Phystech@DataScience

Блок 4: нелинейные модели

Валидация моделей и подбор гиперпараметров

На чем валидируем?

При оценке качества модели нельзя использовать данные, которые использовались для ее обучения.

Разбиваем данные случайно на две части

1. **Обучающие**
   обучаем все наши модели
2. **Валидационные**
   вычисляем метрики качества

Недостатки:

1. Результат сильно зависит от способа разбиения
2. При обучении модели часть данных совсем не используется
3. Небольшое переобучение под валидационную выборку

Как перебираем значения?

Методы оптимизации?

Градиенты по сложным метрикам и моделям не возьмешь.

Остается воспользоваться перебором.

Проводим итерации:

1. Выбрать значение гиперпараметра
2. Посчитать для него метрику качества

Поиск по сетке

Grid Search

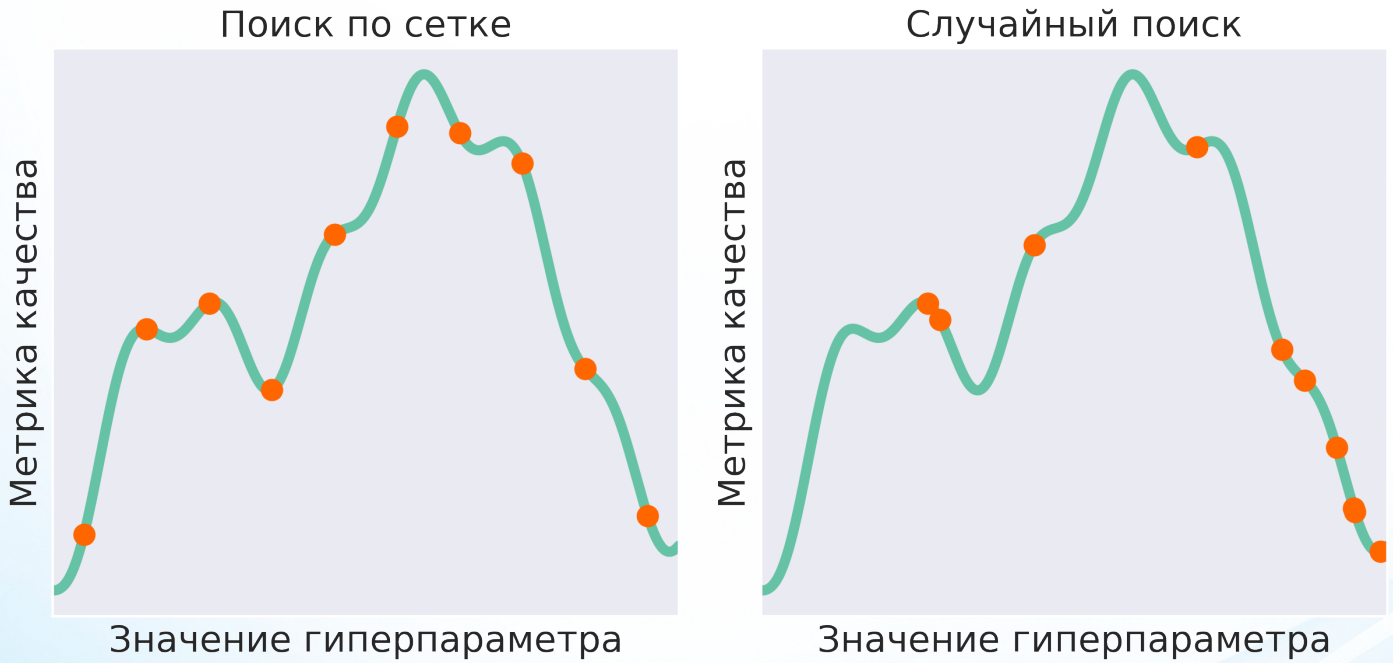
Берем равномерную сетку значений гиперпараметра.

