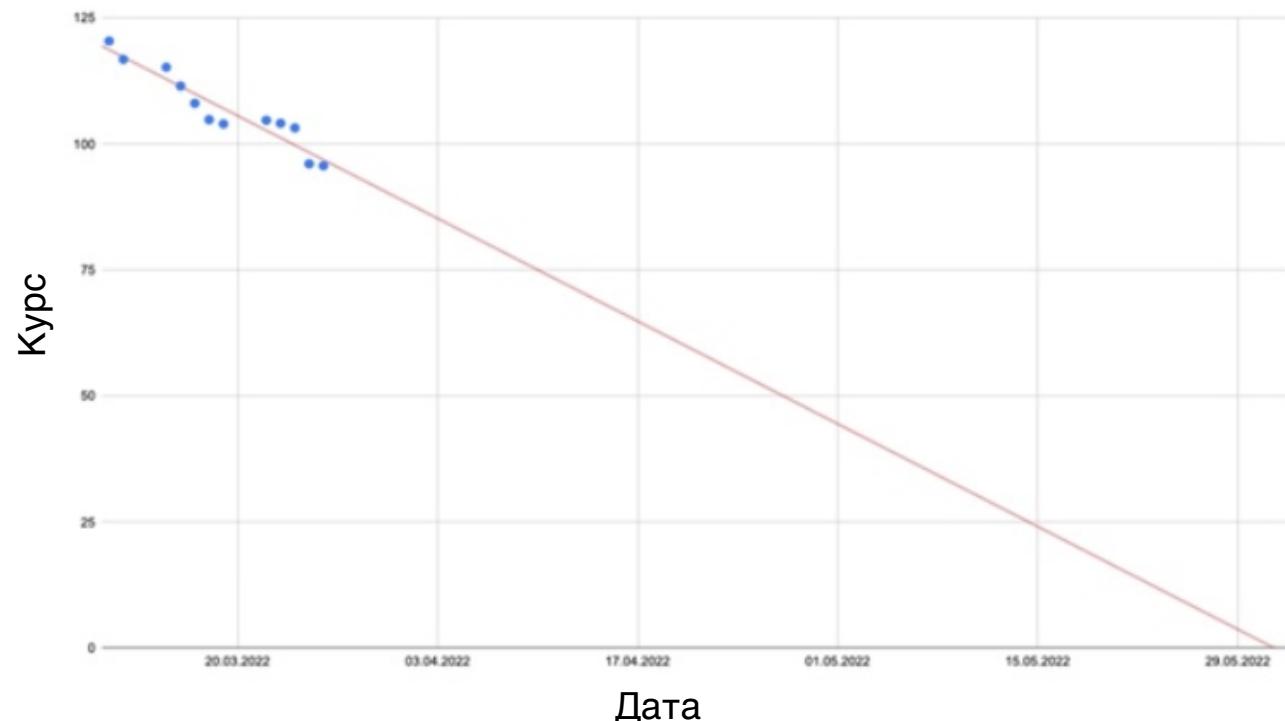
Физкек запись закреплена
26 мар 2022 в 18:53

+ Подписаться

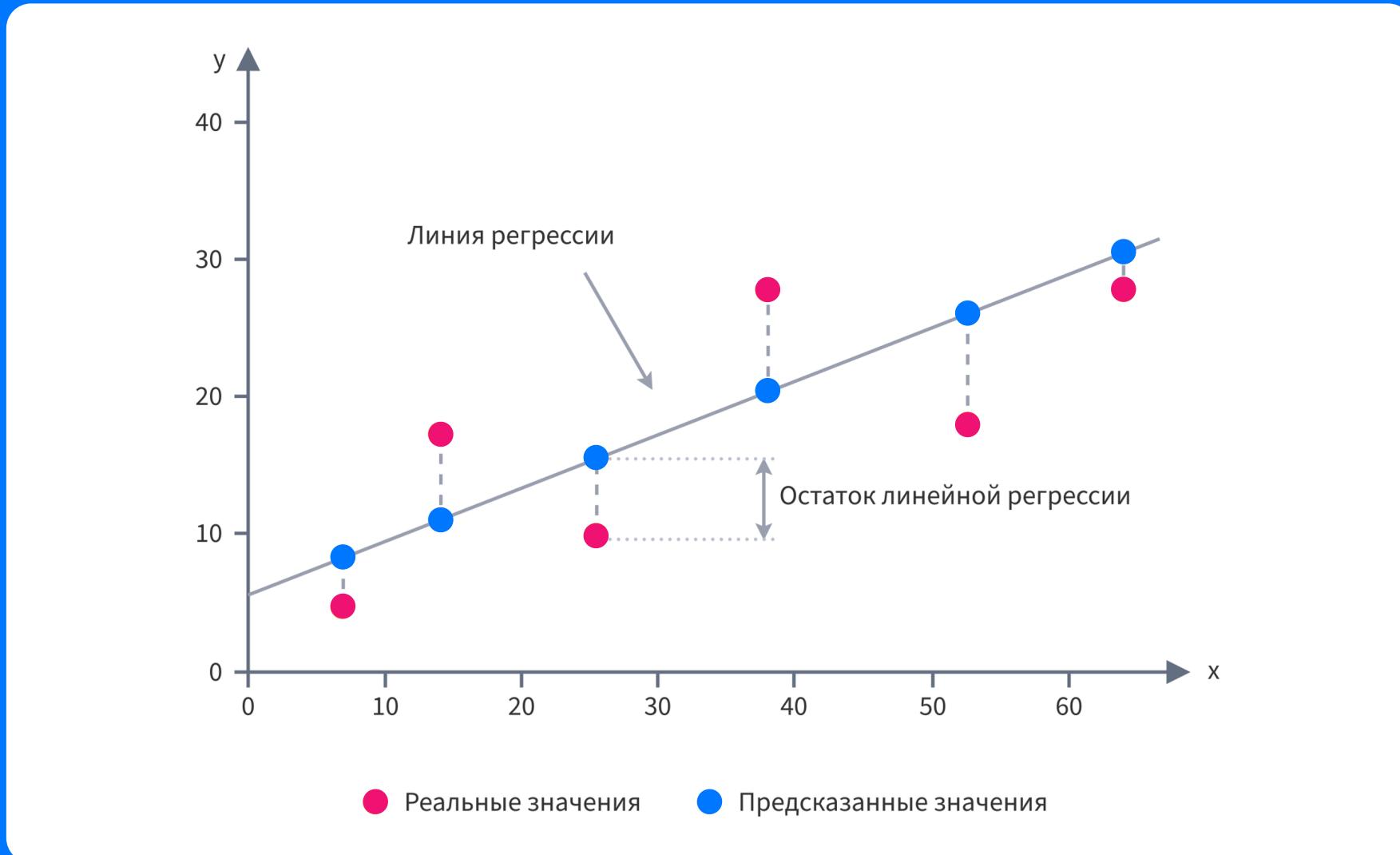
...

примерно к началу лета наши аналитики ожидают бесплатного доллара

Курс доллара США к рублю. Экстраполяция



Остатки (residuals)

