



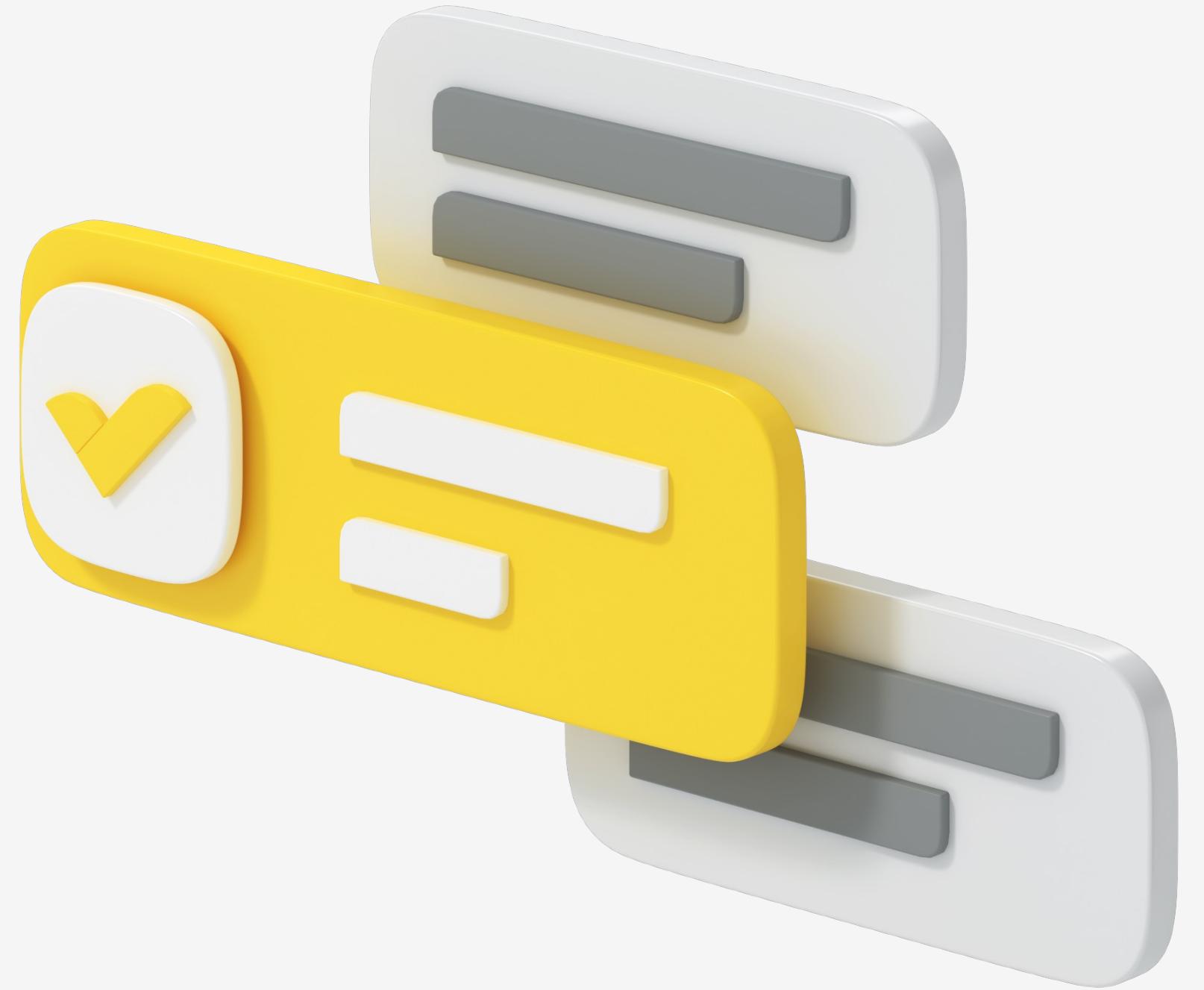
ТИНЬКОФФ

Лекция 11

Признаки: подготовка и отбор

Иван Карпухин

Ведущий исследователь-разработчик





Карпухин Иван

Ведущий исследователь-разработчик

@ i.a.karpukhin@tinkoff.ru

Профессионально занимаюсь машинным обучением
более 8 лет. Опыт:

- Исследования в области computer vision
- Голосовые технологии
- Распознавание лиц и текстов
- Говорящие головы
- Оптимизация NN

