Validating Lag-Llama for Probabilistic Time Series Forecasting in the Indonesian Stock Market: A Comparative Study of Univariate and Multi Time Series

Arbi Haza Nasution [a,*], Anggi Hanafiah [a], Rajalingam Sokkalingam [b], Mohd Sham Mohamad [c], Andry Alamsyah [d] and Winda Monika [e]

[a] *Department of Informatics Engineering, Universitas Islam Riau, Riau 28284, Indonesia*
E-mail: arbi@eng.uir.ac.id; ORCID: https://orcid.org/0000-0001-6283-3217
[b] *Fundamental and Applied Science Department, Universiti Teknologi PETRONAS, Perak 32610, Malaysia*
E-mail: raja.sokkalingam@utp.edu.my
[c] *Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), Lebuh Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, Malaysia*
E-mail: mohdsham@umpsa.edu.my
[d] *School of Economic and Business, Telkom University, Bandung 40257, Indonesia*
E-mail: andrya@telkomuniversity.ac.id; ORCID: https://orcid.org/0000-0001-5106-7561
[e] *Department of Library Information, Universitas Lancang Kuning, Riau 28266, Indonesia*
E-mail: windamonika@unilak.ac.id; ORCID: https://orcid.org/0000-0002-4637-5735

Abstract. Accurately predicting stock prices is crucial for both investors and policymakers. This paper presents the first empirical evaluation of Lag-Llama, a novel probabilistic time series forecasting model, in predicting stock prices on the Indonesian Stock Exchange (IDX). By applying Lag-Llama to univariate and multi-time series forecasts of key IDX stocks, we assess its ability to capture temporal patterns and market volatility, particularly in comparison to established models like DeepAR (RNN) and Temporal Fusion Transformer (TFT). Our results show that, in fine-tuning scenarios, Lag-Llama achieves a Continuous Ranked Probability Score (CRPS) of 0.0195 for the combined BBCA, BMRI, and AMRT stocks, surpassing TFT (CRPS: 0.0179) and DeepAR (CRPS: 0.0270). However, forecasting across broader stock groups (Top 1-9 and Top 10-18 by market cap) presents more variability, with CRPS values rising to 0.0517 for the Top 1-9 stocks. This study demonstrates Lag-Llama's potential as a robust tool for stock price prediction, particularly for select stock groupings, offering enhanced precision and reliability compared to traditional methods.

Keywords: Lag-Llama, Probabilistic Time Series Forecasting, Stock Market Analysis

*Corresponding author. E-mail: arbi@eng.uir.ac.id.

1. Introduction

Recent advancements in machine learning have revolutionized various industries, from agriculture to healthcare and natural language processing. For instance, studies like [59] have demonstrated the effectiveness of machine learning in improving agricultural predictions, while [58] highlights its application in health for detecting critical conditions such as drowsiness. Similarly, [21, 43] showcase how machine learning can be adapted to low-resource language environments, improving sentiment analysis and bilingual lexicon induction for under-resourced languages, respectively. These diverse applications underscore the versatility of machine learning models in addressing complex, dynamic systems.

In the realm of finance, stock market prediction is a critical component of financial decision-making, influencing investment strategies, risk management, and policy formulation. The stock market is characterized by high volatility, with prices fluctuating rapidly due to factors like economic reports, news, and investor sentiment. With its high liquidity and diverse participants, stock market behavior is often shaped by external economic data, where positive news typically drives prices up and negative data leads to declines. While the stock market offers the potential for high returns, it also carries significant risks compared to more stable investments like bonds or savings accounts. Given these complexities, accurate forecasting models are essential for making informed decisions and managing financial risk effectively.

Stock market data exhibits complex characteristics, such as temporal structure, non-stationarity, volatility, high granularity, non-linearity, multivariate nature, and so on [5, 11, 27]. Understanding these characteristics is essential for applying appropriate analysis methods, which leads to more accurate and insightful forecasts and analyses.

The Indonesian Stock Exchange (IDX) showcases a significant component of the financial market in Southeast Asia. The IDX includes diverse market composition sectors, with a strong presence in finance, consumer goods, and mining industries [37]. This diversity can introduce unique sector-specific trends and volatility patterns [11]. Moreover, The IDX is quite sensitive to an economic and political climate where changes in government policies, infrastructure projects, and political stability lead to market reaction and trends, as well as the fluctuation in global commodity prices can heavily influence stock prices on the IDX. Also, IDX is also influenced by regional economic developments in Southeast Asia. These indicators and market performance can differ from those observed in more developed or international stock markets, adding another layer of complexity to market analysis.

Several techniques have been applied to assess portfolios to maximize profits while managing risks and promoting portfolio diversification. Time series analysis techniques are widely used in stock market forecasting to analyze sequential data patterns, helping to understand volatility, identify trends, and predict future prices using historical data [11]. Traditional statistical models, such as ARIMA, Support Vector Machines (SVM), and basic neural networks, have been commonly applied in the context of the Indonesian stock market. However, these models often fall short when faced with the high volatility and complex patterns of financial time series data.

Recent advancements have focused on incorporating deep learning techniques, which show promise for improved prediction accuracy. Models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks have been employed for stock price prediction, offering better handling of temporal dependencies. However, these models still struggle with long-term dependencies and uncertainty, which are critical in financial forecasting. Moreover, their performance can be limited by the challenges inherent in non-stationary data and the high volatility typical of financial markets.

In response to these limitations, probabilistic forecasting models like DeepAR and Temporal Fusion Transformer (TFT) have emerged as more robust alternatives, capable of handling complex temporal patterns and providing probabilistic forecasts. However, even these models face challenges in fully capturing uncertainty, especially in highly volatile markets like the Indonesian Stock Exchange (IDX). In this context, Lag-Llama, a probabilistic time series model leveraging the strengths of large language models (LLMs), offers a novel approach. By incorporating the ability to model uncertainty and long-term dependencies, Lag-Llama has the potential to improve stock price prediction where traditional and even advanced models face limitations.

The primary contribution of this study is the first validation of Lag-Llama, a recently introduced probabilistic time series forecasting model, in the context of stock price predictions for the Indonesian market. With stock markets being inherently volatile and complex, accurate forecasting models are essential for informed decision-making. While established models like DeepAR (RNN) and Temporal Fusion Transformer (TFT) have demonstrated significant promise, they still face limitations in handling long-term dependencies and uncertainty in financial time series data. Lag-Llama, leveraging the strengths of large language models (LLMs), aims to address these challenges by introducing a probabilistic approach that better captures temporal patterns and uncertainty.

This study benchmarks Lag-Llama against these established models, assessing its predictive accuracy in both zero-shot and fine-tuning scenarios. By focusing on the top 18 stocks listed on the IDX Market Cap for Q2 2024, we explore the effectiveness of Lag-Llama in various experimental settings, including different context lengths and learning rates. This research not only introduces Lag-Llama to the financial forecasting domain but also demonstrates its potential as a robust tool for handling the complexities of stock price prediction in emerging markets like Indonesia.

Our contribution in this paper is threefold: (1) the first empirical validation of Lag-Llama in predicting stock prices on the Indonesian Stock Exchange (IDX), offering insights into the applicability of large language models (LLMs) in financial forecasting; (2) a detailed analysis of the impact of different context lengths and learning rates on the performance of Lag-Llama, providing optimization strategies for its application in stock market prediction; and (3) a comprehensive comparison between Lag-Llama and established baseline models (DeepAR and Temporal Fusion Transformer), demonstrating Lag-Llama's potential advantages in handling complex temporal patterns and uncertainty, particularly in multi-time series forecasting.

2. Related Works

2.1. Prediction of International Stock Markets

The historical prediction of world stock markets has evolved significantly over the past several decades, incorporating various methodologies and factors to enhance accuracy. Initially, predictions relied heavily on fundamental data such as interest rates, currency exchange rates, inflation rates, trading volume, and annual returns for listed corporations [45]. However, the last six decades have seen the introduction of new predictive factors, including options strike rates, psychological price barriers, and the impact of national and international crises, such as the COVID-19 pandemic [45]. Machine learning models, particularly Long Short-Term Memory (LSTM) networks, have become increasingly popular due to their ability to capture long-term dependencies in historical data, outperforming traditional models like Random Forest and Linear Regression in terms of accuracy and reliability [24]. The integration of both historical prices and news data through hybrid models combining RNN-LSTM and CNN networks has further improved prediction accuracy, demonstrating the importance of considering multiple data sources [57]. Additionally, regression trees and multifactor models have shown that stock markets are statistically predictable on an economically interesting scale, with factors such as value and momentum playing crucial roles, although their effectiveness can vary over time [61]. A comprehensive review of prediction strategies over the last 50 years highlights the shift towards machine learning methods, which have proven to be highly effective in drawing generalized patterns from input data to produce desired outputs [48]. This continuous innovation in predictive methodologies underscores the importance of adapting to new data and techniques to navigate the complexities of financial markets effectively.

2.2. Prediction of Indonesian Stock Markets

The historical prediction of the Indonesian stock market has seen various methodologies and models being employed to enhance accuracy and reliability. One prominent approach is the use of the Long Short-Term Memory (LSTM) algorithm, which has been applied to predict the closing prices of stocks such as BBRI, demonstrating good model performance with low error rates, indicating its effectiveness in recognizing complex temporal patterns in financial data [56]. Additionally, a comparative study of different deep learning architectures, including CNN, GRU, LSTM, and GCN, identified the TFGRU architecture as the best performer for stock price prediction among 315 companies listed on the Indonesia Stock Exchange (IDX), showcasing the potential of deep learning in financial forecasting [18]. Another significant finding is the predictive power of U.S. stock market skewness on Indonesian stock market returns, where an increase in U.S. market skewness is associated with a decrease in Indonesian market returns in the following month, providing a strategic insight for investors [33]. Furthermore, the development of a stock price prediction website using LSTM for major Indonesian banks has shown high accuracy, with a Mean Absolute Percentage Error (MAPE) of less than 10%, making it a valuable tool for investors [6]. Lastly, risk assessment models like the Stochastic model (Geometric Brownian Motion and Jump Diffusion) have been used to predict the JKSE index, with the Jump Diffusion model achieving a MAPE value of 1.08%, and the Adjusted Expected Shortfall model providing insights into potential loss risks, thus aiding in better risk management [34]. These diverse methodologies highlight the evolution and sophistication of stock market prediction techniques in Indonesia, contributing to more informed investment decisions.

2.3. Stock Market Prediction Techniques

Stock market prediction has seen a wide range of methodologies, evolving from statistical approaches to modern deep learning techniques.

2.3.1. Statistical Approaches

The ARIMA (AutoRegressive Integrated Moving Average) model has been a cornerstone in time series forecasting, providing a baseline for linear trend prediction [4]. Despite its simplicity, ARIMA struggles with non-linear patterns and long-term dependencies.

Time series research with a statistical approach involves analyzing sequences of data points collected or recorded at specific time intervals to extract meaningful statistics and predict future values. This method is particularly valuable in uncertain times, such as during pandemics, wars, and economic fluctuations, where predicting business outcomes becomes challenging. Statistical forecasting methods, such as time series analysis and linear regression, can help organizations smooth out fluctuations, eliminate seasonality, and achieve accurate predictions, as demonstrated in the study of sales budgets in the glass mould manufacturing industry [60]. Time series data can be collected at various intervals, from seconds to years, and is used in diverse fields, including finance, population studies, and sales forecasting [17, 71]. A notable approach in time series research is the use of N-grams from computational linguistics, which allows for the prediction of dynamic and non-stationary time series without requiring extensive preliminary studies or complex parameter tuning. This method is advantageous for its high level of automation and applicability to large complex systems, making it suitable for monitoring and forecasting in content monitoring systems and other areas characterized by trends and cyclicality [29, 70]. Overall, time series analysis remains a prominent statistical tool, leveraging historical data and current trends to provide valuable insights and support decision-making in various domains [17, 29, 60, 70, 71].

2.3.2. Traditional Machine Learning Models

Methods like logistic regression, decision trees, and support vector machines have been applied to stock market prediction. These models, while effective for certain tasks, often require extensive feature engineering and may not fully capture the temporal dependencies in stock data.

