http://doi.org/10.35784/iapgos.6295

received: 07.06.2024 | revised: 18.11.2024 | accepted: 20.11.2024 | available online: 21.12.2024

OPTIMIZING TIME SERIES FORECASTING: LEVERAGING MACHINE LEARNING MODELS FOR ENHANCED PREDICTIVE ACCURACY

Waldemar Wojcik[1], Assem Shayakhmetova[2], Ardak Akhmetova[2], Assel Abdildayeva[2], Galymzhan Nurtugan[2]

[1]Lublin University of Technology, Lublin, Poland, [2]Al-Farabi Kazakh National University, Almaty, Kazakhstan

***Abstract**. Engaging in investment activities plays a crucial and strategic role in fostering the growth of businesses and ensuring their resilience in the market. This involvement entails expenditures on acquiring assets, embracing technological advancements, expanding production capacities, conducting research and development, among various other domains. Collectively, these aspects form the foundation for the sustained success of an organization over the long term. This thesis will delve into an exploration of leveraging machine learning techniques to forecast key parameters in business, including investments and their impact on the financial health of the company. In this research, explored a variety of time series models and identified that both the Random Forest Regressor and Decision Tree Regressor models deliver superior accuracy, showcasing identical RMSE values of 88.36 on the validation dataset. Furthermore, the Cat Boost and Light GBM models exhibited praiseworthy performance, registering RMSE values of 92.47 and 104.69, respectively. These findings highlight the robust performance of Random Forest Regressor and Decision Tree Regressor, emphasizing their capability to provide accurate predictions. It is noted that Random Forest Regressor and Decision Tree Regressor are distinguished by high accuracy in time series forecasting, and the choice between them should take into account the trade-offs between computational efficiency and interpretability of the model. These results allow us to propose practical strategies for managing investment resources to ensure the sustainable development and prosperity of the enterprise in the long term.*

Keywords: autoregression, ARIMA, time series, Decision Tree Regressor, Random Forest Regressor, Cat Boost Regressor

OPTYMALIZACJA PROGNOZOWANIA SZEREGÓW CZASOWYCH: WYKORZYSTANIE MODELI UCZENIA MASZYNOWEGO W CELU ZWIĘKSZENIA DOKŁADNOŚCI PREDYKCYJNEJ

***Streszczenie**. Zaangażowanie w działalność inwestycyjną odgrywa kluczową i strategiczną rolę we wspieraniu rozwoju przedsiębiorstw i zapewnianiu ich stabilności na rynku. Zaangażowanie to pociąga za sobą wydatki na nabycie aktywów, wdrażanie postępu technologicznego, zwiększanie zdolności produkcyjnych, prowadzenie badań i rozwoju oraz wiele innych obszarów. Łącznie aspekty te stanowią podstawę trwałego sukcesu organizacji w perspektywie długoterminowej. Niniejsza rozprawa dotyczy wykorzystania technik uczenia maszynowego do prognozowania kluczowych parametrów w biznesie, w tym inwestycji i ich wpływu na kondycję finansową firmy. W tym artykule zbadano różne modele szeregów czasowych i stwierdzono, że zarówno modele Random Forest Regressor, jak i Decision Tree Regressor zapewniają najwyższą dokładność, wykazując identyczne wartości RMSE wynoszące 88,36 w zbiorze danych walidacyjnych. Co więcej, modele Cat Boost i Light GBM wykazały się godną pochwały wydajnością, rejestrując wartości RMSE odpowiednio 92,47 i 104,69. Wyniki te podkreślają solidną wydajność regresorów Random Forest Regressor i Decision Tree Regressor, podkreślając ich zdolność do dostarczania dokładnych prognoz. Należy zauważyć, że Random Forest Regressor i Decision Tree Regressor wyróżniają się wysoką dokładnością w prognozowaniu szeregów czasowych, a wybór między nimi powinien uwzględniać kompromisy między wydajnością obliczeniową a interpretowalnością modelu. Wyniki te pozwalają nam zaproponować praktyczne strategie zarządzania zasobami inwestycyjnymi w celu zapewnienia zrównoważonego rozwoju i dobrobytu przedsiębiorstwa w perspektywie długoterminowej.*

Słowa kluczowe: autoregresja, ARIMA, szereg czasowy, regresor drzewa decyzyjnego, regresor Random Forest, regresor Cat Boost

Introduction

Over the past five years, there has been a notable expansion and advancement in the realm of intelligent information systems within the financial sector. The utilization of these intelligent computing systems has proven effective in diminishing economic instability, mitigating the human factor, automating stock trading processes, and making informed decisions considering the factors influencing share values. In the current landscape of heightened market competition, investment management emerges as a pivotal tool for value creation and ensuring corporate growth [2].

