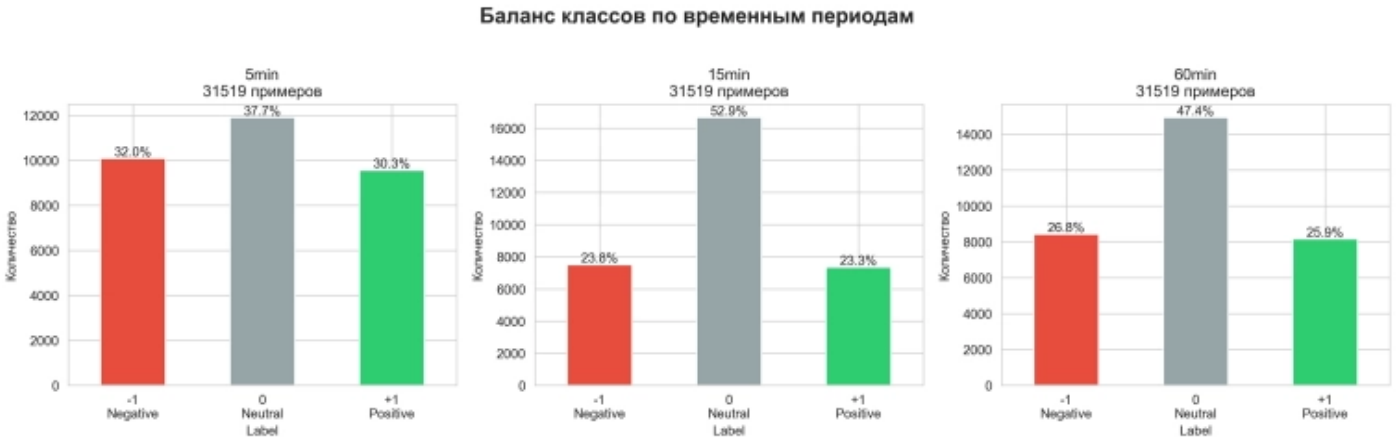
Михаил Шардин
личный блог

17 октября 2025, 06:35

+ Подписаться

Как я погружаюсь в ML и понимаю, что всё начинается с данных

Экспериментирую с ML. Несколько недель ковыряюсь в данных и всё больше понимаю — алгоритмы это не главное. Главная боль — подготовка данных.



Уже несколько недель разбираюсь с машинным обучением. Не ради статьи на Смартлабе — захотелось понять, как это устроено изнутри. И чем глубже погружаюсь, тем больше понимаю: вся «магия» моделей начинается задолго до самого обучения.

Собрать данные — не проблема. Подготовить их — проблема. И особенно сложно — сбалансировать классы.

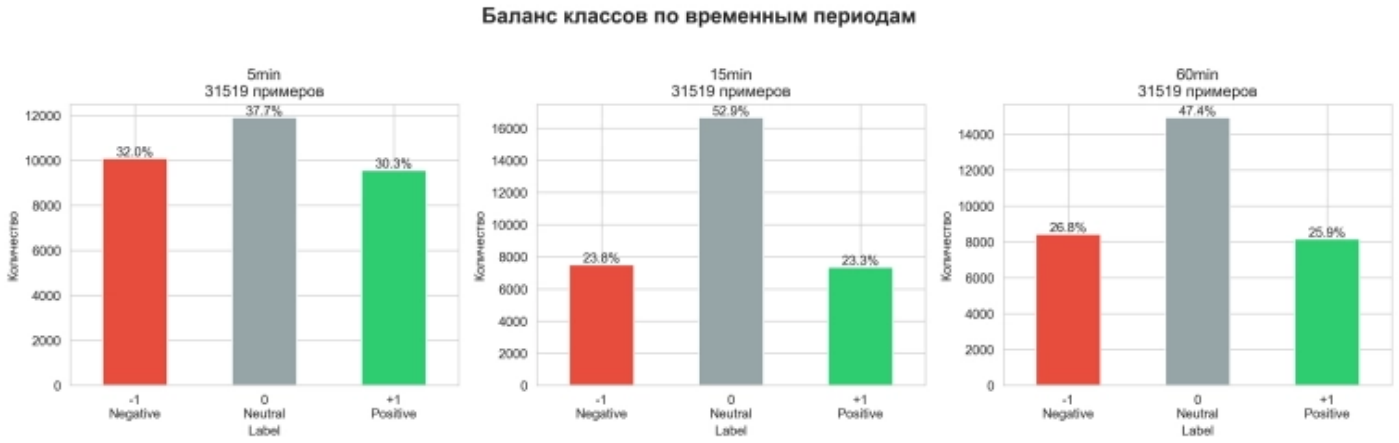
Когда модель учится отличать категории, она должна видеть их примерно в равных долях. Если один класс встречается гораздо чаще других, модель быстро «разучивается думать» и начинает просто угадывать самый частый вариант. В итоге вроде бы всё обучилось, но результат — в мусорку.