Time series research with machine learning, particularly regression models, has shown promising results across various applications. For small-sample, multi-feature time series data, different machine learning models such as KNN, decision tree, random forest, multilayer perceptron (MLP), support vector regression (SVR), and ridge regression have been evaluated. Ridge regression demonstrated superior generalization capabilities, while SVR was found suitable for nonlinear data, and ensemble learning models like random forests outperformed single learners in generalization [68]. Automated Machine Learning (AutoML) tools like AutoGluon, Auto-Sklearn, and PyCaret have also been explored for time series analysis, revealing that their performance is highly dataset-dependent, emphasizing the need for dataset-specific considerations in time series forecasting [63]. In the realm of web analytics, a study involving five years of daily time series data for website traffic measures utilized a voting regression model combining Decision Tree Regression, Multi Linear Regression, and Support Vector Machine Regression, achieving a prediction accuracy of 99.96% and an absolute error of 0.24% [52, 72]. Additionally, exponential smoothing models have been developed for time series data, achieving a good fit with an R^2 of 0.984 and passing residual tests, indicating high reliability with a relative error of 7.53% [12]. These studies collectively highlight the effectiveness of various regression models and AutoML tools in time series research, demonstrating their potential in accurately forecasting trends and managing the complexities inherent in time series data. The ongoing research and development in this field aim to further enhance the application of these models, making them more robust and adaptable to different datasets and forecasting needs.

2.3.3. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) Networks

These models have shown significant improvements in handling sequential data. RNNs and LSTMs can capture temporal dependencies and non-linear patterns, making them suitable for stock market prediction. However, they still face challenges with long-term dependencies.

DeepAR is a deep learning-based approach for probabilistic time series forecasting that leverages autoregressive recurrent networks to model the distribution of future values conditioned on past observations [46]. This method is particularly effective in capturing complex temporal dependencies and generating probabilistic forecasts, which are crucial for decision-making in various domains. Integrating chaotic systems with deep learning, as explored by Jia et al., can enhance the performance of models like DeepAR by leveraging the randomness and sensitivity of chaotic mapping, leading to improved forecasting accuracy and efficiency across diverse datasets [23]. Additionally, the hybrid model combining 1D Convolutional Neural Networks (1DCNN) and Gated Recurrent Unit (GRU) memory cells, as proposed by El Zaar et al., demonstrates the potential of intricate architectures in surpassing traditional models, which can be beneficial for enhancing DeepAR's capabilities in handling volatile market data and other real-world scenarios [67]. Kumar and Haider's examination of encoder-decoder components and the integration of statistical models with neural networks further highlights the advancements in hybrid deep learning models, which can be applied to improve DeepAR's performance in both one-step-ahead and multi-horizon forecasting [28]. Madhusudhanan et al. emphasize the importance of tuning hyperparameters such as context length and validation strategy, which are critical for optimizing the performance of models like DeepAR, as demonstrated through extensive experiments and the creation of the TSBench metadataset [32]. Finally, Miller et al. discuss the challenges and recent advances in deep learning for time series forecasting, including the use of foundation models and knowledge graphs, which can provide DeepAR with a more robust understanding of patterns and domain-specific knowledge, thereby enhancing its predictive capabilities [35]. By integrating these insights, DeepAR can be significantly improved to provide more accurate and reliable probabilistic forecasts.

2.3.4. Attention Mechanisms and Transformers

The introduction of self-attention mechanisms and Transformer models has revolutionized time series forecasting. These models can focus on relevant parts of the input sequence, improving performance in capturing long-term dependencies and complex patterns. Temporal Fusion Transformers (TFT) have emerged as a powerful tool in time series forecasting, addressing various challenges across different domains [30]. TFTs are particularly effective in capturing long-term dependencies and handling complex temporal properties, making them suitable for tasks with limited historical data, such as cold start forecasting for new products. In this scenario, TFTs leverage the assumption that products with similar characteristics exhibit similar time series trajectories, outperforming traditional models like DeepAR and LSTM in terms of forecasting accuracy, though they may be more sensitive to anomalies [65]. For anomaly detection in time series, the Temporal Context Fusion Transformer (TCF-Trans) enhances the robustness of predictions by integrating features from both shallow and deep decoder layers, effectively capturing unusual details while maintaining noise resilience [41]. Additionally, TFTs have been employed in generating synthetic time series data to address data deficiency issues. The Time-Transformer AAE model, which combines the strengths of Temporal Convolutional Networks and Transformers, excels in learning both local and global features, proving advantageous for data augmentation in small and imbalanced datasets [31]. In environmental monitoring, TFTs have been used to predict sewer manhole overflows during heavy rainfall, showing high predictive performance compared to LSTM and DA-RNN models. The study found that using a single measuring point at the sewer network outlet provided better forecasts due to reduced model complexity and high correlation among measurements [9]. Overall, TFTs offer a versatile and robust approach to time series forecasting, anomaly detection, and data generation across various applications

2.3.5. Probabilistic Time Series Forecasting

Probabilistic time series forecasting models provide a distribution of possible future values rather than a single point estimate, offering a more comprehensive view of potential outcomes. Existing time series forecasting models, such as DeepAR and Temporal Fusion Transformer (TFT), have demonstrated the ability to capture complex temporal relationships in financial markets. However, these models face limitations in probabilistic forecasting, particularly in scenarios with high volatility or limited data availability. DeepAR's reliance on recurrent neural networks (RNNs) can result in challenges with long-term dependencies, while TFT's focus on self-attention mechanisms sometimes struggles with uncertainty estimation.

In recent years, large language models (LLMs) have gained prominence in natural language processing (NLP) tasks due to their ability to generate high-quality text and perform complex linguistic tasks, even in low-resource language settings. According to the recent study, LLMs like ChatGPT have demonstrated competitive performance in generating annotations and solving NLP tasks, often approaching or matching human-level quality [36]. This success in handling linguistic uncertainty and context-rich environments has inspired the application of LLMs beyond NLP, such as in time series forecasting. Lag-Llama, leveraging the strengths of large language models in probabilistic forecasting, offers a new approach that aims to address these limitations, providing more accurate and reliable stock price predictions. The Lag-Llama model, a recent addition to this domain, leverages advanced deep learning techniques to enhance forecasting accuracy and robustness [42].

3. Materials and Methods

3.1. Data

The dataset used in this study consists of the top 18 stocks listed on the IDX Market Cap for Q2 2024 (https://idx.co.id). The data includes daily stock prices, which are used to train and evaluate the forecasting models. The dataset is divided into training and testing sets, with the 70% dataset as training set covering a period of 3 February 2020 - 16 February 2023 and the rest of 30% dataset as testing set covering a period of 17 February 2023 - 07 June 2024.

3.2. Experimental Setup

We designed our experiments using the GluonTS framework, a widely-used Python package for probabilistic time series modeling [2]. Our primary focus is on the validation of the Lag-Llama model. We compare its performance against DeepAR and Temporal Fusion Transformer, two well-established models in time series forecasting, under both zero-shot and fine-tuning scenarios. To rigorously test Lag-Llama, we vary context lengths and learning rates, aiming to optimize its performance for univariate and multi-time series forecasting. This experimental setup enables us to thoroughly assess Lag-Llama's ability to capture complex temporal dependencies and its robustness in real-world stock market data.

- **Zero-Shot**: In the zero-shot approach, the models are trained without any prior fine-tuning. Various context lengths are tested to evaluate their impact on model performance. The context lengths used are 32, 64, 128, 256, 512, and 1024.
- **Fine-Tuning**: In the fine-tuning approach, the models are first trained with the initial setup and then fine-tuned using different learning rates. The context lengths used are the same as in the zero-shot approach (32, 64, 128, 256, 512, and 1024), and the learning rates tested are 1e-2, 1e-3, 1e-4, 5e-3, 5e-4, and 5e-5.

3.3. Models

Three models are evaluated in this study:

- **Lag-Llama**: A probabilistic time series forecasting model that leverages advanced deep learning techniques. Lag-Llama is designed to capture long-term dependencies and provide robust forecasts.

- **DeepAR (RNN)**: An autoregressive recurrent neural network model that generates probabilistic forecasts. DeepAR is widely used for time series forecasting due to its ability to capture temporal dependencies.
- **Temporal Fusion Transformer (LSTM with Self-Attention)**: A model that combines LSTM networks with self-attention mechanisms. The Temporal Fusion Transformer is designed to enhance forecasting performance by focusing on relevant parts of the input sequence.

3.4. Evaluation Metrics

The models in this study are evaluated using the Continuous Ranked Probability Score (CRPS), a widely used metric in probabilistic forecasting [14]. CRPS measures how well the predicted probability distribution of future outcomes matches the actual observed values. CRPS, a common metric in the probabilistic forecasting literature [13, 42, 49] is similar to MAE. Unlike traditional metrics like Mean Absolute Error (MAE), which focus on point estimates, CRPS accounts for the full range of possible outcomes, making it particularly useful for probabilistic forecasts.

In essence, CRPS compares the predicted cumulative distribution function (CDF)—which describes the probability of different outcomes—with the actual CDF derived from the observed data. The score provides a single value that reflects how "close" the entire predicted distribution is to the real outcome, not just a single point.

A lower CRPS value indicates a better alignment between the predicted probabilities and the actual outcomes, meaning the model provides more accurate and reliable probabilistic forecasts. This makes CRPS an ideal metric for evaluating models in environments with high uncertainty, such as stock market prediction, where forecasting the range of possible price movements is more informative than predicting a single price.

3.5. Experimental Procedure

- **Data Preprocessing**: The data is preprocessed to handle missing values and normalize the stock prices. The preprocessing steps ensure that the models receive clean and standardized input data.
- **Dataset Selection**: To comprehensively evaluate the performance of the Lag-Llama model, we designed a series of six experiments that examine its predictive capabilities across various stock datasets. The experiments are structured as follows:

  ∗ **Univariate Time Series**:

    ∗ Experiments 1, 2, and 3 evaluate the model's predictions for individual stocks with best predictive performance: BBCA, AMRT, and BMRI, respectively. This allows for a detailed analysis of the model's accuracy at the individual stock level, highlighting its effectiveness in forecasting specific stock movements.

  ∗ **Multi Time Series**:

    ∗ Experiment 4 combines the three stocks with the best predictive performance, offering an understanding of how well Lag-Llama can manage a concentrated portfolio of top-performing stocks.
    ∗ Experiment 5 focuses on the top nine highest-valued stocks, aiming to assess the model's performance on stocks with significant market capitalization.

    * Experiment 6 examines stocks ranked 10th to 18th by market value, providing insight into how the model performs on a different market segment with potentially different characteristics.

- **Model Training**: The models are trained using the training set, with hyperparameters tuned through grid search. The training process involves optimizing the model parameters to minimize the chosen evaluation metrics.
- **Model Evaluation**: The trained models are evaluated on the testing set using the defined metrics. The evaluation process compares the performance of Lag-Llama with the baseline models (DeepAR and Temporal Fusion Transformer) under both zero-shot and fine-tuning scenarios.

4. Results

The results from our experiments reveal that Lag-Llama, particularly in its fine-tuned form, outperforms both DeepAR and Temporal Fusion Transformer in specific scenarios, particularly for predicting key stocks such as BBCA, BMRI, and AMRT. Lag-Llama's probabilistic framework allows for more accurate modeling of uncertainty and temporal dependencies, making it especially effective in the volatile context of the Indonesian stock market. Notably, Lag-Llama demonstrated superior performance in multi-time series forecasting, suggesting its potential as a tool for managing diverse portfolios.

4.1. Baseline Model Performance

When evaluating the performance of a new time series forecasting model like Lag-Llama, it is common to use established models as baselines for comparison. In this case, DeepAR and Temporal Fusion Transformer (TFT) are chosen as baseline models due to their strong track record in time series forecasting as listed in Table 1.

4.2. Zero-Shot Performance

Table 2 presents the CRPS for the zero-shot approach with different context lengths and rope scaled for Lag-Llama.

4.3. Fine-Tuning Performance for Lag-Llama

Table 3 presents the CRPS for the fine-tuning approach with different context lengths and learning rates for Lag-Llama.

Table 1

Baseline Model (DeepAR and Temporal Fusion Transformer) Performance Metrics. The best results are in **bold**.