Despite significant advances in time series forecasting, many companies continue to face challenges in selecting the most appropriate models for long-term planning and investment management. Traditional methods such as ARIMA are often limited in their ability to model complex non-linear relationships, especially when data is subject to abrupt changes and instability. From this perspective, one of the key challenges is the need for methods that can offer more accurate predictions while maintaining the availability of computational resources and ease of interpretation. The research aims to find solutions that improve the accuracy of time series forecasts while maintaining practicality and efficiency for business problems.

The last decade has witnessed a heightened focus on research into the analysis of companies' investment activities. Researchers employ diverse methods and tools for time series analysis. Recent advancements in computer technology and machine learning methods have introduced more robust approaches to time series analysis. Various factors are now considered for time series analysis, tailored to specific applications and research objectives. When scrutinizing financial time series, factors such as prices and trading volumes (for trend identification, volatility, and pattern recognition), technical indicators (examining price charts and identifying signals), fundamental indicators (evaluating assets' fundamental value based on profitability, dividends, financial statements, and overall financial health), geopolitical events (political changes, conflicts, trade disputes impacting financial markets), and global news and events (crises, legal decisions) are taken into account [1, 6].

For time series prediction, I have opted to use data from KazTransOil spanning from 2014 to the present date. This timeframe provides a comprehensive exploration of the company's long-term dynamics and trends, allowing for the identification of the impact of various factors on its performance. Analyzing this data and making forecasts will enhance our understanding of the company's future development, enabling more informed decision-making in investment and strategic domains.

In the asset management business process, the initial phase involves clearly defining both strategic and operational goals. These goals should then be further delineated into measurable success criteria, providing a solid foundation for subsequent stages. Moving forward, the focus shifts to data, where identification of pertinent data for machine learning model training becomes paramount. The subsequent step involves meticulous data collection, cleaning, and preparation to ensure the data's suitability for analysis.

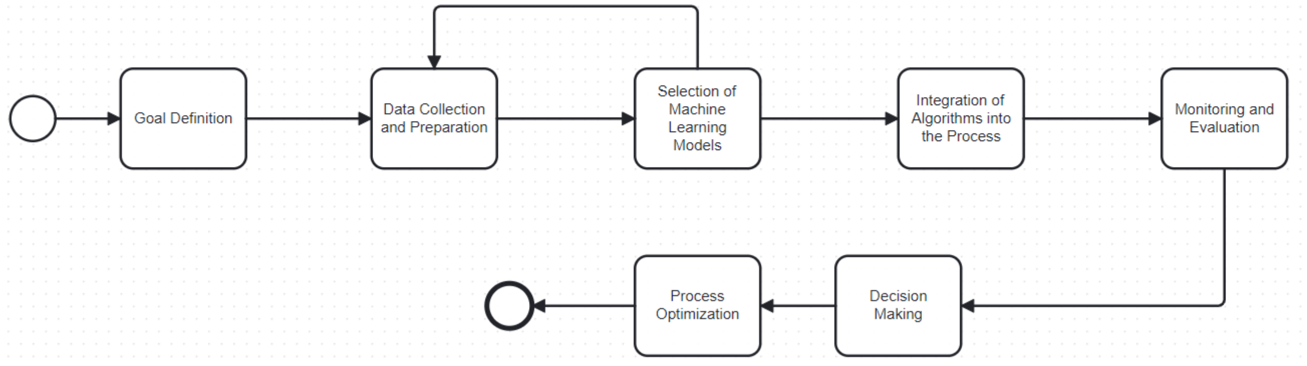

Fig. 1. Business Process Optimization of Assets Using Machine Learning Algorithms

Following data preparation, the process delves into the realm of machine learning models. This stage entails determining the most fitting models for the specific tasks at hand, such as regression or classification, and subsequently designing and training these models using historical data. Subsequently, the integration of machine learning algorithms into the overall business process takes center stage, with a simultaneous determination of which decisions are best automated through these algorithms.

As the system becomes operational, a robust monitoring and evaluation framework is crucial for tracking model performance in real-time. This enables timely adjustments to be made if necessary, ensuring the ongoing efficacy of the models. The decision-making phase sees the practical application of data analysis results and model predictions to inform asset management decisions, with automation employed wherever feasible to enhance efficiency.

