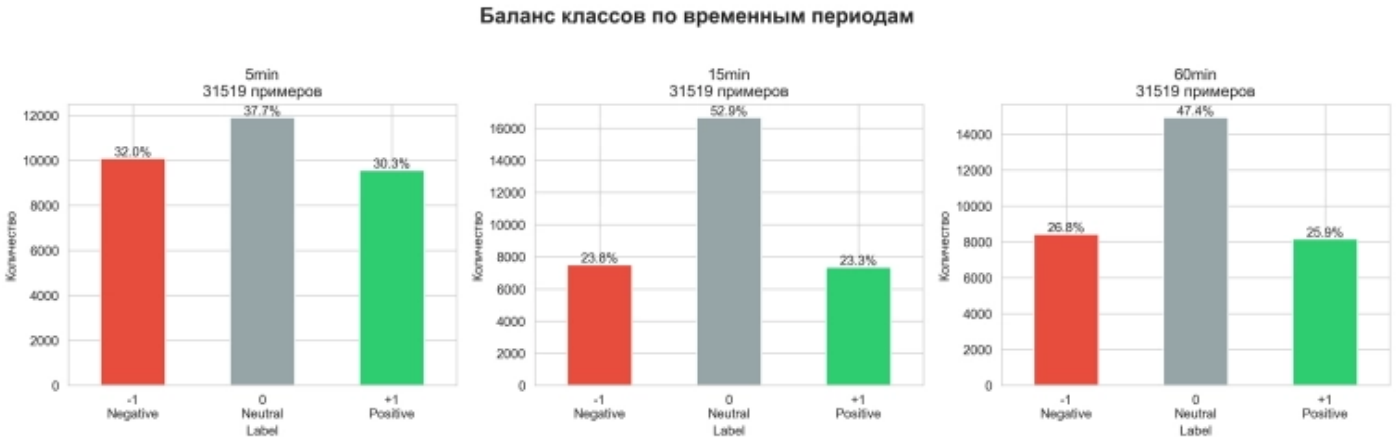
Михаил Шардин
личный блог

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.

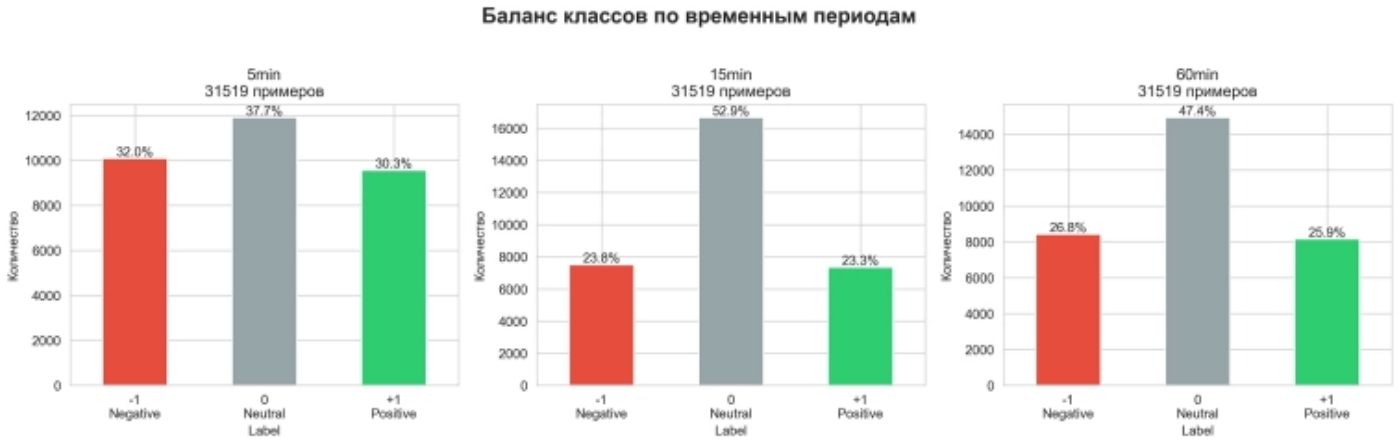
Собрать данные — не проблема. Подготовить их — проблема. И особенно сложно — сбалансировать классы.

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:

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

362 3 511

с 23 января 2019

+HreHDn1F5CZjN...

+ Подписаться

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:

Первичные размещения облигаций — в чем выгода?

Российские эмитенты продолжают активно предлагать новые выпуски облигаций. Участие в первичном размещении может...

БКС Мир инвестиций 21.11.2025

RENI провела звонок для инвесторов и аналитиков по итогам раскрытия результатов за 9М 2025 года

Для удобства вопросы собраны в единый документ, размещенный на нашем сайте в разделе для инвесторов...

Ренессанс страхование 21.11.2025

Инфляция активов против инфляции потребительской – что действительно тормозит снижение ставки?

В публичном поле инфляция традиционно отождествляется с индексом потребительских цен (CPI). Именно этот показатель определяет траекторию ключевой ставки и служит номинальной...

Газпромбанк 20.11.2025