Model	Univariate Time Series			Multi Time Series		
	BBCA	AMRT	BMRI	Best 3	Top 1-9	Top 10-18
DeepAR	**0.0170**	0.0827	**0.1121**	0.0270	0.0677	0.2282
TFT	0.0373	**0.0403**	0.2299	**0.0179**	**0.0517**	**0.1429**

Table 2
Zero-Shot Performance Metrics for Lag-Llama with Different Context Lengths and Rope Scaled. The best results are in **bold**.

Rope Scaled	Context Length	Univariate Time Series			Multi Time Series		
		BBCA	AMRT	BMRI	Best 3	Top 1-9	Top 10-18
False	32	**0.0313**	**0.0383**	**0.0373**	**0.0364**	**0.0890**	0.2660
False	64	0.0383	0.0415	0.0445	0.0395	0.0980	0.2518
True	64	0.0365	0.0422	0.0544	0.0417	0.0927	0.2611
True	128	0.0374	0.0397	0.0377	0.0413	**0.0890**	0.2113
True	256	0.0348	0.0394	0.0466	0.0379	0.0973	0.1912
True	512	0.0459	0.0384	0.1305	0.0681	0.1330	**0.1811**
True	1024	0.0461	0.0435	0.1966	0.1107	0.1515	0.2311

5. Discussion

To better illustrate the performance of the models, we provide several figures depicting results on Table 1, Table 2, and Table 3 into two subsections: Univariate Time Series and Multi Time Series.

5.1. Univariate Time Series

In this section, we evaluate the performance of Lag-Llama on univariate time series forecasting for individual stocks. The focus is on predicting stock prices for three leading companies on the Indonesian Stock Exchange: BBCA, BMRI, and AMRT. Each stock is modeled independently, allowing for an analysis of the model's ability to capture unique trends and behaviors in single-stock data. We examine the accuracy of several approaches, including zero-shot predictions, fine-tuning, Temporal Fusion Transformer (TFT), and DeepAR, with performance metrics such as Continuous Ranked Probability Score (CRPS) used to compare results across different models.

5.1.1. Bank Central Asia (BBCA)

The result comparison for BBCA Stock is depicted in Figure 1. The optimal parameters for each four methods for the BBCA Stock is depicted in Figure 2. The detail of the forecasting results showing the predicted versus actual stock prices for selected stocks are displayed in Figure 3 to Figure 6.

As shown in Figure 2, Lag-Llama Zero-Shot can only outperformed TFT, while Lag-Llama Fine-Tuning can outperformed both TFT and DeepAR. The fine-tuned model (CRPS: 0.0140) delivers the most accurate predictions, with a narrow uncertainty band and close alignment to actual stock prices. DeepAR (CRPS: 0.0170) follows closely in accuracy but slightly overestimates the recovery phase. The zero-shot model (CRPS: 0.0313) captures the trend reasonably well but with higher uncertainty, while TFT (CRPS: 0.0373) overestimates the stock price and shows the least precision among the models.

Table 3
Fine-Tuning Performance Metrics for Lag-Llama with Different Context Lengths and Learning Rates. The best results are in **bold**.

Learning Rate	Context Length	Univariate Time Series			Multi Time Series		
		BBCA	AMRT	BMRI	Best 3	Top 1-9	Top 10-18
1e-2	32	0.0182	0.0498	0.0930	**0.0195**	0.0674	0.3529
	64	**0.0140**	**0.0295**	0.0879	0.2130	0.1220	0.2470
	128	0.0599	0.1072	0.0637	0.0928	0.1088	0.2607
	256	0.0410	0.0491	**0.0431**	0.0445	0.1072	0.2437
	512	0.0231	0.1349	0.0461	0.1919	0.0803	**0.2028**
	1024	0.0294	0.0944	0.0536	0.0367	**0.0628**	0.2843
1e-3	32	0.0860	0.0984	0.0880	**0.0327**	0.2405	0.2494
	64	0.0574	**0.0349**	0.1481	0.1018	0.1050	0.2567
	128	0.0637	0.0538	0.1429	0.1898	0.1050	0.2604
	256	0.0206	0.1336	**0.0467**	0.1119	0.0969	**0.1331**
	512	0.0225	0.1458	0.0898	0.1354	0.1151	0.1541
	1024	**0.0200**	0.1230	0.0696	0.1265	**0.0949**	0.1532
1e-4	32	0.1493	0.0642	0.0676	0.0863	0.0913	0.2342
	64	0.1585	0.0572	0.1416	0.0645	0.1167	0.1847
	128	0.1611	0.0919	0.1285	0.1146	0.0686	**0.1594**
	256	0.1258	0.0523	**0.0371**	0.0649	**0.0594**	0.1663
	512	0.0603	**0.0346**	0.0887	**0.0349**	0.0750	0.2263
	1024	**0.0581**	0.0669	0.0665	0.0362	0.1075	0.2647
5e-3	32	0.0521	0.0782	0.0699	**0.0250**	0.2543	0.2619
	64	0.0333	**0.0222**	0.0664	0.1221	0.4983	0.3374
	128	0.0217	0.0236	0.0582	0.0295	0.4926	0.2486
	256	0.0231	0.0615	0.0469	0.1313	0.1087	0.2693
	512	**0.0155**	0.1135	**0.0439**	0.1361	**0.0977**	0.1798
	1024	0.0471	0.1639	0.0585	0.0273	0.1544	**0.1753**
5e-4	32	0.1494	0.0670	0.1264	**0.0318**	0.2317	0.2629
	64	0.1375	0.0624	0.0765	0.1218	**0.0725**	0.2700
	128	0.1481	**0.0515**	0.1204	0.0464	0.1360	0.1965
	256	0.0263	0.0629	0.0757	0.1117	0.0944	**0.1849**
	512	**0.0233**	0.1371	0.0585	0.1098	0.1198	0.1974
	1024	0.0388	0.0516	**0.0578**	0.0378	0.0838	0.2453
5e-5	32	0.2076	0.0287	0.0673	0.0258	0.0771	0.1995
	64	0.1650	0.0513	0.0720	0.0576	0.0723	0.1873
	128	0.1923	0.0691	**0.0506**	0.0436	0.1192	**0.1690**
	256	0.1630	0.0454	0.0705	0.0367	0.0724	0.1866
	512	**0.0567**	**0.0196**	0.0736	**0.0240**	0.0842	0.2048
	1024	0.0700	0.0552	0.0711	0.0512	**0.0709**	0.2097

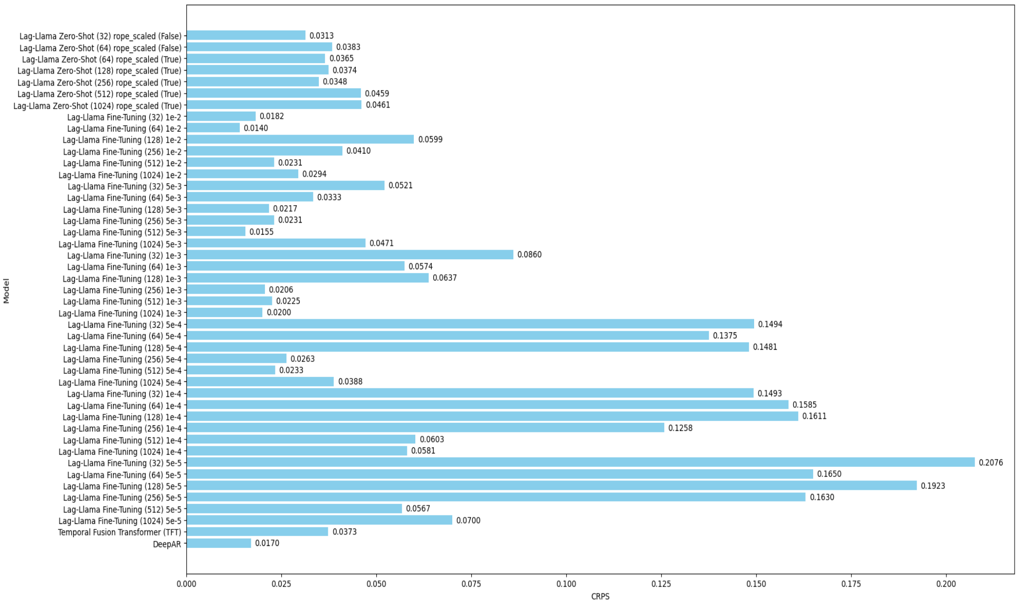

Fig. 1. Result Comparison for Univariate Time Series: BBCA Stock

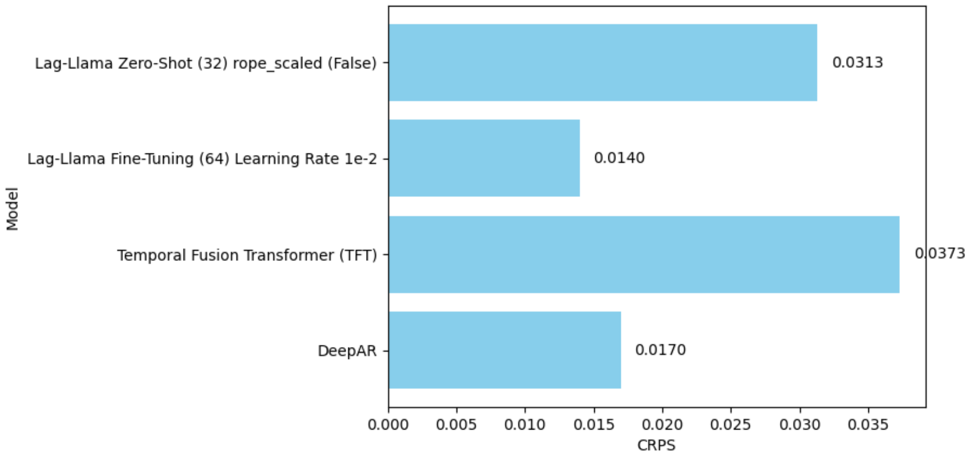

Fig. 2. Optimal Parameters for Univariate Time Series: BBCA Stock

We can further analyse the forecasting results as follow:

- Figure 3 (Lag-Llama Zero-Shot with Rope Scaling False, CRPS: 0.0313): The zero-shot model delivers a reasonable forecast with a moderate uncertainty band. The predictions generally follow the actual stock price trend but display higher uncertainty, especially towards the end of the forecast horizon. The CRPS value of 0.0313 indicates a fair performance, although the model is less accurate compared to the fine-tuned variant. The predictions slightly lag in terms of precision, but the model still captures the overall price movement reasonably well.
- Figure 4 (Lag-Llama Fine-Tuning with Learning Rate 1e-2, CRPS: 0.0140): The fine-tuned model provides the most accurate results, as evidenced by the lowest CRPS value of 0.0140. The uncer-

tainty band is much narrower, and the predictions align closely with the actual stock price. This model captures both the downward trend and minor fluctuations with greater precision, reflecting the benefits of fine-tuning in improving forecast accuracy. The model's predictions are more reliable, offering the best performance among the tested models.

- Figure 5 (TFT Model, CRPS: 0.0373): The Temporal Fusion Transformer (TFT) model performs reasonably but with notable limitations. The forecasted values slightly overestimate the stock price, and the CRPS value of 0.0373 is higher than the other models, indicating that the model is less accurate. Although the uncertainty band is narrow, the predictions do not capture the exact downward trend, suggesting that TFT struggles to precisely forecast the stock movements for BBCA in this instance.
- Figure 6 (DeepAR Model, CRPS: 0.0170): The DeepAR model performs well, with a CRPS value of 0.0170, indicating higher accuracy than the zero-shot and TFT models but slightly less precise than the fine-tuned approach. The uncertainty band is moderately narrow, and the predictions generally align with the actual stock price. The model captures the overall trend, although it slightly overestimates the recovery. Nonetheless, DeepAR provides strong performance in predicting the BBCA stock price.

