

Lomonosov Moscow State University

Anticipating Data Demand in HEP: A Transformer Approach

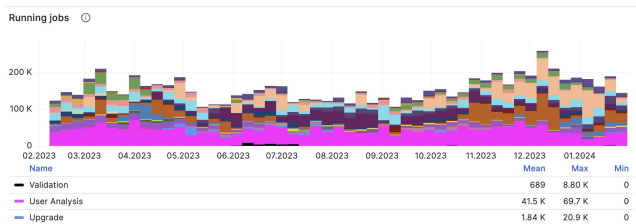
Mikhail Shubin, PHD Student, CMC MSU
Maria Grigorieva, Leading Specialist, RCC MSU
Nina Popova, Associate Professor, CMC MSU

11th International Conference “Distributed Computing and Grid Technologies in Science and Education”
(GRID’2025)

Dubna, July 08, 2025

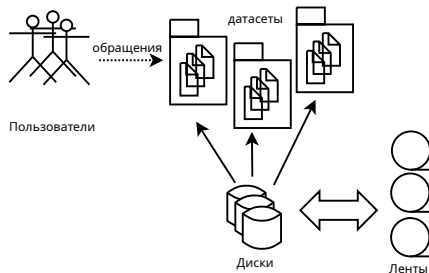
Введение

- Работа посвящена анализу паттернов обращений пользователей к данным в экспериментах НЕР с использованием глубоких моделей ИИ.
- Данные в распределённой системе представлены в виде **наборов логически сгруппированных файлов (datasets)**.
- Пользователи системы запускают **задания (tasks)** по обработке этих датасетов.
- Информация об обращениях заданий к датасетам сохраняется в **журналах**, которые подлежат анализу.
- Разрабатываемые подходы демонстрируются на примере данных эксперимента ATLAS на БАК. Подходы могут использоваться для других аналогичных систем.



Актуальность задачи

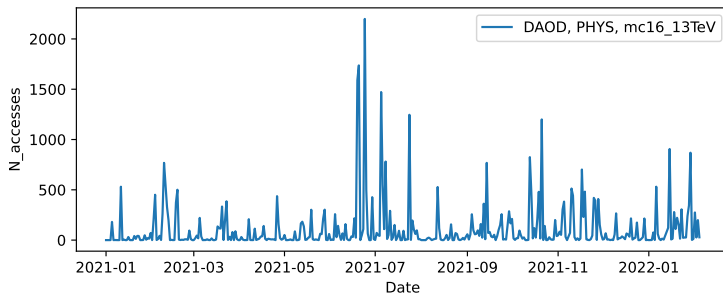
Задача прогнозирования востребованности (**популярности**) данных среди пользователей важна для принятия решений по размещению и кэшированию данных.



- Управление репликами датасетов является ключевым фактором, обеспечивающим эффективное функционирование грид-системы.

Понятие популярности данных

- Под **популярностью датасетов** будем понимать количество обращений к ним за определённый промежуток времени (день, неделя, месяц).
- **Популярность группы датасетов** определим как **суммарное** количество обращений к датасетам данной группы за определённый промежуток времени.
- Рассматриваемая величина представляет собой **временной ряд**



Задача прогнозирования популярности групп датасетов

Пусть $G = \{D_1, D_2, \dots, D_n\}$ — группа датасетов.

$D_i = \{a_{i1}, a_{i2}, \dots, a_{im}, \dots\}$ — временной ряд i -го датасета, где a_{im} — количество обращений к нему за промежуток времени с номером m .