На практике это выглядит так:

- один класс (нейтральный) встречается чуть ли не в половине всех примеров,

Введите текст комментария

После долгих экспериментов и отбраковки однотипных записей я получил такую картину:



Выглядит вроде бы прилично, но за этими цифрами — десятки вариантов фильтрации, пересемплирования и чистки.

Главная проблема — нейтральные примеры. Их очень много, и именно они чаще всего «смазывают» сигналы в данных.

Удалить слишком много — модель потеряет способность видеть естественные переходные состояния.
Удалить слишком мало — она начнёт всё считать нейтральным.

Это в продолжение темы.

Поэтому подбор баланса — это не механическая операция, а прямо сложно.
Ведь машинное обучение — это не про «нажал кнопку и получил результат» как в ChatGPT.
Это скорее про методичное, почти исследовательское выравнивание весов, классов и смыслов.

Если кто-то тоже бьётся с дисбалансом классов — делитесь опытом в комментариях.
Особенно интересны реальные кейсы, где удалось найти золотую середину между балансом и сохранением репрезентативности данных.

Автор:
Михаил Шардин
🔗 Моя онлайн-визитка
📢 Telegram «Умный Дом Инвестора»

17 октября 2025

Михаил Шардин

Пермь

398 4 228

с 23 января 2019

+ Подписаться

+HreHDn1F5CZjN...

24 КОММЕНТАРИЯ

Сначала старые

Михаил
17 октября 2025, 07:00

Никак, потому что не имею дело с классами, а предсказываю распределение доходности

+1

Ho_Chu
17 октября 2025, 07:20

рискнул бы не согласиться с тезисом о неважности алгоритмов

взять хотя бы вчерашний день
из-за произошедших событий придется заново обучать некоторые алгоритмы, хотя другие демонстрируют поразительную стойкость к такого рода явлениям, при этом все используют почти один и тот же входной набор данных

но, кажется мне, что даже новое переобучение не поможет… а вот почему не поможет, — это надо сильно подумать

+1

Максим Павлов
17 октября 2025, 09:54

Смотря какой алгоритм используешь. Но почти в любом алгоритме машинного обучения есть параметр по типу «scale_pos_weight», который делает больше веса в обучении на примеры минорного класса. Но лучший вариант — это добрать выборки)

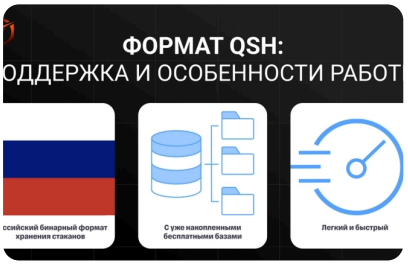
+1

Ещё 7 комментариев

Напишите комментарий...

ОТПРАВИТЬ

Читайте на SMART-LAB:



Формат QSH в OsEngine: поддержка и особенности работы

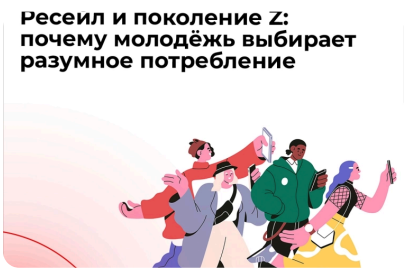
Недавно OsEngine начал поддержку бинарного формата хранения и трансляции данных по стаканам. Это было нужно, чтобы:...

OsEngine 18:01

Промышленная автоматизация — один из ключевых трендов 2026 в ИТ #SOFL_тренды

Сегодня промышленность все чаще смотрит на ИТ как на инструмент для наращивания мощностей. Для российской...

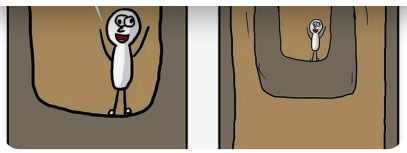
Softline 17:16



🧠 Ресейл и поколение Z: почему молодёжь выбирает разумное потребление

📱 Поколение Z относится к потреблению прагматичнее, чем остальные. Для них важны не громкие слова и статус, а понятна...

МГКЛ 10:00



Вышла статистика рынка труда за декабрь 2025 года, которую Хедхантер публикует ежемесячно, что же там интересного:...

Mozgovik 13.01.2026