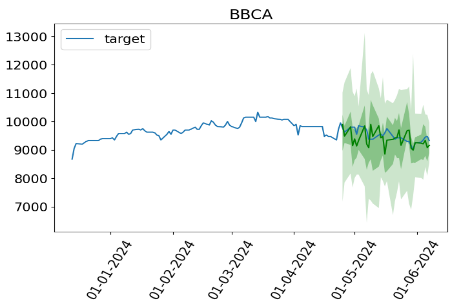

Fig. 3. Optimal Parameters for Univariate Time Series: BBCA Stock with Zero-Shot (32) and Rope Scaled (False)

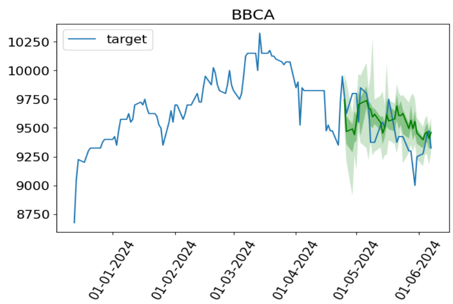

Fig. 4. Optimal Parameters for Univariate Time Series: BBCA Stock with Fine Tuning (64) Learning Rate 1e-2

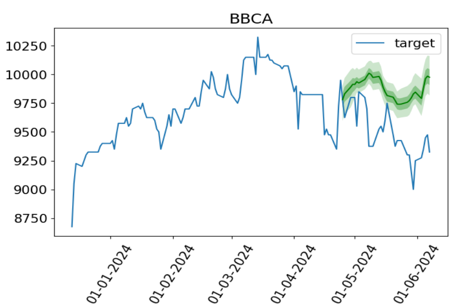

Fig. 5. Optimal Parameters for Univariate Time Series: BBCA Stock with TFT

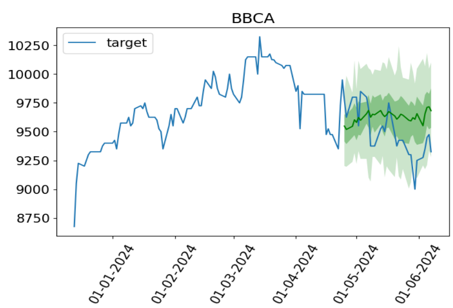

Fig. 6. Optimal Parameters for Univariate Time Series: BBCA Stock with DeepAR

5.1.2. Bank Mandiri (BMRI)

The result comparison for BMRI Stock is depicted in Figure 7. The optimal parameters for each four methods for the BMRI Stock is depicted in Figure 8. The detail of the forecasting results showing the predicted versus actual stock prices for selected stocks are displayed in Figure 9 to Figure 12.

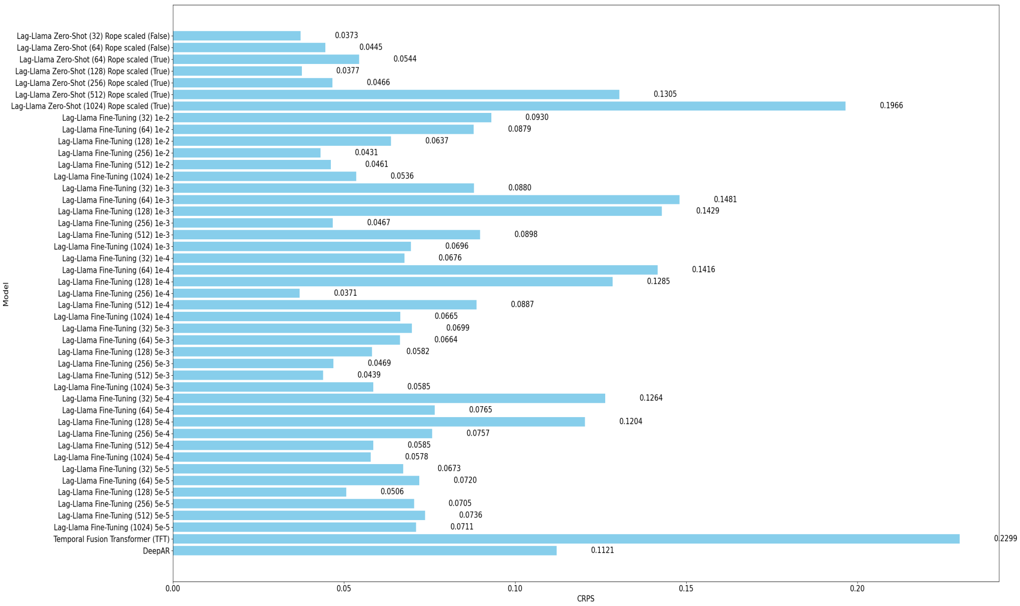

Fig. 7. Result Comparison for Univariate Time Series: BMRI Stock

As shown in Figure 8, both Lag-Llama Zero-Shot and Lag-Llama Fine-Tuning outperformed the baseline models; TFT and DeepAR. The fine-tuned model (CRPS: 0.0371) produces the most accurate predictions, with greater caution reflected in its wider uncertainty band. The zero-shot model (CRPS: 0.0373) also performs well but shows higher confidence with a narrower band, slightly underestimating variability. TFT (CRPS: 0.2299) significantly overestimates prices and shows the poorest performance, while DeepAR (CRPS: 0.1121) mispredicts the trend but with more confidence than TFT.

We can further analyse the forecasting results as follow:

- Figure 9 (Lag-Llama Zero-Shot with Rope Scaling False, CRPS: 0.0373): The zero-shot model delivers reasonably accurate point predictions, capturing the general downward trend in BMRI's stock price. The uncertainty band is relatively narrow, indicating high confidence in the forecast. However, this narrow band may suggest that the model is somewhat overconfident, as it slightly underestimates the recovery toward the end of the period. The low CRPS value of 0.0373 demonstrates that the zero-shot model performs well in terms of probabilistic forecasting, though its confidence may not fully reflect the actual variability in the stock's price.
- Figure 10 (Lag-Llama Fine-Tuning with Learning Rate 1e-4, CRPS: 0.0371): The fine-tuned model produces the most accurate point predictions, as reflected by the lowest CRPS value of 0.0371. However, unlike the zero-shot model, the fine-tuning approach results in a wider uncertainty band. This wider band indicates that the model is more cautious, allowing for greater variability in the possible outcomes. While the model is more precise in its point forecasts, particularly during the

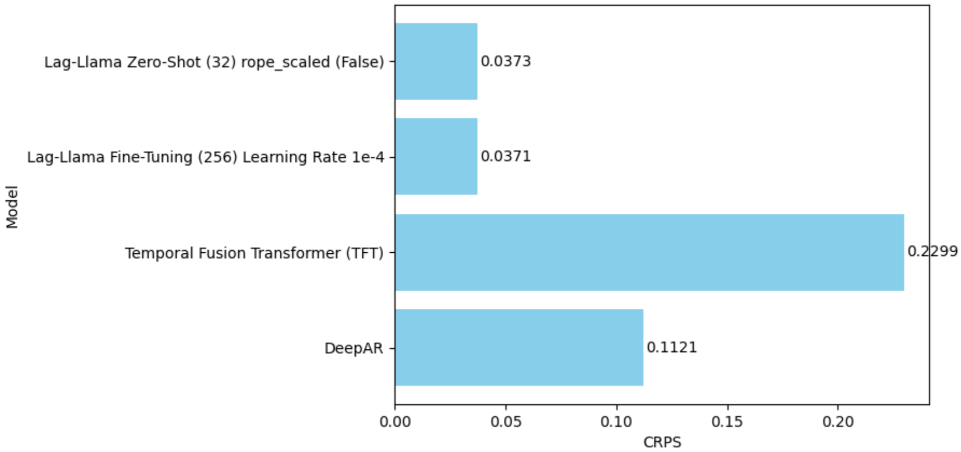

Fig. 8. Optimal Parameters for Univariate Time Series: BMRI Stock

recovery phase, it expresses greater uncertainty, reflecting the complexities of the stock's price movements more effectively than the zero-shot model.

- Figure 11 (TFT Model, CRPS: 0.2299): The Temporal Fusion Transformer (TFT) model significantly overestimates BMRI's stock price, predicting an upward movement that does not occur. The wide uncertainty band suggests that the model is uncertain about its predictions, but the consistently high forecasted values indicate that it struggles to capture the downward trend in the actual stock price. With the highest CRPS value of 0.2299, the TFT model shows poor performance in this scenario, indicating that it is less reliable for predicting BMRI's stock price movements.
- Figure 12 (DeepAR Model, CRPS: 0.1121): The DeepAR model mispredicts the trend direction, forecasting an upward movement while the actual stock price declines. The uncertainty band is narrower than that of the TFT model, suggesting more confidence in its (incorrect) predictions, but the trend forecast remains inaccurate. With a CRPS value of 0.1121, DeepAR performs better than TFT but still struggles to capture the true direction of the stock's price movements, particularly in handling the market reversal.

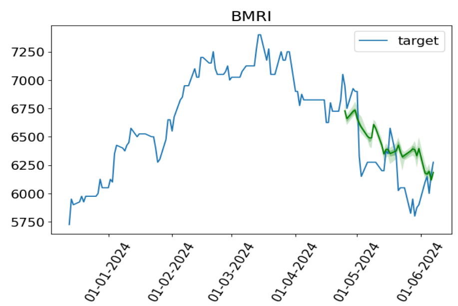

Fig. 9. Optimal Parameters for BMRI Stock with Zero-Shot (32) and Rope Scaled (False)

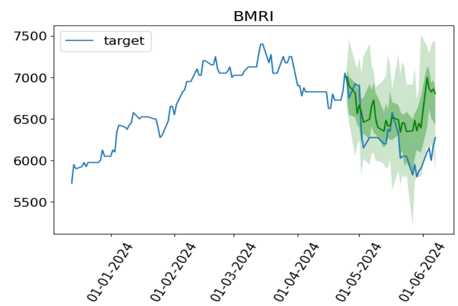

Fig. 10. Optimal Parameters for BMRI Stock with Fine Tuning (256) Learning Rate 1e-4

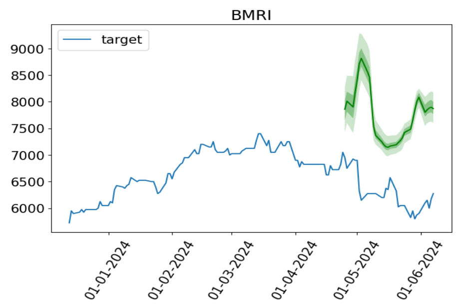

Fig. 11. Optimal Parameters for BMRI Stock with TFT

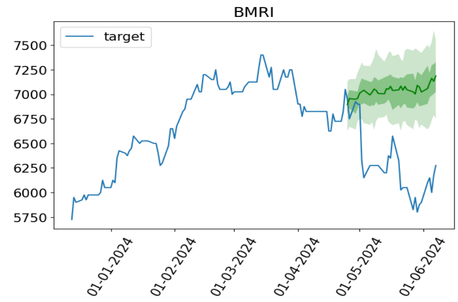

Fig. 12. Optimal Parameters for BMRI Stock with DeepAR

5.1.3. *Alfamart (AMRT)*

The result comparison for AMRT Stock is depicted in Figure 13. The optimal parameters for each four methods for the AMRT Stock is depicted in Figure 14. The detail of the forecasting results showing the predicted versus actual stock prices for selected stocks are displayed in Figure 15 to Figure 18.

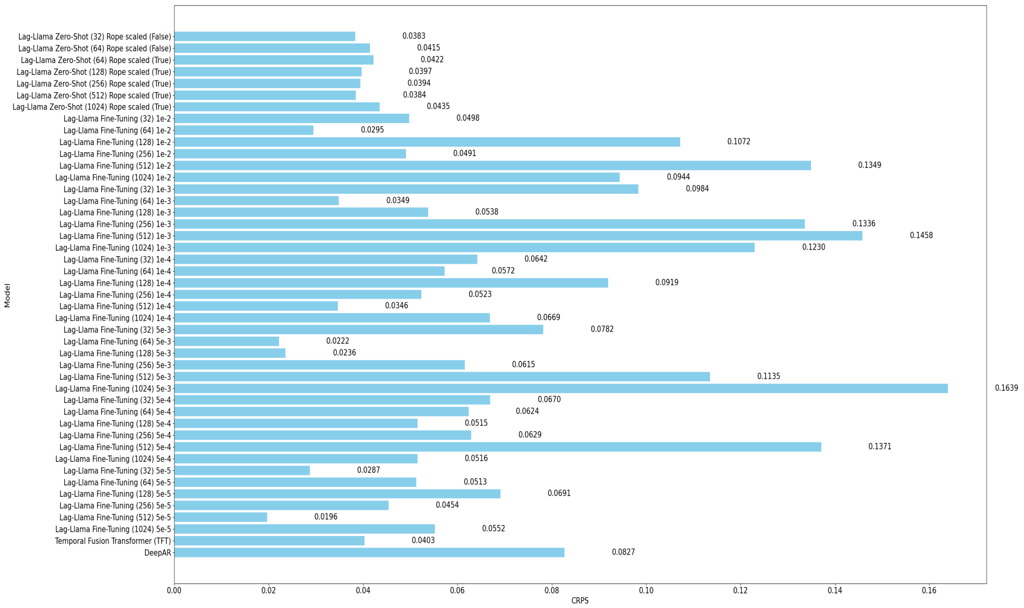

Fig. 13. Result Comparison for Univariate Time Series: AMRT Stock

As shown in Figure 14, both Lag-Llama Zero-Shot and Lag-Llama Fine-Tuning also outperformed the baseline models; TFT and DeepAR. The fine-tuned model (CRPS: 0.0196) provides the most accurate predictions for AMRT stock, with close alignment to the actual prices and a narrow uncertainty band. The zero-shot model (CRPS: 0.0383) performs reasonably but with higher uncertainty. TFT (CRPS: 0.0403) overestimates the stock price and shows higher uncertainty, while DeepAR (CRPS: 0.0827) performs the worst, with large deviations and overestimation throughout the forecast period.

