

ПОРІВНЯЛЬНИЙ АНАЛІЗ ПІДХОДІВ ДО ВИЯВЛЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ НА ОСНОВІ МОДЕЛЕЙ ARIMA, XGBOOST ТА LSTM

Виконала студентка БП-4 ПМ: Мельник Катерина
Науковий керівник: професор, д. ф.-м. н. Крюкова Галина Віталіївна

Актуальність

- Часові ряди — один із найпоширеніших форматів даних у сучасних системах
- Виявлення аномалій критично важливе у багатьох сферах
- Проблема: немає універсального методу — ефективність залежить від характеру даних, типу аномалій та умов застосування
- Питання дослідження: який з підходів забезпечує найкращий компроміс між якістю детекції та часом обчислення на одновимірних рядах?

Мета: порівняльний аналіз методів виявлення аномалій у часових рядах та оцінка їхньої ефективності на реальних даних

Завдання:

1. Систематизувати теоретичні основи часових рядів та типів аномалій
2. Розглянути різні класифікації методів
3. Обрати представників з трьох груп: статистичні методи, машинне навчання, нейронні мережі
4. Провести практичне порівняння за однакових умов

Основні поняття

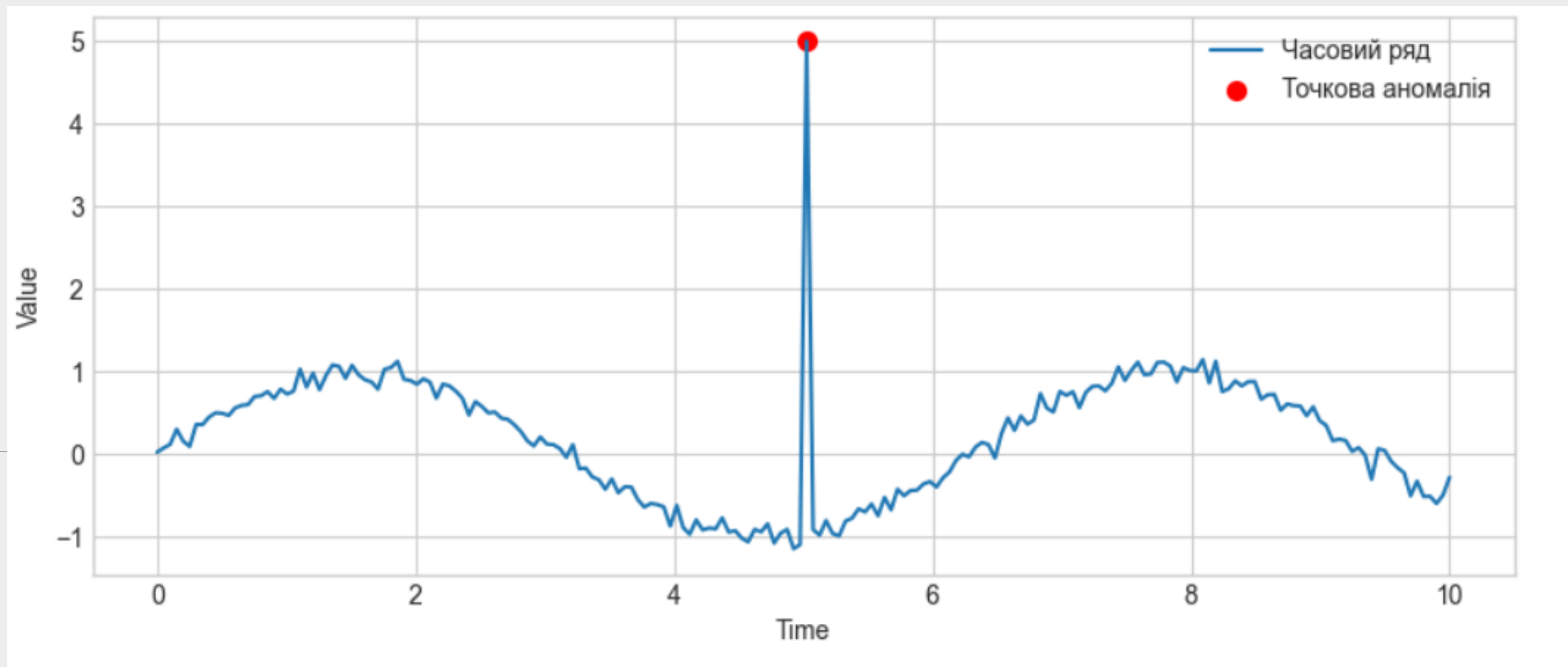
Часовий ряд — впорядкована за часом послідовність спостережень $\{X_t, t \in T\}$, що є реалізацією стохастичного процесу [1].

