https://doi.org/10.31891/csit-2024-1-7
UDC 004.9; 004.8

Kyrylo YEMETS
Lviv Polytechnic National University

TIME SERIES FORECASTING MODEL FOR SOLVING COLD START PROBLEM VIA TEMPORAL FUSION TRANSFORMER

Time series forecasting is an important tool in many businesses. It can range from efficiently allocating resources for web traffic, predicting patient needs for staffing requirements, to forecasting a company's product sales. A particular use case, known as "cold start" forecasting, involves making predictions for time series that have little or no historical data, like a new product just entering the retail market. The key assumption of cold start forecasting is that products with similar characteristics should have similar time series trajectories. In such scenarios, traditional forecasting models that heavily rely on past observations may face challenges, necessitating the development of innovative approaches that can effectively make predictions in the absence of a substantial historical dataset.

In this paper, Temporal fusion transformer neural network architecture was applied for solving cold start time series forecasting task. Modeling of the method was based on the use of a dataset contained in an open repository. After the preprocessing procedures, the dataset has about 370 time series, each of which has different length of series and has one categorical feature. Categorical feature have only 4 types of different values. For model training was performed to search for optimal hyperparameters across such parameters as: number of attention heads, learning rate, dropout percentage and hidden size.Model performed pretty well on this task. For model comparison were chosen metrics: MAE, RMSE, SMAPE. As can be seen from comparison with such popular models as DeepAR and LSTM, the proposed approach demonstrated the smallest forecasting error. Only one downside is that it can have more problems with anomalies in time series than DeepAR. But at the same time still provide interpretability of results.

Keywords: time series, cold start time series prediction, transformer, temporal fusion transformer.

Кирило ЄМЕЦЬ
Національний університет «Львівська політехніка»

МОДЕЛЬ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ДЛЯ ВИРІШЕННЯ ПРОБЛЕМИ "ХОЛОДНОГО СТАРТУ" ЗА ДОПОМОГОЮ TEMPORAL FUSION TRANSFORMER

Прогнозування часових рядів є важливим інструментом у багатьох бізнесах. Воно може варіюватися від ефективного розподілу ресурсів для веб-трафіку, прогнозування потреб пацієнтів для персоналу, до прогнозування продажів продукції компанії. Особливий випадок використання, відомий як прогнозування "холодного старту", передбачає створення прогнозів для часових рядів, які мають мало або зовсім не мають історичних даних, як-от новий продукт, що лише виходить на роздрібний ринок. Ключове припущення прогнозування "холодного старту" полягає в тому, що товари з подібними характеристиками повинні мати схожі траєкторії часових рядів. У таких сценаріях традиційні моделі прогнозування, які сильно залежать від минулих спостережень, можуть зіткнутися з викликами, що вимагає розробки інноваційних підходів, які ефективно прогнозують у відсутності значного історичного набору даних.

У цій статті була застосована архітектура нейронної мережі temporal fusion transformer для розв'язання завдання прогнозування часових рядів з "холодним стартом". Моделювання методу базувалося на використанні набору даних з відкритого репозиторію. Після процедур попередньої обробки, набір даних містив близько 370 часових рядів, кожен з яких мав різну довжину серії та одну категоріальну ознаку. Категоріальна ознака мала лише 4 типи різних значень. Для тренування моделі було проведено пошук оптимальних гіперпараметрів, таких як: кількість голів уваги, швидкість навчання, відсоток викидання та розмір прихованого шару. Модель показала досить хороші результати на цьому завданні. Для порівняння моделей були обрані такі метрики, як MAE, RMSE, SMAPE. Як видно з порівняння з такими популярними моделями, як DeepAR та LSTM, запропонований підхід продемонстрував найменшу помилку прогнозування. Єдиним недоліком є те, що він може мати більше проблем з аномаліями в часових рядах, ніж DeepAR. Але в той же час модель все ще забезпечує інтерпретованість результатів.

Ключові слова: часові ряди, прогнозування часових рядів з холодним стартом, трансформер, темпоральний фузійний трансформер.

Introduction

Time series forecasting is an important tool in many businesses, ranging from more efficient resource allocation for web traffic, electricity demand [1], mortality rates [2], biomedical data [3], predicting patient needs for staffing requirements, to forecasting a company's product sales [4]. One specific use case, known as "cold start" forecasting, involves making predictions for time series without any historical data, such as a new product just released into the retail market.

