Федеральное государственное автономное образовательное учреждение
высшего образования «Московский физико-технический институт
(национальный исследовательский университет)»

Физтех-школа аэрокосмических технологий

Кафедра Аэрофизики летательных аппаратов

Направление подготовки: 09.03.01 Информатика и вычислительная техника (бакалавриат)

Направленность (профиль) подготовки: Компьютерное моделирование

Форма обучения: очная

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

«Алгоритм предиктивного анализа отказов системы видеоаналитики в режиме времени по данным от систем мониторинга»

(бакалаврская работа)

Студент:

Боровец Николай Васильевич

_____

(подпись студента)

Научный руководитель:

Гришин Никита Александрович,

программист ПИШ РПИ

_____

(подпись научного руководителя)

Жуковский

2025

АННОТАЦИЯ

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Обоснование выбора темы и актуальность

Современные системы видеоаналитики используются для мониторинга объектов критической инфраструктуры: аэродромов, транспортных узлов, промышленных площадок. Повышение масштабов данных и требований к задержкам end-to-end-задержки делает необходимым разработку предиктивного анализа отказов на основе потоковых метрик Prometheus.

Цель и задачи исследования

Цель работы: разработать и внедрить метод предиктивного анализа задержек в конвейере видеоаналитики, способный прогнозировать метрику $common\_event\_delay$ и автоматически детектировать аномалии.

Задачи:

1. Анализ структуры и корреляций временных рядов метрик на всех этапах конвейера.
2. Обзор методов прогнозирования и обнаружения аномалий (статистические, ML, DL).
3. Выбор оптимальной модели (трансформер, бустинг, гибрид) и оценка объёма данных и периодичности обучения.
4. Проектирование MLOps-конвейера: feature-engineering, механизм периодического дообучения модели, inference-сервис.
5. Экспериментальная оценка точности и скорости модели на исторических данных.

6. Внедрение системы оповещений и рекомендации по эксплуатации и масштабированию.

Методология и методы исследования

1. Сбор и предобработка потоковых метрик Prometheus.
2. Построение временных рядов и генерация признаков (лаги, статистики, эмбеддинги).
3. Обучение и дообучение моделей (трансформер, LightGBM/CatBoost) с проверкой на перекрёстной валидации.
4. Развёртывание в Docker и измерение latency inference.
5. A/B-тестирование в продуктивной среде.

Теоретическая и практическая значимость

Теоретическая: расширение знаний о гибридных подходах к онлайн-прогнозированию многомерных временных рядов.

Практическая: готовое решение для мониторинга и предупреждения отказов видеоконвейера с SLA конечной метрики $common\_event\_delay$.

1 Общие положения

1.1 Постановка задачи

Пусть $T = \{t_1, t_2, \ldots, t_n\}$ — множество временных меток наблюдений, а d — число метрик, собираемых системой. Для каждой метки t_i формируется вектор значений:

$$\mathbf{x}_i = [m_i^{(1)}, m_i^{(2)}, \ldots, m_i^{(d)}] \in \mathbb{R}^d, \tag{1.1}$$

где компоненты соответствуют:

- метрикам ML-конвейера: $timestamp\_sei$, $time\_delta\_$, $FPS\_$;
- метрикам бэкенда: $ml\_to\_backend\_kafka\_delay$, db_insert_delay;
- метрикам WS-клиента: $common\_event\_delay$, $heartbeat\_*$, $event\_counter$, seq_events_health.