Складові:

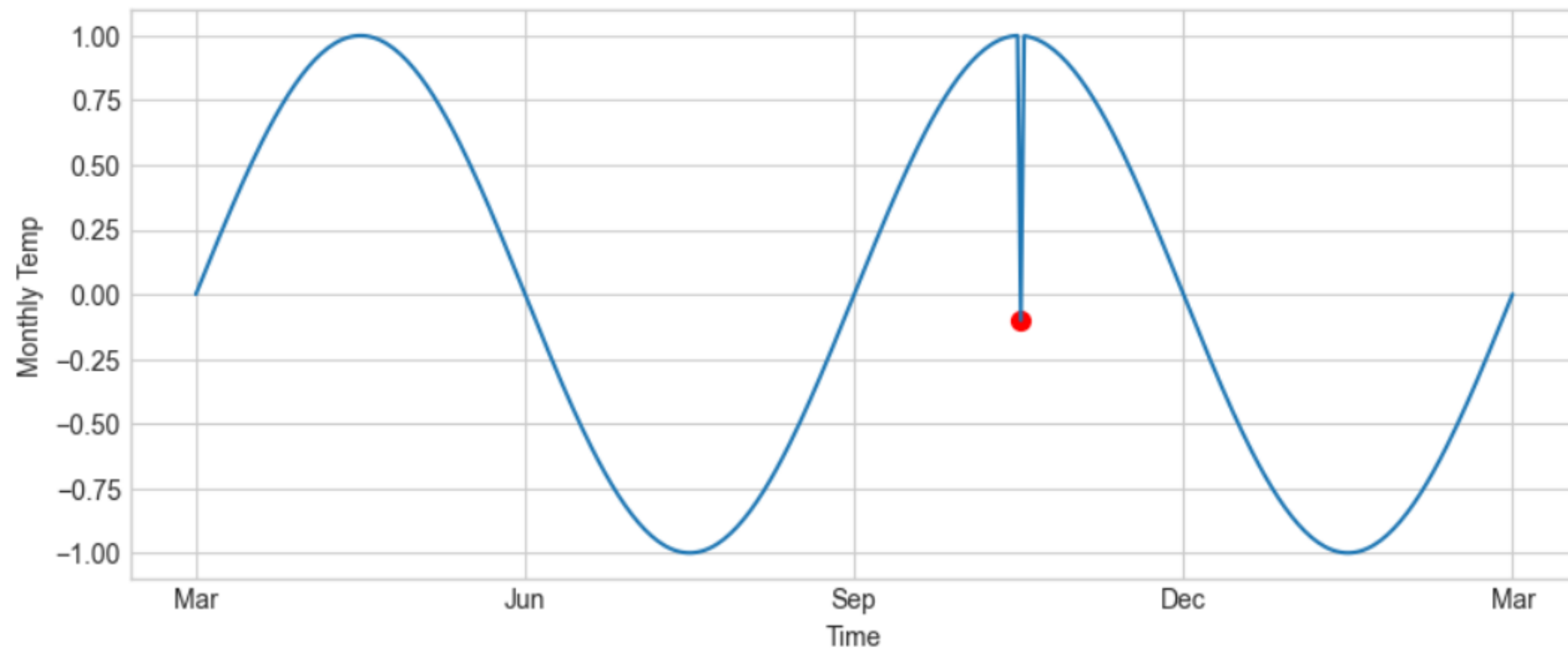
- Тренд — довгостроковий напрям змін
- Сезонність — регулярні коливання з фіксованою періодичністю
- Циклічність — довші нерегулярні коливання
- Залишкова складова — випадкові коливання