Случайный поиск

Random Search

Генерируем случайные гиперпараметры

Сравнение методов перебора значений



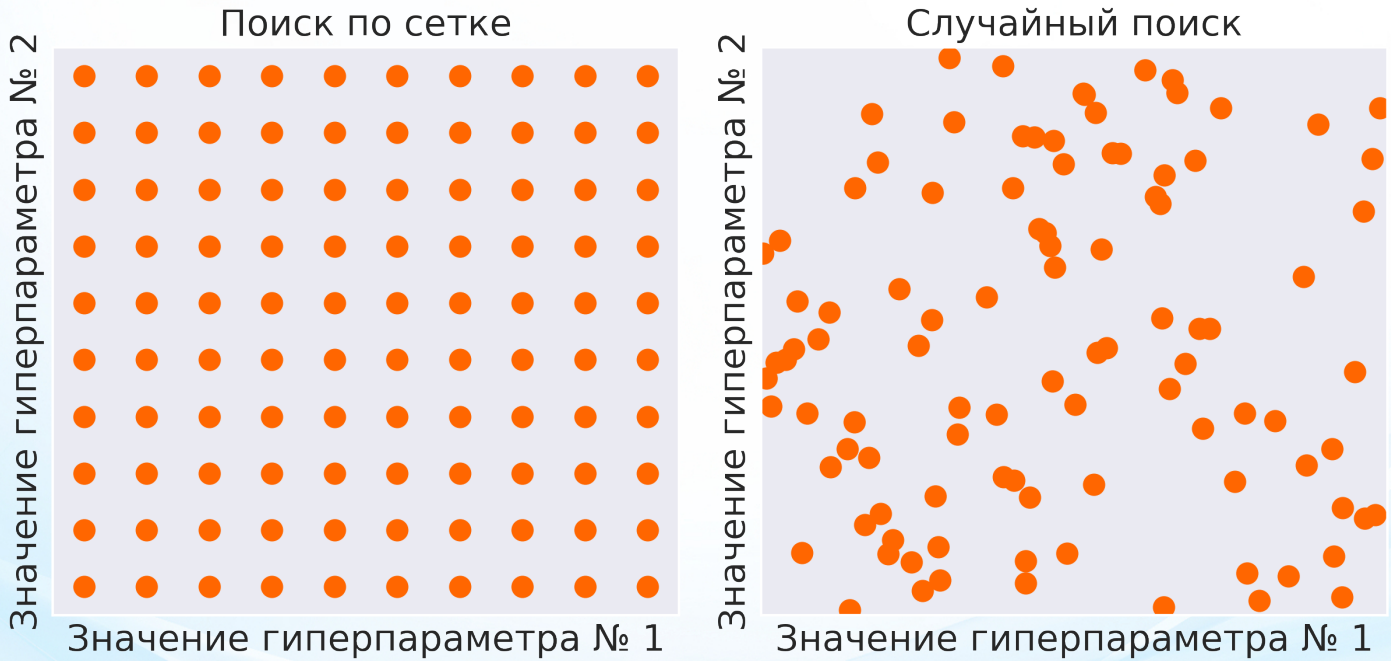
Какую стратегию лучше использовать?

Если параметр один, то Grid Search.

А если несколько гиперпараметров?