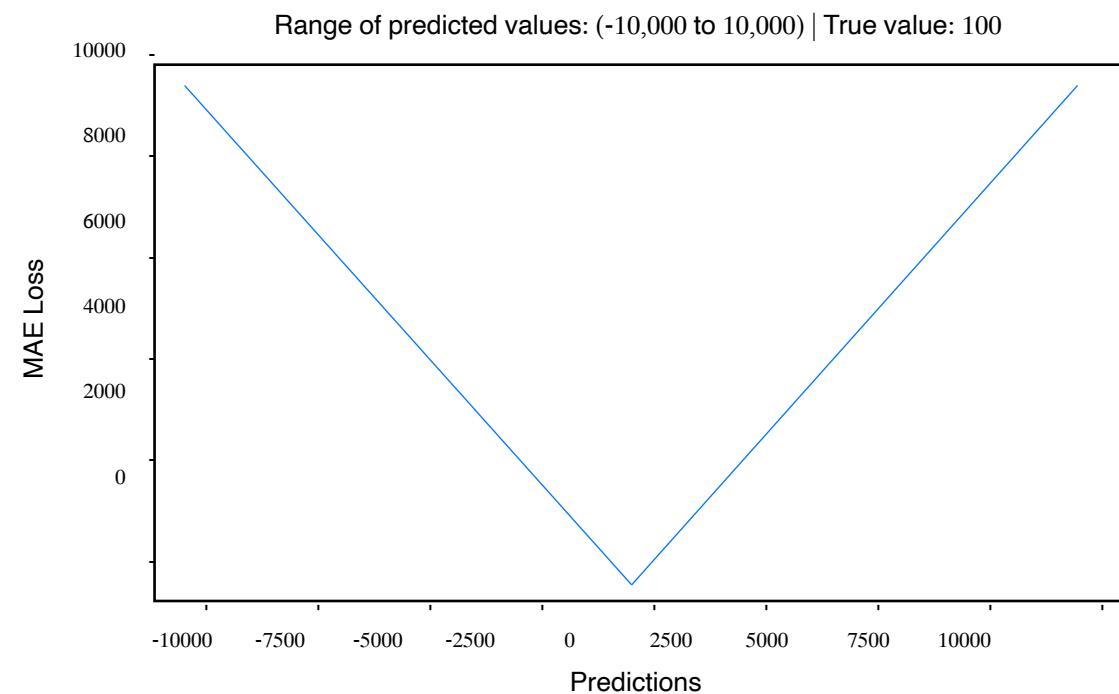


MAE

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

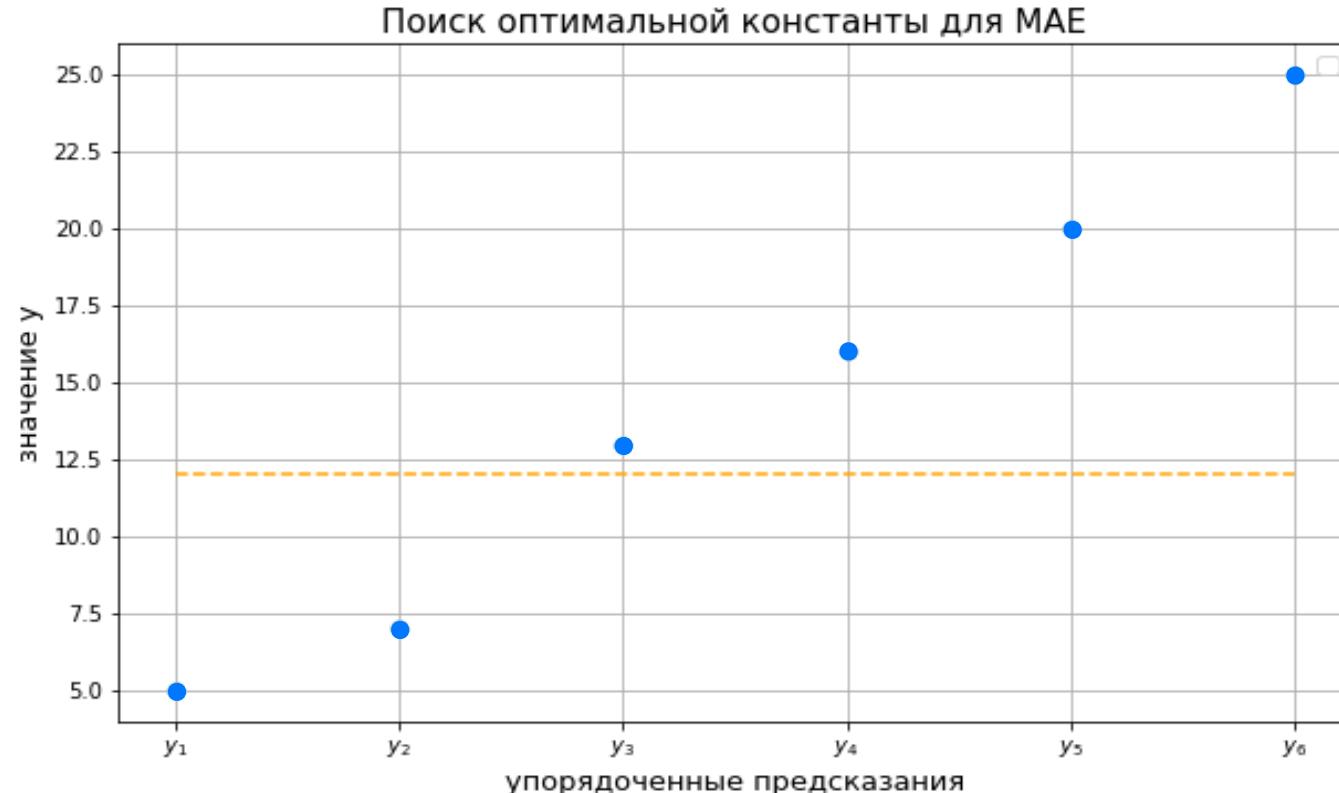
Оптимальное статичное решение – медиана ответов Y.

- Единицы измерения как у таргета
- Сложно интерпретировать
- Нечувствителен к выбросам
- Не дифференцируемая



Оптимальное решение МАЕ

- Красивое доказательство – через производную
- Интуитивно – всегда можно нарисовать картинку

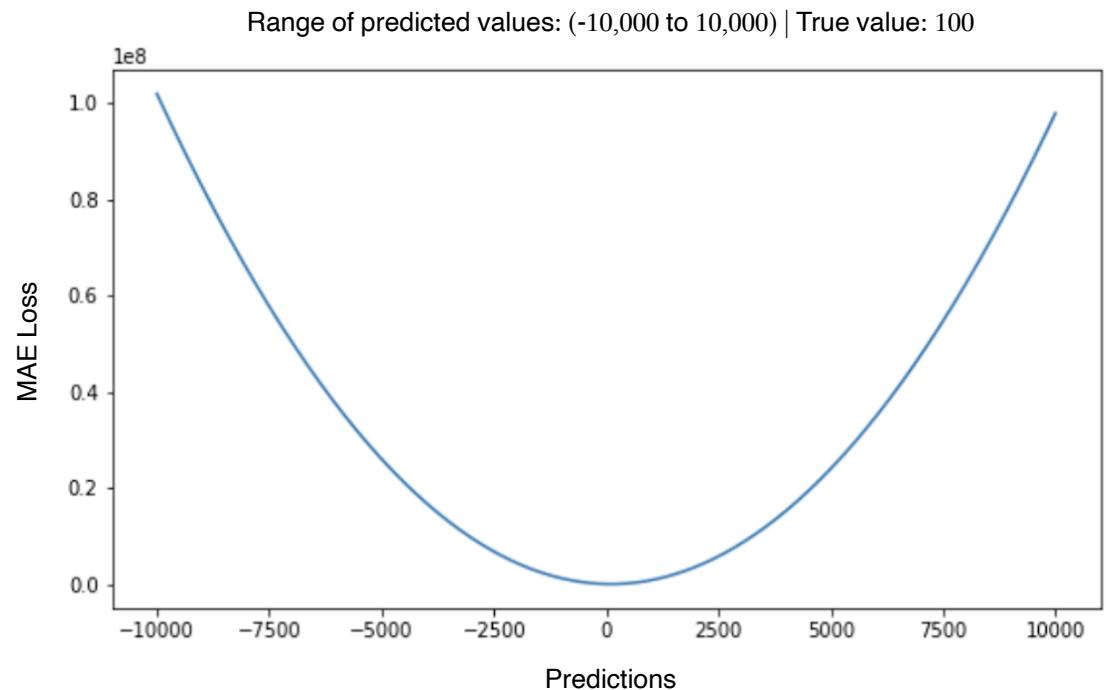


MSE

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Оптимальное статичное решение – среднее ответов Y.

- Дифференцируемая
- Чувствительна к выбросам
- Сложно интерпретировать



RMSE

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

Оптимальное статичное решение – среднее ответов \bar{Y} .

- Дифференцируемая
- Чувствительна к выбросам
- Интерпретация: стандартное отклонение ответа

R-квадрат

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

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$$

- Дифференцируемая
- Чувствительна к выбросам
- Хорошо интерпретируется: насколько наша модель лучше, чем константное решение

MSPE и MAPE

$$\text{MSPE} = \frac{100\%}{N} \sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2$$

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Оптимальное константное решение:

взвешенное среднее таргета

взвешенная медиана таргета

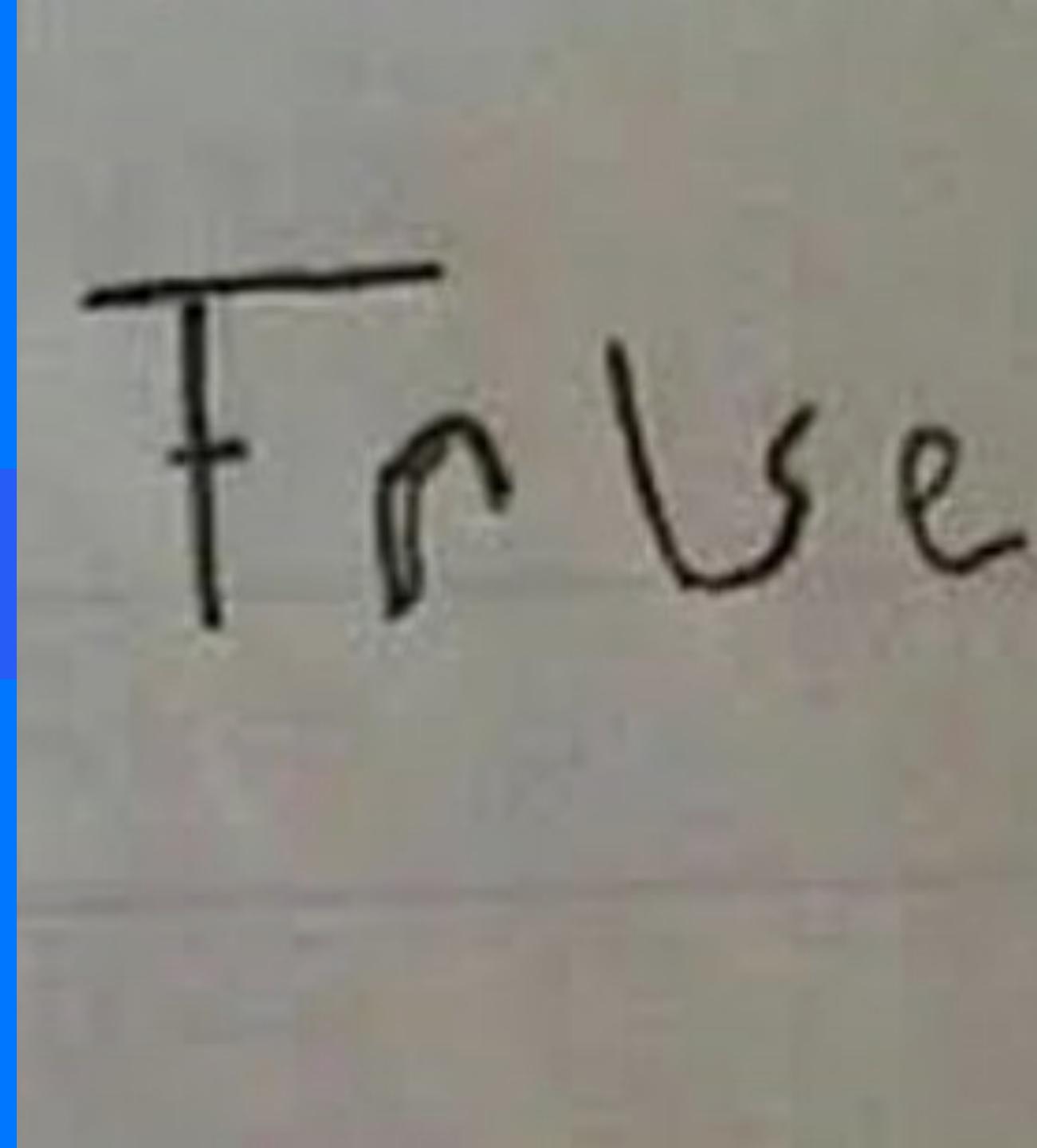
- Присваивают большой вес абсолютно более маленьким объектам => смещены
- Нечувствительны к выбросам
- Хорошо интерпретируются: относительный прирост

Выводы

		Аномальные значения	
		Учитываем	Не учитываем (выбросы)
Интерпретируемость	Важна	RMSE, R-квадрат	MSPE, MAPE
	Не важна	MSE	MAE

Задача классификации

When your binary classification model
outputs 0.5



Accuracy

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N [\hat{y}_i = y_i] \quad - \quad \text{доля правильных предсказаний}$$

`target = np.array([1, 1, 1, 2, 1, 1, 1, 2])` – какое лучшее константное предсказание?

Accuracy

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N [\hat{y}_i = y_i]$$

- ОЧЕНЬ большая чувствительность
к дисбалансу классов

Оптимальное статичное решение – самый популярный класс.