We can further analyse the forecasting results as follow:

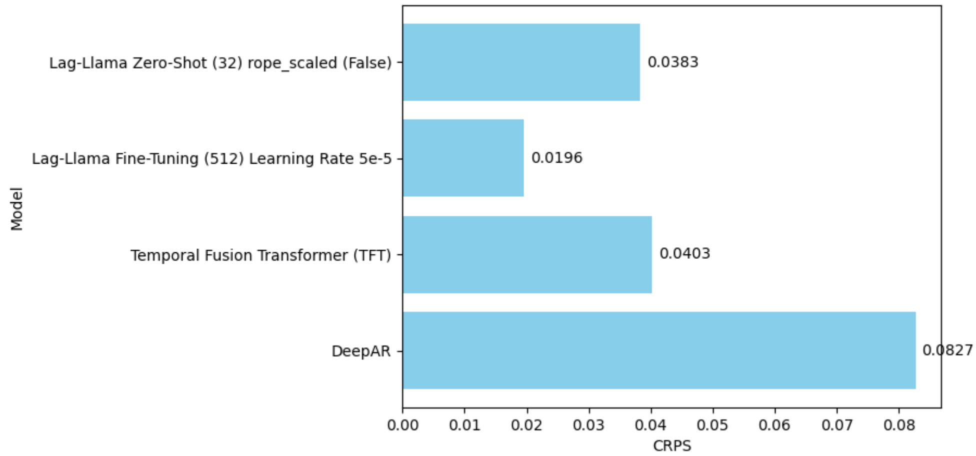

Fig. 14. Optimal Parameters for Univariate Time Series: AMRT Stock

- Figure 15 (Lag-Llama Zero-Shot with Rope Scaling False, CRPS: 0.0383): The zero-shot model demonstrates a reasonable forecast, though with a noticeable level of uncertainty as indicated by the wide uncertainty band. The model manages to capture the general price movement of AMRT but tends to show more uncertainty, especially towards the end of the prediction horizon. The CRPS value of 0.0383 shows that the model performs adequately, though it is outperformed by fine-tuning and DeepAR in this scenario.
- Figure 16 (Lag-Llama Fine-Tuning with Learning Rate 5e-5, CRPS: 0.0196): The fine-tuned model provides the most accurate predictions with a significantly narrower uncertainty band. The predictions closely follow the actual stock prices, with minimal deviation. The CRPS value of 0.0196, the lowest among the models, indicates that fine-tuning significantly improves both the accuracy and confidence of the predictions, offering the most reliable forecast for AMRT stock movements.
- Figure 17 (TFT Model, CRPS: 0.0403): The Temporal Fusion Transformer (TFT) model exhibits slightly higher uncertainty compared to the fine-tuned model, as reflected in its wider uncertainty band. The model overestimates the stock price for a portion of the forecast period, and with a CRPS value of 0.0403, it performs worse than both the zero-shot and fine-tuned models. TFT struggles to capture the price fluctuations as effectively, leading to less precise forecasts.
- Figure 18 (DeepAR Model, CRPS: 0.0827): The DeepAR model has the poorest performance, with a CRPS value of 0.0827, significantly higher than the other models. The model's uncertainty band is quite wide, reflecting low confidence in the predictions. Additionally, DeepAR overestimates the stock price trajectory throughout the forecast period, leading to large deviations from the actual stock prices. The model's predictions are less reliable for AMRT stock compared to the other approaches.

5.2. *Multi Time Series*

This section explores the performance of Lag-Llama in a multi time series forecasting scenario, where the model simultaneously predicts the stock prices of BBCA, BMRI, and AMRT. By combining the datasets of multiple stocks, we aim to assess the model's ability to generalize across different financial instruments and improve prediction accuracy by leveraging shared temporal patterns. We also examines

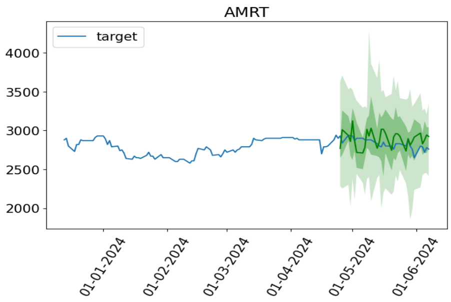

Fig. 15. Optimal Parameters for AMRT Stock with Zero-Shot (32) and Rope Scaled (False)

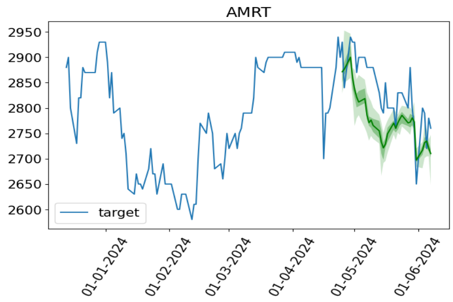

Fig. 16. Optimal Parameters for AMRT Stock with Fine Tuning (512) Learning Rate 5e-5

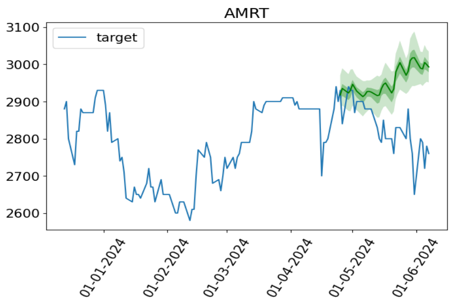

Fig. 17. Optimal Parameters for AMRT Stock with TFT

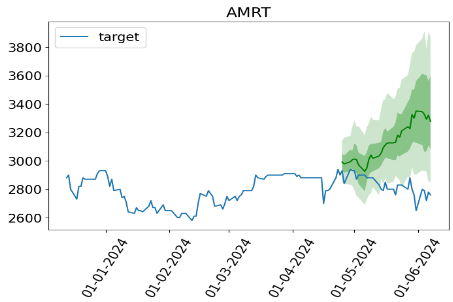

Fig. 18. Optimal Parameters for AMRT Stock with DeepAR

the top nine highest-valued stocks, aiming to assess the model's performance on stocks with significant market capitalization, and examines stocks ranked 10th to 18th by market value, providing insight into how the model performs on a different market segment with potentially different characteristics. Similar to the univariate analysis, we evaluate the zero-shot, fine-tuning, TFT, and DeepAR models, comparing their performance based on CRPS to determine which approach is most effective in a multi-stock forecasting environment.

5.2.1. Combined Three Stocks with the Best Predictive Performance

The result comparison for Multi Time Series: BBCA, AMRT, and BMRI Stocks is depicted in Figure 19. The optimal parameters for each four methods for the Multi Time Series: BBCA, AMRT, and BMRI Stocks is depicted in Figure 20. The detail of the forecasting results showing the predicted versus actual stock prices for selected stocks are displayed in Figure 15 to Figure 18.

For multi time series forecasting of BBCA, BMRI, and AMRT stocks, the TFT model (CRPS: 0.0179) unexpectedly provides the most accurate probabilistic forecasts with the narrowest uncertainty bands, closely followed by the fine-tuning approach (CRPS: 0.0195), which also demonstrates excellent alignment with the actual data. DeepAR performs better than zero-shot (CRPS: 0.0364) but not as well as TFT or fine-tuning, capturing trends with moderate accuracy (CRPS: 0.0270).

We can further analyse the forecasting results as follow:

- Figure 21 (Lag-Llama Zero-Shot, CRPS: 0.0364): The zero-shot model performs adequately, with predictions that generally follow the actual trends of the combined stocks but with visible deviations

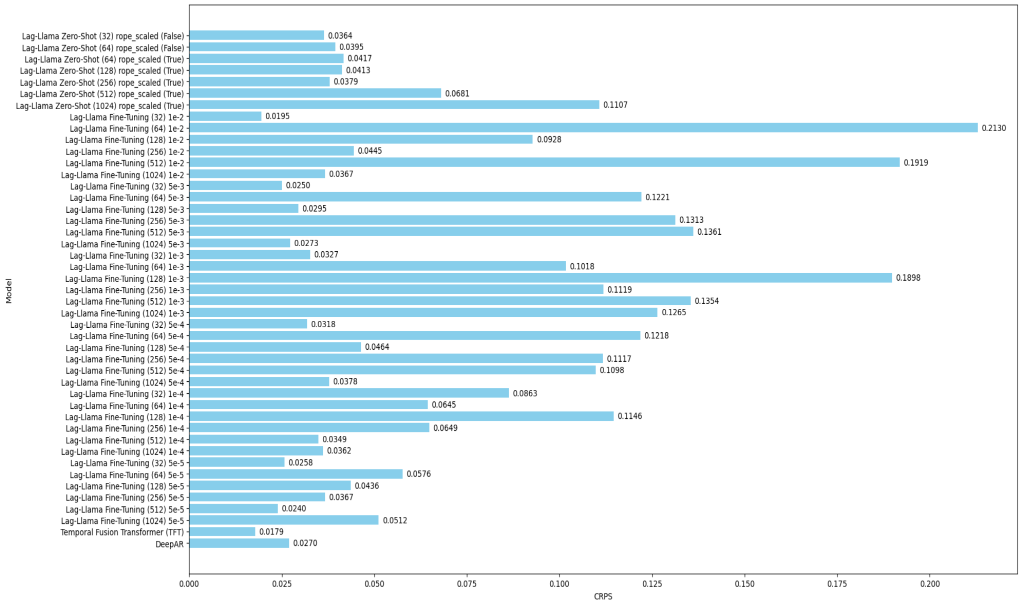

Fig. 19. Result for Multi Time Series: BBCA, AMRT, and BMRI Stocks

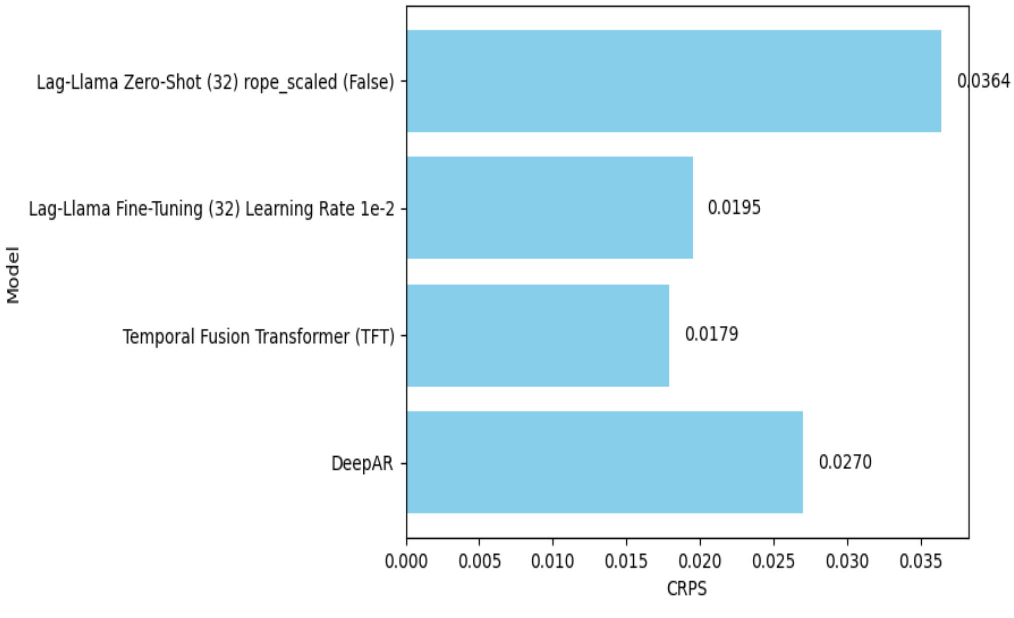

Fig. 20. Optimal Parameters for Multi Time Series: BBCA, AMRT, and BMRI Stocks

and a somewhat broad uncertainty band. The model captures major trends but with less precision, reflected in a CRPS of 0.0364.