The main assumption of "cold start" forecasting is that products with near the same characteristics should have similar time series trajectories. This assumption allows forecasting "cold start" time series about products without any historical data, as illustrated in the following Figure 1. In a general case, it might not be just one categorical feature but also a dependent time series.

In time series forecasting are pretty popular classic methods, such as the Autoregressive Integrated Moving Average (ARIMA)[5] or Exponential Smoothing (ES)[6], heavily rely on the history of values for each individual

product, making them inefficient for "cold start" forecasting [7]. The classical approach to solving the cold start problem in time series without using models is the mean, or weighted average over samples with similar or identical characteristics. However, the results of using this approach are not always satisfactory.

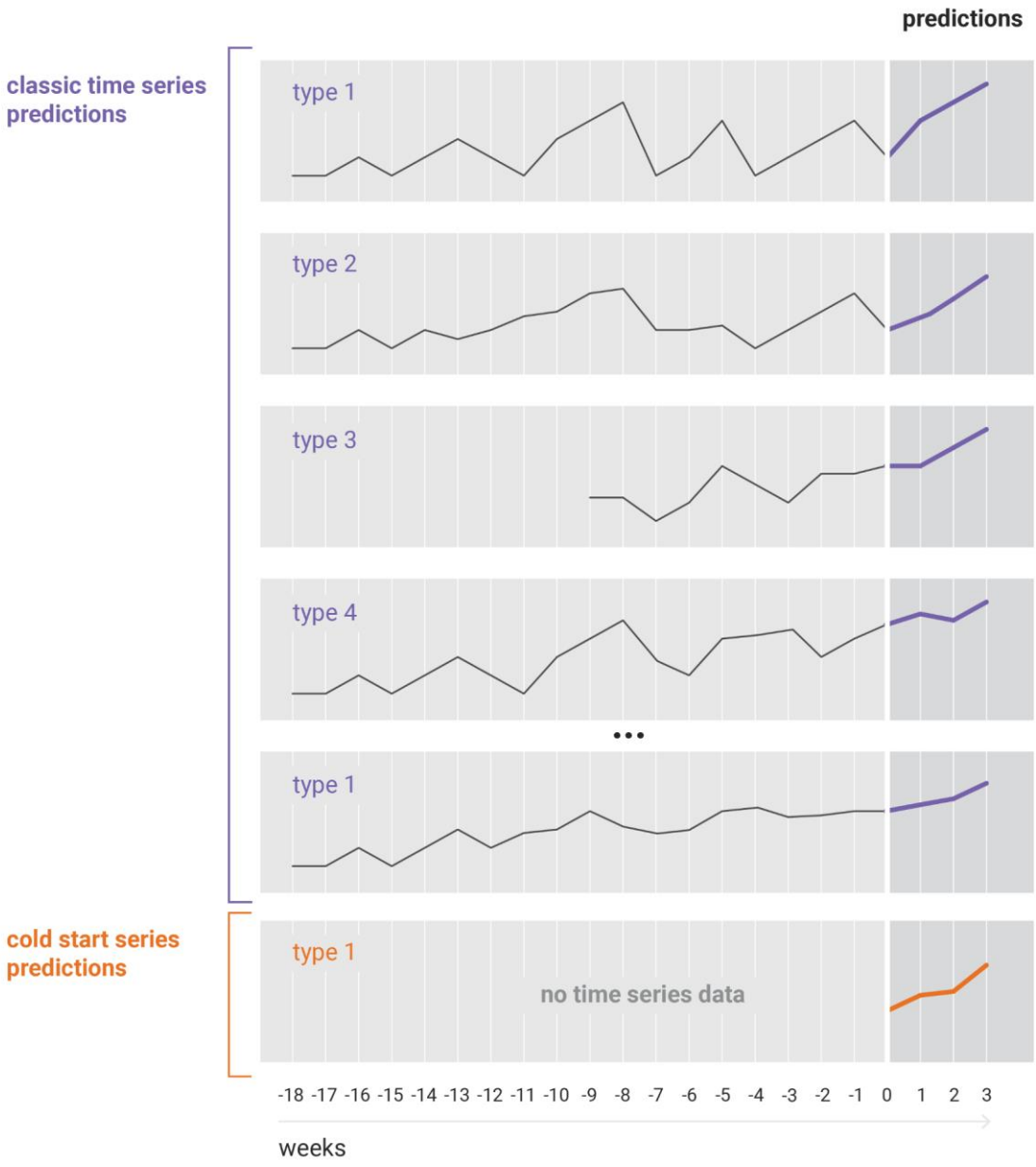


Fig. 1. Chart that show difference between cold start time series prediction and classic