Confusion matrix

True – совпадение ответов

		$\hat{Y} = 0$ NEGATIVE	$\hat{Y} = 1$ POSITIVE
$Y = 0$ NOT PREGNANT	TRUE NEGATIVE	FALSE POSITIVE	Positive – значение предсказания
	TYPE 1 ERROR	...	
$Y = 1$ PREGNANT	FALSE NEGATIVE	TRUE POSITIVE	Positive – значение предсказания
	TYPE 2 ERROR	You're pregnant	

TRUE NEGATIVE

You're not pregnant

FALSE POSITIVE

You're pregnant

TRUE POSITIVE

You're pregnant

FALSE NEGATIVE

You're not pregnant

TYPE 2 ERROR

Confusion matrix

		Predicted		
		Positive	Negative	
Actual	Positive	True Positive (TP)	False Negative (FN)	Sensitivity or Recall or True Positive Rate = $TP/(TP+FN)$
	Negative	False Positive (FP)	True Negative (TN)	Specificity or True Negative Rate = $TN/(TN+FP)$
	Precision or Positive Predictive Value = $TP/(TP+FP)$	Negative Predictive Value = $FN/(FN+TN)$	Accuracy = $TP+TN/TP+TN+FP+FN$	

Confusion matrix

How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

relevant elements

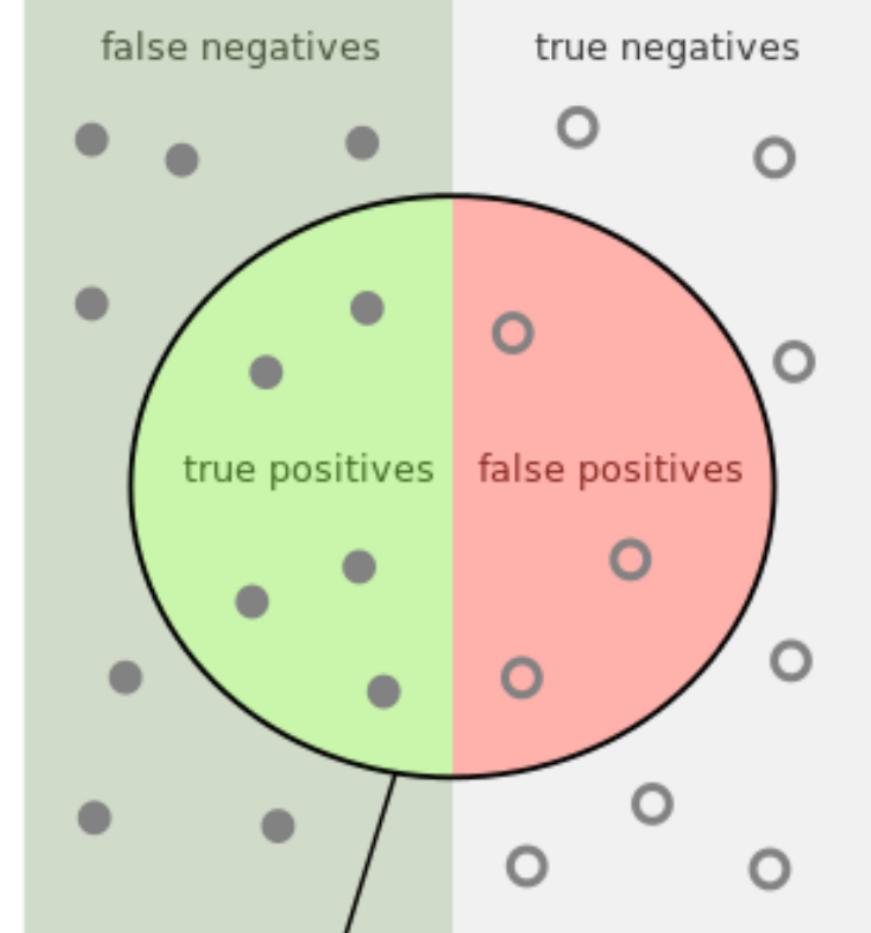
false negatives

true negatives

true positives

false positives

selected elements

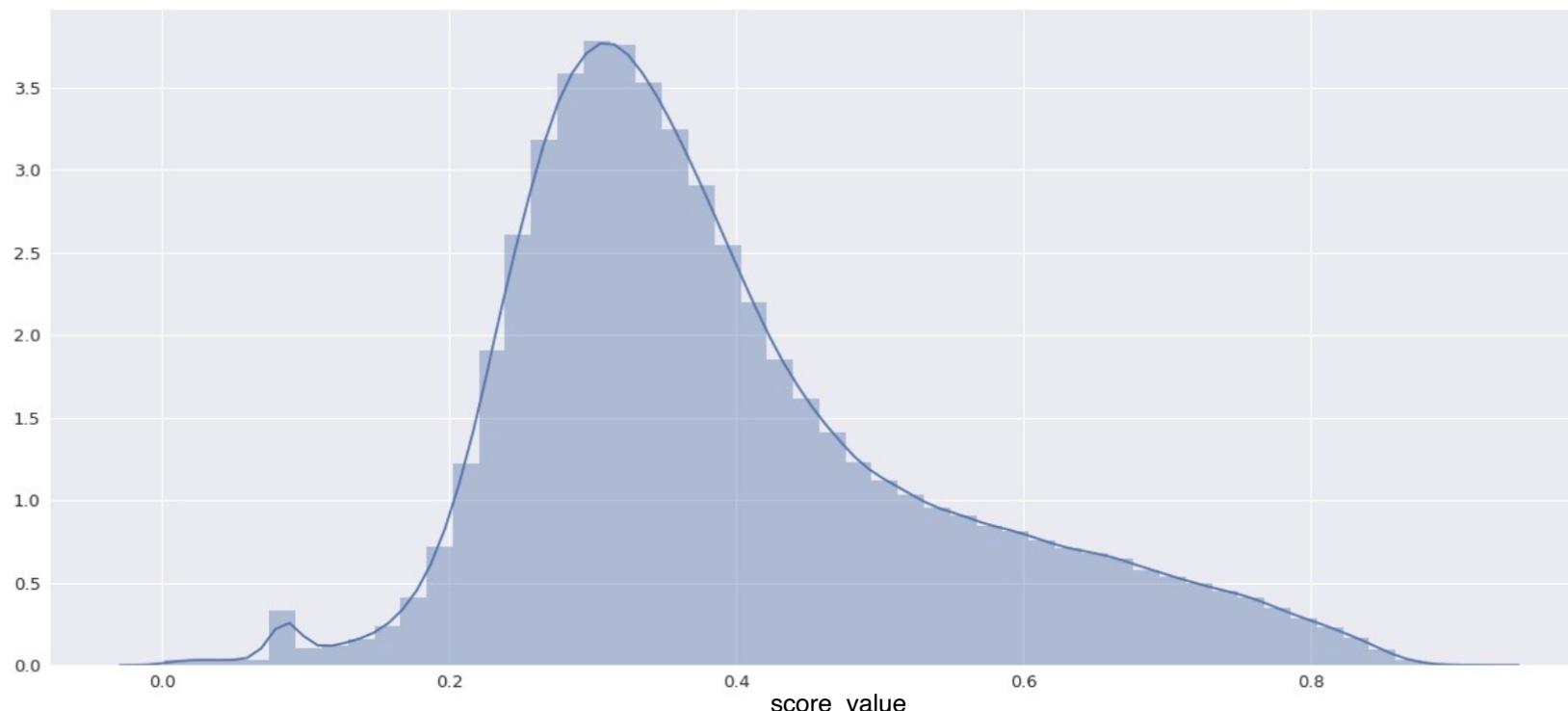


Soft target

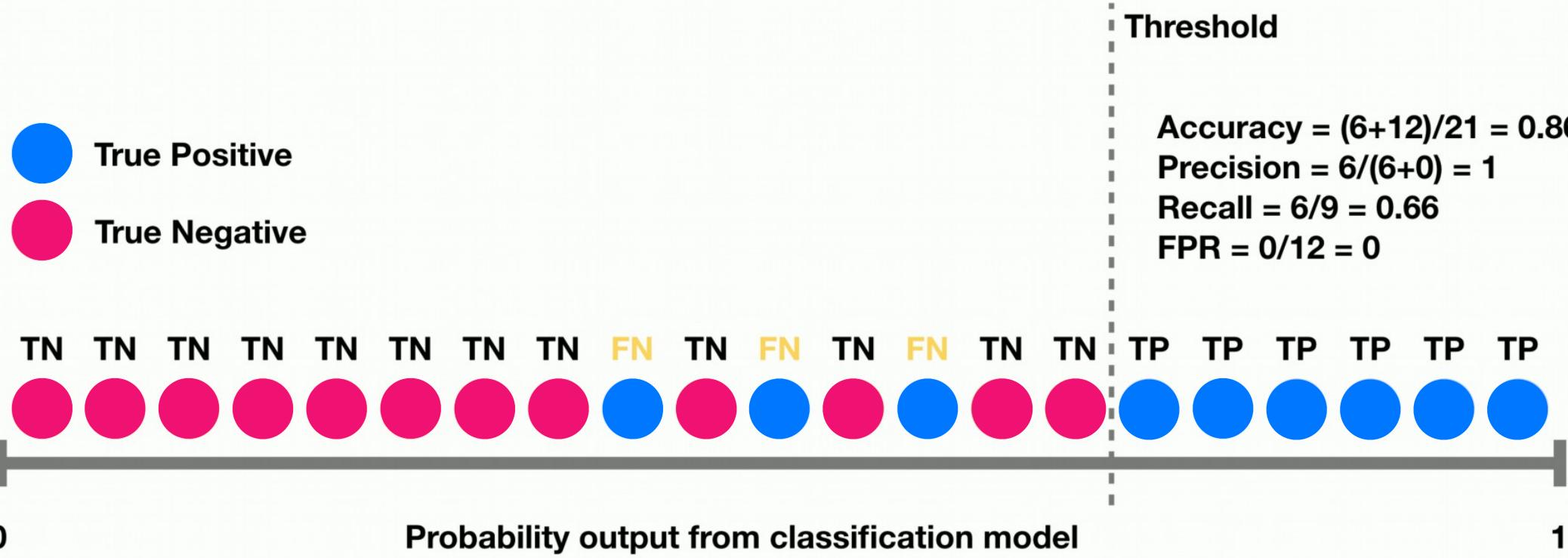
$$a : X \rightarrow Y, Y \in (0,1)$$

- алгоритм предсказывает значение от 0 до 1 (например, вероятность принадлежности к положительному классу)

Распределение предсказания на тестовом множестве



Confusion matrix



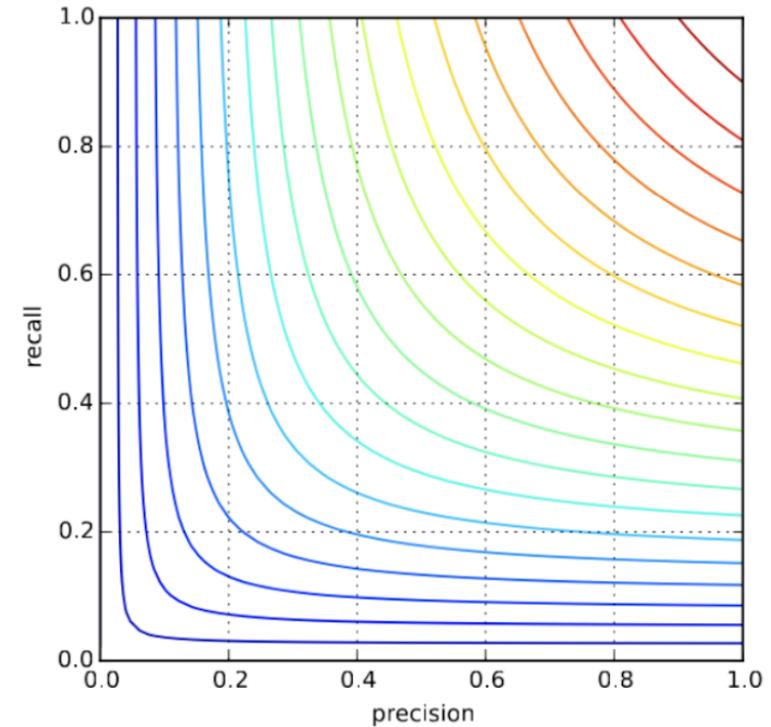


F-measure

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

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{true positive}}{(1 + \beta^2) \cdot \text{true positive} + \beta^2 \cdot \text{false negative} + \text{false positive}}$$