Цели вебинара

01

Научиться работать
с признаками разной
природы

02

Научиться адаптировать
признаки для ML
алгоритмов

03

Получить
практический опыт
отбора признаков

Маршрут вебинара

01



Подготовка
данных

02

Генерация
признаков

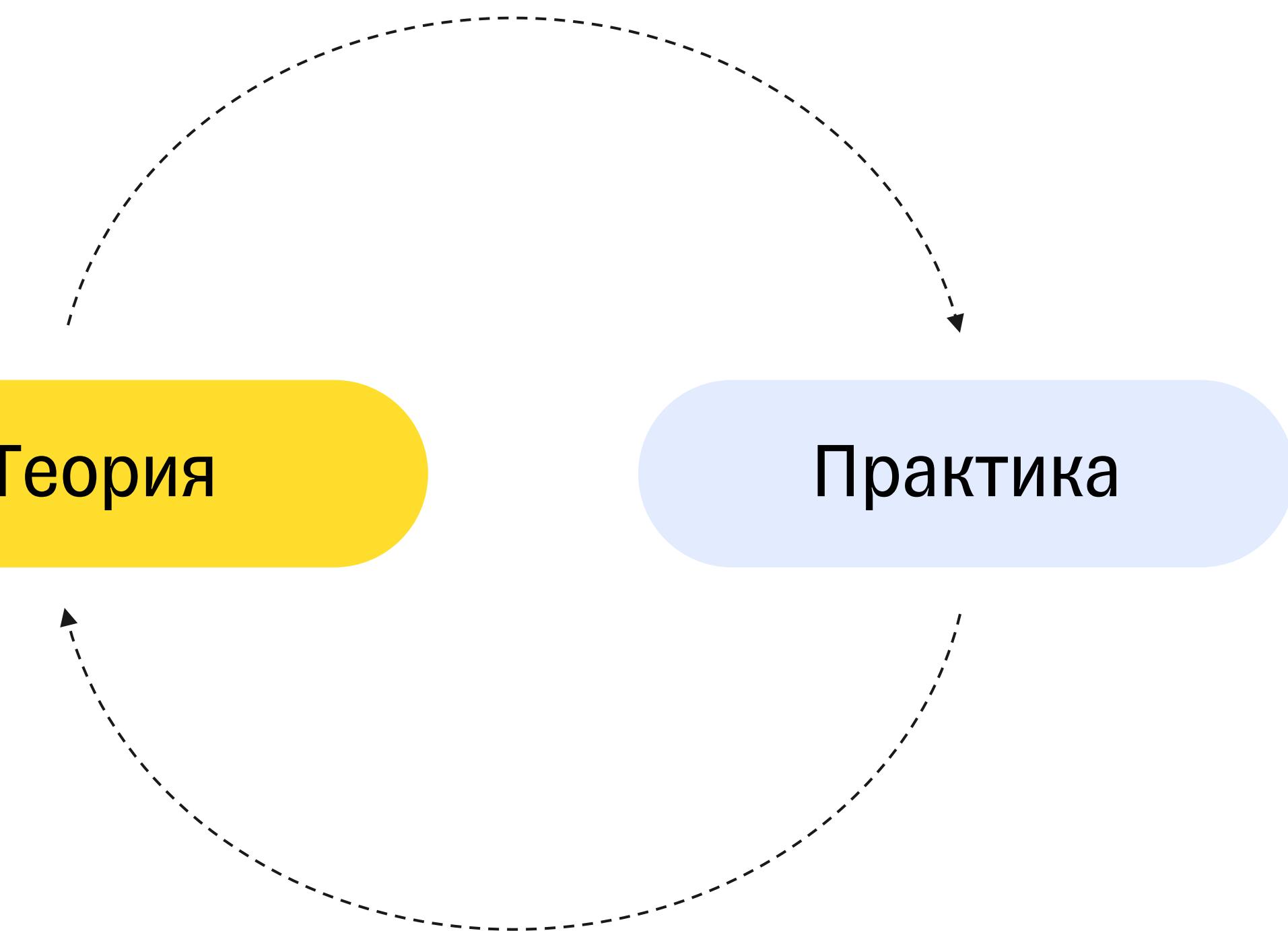
03

Отбор
признаков

формат

Теория

Практика



формат: практика



Тесты по QR или ссылкам



Задания “на бумаге”



Задания по программированию



ТИНЬКОФФ

Подготовка данных

Задание

Тест по основам pandas



Ссылка в чате



5 минут



Анонимно

Пропуски в данных

	state	county	community		communityname	fold	population
0	8	NaN		NaN	Lakewoodcity	1	0.19
1	53	NaN		NaN	NaN	1	0.00
2	24	NaN		NaN	Aberdeentown	1	0.00
3	34	5.0	81440.0	Willingborotownship		1	0.04
4	42	95.0	6096.0	Bethlehemptownship		1	0.01

Числовые признаки

Что можно сделать с числовыми признаками?

	state	county	community	communityname	fold	population
0	8	NaN	NaN	Lakewoodcity	1	0.19
1	53	NaN	NaN	NaN	1	0.00
2	24	NaN	NaN	Aberdeentown	1	0.00
3	34	5.0	81440.0	Willingborotownship	1	0.04
4	42	95.0	6096.0	Bethlehemptownship	1	0.01

Числовые признаки

Что можно сделать с числовыми признаками?



Удалить

- Признаки
- Элементы

Числовые признаки

Что можно сделать с числовыми признаками?



Удалить



Что-то подставить (англ. Imputation)

- Признаки
 - Элементы
- Среднее / медиану / моду
 - Обучить регрессию на элементах без пропусков и предсказать

Категориальные признаки

Что можно сделать с категориальными признаками?

	state	county	community	communityname	fold	population
0	8	NaN	NaN	Lakewoodcity	1	0.19
1	53	NaN	NaN	NaN	1	0.00
2	24	NaN	NaN	Aberdeentown	1	0.00
3	34	5.0	81440.0	Willingborotownship	1	0.04
4	42	95.0	6096.0	Bethlehemptownship	1	0.01

Категориальные признаки

Что можно сделать с категориальными признаками?



Заменить на самое
частое значение (мода)



Ввести специальную
категорию “n/a”



Обучить модель
на данных без пропусков
и предсказать

Для любых признаков

Можно добавить бинарный признак “column_x_is_missing”

	state	county	county_is_missing	communityname
0	8	NaN	True	Lakewoodcity
1	53	NaN	True	Tukwilacity
2	24	NaN	True	Aberdeentown
3	34	5.0	False	Willingborotownship
4	42	95.0	False	Bethlehemtownship

Кодирование категориальных признаков

Большинство алгоритмов работает с числами. Как закодировать строки?

	age	job	marital	education	default	balance
13932	57	admin.	divorced	secondary	no	658.00000
9894	37	blue-collar	married	secondary	no	1362.26877
39946	35	technician	divorced	secondary	no	2823.00000
9217	35	admin.	married	secondary	no	1362.26877
4124	38	services	single	tertiary	no	1362.26877

Кодирование категориальных признаков

Если возможных значений очень много, то NLP)

Кодирование категориальных признаков

Если возможных значений очень много, то NLP)

Иначе:

- One-hot кодирование

Признак
Дом
Самолет
Дом



Дом	Машина	Самолет
1	0	0
0	0	1
1	0	0

Кодирование категориальных признаков

Если возможных значений очень много, то NLP)

