Аугментация данных

Проблема: Недостаток данных в ML

Причины проблемы

Сбор и разметка данных требуют значительных временных и финансовых затрат. В специализированных доменах (медицина, юриспруденция) данные ограничены или конфиденциальны. Редкие события и дисбаланс классов усугубляют проблему.

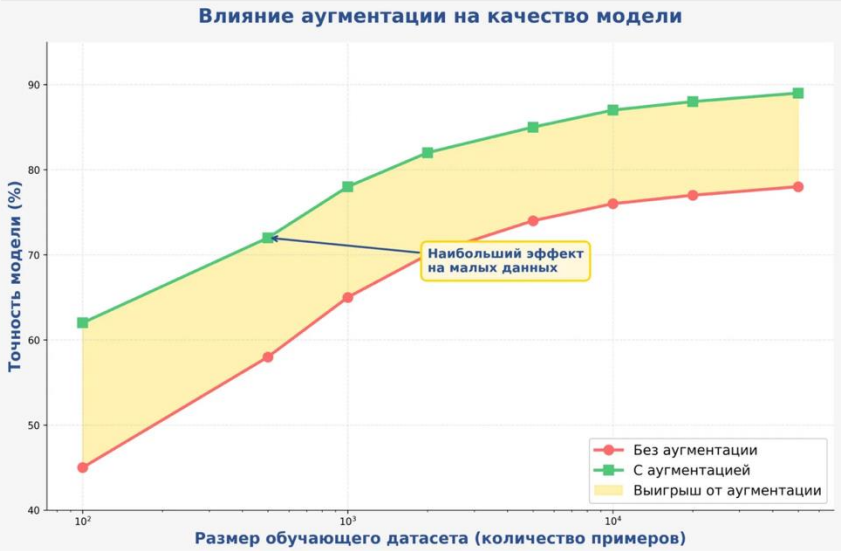
Последствия

Переобучение: модель запоминает обучающую выборку вместо обучения общим закономерностям. **Плохая генерализация:** низкая производительность на новых данных. **Высокая вариативность:** нестабильные результаты.

Решение

Аугментация данных позволяет искусственно увеличить размер и разнообразие обучающей выборки без затрат на реальный сбор данных. Это экономически эффективный способ повышения качества моделей.

Определение аугментации данных



Формальное определение

Аугментация данных — это процесс искусственного увеличения объема и разнообразия обучающих данных путем применения к исходным примерам различных преобразований, которые сохраняют семантику и метки данных, но создают новые вариации.

Ключевые принципы

1. Сохранение меток: Преобразования не должны изменять класс или целевую переменную исходного примера.

2. Реалистичность: Синтетические данные должны быть похожи на реальные и встречаться в практических условиях.

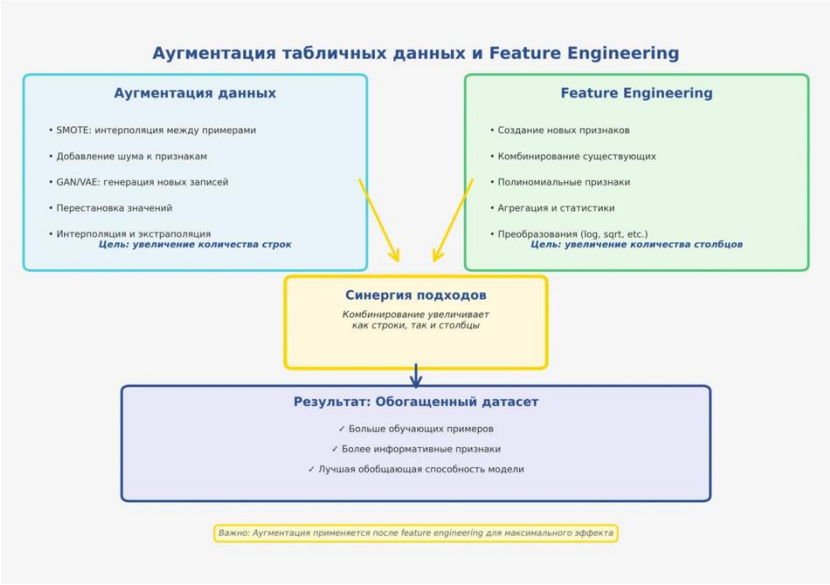
3. Разнообразие: Преобразования должны создавать достаточную вариативность для обобщения модели.