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

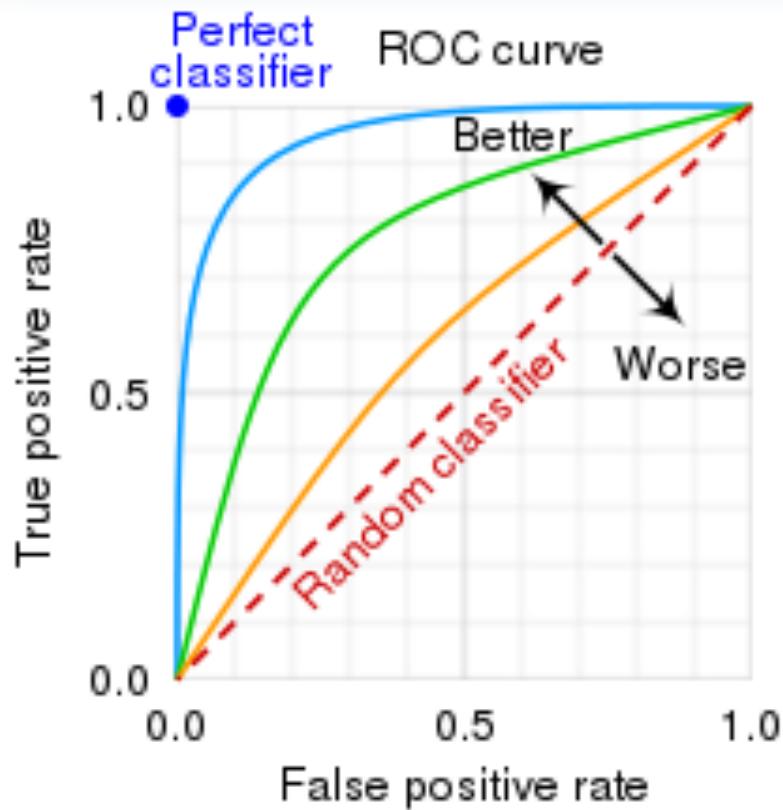


Стремится к нулю, если precision или recall близки к нулю.

Коэффициент β отвечает за пропорцию между метриками.

AUC-ROC

Area Under the ROC curve определяет долю правильно отранжированных пар



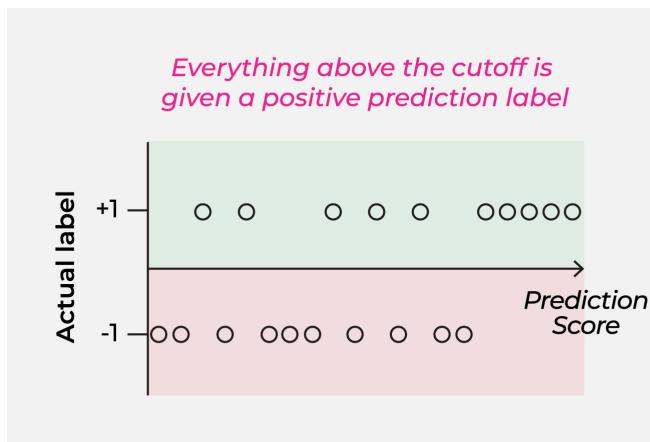
$$\text{TPR (sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR (1-specificity)} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

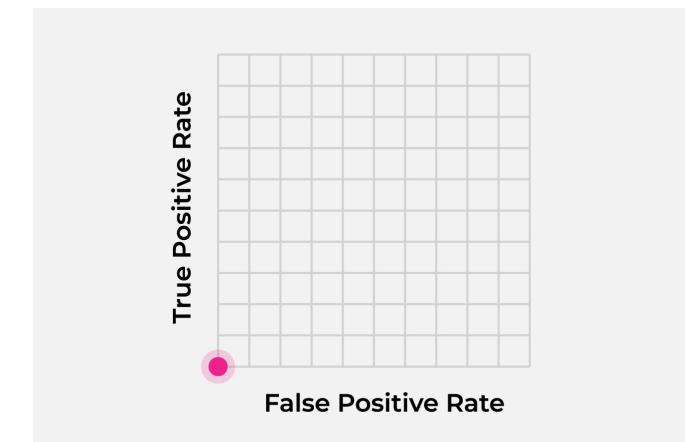
AUC-ROC

True Positive Rate: 0/10

False Positive Rate: 0/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	0	10
	NEG (-)	0	10

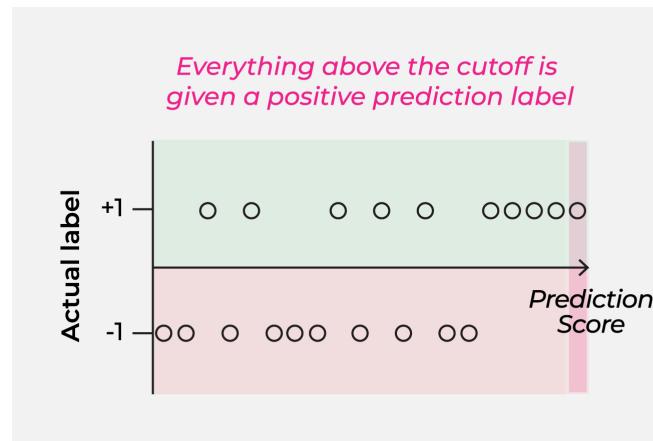


<https://arize.com/blog/what-is-auc/>

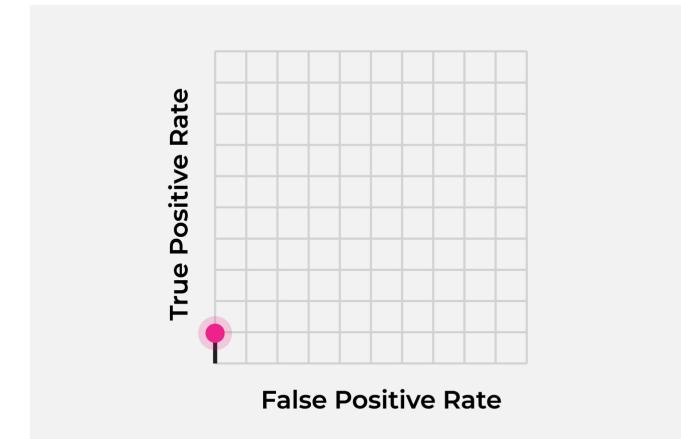
AUC-ROC

True Positive Rate: 1/10

False Positive Rate: 0/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	1	9
	NEG (-)	0	10

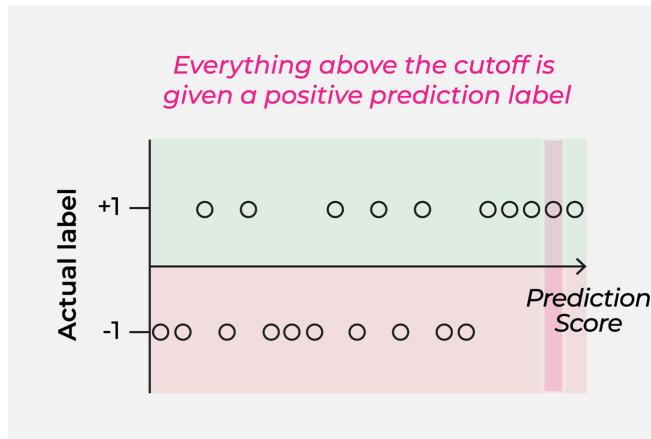


<https://arize.com/blog/what-is-auc/>

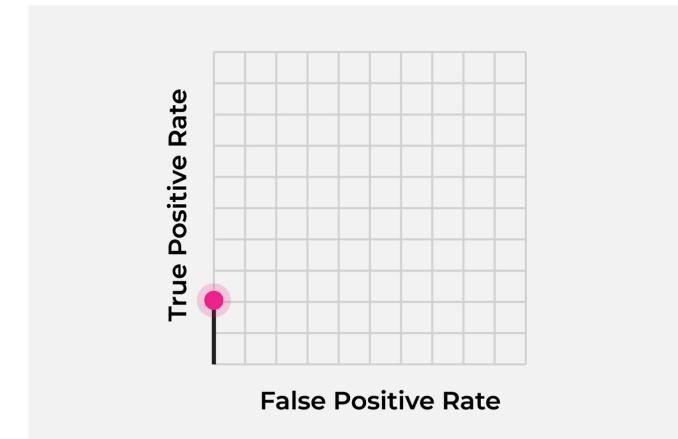
AUC-ROC

True Positive Rate: 2/10

False Positive Rate: 0/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	2	8
	NEG (-)	0	10

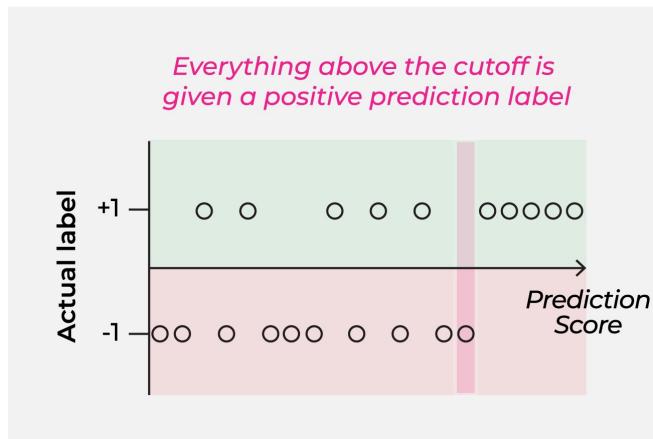


<https://arize.com/blog/what-is-auc/>

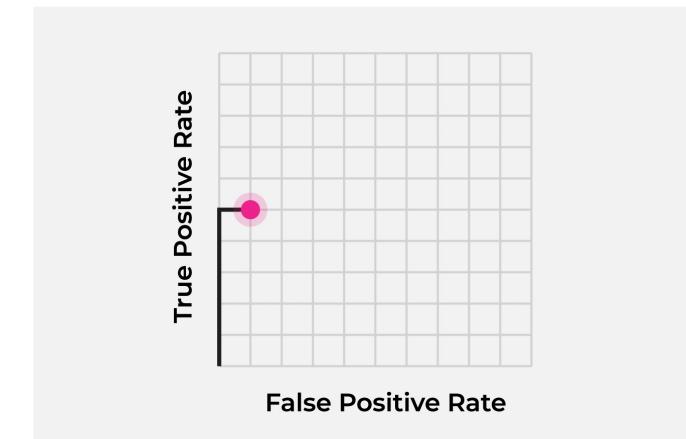
AUC-ROC

True Positive Rate: 5/10

False Positive Rate: 1/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	5	5
	NEG (-)	1	9