Определим скользящее окно длины L и шаг s. Каждое «окно»:

$$X_k = [\mathbf{x}_{t_k - L + 1}, \ldots, \mathbf{x}_{t_k}] \in \mathbb{R}^{L \times d}. \tag{1.2}$$

Целевая переменная — значение end-to-end-задержки в следующий момент:

$$y_k = common\_event\_delay(t_k + \Delta), \quad \Delta = 15 \text{ с}. \tag{1.3}$$

Обучающая выборка:

$$\mathcal{N} = \{(X_k, y_k)\}_{k=1}^{N}. \tag{1.4}$$

Существует неизвестная функция:

$$f^* : \mathbb{R}^{L \times d} \to \mathbb{R}, \tag{1.5}$$

и задача состоит в построении алгоритма $A : \mathbb{R}^{L \times d} \to \mathbb{R}$ такого, что

$$|A(X_k) - f^*(X_k)| \leq \varepsilon, \quad \forall k. \tag{1.6}$$

Требования к алгоритму A:

1. Минимизировать среднеквадратическую ошибку MSE на валидационной выборке.
2. Обеспечить $latency\_inference(A) < 1$ с в Docker-контейнере.
3. Поддерживать периодическое дообучение (warm-start, frozen-слои, адаптеры/LoRNA).
4. Гибкость частоты прогнозов: онлайн или с фиксацией интервала (например, 30 мин).

Входные данные: многомерный временной ряд $X \in \mathbb{R}^{n \times d}$ (тип FLOAT64), формируемый из Prometheus по 30-секундным срезам (≈90,644 точки за 16 дней).

Формализация задачи: построить алгоритм A, приближающий f^*, и удовлетворяющий указанным ограничениям по точности и latency.

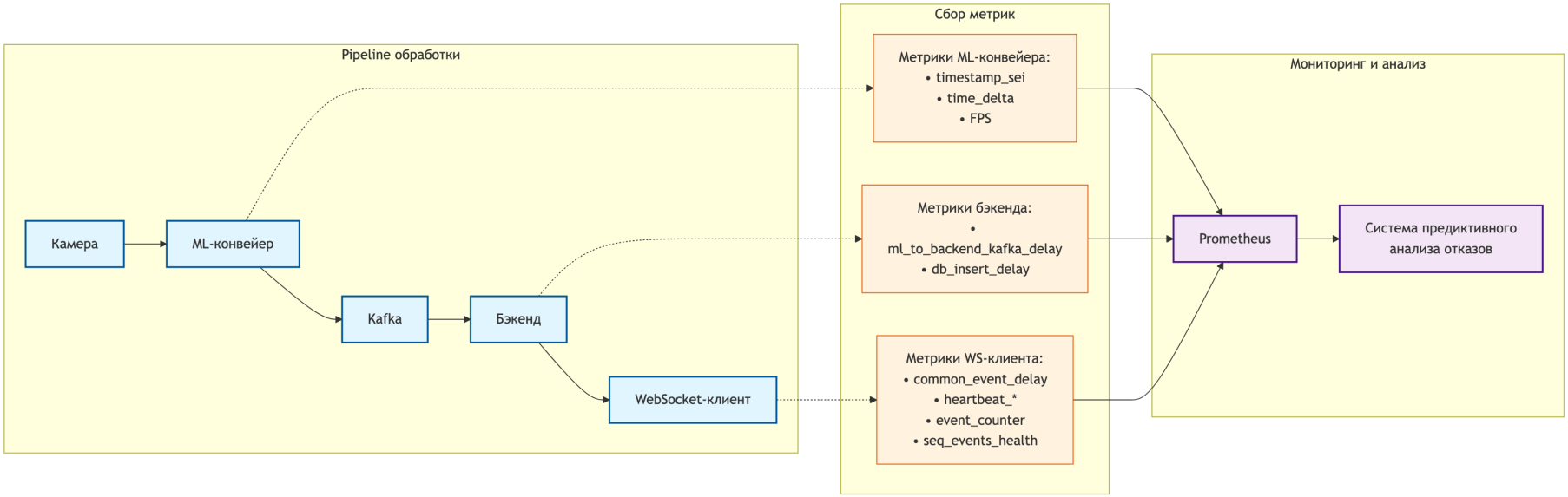

Рисунок 1.1 — Схема видеоконвейера и точки сбора метрик

2 Глава n

2.1 Секция n

3 Глава n

3.1 Секция n

4 Глава n

4.1 Секция n

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1) First

2) Second

3) Third