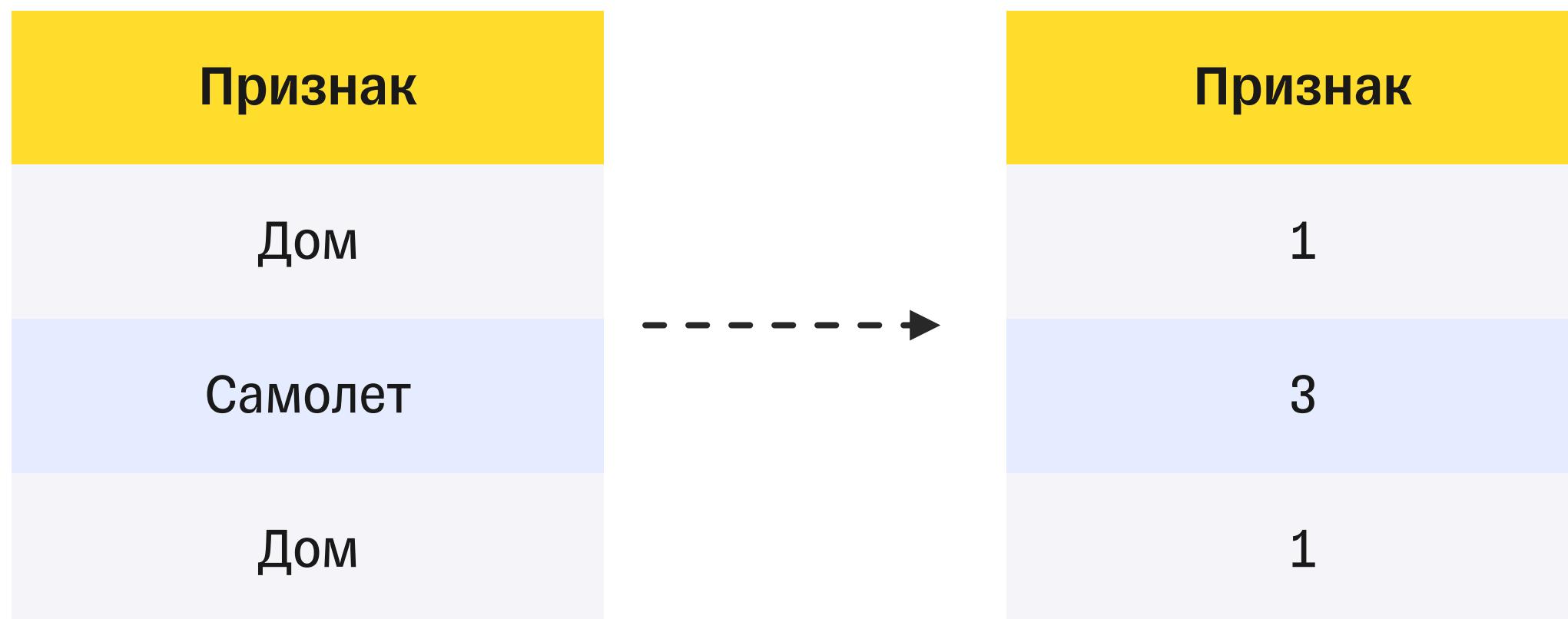
Иначе:

- One-hot кодирование
- Сгруппировать числовые признаки по категориальным

Владение	Доход	Средний доход для владения	Доход
Дом	10	15	10
Самолет	1000	1000	1000
Дом	20	15	20

Кодирование категориальных признаков

Почему бы не отобразить категории в числа от 1 до N?



Кодирование времени и дат



Выделить числовой день недели (1-7)

7 июля 2022 → 4 (четверг)



Кодирование времени и дат



Выделить числовой день недели (1-7)

7 июля 2022 → 4 (четверг)



Бинарный признак выходного дня

7 июля 2022 → 0 (False)



Кодирование времени и дат



Выделить числовой день недели (1-7)

7 июля 2022 → 4 (четверг)



Бинарный признак выходного дня

7 июля 2022 → 0 (False)



Периодические признаки (\sin и \cos)

7 июля 2022 → $\cos(2\pi * 4 / 7)$, $\sin(2\pi * 4 / 7)$



Кодирование геоданных

Если даны широта и долгота

- Страна
- Город

```
In [137]: import reverse_geocoder as revgc  
revgc.search((40.74482, -73.94875))
```

```
Out[137]: [{"lat": "40.74482",  
            "lon": "-73.94875",  
            "name": "Long Island City",  
            "admin1": "New York",  
            "admin2": "Queens County",  
            "cc": "US"}]
```