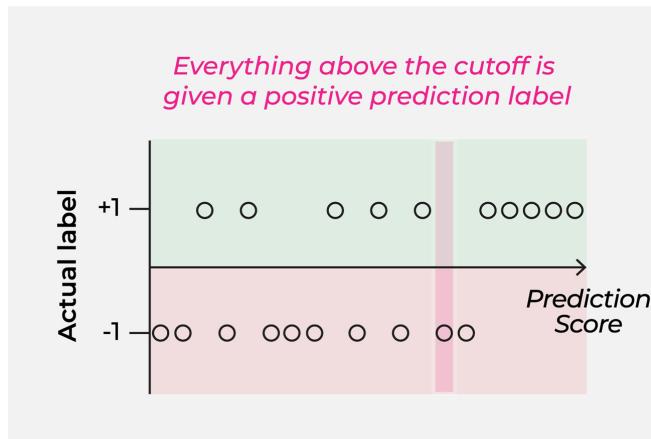


<https://arize.com/blog/what-is-auc/>

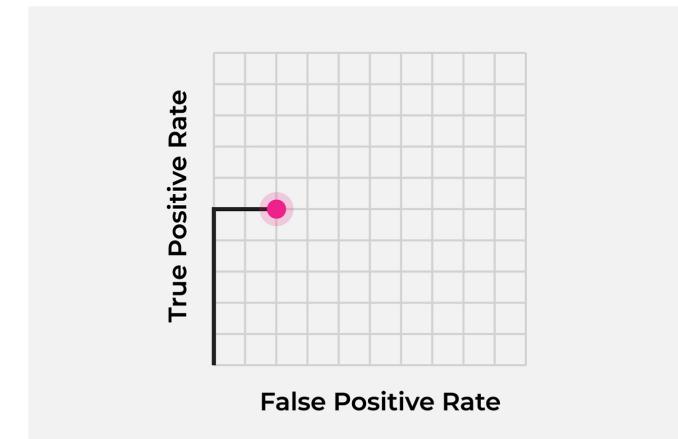
AUC-ROC

True Positive Rate: 5/10

False Positive Rate: 2/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	5	5
	NEG (-)	2	8

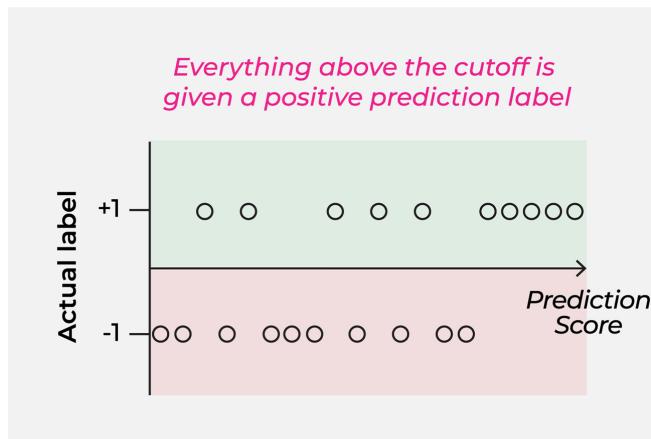


<https://arize.com/blog/what-is-auc/>

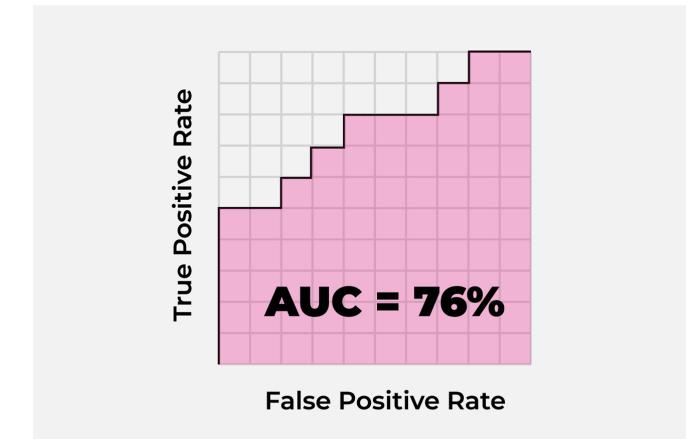
AUC-ROC

True Positive Rate: 10/10

False Positive Rate: 10/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	10	0
	NEG (-)	10	0

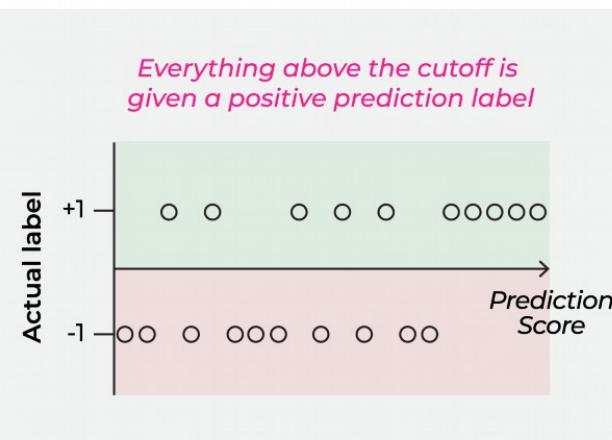


<https://arize.com/blog/what-is-auc/>

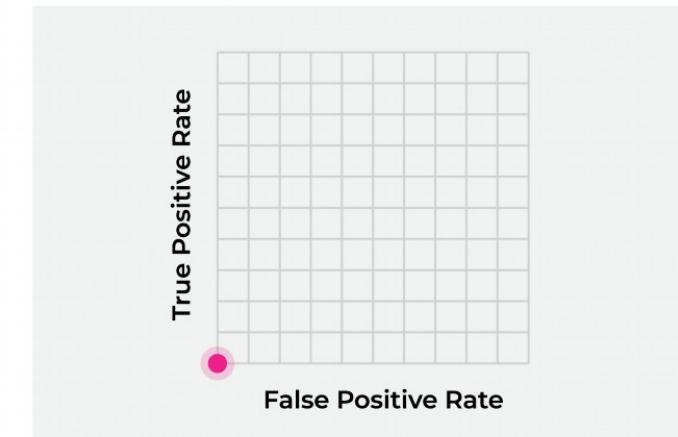
AUC-ROC

True Positive Rate: 0/10

False Positive Rate: 0/10



		Prediction label	
		POS (+)	NEG (-)
Actual label	POS (+)	0	10
	NEG (-)	0	10

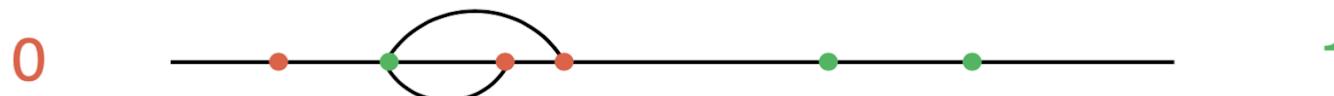


<https://arize.com/blog/what-is-auc/>

AUC-ROC

*только пары с разными метками

$$\text{AUC} = \frac{\# \text{ correctly ordered pairs}}{\text{total number of pairs}} = 1 - \frac{\# \text{ incorrectly ordered pairs}}{\text{total number of pairs}}$$
$$= 1 - 2 / 9 = 7 / 9$$



Соответственно, ROC-AUC слабо чувствительна на несбалансированных выборках.
Очень интересный [тест](#)

AUC-ROC

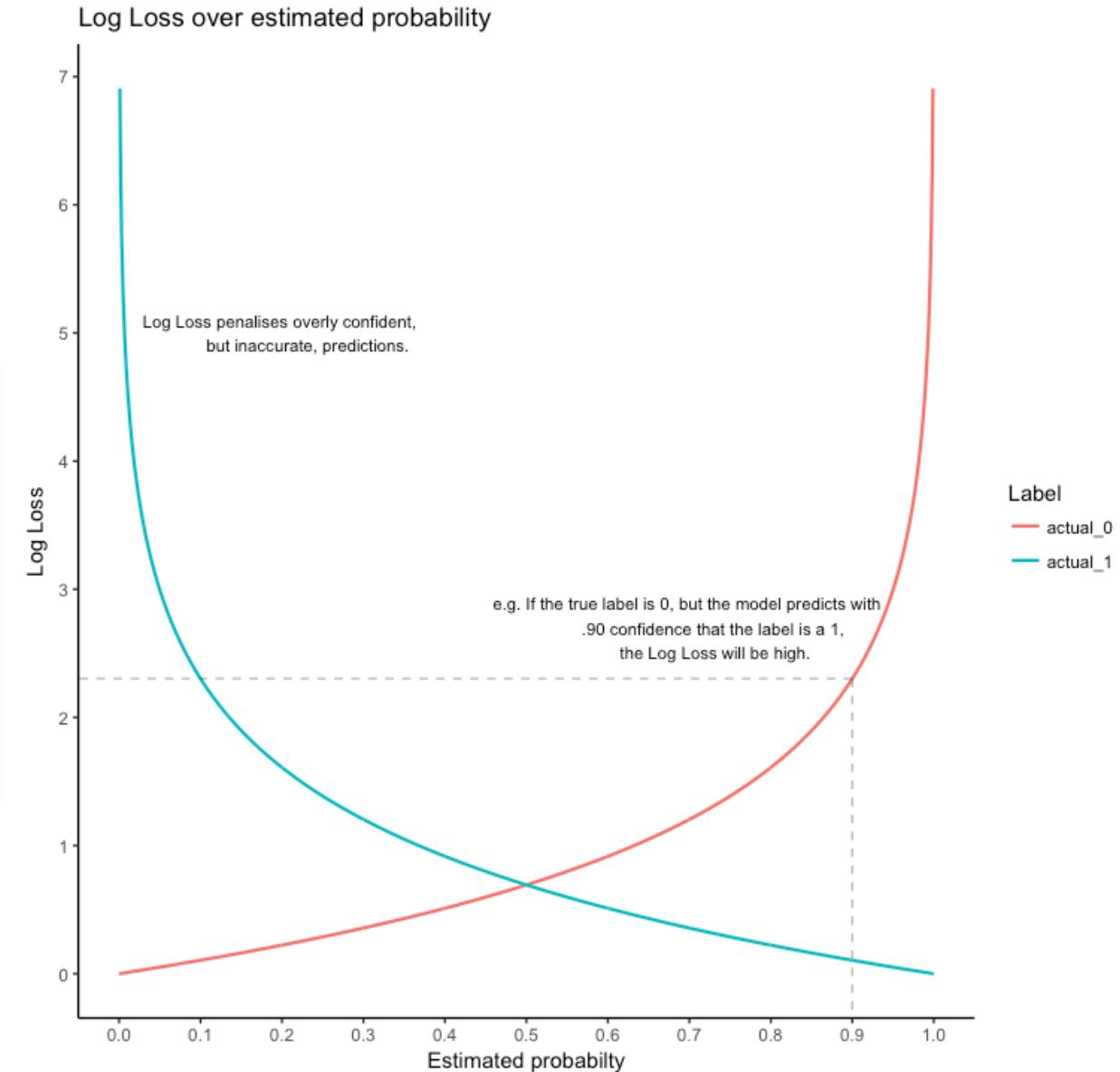
$$\text{LogLoss} = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

- Выводится из метода максимального правдоподобия
- В ходе построения алгоритма данное выражение минимизируется

LogLoss

При неправильном предсказании значение функционала уходит в бесконечность.