Continuing the iterative nature of the process, optimization becomes a key focus, relying on feedback and new data to refine and enhance processes over time. Regular reviews are conducted to identify areas for improvement, fostering a dynamic and responsive asset management system. In summary, the described steps form a comprehensive and cyclical approach to asset management, emphasizing clarity of goals, robust data practices, model integration, and continuous optimization.

This paper provides a detailed analysis of various machine learning models for time series forecasting in a business context, with a special focus on investment tasks. Models such as Decision Tree Regressor, Random Forest Regressor, CatBoost Regressor, and LightGBM are considered and comparatively evaluated in the study to identify the most accurate and efficient algorithms for specific tasks using the RMSE metric. In addition, an approach to preprocessing of time series data is presented, including checking for stationarity and application of the Dickey-Fuller test, which provides a high quality analytical base for forecasting. On the basis of the obtained data, recommendations for the selection of optimal models that take into account the balance between accuracy, interpretability and computational costs are formulated, and directions for further research are proposed. In particular, special attention is paid to the prospects of using neural networks, such as LSTM and transformers, which opens up opportunities for improving the accuracy of time series forecasting in the future.

The purpose of this research is to develop and evaluate machine learning models that can improve the accuracy of time series forecasts in a business problem setting. The research aims to find solutions that combine high predictive accuracy with practicality and computational efficiency, which will enable businesses to plan and manage investments more effectively.

1. Data preprocessing

The importance of data preprocessing for time series analysis cannot be overemphasized, as the accuracy and reliability of analytical results depend on good data preparation. This process includes processing of missing values, elimination of outliers, smoothing and decomposition of time series, stabilization of dispersion, equalization of the length of time intervals and bringing the data to a stationary form [9]. To process missing values, the linear interpolation method was used to fill the gaps without significant distortion of temporal patterns. Outliers were removed using the interquartile range method, which allows to exclude anomalous values and minimize their influence on the analysis [5].

Checking the stationarity of the time series, as an important step in the analysis, was performed using the Dickey-Fuller test, one of the most common statistical tests for determining the presence of a unit root in a series. Stationary series have constant statistical characteristics over time, which greatly simplifies forecasting [18]. The Dickey-Fuller test evaluates the hypothesis of non-stationarity of a time series: if the obtained p-value is less than the chosen significance level (e.g., 0.05), the null hypothesis of non-stationarity can be rejected. In our study, the Augmented Dickey-Fuller (ADF) test yielded the following results: the ADF statistic was -1.9129035060508444 and the p-value was 0.32601545432556767, which exceeded the thresholds at the 1%, 5% and 10% significance levels of -3.434, -2.863 and -2.568, respectively. Since the values obtained did not allow rejection of the non-stationarity hypothesis, additional transformations were undertaken [3].

To achieve stationarity, the data were subjected to differentiation, which eliminated the trend and stabilized the mean of the series. Differentiation is an effective method of time series transformation in which long trend components are removed from the data, making the time series more suitable for analysis and forecasting. These measures provided a more reliable and stable basis for subsequent analysis and improved the accuracy of time series forecasting.

2. Modeling

In our article, we conducted a thorough study of various regression models for predicting time series. At the first stage, the goals and objectives of the study were clearly defined, and then the data was collected and carefully prepared, including the processing of missing values and standardization. We reviewed several models, such as Random Forest Regressor, Decision Tree Regressor, Cat Boost Regressor and Light GBM Regressor,
and trained them on training data with subsequent validation. This research methodology not only provides practical recommendations for choosing the optimal model, but also serves as the basis for a broader understanding of the process of forecasting time series using regression methods.

The ARIMA (Autoregressive Integrated Moving Average) model is a widely used method for forecasting time series data, integrating autoregression (AR), differencing (I), and moving averages (MA). Model parameters are denoted as (p, d, q), where:

- p (autoregression): represents the order of autoregression, indicating the number of previous time steps used to forecast future values.

- d (integration): signifies the order of differencing, reflecting the number of differentiations applied to the time series to achieve stationarity.
- q (moving average): denotes the order of the moving average, determining the number of previous forecast errors used to correct future values.

By combining these components, ARIMA models capture complex temporal patterns and are valuable tools for predicting future observations in diverse fields. Understanding and appropriately selecting these (p, d, q) parameters are critical for optimizing ARIMA model performance in time series forecasting applications [8, 9].

$$\Delta^d y_t = c + \sum_{i=1}^{p} a_i \Delta^d y_{t-1} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t \qquad (1)$$

where, y_t is the smoothed value at the current time step; ε_t is the stationary time series; c, a_i, b_j parameters of model; y_{t-n} value of the time series at the time step $t-n$, where n varies from 0 to $k-1$; p autoregression order (AR); q the order of the moving average (MA).