Математическая формулировка

Связь аугментации и Feature Engineering

Feature Engineering: расширение по горизонтали

Feature Engineering увеличивает количество **столбцов** (признаков) через создание новых информативных характеристик. Комбинирование существующих признаков, полиномиальные преобразования, агрегации и статистики обогащают представление данных для модели.



Аугментация: расширение по вертикали

Аугментация табличных данных увеличивает количество **строк** (примеров) в датасете. Методы как SMOTE, добавление шума или генеративные модели создают новые синтетические записи, сохраняя структуру признаков. Это помогает модели лучше обобщать и справляться с несбалансированностью классов.

Ключевая идея

Оптимальная стратегия: сначала применить **Feature Engineering** для создания информативных признаков, затем использовать **аугментацию** для увеличения количества примеров. Это создает обогащенный датасет с максимальной информативностью.

Аугментация как этап ML пайплайна

Feature Engineering

Создание новых признаков из исходных данных: комбинирование столбцов, полиномиальные признаки, агрегации, преобразования.

Результат: **n строк, m+k столбцов**

где k — количечство новых призчнаков.

Аугментация данных

Применяется **после** FE к обогащенным признакам. SMOTE, добавление шума, генеративные модели создают синтетические примеры.

Результат: **n×α строк, m+k столбцов**

где α — коэффициент увеличения.

Обучение и валидация

Модель обучается на аугментированных данных с новыми признаками. **Критично:** валидация и тестирование проводятся только на **исходных** (не аугментированных) данных для объективной оценки.

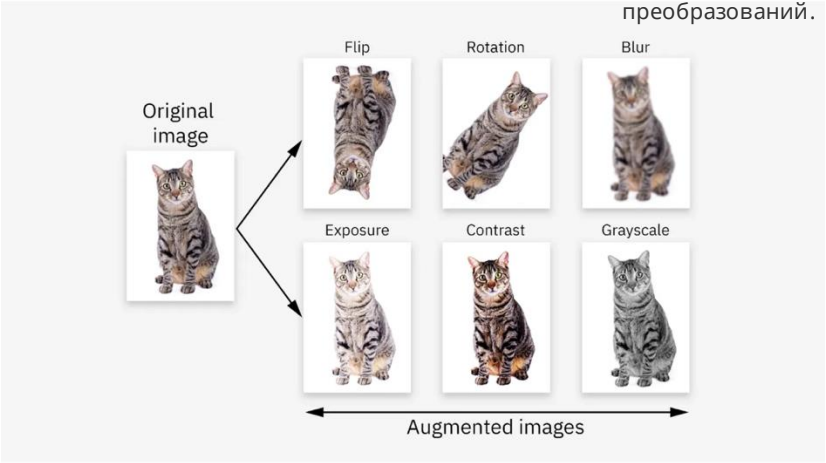
Аугментация в компьютерном зрении

Геометрические преобразования

Поворот, отражение, масштабирование изменяют ориентацию и размер объекта. Обрезка фокусируется на частях изображения. Сдвиг и деформация имитируют небольшие смещения и искажения, характерные для реальных условий съемки.

Фотометрические преобразования

Изменение яркости, контраста и насыщенности имитирует различные условия освещения. Модификация цветового баланса адаптирует модель к разным настройкам камер. Добавление шума повышает устойчивость к низкому качеству изображений.