Многомерный случай

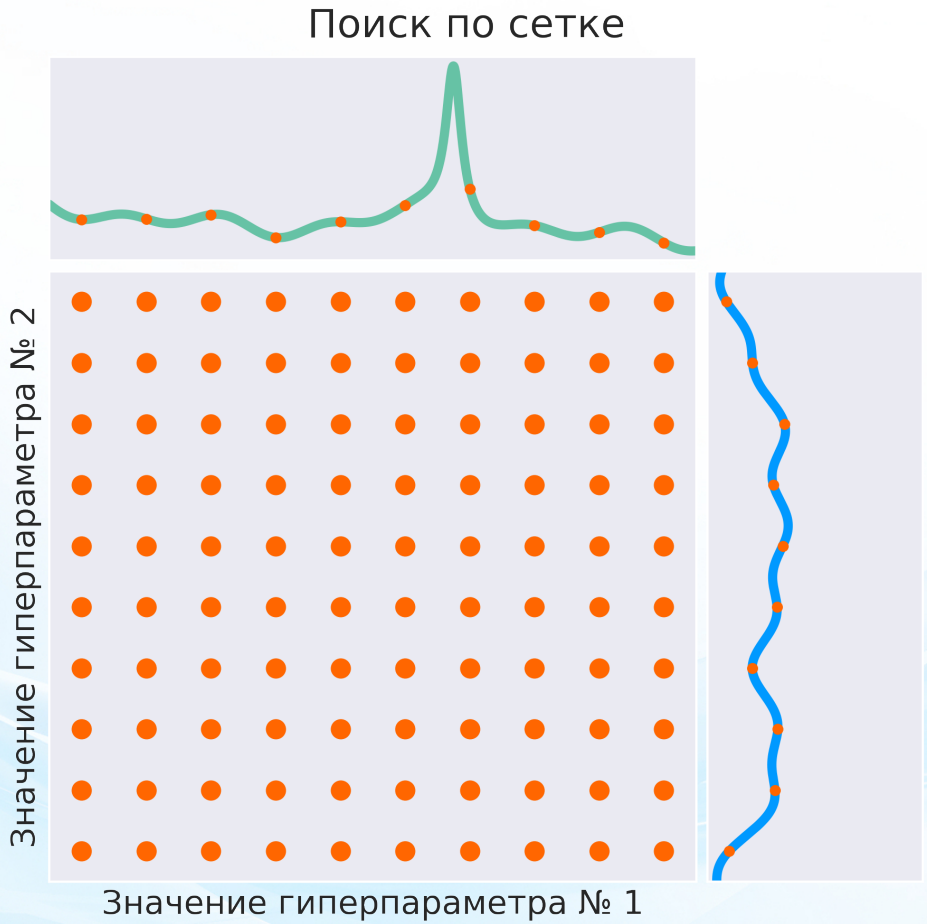
Какую стратегию лучше использовать?



Оказывается, лучше Random Search

Многомерный случай

Рассмотрим функцию $F(x, y) = g(x) + h(y)$,
где $g(x)$ имеет узкий пик, а $h(y)$ меняется слабо.



Мы рассмотрели слишком мало значений каждого гиперпараметра.

Многомерный случай

Рассмотрим функцию $F(x, y) = g(x) + h(y)$,
где $g(x)$ имеет узкий пик, а $h(y)$ меняется слабо.

Для каждого гиперпараметра мы рассмотрели достаточно значений.

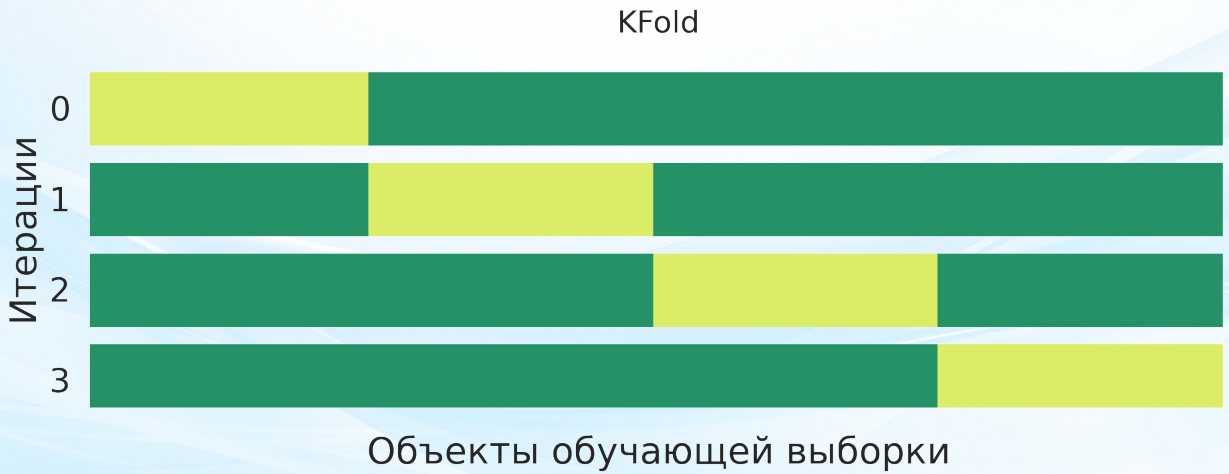
Кросс-валидация: **KFold**

KFold — метод оценки качества модели, при котором обучающая выборка делится на k частей, или **фолдов**.