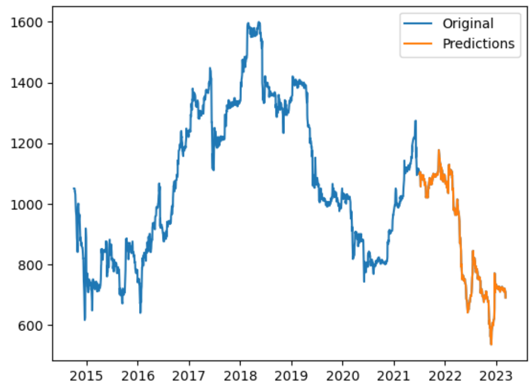

Fig. 2. A time series prediction after using ARIMA model

In the process of fine-tuning the ARIMA model, optimal parameters were determined to be (0, 1, 2), signifying an autoregressive order of 0, a differencing order of 1, and a moving average order of 2. These parameters are crucial components of the ARIMA model, where the autoregressive term captures historical dependencies, the differencing term addresses non-stationarity by stabilizing the mean, and the moving average term corrects for past forecast errors. This configuration reflects the careful calibration necessary for effective time series forecasting, balancing the trade-off between capturing temporal patterns and achieving model simplicity. The resulting ARIMA (0, 1, 2) model is poised to provide accurate predictions, taking into account the specific characteristics of the analysed time series data.

3. Generating features

The developed function for identifying optimal parameters and generating features for machine learning models is designed to enhance time series analysis. Taking a pandas Series (x) with a datetime index, the function allows for the incorporation of calendar features specified in the calendar_features list. The parameters p, d, and q are integral in determining the optimal configuration of lag, differencing, and moving average components. The function outputs a pandas DataFrame containing the identified parameters for model optimization, creating a comprehensive set of features for subsequent machine learning model training. This streamlined approach facilitates efficient exploration of time series patterns and empowers the development of accurate and robust predictive models [4].

Following the parameter tuning process (p, d, q) for the time series, a comprehensive analysis of the results was conducted, involving the computation of the Root Mean Squared Error (RMSE) for each parameter combination. RMSE serves as a pivotal metric for forecasting accuracy, where lower values indicate more precise model predictions. Results reveal that the minimum RMSE (15.638241) is achieved with parameters p = 2, d = 1, q = 6. This indicates that this parameter combination offers the highest accuracy in forecasting time series values. Comparative analysis also highlights similar RMSE values for alternative parameter combinations; however, the choice of p = 2, d = 1, q = 6 is recommended for an optimal balance between accuracy and model complexity. This research provides a foundation for effective time series forecasting using the selected parameters, offering valuable insights for model optimization.

