1. Понимание данных

- **Идентификация и описание датасета:**

В данном проекте используется датасет elb_request_count_8c0756.csv из репозитория NAB (Numenta Anomaly Benchmark), адаптированного для задач обнаружения аномалий. Конкретно, этот файл содержит данные счетчика запросов к Elastic Load Balancer (ELB) в AWS CloudWatch. Датасет состоит из двух столбцов:

* `timestamp`: Метка времени, указывающая на момент сбора данных. Формат данных – временной ряд.

* `value`: Количество запросов к ELB за 5-минутный интервал, зафиксированное в соответствующий момент времени.

- **Источник, размеры и ключевые признаки:**
    - **Источник:** Репозиторий NAB (Numenta Anomaly Benchmark), который включает в себя различные наборы данных временных рядов для тестирования алгоритмов обнаружения аномалий. В данном случае, данные эмулируют реальные метрики производительности облачных сервисов, собранные AWS CloudWatch.
    - **Размеры:** Датасет содержит 4032 записи, охватывающие временной период с 10 апреля 2014 года по 24 апреля 2014 года.
    - **Ключевые признаки:** Основным признаком является value, представляющий собой количество запросов. timestamp служит индексом временного ряда и ключом для анализа временных паттернов.
- **Значение столбцов и их значимость для анализа:**
    - timestamp: Критически важен для анализа временных рядов, поскольку порядок данных и временная привязка являются ключевыми аспектами. Позволяет отслеживать изменения value во времени и выявлять аномалии в контексте временной динамики.
    - value: Основная метрика, для которой ищется аномальное поведение. Резкие отклонения от ожидаемых значений value могут сигнализировать о проблемах в работе веб-приложения (например, DDoS-атака, сбои, аномальная нагрузка).
- **Проблемы качества данных, выявленные в Notebook (на основе разведочного анализа):**

    - **Пропуски:** Notebook явно показывает отсутствие пропусков (df.isnull().sum()).
    - **Выбросы:** Статистический анализ (df.describe(), boxplot) демонстрирует наличие выбросов (максимальное значение значительно превышает 75% квартиль). Это типично для данных мониторинга производительности, где кратковременные всплески нагрузки могут быть нормальным явлением, но и потенциальными аномалиями.
    - **Дисбаланс:** Распределение данных (гистограмма) смещено вправо, что указывает на то, что большинство значений сосредоточено в нижней части диапазона, а более высокие значения встречаются реже. Это также характерно для метрик запросов, где периоды низкой активности преобладают над периодами пиковой нагрузки.
- **Ключевые статистические свойства датасета (из разведочного анализа):**
    - count: 4032 (количество записей)
    - mean: 61.84 (среднее количество запросов)
    - std: 56.66 (стандартное отклонение, указывающее на значительную изменчивость данных)
    - min: 1.00 (минимальное значение)
    - 25%: 15.00 (первый квартиль, 25% значений ниже этого уровня)
    - 50%: 48.00 (медиана, 50% значений ниже этого уровня)
    - 75%: 89.00 (третий квартиль, 75% значений ниже этого уровня)
    - max: 656.00 (максимальное значение, подтверждающее наличие выбросов)

2. Документация по предобработке данных

- **Шаги по очистке данных:**
    - **Проверка на пропуски:** Явных шагов по обработке пропусков в Notebook нет, поскольку анализ показал их отсутствие.
    - **Преобразование типа данных:** Столбец timestamp был преобразован в формат datetime с помощью pd.to_datetime(). Это необходимо для корректной работы с временными рядами и извлечения временных признаков.

    - **Установка временного индекса:** Столбец timestamp установлен в качестве индекса DataFrame (df.set_index('timestamp', inplace=True)). Это стандартная практика для временных рядов в pandas, облегчающая их анализ и визуализацию.

- **Обоснование трансформаций, кодирования и нормализации:**
    - **Преобразование времени в datetime и установка индекса:** Необходимы для временного анализа и использования pandas для работы с временными рядами.
    - **MinMaxScaler:** Применен для масштабирования значений столбца value в диапазон [-1, 1]. Это стандартная практика для нейронных сетей, помогающая улучшить процесс обучения, ускорить сходимость и предотвратить проблемы с численными значениями. Диапазон [-1, 1] выбран для соответствия диапазону активационной функции tanh, часто используемой в LSTM.
- **Подходы к Feature Engineering и их обоснование:**
    - **Временные признаки:** Созданы признаки hour, day, dayofweek, month на основе временного индекса. Обоснование:
        - **Учет временных паттернов:** Временные ряды запросов к веб-приложениям часто демонстрируют сезонность (например, почасовая, дневная, недельная, месячная). Эти признаки позволяют модели LSTM учитывать эти паттерны при прогнозировании и обнаружении аномалий.
        - **Почасовая активность (hour):** Позволяет модели улавливать закономерности, связанные с временем суток (например, пики нагрузки в рабочее время, снижение ночью).
        - **Дни недели (dayofweek):** Позволяет модели учитывать различия в активности в будние и выходные дни.
        - **Месяц (month):** Хотя временной период датасета невелик (несколько дней), включение месяца может быть полезно для более долгосрочных прогнозов или при анализе более длительных временных рядов.
- **Методология разделения данных на обучающую и тестовую выборки:**
    - **train_test_split с test_size=0.2 и shuffle=False:**

        - **test_size=0.2:** 20% данных выделено для тестовой выборки, что является распространенным соотношением для оценки качества модели.
        - **shuffle=False:** Перемешивание данных отключено (shuffle=False). Обоснование: Важно сохранить временной порядок данных временных рядов. Перемешивание нарушит временную структуру и сделает оценку модели некорректной для задач прогнозирования временных рядов. Тестовая выборка должна представлять собой последние 20% данных, имитируя сценарий прогнозирования на "новых" данных, следующих за обучающим периодом.