Продвинутые методы

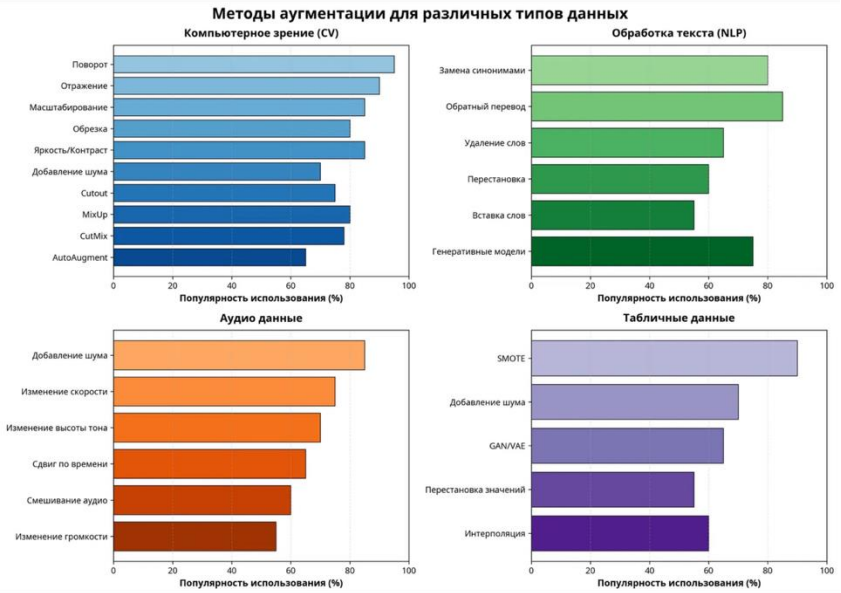
Cutout/Random Erasing случайно удаляет части изображения, заставляя модель не фокусироваться на одной детали.

Mixup & CutMix смешивают два изображения и их метки для создания более плавных границ принятия решений.

Автоматическая аугментация

AutoAugment и **RandAugment** — алгоритмы, которые автоматически находят оптимальную стратегию аугментации для конкретного набора данных, избавляя от необходимости ручного подбора преобразований.

Аугментация для текста и аудио



Аугментация текста (NLP)

Замена синонимами: Замена слов на их синонимы без потери смысла (например, с помощью WordNet или языковых моделей).

Обратный перевод (Back-Translation): Перевод текста на другой язык и обратно для получения новой формулировки с сохранением смысла.

Случайные операции: Вставка, удаление или перестановка слов и предложений для создания вариаций текста.

Генеративные модели: Использование больших языковых моделей (LLM) для генерации парафраз и семантически эквивалентных текстов.

Аугментация аудио

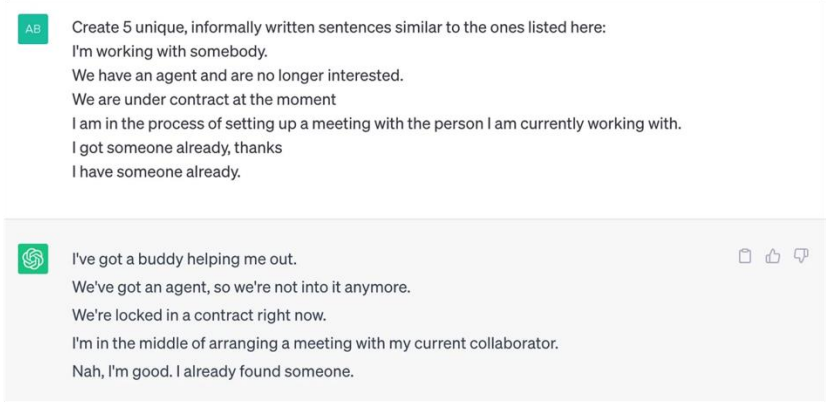
Добавление шума: Наложение фоновых шумов (улица, другие голоса) для повышения робастности модели к реальным условиям.

Изменение скорости и высоты тона: Time/Pitch Shift имитирует разную скорость речи и тембр голоса говорящих.

Сдвиг по времени: Небольшое смещение аудиосигнала во времени для создания вариаций без изменения содержания.

Смешивание аудио: Наложение одного аудио на другое с разными коэффициентами для имитации многоголосых сцен.

Применение LLM в аугментации данных



⭐ Преимущества

Масштабирование малых датасетов в **Большие** . Работа со специализированными доменами (медицина, юриспруденция), где разметка дорогая. Контролируемая генерация с заданными свойствами. Автоматическая верификация и фильтрация качества.

Ограничения

Качество может уступать человеческой разметке. Высокие вычислительные затраты. Необходимость валидации сгенерированных данных. Проблемы с семантически близкими классами.