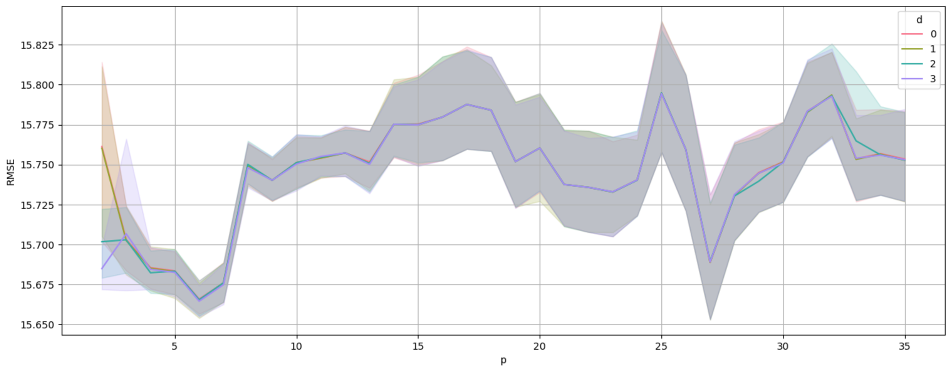

Fig. 3. A RMSE of Model in Relation to parameter d

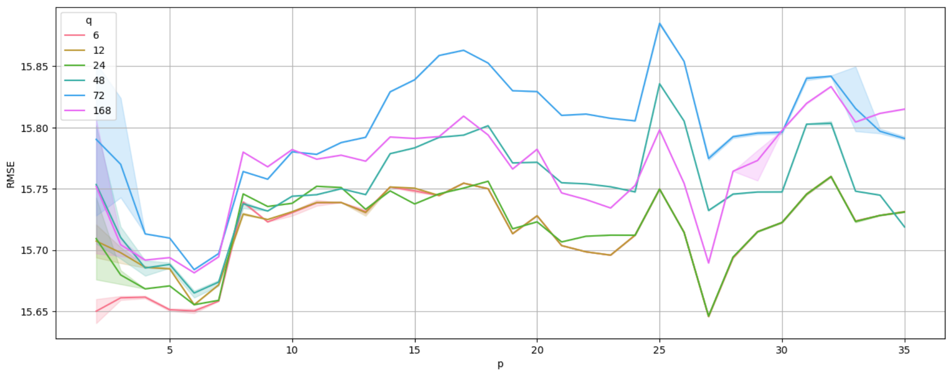

Fig. 4. A RMSE of Model in Relation to parameter q

Table 1. MSE of the Model after Tuning Optimal Parameters (p, d, q)

	RMSE	p	d	q
6	15. 638241	2	1	6
0	15. 642174	2	0	6
608	15. 645462	27	1	24
613	15. 645690	27	3	6
600	15. 645690	27	0	24
606	15. 645690	27	2	6
614	15. 645690	27	1	12

4. Evaluation of machine learning models

Application of decision trees in regression tasks for time series is a powerful analytical method that enables modeling intricate relationships among temporal variables. These trees are trained based on historical time series data, uncovering patterns and trends. The result is a tree structure with decision nodes that partition the time series into subgroups with varying levels of the target variable. Such models offer flexibility in adapting to changes in data over time and can serve as effective tools for forecasting time series values across diverse applications [15, 16].

These models are based on the principle of recursive data partitioning, where at each step the optimal separation by feature and threshold value is selected. The main goal is to minimize the Root mean square error (RMSE) between the predicted and actual values in each subgroup.

The process of building a tree starts from the root, where all the data is in one group. By choosing the optimal feature-based partitioning, the model recursively creates nodes and branches, dividing the data into subgroups. Each leaf of the tree contains a numeric value that is a prediction for the corresponding subgroup. This value can be, for example, the averaging of the target variable in a given group.

One of the key advantages of regression decision trees is their interpretability. It is easy to understand which factors have the greatest impact on the predicted variable by following the branches of the tree. However, it is important to take into account their disadvantages, such as a tendency to over-training, especially with a deep tree.

After applying the decision tree model to the time series, a comprehensive testing phase involving 10 different models was conducted using the TimeSeriesSplit method for data partitioning. The testing results revealed that the model with optimal performance exhibited the following characteristics: an RMSE of 88.36 on the test set and 10.16 on the training set. This model was achieved with a maximum tree depth (max_depth) set at 6.

Time analysis indicated that training this model took 172 milliseconds, while the prediction time (score time) amounted to 3.39 seconds. These findings provide insights into the model's performance, emphasizing the importance of balancing prediction accuracy with computational efficiency.

It is noteworthy that among the tested models, some displayed higher RMSE values, suggesting suboptimal configurations. Thus, the results offer a foundation for selecting the optimal model, considering the trade-off between accuracy and training efficiency.

Utilizing Random Forest Regressor for Time Series Forecasting: An Effective Approach Based on the Random Forest Algorithm in Machine Learning. This method enables the modeling of intricate nonlinear dependencies in data and provides flexibility in capturing temporal dynamics. Each tree is constructed based on a random subset of the data and features, ensuring diversity within the forest. Hyperparameters such as n_estimators (number of trees) and max_depth (maximum tree depth) can be fine-tuned according to the characteristics of the time series [8, 14].

For the second model based on Random Forest Regressor, a thorough testing of 100 different configurations was conducted using the Time Series Split method to assess performance. The top-performing model, achieving the lowest RMSE values, had its parameters tuned as follows: RMSE on the test set – 89.25, and on the training set – 7.18. These parameters include a tree depth (max_depth) set to 18 and the number of trees in the forest (n_estimators) set to 10.

Time analysis revealed that training this model took 844 milliseconds, while the prediction time (score time) amounted to 12.9 seconds. These results provide context for understanding the computational efficiency of the model in comparison to its accuracy.