Вопросы



Задание

Подготовить данные
для классификации



**Ссылка на ноутбук в чате
(1-data-preparation)**



30 минут



**Можно шарить экран
и задавать вопросы**

Перерыв

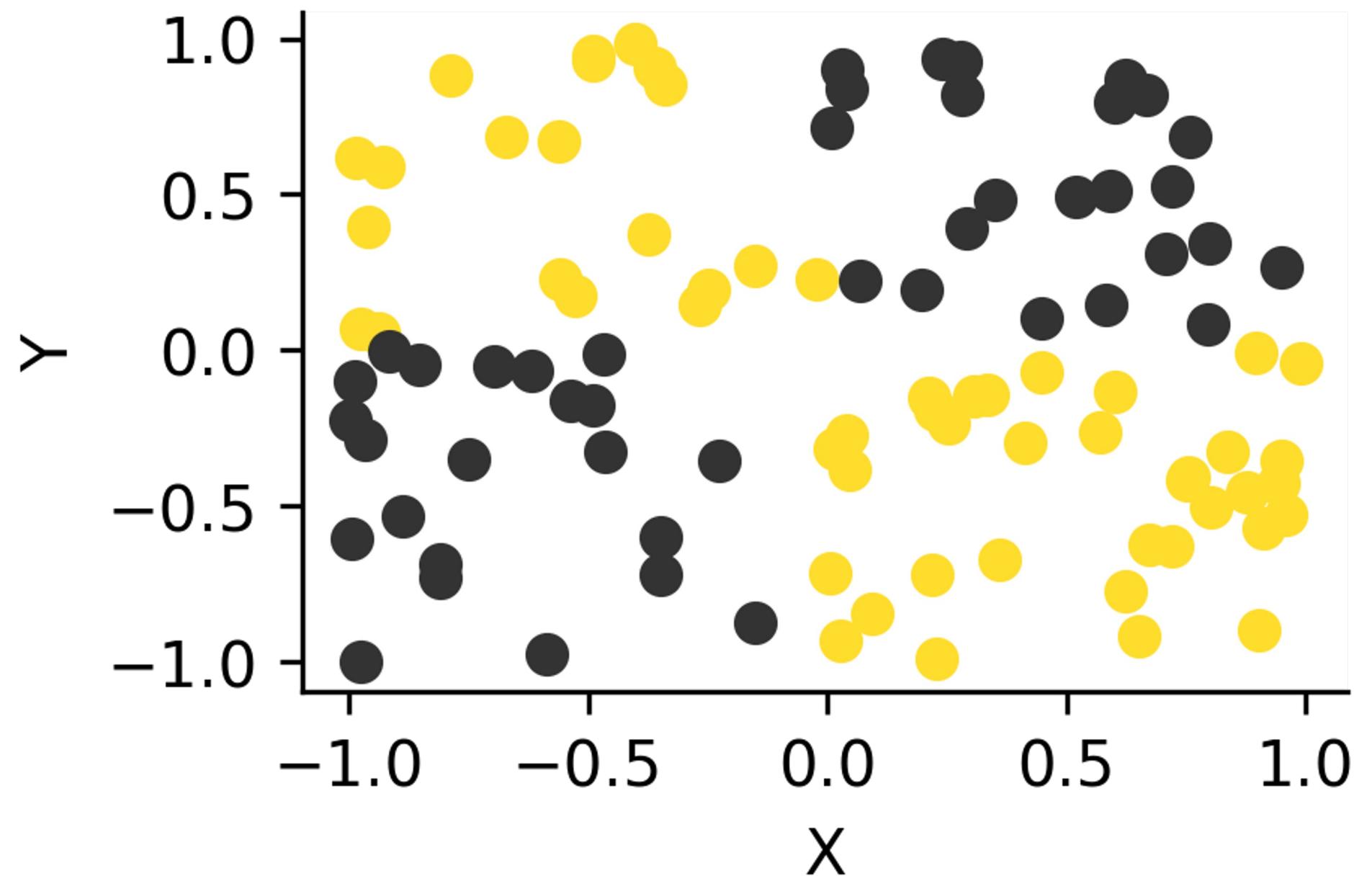


ТИНЬКОФФ

Генерация признаков

Мотивация

- Использование простых моделей
в том числе линейных



Мотивация

- Использование простых моделей
в том числе линейных
- Внесение априорных знаний



$$a = \frac{F}{m}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Примеры

- Полиномы

XOR и линейные модели

$$\{x, y\} \rightarrow \{x, y, x^2, xy, \frac{x}{y}, y^3\}$$

Примеры

- Полиномы

XOR и линейные модели

$$\{x, y\} \rightarrow \{x, y, x^2, xy, \frac{x}{y}, y^3\}$$

Почему не добавляем
суммы?



Примеры

- Полиномы
- Периодические функции

[Даты / время](#)

$$\sin\left(\frac{2\pi x}{T}\right)$$

$$\cos\left(\frac{2\pi x}{T}\right)$$



Примеры

- Полиномы
- Периодические функции
- Ограничения

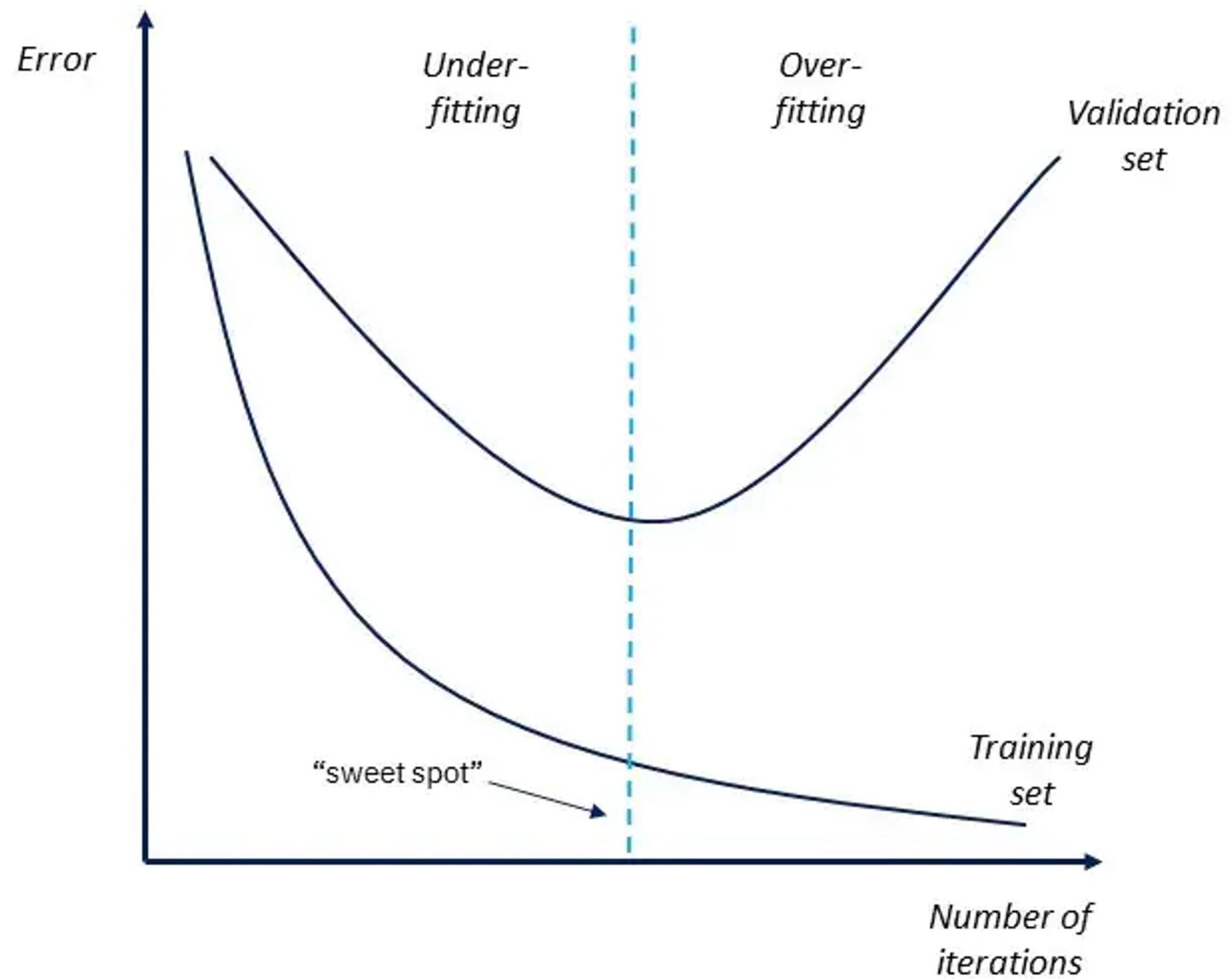
[Баланс кредитной карты](#)



$$\max(x, 0)$$

Примеры

- Полиномы
- Периодические функции
- Ограничения
- ...