После чего производится k **итераций**:

1. модель обучается на совокупности всех фолдов, кроме фолда с номером j
2. обученная модель оценивается на оставшемся j-ом фолде

Таким образом мы получаем k оценок качества.
Итоговая метрика считается как среднее полученных оценок.

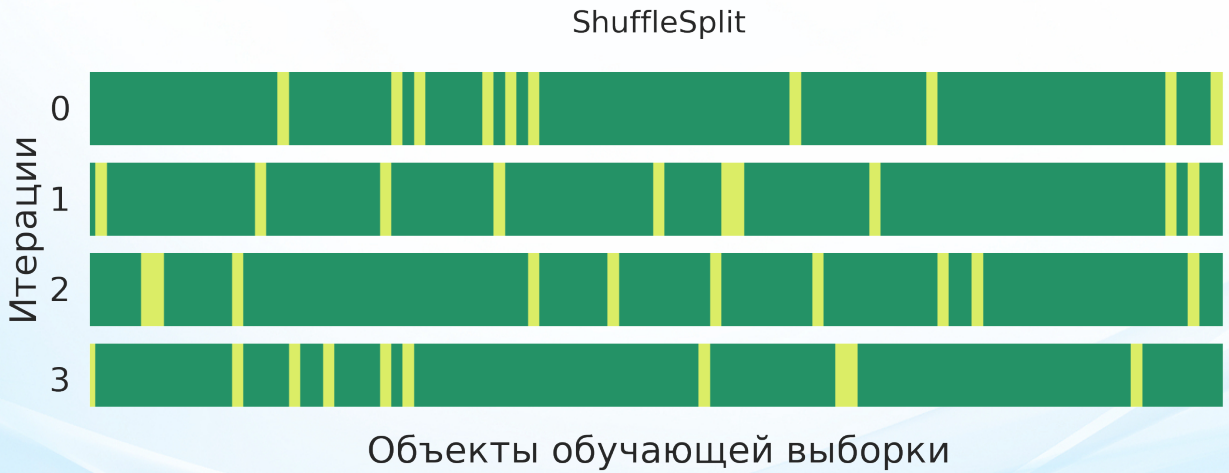


Кросс-валидация: **ShuffleSplit**

После чего производится k **итераций**:

1. делим выборку на две части случайным образом
2. модель обучается на первой части
3. обученная модель оценивается на второй части

Итоговая метрика считается как среднее полученных оценок.



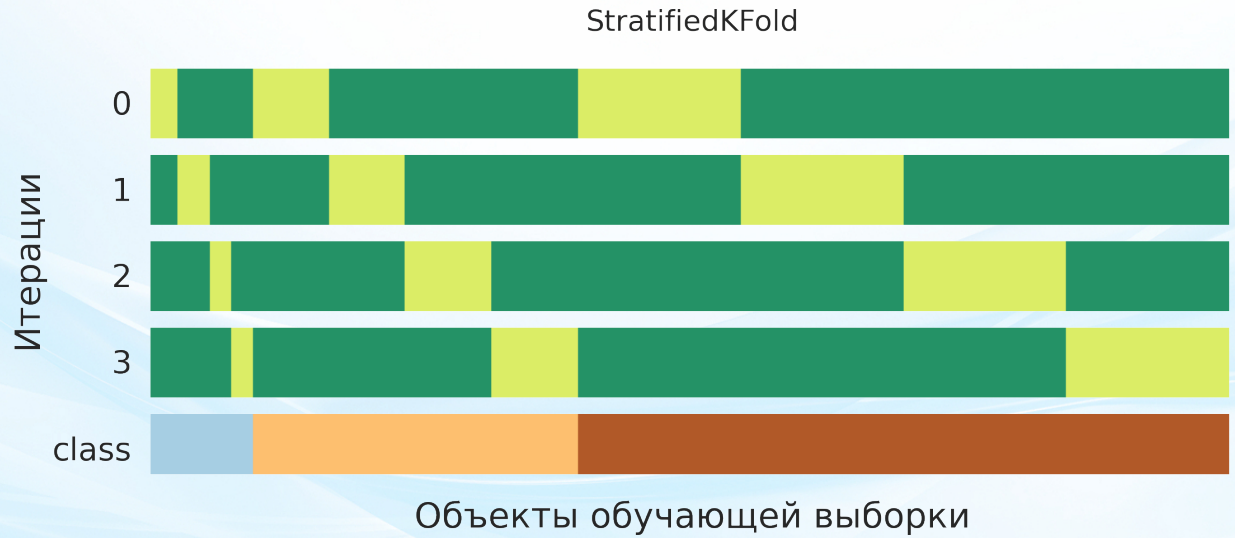
1. **Преимущества**: более четкий контроль над количеством итераций и разбиением на train и test.
2. **Недостаток**: иногда распределение таргета в обучающей и тестовой выборке может быть слишком разным.