- Figure 22 (Lag-Llama Fine-Tuning, CRPS: 0.0195): The fine-tuned model exhibits the highest accuracy among the approaches, with a CRPS of 0.0195. It produces a tighter uncertainty band and closely matches the actual stock price movements across all three stocks, indicating a significant improvement in prediction accuracy through fine-tuning.

- Figure 23 (TFT, CRPS: 0.0179): Surprisingly, the TFT model performs best in terms of CRPS, indicating the highest precision in probabilistic forecasting among the evaluated models. It manages to closely track the actual movements with a narrow uncertainty band, showing robust performance in multi time series forecasting.
- Figure 24 (DeepAR, CRPS: 0.0270): DeepAR shows good performance with a CRPS of 0.0270, offering more accurate predictions than the zero-shot model but less so than TFT and fine-tuning. Its uncertainty band is moderate, and it captures the overall trends effectively, though with some overestimations in peak values.

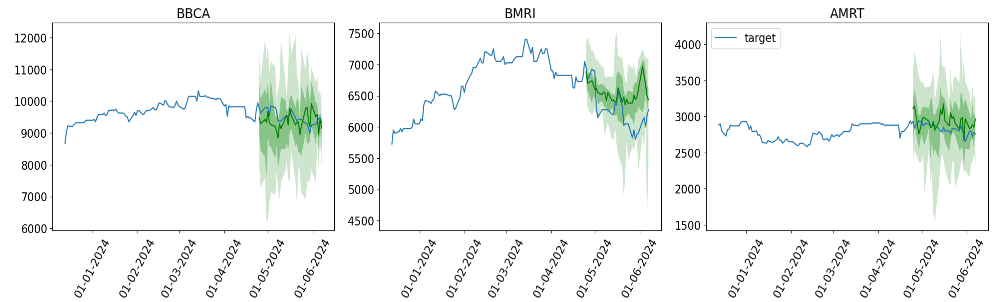

Fig. 21. Optimal Parameters for Multi Time Series: BBCA, AMRT, and BMRI Stocks with Zero-Shot (32) and Rope Scaled (False)

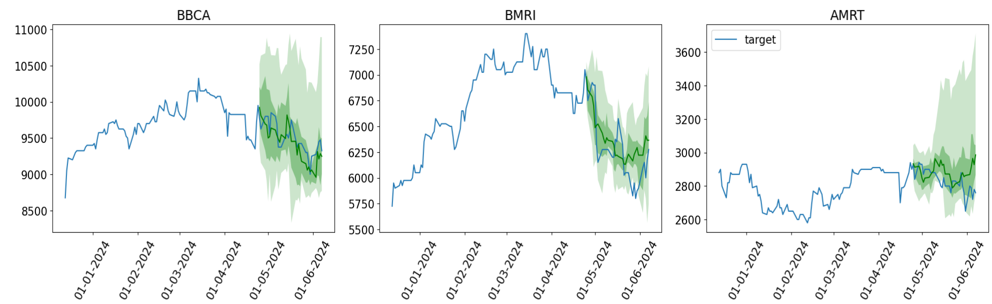

Fig. 22. Optimal Parameters for Multi Time Series: BBCA, AMRT, and BMRI Stocks with Fine Tuning (32) Learning Rate 1e-2

5.2.2. Top Highest-Valued Stocks

For the top 1-9 stocks by market capitalization shown in Figure 25, the best-performing model is TFT with a CRPS of 0.0517, followed by the fine-tuning model (CRPS: 0.0594), zero-shot (CRPS: 0.0890), and DeepAR (CRPS: 0.0677). While these results are relatively strong, none of the models outperform the fine-tuned and TFT models in the combined BBCA, BMRI, and AMRT dataset. The CRPS values are slightly higher across the board, indicating that the models face more difficulty in generalizing and predicting across the larger, high-market-cap stocks.

For the top 10-18 stocks by market capitalization shown in Figure 26, the models perform worse overall compared to the top 1-9 stocks. The best result comes from the fine-tuned model with a CRPS of 0.1331, followed by zero-shot (CRPS: 0.1811), TFT (CRPS: 0.1429), and DeepAR (CRPS: 0.2282).

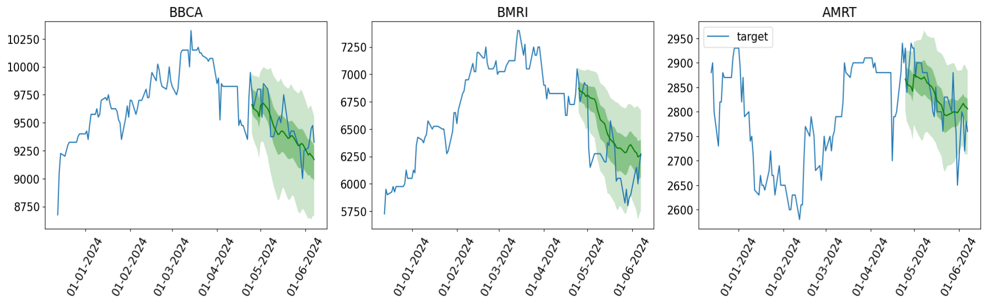

Fig. 23. Optimal Parameters for Multi Time Series: BBCA, AMRT, and BMRI Stocks with TFT

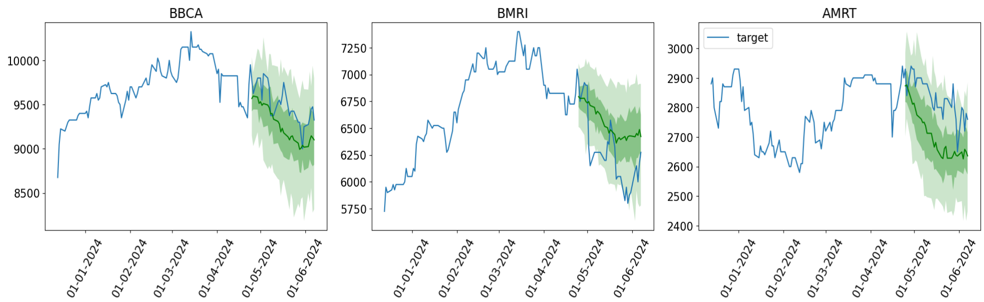

Fig. 24. Optimal Parameters for Multi Time Series: BBCA, AMRT, and BMRI Stocks with DeepAR

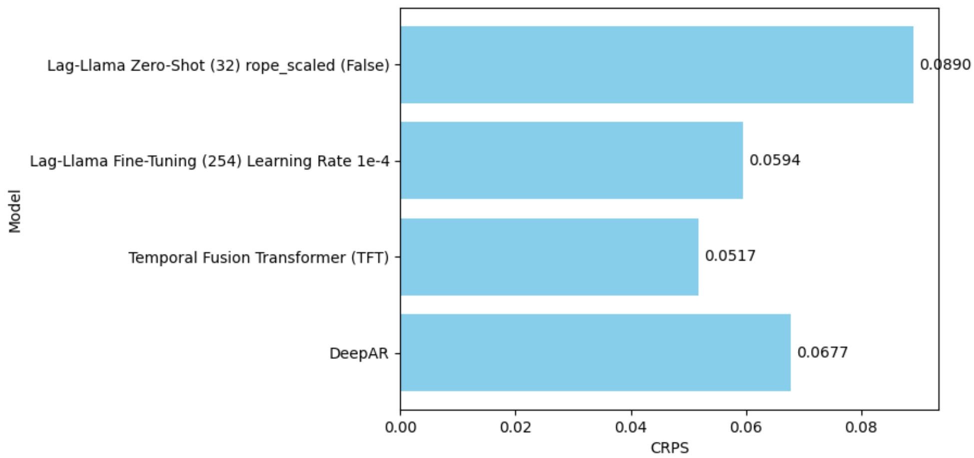

Fig. 25. Optimal Parameters for Multi Time Series: Top 1-9 Stocks by Market Cap

The higher CRPS values here indicate that predicting stock prices in this mid-tier market cap group is more challenging for all models.

When comparing these results with the combined BBCA, BMRI, and AMRT stocks, the models perform significantly better on the latter dataset. The TFT model achieved a CRPS of 0.0179 for BBCA, BMRI, and AMRT combined, which is much lower than the best CRPS obtained for the top 1-9 stocks (0.0517) and top 10-18 stocks (0.1331). Similarly, the fine-tuning model for the combined dataset had a

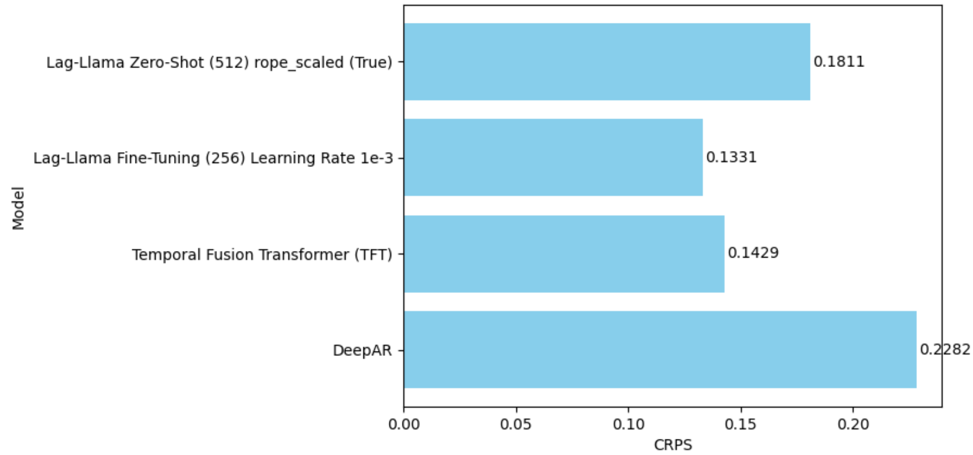

Fig. 26. Optimal Parameters for Multi Time Series: Top 10-18 Stocks by Market Cap

CRPS of 0.0195, much better than its performance for the top 1-9 stocks (0.0594) and top 10-18 stocks (0.1331).

These results indicate that the models perform better when forecasting the combined dataset of BBCA, BMRI, and AMRT compared to forecasting larger groups of top-market-cap stocks. This suggests that the selected combination of BBCA, BMRI, and AMRT likely offers more predictable, less volatile trends, or that the model benefits from focusing on a smaller set of highly correlated stocks rather than trying to generalize across a broader, more diverse stock group.

The combined BBCA, BMRI, and AMRT dataset produces significantly better forecasting results across all models, with much lower CRPS values than those observed in the top 1-9 and top 10-18 stocks by market cap. The TFT and fine-tuned models perform particularly well on the combined dataset, indicating a better ability to capture shared temporal patterns within this smaller group of stocks, whereas the broader market cap stocks introduce more variability, making them harder to predict accurately.

5.3. Summary of Findings

This study aimed to evaluate the performance of the Lag-Llama probabilistic time series forecasting model in predicting stock prices of Indonesian stock market, specifically the top 18 stocks listed on the IDX Market Cap for Q2 2024. We compared Lag-Llama's performance with two baseline models, DeepAR (RNN) and Temporal Fusion Transformer (LSTM with self-attention), under both zero-shot and fine-tuning scenarios with univariate and multi time series forecasting approaches. The experiments were conducted on three key stocks—BBCA, BMRI, and AMRT—as well as broader groups of stocks based on market capitalization (Top 1-9 and Top 10-18 by market cap).