Обычно больше – лучше,
но следите за переобучением

Нормировка признаков

**Обучение моделей зависит
от масштаба признаков**

Пример – регуляризация

$$\mathcal{L}(\theta) = Err(\theta) + \frac{1}{2} \|\theta\|^2$$

Масштаб разных признаков
разный



Нормировка признаков



**Обучение моделей зависит
от масштаба признаков**

Пример – регуляризация

$$\mathcal{L}(\theta) = Err(\theta) + \frac{1}{2} \|\theta\|^2$$

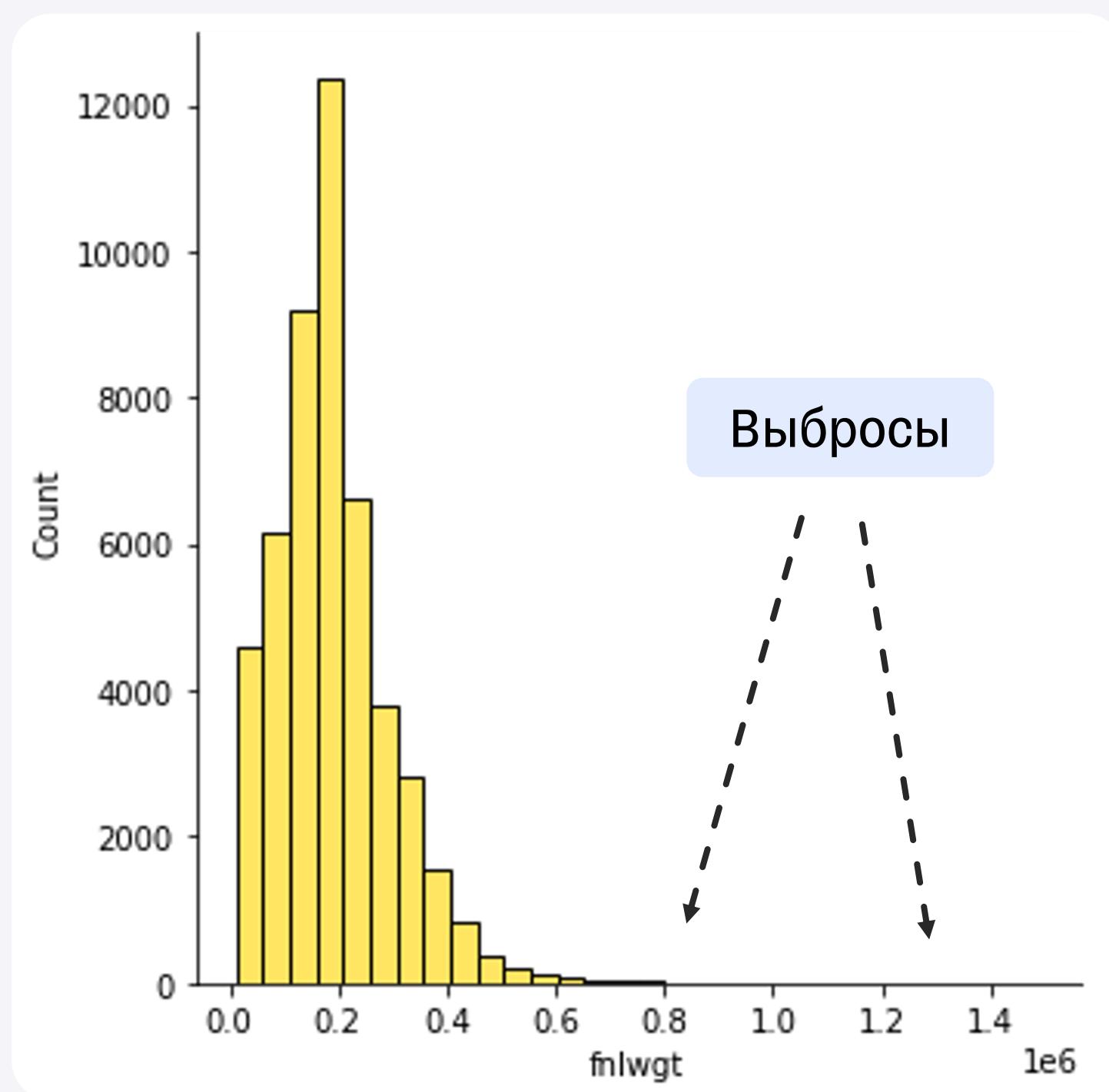
Масштаб разных признаков
разный

Решение – нормировка:

$$x_{new} = \frac{x - \bar{E}x}{\sqrt{\text{Var } x}}$$

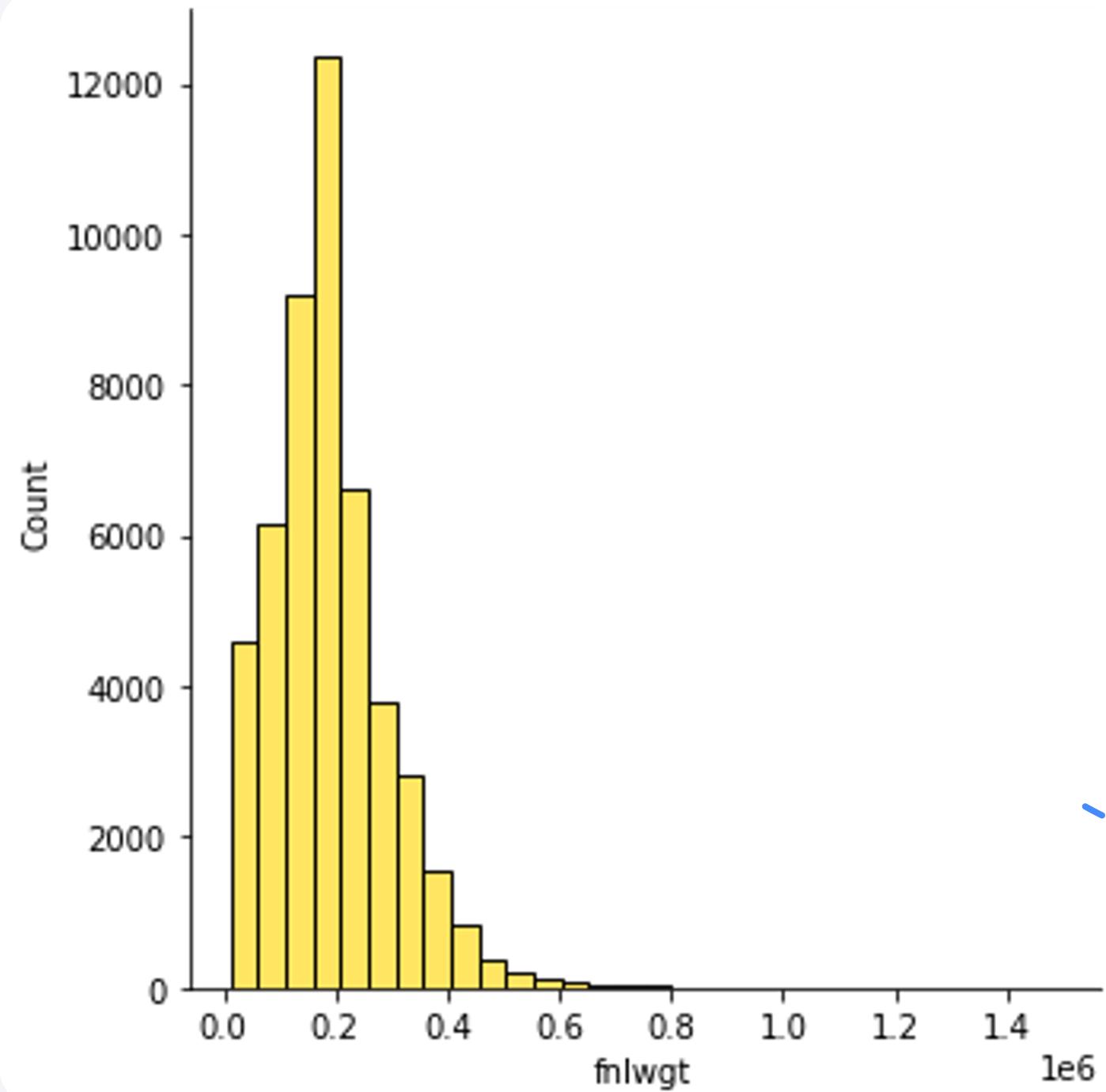
Отсечение выбросов

Признаки для некоторых элементов могут быть слишком большими или слишком маленькими



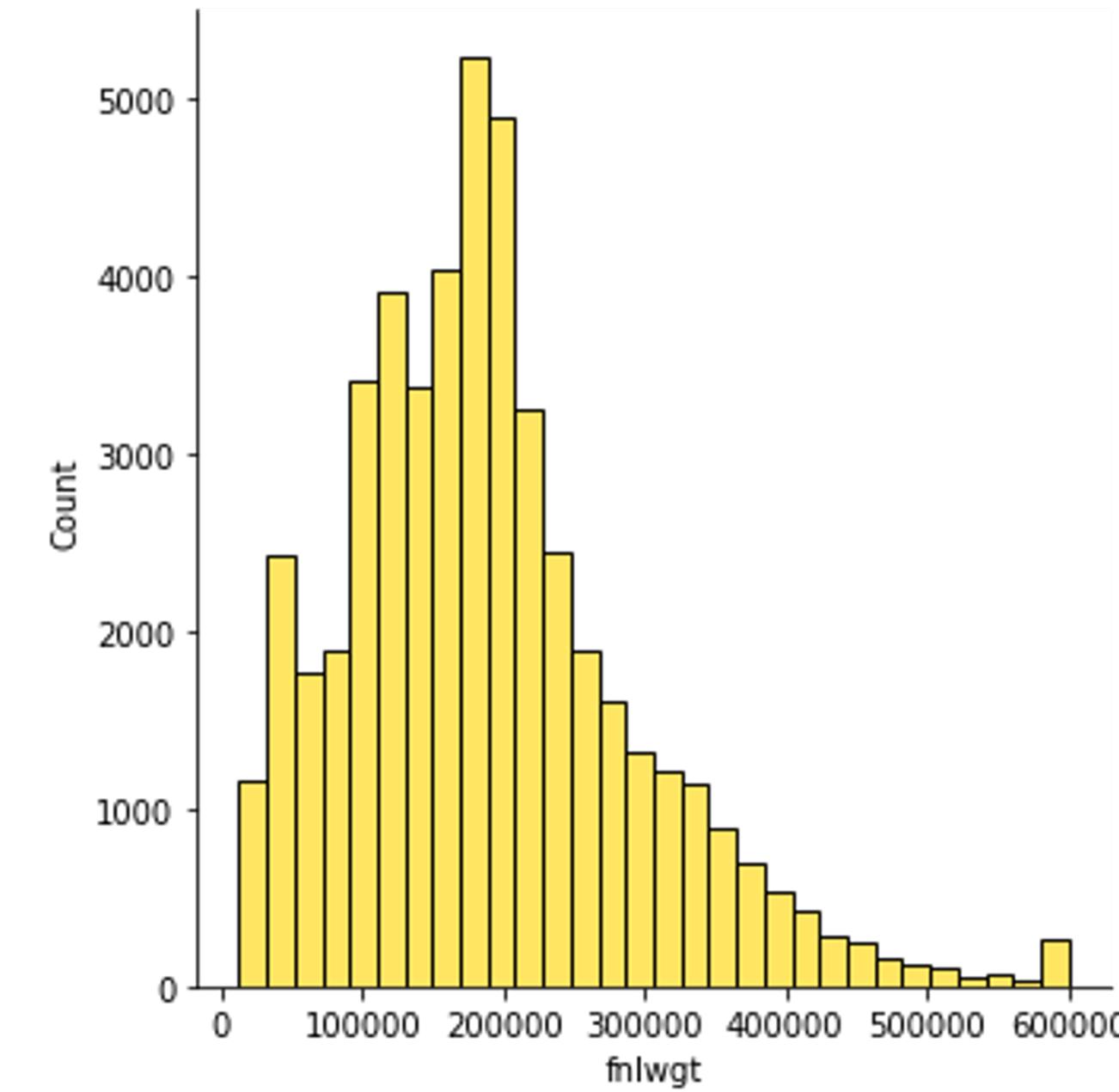
Отсечение выбросов

Признаки для некоторых элементов могут быть слишком большими или слишком маленькими



Решение

Ограничить диапазон значений



Вопросы



Задание

Улучшить качество путем
добавления признаков



**Ссылка на ноутбук в чате
(2-feature-engineering)**



20 минут



**Можно шарить экран
и задавать вопросы**



ТИНЬКОФФ

Отбор признаков

Мотивация



Признаков может
быть много

Какие размечать в будущем?

	age	balance	day	month	duration	campaign	pdays	previous	job_admin.	job_blue-collar	...	education_secondary	education_tertiary
13932	57	658.00000	10	jul	724	1	-1	0	1	0	...	1	0
9894	37	1362.26877	9	jun	63	1	-1	0	0	1	...	1	0
39946	35	2823.00000	2	jun	102	4	96	2	0	0	...	1	0
9217	35	1362.26877	5	jun	247	1	-1	0	1	0	...	1	0
4124	38	1362.26877	19	may	138	1	-1	0	0	0	...	0	1

Мотивация



Признаков может быть много

Какие размечать в будущем?