Стратифицированная кросс-валидация

Ограничение: задача классификации.

Если имеется сильный дисбаланс классов, то объекты одного из классов могут просто не попасть в фолды для обучения.

При стратифицированном разбиении происходит разделение на фолды отдельно **внутри** каждого класса.

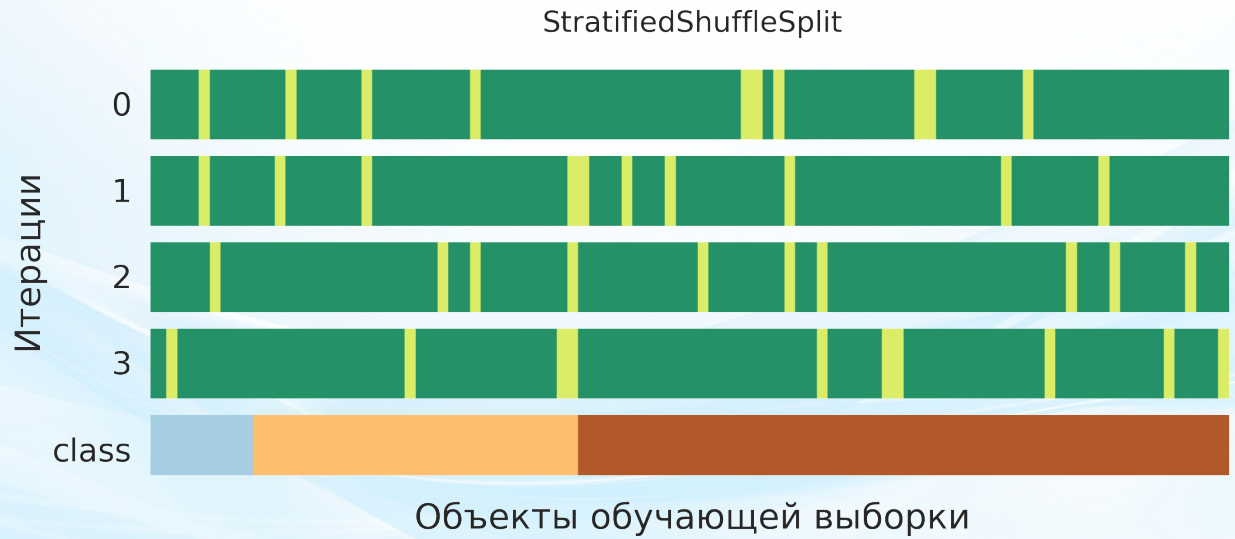


Стратифицированная кросс-валидация

Ограничение: задача классификации.

Если имеется сильный дисбаланс классов, то объекты одного из классов могут просто не попасть в фолды для обучения.

При стратифицированном разбиении происходит разделение на фолды отдельно **внутри** каждого класса.