State-of-the-arts

Modern time series forecasting methods for solving the cold start problem are based on the use of artificial neural networks. Examples of quite productive models used to solve this type of problem are DeepAR[8], LSTM, LSTM+GNN[9]. Let's take a closer look at them.

DeepAR[10] is a forecasting method developed primarily for time series data, using deep learning techniques. It's notable for its ability to model complex patterns in time series data and to handle large datasets with multiple correlated series. It adopts a probabilistic forecasting method, offering a distribution of possible future values rather than a single-point forecast. This probabilistic nature is especially beneficial for scenarios demanding risk management and planning under uncertainty. At its core, DeepAR utilizes Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks. LSTMs are adept at capturing long-term dependencies in time series data, addressing a common challenge in time series forecasting. The model's autoregressive nature means it uses past values of the target variable to predict future values, blending traditional time series approaches like ARIMA with the advanced capability of deep learning to model complex non-linear relationships. DeepAR stands

out in its ability to incorporate both static (time-invariant) and dynamic (time-variant) covariates, allowing the inclusion of additional influential information, such as economic indicators or day of the week. A key advantage of DeepAR is its scalability and efficient parallel training. Designed to train on many related time series simultaneously, it leverages similarities across these series for improved forecast accuracy and efficient handling of large datasets. Furthermore, the model is equipped to handle time series with missing values and those of varying lengths, which is a common challenge in real-world data. In essence, DeepAR's integration of deep learning with time series forecasting principles makes it a potent tool for complex forecasting tasks, surpassing traditional models in handling non-linear patterns and multiple, correlated time series.

DeepAR, a probabilistic forecasting method, relies on historical data to model complex patterns in time series. In cold start situations, where limited historical data is available, DeepAR may struggle to capture the underlying dynamics of the series, leading to less accurate forecasts. Additionally, its probabilistic nature, while advantageous in estimating uncertainty, may be compromised in cold start scenarios due to insufficient data to accurately model the distribution of future values.

Long Short-Term Memory (LSTM) networks [11], a type of Recurrent Neural Network (RNN) architecture, were introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997. They were designed to overcome the limitations of traditional RNNs, especially in capturing long-term dependencies in sequential data. The core innovation of LSTMs lies in their specialized structure, which includes gates that regulate the flow of information. These gates - the forget gate, input gate, and output gate - collaboratively decide what information to retain or discard, how to update the cell state with new information, and what to output based on the current input and the previous cell state. This structure enables LSTMs to make selective decisions about retaining and passing on information, which is crucial for processing sequences over long periods. The cell state in LSTMs acts as a conveyor belt, running straight through the entire chain of the model with minimal changes. This unique feature allows them to carry information across many time steps, effectively addressing the long-term dependency problem that plagues traditional RNNs. LSTMs have found extensive applications in various fields. They are a staple in Natural Language Processing for tasks like machine translation, text generation, and speech recognition. In Time Series Analysis, they are used for forecasting and anomaly detection, applicable in domains like finance and weather prediction. Moreover, LSTMs are employed in processing sequential data, such as in video analysis and music composition. The advantages of LSTMs are significant. They are adept at learning and remembering over long sequences, a critical factor in processing sequential and time-series data. This flexibility to handle not just individual data points but entire sequences makes them versatile for a range of applications.

However, LSTMs come with their challenges. They are computationally intensive due to their complex internal structure. Training LSTMs can be a demanding process, often requiring extensive hyperparameter tuning to optimize performance.

LSTM networks, known for their ability to capture long-term dependencies in sequential data, also face difficulties in cold start scenarios. The lack of sufficient historical data can hinder the LSTM's ability to learn and remember important patterns, resulting in poor forecast accuracy. Moreover, the complexity of LSTM architectures makes them computationally intensive and prone to overfitting, especially when trained on limited data.