Key findings include:

- Univariate Time Series Forecasting:
  - ∗ The fine-tuned model consistently provided the most accurate predictions for individual stocks, particularly for BBCA, BMRI, and AMRT, as shown by the low CRPS scores.
  - ∗ TFT and DeepAR also performed well in certain cases but were generally outperformed by the fine-tuning approach, with TFT overestimating in some scenarios and DeepAR occasionally mis-predicting trend direction.

- Multi Time Series Forecasting (BBCA, BMRI, and AMRT):
  - ∗ The multi time series forecasting results demonstrated that combining stocks significantly improved model performance. The TFT and fine-tuned models achieved the lowest CRPS values, indicating that the models were able to capture shared patterns across the three stocks, resulting in more accurate forecasts.
  - ∗ The combined BBCA, BMRI, and AMRT dataset outperformed both the top 1-9 and top 10-18 stock groups by market cap, showing that focusing on a smaller, more correlated set of stocks led to better predictive accuracy.
- Market Cap Group Predictions:
  - ∗ When forecasting across the broader top 1-9 and top 10-18 stock groups, the models performed worse, with higher CRPS values compared to the combined BBCA, BMRI, and AMRT dataset. The top 10-18 stocks by market cap were particularly challenging to predict, as reflected by the higher CRPS values.
  - ∗ TFT remained the best-performing model for the top 1-9 stocks, but none of the models reached the predictive accuracy observed in the combined BBCA, BMRI, and AMRT dataset.

5.4. *Implications*

The findings have significant implications for investors and policymakers:

- **For Investors**:
  - ∗ Improved Forecasting Accuracy for Targeted Investments: The results show that the Lag-Llama model, particularly in its fine-tuned and multi time series configurations, can provide highly accurate stock price predictions for select stocks like BBCA, BMRI, and AMRT. Investors can leverage these models to make more informed decisions, particularly when focusing on a smaller group of stocks. This ability to predict stock price trends with higher accuracy can lead to better investment strategies, risk management, and portfolio optimization. Investors who focus on high-potential or stable stocks, such as the ones used in this study, can benefit significantly from these predictive tools.
  - ∗ Customization of Forecasting Models: The findings demonstrate that different models perform better for specific stock groups, indicating that investors should customize their forecasting models based on the characteristics of their target investments. For instance, multi time series models are more effective for smaller, highly correlated groups of stocks, which can be particularly useful for investors aiming to build specialized portfolios in sectors like finance or retail. In contrast, broader stock groups introduce more variability, making them harder to predict, suggesting that investors may need to apply different strategies or models when dealing with more diverse or volatile assets.
  - ∗ Enhanced Investment in Stable Markets: The study highlights that the combined dataset of BBCA, BMRI, and AMRT, representing key players in stable sectors, yielded the best predictive results. This implies that investors focusing on stable or established stocks in the Indonesian market could achieve more consistent returns by utilizing advanced forecasting models like Lag-Llama. Predicting price movements in well-established stocks can help investors minimize risks and capitalize on predictable trends, making these models highly valuable for long-term and low-risk investment strategies.

- **For Policymakers**:
  * Market Stability and Predictability: The ability of these models to accurately predict stock prices, particularly for stable and influential stocks like BBCA, BMRI, and AMRT, suggests that policymakers can use such models to assess market stability and identify early warning signs of volatility or unexpected market shifts. Accurate predictive models could support regulatory bodies in monitoring market dynamics, ensuring smoother operation of stock exchanges, and implementing timely interventions to stabilize markets when needed.
  * Data-Driven Policy Formulation: Predictive models such as Lag-Llama, especially when combined with multi time series forecasting, provide deep insights into market behavior. Policymakers can use these tools to understand how different sectors interact, assess market trends, and craft policies to encourage economic stability. For instance, by observing price patterns in key industries like banking (BBCA, BMRI) and retail (AMRT), policymakers can make informed decisions about sector-specific regulations, taxation, and incentives to support sustainable economic growth.
  * Support for Long-Term Market Development: With the demonstrated accuracy in predicting the behavior of large, well-established stocks, policymakers can focus on creating a favorable environment for investors in these sectors. By encouraging investment in stable stocks with predictable trends, they can promote long-term market development and investor confidence. In addition, these models could guide the formation of regulations that protect investors from extreme market volatility, thereby fostering a more resilient and predictable stock market.

5.5. Limitations

While the study presents promising results, several limitations should be noted:

- **Focus on a Limited Set of Stocks**: While this study provides valuable insights into the forecasting capabilities of the Lag-Llama model, the analysis focuses on a specific subset of the Indonesian stock market, namely BBCA, BMRI, AMRT, and the top 18 stocks by market capitalization. This limits the generalizability of the findings to other sectors, smaller-cap stocks, or markets outside of Indonesia. As a result, the model's performance on less liquid or highly volatile stocks remains unexplored.
- **Model Complexity and Computational Resources**: The fine-tuning of the Lag-Llama model and the use of sophisticated architectures such as Temporal Fusion Transformer (TFT) and DeepAR require substantial computational resources and time. This complexity may present a barrier for retail investors or smaller institutions that do not have access to the required infrastructure. Additionally, the fine-tuning process might not be feasible for real-time predictions in fast-moving markets, limiting the model's utility in highly dynamic trading environments.
- **Lack of Macroeconomic and External Data**: The current analysis does not incorporate external macroeconomic indicators or global market influences that could affect stock prices. Factors such as interest rates, inflation, foreign exchange rates, or geopolitical events, which can play a significant role in stock market movements, were not considered. This could limit the model's ability to accurately predict sudden market shifts driven by external factors not present in the historical stock price data.

5.6. Future Research

Future research directions include:

- **Expansion to a Broader Set of Stocks and Markets**: Future research should expand the analysis to a wider set of stocks, including small- and mid-cap stocks, and potentially other emerging markets beyond Indonesia. This will help evaluate the generalizability of the Lag-Llama model and other forecasting models across different sectors, industries, and economic environments. Investigating model performance in more volatile or less liquid stocks could uncover further insights into its robustness.
- **Incorporating External and Macro-level Data**: Integrating macroeconomic variables, such as interest rates, GDP growth, inflation, or oil prices, into the forecasting models could enhance the model's ability to predict stock movements driven by external shocks or economic changes. Additionally, incorporating sentiment analysis from news or social media could provide a more holistic view of market behavior. This could lead to improved predictive performance, especially during periods of heightened uncertainty.
- **Real-Time and High-Frequency Predictions**: Another area of future research could focus on adapting the models for real-time or high-frequency forecasting. By developing more efficient versions of these models that can process large volumes of data in real-time, researchers could explore how these models perform in fast-paced trading environments. This would also include evaluating the trade-offs between model accuracy and computational speed to ensure feasibility in real-world applications.
- **Exploring Hybrid Models**: Combining machine learning models like Lag-Llama with traditional econometric models could lead to hybrid approaches that capitalize on the strengths of both methodologies. Future research could investigate how these hybrid models perform in comparison to standalone models, potentially offering more reliable forecasts with enhanced interpretability.
- **Robustness Testing Under Market Stress Conditions**: A valuable future research direction would be to test the robustness of the models under market stress conditions, such as during economic downturns, financial crises, or pandemics. Understanding how these models perform under extreme volatility or liquidity constraints could help determine their utility in real-world financial decision-making, where risk management during crises is a critical component.
- **Interpretable AI in Finance**: Future research should also focus on improving the interpretability of complex models like Lag-Llama, TFT, and DeepAR in financial contexts. Investors and policymakers alike benefit from models that provide not only accurate forecasts but also clear reasoning behind the predictions. Developing techniques to enhance the transparency and interpretability of these models could lead to wider adoption and trust from users.

6. Conclusion

This study presents the first empirical validation of the Lag-Llama model for probabilistic time series forecasting in the Indonesian stock market, focusing on both univariate and multi-time series forecasting approaches. The results demonstrate that fine-tuning Lag-Llama, particularly in multi-time series scenarios, significantly improves forecasting accuracy, especially for a targeted set of stocks such as BBCA, BMRI, and AMRT. These findings establish Lag-Llama as a promising tool for handling complex temporal patterns and uncertainty in stock price predictions, outperforming traditional models like DeepAR and Temporal Fusion Transformer in specific contexts.

This study not only validates the effectiveness of Lag-Llama in real-world financial applications but also highlights its potential to inform more accurate investment strategies and policymaking. The ability to model uncertainty and capture long-term dependencies makes Lag-Llama a valuable addition to the growing field of probabilistic time series forecasting. Future work should expand its application to broader market segments and incorporate external macroeconomic factors to further enhance its predictive power.

While the fine-tuned and TFT models performed exceptionally well on the combined dataset of BBCA, BMRI, and AMRT, broader market capitalization groups, such as the top 1-9 and top 10-18 stocks, introduced more variability and resulted in higher prediction errors. This indicates that forecasting performance is enhanced when focusing on a smaller, more correlated set of stocks, as opposed to attempting to generalize across larger and more diverse groups.

The implications of these findings are significant for investors and policymakers. Investors can leverage these predictive models to make more informed and precise decisions, particularly when focusing on stable or well-established stocks. Policymakers, on the other hand, can utilize these tools to monitor market stability and formulate data-driven policies aimed at fostering a stable financial environment.

However, this study also highlights several limitations, including the narrow focus on a specific set of stocks and the need for computational resources to fine-tune the models. Future research should expand the scope to include a broader range of stocks, incorporate macroeconomic factors, and explore the application of hybrid models to further enhance prediction accuracy and robustness.

In conclusion, the Lag-Llama model, especially when fine-tuned, shows great promise in improving stock price forecasting accuracy in the Indonesian market. With further refinements and broader applications, these advanced models have the potential to become powerful tools for both investors and policymakers in managing risks and making informed financial decisions.

7. Acknowledgment

This research was funded by Universitas Islam Riau under Joint Research Project UTP-UMP-TelU-UIR (1058/KONTRAK/DPPM-UIR-12-2022).

References

[1] Z. Alameer, A. Fathalla, K. Li, H. Ye and Z. Jianhua, Multistep-ahead forecasting of coal prices using a hybrid deep learning model, *Resources Policy* **65** (2020), 101588.

[2] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D.C. Maddix, S. Rangapuram, D. Salinas, J. Schulz et al., Gluonts: Probabilistic and neural time series modeling in python, *Journal of Machine Learning Research* **21**(116) (2020), 1–6.

[3] K.A. Althelaya, E.-S.M. El-Alfy and S. Mohammed, Evaluation of bidirectional LSTM for short-and long-term stock market prediction, in: *2018 9th international conference on information and communication systems (ICICS)*, IEEE, 2018, pp. 151–156.

[4] A.A. Ariyo, A.O. Adewumi and C.K. Ayo, Stock price prediction using the ARIMA model, in: *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, IEEE, 2014, pp. 106–112.

[5] V. Azevedo, G.S. Kaiser and S. Mueller, Stock market anomalies and machine learning across the globe, *Journal of Asset Management* **24**(5) (2023), 419–441.

[6] K. Bagastio, R. Oetama and A. Ramadhan, Development of stock price prediction system using Flask framework and LSTM algorithm, *Journal of infrastructure, policy and development* (2023). doi:10.24294/jipd.v7i3.2631.

[7] M. Bildirici, N. Guler Bayazit and Y. Ucan, Analyzing crude oil prices under the impact of COVID-19 by using LSTAR-GARCHLSTM, *Energies* **13**(11) (2020), 2980.