Основные подходы

Prompt-based generation: использование промптов для контролируемой генерации данных. **Few-shot learning:** создание новых примеров на основе нескольких исходных. **Instruction-tuning:** обучение моделей на синтетических инструкциях от GPT-4.

Применение в задачах

Диалоговые системы: генерация медицинских диалогов, эмоциональной поддержки. **Классификация:** создание примеров для заданных классов. **Reasoning:** генерация цепочек рассуждений (Chain-of-Thought). **Multilingual:** кросс-языковая аугментация.

Табличные данные и лучшие практики

Аугментация табличных данных

SMOTE (Synthetic Minority Over-sampling Technique) генерирует синтетические примеры для миноритарного класса путем интерполяции между существующими примерами в пространстве признаков.

Добавление шума — внесение небольшого случайного шума в числовые признаки для создания вариативности данных.

Генеративные модели (VAE, GAN) обучаются на существующих данных для генерации новых, статистически похожих записей с сохранением корреляций между признаками.

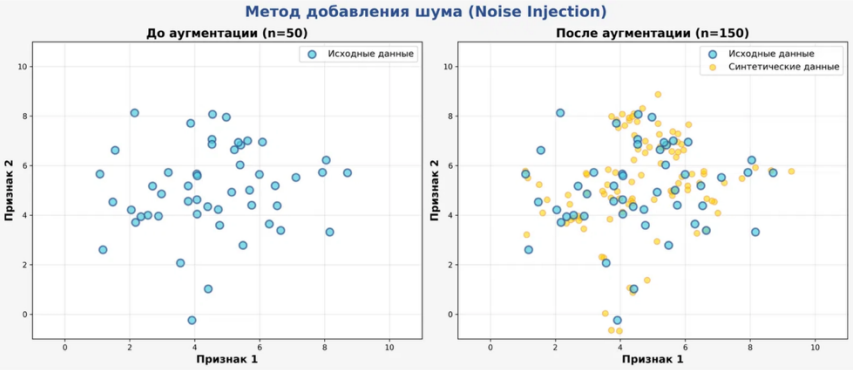
Лучшие практики

Разделение данных: Аугментированные данные только в обучающей выборке. Валидация и тестирование — строго на исходных данных.

Контроль силы преобразований: Избегайте чрезмерных искажений, которые создают нереалистичные данные (например, поворот цифры 6 превратит ее в 9).

Соответствие домену: Выбирайте преобразования, которые соответствуют реальным искажениям в вашей задаче и области применения.

Метод добавления шума (Noise Injection)



Описание метода

Простой и эффективный метод аугментации: к числовым признакам добавляется небольшой гауссовский шум. Это создает вариации исходных данных, сохраняя их статистические свойства и улучшая робастность модели.