But from neural network architectures for forecasting time series with a cold start, transformers weren't in use. [9]. That is why, this paper aims to investigate the efficiency of using the Temporal fusion transformer neural network architecture for solving cold start time series forecasting tasks.

Materials and methods

One of the most advanced architecture based on transformers for time series forecasting is the Temporal Fusion Transformer[13].

The Temporal Fusion Transformer (TFT) is a sophisticated model for time series forecasting that integrates several components to handle different types of input data and temporal relationships. Its architecture includes gating mechanisms that allow the model to adapt its complexity based on the data, variable selection networks for identifying relevant input variables, and static covariate encoders that incorporate static information into the network.

The model employs a sequence-to-sequence approach for local processing and an interpretable multi-head attention mechanism for learning long-term dependencies. This attention mechanism is a modified version of the standard multi-head attention found in transformer models, designed to improve interpretability. Additionally, the TFT uses a temporal fusion decoder, which leverages a series of layers to learn temporal relationships within the data.

Finally, the TFT generates forecasts using quantile outputs, allowing for the prediction of various percentiles at each time step, which is crucial for probabilistic forecasting. This architecture makes the TFT a powerful tool for multi-horizon forecasting tasks, capable of handling complex datasets with a mix of static and time-varying inputs.

The major constituents of TFT are:

- Gating Mechanisms: Variable Selection Networks: These networks are designed to automatically identify and select the most relevant input features at each time step. This feature selection is critical for enhancing

the model's performance, as it ensures that only the most informative parts of the data are used for forecasting.

- Static Covariate Encoders: These encoders handle static, time-invariant data. Unlike dynamic data that changes over time, static data remains constant. Incorporating this type of data effectively into the model is essential for enriching the context of the forecasts and improving overall prediction accuracy.
- Temporal Processing: This refers to how the model processes and learns from time-series data. It involves understanding and capturing both short-term and long-term temporal relationships within the data. This processing is key to effectively predicting future values based on past trends and patterns.
- Prediction Intervals: The model's ability to provide prediction intervals is a significant aspect. Instead of providing a single point forecast, it offers a range of possible future values, usually in the form of quantiles. This probabilistic forecasting approach is particularly useful for risk assessment and decision-making under uncertainty.

Diagram of model architecture represented in Fig. 2.