Удалив шум
повысим точность

	age	balance	day	month	duration	campaign	pdays	previous	job_admin.	job_blue-collar	...	education_secondary	education_tertiary
13932	57	658.00000	10	jul	724	1	-1	0	1	0	...	1	0
9894	37	1362.26877	9	jun	63	1	-1	0	0	1	...	1	0
39946	35	2823.00000	2	jun	102	4	96	2	0	0	...	1	0
9217	35	1362.26877	5	jun	247	1	-1	0	1	0	...	1	0
4124	38	1362.26877	19	may	138	1	-1	0	0	0	...	0	1

Мотивация



Признаков может быть много

Какие размечать в будущем?



Удалив шум повысим точность



Почему модель предсказывает то, что предсказывает?

	age	balance	day	month	duration	campaign	pdays	previous	job_admin.	job_blue-collar	...	education_secondary	education_tertiary
13932	57	658.00000	10	jul	724	1	-1	0	1	0	...	1	0
9894	37	1362.26877	9	jun	63	1	-1	0	0	1	...	1	0
39946	35	2823.00000	2	jun	102	4	96	2	0	0	...	1	0
9217	35	1362.26877	5	jun	247	1	-1	0	1	0	...	1	0
4124	38	1362.26877	19	may	138	1	-1	0	0	0	...	0	1

Подходы

- ➡ Анализ вариативности признаков
- ➡ Анализ признаков с учетом меток
- ➡ Анализ признаков с учетом модели



Анализ разброса

Можно оценить разброс
признака через дисперсию

$$s = \text{Var } x = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Анализ разброса

Можно оценить разброс признака через дисперсию

$$s^2 = \text{Var } x = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

→ **Варианты:**

Дисперсия равна 0 →
признак одинаковый
для всех примеров

Дисперсия мала →
признак почти не меняется,
или проблема с масштабом

Дисперсия большая →
признак меняется сильно

ANOVA

ANalysis Of VAriance (или дисперсионный анализ)

Признак X1 для элементов разных классов

Класс 1	Класс 2	Класс 3
10	6	5
5	8	11
4	12	7

- Большой разброс внутри каждого класса
- Примерно одинаковый разброс между классами
- Признак менее важный

ANOVA

ANalysis Of VAriance (или дисперсионный анализ)

Признак X1 для элементов разных классов

Класс 1	Класс 2	Класс 3
10	6	5
5	8	11
4	12	7

Признак X2 для элементов разных классов

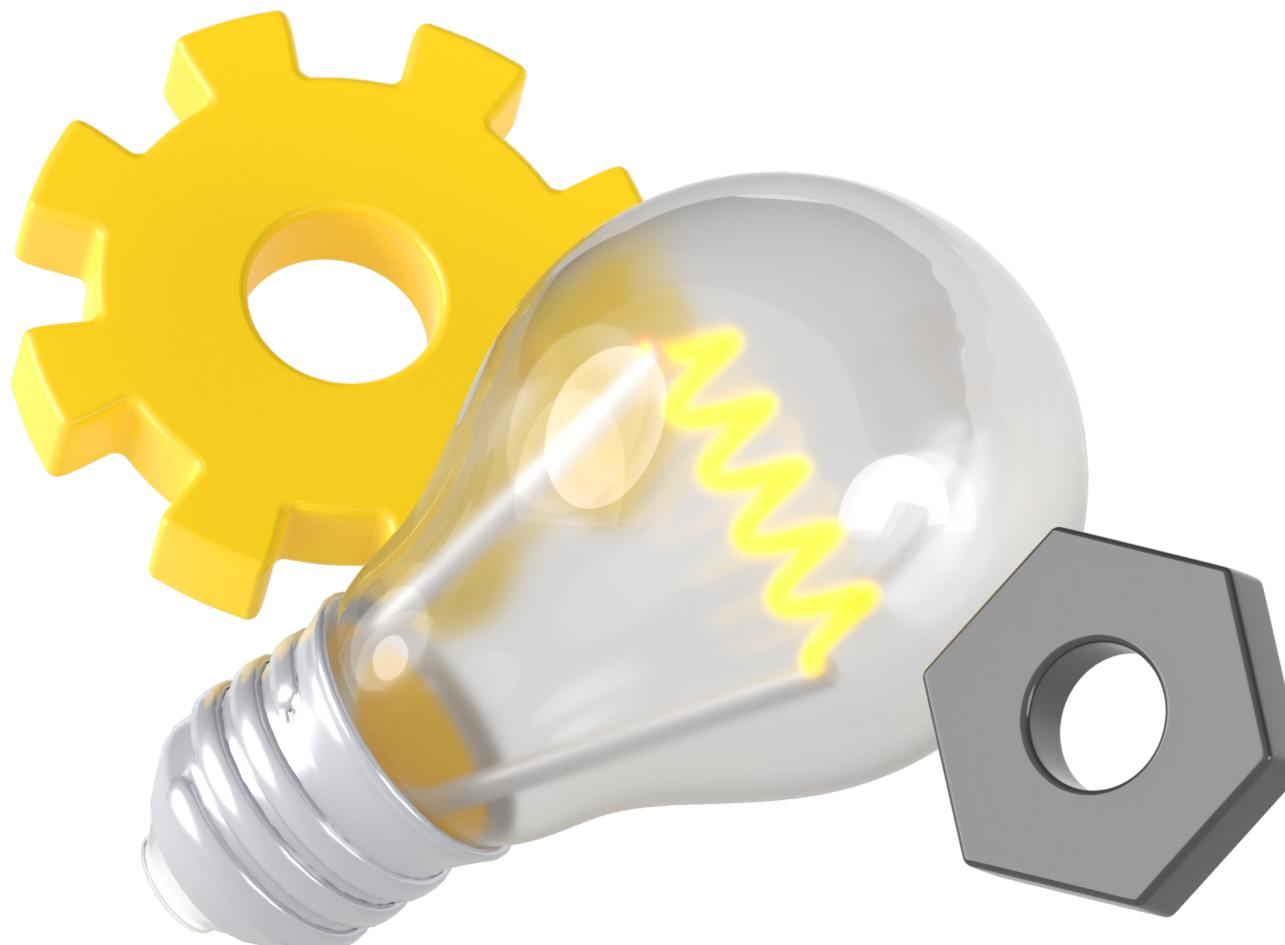
Класс 1	Класс 2	Класс 3
10	5	12
11	3	15
9	4	13