[8] A. Boudhaouia and P. Wira, A real-time data analysis platform for short-term water consumption forecasting with machine learning, *Forecasting* **3**(4) (2021), 682–694.
[9] B. Burrichter, J.K. da Silva, A. Niemann and M. Quirmbach, A Temporal Fusion Transformer Model to Forecast Overflow from Sewer Manholes during Pluvial Flash Flood Events, *Hydrology* (2024). doi:10.3390/hydrology11030041.
[10] H. Chunhua, Prediction method of oil production in new wells based on long and short term memory neural network [J], *Oil and Gas Geology and Recovery* **26**(3) (2019), 105–110.
[11] R. Corizzo and J. Rosen, Stock market prediction with time series data and news headlines: a stacking ensemble approach, *Journal of Intelligent Information Systems* **62**(1) (2024), 27–56.
[12] H. Cui, Y. Zhou, J. Li, Z. Chang and Y.-C. Luo, Research and Application Based on Time Series Analysis with Least Squares Regression, 2023. doi:10.1109/icsece58870.2023.10263463.
[13] S. Feng, C. Miao, Z. Zhang and P. Zhao, Latent diffusion transformer for probabilistic time series forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 2024, pp. 11979–11987.
[14] T. Gneiting and A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, *Journal of the American statistical Association* **102**(477) (2007), 359–378.
[15] J. GU, M. Zhou, Z. LI et al., Prediction method of oil well production based on data mining based on long-term and short-term memory network model [J], *Special Oil and Gas Reservoirs* **26**(2) (2019), 77–81.
[16] C. Guo, Q. Ge, H. Jiang, G. Yao and Q. Hua, Maximum power demand prediction using fbprophet with adaptive Kalman filtering, *IEEE Access* **8** (2020), 19236–19247.
[17] P. Haksun Li, Time-Series Analysis, 2022. doi:10.1007/978-1-4842-6797-4$_1$5.
[18] A.T. Haryono, R. Sarno and K.R. Sungkono, Stock price forecasting in Indonesia stock exchange using deep learning: a comparative study, *International Journal of Electrical and Computer Engineering* (2024). doi:10.11591/ijece.v14i1.pp861-869.
[19] G.P. Herrera, M. Constantino, B.M. Tabak, H. Pistori, J.-J. Su and A. Naranpanawa, Data on forecasting energy prices using machine learning, *Data in brief* **25** (2019), 104122.
[20] G.P. Herrera, M. Constantino, B.M. Tabak, H. Pistori, J.-J. Su and A. Naranpanawa, Long-term forecast of energy commodities price using machine learning, *Energy* **179** (2019), 214–221.
[21] D. Indriani, A.H. Nasution, W. Monika and S. Nasution, Towards a Sentiment Analyser for Low-resource Languages, in: *Proceedings of International Conference on Smart Computing and Cyber Security*, P.K. Pattnaik, M. Sain, A.A. Al-Absi and P. Kumar, eds, Springer Singapore, Singapore, 2021, pp. 109–118. ISBN 978-981-15-7990-5.
[22] B.K. Jha and S. Pande, Time series forecasting model for supermarket sales using FB-prophet, in: *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 2021, pp. 547–554.
[23] B. Jia, H. Wu and K. Guo, Chaos theory meets deep learning: A new approach to time series forecasting, 2024. doi:10.1016/j.eswa.2024.124533.
[24] A. Josey and A. N, Stock Market Prediction, *Indian Journal of Data Mining (IJDM)* (2024). doi:10.54105/ijdm.a1641.04010524.
[25] E.A. Kadir, H.T. Kung, A.H. Nasution, H. Daud, A.A. AlMansour, M. Othman and S.L. Rosa, Fires Hotspot Forecasting in Indonesia Using Long Short-Term Memory Algorithm and MODIS Datasets, in: *Vegetation Fires and Pollution in Asia*, Springer, 2023, pp. 589–602.
[26] S. Karasu, A. Altan, S. Bekiros and W. Ahmad, A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series, *Energy* **212** (2020), 118750.
[27] R. Kumar, P. Kumar and Y. Kumar, Multi-step time series analysis and forecasting strategy using ARIMA and evolutionary algorithms, *International Journal of Information Technology* **14**(1) (2022), 359–373.
[28] V. Kumar and A. Haider, A survey of deep learning techniques for time-series forecasting, 2024. doi:10.55524/csistw.2024.12.1.72.
[29] D. Lande, V.V. Yuzefovych and Y. Tsybulska, Linguistic Approach to Time Series Forecasting, *arXiv.org* (2022). doi:10.48550/arXiv.2207.00985.
[30] B. Lim, S. Arık, N. Loeff and T. Pfister, Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, *International Journal of Forecasting* **37**(4) (2021), 1748–1764. doi:https://doi.org/10.1016/j.ijforecast.2021.03.012. https://www.sciencedirect.com/science/article/pii/S0169207021000637.
[31] Y. Liu, S. Wijewickrema, A. Li, C. Bester, S. O'Leary and J. Bailey, Time-Transformer: Integrating Local and Global Features for Better Time Series Generation, *arXiv.org* (2023). doi:10.48550/arxiv.2312.11714.
[32] K. Madhusudhanan, S. Jawed and L. Schmidt-Thieme, Hyperparameter Tuning MLP's for Probabilistic Time Series Forecasting, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2024, pp. 264–275.
[33] M.S. Maksar, W.S. Firdani, I. Rabbani, Y. Swastika and R.C. Laksono, The Predictive Ability of U.S. Stock Market Skewness on Indonesian Stock Market Returns, 2024. doi:10.47191/jefms/v7-i5-71.
[34] D.A.I. Maruddani and T. Trimono, Implementation of stochastic model for risk assessment on indonesian stock exchange, *Media statistika* (2023). doi:10.14710/medstat.15.2.151-162.

[35] J.A. Miller, M. Aldosari, F. Saeed, N.H. Barna, S. Rana, I. Arpinar and N. Liu, A Survey of Deep Learning and Foundation Models for Time Series Forecasting, *arXiv.org* (2024).

[36] A.H. Nasution and A. Onan, ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks, *IEEE Access* **12** (2024), 71876–71900. doi:10.1109/ACCESS.2024.3402809.

[37] K. Olorunnimbe and H. Viktor, Deep learning in the stock market—a systematic survey of practice, backtesting, and applications, *Artificial Intelligence Review* **56**(3) (2023), 2057–2109.

[38] E.h. Ouassou and H. Taya, Forecasting regional tourism demand in morocco from traditional and AI-based methods to ensemble modeling, *Forecasting* **4**(2) (2022), 420–437.

[39] S.H. Park, B. Kim, C.M. Kang, C.C. Chung and J.W. Choi, Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture, in: *2018 IEEE intelligent vehicles symposium (IV)*, IEEE, 2018, pp. 1672–1678.

[40] J. Peng, Z. Li and B.M. Drakeford, Dynamic characteristics of crude oil price fluctuation—from the perspective of crude oil price influence mechanism, *Energies* **13**(17) (2020), 4465.

[41] X. Peng, Y. Lin, Y. Chen, P. Fan and Z. Lin, TCF-Trans: Temporal Context Fusion Transformer for Anomaly Detection in Time Series (2023). doi:10.3390/s23208508.

[42] K. Rasul, A. Ashok, A.R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M.J.D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen et al., Lag-llama: Towards foundation models for probabilistic time series forecasting, *Preprint* (2024).

[43] K. Resiandi, Y. Murakami and A.H. Nasution, Neural Network-Based Bilingual Lexicon Induction for Indonesian Ethnic Languages, *Applied Sciences* **13**(15) (2023). doi:10.3390/app13158666. https://www.mdpi.com/2076-3417/13/15/8666.

[44] M. Roondiwala, H. Patel and S. Varma, Predicting stock prices using LSTM, *International Journal of Science and Research (IJSR)* **6**(4) (2017), 1754–1756.

[45] N.R. Sabri, The Reliability of Prediction Factors, for the World Stock Markets, *Theoretical Economics Letters* (2021). doi:10.4236/TEL.2021.113030.

[46] D. Salinas, V. Flunkert, J. Gasthaus and T. Januschowski, DeepAR: Probabilistic forecasting with autoregressive recurrent networks, *International journal of forecasting* **36**(3) (2020), 1181–1191.

[47] G.C. Santos, F. Barboza, A.C.P. Veiga and M.F. Silva, Forecasting Brazilian Ethanol Spot Prices Using LSTM, *Energies* **14**(23) (2021), 7987.

[48] P.K. Sarangi, Muskaan, S. Singh and A.K. Sahoo, A Study on Stock Market Forecasting and Machine Learning Models: 1970–2020, 2022. doi:10.1007/978-981-16-1740-9$_4$2.

[49] O. Shchur, A.C. Turkmen, N. Erickson, H. Shen, A. Shirkov, T. Hu and B. Wang, AutoGluon–TimeSeries: AutoML for probabilistic time series forecasting, in: *International Conference on Automated Machine Learning*, PMLR, 2023, pp. 9–1.

[50] A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Physica D: Nonlinear Phenomena* **404** (2020), 132306.

[51] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-c. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, *Advances in neural information processing systems* **28** (2015).

[52] D. Sikka and C.V.V. Kumar, Website Traffic Time Series Forecasting Using Regression Machine Learning, 2023. doi:10.1109/CSNT57126.2023.10134631.

[53] W. Sun and J. Zhang, Carbon price prediction based on ensemble empirical mode decomposition and extreme learning machine optimized by improved bat algorithm considering energy price factors, *Energies* **13**(13) (2020), 3471.

[54] M.A.I. Sunny, M.M.S. Maswood and A.G. Alharbi, Deep learning-based stock price prediction using LSTM and bi-directional LSTM model, in: *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, IEEE, 2020, pp. 87–92.

[55] D. Swain, P.K. Pattnaik and P.K. Gupta, *Machine Learning and Information Processing*, Springer, 2021.

[56] T.H. Sworo and A. Hermawan, Analysis and Prediction of Indonesia Stock Exchange (IDX) Stock Prices Using Long Short Term Memory (LSTM) Algorithm, 2024. doi:10.32996/jcsts.2024.6.2.16.

[57] J. Tang and X. Chen, Stock Market Prediction Based on Historic Prices and News Titles, 2018. doi:10.1145/3231884.3231887.

[58] S. Tarafder, N. Badruddin, N. Yahya and A.H. Nasution, Drowsiness Detection Using Ocular Indices from EEG Signal, *Sensors* **22**(13) (2022). doi:10.3390/s22134764. https://www.mdpi.com/1424-8220/22/13/4764.

[59] P.W. Titisari, A.H. Nasution, Elfis and W. Monika, Toward Crops Prediction in Indonesia, in: *Proceedings of 3rd International Conference on Smart Computing and Cyber Security*, P.K. Pattnaik, M. Sain and A.A. Al-Absi, eds, Springer Nature Singapore, Singapore, 2024, pp. 207–216. ISBN 978-981-97-0573-3.

[60] I. Ursu, Customer Demand Forecast Using Time Series Approach, *Journal of Eastern Europe research in business economics* (2024). doi:10.5171/2024.915300.

[61] E.H. Verbiest, Stock Return Prediction by History Mapping, *Social Science Research Network* (2011). doi:10.2139/SSRN.1963679.

[62] X. Wan, Y. Zou, J. Wang and W. Wang, Prediction of shale oil production based on Prophet algorithm, in: *Journal of Physics: Conference Series*, Vol. 2009, IOP Publishing, 2021, p. 012056.
[63] G. Westergaard, U. Erden, O.A. Mateo, S.M. Lampo, T.C. Akinci and O. Topsakal, Time Series Forecasting Utilizing Automated Machine Learning (AutoML): A Comparative Analysis Study on Diverse Datasets, *Information* (2024). doi:10.3390/info15010039.
[64] K. Yadav, M. Yadav and S. Saini, Stock values predictions using deep learning based hybrid models, *CAAI Transactions on Intelligence Technology* **7**(1) (2022), 107–116.
[65] K. Yemets, Time series forecasting model for solving cold start problem via temporal fusion transformer, *Kompûternì sistemi ta ìnformacìjnì tehnologìï* (2024). doi:10.31891/csit-2024-1-7.
[66] Y. Yu, X. Si, C. Hu and J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, *Neural computation* **31**(7) (2019), 1235–1270.
[67] A.E. Zaar, A. Mansouri, N. Benaya, A.E. Allati and T. Bakir, A Contribution to Time Series Analysis and Forecasting Using Deep Learning Approaches, 2024. doi:10.1109/iccad60883.2024.10553976.
[68] Z. Zheng, Y. Yang, J. Zhou and F. Gu, Research on Time Series Data Prediction Based on Machine Learning Algorithms, 2024. doi:10.1109/iccect60629.2024.10546215.
[69] E. Zunic, K. Korjenic, K. Hodzic and D. Donko, Application of facebook's prophet algorithm for successful sales forecasting based on real-world data, *arXiv preprint arXiv:2005.07575* (2020).
[70] Linguistic Approach to Time Series Forecasting, 2022. doi:10.48550/arxiv.2207.00985.
[71] Time-Series Analysis, *Advanced technologies and societal change* (2023). doi:10.1007/978-981-99-3157-6$_6$.
[72] Website Traffic Time Series Forecasting Using Regression Machine Learning, 2023. doi:10.1109/csnt57126.2023.10134631.