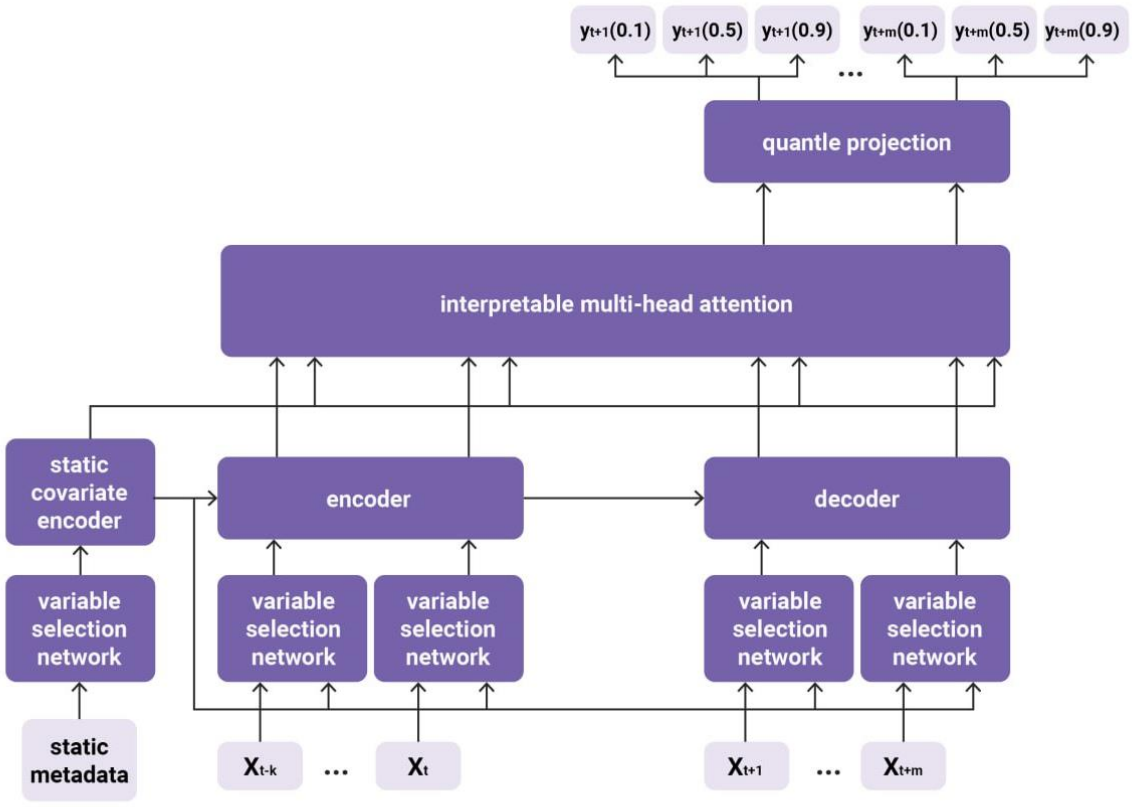


Fig. 2 Temporal Fusion Transformer architecture

The Temporal Fusion Transformer (TFT) is particularly effective for cold start time series forecasting due to several key aspects of its design. Firstly, its capability to incorporate and leverage different types of data inputs, including static and known future covariates, makes it highly adaptable in scenarios with limited historical data. The integration of these varied data types allows for a more comprehensive analysis, compensating for the lack of extensive past time series data. Furthermore, the model's variable selection networks are instrumental in identifying and focusing on the most relevant features available. This feature is crucial in cold start scenarios, where the available data might not be abundant but is still critical for making accurate predictions. Another significant aspect of TFT is its approach to probabilistic forecasting. By generating prediction intervals instead of single-point forecasts, TFT provides a range of possible future outcomes. This is particularly valuable in new and uncertain time series scenarios common in cold start problems, where the ability to quantify forecast uncertainty is vital for informed decision-making. TFT's inherent flexibility and adaptability, stemming from its architecture, enable it to handle the unique challenges presented by cold start forecasting. It can adapt to different data environments efficiently, making it suitable for situations where traditional models, which rely on extensive historical data, might not perform well.

In summary, the Temporal Fusion Transformer's design and capabilities, including its integration of various data types, variable selection mechanisms, and probabilistic forecasting approach, make it an effective tool for tackling the challenges of cold start time series forecasting.

Modeling, results, and comparison.

In the study, a synthetic dataset was utilized, based on electricity usage, comprising hourly time series for 370 distinct items, each identified by an item_id ranging from 0 to 369. This synthetic dataset also assigns each item_id a static feature, a characteristic that remains constant over time. The objective in training the TFT model is to understand the standard behavior patterns of items that are similar and apply this knowledge to forecast for new items (item_id 370–373), which lack historical time series data.

Also in this dataset, a "cold start" forecasting approach is used with only one static characteristic, in practice the presence of informative and high-quality static characteristics is key to successful "cold start" prediction.

Time series require that static characteristics be represented in numeric format. This can be achieved by applying LabelEncoder() to our static characteristic, where encoding is performed according to the following scheme A=0, B=1, C=2, D=3. Only one static characteristic will be input to the prediction network.

Model training was accompanied by a stage of optimizing its hyperparameters. The search was carried out using the optuna library. TPEsampler was chosen for sampler, and Hyperband for pruner. The parameters that were optimized are the number of attention heads from 1 to 4, learning rate from 0.001 to 0.1, dropout percentage from 0.1 to 0.3 and hidden size from 8 to 128.

To evaluate the effectiveness of the model under study, a number of performance indicators were used, in particular: MAE (mean average error), RMSE (root mean squared error), SMAPE (symmetric mean absolute percentage error). The formulas for these metrics are presented below:

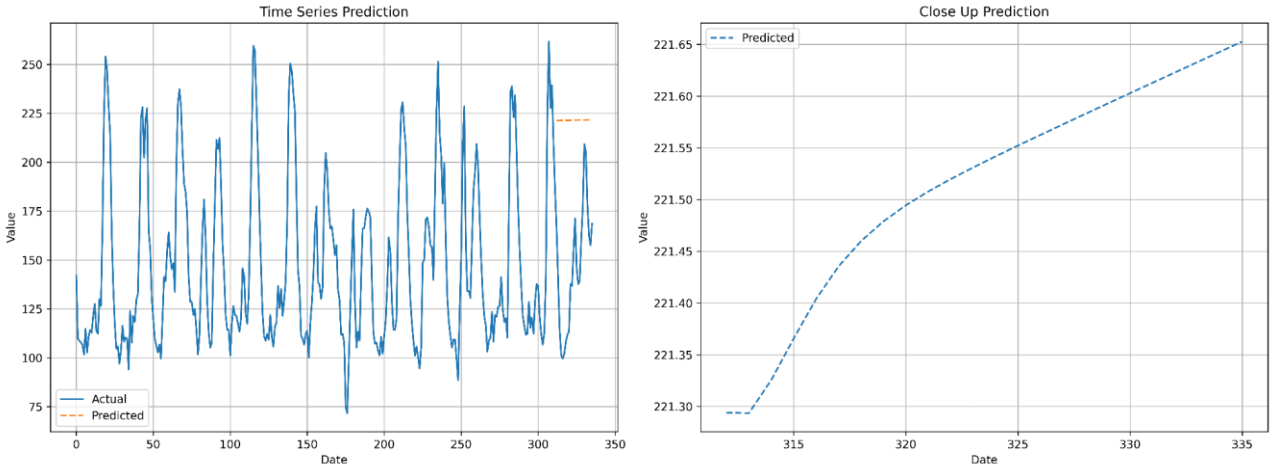
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |actual - forecast| \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(forecast - actual)2}{n}} \quad (2)$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|forecast - actual|}{(|forecast| - |actual|)/2} \quad (3)$$

where n — number of time serieses, forecast — value from model for each time series, actual — ground truth value for each time series.

The simulation was carried out on the basis of the dataset described above. The results of the model under study for predicting time series with a cold start demonstrated decent results. In particular, the MAE error is 165.15. To visually evaluate the results of the TFT, Figure 2 shows graphs. In particular, the left graph shows the prediction of the time series taking into account only its category, where the blue graph is the real one, and the orange one is the prediction from the moment of cutting off the time data. The right graph shows a detailed result of predicting for a new product taking into account historical data on similar existing products.