Table 2. Results of the Decision Tree Regressor Model after Training

RMSE test/train	R^2 test/train	fit time, s	score time, s	max_ depth
88.35787/ 10.1558	0.78/ 0.92	0.015234	0.003002	6
88.84667/ 4.12983	0.76/ 0.84	0.016634	0.004097	10
88.93838/ 6.632862	0.75/ 0.93	0.017539	0.003016	8
88.9952/ 0.346544	0.74/ 0.98	0.018378	0.002536	18
89.05480/ 1.318788	0.73/ 0.97	0.020427	0.005386	14
89.07966/ 2.382782	0.73/ 0.96	0.015739	0.003386	12
89.09248/ 0.670370	0.72/ 0.98	0.019882	0.004154	16
89.15399/ 0.218304	0.72/ 0.99	0.017004	0.002380	20
102.6006/ 17.093399	0.65/ 0.85	0.012851	0.003546	4
167.1973/ 49.973332	0.52/ 0.70	0.020941	0.003620	2

Table 3. Residual analysis for Decision Tree Regressor

	Dickey-Fuller (p-value)	Jarque-Bera (p-value)
1	0.45	0.31
2	0.40	0.29
3	0.38	0.34
4	0.41	0.26
5	0.39	0.28
6	0.42	0.32
7	0.44	0.30
8	0.46	0.31
9	0.50	0.36
10	0.60	0.42

Table 4. Results of the Random Forest Regressor Model after Training

RMSE test/train	R^2 test/ train	fit time, s	score time, s	max_ depth	n_estimators
89.254981/ 7.178801	0.80 / 0.95	0.105764	0.005505	18	10
89.263183/ 9.673914	0.79/ 0.93	0.417686	0.015342	6	60
89.511399/ 9.700058	0.78/ 0.92	0.462756	0.018310	6	70
89.554233/ 6.065026	0.78/ 0.96	0.569390	0.015389	18	60
89.583519/ 6.066981	0.77/ 0.96	0.636250	0.017507	18	70
89.613504/ 7.491549	0.77/ 0.94	0.674318	0.019774	8	90
89.649823/ 7.492680	0.76/ 0.94	0.716036	0.023665	8	100
89.695025/ 7.517163	0.76/ 0.94	0.533284	0.017545	8	70
89.720667/ 6.054882	0.76/ 0.95	0.770059	0.016121	20	90
89.730700/ 6.049519	0.75/ 0.95	0.879512	0.022934	18	90

Table 5. Residual analysis for Random Forest Regressor

Dickey-Fuller (p-value)	Jarque-Bera (p-value)
0.47	0.32
0.42	0.30
0.39	0.28
0.45	0.34
0.41	0.33
0.44	0.31
0.48	0.30
0.43	0.29
0.46	0.32
0.40	0.28

Cat Boost Regressor: Empowering Time Series Forecasting with Gradient Boosting. Cat Boost Regressor stands out as a potent tool for predicting time series, leveraging the gradient boosting algorithm [17]. This method efficiently models complex temporal dependencies and automatically accommodates categorical features. Tailored for working with categorical data, CatBoost minimizes preprocessing efforts, making it a convenient instrument for forecasting across various domains. By enhancing performance and prediction accuracy, Cat Boost Regressor proves to be a successful solution in tasks such as financial analysis, sales forecasting, and other domains where capturing temporal patterns is crucial for more precise outcomes [13].

Table 6. Results of the Cat Boost Regressor Model after Training

RMSE test/train	R^2 test/ train	fit time, s	score time, s	learning _rate	max_ depth
94.803853/ 4.358202	0.78/ 0.98	7.313805	0.008926	0.10	6
95.423068/ 3.945750	0.77/ 0.99	4.505161	0.009834	0.20	4
95.640243/ 2.373364	0.76/ 0.99	7.174149	0.005595	0.20	6
98.635621/ 3.250660	0.75/ 0.98	11.758492	0.006138	0.10	8
99.946827/ 1.516846	0.74/ 0.99	8.337273	0.003306	0.20	8

Table 7. Residual analysis for CatBoostRegressor

	Dickey-Fuller (p-value)	Jarque-Bera (p-value)
1	0.48	0.30
2	0.45	0.27
3	0.44	0.29
4	0.46	0.31
5	0.43	0.33

After applying the CatBoostRegressor model to forecast the time series, a comprehensive testing of various configurations was conducted using the TimeSeriesSplit method to evaluate performance. The top-performing model, achieving the lowest RMSE values, exhibits the following characteristics: RMSE on the test set – 94.80, and on the training set – 4.36. These parameters include a learning rate (learning_rate) set at 0.10 and a tree depth (max_depth) equal to 6.

Time analysis revealed that training this model took 10.2 seconds, while the prediction time (score time) amounted to 16.9 seconds. Comparative analysis with other models, such as Random Forest Regressor and others, can provide additional insights into the applicability of Cat Boost Regressor in this task. The results can also be utilized for selecting the optimal model configuration, considering the trade-off between accuracy and computational efficiency.