3. Документация по моделированию

- **Используемые алгоритмы машинного обучения/статистические методы:**
    - **LSTM (Long Short-Term Memory) сеть:** Использована как основная модель для прогнозирования временных рядов и обнаружения аномалий. Обоснование:
        - **Работа с последовательностями:** LSTM - это тип рекуррентной нейронной сети (RNN), специально разработанный для обработки последовательных данных, таких как временные ряды.
        - **Учет долгосрочных зависимостей:** LSTM способна запоминать информацию на длительных временных интервалах, что важно для улавливания сложных временных паттернов и зависимостей в данных запросов к веб-приложениям.
        - Э**ффективность для прогнозирования временных рядов:** LSTM хорошо зарекомендовали себя в задачах прогнозирования временных рядов и обнаружения аномалий в таких данных.
- **Обоснование выбора модели и гиперпараметров:**
    - **Модель LSTM:** Выбор LSTM обусловлен ее способностью обрабатывать последовательные данные и улавливать временные зависимости, что делает ее подходящей для прогнозирования временных рядов и обнаружения аномалий.
    - **Гиперпараметры LSTM:**

        - input_size=1: Один входной признак (value).
        - hidden_layer_size=100: Размер скрытого слоя LSTM. Выбран эмпирически, достаточно большой для улавливания сложных паттернов, но не чрезмерно большой, чтобы избежать переобучения.
        - output_size=1: Один выходной признак (прогнозируемое значение value).
        - num_layers=2: Два LSTM слоя. Увеличение количества слоев позволяет модели учить более сложные представления, но также увеличивает сложность модели и риск переобучения. Два слоя - разумный компромисс.
    - **Параметры обучения:**
        - criterion = nn.MSELoss(): Функция потерь - среднеквадратичная ошибка (MSE). Подходит для задач регрессии, где нужно минимизировать разницу между предсказанными и реальными значениями.
        - optimizer = optim.Adam(model.parameters(), lr=0.001): Оптимизатор Adam с learning rate 0.001. Adam - популярный и эффективный оптимизатор для нейронных сетей. Learning rate 0.001 - типичное значение для Adam.
        - seq_length = 24: Длина последовательности (временного окна) для прогнозирования. Выбрана равной 24 часам, что позволяет модели учитывать суточную сезонность.
        - batch_size = 32: Размер батча при обучении. Типичное значение, балансирующее между скоростью обучения и устойчивостью сходимости.
        - num_epochs = 50: Количество эпох обучения. Выбрано эмпирически, достаточное для сходимости модели, как видно из графика динамики функции потерь.
- **Метрики оценки качества модели и их обоснование:**
    - **MSE (Mean Squared Error - Среднеквадратичная ошибка):** Основная метрика оценки качества модели. Обоснование:
        - **Чувствительность к величине ошибки:** MSE квадратично увеличивает вес больших ошибок, что важно

для обнаружения аномалий, которые часто характеризуются резкими отклонениями от нормы.

- **Простота интерпретации:** MSE легко интерпретировать как среднюю квадратичную разницу между предсказанными и реальными значениями.

4. Интерпретация результатов

- **Бизнес-ориентированное описание результатов работы модели:**
  - Модель LSTM, обученная на исторических данных о количестве запросов к ELB, успешно обнаруживает аномалии в тестовой выборке.
  - Обнаружено 6 аномалий из 802 точек данных (0.75%), что свидетельствует о высокой точности модели и ее способности выделять действительно редкие и значимые отклонения от нормального поведения.
  - Средняя ошибка предсказания (MSE) на тестовой выборке составляет 0.032, что указывает на достаточно хорошее качество прогнозирования модели.
  - Порог аномалий (0.397) установлен автоматически на основе распределения ошибок предсказания и может использоваться для автоматического обнаружения аномалий в реальном времени.
- **Интерпретация метрик качества модели:**
  - **MSE = 0.032:** Низкое значение MSE говорит о том, что модель достаточно точно прогнозирует количество запросов. Однако, абсолютное значение MSE зависит от масштаба данных (после MinMaxScaler значения находятся в диапазоне [-1, 1]). Более важным является использование MSE для сравнения моделей и для определения порога аномалий.
  - **0.75% аномалий:** Низкий процент обнаруженных аномалий (менее 1%) указывает на то, что модель не склонна к "ложной тревоге" и выделяет только наиболее значимые отклонения. Это важно для практического применения, чтобы избежать перегрузки операторов большим количеством ложных срабатываний.
- **Выводы на основе технического анализа:**
  - Обнаруженные аномалии, особенно топ-10 аномалий с наибольшей MSE, требуют детального анализа и проверки. Они

могут указывать на реальные проблемы в работе веб-приложения, такие как:

- **DDoS-атаки:** Резкий всплеск запросов (как в аномалии №1 с value=656) может быть признаком DDoS-атаки.
- **Сбои или ошибки в приложении:** Аномально низкие значения запросов или резкие спады могут указывать на проблемы с доступностью или производительностью приложения.
- **Аномальная активность пользователей:** Необычные паттерны запросов могут быть связаны с аномальным поведением пользователей или ботов.

- **Распределение аномалий по часам и дням недели:** Анализ countplot показывает, что аномалии чаще встречаются в определенные часы и дни недели. Это может указывать на закономерности в аномальном поведении, связанные с бизнес-циклами или внешними факторами. Например, если большинство аномалий приходится на ночное время выходных дней, это может быть менее критично, чем аномалии в рабочее время будних дней.
  -