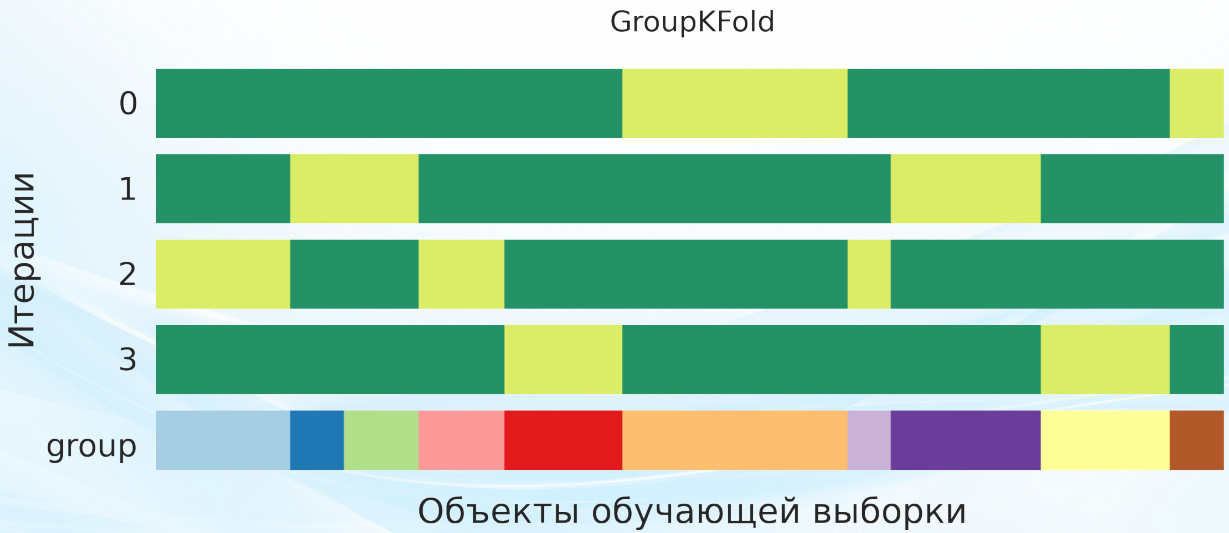


Групповая кросс-валидация

Пусть в данных есть **группы независимых объектов**.
Например, объекты выборки — показатели уровня сахара в крови, причем на одного человека приходится по несколько записей.

Основная особенность: записи, относящиеся к одному человеку, зависимы, их нельзя разделять.

В таком случае разбиение нужно производить по группам.

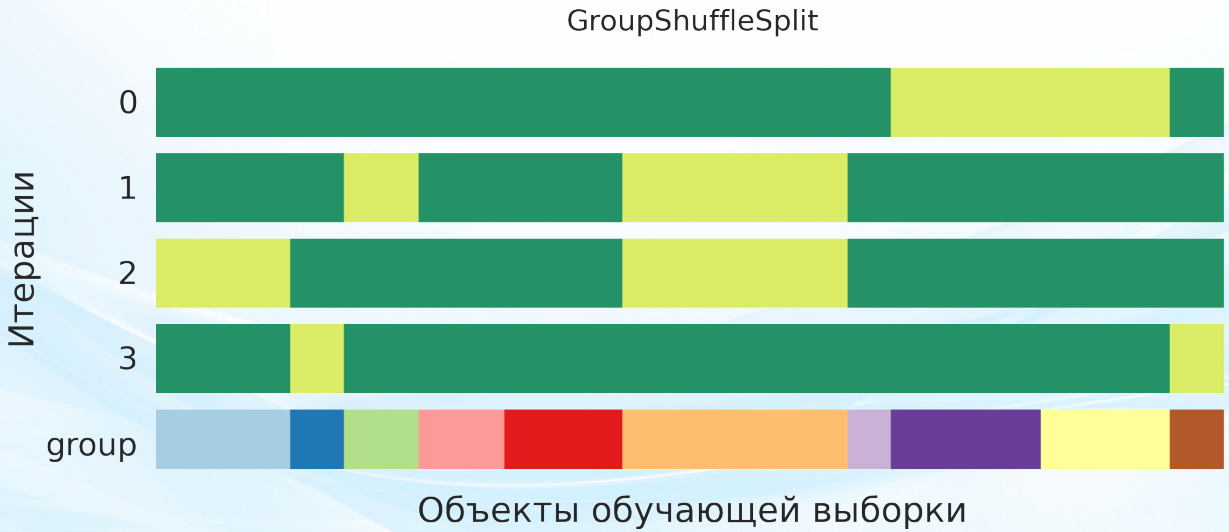


Групповая кросс-валидация

Пусть в данных есть **группы независимых объектов**.
Например, объекты выборки — показатели уровня сахара в крови, причем на одного человека приходится по несколько записей.