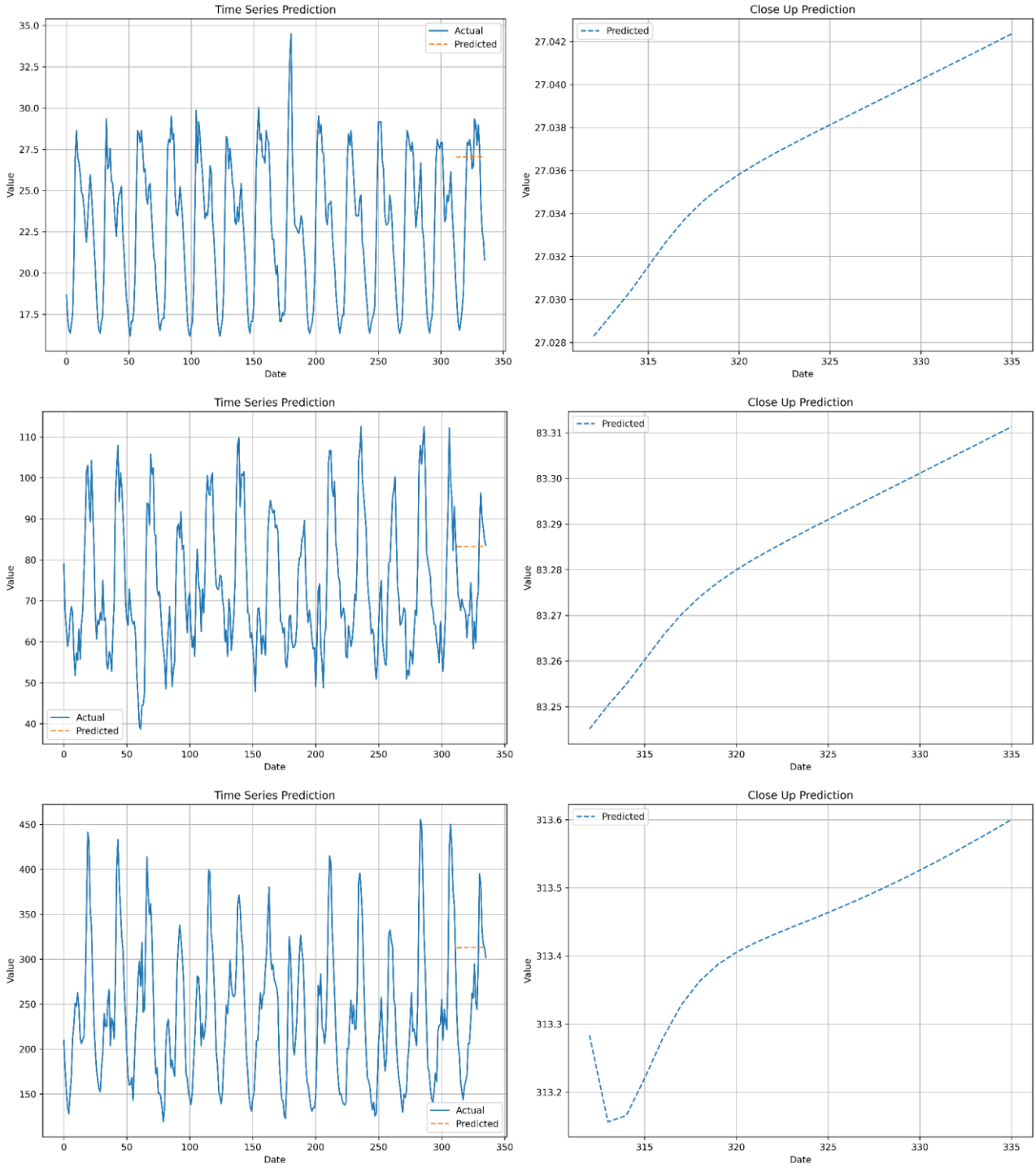


Fig. 3 Predictions of TFT and close up predictions

For comparing the efficiency of TFT, popular models for solving this problem were chosen: DeepAR, LSTM.

The results of this comparison are shown in Table 1. To evaluate the effectiveness of the investigated approach, Table 1 summarizes the values of metrics (1)-(3) of all methods investigated in this work. Table 1 shows that the Baseline method (average value for samples with similar or identical characteristics) demonstrates the worst forecast result. The use of LSTM showed a slight improvement according to MAE and SMAPE. The DeepAR architecture was next in terms of accuracy. The most accurate forecast results are demonstrated by the investigated TFT architecture. In particular, it shows more than 22% less MAE error compared to the Baseline method and 20% less MAE value than DeepAR.

Table 1

Comparison of metrics of different models

Method/Performance indicator	Baseline	LSTM	Deep-AR	TFT(our)
MAE	213.9187	210.9327	206.6158	165.1492
RMSE	871.3490	854.2674	828.1398	1022.9039
SMAPE	0.5387	0.4523	0.4138	0.4103

The results presented in Table 1 indicate that the Temporal Fusion Transformer (TFT) model outperforms the baseline, LSTM, and DeepAR models in terms of Mean Absolute Error (MAE) and Symmetric Mean Absolute Percentage Error (SMAPE). Specifically, the TFT model achieves the lowest MAE of 165.1492, which is a significant improvement over the other models, indicating its superior accuracy in predicting the central tendency of the time series. Additionally, the TFT model also has the lowest SMAPE value of 0.4103, suggesting that it provides more accurate percentage predictions, which is particularly useful in scenarios where the scale of the data varies.

However, it is important to note that the TFT model has a higher Root Mean Squared Error (RMSE) value of 1022.9039 compared to the DeepAR model's RMSE of 828.1398. This suggests that while the TFT model is generally more accurate in its predictions, it may be more sensitive to anomalies and outliers in the data, leading to larger errors in some cases.

Overall, the TFT model is considered the best model based on these results because it consistently provides more accurate predictions in terms of both MAE and SMAPE. Its ability to handle complex temporal relationships and incorporate multiple input features effectively contributes to its superior performance. However, its sensitivity to outliers should be taken into consideration when applying it to datasets with significant anomalies.

Conclusions

In this paper, the urgent task of cold start time series forecasting was solved in many areas. The author investigated the effectiveness of using TFT to solve it. The basic architecture of TFT, the principles of its operation and the advantages during application for solving the set task are described.