Временной ряд группы датасетов определяется следующим образом:

$$\left\{ \sum_i^n a_{i1}, \sum_i^n a_{i2}, \dots, \sum_i^n a_{im}, \dots \right\}$$

Задача: По заданным m историческим значениям временного ряда найти k будущих значений:

$$\left(\sum_i^n a_{i1}, \sum_i^n a_{i2}, \dots, \sum_i^n a_{im} \right) \longrightarrow \left(\sum_i^n a_{i,m+1}, \dots, \sum_i^n a_{i,m+k} \right)$$

- Задача прогнозирования популярности групп данных на ежедневных и еженедельных временных рядах базовыми и нейросетевыми методами машинного обучения.

$$\left(\sum_i^n a_{i1}, \sum_i^n a_{i2}, \dots, \sum_i^n a_{im} \right) \longrightarrow \left(\sum_i^n a_{i,m+1}, \dots, \sum_i^n a_{i,m+k} \right)$$

(регрессия)

- Задача прогнозирования популярности отдельных датасетов путём разделения датасетов на популярные и непопулярные.

$$\{a_{i1}, a_{i2}, \dots, a_{im}\} \longrightarrow \{0, 1\}$$

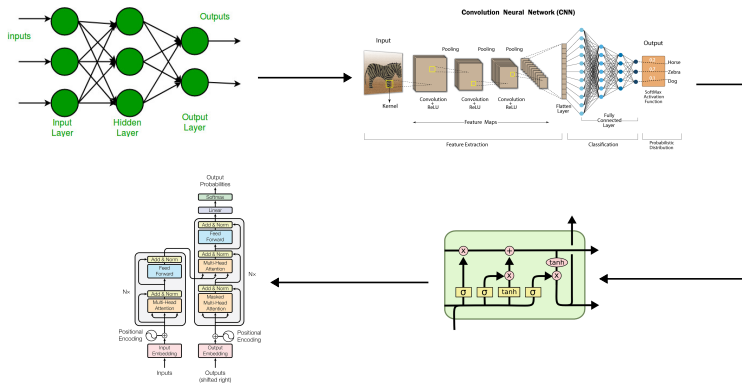
(бинарная классификация)

- M.A. Grigorieva, N.N. Popova, D.A. Vartanov, M.V. Shubin, **Exploring Hierarchical Forecasting of Data Popularity in High-Energy Physics Experiments**. Lobachevskii Journal of Mathematics, 2023, Vol. 44, No. 8, pp. 3076-3090.
- M.A. Grigorieva, N.N. Popova, D.A. Vartanov, M.V. Shubin, **Methodology of Data Popularity Forecasting in High-Energy Physics Experiments on Unbalanced and Irregular Time-series Data**. Lobachevskii Journal of Mathematics, 2024, Vol. 45, No. 7, pp. 3072-3084.

Методы прогнозирования

- Статистические:
 - Авторегрессия и производные модели (ARIMA, Seasonal ARIMA)
 - Facebook Prophet
- Классические методы машинного обучения
 - Логистическая регрессия
 - Случайный лес
 - SVM
 - XGBoost
 - CATBoost
- Нейросетевые методы
 - RNN
 - LSTM
 - GRU
 - Трансформеры: Informer, Autoformer, FEDformer, Chronos T5, Cross-Attention-only Time Series transformer (CATS)

Этапы развития нейронных сетей



Многослойный перцептрон, Свёрточная нейронная сеть, Рекуррентная сеть LSTM, Трансформер

Transformer-based архитектуры для временных рядов

- Informer — для прогнозирования длинных последовательностей данных, с использованием разреженного механизма внимания.
- Autoformer — модель, сочетающая идеи автокодировщиков и механизмов внимания. Учитывает сезонные и трендовые компоненты.
- FEDformer — (Frequency Enhanced Decomposition Transformer) использует частотное разложение для выделения различных компонентов временных рядов.
- CATS¹ — Cross-Attention-only Time Series transformer, — модель, в которой оставлен только механизм внимания, связывающий энкодер и декодер. Отличается легковесностью.
- Chronos² T5 — предобученное семейство моделей от Amazon основанных на языковой архитектуре T5

¹ Kim, Dongbin et al, Are Self-Attentions Effective for Time Series Forecasting? // Advances in Neural Information Processing Systems (2024).