Типи аномалій

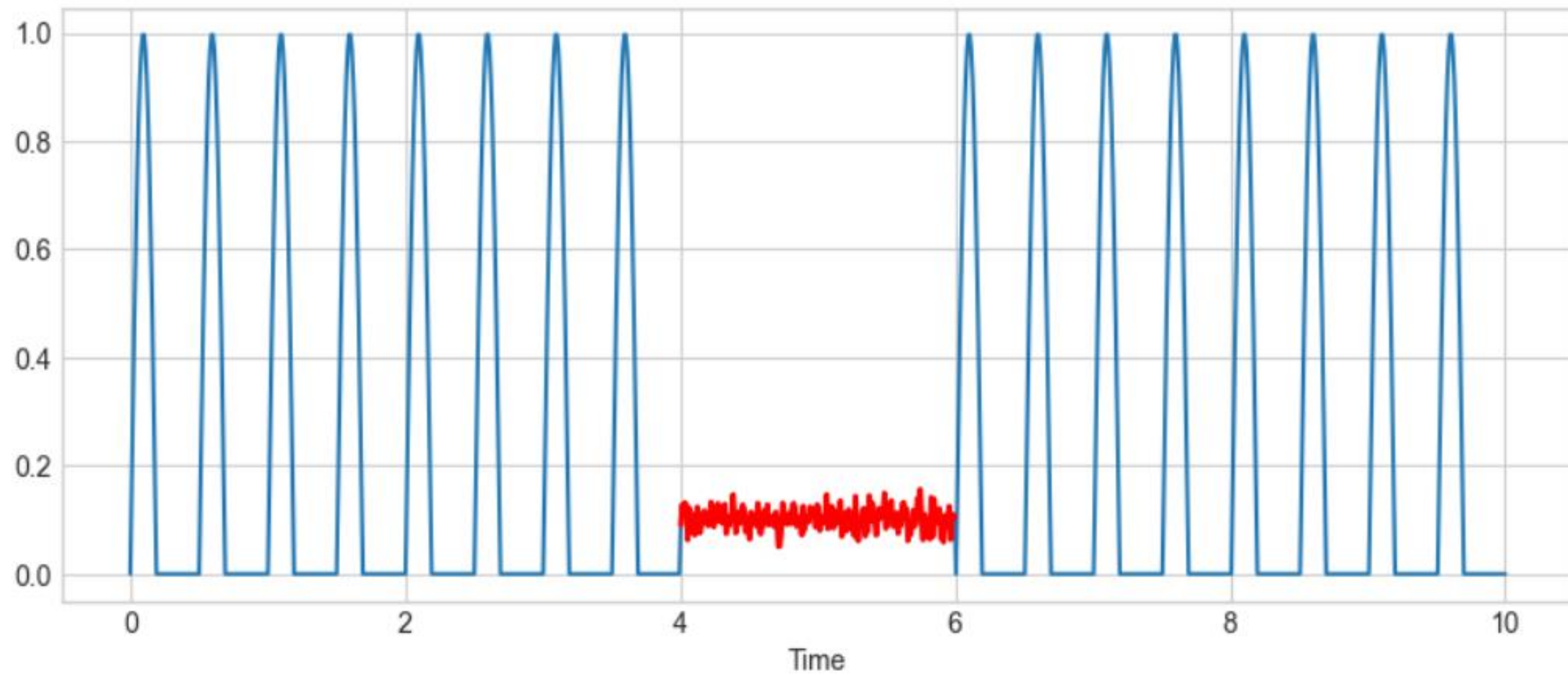
Точкові аномалії — окреме спостереження різко відрізняється від решти[2]



Контекстуальні аномалії — значення нормальне в одному контексті, але аномальне в іншому[2]



Колективні аномалії — аномальні підпоследовності точок[2]