Основная особенность: записи, относящиеся к одному человеку, зависимы, их нельзя разделять.

В таком случае разбиение нужно производить по группам.



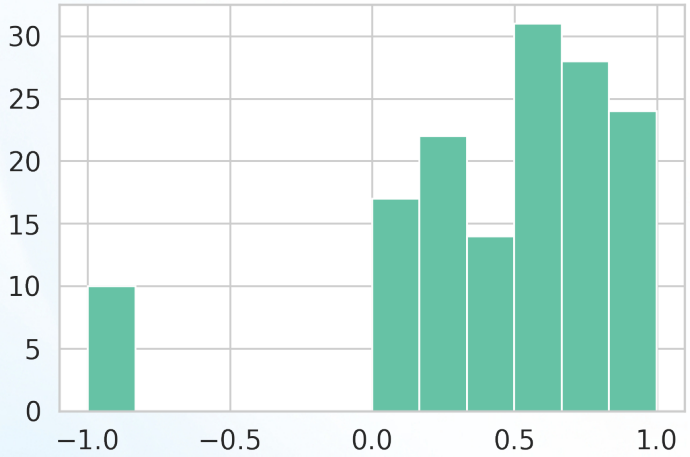
Обработка пропусков в данных

Что может быть пропуском?

Пропуском может быть:

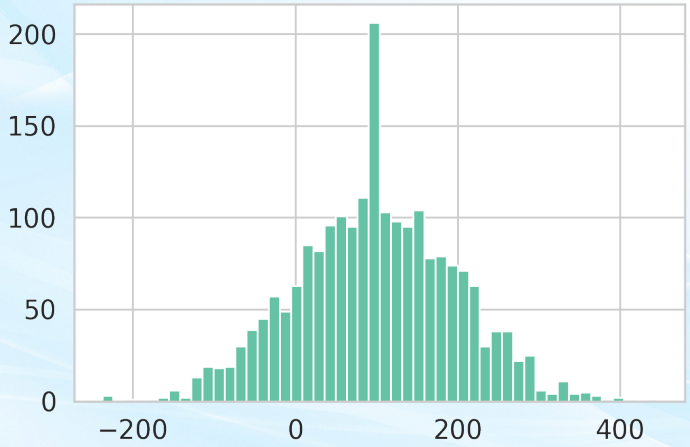
- NaN
- "nan"
- Пустая строка
- -
- ?
- -1
- 1000000
- -99999
- 999

Как понять что является пропуском?



Посмотрим на гистограмму.

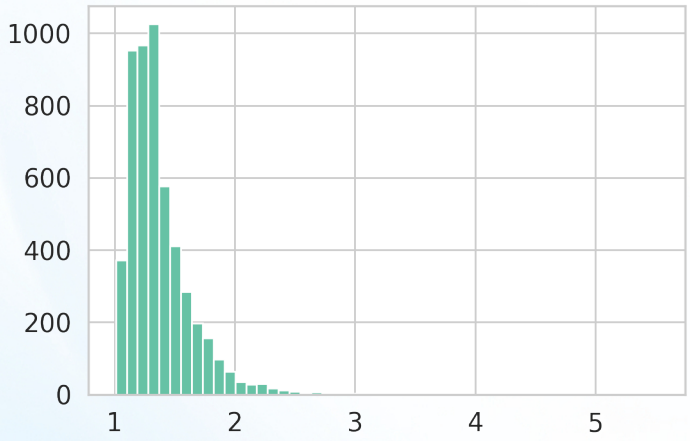
Все пропущенные значения заменены на -1.



А что произошло здесь?

Пропущенные значения заменены на среднее значение признака.

Как понять что является пропуском?