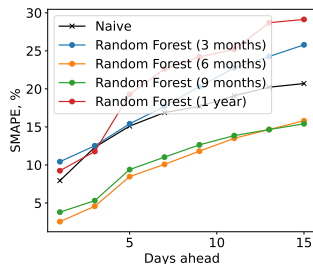
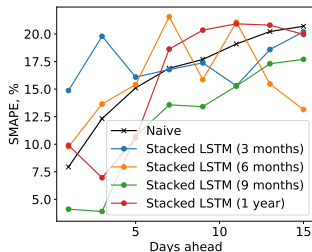
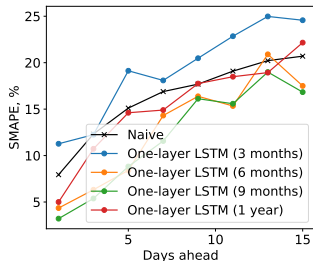
² Ansari, Abdul Fatir et al, Chronos: Learning the Language of Time Series // Transactions on Machine Learning Research (2024).

Кросс-валидация предиктивных моделей на примере группы PHYS

Длина общего временного ряда: 2,5 года начиная с 2021 года (853 дня).

Предобработка: сглаживание Moving Average, логарифмирование.

Результаты приводятся на логарифмической шкале.

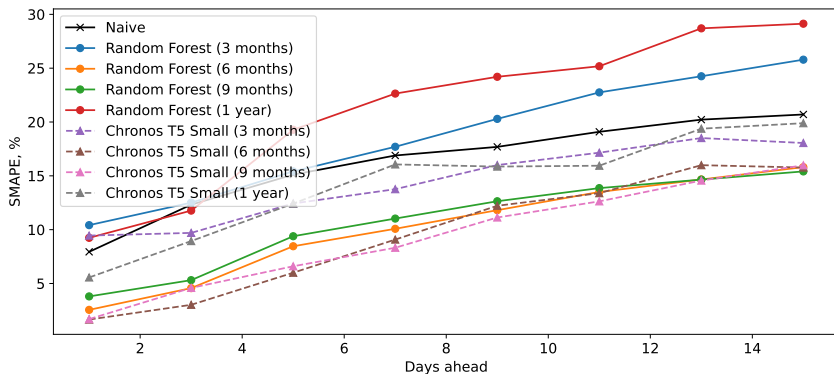


- Устойчивое превосходство по сравнению с наивным методом для выборки длиной 6-9 месяцев.
- Лучший результат показала модель на основе Случайного Леса.

$$SMAPE = \frac{100}{n} \sum_{t=1}^n \frac{|Forecast_t - Actual_t|}{(|Forecast_t| + |Actual_t|)/2}$$

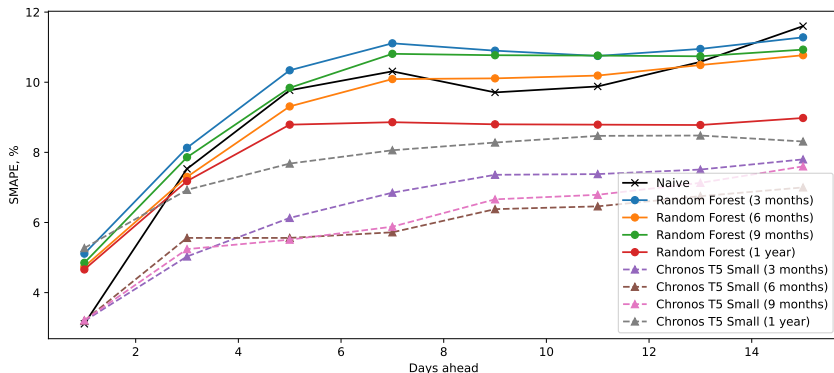
(Symmetric Mean Absolute Percentage Error)

Кросс-валидация модели Chronos T5 в сравнении с Random Forest (PHYS)