Розглянуті групи методів за підходом до моделювання даних

Статистичні методи

явна модель
або розподіл,
задані
заздалегідь

ARIMA

Машинне навчання

закономірності з
даних без
припущень про
розподіл

XGBoost

Нейронні мережі

спеціалізовані
архітектури з
власним механізмом
моделювання

LSTM

ARIMA

Поєднує авторегресію, ковзне середнє та диференціювання. Описує значення стаціонарного ряду як лінійну комбінацію p попередніх значень та q попередніх помилок прогнозування[3]:

$$X'_t = c + \underbrace{\sum_{i=1}^p \varphi_i X'_{t-i}}_{\text{AR-частина}} + \underbrace{\sum_{j=1}^q \theta_j \varepsilon_{t-j}}_{\text{MA-частина}} + \varepsilon_t$$

Переваги

- невелика кількість параметрів
- інтерпретовані коефіцієнти
- не потребує обробки ознак

Недоліки

- описує лише лінійні залежності

XGBoost

Будує ансамбль регресійних дерев, кожне з яких виправляє похибки попередніх[4]:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F}$$

Цільова функція:

$$\mathcal{L} = \underbrace{\sum_{i=1}^n l(\hat{y}_i, y_i)}_{\text{Функція втрат}} + \underbrace{\sum_{k=1}^K \Omega(f_k)}_{\text{Регуляризація}}$$

Переваги

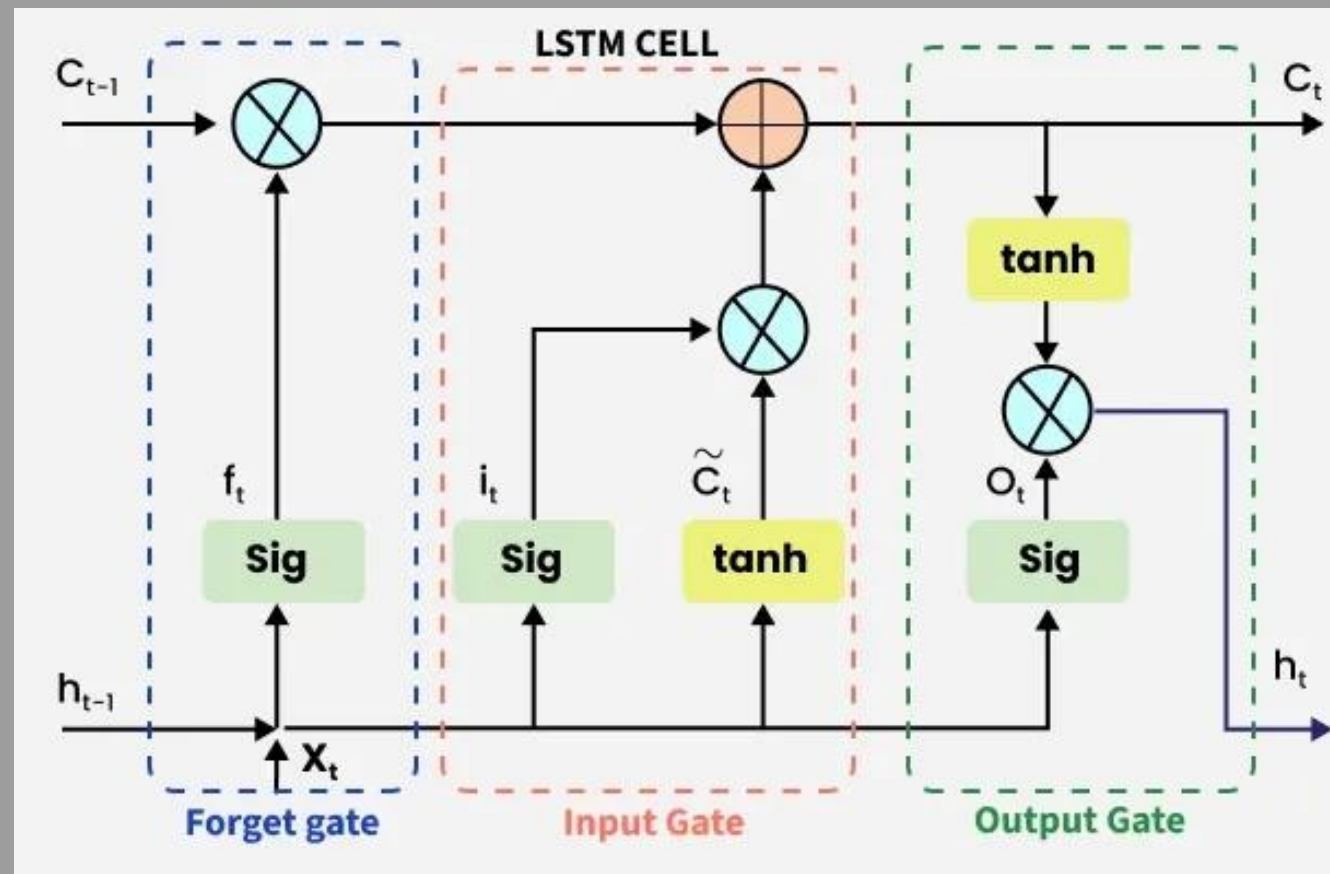
- моделює нелінійні залежності
- вбудована регуляризація
- паралельне обчислення при побудові дерев