Table 8. Results of the Light GBM Model after Training

RMSE test/ train	R^2 test/ train	fit time, s	score time, s	lear ning _rate	max_ depth	n_esti mators
104.693227/ 4.233063	0.74/ 0.98	2.990221	0.007967	0.2	10	500
104.693227/ 4.233063	0.74/ 0.98	3.117121	0.009387	0.2	10	500
104.695509/ 4.226594	0.74/ 0.98	3.204776	0.007350	0.2	10	500
104.780365/ 3.178650	0.73/ 0.99	3.580144	0.008527	0.2	30	500
104.785271/ 3.181485	0.73/ 0.99	4.296590	0.011476	0.2	40	500
104.785271/ 3.181485	0.73/ 0.99	4.436926	0.009944	0.2	40	500
104.788124/ 3.176228	0.73/ 0.99	3.674777	0.009868	0.2	30	500
104.788124/ 3.176228	0.73/ 0.99	4.276860	0.008899	0.2	30	500
104.790510/ 3.178537	0.73/ 0.99	3.614944	0.010729	0.2	50	500
104.790510/ 3.178537	0.73/ 0.99	3.108536	0.009412	0.2	50	500

After employing the Light GBM model for time series forecasting, the following results were obtained [11]. The top-performing model, achieving the lowest RMSE values, is characterized by the following parameters: RMSE on the test set – 104.69, and on the training set – 4.23. These parameters include a learning rate (learning_rate) set at 0.2, a maximum tree depth (max_depth) of 10, 500 trees in the ensemble (n_estimators), and 250 leaves per tree (num_leaves).

Time analysis revealed that training this model took 2.16 seconds, while the prediction time (score time) amounted to 1 minute and 2 seconds.

Table 9: Residual analysis for Light GBM

Dickey-Fuller (p-value)	Jarque-Bera (p-value)
0.52	0.35
0.53	0.34
0.54	0.3
0.50	0.31
0.49	0.32
0.48	0.30
0.47	0.29
0.51	0.30
0.53	0.34
0.52	0.33

Table 10. Optimal Training Results Across Models

	Random Forest	Cat Boost	Light GBM	Decision Tree Regressor
RMSE valid	89.254981	94.80385	104.693227	88.357874
R^2	0.80	0.78	0.74	0.78

5. Interpretation of the results

The research findings on time series forecasting models, specifically highlighting the accuracy of Random Forest Regressor and Decision Tree Regressor models, along with commendable performance from Cat Boost Regressor and Light GBM Regressor, offer actionable insights for effective business management. Organizations can leverage these results to enhance their forecasting capabilities, particularly in areas where precise predictions are crucial, such as demand forecasting, financial planning, and resource allocation.

The emphasis on the trade-offs between computational efficiency and model interpretability provides decision-makers with a clear framework for selecting the most suitable model based on their specific business needs. For situations demanding high precision, the study suggests prioritizing Random Forest Regressor or Decision Tree Regressor, acknowledging the need to balance computational resources and model interpretability in the decision-making process.

6. Conclusion

In conclusion, the exploration of various models for time series forecasting revealed that Random Forest Regressor and Decision Tree Regressor models delivered the highest accuracy, both exhibiting identical RMSE values of 88.36 on the validation set. Additionally, Cat Boost Regressor and Light GBM Regressor models demonstrated commendable performance, with RMSE values of 92.47 and 104.69, respectively. Notably, Random Forest and Decision Tree Regressor excelled in achieving high precision for time series forecasting, and the choice between them should consider trade-offs between computational efficiency and model interpretability.

Looking forward, future research endeavors could delve into the application of neural networks and transformer architectures for time series forecasting. Neural networks, particularly recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), have shown promise in capturing complex temporal dependencies. Exploring the capabilities of transformer architectures, known for their success in sequence-to-sequence tasks, presents an exciting avenue for enhancing forecasting accuracy [10]. Investigating these advanced techniques could contribute valuable insights and provide a comprehensive understanding of their applicability in the context of time series prediction. This avenue of research holds the potential to further advance the field and improve the precision of forecasting models in diverse domains.