> Лечится ограничением значений функции.



Практическая часть

Открываем jupyter notebook)



Classification report

```
from sklearn.metrics import classification_report  
  
print(classification_report(y_test,y_hat_test))
```

		precision	recall	f1-score	support
	0	0.74	0.84	0.79	12733
	1	0.96	0.92	0.94	48532
micro avg		0.91	0.91	0.91	61265
macro avg		0.85	0.88	0.86	61265
weighted avg		0.91	0.91	0.91	61265

Оценка качества моделей



Как выбрать лучшую модель?

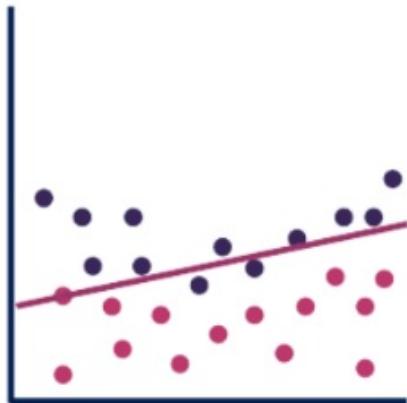
При построении модели есть 2 пути:

1

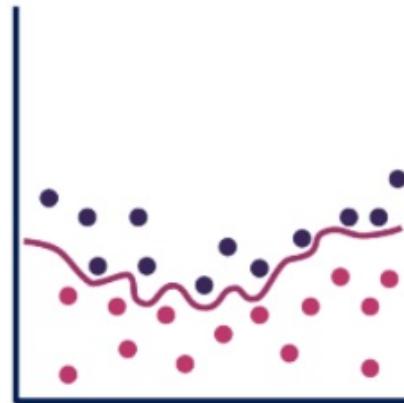
построить алгоритм улавливающий общие закономерности данных ([путь воина](#))

2

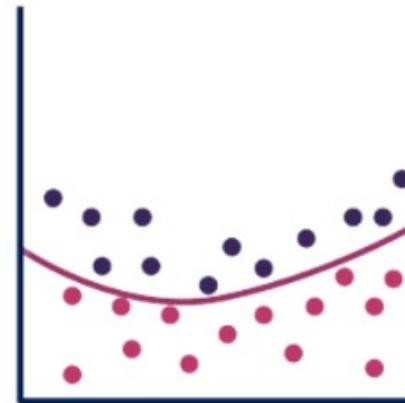
запомнить ответы обучающей выборки



Underfitting



Overfitting



Balanced

Как построить обучение?

Нужно разбить выборку на несколько частей

Train
(~70-80 % выборки)

Validation
(~10-20 % выборки)

Test
(все, что осталось)

Строим набор моделей-кандидатов

Выбираем модель

Проверяем, что все хорошо

Как делить выборку?

Shuffle and split

Перемешиваем и делим в желаемой пропорции.

Для возможности повторить результаты важно
фиксировать `random_state`.

[Пример использования](#)

Как делить выборку?

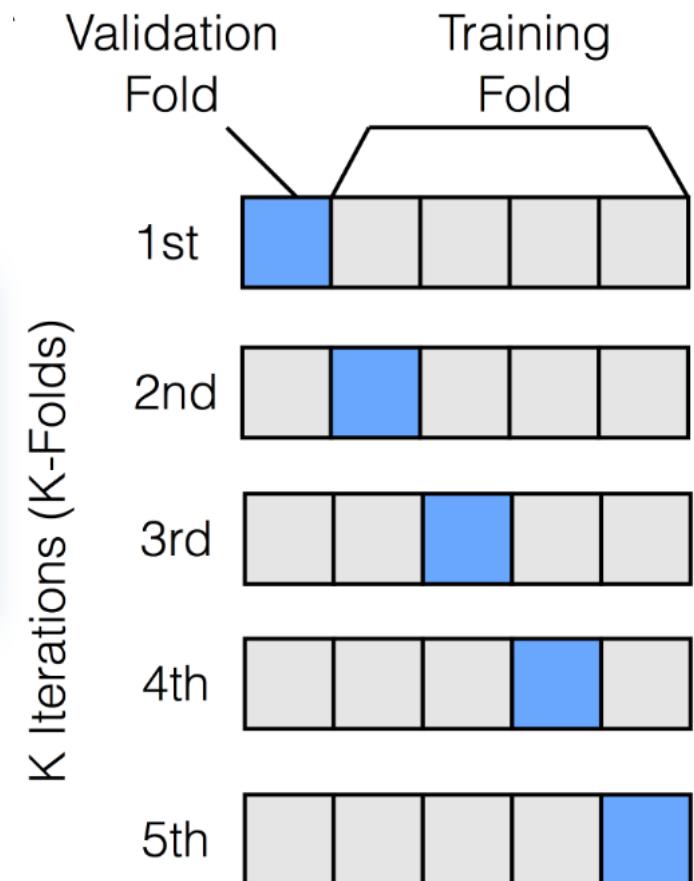
K-fold

Разбиваем датасет на K равных частей, затем строим K моделей где в качестве Test берем одну из частей, а все остальные используем как Train. На Train обучаем модель, на Test оцениваем качество.

Особенности:

- Используем все данные как для построения моделей, так и для оценки качества
- Один из наиболее популярных методов оценки качества моделей

Пример применения



Как делить выборку?

Leave-one-out

Экстремальный случай K-fold, когда К равно
числу сэмплов в наборе данных.

Особенности:

- Модель на датасете без одного сэмпла
практически идентична модели на полном
датасете
- Может быть эффективно посчитан для
некоторых видов моделей

Как делить выборку?

Repeated K-fold/Shuffle and split

Повторяем N раз построение модели с разными
разбиениями (можно просто менять `random_state`)

Особенности:

- В N раз увеличивает сложность построения
- Помогает на маленьких выборках

Как делить выборку?

Stratified K-fold/Shuffle and split

Сохраняем некоторые распределения из исходной выборки при разбиениях

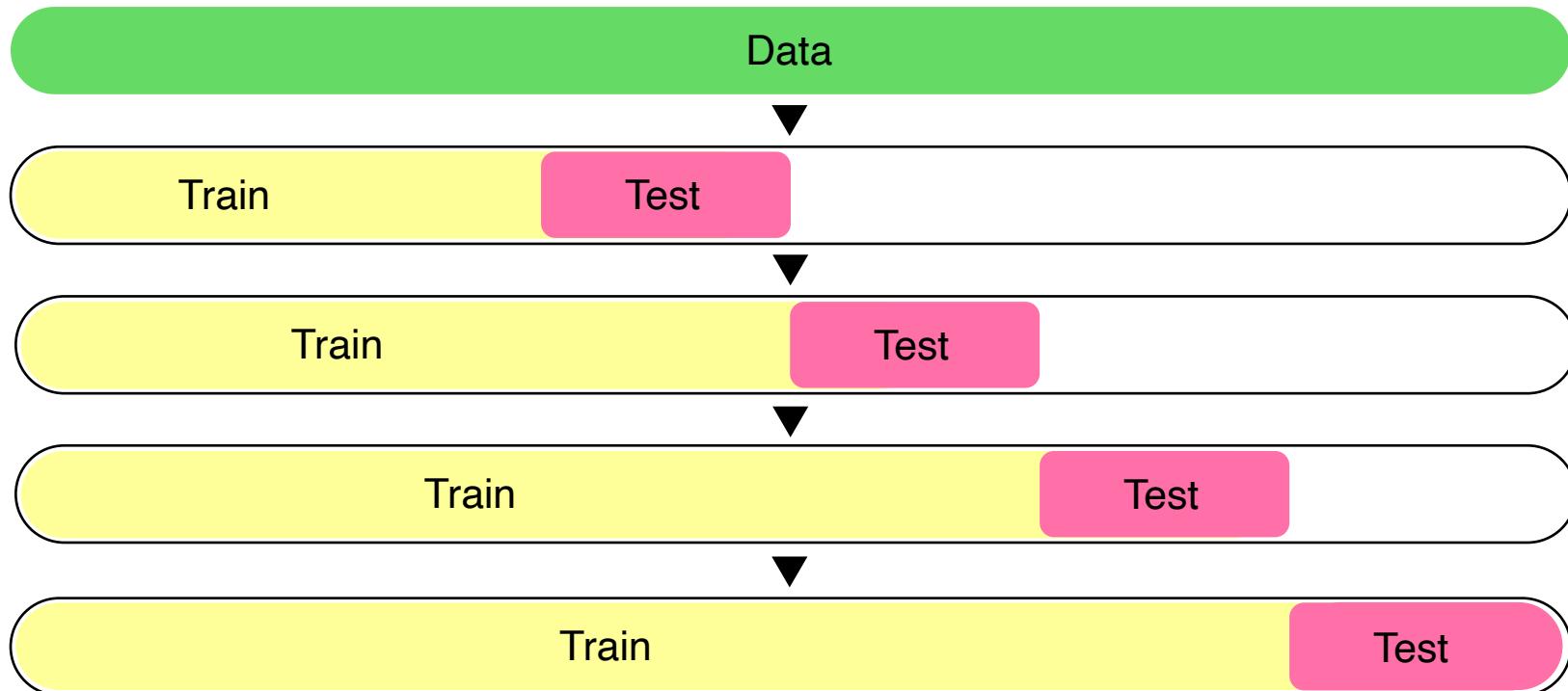
Например:

- Распределение признака (пол/возраст/дата и тд)
- Распределение таргета

Разбиение временных рядов

Разбиваем датасет на K частей, у каждой из которых test лежит правее на временной оси, чем train.