Недоліки

- потребує формування ознак з ряду
- чутливий до вибору гіперпараметрів

LSTM

Рекурентна нейронна мережа зі спеціальним механізмом керування пам'яттю. Три типи воріт визначають, яку інформацію зберігати, передавати далі, а яку відкидати.[5]



$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$

Переваги

- моделює нелінійні і довгострокові залежності
- не потребує формування ознак

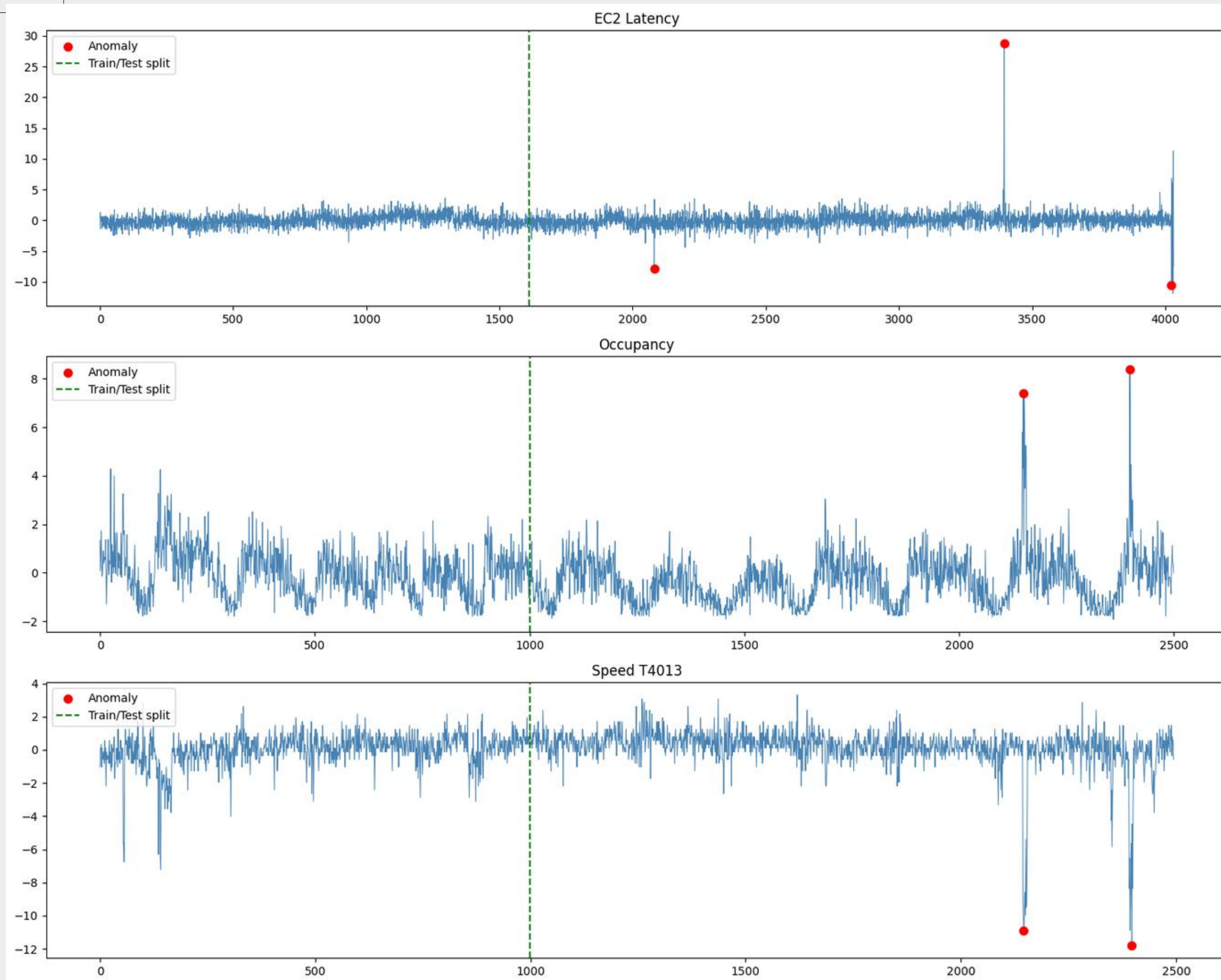
Недоліки

- складна внутрішня структура вимагає значних обчислювальних ресурсів
- низька інтерпретованість

Процедура порівняння

1. Модель навчається на нормальних даних
2. Прогноз для тестової вибірки
3. Обчислення абсолютної похибки: $e_t = |x_t - \hat{x}_t|$
4. Поріг за правилом k-сигм[6]: $T = \mu_e + k \cdot \sigma_e$
5. Точка з похибкою $e_t > T \rightarrow$ аномалія

Експериментальні дані[7]



Метрики оцінювання

F2-score[6]

$$F_2 = \frac{5 \cdot \text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}}$$

Час обчислення

$$t = t_{\text{навч}} + t_{\text{детекції}}$$

ROC-AUC[6]

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t)) dt$$

F2-score

Dataset	ARIMA	XGBoost	LSTM
EC2 Latency	0.7143	0.7500	0.7500
Occupancy	1.0000	1.0000	1.0000
Speed T4013	0.7143	0.8333	0.7692

ROC-AUC

Датасет	ARIMA	XGBoost	LSTM
EC2 Latency	0.9988	0.9990	0.9990
Occupancy	1.0000	1.0000	1.0000
Speed T4013	0.9973	0.9993	0.9997

Час обчислення

Датасет	ARIMA	XGBoost	LSTM
EC2 Latency	3.75	0.34	52.42
Occupancy	3.72	0.13	24.07