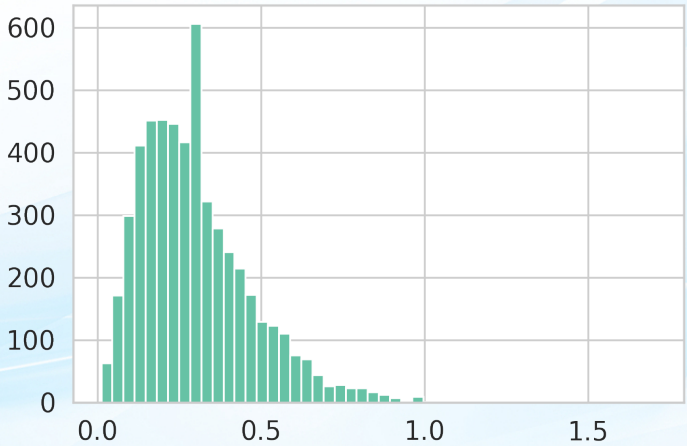
Что можно понять здесь?

Хмм, ничего не понятно...

Прологарифмируем

значения признака.

Теперь пропуски отчетливо видны.



Какие бывают пропуски?

Время	8:00	9:00	10:00	11:00	12:00
Температура возд.	21.4	22.1	**NaN**	24.2	25.5

Знаем: температура воздуха всегда есть :)

Возможные причины пропуска:

- Метеоролог был пьян.
  - События "метеоролог пьян" нет в датасете
    $\Longrightarrow$ *абсолютно случайный пропуск*.
  - Событие "метеоролог пьян" есть в датасете
    $\Longrightarrow$ *случайный пропуск*.
- Перегрелось оборудование
  $\Longrightarrow$ *неслучайный пропуск*.

Какие бывают пропуски?

- **Missing Completely at Random**

  Событие "признак пропущен" не зависит ни от других признаков, ни от значения пропущенного признака.

- **Missing at Random**

  Событие "признак пропущен" не зависит от значения пропущ. признака, но зависит от значения других признаков.

- **Missing not at Random**

  Событие "признак пропущен" зависит от значения пропущенного признака.

Что делать с пропусками? Случайные пропуски.

- Удалить все строки или столбцы с пропущенными значениями.
- Использовать наиболее вероятное значение признака.

  Среднее или медиана для вещественных переменных,
  для категориальных — самое частое значение.
  Неплохо работает на линейных моделей и нейросетей.
- Обучить модель предсказывать пропущенные значения.

  Самые популярные варианты — Linear Regression и KNN.
- Multiple Imputation — обучить несколько разных моделей

  предсказывать пропуски и усреднить их результаты.
- Использовать модели, умеющие работать с пропусками.

  Например, можно считать $X^T X$ и $X^T Y$ только по полным парам

$$\frac{1}{n}\left(X^T X\right)_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, x_{ik} \approx \frac{1}{n_{jk}}\sum_{i=1}^{n} x_{ij}\, x_{ik}\, I\{x_{ij} \text{ и } x_{ik} \text{ не пропущены}\},$$

$$\frac{1}{n}\left(X^T Y\right)_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, y_i \approx \frac{1}{n_j}\sum_{i=1}^{n} x_{ij}\, y_i\, I\{x_{ij} \text{ не пропущено}\}.$$

где n_{jk} — число полных пар (x_{ij}, x_{ik}); n_j — число заполненных x_{ij}.

Что делать с пропусками? Неслучайные пропуски.

- Завести отдельный бинарный признак: $I\{x_j$ — пропущено$\}$.
- Для категориальных признаков завести новую категорию.
- Закодировать каким-то значением, не встречающимся в данных.

  Хорошо работает для моделей на основе деревьев
  т.к. позволяет сделать разделение на пропущенные и непропущенные.
- Использовать модели, умеющие работать с пропусками.

Можно ли их просто удалить?

Нет. Если NaN только для больших знач. T, то распр. будет другим.

Куда отнести Missing at Random?

- Если мы изучаем природу, то пьянство метеоролога
  не должно на нее влиять. Можно считать случайным пропуском.
- Если мы изучаем метеоролога, то это неслучайный пропуск.


ВСЁ!