[Пример использования](#)



Работа с признаками

• • • •



Извлечение признаков

- File123.csv – сырой набор численных и не только признаков
- Sample – объект из обучающей выборки, представлен вектором
- Feature extraction – превращение данных, специфических для предметной области, в понятные для модели векторы

Типы признаков

- Бинарные – принимают значение 0/1 (флаг подключения услуги)
- Категориальные – их множество значений конечно
- Количественные – их множество значений вся числовая прямая
- Порядковые – их множество значений конечно и упорядочено

Примеры признаков

Числовые

Текстовые данные

Дата и время

Картинки

Геоданные

Звук

Временные ряды

Графы

Геоданные

- Количество объектов в некотором радиусе
- Расстояние до важных объектов
- Что находится по заданной координате
- Признаки маршрута между несколькими точками



education

Дата и время

- Абсолютное время
- Периодичность
- Временной интервал до особого события



Временные ряды

- Среднее значение/медиану/дисперсию за период
- Тренд за период
- Количество пиков
- Коэффициенты асимметрии и эксцесса



Извлечение признаков

Год выпуска	2011
Пробег	98 000 км
Кузов	Внедорожник 5 дв.
Цвет	Белый
Двигатель	6.2 л / 409 л.с. / бензин
Коробка	Автоматическая
Привод	Полный
Руль	Левый
Состояние	Не требует ремонта
Владельцы	3 владельца
ПТС	Оригинал
Владение	9 месяцев
Таможня	Растаможен
VIN	XWFS47EF*CO****62
Автокод	Без ограничений

[Характеристики модели в каталоге](#)



Ещё
2 фото

Преобразование признаков

• • • •



Зачем нужны преобразования?

1

Чтобы конкретный алгоритм машинного обучения их правильно интерпретировал

2

Чтобы конкретный алгоритм машинного обучения эффективно находил взаимосвязи

3

Чтобы внести априорные знания о наборе данных или свойствах признаков

Нормализация

Нормализация — это преобразование данных к неким безразмерным единицам. Иногда — в рамках заданного диапазона, например, [0..1] или [-1..1]. Иногда — с какими-то заданным свойством, как, например, стандартным отклонением равным 1.

Ключевая цель нормализации — приведение **различных данных** в самых разных единицах измерения и диапазонах значений **к единому виду**, который позволит сравнивать их между собой или использовать для расчёта схожести объектов. На практике это необходимо, например, для кластеризации и в некоторых алгоритмах машинного обучения.

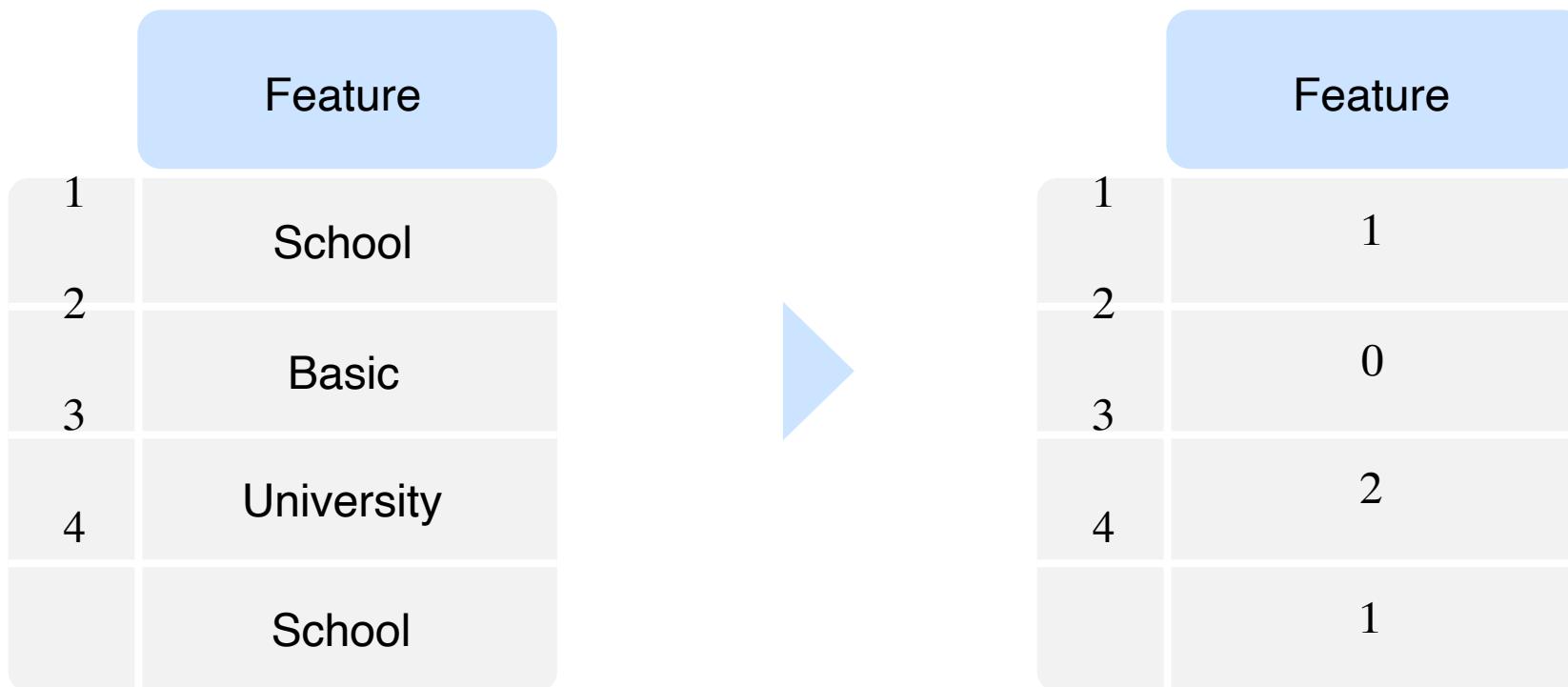
Примеры нормализации

Стандартизация	Масштабирование	Монотонные преобразования	Бинаризация
Для каждого признака в наборе вычитаем среднее и делим на стандартное отклонение.	Значения каждого признака в наборе приводят к диапазону [0,1].	Применение монотонного преобразования к признаку (например: логарифмирование, возведение в степень)	Область значений количественного или порядкового признака делим на N участков и представляем в виде N бинарных признаков.

Преобразование категориальных признаков

- Замена на порядковый признак
- One-hot encoding
- Binning and woe (только для бинарной классификации)
- Hashing trick

Замена порядковым признаком



One-hot-encoding

The diagram illustrates the process of one-hot encoding. On the left, a table shows a single categorical feature across five rows. A blue arrow points to the right, indicating the transformation into three binary features (F=School, F=Basic, F=University) shown in a second table.

Feature

	1	2	3	4	
	School				School
1					
2	Basic				
3					
4	University				

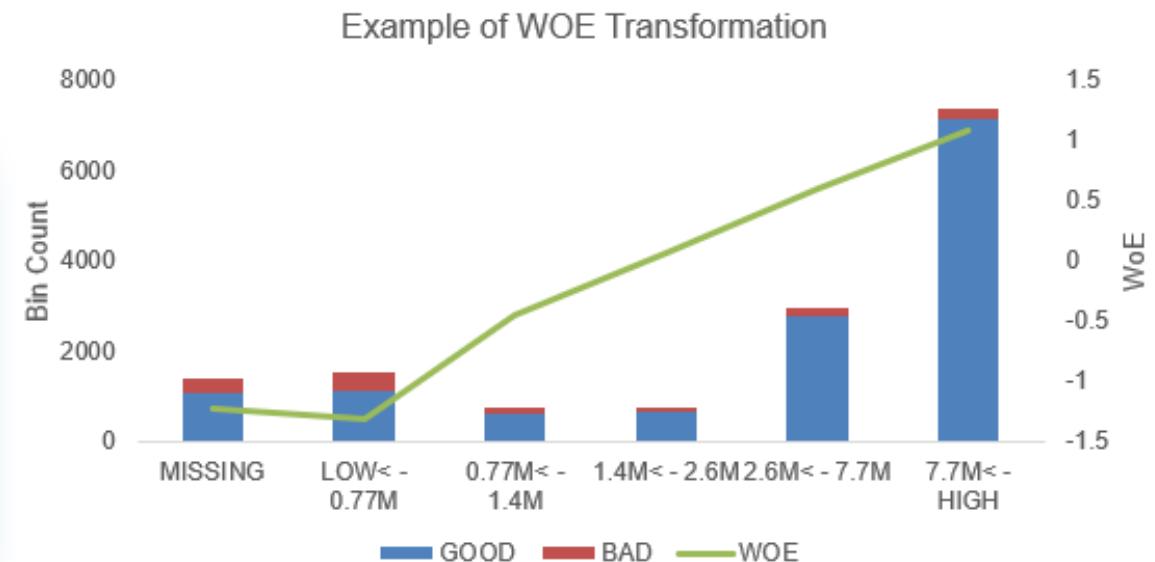
F=School F=Basic F=University

	1	2	3	4	
1	1		0	1	
2	0	1	0	0	
3	0	0	1		
4	1	0	0	0	

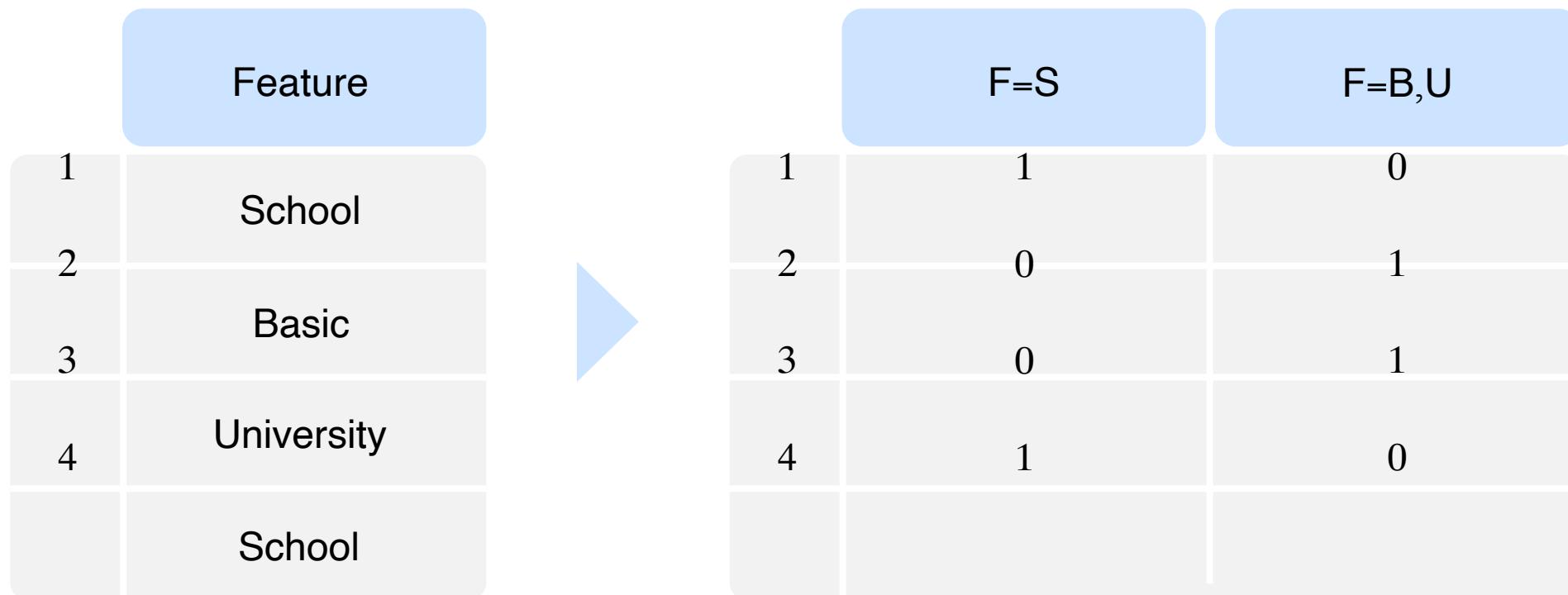
Binning and woe

$$WOE = \ln\left(\frac{\text{Event}\%}{\text{Non Event}\%}\right)$$

- Подсчет WOE каждой категории
- Объединение бинов с близким WOE для максимизации разницы между группами
- Замена значения признака на WOE в бине