- Большой разброс внутри каждого класса
- Примерно одинаковый разброс между классами
- Признак менее важный

- Маленький разброс внутри каждого класса
- Большой разброс между классами
- Признак более важный

ANOVA

ANalysis Of VAriance (или дисперсионный анализ)



Идея:

Сравнить разброс признаков внутри каждого класса с общим разбросом признака в корпусе

Для ANOVA нужны метки классов

Жадный алгоритм



1. Начать с пустого множества признаков
2. Добавить по очереди каждый признак
3. Выбрать признак, который дает макс. качество
4. Повторить с 2.

Жадный алгоритм



1. Начать с пустого множества признаков
2. Добавить по очереди каждый признак
3. Выбрать признак, который дает макс. качество
4. Повторить с 2.

Проблемы:

- Долго, если данных много
- Зависит от конкретной модели

Линейные модели

Предсказание:

$$y = w_0 + x_1w_1 + x_2w_2 + \dots x_nw_n$$

Модуль весов говорит о значимости признаков?



Линейные модели

Предсказание:

$$y = w_0 + x_1w_1 + x_2w_2 + \dots x_nw_n$$

Модуль весов говорит о значимости признаков?

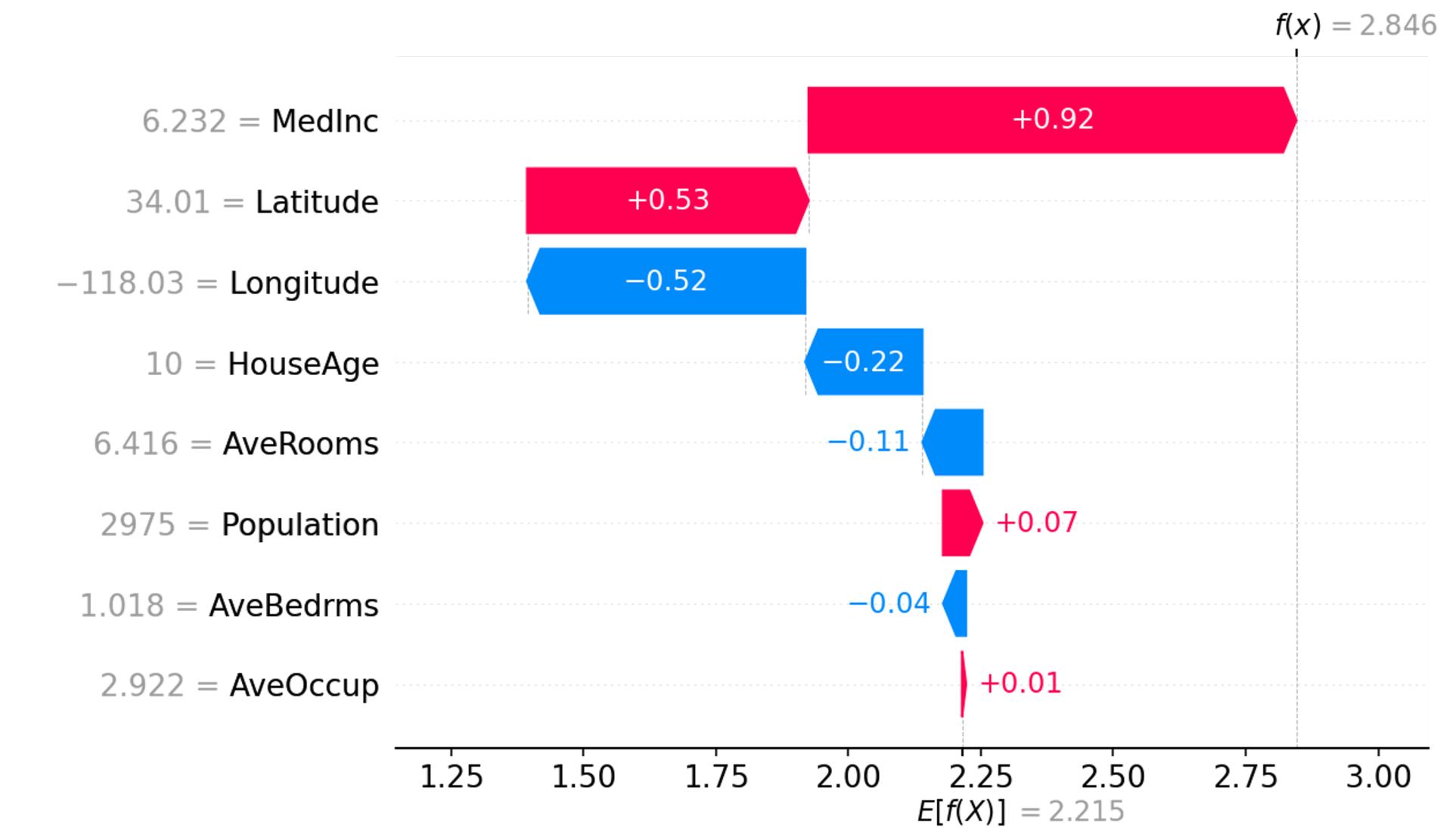
Говорит, если масштаб признаков совпадает

Интерпретация ответа при помощи SHAP

Shapley Additive Explanation Values

Позволяет оценить вклад признаков
в предсказание модели

- Зависит от модели
- Зависит от примера
- Значимость признаков суммируется в 1



Вопросы



Демо ноутбук

3-feature-selection.ipynb





ТИНЬКОФФ

Резюме

Что мы изучили



Способы подготовки
данных



Способы генерации
полезных признаков



Отбор лучших
признаков и оценка
полезности

Ссылки

Пропуски в данных:

<https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

Отбор признаков методом ANOVA:

<https://youtu.be/lTf4vHhyGrc>
https://youtu.be/q48uKU_KWas

Отбор признаков для обученной модели:

<https://habr.com/ru/company/ods/blog/599573/>



ТИНЬКОФФ

Q & A

Спасибо ;)



ТИНЬКОФФ

Он такой один