Speed T4013	4.91	0.07	24.32

Висновки

- Усі три методи забезпечують високу якість детекції аномалій
- ARIMA не поступається XGBoost та LSTM за показниками детекції
- XGBoost демонструє найкраще поєднання точності та швидкості
- Для одновимірних рядів з точковими аномаліями застосування методів машинного навчання та нейронних мереж не показало істотної переваги над статистичним підходом

Джерела

1. Shumway R.H., Stoffer D.S. Time Series Analysis and Its Applications: With R Examples. — 3rd ed. — Springer, 2011.
2. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. — 2009. — Vol. 41, No. 3. — P. 1–58.
3. Moschini G., Houssou R., Bovay J., Robert-Nicoud S. Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model // Engineering Proceedings. — 2021. — Vol. 5, No. 1. — P. 56.
4. Zhang L., Bian W., Qu W., Tuo L., Wang Y. Time Series Forecast of Sales Volume Based on XGBoost // Journal of Physics: Conference Series. — 2021.— Vol. 1873. — Art. 012067.
5. What is LSTM - Long Short Term Memory?— GeeksforGeeks. — Режим доступу:
<https://www.geeksforgeeks.org/deep-learning/deep-learning-introduction-to-long-short-term-memory/>
6. Komadina A., Martinić M., Groš S., Mihajlović Ž. Comparing Threshold Selection Methods for Network Anomaly Detection // IEEE Access. — 2024. — Vol. 12. — P. 124944–124972.
7. Ahmad S., Lavin A., Purdy S., Agha Z. Numenta Anomaly Benchmark. — Режим доступу:
<https://github.com/numenta/NAB>

Дякую за увагу!