Hashing trick



Задача

Алгоритм: k ближайших соседей с евклидовым расстоянием

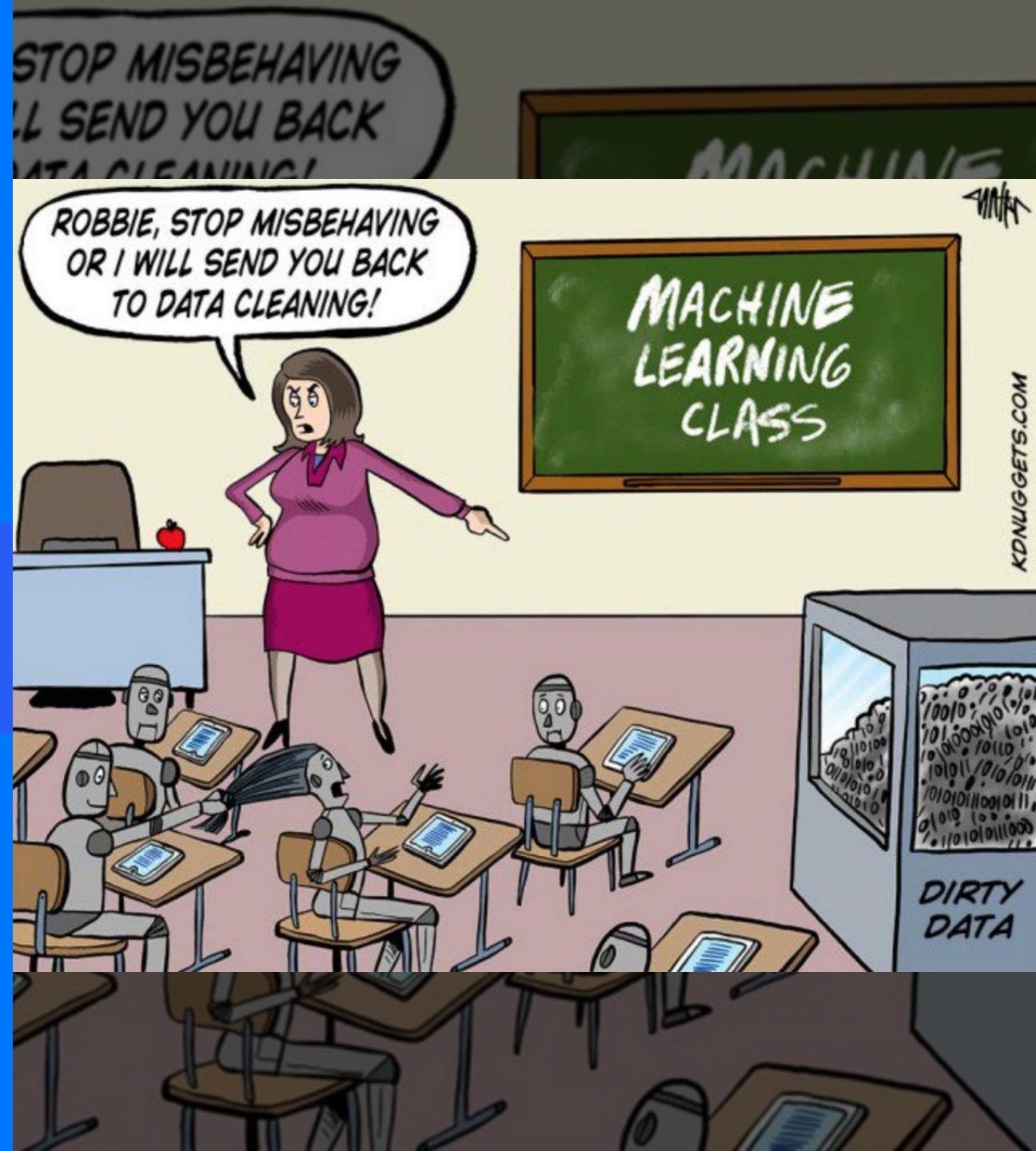
Признаки:

1. категория кинотеатра [1..43]
2. день недели [0..6]
3. час суток [0..23]
4. цена билета [100..1000]

Целевая переменная: заполненность зала в % Что делать?

Практическая часть

Открываем jupyter notebook)



Очистка данных

- Удаление или преобразование пропущенных значений
- Удаление дублей
- Удаление/обработка выбросов

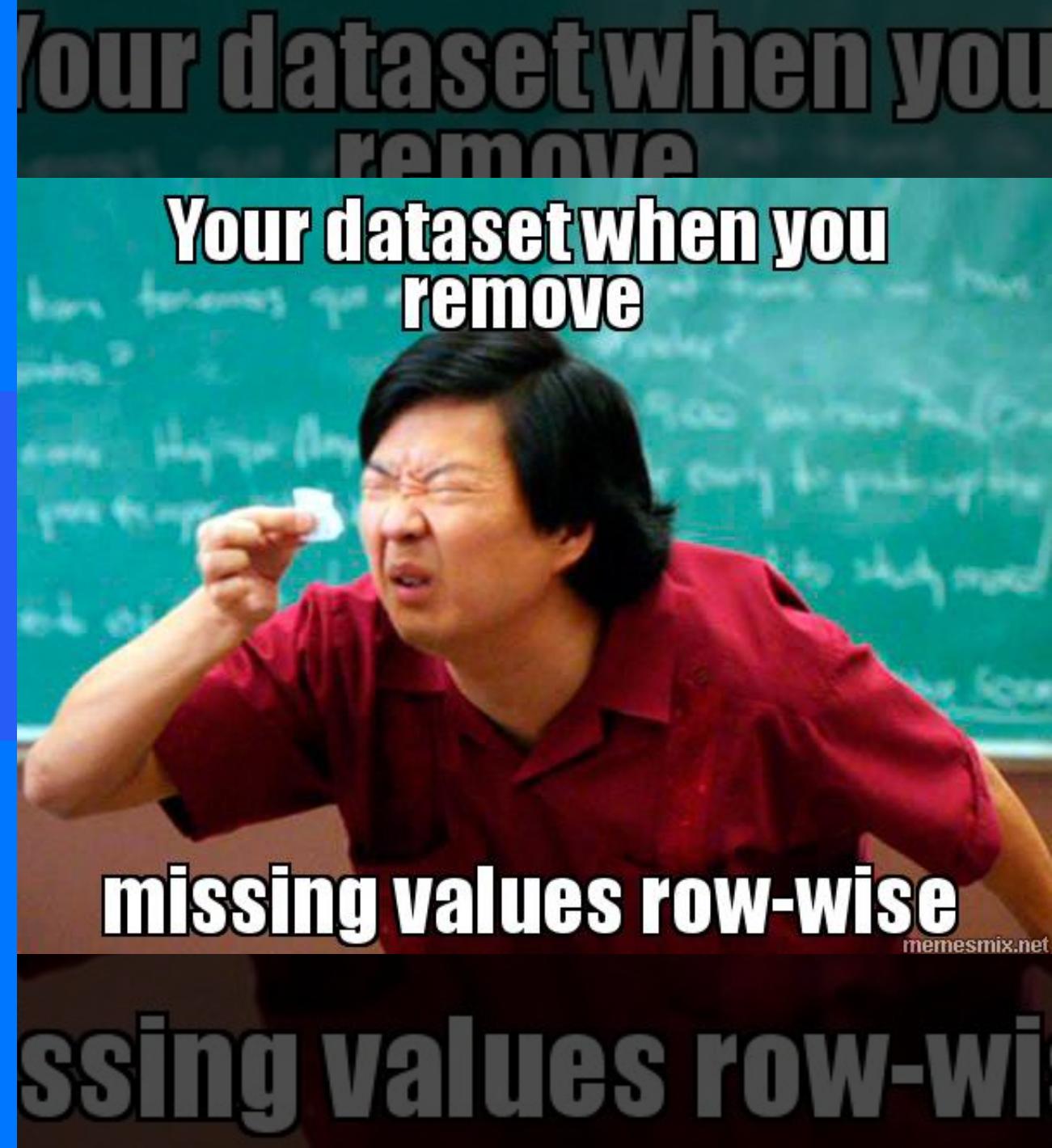


Пропуски

- Ошибки при записи
- Ошибки при сборке данных
- Невозможность сбора



education



Обработка пропусков

Замена на более вероятное значение

Замена на случайное значение
из распределение признака

Замена средним/медианой

Замена средним значением целевой
переменной на пропущенных значениях
признака

Отбор признаков

Выбор признаков, имеющих наиболее тесные взаимосвязи с целевой переменной

- Уменьшает переобучение
- Повышает точность
- Ускорение обучения

1

Однофакторный анализ

2

Рекурсивное исключение признаков

3

PCA

4

Отбор на основе важности признаков

Методы отбора признаков

Отбор признаков по взаимосвязи с целевой переменной. Например по качеству модели на 1 признаке, стабильности, корреляции с таргетом и тд.

1

Однофакторный анализ

2

Рекурсивное исключение признаков

3

PCA

4

Отбор на основе важности признаков

Методы отбора признаков

Отбор признаков с помощью recursive feature elimination.

1

Однофакторный анализ

2

Рекурсивное исключение признаков

3

PCA

4

Отбор на основе важности признаков

Методы отбора признаков

Метод главных компонент позволяет уменьшить количество используемых признаков с помощью взятия проекций пространства признаков на подпространство меньшей размерности.

1

Однофакторный анализ

2

Рекурсивное исключение признаков

3

PCA

4

Отбор на основе важности признаков

Методы отбора признаков

По коэффициентам построенной модели/по feature importance некоторых алгоритмов.

1

Однофакторный анализ

2

Рекурсивное исключение признаков

3

PCA

4

Отбор на основе важности признаков

Слайд для ваших вопросов

**THANK YOU FOR YOUR ATTENTION,
YOU CAN CLAP NOW**



**IF YOU HAVE ANY QUESTIONS, PLEASE
ASK MY FRIEND GOOGLE**

Слайд для вашего отзыва
;)

Поставьте хорошую
Оценку, пожалуйста