Modeling was performed using a dataset of electricity consumption and transformed it to a dataset for cold start prediction by cutting historical data for train samples. Dataset consists of 370 time series with 1 category that includes 4 types of consumption.

For comparing results on this dataset were chosen metrics MAE, RMSE, SMAPE. Models for comparing results were LSTM, DeepAR and simple mean for each category (який обрано як бейслайн метод).

For training Temporal Fusion Transformer selection procedures were carried out for best hyperparameters and after that trained it with early stopping callback.

Results of modeling showing that in comparing the Temporal Fusion Transformer (TFT), DeepAR, and LSTM models for cold start time series predictions, TFT stands out due to its advanced architecture and high performance. TFT's ability to integrate various data types and employ probabilistic forecasting makes it particularly effective in scenarios with limited historical data. DeepAR and LSTM, while powerful in their own right, may not match the adaptability and accuracy of TFT in cold start conditions. Looking forward, the superiority of TFT in handling complex, data-sparse environments suggests a significant potential for more accurate and reliable forecasting in various industries, driving smarter decision-making and better resource allocation in the future.

Acknowledgment. This research is supported by the EURIZON Fellowship Program: “Remote Research Grants for Ukrainian Researchers”, grand № 138.

References

[1] Kholiavka, Y., & Parfenenko, Y. (2023). FORECASTING PEAK LOAD ON THE POWER GRID. Computer Systems and Information Technologies, (3), 12–22. https://doi.org/10.31891/csit-2023-3-2

[2] POPOVYCH, A. ., & YAKOVYNA, V. (2022). COVID-19 MORTALITY PREDICTION USING MACHINE LEARNING METHODS. Computer Systems and Information Technologies, (2), 104–111. https://doi.org/10.31891/csit-2022-2-12

[3] IZONIN, I. (2023). AN UNSUPERVISED-SUPERVISED ENSEMBLE TECHNOLOGY WITH NON-ITERATIVE TRAINING ALGORITHM FOR SMALL BIOMEDICAL DATA ANALYSIS. Computer Systems and Information Technologies, (4), 67–74. https://doi.org/10.31891/csit-2023-4-9

[4] Hovorushchenko T, Medzatyi D, Voichur Y, Lebiga M. Method for forecasting the level of software quality based on quality attributes. 2023, Journal of Intelligent & Fuzzy Systems, 1-15. https://doi.org/10.3233/JIFS-222394.

[5] Hussan A, Yamur K. A, Jan L. Time Series Forecasting using ARIMA Model. ADVCOMP 2018. 1-4.

[6] Hyndman, R.J., & Athanasopoulos, G. Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. 2018. URL: https://otexts.com/fpp2/expsmooth.html.

[7] Christopher X, Alex T, Alec G, Alec G, Emily F. A Unified Framework for Long Range and Cold Start Forecasting of Seasonal Profiles in Time Series. arXiv. 2017. https://doi.org/10.48550/arXiv.1710.08473.

[8] Ivan C, Wenming Y. Build a cold start time series forecasting engine using AutoGluon. 2022. Amazon SageMaker, Artificial Intelligence. URL: https://aws.amazon.com/blogs/machine-learning/build-a-cold-start-time-series-forecasting-engine-using-autogluon/

[9] Zahra F, Minh H, Elena Z, Zamir S, Xiaojun D. Mitigating Cold-start Forecasting using Cold Causal Demand Forecasting Model. arXiv 2023. https://doi.org/10.48550/arXiv.2306.09261.

[10] David S, Valentin F, Jan G. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. arXiv 2019. arXiv https://doi.org/10.48550/arXiv.1704.04110.

[11] Sepp H, Jürgen S. Long short-term memory. Neural computation 9, 8 (1997), p.1735–1780.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser,I. Polosukhin, Attention is all you need, in: NIPS, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[13] Bryan L, Sercan A, Nicolas L, Tomas P. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. 2020. arXiv. URL: https://doi.org/10.48550/arXiv.1912.09363.

Kyrylo YEMETS **Кирило ЄМЕЦЬ**	Ph.D. student (Computer Science), at the Department of Artificial Intelligence, Lviv Polytechnic National University https://orcid.org/0000-0002-5157-9118 e-mail: kyrylo.v.yemets@lpnu.ua	Аспірант кафедри систем штучного інтелекту Національного університету «Львівська політехніка»