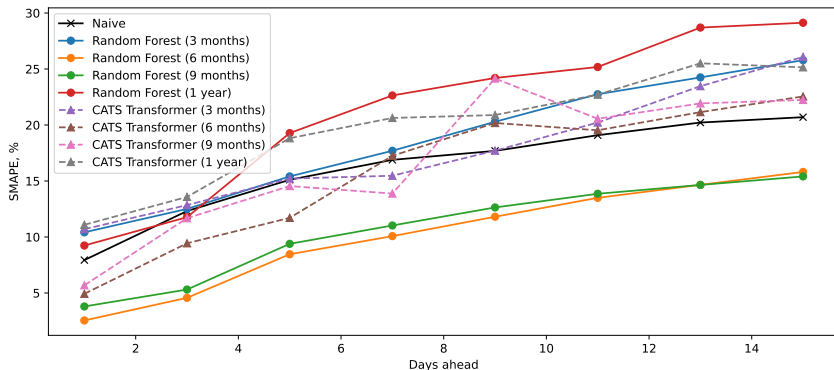
- Модель Chronos T5 превосходит наивный метод при любой длине входного временного ряда.
- На горизонтах предсказания 1-7 дней при длине входного временного ряда 6-9 месяцев трансформерная модель показывает себя немного лучше случайного леса.

Кросс-валидация модели Chronos T5 в сравнении с Random Forest (TOPQ)



- Chronos T5 демонстрирует лучший результат по сравнению со случайным лесом и наивным методом

Кросс-валидация модели CATS в сравнении с Random Forest (PHYS)



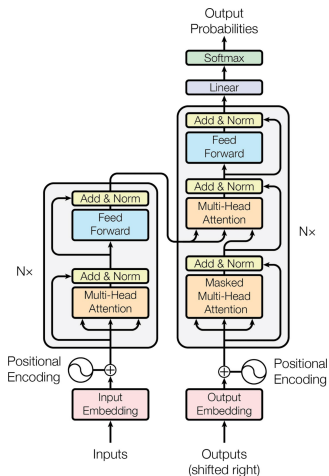
- Модель CATS превосходит наивный метод только для случая с длиной временного ряда 6 месяцев и на малых горизонтах предсказания, показывая себя хуже случайного леса
- Требуется дальнейшая настройка гиперпараметров обучения и модели.

Заключение

- Апробирован подход на основе трансформеров для прогнозирования популярности групп данных с различными параметрами длины обучающего временного ряда и горизонта предсказания.
- На примере предобученной трансформенной модели Chronos T5 продемонстрировано преимущество трансформенных моделей по сравнению с методами LSTM и Random Forest.
- Для обучения трансформерных моделей с нуля в этом подходе требуется настройка гиперпараметров архитектуры и обучения с учётом специфики временных рядов.
- Дальнейшие планы — автоматизация выбора модели, настройки её гиперпараметров и проведение обучения для указываемых пользователем данных.

Backup Slides

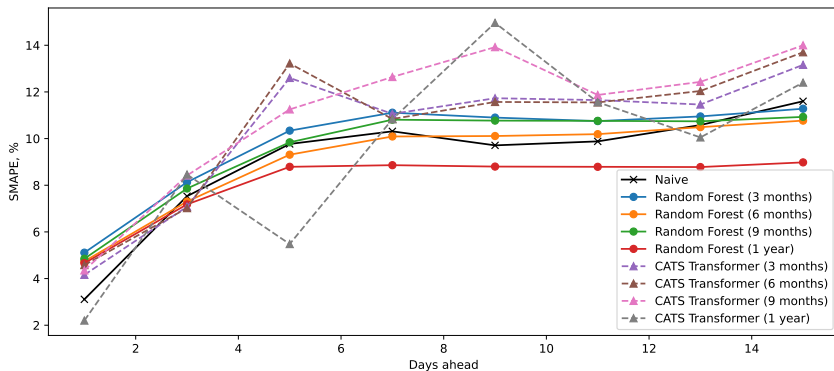
Архитектура трансформера



- Трансформер состоит из двух частей: **энкодера** и **декодера**.
- Энкодер и декодер состоят из N последовательных блоков, в каждом из которых содержатся слои с **механизмом внимания** и **полносвязные** слои.

Su, L., Zuo, X., Li, R. et al. A systematic review for transformer-based long-term series forecasting. Artif Intell Rev 58, 80 (2025).

Кросс-валидация модели CATS в сравнении с Random Forest (TOPQ)



- Модель CATS не показывает устойчивое превосходство наивного метода.
- Требуется дальнейшая настройка гиперпараметров модели.