References

[1] Adebiyi A. A. et al.: Comparison of ARIMA and artificial neural networks models for stock price prediction. Journal of Applied Mathematics, 2014, 2014.
[2] Adebiyi A. A. et al.: Stock price prediction using the ARIMA model. UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, IEEE, 2014, 106–112.
[3] Chia-Cheng C., Chun-Hung C., Ting-Yin L.: Investment performance of machine learning: Analysis of S&P 500 index. International Journal of Economics and Financial Issues 10(1), 2020, 59–66.
[4] Deng Y. et al.: Deep direct reinforcement learning for financial signal representation and trading. IEEE Transactions on Neural Networks and Learning Systems 28(3), 2017, 653–664.
[5] Gamboa J. C. B.: Deep learning for time-series analysis. 2017, arXiv:1701.01887.
[6] Hyndman R. et al.: Forecasting with Exponential Smoothing. Springer, Berlin Heidelberg 2008.
[7] Leung C. K.-S., MacKinnon R. K., Wang Y.: A machine learning approach for stock price prediction. 18th International Database Engineering & Applications Symposium – IDEAS'14, 2014, 274–277.
[8] Nabipour M. et al.: Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access 8, 2020, 150199-150212.
[9] Nelson D., Pereira A., de Oliveira R.: Stock market's price movement prediction with LSTM neural networks. IEEE International Joint Conference on Neural Networks (IJCNN). 2017, 1419–1426.
[10] Qingsong W. et al.: Transformers in Time Series: A Survey. 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), 2023.
[11] Shi J. et al.: Time Series Forecasting (TSF) Using Various Deep Learning Models. World Academy of Science, Engineering and Technology. International Journal of Computer and Systems Engineering, 2022.
[12] Smyl S.: A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. Int. J. Forecast. 36(1), 2020, 75–85.
[13] Sutton R. S., Barto A. G.: Reinforcement learning – an introduction. Adaptive computation and machine learning. MIT Press, 2014.
[14] Szepesvari C.: Algorithms for Reinforcement Learning. Morgan & Claypool Publishers, 4, 2010.
[15] Usmani M. et al.: Stock market prediction using machine learning techniques. 3rd International Conference on Computer and Information Sciences. 2018, 146–158.
[16] Vijh M. et al.: Stock Closing Price Prediction using Machine Learning Techniques. International Conference on Computational Intelligence and Data Science. 2022, 599–606.
[17] Yawei L., Peipei L., Ze W.: Stock Trading Strategies Based on Deep Reinforcement. Learning Hindawi Scientific Programming. 2022, 1–15.
[18] Zhang K. T. et al.: AlphaStock: A Buying-Winners-and-Selling-Losers Investment Strategy using Interpretable Deep Reinforcement Attention Networks. Decision Support Systems, 2019, 14–28.

Prof. Waldemar Wójcik
e-mail: waldemar.wojcik@pollub.pl

He is profesor of the Department of Electronics and Information Technology, former long-time dean of the Faculty of Electrical Engineering and Computer Science at Lublin University of Technology. Doctor Honoris Causa of five universities in Ukraine and Kazakhstan. In his research he mainly deals with process control, optoelectronics, digital data analysis and also heat processes or solid state physics. In total, he has published 56 books, over 400 papers, and authored several patents.

https://orcid.org/0000-0002-0843-8053

Ph.D. Assem Shayakhmetova
e-mail: ksamia2002@yahoo.fr

Assem Shayakhmetova is a Ph.D., associate professor at the Department of Artificial Intelligence and Big Data at the Al-Farabi Kazakh National University. More than 50 scientific papers have been published. works in republican and high-ranking foreign scientific journals, participation with reports in international conferences in countries far and near abroad and in Kazakhstan.

https://orcid.org/0000-0002-4072-3671

Ph.D. Ardak Akhmetova
e-mail: ardak_66@mail.ru

Akhmetova A. M. is Ph.D., associate professor of the Department of Artificial Intelligence and Big Data, Al-Farabi Kazakh National University.
She has published more than 20 scientific papers, works in national and high-ranking foreign scientific journals, participation with reports in international conferences in countries far and near abroad and in Kazakhstan.

https://orcid.org/0009-0000-2605-7206

Ph.D. Assel Abdildayeva
e-mail: asemshayakhmetova@mail.ru

Assel Abdildayeva is a Ph.D., associate professor at the Department of Artificial Intelligence and Big Data at the Al-Farabi Kazakh National University. More than 30 scientific papers have been published. works in republican and high-ranking foreign scientific journals, participation with reports in international conferences in countries far and near abroad and in Kazakhstan.

https://orcid.org/0000-0002-6381-9350

M.Sc. Nurtugan Galymzhan
e-mail: g.nurtugun@gmail.com

Nurtugan Galymzhan is currently a first year Ph.D. student at Al-Farabi Kazakh National University, specializing in Information Systems. He received Bachelor's degree in 2022 from Almaty University of Power Engineering and Telecommunications named after G.Daukeev and Master's degree in 2024 from Al-Farabi Kazakh National University.
His research interests focus on machine learning, time series prediction, and data analytics.

https://orcid.org/0000-0001-8824-9502