Параметры и рекомендации

Уровень шума: 3-10% от стандартного отклонения признака

Количество копий: 1-3 синтетических примера на каждый исходный

Применение: только к числовым признакам

Преимущества: простота, скорость, хорошая регуляризация

Недостатки: может нарушить доменные ограничения

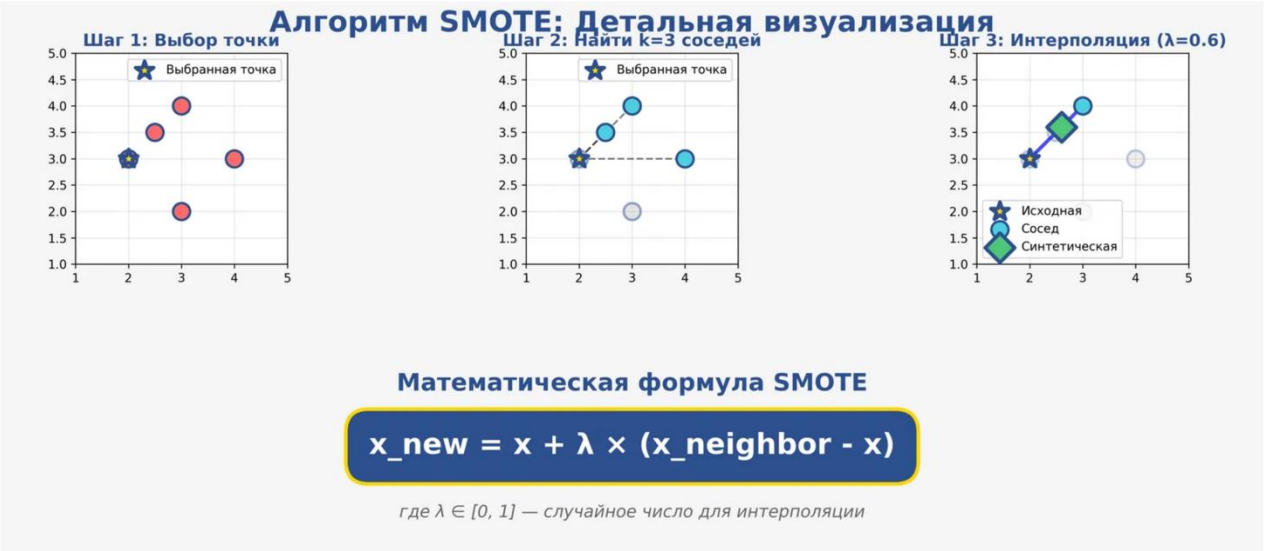
Математическая формула

$$x'_i = x_i + \varepsilon, \text{ где } \varepsilon \sim N(0, \sigma^2)$$

σ — стандартное отклонение шума, обычно $\sigma = \alpha \times \text{std}(x_i)$, где $\alpha \in [0.01, 0.1]$ — коэффициент силы шума.

Метод SMOTE: Теория и алгоритм

SMOTE (Synthetic Minority Over-sampling Technique) — метод создания синтетических примеров путем интерполяции между существующими точками данных и их ближайшими соседями в признаковом пространстве.



Метод ADASYN: Адаптивная синтетическая выборка

Основная идея

ADASYN (Adaptive Synthetic Sampling) — улучшенная версия SMOTE, которая адаптивно определяет количество синтетических примеров для каждой точки данных в зависимости от сложности её окружения. Чем сложнее граница класса, тем больше примеров генерируется.

Преимущества

- Лучше обрабатывает сложные границы классов
- Адаптируется к локальному распределению данных
- Снижает риск переобучения на простых областях
- Более эффективное использование синтетических данных

Отличия от SMOTE

SMOTE: генерирует одинаковое количество примеров для всех точек

ADASYN: фокусируется на сложных областях (границах классов)

Плотность генерации: $r_i = \Delta_i / K$, где Δ_i — количество примеров другого класса среди K соседей

Когда использовать ADASYN

Задачи с сильно несбалансированными классами и сложными нелинейными границами решений

Сравнение методов аугментации

Выбор метода аугментации зависит от характеристик данных, вычислительных ресурсов и требований к качеству. Каждый метод имеет свои преимущества и области применения.

Сравнение методов аугментации табличных данных

Метод	Сложность	Качество данных	Скорость	Применение
SMOTE	Средняя	Высокое	Средняя	Несбалансированные классы
ADASYN	Высокая	Очень высокое	Медленная	Сложные границы классов
Noise Injection	Низкая	Среднее	Быстрая	Общая регуляризация
Комбинированный	Высокая	Очень высокое	Медленная	Максимальное качество

Несбалансированность

Используйте **SMOTE** или **ADASYN** для балансировки классов

Скорость важна

Выбирайте **Noise Injection** для быстрой аугментации

Максимальное качество

Применяйте **комбинированный подход** для лучших результатов

Пайплайн аугментации табличных данных

Правильная последовательность этапов критически важна для получения объективных результатов. Аугментация применяется только к обучающей выборке после разделения данных и создания признаков.

Заключение

Улучшение моделей

Аугментация позволяет не только увеличить объем данных, но и повысить их разнообразие, что напрямую ведет к улучшению обобщающей способности модели и ее устойчивости к новым, невиданным ранее данным.

$ Экономическая эффективность

В ситуациях, когда сбор и разметка новых данных дороги или невозможны, аугментация становится наиболее эффективным способом повышения качества модели с минимальными затратами.

Ключевой вывод

Успешная аугментация требует глубокого понимания данных, доменных знаний и итеративного подхода к подбору методов и их параметров. Это не просто набор техник, а мощный инструмент в арсенале специалиста по данным.